Arturo Javier Kolster Borges // N°USP: 16285381
(SCC 0802 - Prática em Ciência de Dados III)

· GROOVE GURU ·

(Audio Analysis, Music Mixing and Production Assistant Tool)

1. Abstract

This project presents an audio analysis and music production assistant tool designed to assist music producers and audio enthusiasts in the process of audio feature exploration and music creation. The system automatically analyzes uploaded audio files, extracting key musical characteristics and semantically mapping them for easier understandability. It generates specific, actionable mixing recommendations with implementation-ready parameters as well as providing stem separation, MIDI mapping, and follow-up music sample generation based on the audio sample features and the user textual prompt request.

2. Introduction

Digital audio workstations (DAWs) have democratized music production making it more accessible to creators worldwide. Yet professional mixing remains a technically challenging task, requiring expertise and deep understanding of frequency relationships and balance, dynamic processing control, spatial positioning and effects processing.Choices that can feel overwhelming without proper guidance. Traditional mixing expertise requires years of experience to understand how technical adjustments translate into musical improvements, creating a steep learning curve that often discourages newcomers.

This project addresses these challenges by developing an intelligent production assistant that analyzes audio characteristics and provides targeted guidance tailored to each track's unique properties.The system extracts 36 audio features across five analytical domains—musical, spectral, dynamic, harmonic-percussive, and spatial analysis—transforming technical measurements into human-interpretable recommendations. By integrating audio features analysis, deep learning and large language models within a unified web interface, the application serves as both a practical mixing tool and an educational resource.

The goal is to help users understand not just what to adjust, but why specific techniques benefit their particular audio material, providing long-term learning alongside immediate results.

3. Workflow Methodology

3.1. Audio Features Analysis

This foundational module performs comprehensive audio characterization by extracting 36 distinct features across five analytical domains. The module processes raw audio data (mono or stereo) and sample rate information to generate a feature vector suitable for machine learning applications in music mixing and recommendation systems.

Musical Analysis: The system identifies fundamental musical properties using established music theory algorithms. The Krumhansl-Schmuckler algorithm is used to detect musical keys (e.g., C Major, A Minor) by analyzing chroma features - a 12-dimensional representation of pitch distribution that captures harmonic content regardless of octave. Tempo analysis identifies beats per minute through onset detection, which locates moments when new musical events (notes, beats) begin in the audio signal.

Spectral Analysis: Frequency domain analysis characterizes the timbral qualities of audio. Thirteen Mel-Frequency Cepstral Coefficients (MFCCs) represent the spectral envelope and are segmented into three meaningful frequency bands: bass , midrange , and treble. Additional spectral features include centroid (brightness measure), rolloff (frequency energy distribution), spectral contrast (peak ratios), and zero-crossing rate (signal periodicity indicator).

Dynamic Analysis: Loudness and dynamic range characteristics are quantified through multiple metrics. RMS (Root Mean Square) energy provides average loudness measurements converted to decibel scale for perceptual relevance. Crest factor quantifies the ratio between peak and average levels, indicating compression levels and transient content. Dynamic range assessment uses percentile-based analysis to reveal overall loudness variation.

Harmonic-Percussive Separation: Advanced source separation techniques decompose audio into melodic and rhythmic components. The HPSS (Harmonic-Percussive Source Separation) algorithm isolates harmonic content (sustained tones, melodies) from percussive elements (transients, drums). Tonnetz features analyze harmonic relationships between pitches in a tonal space representation.

Spatial Analysis: For stereo recordings, channel correlation analysis measures spatial characteristics. Stereo width calculation quantifies how spread apart the left and right channels sound, providing crucial information for mixing decisions and mono compatibility.

Derived Features: Specialized mixing-oriented features include arrangement density (spectral complexity measure), punch factor (low-frequency impact analysis), loopability score (auto-correlation-based repetition assessment), and sub-bass energy (very low frequency content below 60Hz).

3.2 .Semantic Mapping

This module transforms raw numerical audio features into human-interpretable semantic descriptions, bridging the gap between technical measurements and musical understanding. The system processes the 36 features from Module 1 and converts them into meaningful descriptors that non-technical users can comprehend.

Feature Normalization: Raw feature values are normalized to 0-1 scales using predefined ranges.Each normalized value receives intensity labels (very_low, low, moderate, high, very_high) and semantic levels (minimal, subtle, moderate, full, strong, intense, dominant).

Semantic Interpretation: The module maps technical measurements to musical descriptors. For example, Spectral centroid (frequency brightness) becomes "mellow" when it's a low value, or "dominant." when the value is higher. This allows users to understand the track analized audio features rather than just seeing raw numbers.

Contextual Analysis: Four high-level characteristics are derived: production quality (poor/fair/good/excellent), musical complexity (simple/moderate/complex), listening context (background/casual/active_listening/critical), and emotional impact (boring/neutral/engaging/captivating). These provide immediate track assessment without technical knowledge.

Feature Relationships: The system analyzes coherence between related features. Tempo-energy coherence checks if fast tempos align with high energy levels. Spectral-timbral coherence examines frequency distribution consistency. Harmonic-percussive balance indicates whether melodic or rhythmic elements dominate.

3.3. Mixing Base Recommendations

This module generates specific audio mixing recommendations based on analyzed features, translating technical measurements into actionable production advice. The system produces prioritized suggestions with confidence scores, technical parameters, and reasoning explanations for equalization, compression, spatial processing, and temporal effects.

Recommendation Generation: The module processes normalized features to identify mixing opportunities and problems. Each recommendation includes confidence scores (0-1), priority levels (1-3), detailed reasoning based on feature analysis, and specific technical parameters for implementation. The system focuses on five core mixing domains: equalization for frequency balance, compression for dynamic control, spatial processing for stereo enhancement, temporal effects for depth and ambiance, and saturation for harmonic enrichment.

Equalization Recommendations: Frequency response optimization based on spectral analysis and tonal balance. For example, low spectral centroid values generate high-frequency enhancement suggestions with specific boost frequencies and gain amounts. Mid-range excess triggers clarity recommendations, with targeted cuts and presence boosts. The system provides exact frequency bands, gain adjustments, and filter types for direct DAW implementation.

Spatial Processing Recommendations: Stereo width and positioning enhancement using stereo field analysis. For example, narrow stereo images generate widening

recommendations through mid-side processing and stereo delay techniques. The module suggests specific delay times, processing percentages, and frequency ranges for dimensional enhancement. Panning recommendations position elements strategically across the stereo field based on instrumental content analysis.

Temporal Effects Recommendations: Reverb and delay processing adapted to tempo and arrangement density characteristics. For example , fast tempos receive tight reverb suggestions with short decay times to maintain rhythmic clarity. Slow tempos allow longer, more atmospheric processing. Parameters include reverb type selection, decay times, pre-delay settings, frequency filtering, and mix percentages tailored to track characteristics.

Dynamic Processing Recommendations: Compression and dynamic control based on crest factor and onset density analysis. For example, over-compressed material receives expansion or parallel processing suggestions. Under-controlled dynamics generate compression recommendations with specific ratios, attack/release times, and threshold settings. The system adapts suggestions based on musical style and arrangement complexity.

Priority and Confidence System: Algorithmic confidence scoring based on feature certainty and audio production principles. Priority levels help users focus on critical improvements first: essential recommendations (1) address fundamental problems, important suggestions (2) provide significant enhancement, and creative options (3) offer artistic refinement. High confidence scores indicate strong technical justification, while lower scores suggest creative possibilities.

All recommendations include implementation-ready parameters compatible with digital audio workstations. The system provides natural language explanations alongside technical specifications, making professional mixing techniques accessible to users regardless of technical expertise.

3.4. Enhanced LLM Analysis and Chatbot

This module creates enhanced mixing recommendations based on the audio features, semantic mapping and base mixing recommendations.The system uses large language models (LLM) to synthesize the audio analysis and derived features into professional mixing strategies. Then all of this context is maintained through an interactive chatbot interface for real-time consultation.

Intelligent Enhancement Process: The module analyzes the base mixing recommendations to identify priority classifications and potential conflicts with measured audio features. High-priority suggestions receive expanded technical parameters, while the system flags problematic combinations like for example brightness enhancement on already bright tracks (spectral centroid above 3000Hz). The LLM generates creative

processing ideas that complement rather than duplicate existing recommendations, organizing related techniques logically.

Feature-Driven Decision Making: Advanced prompt engineering ensures all suggestions are based on the analyzed audio features rather than on generic advice. The system evaluates tracks with thresholds, for example: frequency brightness below 800Hz indicates dull sound needing enhancement, while excessive brightness above 3000Hz requires gentle reduction. Dynamic range below 10dB suggests over-compression, while values above 25dB indicate healthy volume variation.

Interactive Chatbot: The conversational interface maintains complete analytical context throughout user sessions, providing mixing advice through natural dialogue. Users can ask questions about specific techniques ("How should I adjust the bass?") or general guidance ("What's most important to fix first?"), receiving responses based on their track's unique characteristics rather than general mixing principles.All advice includes implementation-ready parameters with clear explanations of why each adjustment benefits the specific track.

The module bridges advanced audio analysis with practical applications. Providing specific parameter suggestions compatible with standard mixing software. Technical terminology is explained in accessible language, making mixing techniques understandable regardless of the user expertise level.

3.5. AI Music Generation Prompt Module

This module generates improved prompts for the Meta MusicGen model combining the analyzed audio features with the user prompt request. Enabling the user to create new musical content that matches or complements their analyzed tracks.

Feature-to-Prompt Translation: The system extracts the key essential musical elements from the semantic mapping of the audio feature analysis. Extracting the most important features that will be understanded and interpreted effectively by the AI model.It uses the exact tempo (BPM),the musical key, the timbral characteristics, the instrumentation type, the dynamic impression, and the emotional mood.

Prompt Construction: The module creates concise 80-word prompts combining analyzed characteristics with user creative requests. Rather than generic descriptions, the system generates specific musical instructions like "energetic 140 BPM track in E major with bright synth leads and punchy drums" instead of "fast happy song." User requests are seamlessly integrated while maintaining musical coherence with reference audio.

Generation Logic: Prompt structure adapts based on dominant track characteristics. For example dominant harmonic content audio samples will focus on the melodic elements, while percussive tracks will focus on rhythmic patterns. It also integrates a simple fallback prompt when external services fail, so the music model generation can keep on working.

Users can generate complementary tracks, style variations, or inspired compositions by balancing the technical accuracy with the artistic flexibility. Combining the audio features analysis of the track with the user prompt request.

3.6. AI Music Generation Module

This module implements Metas's MusicGen models to create new musical content based on analyzed audio features and user-generated prompts. The system supports both text-to-music generation and melody-conditioned generation, enabling users to generate original compositions that maintain coherence with their reference tracks.

Multi-Model Architecture: The module supports two MusicGen variants through a unified interface. The "musicgen-small" model generates music from text prompts alone, while "musicgen-melody" incorporates reference audio as melodic conditioning. Model selection adapts to user workflow needs - pure creative generation versus reference-guided composition.

Audio Processing Pipeline: Input audio undergoes standardized preprocessing including mono conversion, 32kHz resampling, normalization, and validation checks for NaN values. The system ensures compatibility with MusicGen's expected input format while preserving audio quality and preventing generation errors from malformed data.

Intelligent Generation Process: Text prompts are tokenized and validated to ensure the model's correct functionality, preventing out-of-vocabulary errors that could break the generation. For melody-conditioned generation, reference audio is processed alongside text prompts, to guide the harmonic and melodic content of the output while allowing creative variations based on the prompt.

Quality Assurance: The module implements comprehensive tensor validation, checking data types, shapes, and numerical stability before model inference as well as audio post-processing normalization and safe file output, ensuring consistency.

The system bridges between the audio sample analysis with the AI generation, enabling the creation of new creative music follow-ups based on the audio sample inputs.

3.7. Audio Stem Separation Module

This module implements advanced source separation using Meta's Demucs models to isolate individual musical elements (vocals, drums, bass, other instruments) from mixed audio recordings.

Multi-Model Source Separation: The module supports multiple Demucs variants optimized for different use cases. The standard "htdemucs" provides high-quality 4-stem separation, while "htdemucs_6s" isolates six sources including dedicated piano and guitar

stems. The "mdx_extra" hybrid model offers enhanced separation quality for complex musical arrangements.

Optimized Processing: Audio is preprocessed to match Demucs requirements, including 44.1kHz resampling , normalization (adjusting volume levels to prevent distortion), and stereo enforcement for mono sources (ensuring two-channel output even from single-channel input). The separation process uses optimized parameters including shift augmentation (analyzing the audio from multiple starting points to improve accuracy), overlap processing (examining overlapping sections to ensure smooth transitions), and split processing (dividing long audio into manageable chunks) in order to maximize separation quality while managing computational resources efficiently.

Output Management: The system evaluates separated stems using RMS energy analysis, saving only stems with meaningful audio content (RMS > 0.005) while discarding silent or trivial components. Each stem receives fade-in/fade-out processing to prevent audio clipping. Each stem is then saved in high-quality format as well as being categorized and renamed based on its nature. For example "drums" for rhythmic elements or "melodic" for harmonic content.

The new stem separation brings on the possibility to analyze each of them separately using the app workflow. So further drum processing, vocal enhancement or melodic adjustment, can be explored based on the separated stems rather than mixed content.

3.8. MIDI Mapping Module

This module converts the stems audio content into MIDI format by extracting melodic and rhythmic elements from mixed recordings. MIDI (Musical Instrument Digital Interface) is a technical standard that represents musical information as digital data, storing note pitches, timing, duration, and velocity rather than actual audio waveforms, making it ideal for music composition and analysis. The system implements two specialized approaches: melodic transcription for harmonic content and experimental drum mapping for rhythmic elements.

Melodic Transcription: The module utilizes Spotify's Basic Pitch model (ICASSP 2022) for high-quality melody-to-MIDI conversion. The system implements optimized parameter configurations including onset threshold for note detection sensitivity, frame threshold for sustained note accuracy, and frequency range filtering to focus on musically relevant content.

Experimental Drum Transcription: The module addresses the challenge of mapping percussive elements into MIDI format through a machine learning approach, as Spotify’s model is not trained for this specific task. Since drum sounds are primarily onset-based rather than harmonic, it is a hard task map to MIDI. The system uses onset detection combined with spectral clustering to identify distinct drum types. Audio segments around

each detected onset are analyzed using MFCC coefficients, spectral centroid, and RMS energy features to characterize timbral properties. K-means clustering groups similar percussive sounds, which are then mapped to standard general MIDI drum notes (kick: 36, snare: 38, hi-hat: 42) based on average energy levels.

The MIDI mapping capability enables users to extract musical notation from their audio analysis, opening possibilities for further musical composition work using the generated MIDI data alongside the original audio features.

3.9. User Interface Module

This module provides a simple web-based interface built with Gradio. That unifies all the analysis and generation capabilities into an intuitive workflow targeted to music producers and audio enthusiasts. The system presents the complex audio analysis data in a user-friendly more readable format.

Analysis Display: After the audio is uploaded and analyzed, the interface presents results through different organized tabs. It displays the key essential features for music production (tempo, key, loudness, brightness, dynamic range) with their semantic interpretations, translating technical measurements into a more understandable format. The second tab displays the base mixing recommendations organized by priority level and confidence scores, providing targeted suggestions for the mixing effects processing chain. The third tab presents the AI-enhanced mixing strategies that identify gaps and suggest comprehensive processing chains with specific parameters and optimal processing order.

Integrated Chatbot Assistant: Following the analysis tabs, the mixing assistant chatbot initializes with the current track's analysis data. Enabling users to ask questions about the mixing processing chain , allowing further explanation of the audio sample and being able to provide new recommendations.

Additional Processing Modules: Two supplementary modules extend the workflow capabilities. One for the stems separation and MIDI mapping module, isolating individual musical elements and storing them in the Google Colab session. And another one for the music generation module, with a user input text box, the generated prompt using the LLM, and a visualization of the generated audio sample.

The interface successfully transforms the raw output data in a more user-friendly format.

4. Conclusion

This project successfully integrates advanced audio analysis with artificial intelligence to create a comprehensive music production assistant. The modular architecture transforms raw audio data into actionable guidance through feature extraction, semantic interpretation, and implementation-ready mixing recommendations. The system combines multiple state-of-the-art models, Demucs for source separation, Basic Pitch for MIDI melody transcription, an experimental drum MIDI transcription, and MusicGen for music composition, creating a complete creative workflow within an accessible web interface. With the possibility of using the workflow to re-analyze the separated stems and the music generated follow-ups resulting in a loopable application.

5. Future Implementation and Deployment

The final goal is to port the application from a notebook to a full functional robust web application. First by uploading the notebook to Hugging Face Spaces with the existing Gradio interface. For an easy deployment that will serve as a beta testing platform for user feedback collection and system refinement.The next step would be to build a full-scale web application with a robust Python backend API handling coupled with a modern responsive frontend. This production-ready platform will implement proper request handling, user sessions, and scalable processing queues to support concurrent users while providing enhanced visualization and workflow management capabilities.

This further implementation ensures that advanced audio analysis can be reached by the community of music creators, regardless of technical background or available resources.