SAPPHIRE

Semantic and Acoustic Perceptual Holistic Integration REtrieval

A Perceptually-Driven Framework for Music Similarity Retrieval: Project Proposal and Implementation Plan

Prakhar Singhal, Kushal Shah, and Aditya Raj Singh

September 2, 2025

Contents

Abstract

The field of Music Information Retrieval (MIR) is fundamentally challenged by the ”perceptual gap”—the disconnect between objective computational analysis and the subjective, nuanced nature of human music perception. Traditional systems, reliant on uni-modal acoustic features and rigid categorical labels, have reached a performance ceiling, failing to capture the intricate cognitive and emotional relationships that define a listener’s experience. This proposal outlines a comprehensive, multi-modal framework designed to bridge this gap. Our approach is explicitly grounded in principles of psychoacoustics and cognitive science, moving beyond what can be measured to what is truly perceived. The solution architecture leverages a modular, multi-modal design that integrates acoustics, lyrical semantics, song structure, and granular vocal phonetics. The core of our methodology is the application of Cross-Modal Contrastive Learning (CMCL), powered by Large Language Models (LLMs) for scalable data synthesis, to create a unified, perceptually-aligned embedding space. This document details a four-phase implementation plan, beginning with foundational research and feature identification, proceeding through data corpus development and statistical analysis, culminating in model development, and validated by a rigorous Human-in-the-Loop (HITL) evaluation paradigm. The project aims to deliver not just a more accurate retrieval system, but a new, human-centric standard for music discovery and recommendation.

1 The Problem: Bridging the ”Perceptual Gap” in Music Similarity

1.1 The Disconnect Between Computation and Cognition

The core mission of Music Information Retrieval (MIR) is to develop systems that organize, search, and recommend music in a manner that aligns with human intuition. This endeavor is hindered by the persistent **”perceptual gap”** [1], a term describing the fundamental discrepancy between a machine’s analysis of a signal’s physical properties (e.g., frequency, amplitude) and the human brain’s subjective interpretation of those properties as perceptual attributes (e.g., pitch, loudness). As noted by Slaney, “The features that are easy to extract computationally are not the features that are most important for human perception”.

This relationship is profoundly non-linear. The perception of loudness, for instance, is frequency-dependent, a phenomenon mapped by the Fletcher-Munson curves [2]. Similarly, the perception of pitch relies not just on the fundamental frequency but also on harmonic content, as illustrated by the ”missing fundamental” effect [3]. A model built on a brute-force combination of generic audio features like Mel-Frequency Cepstral Coefficients (MFCCs), while powerful, is demonstrably inferior for high-level tasks to one built on features specifically engineered to approximate human perceptual dimensions [4].

1.2 The Failure of Traditional MIR Paradigms

Current approaches are constrained by two critical methodological flaws that prevent them from bridging this gap:

1. **Reliance on Static, Uni-Modal Labels:** Much of MIR has relied on predefined categorical labels like genre or mood. This is not scalable and fails to capture music’s subjective qualities. More critically, the low inter-rater agreement among human annotators

[5] means the ”ground truth” is unstable. As noted in analyses of MIREX evaluations, “an algorithm cannot be expected to perform better on a task... than the consensus of human evaluators” [6]. This has created a natural **performance ceiling** for supervised algorithms, which cannot surpass the inconsistent human consensus they are trained to emulate.

2. **Ignoring Multi-Modal Complexity:** Human similarity judgment is a holistic, or ”gestalt,” process, integrating vocal timbre, lyrical meaning, song structure, and rhythmic feel. As Levitin argues, musical categories are ”fuzzy and depend on a large number of factors” [7]. Uni-modal systems that focus only on acoustics are fundamentally incapable of capturing these essential cross-modal relationships that listeners implicitly understand.

The refined problem, therefore, is not merely to improve accuracy against flawed benchmarks, but to **shift the paradigm from modeling static labels to modeling the dynamic, multi-faceted process of human perception itself.**

2 Proposed Approach: A Human-Centric, Multi-Modal Framework

We propose a novel framework that directly addresses the perceptual gap by grounding its architecture in the principles of human music perception. Our approach is built on three foundational pillars:

- **Psychoacoustically-Informed Feature Engineering:** We will prioritize features with strong, verifiable perceptual correlates over generic low-level descriptors. This involves moving from a signal-processing-centric view to a human-centric one, asking not just ”what is in the signal?” but ”what does a listener hear?”

- **Holistic Multi-Modal Integration:** Inspired by the ”transformational theory” of musical similarity, which posits that similarity is a function of the cognitive effort needed to transform one piece into another [8], our system will fuse four critical modalities—Acoustics, Lyrics, Structure, and Vocal Performance—into a single, unified representation that captures the music’s holistic identity.

- **Advanced, Perceptually-Aligned Learning:** We will move away from supervised learning on flawed labels and instead use state-of-the-art self-supervised techniques to learn a shared embedding space that aligns with the complex, non-linear nature of human similarity judgments.

3 The Solution: Architectural Design and Methodology

3.1 Core Architectural Principles

The proposed system is a modular, multi-modal framework designed to achieve a semantically aligned, unified embedding space.

- **Modularity:** Separate, specialized encoders will be used for each input modality (e.g., a CNN/Transformer for audio, a Transformer for text). This allows for optimal, modality-specific feature extraction before fusion [9].

- **Semantic Alignment:** The core objective is to map representations from different modalities into a shared latent space where semantically related concepts are proximate. This ensures that a sad song's audio and its somber lyrics map to a similar region in the embedding space.

- **Unified Embedding:** The final output for any given song is a single, rich vector. This unified representation captures a holistic view of the music, enabling complex, cross-modal retrieval tasks (e.g., query by humming, text, or example).

3.2 Advanced Learning Paradigm

Our methodology is inspired by recent breakthroughs in self-supervised and multi-modal learning.

- **Cross-Modal Contrastive Learning (CMCL):** We will employ a dual-encoder architecture trained with a contrastive loss function (e.g., InfoNCE). The model learns by minimizing the distance between ”positive pairs” (a song's audio and its lyrical description) while maximizing the distance from ”negative pairs” [10]. This approach, inspired by models like CLIP [11] and successfully applied to music in CrossMuSim [12], learns robust, perceptually meaningful representations by leveraging natural language as a rich supervisory signal, which is far more expressive than one-hot encoded labels.

- **Large Language Models (LLMs) for Data Synthesis:** A primary bottleneck in multi-modal MIR is the scarcity of high-quality, paired text-music data. We will overcome this by using a state-of-the-art LLM (e.g., GPT-4) as a sophisticated data synthesis engine. This is a proven strategy, with recent work like CLaMP demonstrating that using LLMs to ”filter, correct, and enrich noisy web-mined datasets” significantly boosts retrieval performance [13]. Through carefully engineered prompts, the LLM will generate contextually rich, diverse, and stylistically-aware descriptions for a vast corpus of music, providing the scalable data needed to train our CMCL model effectively.

4 Detailed Implementation Roadmap

We propose a four-phase implementation plan, designed to move systematically from foundational research to a validated, human-aligned system.

4.1 Phase 1: Foundational Research and Feature Identification

Step 1.1: Comprehensive Literature Review A systematic review across: (1) Psychoacoustics and cognitive musicology (Lerdahl & Jackendoff, Deutsch) to inform feature selection; (2) State-of-the-art MIR techniques (e.g., SI-PLCA for structure, C-NMF for segmentation); (3) Multi-modal deep learning (e.g., attention mechanisms, contrastive learning).

Step 1.2: Identification and Verification of Perceptually Salient Features We will identify a candidate set of features and establish a protocol for verifying their perceptual relevance through small-scale user studies and correlation analysis with existing perceptual datasets (e.g., MagnaTagATune).

Table 1: Candidate Perceptual Feature Domains

Domain	Candidate Features for Investigation
Acoustic	Perceptually-weighted MFCCs, Chroma (tonal), Loudness Curves (ISO 226)
Rhythmic	Beat/Tempo, Swing/Groove parameters, Rhythmic density, Micro-timing
Melodic	Pitch contour (CREPE), Vibrato/Intonation, Melodic motifs, Raga/Scale characteristics
Lyrical	Semantic vectors (BERT), Sentiment arcs, Phonetic patterns (phoneme rate, type)
Structural	Computationally derived segments, Harmonic progression, Repetition (SI-PLCA)

4.2 Phase 2: Data Corpus Development and Statistical Analysis

Step 2.1: Dataset Aggregation We will aggregate key datasets: **SingStyle111** [14] for vocal/phonetic analysis, **SALAMI** [15] for structural analysis, and the **Million Song Dataset (MSD)** [16] as our base corpus.

Step 2.2: LLM-Powered Data Synthesis Pipeline An automated pipeline will be developed. We will employ a Chain-of-Thought or few-shot prompting strategy, providing the LLM with audio features (e.g., tempo, key, energy) and metadata to generate rich textual descriptions covering timbre, mood, instrumentation, and lyrical themes.

Step 2.3: Statistical Analysis of Perceptual Features A deep statistical analysis will be performed on the features identified in Phase 1 across the aggregated corpus. This will involve:

- **Correlation and Mutual Information Analysis:** To identify redundancies and key inter-modal relationships (e.g., between harmonic complexity and lyrical sentiment).
- **Dimensionality Reduction:** Using Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize feature clusters and validate their ability to separate genres or styles.

4.3 Phase 3: Multi-Modal Model Development and Training

Step 3.1: Model Architecture Selection We will design a dual-encoder CMCL architecture. The audio encoder will be a pre-trained, state-of-the-art model like **PANNs** [17], which uses a CNN-based architecture proven effective for audio pattern recognition. The text encoder will be a standard Transformer model like **BERT** [18]. We will investigate cross-attention mechanisms to allow for richer fusion between modality embeddings before the final projection head.

Step 3.2: Implementation and Training The model will be implemented in PyTorch. It will be trained on the LLM-synthesized dataset using the InfoNCE loss. Training will be conducted on a distributed GPU cluster to handle the large dataset size.

Step 3.3: Iterative Refinement and Hyperparameter Tuning The model will undergo extensive hyperparameter tuning (learning rate, batch size, embedding dimensionality, temperature scaling for the loss function) using a validation set to optimize retrieval performance.

4.4 Phase 4: Perceptual Evaluation and HITL Validation

Step 4.1: Objective and Subjective Metric Definition We will use a hybrid evaluation approach:

- **Objective Metrics:** Standard retrieval metrics (recall@k, mAP) alongside perceptual metrics like Multi-Scale Structural Similarity (MS-SSIM) [19] applied to spectrograms to measure timbral and structural similarity.
- **Subjective Metrics:** Rigorous A/B listening tests where human subjects rate the similarity of song pairs retrieved by our model versus a baseline (e.g., a standard content-based filtering model).

Step 4.2: Human-in-the-Loop (HITL) Feedback System The ultimate goal is to move beyond static evaluation. We will design a system to continuously refine the model using human feedback, inspired by Reinforcement Learning from Human Feedback (RLHF) [20]. User feedback will be used to fine-tune the embedding space, allowing the model to adapt and personalize its understanding of perceptual similarity over time.

5 Sprint-wise Work Division Plan

The four-phase roadmap is broken down into an agile, six-sprint implementation plan. The following table outlines the focus, key tasks, and deliverables for each two-week sprint.

Table 2: Sprint-wise Implementation Plan

Sprint	**Focus**	**Key Tasks**	**Deliverable**
1 (Weeks 1-2)	Phase 1: Foundation	- Conduct Literature Review - Prepare a initial Perceptual Feature List	Feature Specification Document
2 (Weeks 3-4)	Phase 2: Data Corpus	- Aggregate Preprocess Datasets - Implement Feature Extraction Scripts - Run Extraction on Full Corpus	Processed Dataset with Features
3 (Weeks 5-6)	Phase 2: Synthesis	- Finalize the Perceptual Feature List - Generate Text Descriptions - Perform Statistical Analysis	Paired (Audio, Text) Dataset
4 (Weeks 7-8)	Phase 3: Baseline Model	- Implement CMCL Architecture - Build Data Loaders - Train Baseline on Sample Data	Working End-to-End Training Pipeline
5 (Weeks 9-10)	Phase 3: Optimization	- Launch Full-Scale Training - Run Hyperparameter Tuning - Analyze Training Curves	Trained Model Checkpoint
6 (Weeks 11-12)	Phase 4: Validation	- Evaluate with Objective Metrics - Conduct Subjective A/B Listening Tests - Design HITL System Architecture	Final Evaluation Report & HITL Design Doc

6 Expected Outcomes and Impact

Project *SAPPHIRE* is framework for music similarity that is explicitly grounded in the cognitive and psychoacoustic principles of human perception. Complementing this, it introduces a new evaluation paradigm centered on Human-in-the-Loop (HITL) validation, designed to push the field beyond its reliance on static benchmarks and toward the development of truly dynamic and adaptive systems. Thus we hope to make significant contribution to the field.

References

[1] M. Slaney, *The perceptual gap in music information retrieval*. In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008.

[2] H. Fletcher and W. A. Munson, *Loudness, its definition, measurement and calculation*. The Journal of the Acoustical Society of America, 1933.

[3] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Brill, 2012.

[4] B. Logan, *Mel frequency cepstral coefficients for music modeling*. In Proc. of the International Symposium on Music Information Retrieval (ISMIR), 2000.

[5] D. P. W. Ellis, *Classifying music audio with timbral and chroma features*. In Proc. of the International Society for Music Information Retrieval Conference (ISMIR), 2012.

[6] J. S. Downie, *The music information retrieval evaluation exchange (MIREX)*. D-Lib Magazine, 2010.

[7] D. J. Levitin, *Memory for musical attributes*. In Music perception and cognition, 1999.

[8] J. B. Tenenbaum and F. L. T. Griffiths, *A rational basis for multi-modal similarity*. In Advances in Neural Information Processing Systems (NIPS), 2001.

[9] T. Baltrušaitis, C. Ahuja, and L. P. Morency, *Multimodal machine learning: A survey and taxonomy*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

[10] A. van den Oord, Y. Li, and O. Vinyals, *Representation learning with contrastive predictive coding*. arXiv preprint arXiv:1807.03748, 2018.

[11] A. Radford et al., *Learning Transferable Visual Models From Natural Language Supervision*. In Proc. of the International Conference on Machine Learning (ICML), 2021.

[12] Y. Li, J. Liu, and J. a. Fu, *CrossMuSim: A Cross-Modal Framework for Music Similarity Retrieval with LLM-Powered Text Description Sourcing and Mining*. arXiv preprint arXiv:2308.10118, 2023.

[13] S. Wu et al., *CLaMP: Contrastive language-music pre-training for cross-modal symbolic music information retrieval*. In Proc. of the International Society for Music Information Retrieval Conference (ISMIR), 2023.

[14] Y. Wu et al., *SINGSTYLE111: A MULTILINGUAL SINGING DATASET WITH STYLE TRANSFER*. In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024.

[15] J. J. Smith et al., *Design and creation of a large-scale database of structural annotations*. In Proc. of the International Society for Music Information Retrieval Conference (ISMIR), 2011.

[16] T. Bertin-Mahieux et al., *The Million Song Dataset*. In Proc. of the International Society for Music Information Retrieval Conference (ISMIR), 2011.

[17] Q. Kong et al., *PANNs: Large-scale pretrained audio neural networks for audio pattern recognition*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020.

[18] J. Devlin et al., *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805, 2018.

[19] Z. Wang, E. P. Simoncelli, and A. C. Bovik, *Multiscale structural similarity for image quality assessment*. In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003.

[20] P. F. Christiano et al., *Deep reinforcement learning from human preferences*. In Advances in Neural Information Processing Systems (NIPS), 2017.