System Overview

The system is a content-based music recommender designed to retrieve acoustically similar tracks using spectral descriptors and cosine similarity. Audio material is drawn from the FMA Small dataset, where approximately forty tracks are sampled, and thirty-second excerpts are rendered as WAV files at a uniform sampling rate of 22,050 Hz. Each clip undergoes short-time Fourier analysis using a 2,048-point FFT with a 512-sample hop and Hann window, generating a magnitude spectrogram that captures the evolving frequency content of the signal. From this, a mel-spectrogram is derived and converted to a log-mel representation, which serves as the input for computing a bank of mel-frequency cepstral coefficients (MFCCs).

These MFCCs, alongside traditional spectral statistics centroid, roll-off, flatness, and zero-crossing rate plus a tempo estimate derived from onset strength, are aggregated across time using both means and standard deviations. The resulting features form a fixed-length vector that summarizes the timbral and rhythmic character of each track. The feature matrix is cleaned by replacing infinite values, imputing missing entries with column means, and standardizing to zero mean and unit variance to prepare for cosine-based similarity computation.

Similarity between tracks is quantified via cosine distance in the standardized feature space, yielding a dense similarity matrix where diagonal entries are zeroed to prevent trivial self-matches. For every seed track, the system ranks all other tracks by descending similarity and returns the top-K nearest neighbors. An ablation analysis isolates the contribution of timbral versus broader spectral features by comparing an MFCC-only model to a combined model that concatenates cepstral and spectral statistics with tempo.

Performance Evaluation

Evaluation centers on both **feature stability** and **semantic neighborhood coherence**. Feature stability measures the internal robustness of the MFCC representation by drawing two random five-second segments from each track, computing their mean cepstral vectors, and calculating the Pearson correlation between them. This approach tests whether the system's spectral features remain consistent under temporal variation within a single song. High correlation indicates that the representation encodes core timbral identity rather than transient noise or production artifacts.

Semantic quality is assessed through **Artist@K**, **Genre Purity@K**, and **Mean Reciprocal Rank (MRR)** metrics. Artist@K quantifies the proportion of seed tracks whose top-K neighbors include at least one by the same artist, capturing artist-level cohesion in retrieval. Genre Purity@K calculates the average fraction of neighbors sharing the seed's genre label, evaluating stylistic homogeneity within each local cluster. MRR reports the expected inverse rank of the first same-artist occurrence within the top fifty candidates, reflecting how quickly related material surfaces in ranked recommendations.

Across the evaluated subset, feature stability computed as the correlation between MFCC means from random segments achieved **0.969**, confirming high temporal reliability in cepstral representations. Artist@10 reached **0.226**, while Genre Purity@10 attained **0.648**, indicating stronger alignment at the genre level than at the artist level. The mean reciprocal rank for same-artist retrieval was **1.000**, showing that when same-artist tracks were present, they consistently appeared at the top of the ranking.

An ablation experiment further compared MFCC-only and combined feature models. The MFCC-only configuration yielded an Artist@K score of **0.951**, slightly surpassing **0.949** for the combined variant. This result suggests that cepstral features alone dominate the similarity structure on the current dataset size, capturing most of the discriminative spectral cues needed for perceptually coherent retrievals.

Limitations

The principal limitations stem from sample size and class imbalance, both of which can inflate artist-based metrics and obscure generalization trends. With only a small subset of tracks per genre and artist, stability estimates are more reliable than semantic ones. Scaling the dataset to several hundred or thousand clips, distributed evenly across genres, would reduce variance and improve statistical confidence. Future directions include incorporating dimensionality-reduction techniques such as PCA or UMAP for compact embeddings, employing efficient nearest-neighbor indexing (e.g., Faiss or Annoy) to enhance scalability, and experimenting with metric-learning objectives to refine the geometry of the embedding space.