1 Segment-Level Feature Representation

Each beat-aligned audio segment is represented by a fixed-dimensional feature vector derived from its short-time Fourier transform (STFT). All audio is processed at a fixed sampling rate of 22,050 Hz, and identical STFT parameters are used across all segments to ensure comparability and reproducibility.

Let $X(f,t)$ denote the complex-valued STFT of a given audio segment, where f indexes frequency bins and t indexes time frames. We first compute the corresponding power spectrogram

$$P(f,t) = |X(f,t)|^2. \tag{1}$$

A mel filterbank with $M = 40$ mel frequency bands is then applied to $P(f,t)$ to obtain a mel-spectrogram $M(m,t)$, where $m \in \{1,\ldots,40\}$ indexes mel bands. This transformation provides a perceptually motivated frequency representation while preserving spectral energy information.

To obtain a fixed-dimensional embedding for each segment, the mel-spectrogram is temporally averaged:

$$v_m = \frac{1}{T}\sum_{t=1}^{T} M(m,t), \tag{2}$$

yielding a feature vector $\mathbf{v} \in \mathbb{R}^{40}$, where T denotes the number of STFT frames in the segment.

All feature vectors are non-negative by construction. Prior to similarity computation, each vector is ℓ_2-normalized to unit norm. This normalization ensures that cosine similarity reflects angular similarity in feature space rather than absolute energy differences.

The feature construction described above is fixed throughout all experiments and is used consistently for similarity graph construction and subsequent quantum-inspired dynamics. Any modification to this feature definition would invalidate direct comparisons across experiments.