Sprint 2 Report — Project SAPPHIRE

Semantic and Acoustic Perceptual Holistic Integration REtrieval

Prakhar Singhal Kushal Shah Aditya Raj Singh

September 18, 2025

Abstract

This document reports the work completed during **Sprint 2** of Project SAPPHIRE. This sprint's goals were centered on establishing a robust data foundation for the project. Key achievements include the finalization of a standardized data format, the identification and selection of suitable datasets, the implementation of a comprehensive preprocessing pipeline for quality assurance, and the extraction of a rich, multi-modal feature set. The design and implementation follow the project's core perceptually-driven framework, grounding our work in established literature and the project's original proposal.

1 Sprint Objectives

The primary goals for this sprint were to:

- Finalize a standard for raw data (audio-lyrics pairing) and associated metadata.
- Identify and evaluate candidate datasets supplying paired audio and lyrics.
- Select and filter the most suitable dataset based on quality, licensing, and completeness.
- Implement a preprocessing pipeline to clean, normalize, and curate a high-quality subset.
- Extract a comprehensive suite of perceptual, acoustic, and lyrical features.

2 Data Curation and Standardization

A significant portion of this sprint was dedicated to identifying, evaluating, and standardizing the data that will form the backbone of our research.

2.1 Data Standard (Finalised)

We adopted a minimal, strict standard for each record to ensure consistency and simplify downstream processing:

Audio: A single audio file per track in WAV (lossless) or 320 kbps MP3 format. Filename: `<dataset>-<trackid>.wav`.

Lyrics: A UTF-8 encoded plain text file, timestamp-aligned where possible. Filename: `<dataset>-<trackid>.txt`.

Metadata: A JSON file containing essential fields such as `title`, `artist`, `year`, `language`, `track_id`, and `source_dataset`. Filename: `<dataset>-<trackid>.json`.

Quality Flags: Additional metadata fields for quality assessment, including `audio_snr`, `lyrics_completeness`, `alignment_available`, and `dynamic_range`.

2.2 Candidate Datasets Considered

Based on the project plan and literature review, we evaluated several candidate datasets:

- **Zalo AI Challenge 2022 (Lyric Alignment):** A dataset focused on Vietnamese song-lyric alignment. Link
- **DALI:** A large-scale dataset containing synchronized audio, lyrics, and melodies. Link
- **SINGSTYLE111:** A multilingual singing dataset useful for vocal and phonetic analysis.
- **SALAMI:** A dataset with structural annotations (e.g., verse, chorus) valuable for segmentation.
- **Million Song Dataset (MSD):** A vast corpus providing metadata and pre-computed features for a million contemporary tracks.

2.3 Dataset Filtering and Final Decision

Each dataset was evaluated on modality completeness (audio+lyrics), licensing, data quality, and relevance.
Final Decision: We have chosen to proceed with the **Zalo2022 Lyric Alignment dataset** for the initial phase.

Rationale: This dataset was selected due to its high degree of completeness, providing well-paired audio and lyric data. Its straightforward structure and accessibility (hosted on Kaggle) greatly simplify the initial data ingestion and preprocessing steps, allowing us to focus on feature extraction and model development. Other datasets like DALI and MSD remain candidates for future expansion.

3 Preprocessing and Quality Assurance Pipeline

To ensure the quality of our data, we implemented a multi-stage preprocessing pipeline.

3.1 Implemented Steps

1. **Audio Normalization:**
    - Resample all audio to a standard 44.1 kHz.
    - Convert to mono for consistent feature extraction.
    - Apply loudness normalization to a target of -23 LUFS.
    - Compute quality metrics (SNR, clipping) and flag low-quality files.
2. **Lyrics Cleaning:**
    - Strip any HTML/XML markup and normalize whitespace.
    - Ensure UTF-8 encoding.
    - Detect language and flag samples with incomplete lyrics.
3. **Metadata Consolidation:**
    - Generate a canonical JSON metadata file for each track, recording its source and quality flags.

3.2 High-Quality Subset Selection

We defined a set of criteria to filter for a high-quality subset for initial experiments:

- `audio_snr` $\geq$ `SNR_THRESHOLD`
- `lyrics_completeness` ≥ 0.9
- Successful extraction of all required features.

Note: Specific threshold values will be determined during experimental validation.

4 Feature Extraction Framework

Our Python-based extraction script performs a deep, multi-modal analysis of each audio file, producing a rich set of features that describe the song's acoustic, rhythmic, harmonic, and lyrical characteristics.

4.1 Acoustic & Timbral Features

These features describe the sonic texture and "color" of the sound.

MFCCs: Mel-Frequency Cepstral Coefficients that provide a robust fingerprint of the audio's timbre.

Spectral Centroid: Indicates the "brightness" of a sound.

Spectral Bandwidth & Spread: Measures the frequency range, corresponding to the "richness" of the sound.

Spectral Rolloff: Represents the frequency below which most of the spectral energy lies.

Spectral Flatness: Distinguishes between tonal (music-like) and noisy (hiss-like) sounds.

Spectral Flux: Measures the rate of change in the power spectrum, indicating timbral variation.

Zero-Crossing Rate: The rate at which the signal changes sign, correlated with percussive sounds.

4.2 Rhythmic Features

These features analyze the tempo, beat, and underlying pulse of the music.

Tempo: The primary speed of the music in Beats Per Minute (BPM).

Beat & Onset Information: Timestamps of the main beats and the start of individual musical events.

Rhythm Regularity: A measure of how steady and consistent the beat is.

Syncopation Index: A metric to quantify how "off-beat" or rhythmically complex the track is.

4.3 Harmonic & Tonal Features

These features describe the melodic and chordal content.

Chroma Features: A 12-element vector representing the intensity of each musical pitch class (C, C#, etc.).

Key Estimation: Predicts the song's musical key (e.g., C Major) with a confidence score.

Tonal Complexity: A measure of the harmonic richness of the track.

4.4 Lyrical Features

When lyrics are provided, comprehensive Natural Language Processing (NLP) is performed.

Structural Metrics: Counts of words, unique words, and sentences.

Readability Scores: Flesch Reading Ease to determine linguistic complexity.

Sentiment Analysis: Scores for positive, negative, and neutral emotional tone, plus a compound score.

Thematic Content: Identifies key words to provide insight into the song's topic.

Repetition Analysis: Measures lyrical repetitiveness by comparing unique vs. total word counts.

4.5 Audio Quality Metrics

These features provide an objective assessment of the recording's technical quality.

Signal-to-Noise Ratio (SNR): An estimate of the signal level compared to background noise.

Dynamic Range: The difference in decibels between the quietest and loudest parts.

RMS Energy: The root-mean-square energy, corresponding to perceived loudness.

5 Implementation and Repository Structure

All scripts for preprocessing and feature extraction are version-controlled in a GitHub repository.

```
sapphire-sprint2/
 data/
   raw/
   processed/
   manifests/
 src/
   preprocess/
   features/
   utils/
 notebooks/
 README.md
```

Repository URL: github.com/prakhar479/SAPPHIRE

6 Next Steps

The subsequent sprint will focus on:

- In-depth exploratory data analysis (EDA) and visualization of the extracted features.
- Implementing an LLM-based pipeline to generate textual descriptions from features.
- Creating standardized dataset splits (train/validation/test) for model development.

References

[1] P. Singhal, K. Shah, and A. R. Singh, "Semantic and Acoustic Perceptual Holistic Integration REtrieval: A Perceptually-Driven Framework for Music Similarity Retrieval: Project Proposal and Implementation Plan," Sept. 2025. (Project document).

[2] M. Slaney, "The perceptual gap in music information retrieval," ICASSP, 2008.

[3] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," JASA, 1933.

[4] Y. Wu et al., "SINGSTYLE111: A MULTILINGUAL SINGING DATASET WITH STYLE TRANSFER," ICASSP, 2024.

[5] J. J. Smith et al., "Design and creation of a large-scale database of structural annotations," ISMIR, 2011.

[6] T. Bertin-Mahieux et al., "The Million Song Dataset," ISMIR, 2011.