Weak Tokenization: A Preliminary Study of Dynamic Audio Chunking for Irregular Music Generation

LLM4MA

Weixi Zhai
Quanzhou Normal University
wishzhai@gmail.com

1.Introduction

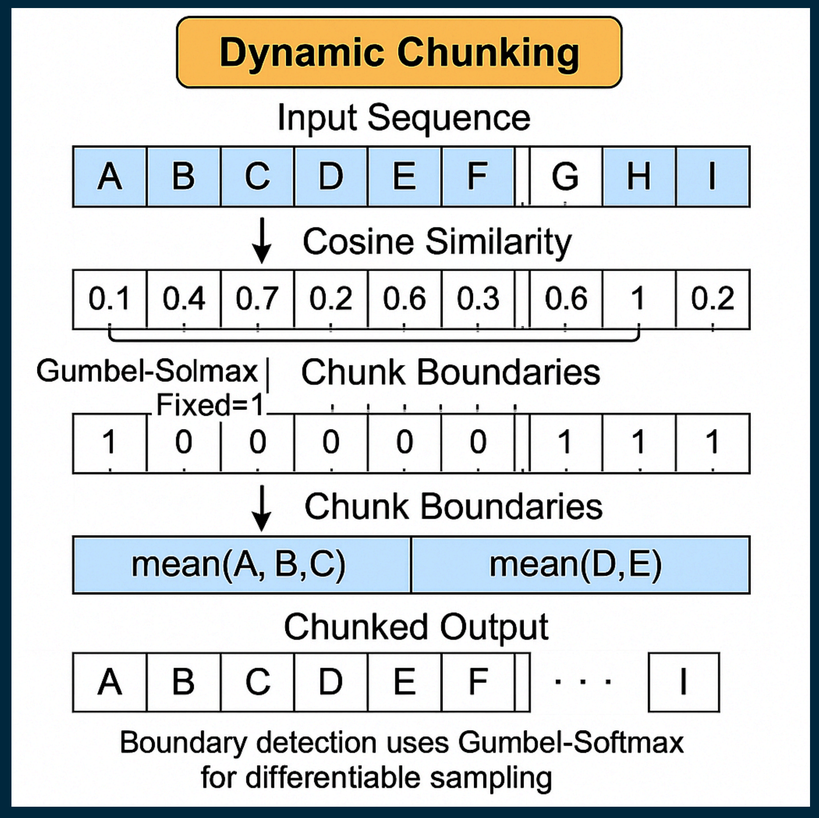
Technology provides the tools, culture guides their use. IDM exemplifies how fixed, grid-based tokenization leads to stylistic convergence. We introduce Dynamic Chunking, an audio-driven, content-adaptive segmentation method that enables LLMs to reason about music non-linearly. By breaking free from uniform grids, our approach restores creative irregularity and aesthetic diversity, allowing generative models to capture the anti-linear structures central to IDM's expressive power.

Our Contribution

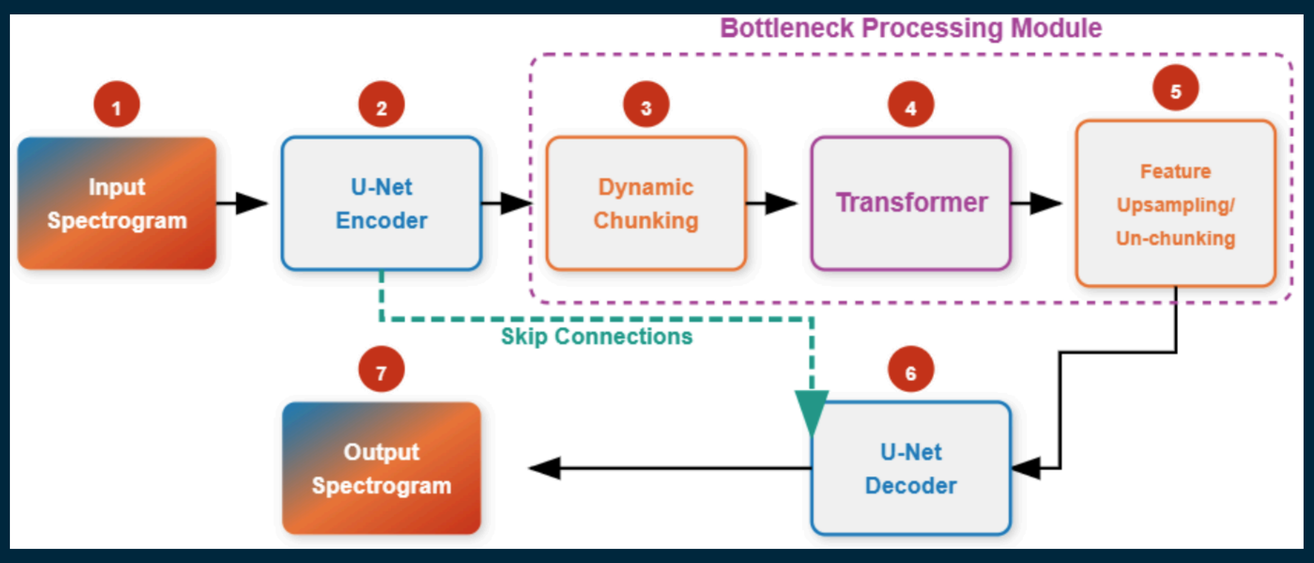
1. A Dynamic Audio Chunking⚙️〰️:
- Inspired by tokenizer-free NLP models.
- Investigates a generative architecture where segmentation is learned directly from audio features.
- Enables musically-aware, content-driven units rather than fixed temporal grids.

2. A Learnable Objective for Cognitive Complexity (L-Score)🎯📊:
- Introduces the L-Score, a multi-dimensional complexity metric.
- Spans timbral, rhythmic, and structural axes.
- Aims to replace conventional notions of "pleasantness" with a learnable proxy for listener challenge and aesthetic tension, by matching target distributions of L-Score.

2.Methodology



Dynamic Audio Chunking pipeline. Cosine similarity guides boundary prediction, Gumbel-Softmax sampling enables differentiable segmentation, and mean pooling produces compressed chunk embeddings.



Overview of the model architecture. The model utilizes a U-Net structure where the core bottleneck module (within the purple dashed box) consists of three components: Dynamic Chunking, a Transformer, and a Feature Upsampling/Un-chunking layer. The output from the encoder is processed by this bottleneck module and then reconstructed into the final output spectrogram by the decoder, which is aided by skip connections.

3.Experiment&Result.

L-Score Dimension	Target	Uncon	Seed
Timbral Complexity	0.4	**0.409**	0.502
Rhythmic Density	0.8	*2.962*	*6.667*
Rhythmic Irregularity	0.6	0.739	0.067
Structural Complexity	0.5	**0.496**	0.571

We trained our model using curriculum learning over 50 epochs, starting with a 10-epoch reconstruction warmup, followed by 40 epochs gradually guided by the L-Score. Quantitative evaluation shows strong alignment with spectral and structural targets (timbral 0.409 vs 0.4, structural 0.496 vs 0.5), indicating the L-Score effectively guides these aspects.
However, rhythmic control remains a major limitation: generated audio exhibits oversaturated density (2.962–6.667 vs 0.8) and chaotic micro-events, failing to capture IDM's deliberate irregularity. Qualitative analysis highlights limited musical coherence and strong dependence on seed material.
In summary, while our framework successfully models spectral and structural complexity, it struggles with temporal regulation and autonomous generation, underscoring challenges in producing anti-functional, non-linear musical aesthetics.

Code: