Context-Aware Generative Audio Model for Maqam Music

(Rewritten Summary)

Charbel El Haddad

Abstract

This work investigates the construction of a context-aware generative system tailored for Maqam music, emphasizing Oud recordings. Early analysis highlighted the limitations of frequency-based classification and motivated a shift toward diffusion-based generative modeling. A U-Net-driven Denoising Diffusion Probabilistic Model (DDPM) was adopted and later extended through a latent diffusion framework and a tokenized generation strategy to improve coherence over long durations.

Part I: Diffusion-Based Audio Modeling

Phase 1: Maqam Analysis and Audio Extraction

Audio Feature Extraction

The initial phase focused on extracting robust time-frequency representations of the audio signal Y using the Short-Time Fourier Transform (STFT):

$$F_{db} = \text{STFT}(Y) \in \mathbb{R}^{m \times n}.$$

A nonlinear thresholding stage filtered out weak frequencies through $\theta_{\alpha,n}(x)$, followed by smoothing using a Gaussian kernel $\mathcal{N}_\epsilon(\mu, \sigma)$ to obtain a refined spectrogram F'_{mod}.

- **Frequency Isolation:** Threshold filtering retained only dominant components above α dB.
- **Spectral Smoothing:** Convolution with a Gaussian kernel reduced noise and provided a clearer harmonic structure.

Initial Classification and Problem Formulation

Dominant frequencies were examined via the global FFT to infer Maqam characteristics. While informative, this approach was hindered by the Oud's continuous spectrum and sensitivity to micro-timing variations, producing unreliable estimations. Focusing on the opening and closing thirds of the recording helped mitigate modulation-induced inconsistencies.

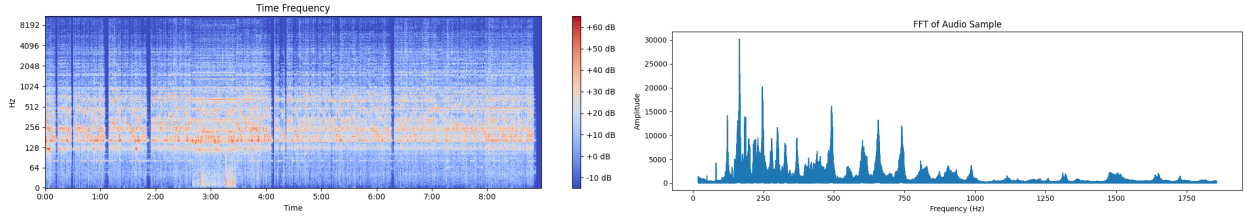

Figure 1: Spectrogram and FFT visualizations highlighting dominant harmonic structures.

Phase 2: Early Generative Approaches

Gaussian Model Experimentation

A PCA-driven Gaussian generative model of the form

$$x = \hat{\mu} + EA^{1/2}z, \quad x \sim \mathcal{N}(\hat{\mu}, \Sigma)$$

was examined. While useful for low-dimensional patterns, it lacked the expressive capability required for long-term musical structure.

Diffusion Model and Context-Aware Design

The study progressed to Denoising Diffusion Probabilistic Models (DDPMs), which offered a more powerful probabilistic framework. A multi-stage pipeline was adopted:

1. Convolutional autoencoder for latent compression.
2. U-Net-based diffusion model operating in latent space.
3. Decoder for audio reconstruction.

To preserve long-range musical coherence, a sliding-window strategy ensured that consecutive segments shared contextual overlap.

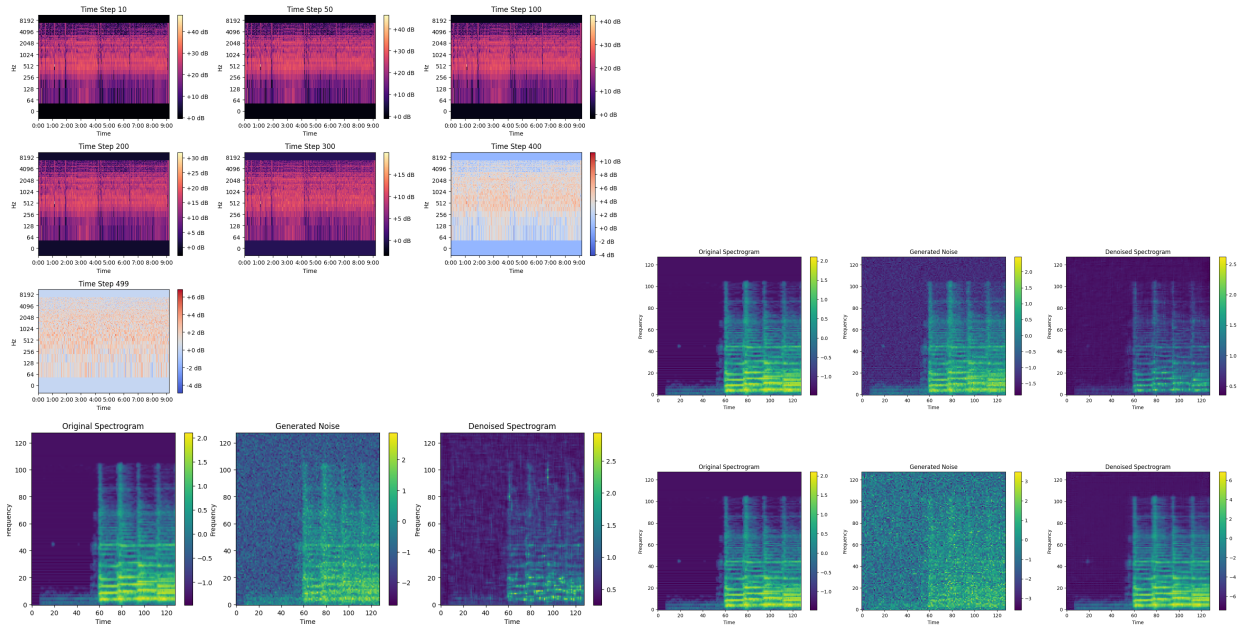

Figure 2: Diffusion training visualizations: noising, denoising, loss behavior, and reconstructed outputs.

Phase 3: U-Net Implementation and Scaling Considerations

Training Stability

A fully convolutional U-Net with channel widths (32, 64, 128, 256, 512) was implemented. Multiple optimization schemes and noise schedules were evaluated using MSE and MAE losses to stabilize convergence.

Scaling Limitations and Middle-Part Hallucination

Models operating on large inputs (> 32k dimensions) consistently failed to reach meaningful convergence. Generated samples often retained coherent openings and closings while producing repetitive or unstable content mid-sequence, a phenomenon referred to as *middle-part hallucination*.

Latent Diffusion and Tokenized Audio Generation

A latent diffusion approach alleviated the scaling bottleneck. A convolutional autoencoder reduced audio samples to latent grids near 128×128, enabling efficient diffusion training.

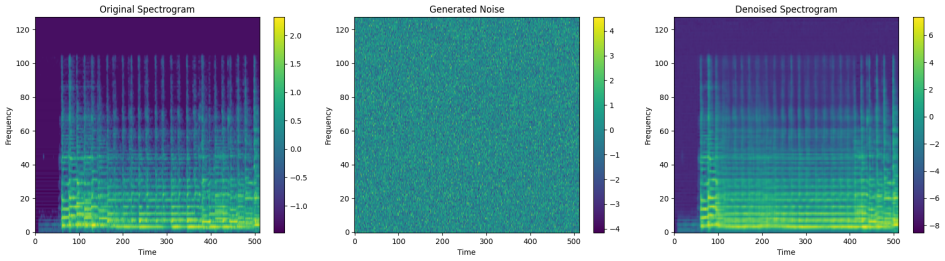

Figure 3: Training behavior and structure of the latent diffusion system.

Short latent segments (e.g., 6-second fragments) were generated independently and stitched together via overlapping windows to maintain continuity across the full sequence.

References

- GitHub: https://github.com/CHCICH/Music_diffusion
- J. Ho et al. *Denoising Diffusion Probabilistic Models*, NeurIPS 2020.
- A. Nichol and P. Dhariwal. *Improved DDPM*, ICML 2021.

Part II: Recurrent Neural Networks (LSTM & RNN Models)

Phase 4: Literature Review and Model Foundations

Background and Motivation

Following the diffusion-based experimentation, attention shifted toward recurrent neural architectures to explore their potential for symbolic and low-dimensional music generation tasks. LSTM networks, with their explicit recurrence and memory gating, offer an alternative perspective on temporal generation, especially when audio can be represented as sequences of discrete musical tokens rather than high-dimensional spectrograms.

A targeted literature review was conducted to identify the most effective sequence-modeling strategies for music:

- **MusicLM (Google Research):** A hierarchical sequence-to-sequence architecture capable of high-fidelity music generation was examined for its tokenization strategies and multi-level modeling capabilities.
- **Transformer-based Music Models:** Additional transformer-focused resources were studied to understand contemporary token-based pipelines, including spectrogram encoders and autoregressive decoders outputting MIDI-like symbolic tokens.

These references informed the subsequent design of an LSTM-based system aimed at learning temporal structure from symbolic music representations.

Phase 5: Initial LSTM Prototyping and Infrastructure Challenges

Prototype Experiments

Initial attempts focused on building a basic LSTM generator aligned with existing literature. Early trials were limited by hardware constraints; both personal and laboratory systems experienced GPU failures or incompatibility, preventing reliable training. These issues motivated a redesign of the codebase to reduce memory overhead and streamline training once hardware became available.

Phase 6: Pipeline Reconstruction and Data Processing Refinement

Structural Rewrite of the LSTM Codebase

A full rewrite of the LSTM pipeline was undertaken to improve modularity and correctness. Particular focus was placed on clarifying the relationship between input tokenization, temporal ordering, and batch construction, as these elements directly influence model stability.

Corrections to the Data Processing Pipeline

During reconstruction, several issues were identified in the earlier preprocessing logic:

- The JSON dataset encodes music as **continuous note-by-note sequences**, not isolated feature vectors.
- The previous implementation treated each vector as an independent token, disrupting temporal continuity.

A revised data-processing system now constructs sequences that accurately reflect the original musical structure, ensuring proper temporal dependencies for recurrent learning.

Dataset and DataLoader Integration

A structured PyTorch `Dataset`/`DataLoader` framework was implemented. It provides:

- Correct sequence slicing,
- Efficient batching and shuffling,
- Compatibility with sliding-window generation approaches,
- Scalability for long-token sequences (e.g., 16k tokens with 64-note windows).

This refactor established a stable foundation for repeated experiments and tuning.

Phase 7: Tokenization Experiments and Model Optimization

Token Representation Analysis

Two principal input-token formats were evaluated:

- **2D Tokens:** $(\text{duration}, \text{note})$, preserving rhythm information but introducing duration-related noise.

- **1D Tokens:** Sampling note values at one-second intervals, yielding cleaner phonetic outputs and more stable training dynamics.

The 1D representation demonstrated superior performance, largely due to its reduced sensitivity to noisy or irregular duration encoding.

Training Behavior and Hyperparameter Exploration

Multiple optimization strategies were tested, including variations in learning rate, optimizer choice, and loss function. Cross-entropy loss consistently provided the highest-quality generations, matching its suitability for discrete token prediction.

Experiments on Bach datasets showed that the LSTM occasionally produced coherent musical phrasing. However, the output quality depended strongly on:

- The initial random seed,
- Preprocessing precision,
- Stability of the sliding-window construction,
- Length of the training sequence.

Computational Limitations

A major constraint was training time. On available hardware, a single batch trained for 20 epochs required approximately 90 minutes, significantly slowing the experimentation cycle.

Overall Summary of the RNN/LSTM Era

The LSTM-based exploration provided valuable insights into symbolic and token-based music modeling. While recurrent architectures can capture short-term temporal patterns and occasionally produce musically coherent phrases, they struggle with:

- Long-range dependencies,
- Large-token vocabularies,
- Extended sequence generation without drift.

These constraints motivated the transition back toward diffusion-transformer hybrids and tokenized latent-generation architectures with stronger long-context modeling capacities.

What's Next: Toward Diffusion Transformers (DiTs)

The next stage of the project focuses on unifying the strengths of both explored paradigms—diffusion models and recurrent/attention-based sequence models—through the adoption of **Diffusion Transformers (DiTs)**. This architecture introduces an attention-driven generative mechanism within the diffusion process, effectively merging the token-awareness of transformer models with the stability and expressivity of diffusion-based generation.

Motivation

Both previous eras revealed valuable insights:

- Diffusion models excel in high-fidelity generation but struggle with long-range temporal structure without explicit context mechanisms.
- LSTM and token-based models capture symbolic structure but lack the generative power needed for high-dimensional audio modeling.

Diffusion Transformers provide a natural synthesis of these strengths. They operate entirely in a latent space, allowing:

- **Attention-based reasoning over long contexts**,
- **Diffusion-based denoising for high-quality generation**, and
- **Flexible conditioning schemes** through tokenization and embeddings.

This combination enables modeling of long-form musical sequences while preserving fine-grained spectral detail.

Tokenization in Latent Space

The proposed framework follows a latent-diffusion approach in which raw audio is compressed into a structured latent representation. Instead of operating directly on spectrogram pixels, DiTs manipulate:

$$z \in \mathbb{R}^{h \times w \times c}$$

as a sequence of token-like latent patches.

Attention layers then provide the capacity to:

- reference musical motifs across long durations,
- maintain global structure such as tonal centers,
- stabilize transitions between phrases using cross-token context.

Custom Positional Encoding for Music

A key future direction involves designing **custom positional encodings** suited specifically for Maqam and microtonal musical structure. Instead of relying solely on classical diffusion timestep embeddings, the model can incorporate additional conditioning dimensions:

- **Time-based positional encoding:** corresponding to the latent sequence index;
- **Diffusion-step encoding:** representing noising depth;
- **Ornamentation-aware encoding:** encoding vibrato, slides, trills, and other stylistic elements characteristic of Maqam;
- **Stylistic-conditioning vectors:** representing performance attributes, timbre variations, or improvisational gestures.

Embedding these stylistic cues into the positional encoding framework allows the model to interact with higher-level musical semantics. This opens the possibility of studying how ornamentation affects learned latent spaces and how such signals influence both phonetic realism and creative expressivity.

Expected Benefits

The DiT framework is expected to offer several advantages:

- Improved global coherence across long sequences due to attention mechanisms;
- Higher-fidelity audio generation maintained by the diffusion backbone;
- Greater control over musical style and ornamentation through custom embeddings;
- A unified latent domain enabling seamless integration with the existing sliding-window and tokenization strategies.

Research Outlook

Future work will involve:

- Constructing a latent tokenizer compatible with the current autoencoder,
- Implementing a small-scale DiT as a proof of concept,
- Testing ornamentation-aware positional encodings,
- Studying the influence of stylistic conditioning on creativity, phonetic clarity, and Maqam authenticity.

This next phase represents a convergence of the project's two principal research paths, aiming to build a generative model that is both structurally aware and musically expressive.