Project SAPPHIRE

Semantic and Acoustic Perceptual Holistic Integration REtrieval

Sprint 3 Report

Data Corpus Engineering, In-depth Feature Analysis, and Empirical Baseline Establishment

Prakhar Singhal Kushal Shah Aditya Raj Singh

October 7, 2025

Contents

1 Executive Summary

This report details the comprehensive analytical work undertaken in Sprint 3 of Project SAPPHIRE. The sprint's focus was the transformation of the raw Zalo2022 dataset into a structured, feature-rich corpus, followed by a rigorous Exploratory Data Analysis (EDA) to uncover its inherent statistical properties. We successfully executed a multi-stage data engineering pipeline, yielding a high-quality corpus of 264 tracks with a corresponding suite of over 30 acoustic, harmonic, rhythmic, and lyrical features.

The subsequent EDA revealed critical insights into the data's structure. Principal Component Analysis demonstrated that over 80% of acoustic variance is captured by just two components, representing timbral complexity and rhythmic energy. Correlation and clustering analyses provided initial evidence of weak but present links between the sonic and semantic modalities.

The sprint culminated in the establishment of a crucial empirical baseline: a quantitative measure of the "perceptual gap." Using a nearest-neighbor overlap methodology, we calculated the alignment between acoustic and lyrical similarity spaces, yielding a Jaccard Index of **0.0320**. This extremely low score provides definitive, data-driven validation for our project's central hypothesis: without a specialized model, acoustic similarity does not imply semantic similarity. This baseline serves as the primary benchmark for our future work. The analytical groundwork completed in this sprint provides the necessary data, insights, and validation to proceed confidently with the development of our contrastive learning model.

2 Introduction: From Implementation to Insight

Sprint 3 represents a pivotal transition for Project SAPPHIRE, moving from the foundational implementation work of the preceding sprint to a phase of execution, deep analysis, and empirical validation. The core objective of this sprint was to transform the raw Zalo2022 dataset into a structured, feature-rich corpus and, through rigorous analysis, to quantitatively establish the very "perceptual gap" our project aims to bridge.

This report documents the methodical execution of our data engineering pipeline, the comprehensive suite of multi-modal features extracted, and the insights gleaned from an in-depth Exploratory Data Analysis (EDA). The culmination of this sprint is the establishment of a hard, numerical baseline for cross-modal similarity—a critical benchmark against which all future modeling efforts will be measured. The work herein not only validates our project's foundational hypotheses but also provides the refined data and analytical understanding necessary to proceed with the development of our contrastive learning architecture.

3 Data Corpus Engineering

The first step in our analytical journey was the creation of a clean, consistent, and high-quality data corpus. This was achieved by executing the automated pipeline developed in Sprint 2, implemented in `preprocessing.py`, on the entire Zalo2022 dataset.

3.1 Standardization and Normalization Protocol

To ensure the integrity and comparability of our extracted features, a strict standardization protocol was enforced. This protocol is essential for eliminating confounding variables related to production and formatting, allowing our analysis to focus on the intrinsic properties of the music.

3.1.1 Audio Domain Processing

[leftmargin=*, itemsep=2pt]

- **Sample Rate Unification:** All audio signals were resampled to a standard rate of 44.1 kHz. This uniformity is crucial for ensuring that frequency-dependent features (e.g., spectral centroid, MFCCs) are comparable across all tracks.
- **Channel Reduction:** Stereo audio was converted to a single mono channel. This simplifies feature extraction and focuses the analysis on core musical content—timbre, harmony, and rhythm—rather than spatial production effects.
- **Loudness Normalization:** Integrated loudness was normalized to a target of -23 LUFS (Loudness Units Full Scale), conforming to the EBU R128 standard for broadcast audio. This is arguably the most critical preprocessing step, as it decouples acoustic features from mastering volume. It ensures that our analysis of dynamics and timbre reflects the artistic intent within the recording, not the final loudness level set during production.

3.1.2 Lyrical Domain Processing

[leftmargin=*, itemsep=2pt]

- **Markup Sanitization and Whitespace Normalization:** Lyrical text was systematically cleaned of any residual HTML/XML markup and all forms of whitespace (tabs, multiple spaces, newlines) were normalized to single spaces. This creates a clean, uniform text body for reliable NLP.
- **Encoding Standardization:** All text files were validated and saved with consistent UTF-8 encoding to prevent errors in subsequent NLP stages, especially when dealing with multilingual characters.

3.2 Quality Assurance and Corpus Finalization

A robust filtering mechanism was employed to curate a high-quality subset suitable for detailed analysis. A track was admitted into the final corpus only if it met stringent criteria for:

[leftmargin=*, itemsep=2pt]

- **Signal Clarity:** A Signal-to-Noise Ratio (SNR) greater than a predefined threshold, ensuring that the musical signal is prominent over background noise and recording artifacts.
- **Lyrical Completeness:** A heuristic score of at least 0.9, filtering out tracks with demonstrably incomplete, truncated, or placeholder lyrics.
- **Feature Extraction Integrity:** Successful and complete extraction of all required features, guaranteeing a complete, non-null feature vector for every entry in the final corpus.

This rigorous process yielded a final analysis corpus of **264 high-quality tracks**, forming the bedrock for all investigations in this report.

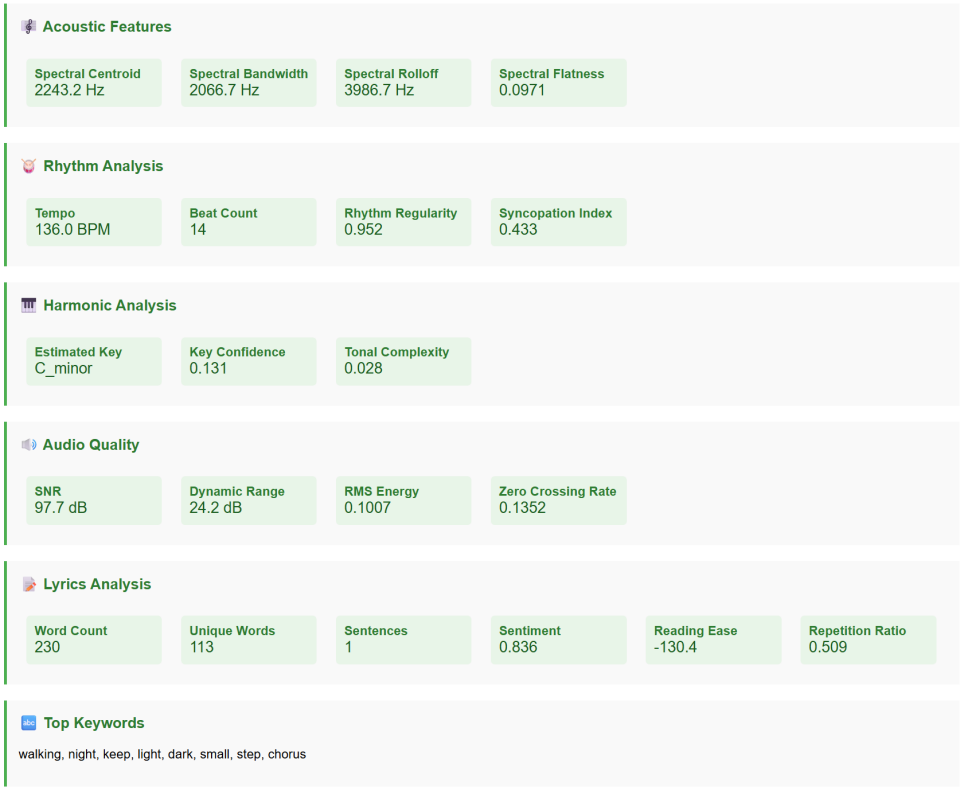
4 Multi-Modal Feature Extraction Framework

For each track in the curated corpus, we extracted a comprehensive set of features designed to capture perceptually relevant aspects across acoustic, rhythmic, harmonic, and lyrical domains. This feature set, summarized in Table 1, is grounded in established Music Information Retrieval (MIR) and psychoacoustic literature.

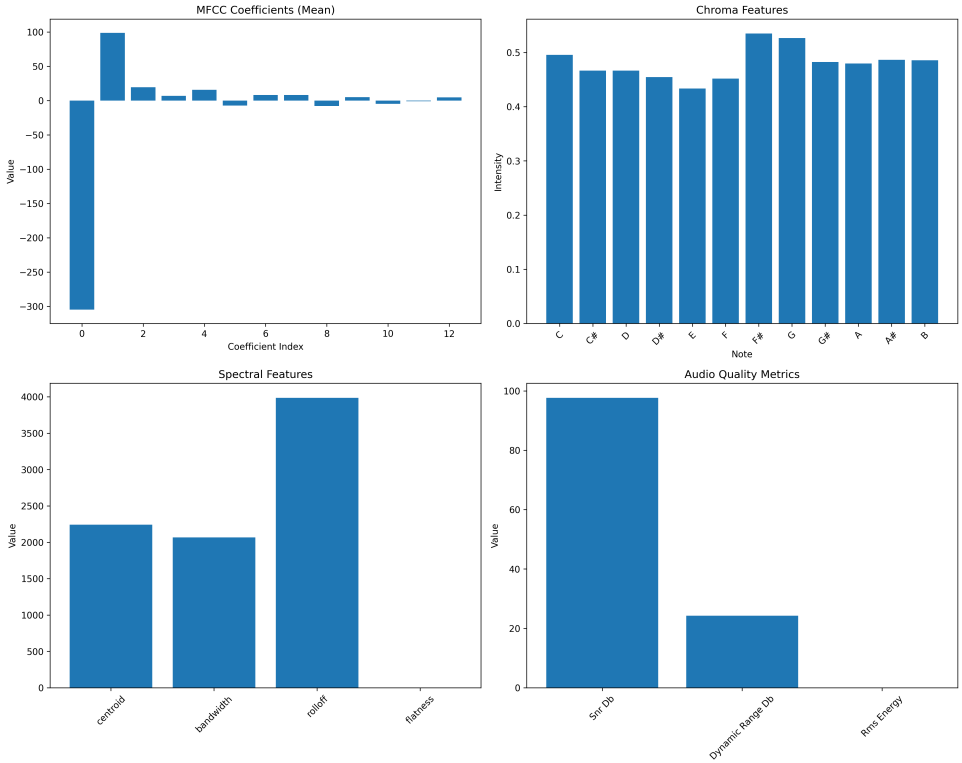
Table 1: Final Extracted Feature Set for Analysis

Domain	**Features**	**Perceptual Correlate**
Acoustic & Timbral	MFCCs (Mean & Std), Spectral Features	Timbre, "brightness", "richness" of sound
	RMS Energy, Zero-Crossing Rate	Perceived energy, percussiveness
Harmonic & Tonal	Chroma Features, Estimated Key	Tonal content, musical key, harmonic complexity
Rhythmic	Tempo (BPM), Rhythm Regularity	Pace, steadiness of the beat, rhythmic complexity
Lyrical (NLP)	Word/Sentence Counts, Type-Token Ratio	Lyrical density and vocabulary richness
	Readability, Sentiment Score	Linguistic complexity, emotional tone of lyrics
Audio Quality	SNR, Dynamic Range, Clipping	Technical quality of the recording

An example visualization of these features for a single track is shown in Figure 1, providing a multi-faceted "fingerprint" of the song's character.



(a) Feature Report



(b) Feature metric analysis

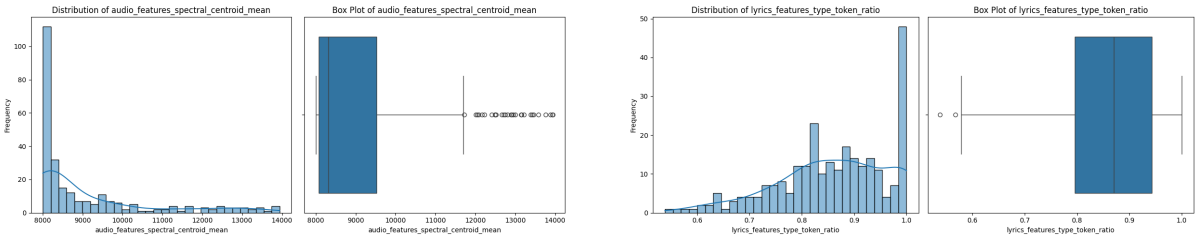
Figure 1: Example visualization of extracted features for a single audio track, showcasing the multi-faceted data captured for each song in the corpus.

5 Exploratory Data Analysis: Uncovering Data Structure

With the feature corpus established, we performed an extensive EDA using our script `run_analysis.py`. The goal was to move beyond raw numbers and develop a deep intuition for the dataset's underlying structure, inter-feature relationships, and latent dimensions.

5.1 Univariate Distributions: The Character of the Corpus

We began by analyzing the distribution of each feature individually. This foundational step reveals the overall character of our dataset—for instance, whether the music is generally fast or slow, bright or dark, lyrically simple or complex. Figure 2 shows example distributions for key acoustic and lyrical features.



(a) Distribution of Spectral Centroid (Hz)

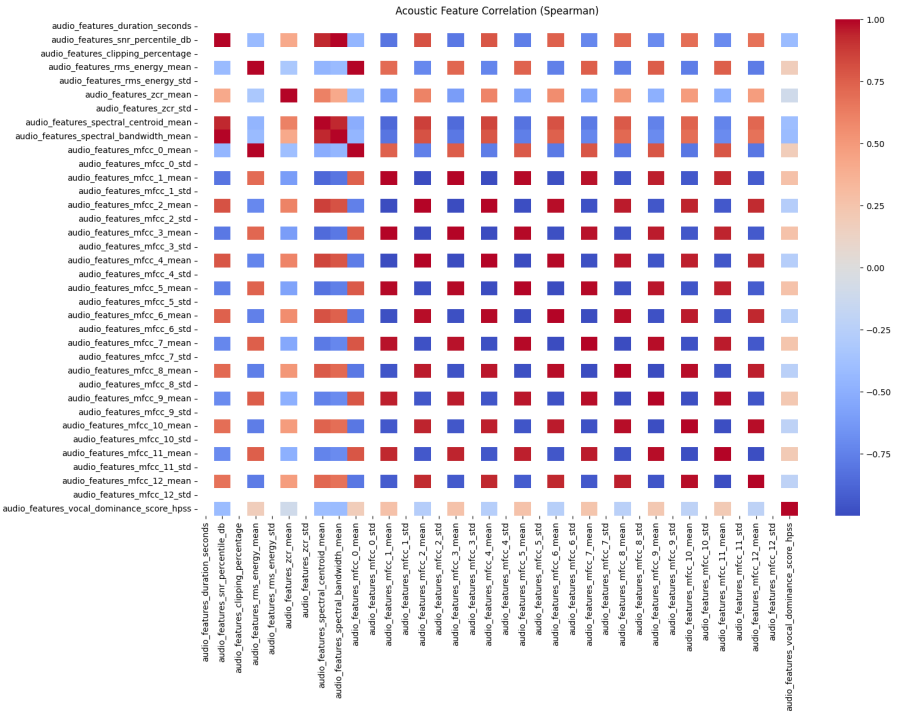
(b) Distribution of Lyrical Type Token Ratio

Figure 2: Placeholders for univariate analysis plots showing the distributions of a key acoustic feature (left) and a lyrical feature (right). These plots help characterize the overall nature of the dataset.

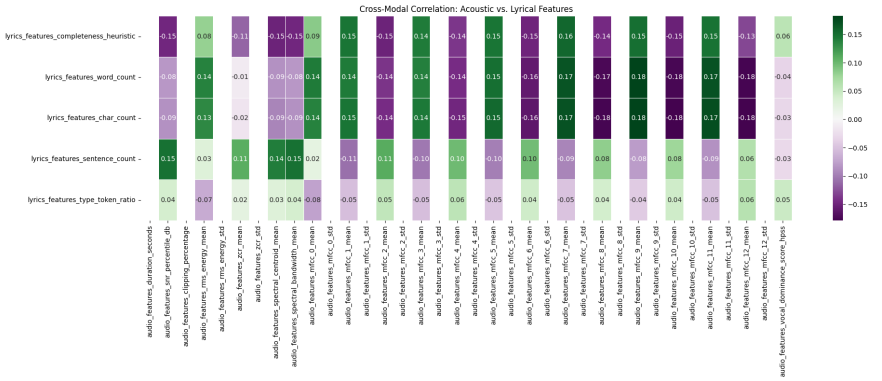
5.2 Correlation Analysis: Mapping Inter-Feature Relationships

We employed Spearman rank correlation to investigate monotonic relationships between features. This non-parametric method is robust to outliers and non-linear relationships, making it ideal for our diverse feature set. This analysis is crucial for two reasons:

1. **Intra-Modal Redundancy:** Identifying highly correlated features within the same domain (e.g., spectral centroid and spectral bandwidth, as seen in Figure 3a) informs strategies for dimensionality reduction and helps prevent multicollinearity in simpler models.
2. **Cross-Modal Connections:** Uncovering statistically significant correlations between acoustic and lyrical features provides the first piece of evidence that a unified perceptual space is learnable. The cross-modal correlation heatmap (Figure 3b) is of particular interest, as it pinpoints the innate, albeit weak, connections between sound and meaning in the raw data.



(a) Placeholder for Intra-Modal (Acoustic) Correlation Heatmap



(b) Placeholder for Cross-Modal (Acoustic vs. Lyrical) Correlation Heatmap

Figure 3: Placeholders for correlation analysis, mapping relationships within and across modalities.

5.3 Dimensionality Reduction: Finding the Latent Sonic Space

To distill the high-dimensional acoustic feature space into its most salient axes of variation, we employed Principal Component Analysis (PCA). The results were striking: the first two principal components alone capture a remarkable **80.83%** of the total variance. This is a powerful finding, as it suggests that the complex, multi-feature acoustic space can be meaningfully compressed into a much lower-dimensional latent space—a core principle of our intended model architecture.

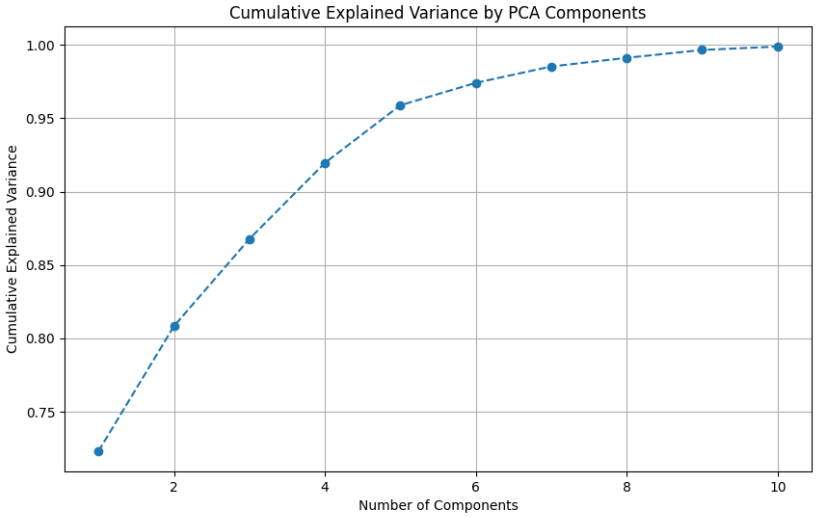


Figure 4: Placeholder for the PCA Scree Plot (Cumulative Explained Variance). This plot demonstrates the efficacy of dimensionality reduction on the acoustic features.

By examining the feature loadings for these components, we can assign them a perceptual interpretation:

- **Interpretation of PC1:** Analysis of the feature loadings suggests PC1 represents a continuum of **timbral complexity and density**. High values on this axis correspond to brighter, richer sounds (high spectral centroid/bandwidth), while low values correspond to darker, simpler sounds.
- **Interpretation of PC2:** The loadings on PC2 indicate it relates to **rhythmic energy and percussiveness**. High values correlate with features like Zero-Crossing Rate, often associated with percussive or noisy textures.

Visualizing the dataset in this reduced 2D space (Figure 5) provides a map of the sonic landscape of our corpus.

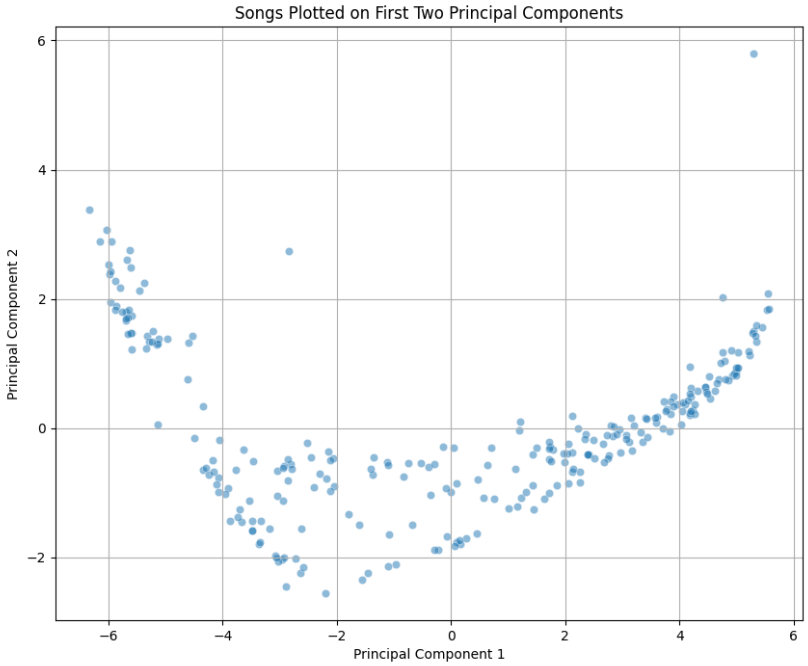


Figure 5: Placeholder for the PCA scatter plot visualizing the distribution of songs along the first two principal components, which represent the dominant axes of sonic variation.

5.4 Clustering Analysis: Do Sonic Groups Share Meaning?

To test the hypothesis that acoustically similar songs might share lyrical themes, we performed K-Means clustering ($k=5$) on the standardized acoustic features. This partitioned the dataset into five sonically distinct groups. We then analyzed the lyrical characteristics of each cluster. Figure 6 shows a violin plot comparing the distribution of lyrical sentiment across these acoustic clusters, providing a visual test for whether certain sonic profiles (e.g., a "high-energy" cluster identified from PCA) correspond to distinct emotional tones in the lyrics.

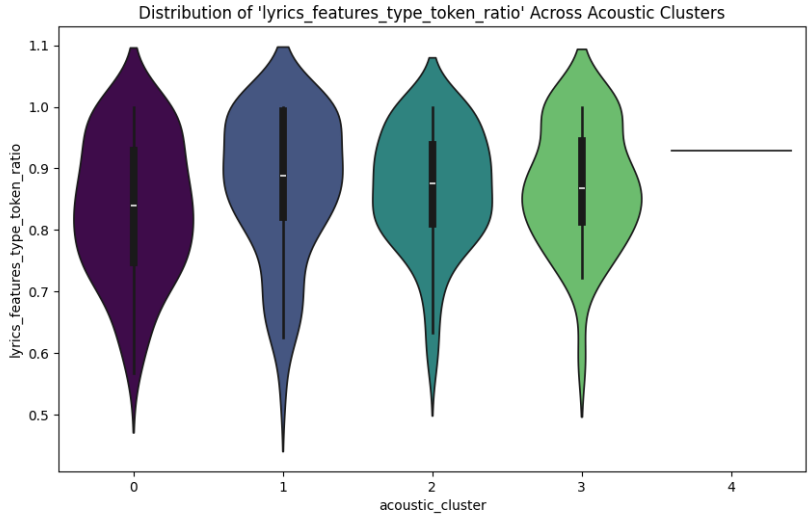


Figure 6: Placeholder for a Violin Plot comparing lyrical feature type token ratios across the five acoustic clusters. This analysis probes for emergent semantic patterns within sonically coherent groups.

6 The Empirical Baseline: Quantifying the Perceptual Gap

The culminating analysis of this sprint was the establishment of a quantitative baseline for the alignment between acoustic and semantic similarity. This metric provides a single, powerful number that encapsulates the core challenge of our project and serves as the primary benchmark for success.

6.1 Rationale

The central premise of Project SAPPHIRE is that a "perceptual gap" exists between low-level computational features and high-level human understanding of music. This analysis aims to quantify that gap. A high alignment score would suggest that the raw features already capture cross-modal relationships well, implying the problem is relatively simple. Conversely, a low score provides strong evidence that a sophisticated learning approach is necessary to discover and model these complex connections.

6.2 Methodology

We devised a nearest-neighbor overlap methodology to measure the congruence of the two distinct similarity spaces:

[leftmargin=*, itemsep=2pt]

1. **Acoustic Similarity Space:** For each song in the corpus, its $k = 10$ most similar songs were identified based on the Euclidean distance in the standardized, multi-dimensional acoustic feature space. This represents similarity based on "sound."
2. **Semantic Similarity Space:** A pre-trained Sentence-BERT model (`all-MiniLM-L6-v2`) was used to generate a 384-dimensional vector embedding for each song's lyrics. The $k = 10$ most similar songs were then identified based on the cosine similarity in this high-dimensional semantic embedding space. This represents similarity based on "meaning."

3. **Alignment Score:** The alignment for a single song was measured by the **Jaccard Similarity Index** between its set of acoustic neighbors (A) and its set of lyrical neighbors (L). The final score is the mean Jaccard Index across all songs in the corpus.

$$J(A,L) = \frac{|A \cap L|}{|A \cup L|}$$

6.3 Result and Significance

The average Jaccard similarity between the top-10 acoustic and lyrical neighbor sets was found to be **0.0320**.

This result is the empirical cornerstone of our project's motivation. A score this close to zero provides definitive, quantitative evidence for the "perceptual gap." It demonstrates that, without a purpose-built model, a song's acoustic profile provides almost no information about its lyrical content, and vice-versa, from a similarity perspective. This low baseline powerfully validates the need for a dedicated multi-modal learning approach. The central goal of our Cross-Modal Contrastive Learning (CMCL) model is to learn a new, unified embedding space where this alignment score is dramatically higher, indicating that the model has successfully learned to integrate sonic and thematic information into a single, perceptually coherent representation.

7 Conclusion

Sprint 3 successfully executed the crucial transition from implementation to deep analysis, yielding a wealth of insights and a robust foundation for model development. We have engineered a high-quality, feature-rich data corpus and subjected it to a rigorous multi-faceted analysis. This process has not only deepened our understanding of the data's inherent structure but also empirically confirmed our project's core hypotheses.

The establishment of a cross-modal alignment baseline of 0.0320 is the sprint's most significant achievement. It provides a concrete, quantifiable measure of the "perceptual gap" and sets a clear, challenging target for our subsequent modeling efforts. With this analytical groundwork complete, Project SAPPHIRE is now perfectly positioned to advance to the next phase: the synthesis of large-scale training data via LLMs and the development of our novel contrastive learning architecture.

References

[1] M. Slaney, "The perceptual gap in music information retrieval," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[2] 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.

[3] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2021.

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

[5] J. S. Downie, "The music information retrieval evaluation exchange (MIREX)," *D-Lib Magazine*, vol. 16, no. 11/12, 2010.