An Evaluation of Generative Audio Architectures Under Consumer-Hardware Constraints

---

Idea

Complete a controlled, empirical comparison of three dominant deep learning sequence paradigms trained on physical, continuous performance control curves, these include fundamental frequency, RMS energy, and note onset indicators. In particular, we will be investigating the concept of gesture-to-audio timbre transfer under strict computational constraints (in this case it will mainly be a 3080 with 10GB of VRAM. The goal is to see which architecture enables high-fidelity instrument synthesis on consumer-grade hardware These architectures include:

- Diffusion Transformers (DiT)
    - Operating in continuous latent spaces via iterative de-noising, this establishing the benchmark in acoustic fidelity
- State Space Models (Mamba)
    - Operating in continuous latent spaces via selective linear recurrence, optimized for low-latency generation
- Masked Non-Autoregressive Transformers (Non-AR)
    - Operating in discrete token spaces via parallel iterative masking, optimized for rapid batch generation

Main Questions

1. Perceptual Quality under VRAM Constraints
    a. How do continuous-diffusion (DiT), continuous-recurrent (Mamba), and discrete-masked (Non-AR) modeling paradigms compare in terms of reconstructed perceptual quality and objective spectral accuracy when trained under identical local VRAM constraints
2. Synthesis Efficiency and Resource Footprint
    a. What is the relationship between model architecture, local generation throughput (Real-Time Factor), and peak GPU memory utilization during local inference on consumer-grade hardware?
3. Contour Adherence and Expressive Alignment
    a. To what extent does the underlying sequence processing mechanism (denoising, recurrence, or parallel masking) affect a model's ability to accurately follow continuous, highly dynamic physical control gestures (pitch slides, micro-dynamics, and sharp transients)?

Plan

The plan is to use a solo violin as our primary target instrument, with later adding other instruments for evaluation (possible options being flute and trumpet). To facilitate multi-instrument synthesis within a single architecture, all models will be conditioned on a categorical instrument identifier alongside the continuous performance contours.

We will be evaluating these models across four key axes:

1. Perceptual Audio Quality
    a. Fréchet Audio Distance (FAD)
        i. To measure general acoustic realism against reference datasets.
    b. Multi-Scale Spectral Loss (MSSL)
        i. To measure spectral reconstruction accuracy.
    c. MUSHRA (Multi-Stimulus Test with Hidden Reference and Anchor)
        i. To conduct subjective, human-in-the-loop listening tests evaluating overall timber naturalness
2. Physical Control Alignment
    a. Gross Pitch Error (GPE):
        i. To evaluate how accurately the model follows the input frequency contour
    b. Energy Correlation
        i. To measure the mathematical alignment between the target volume envelope and the synthesized output envelope.
3. Acoustic Consistency & Artifact Detection
    a. Spectral Flux
        i. To track the frame-to-frame rate of spectral change, allowing us to mathematically detect unnatural phase jumps or sudden "digital buzz."
    b. Envelope Jitter
        i. To measure micro-amplitude fluctuations, specifically catching phase mismatches or amplitude stuttering at generation boundaries.
4. Computational Efficiency
    a. Real-Time Factor (RTF)
        i. To measure generation speed (seconds of audio generated per wall-clock second).
    b. Peak VRAM
        i. To track memory consumption during training and inference passes on our target RTX 3080 hardware.

These multi-dimensional metrics will be compiled to plot the Quality-vs-Compute Pareto Frontiers for all three architectures, revealing the exact performance-to-cost tradeoffs of modern generative audio backbone