# Harmonic Characterization of Musical Audio

Jesse Hautala
hautala.j@northeastern.edu

2024-02-06

**Abstract**

This project aims to synthesize various techniques for musical audio signal processing into a hierarchical feature extraction pipeline, including note extraction, chord detection, and topic analysis from a corpus of musical pieces. The ultimate goal is to develop a robust system for song similarity search, music categorization, and playlist generation, on the basis of intrinsic features of musical audio data.

# 1 Introduction

## 1.1 Background

Audio signal processing, particularly in music, is a challenging yet rewarding field. The ability to understand and categorize music computationally opens numerous possibilities for music recommendation, generation, and analysis. Advanced tools like Melodyne have set a high standard in the industry for note extraction and manipulation, offering detailed segmentation of notes and their harmonics alongside a suite of editing capabilities (e.g. pitch, vibrato, amplitude, and duration).

This project aims to identify and utilize existing open source tools that are well suited to hierarchical feature extraction and provide a novel and intuitive solution for navigating a music library. While Melodyne offers a remarkable capacity for note extraction and manipulation, our goal is to achieve a similar level of precision in note extraction within an open source framework, focusing primarily on the extraction aspect rather than the manipulation of audio signals. We will proceed to build higher-level features from there and hopefully discover meaningful patterns through the lens of western harmony.

## 1.2 Problem Statement

The project focuses on the extraction of musical features from complex audio signals, the detection of chords, and the analysis of musical corpora. Despite the advancements in deep learning for audio processing, achieving high accuracy in tasks like note extraction and song similarity search remains challenging, especially in the context of complex musical compositions.

As a matter of expedience, this initial work is primarily focused on western "functional harmony" in the context of a twelve tone equal-tempered tuning system. Hopefully some of the tools and techniques explored and developed in the scope of this project will also be extensible to a more comprehensive analytical framework.

# 2 Literature Review (Related Works)

## 2.1 General Overview

A few key works provide the foundation for our understanding of the state of the art:

- Y. Bayle's compilation of resources [3] offers an extensive list of deep learning techniques applied to music, only recently transitioning to unmaintained status (around 2023-12-15).
- H. Purwins et al. [17] provide a comprehensive overview of deep learning applications in audio signal processing.
- Parekh et al. [16] addresses interpretability of NN solutions, utilizing non-negative matrix factorization (NMF).

## 2.2 Reference Annotations and Tools

To inform the process of generating annotations for notes and chords, we will review extant annotation schemas, including the Reference Annotations of the Centre for Digital Music [18], particularly the *Reference Annotations: The Beatles* [19] metadata.

## 2.3 Note Extraction

- Bay et al. [2] compare multiple techniques in *Evaluation of multiple-f0 estimation and tracking systems*.
- In one of the older papers we came across, Tolonen and Karjalainen [21] present an efficient technique.
- In another older paper, also focused on efficiency, Klapuri [10] introduces the "salience spectrum".
- Barbedo at al. [1] extend Klapuri's work, in an iterative manner, that seems intuitive to me.
- Heittola et al. [8] introduce non-negative matrix factorization (or NMF).
- Bittner et al. [5] introduce deep learning methods.
- In a more recent work, Won et al. [22] use a "harmonic filter" in front of a convolutional neural net.
- Mariotte et al. [13] perform comprehensive audio segmentation, again also using NMF, but not limited to musical audio (perhaps a bit too general, but involving interesting techniques).
- Perhaps a bit specific for our purposes, Gómez et al. [7] approach fundamental extraction for flamenco vocals in the context of guitar accompaniment (so, a bit more complex than a monophonic sound source, but not as complex as arbitrary full ensemble context).

### 2.4 Chord Detection

- Mauch et al. [14] focus on improving recognition of particularly challenging chords.
- Jacoby et al. [9] address many different encodings for functional harmonic concepts (e.g. function theory, root theory, and figured bass).
- De Haas et al. [6] introduce the HarmTrace system of chord labeling.
- I have not worked out how to access the paper yet but Magalhaes and De Haas [11] report that they developed a powerful Haskell model for harmonic analysis.

## 3 Methodology (Proposed Algorithms)

This section outlines the steps and algorithms proposed for the audio signal processing project, focusing on feature extraction, similarity metrics, and applications.

### 3.1 Feature Extraction

At a high level, we aim to accomplish the following hierarchical feature extraction:

raw audio → frequencies/tones (e.g. via FFT)
→ fundamentals/notes
→ functional harmony/chords and sequences
→ topics/genres (e.g. via LDA)

1. In order to extract notes, we analyze harmonic information to identify fundamental tones. We may be able to distinguish them from overtones based on the quirks of equal tempered tuning (see Figure 1 for a visualization of the discrepancy between 12-tet and pure harmonics). We will likely require high resolution on the frequency axis, if we are to distinguish between natural harmonics and tempered tones. If high resolution Short-Time Fourier Transform algorithms are prohibitively slow, perhaps we can use QISP [20]. See Figure 2 for some preliminary test results, rendered in a custom visualization we developed for comparing spectrograms to a musical scale.
2. From note information, we can infer the key and reference pitch, to provide context for the harmonic information. These may be useful metadata in themselves for other processes, but in this pipeline we expect they will mainly be "factored out" so we can focus on functional harmony independently.
3. Given a collection of notes, we can analyze intervalic relationships to derive harmonic functions and identify chords, which will be the primary terms in our subsequent analysis. A degree of reductionism is inherent in functional analysis; our selection of how to encode chordal information will largely determine the capacity of models to detect similarities and differentiate pieces of music.
4. Finally, we can analyze all the chords and sequences thereof in a corpus, to perform topic extraction. Such high-level features can be used in subsequent calculation of similarity metrics between pieces of music and similarity-based search functions.

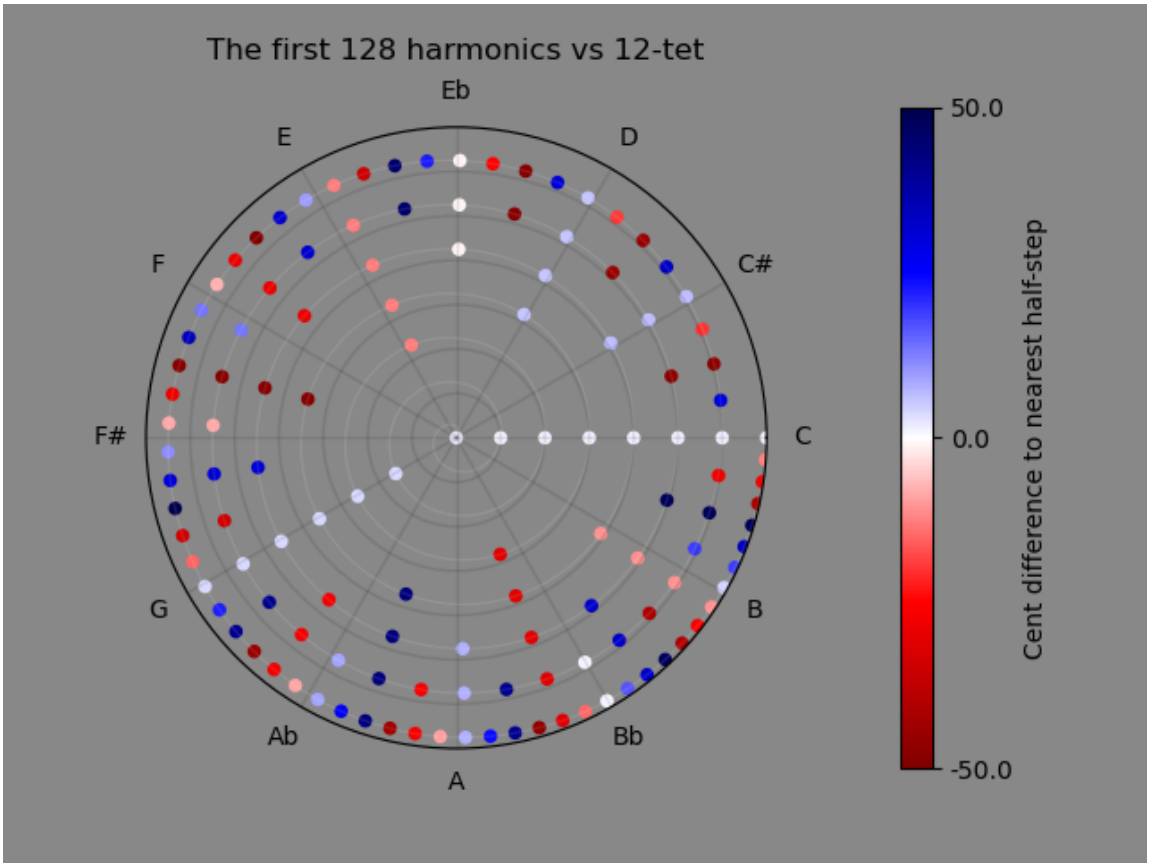


Figure 1: An illustration of the relationship between equal tempered tuning and natural harmonics

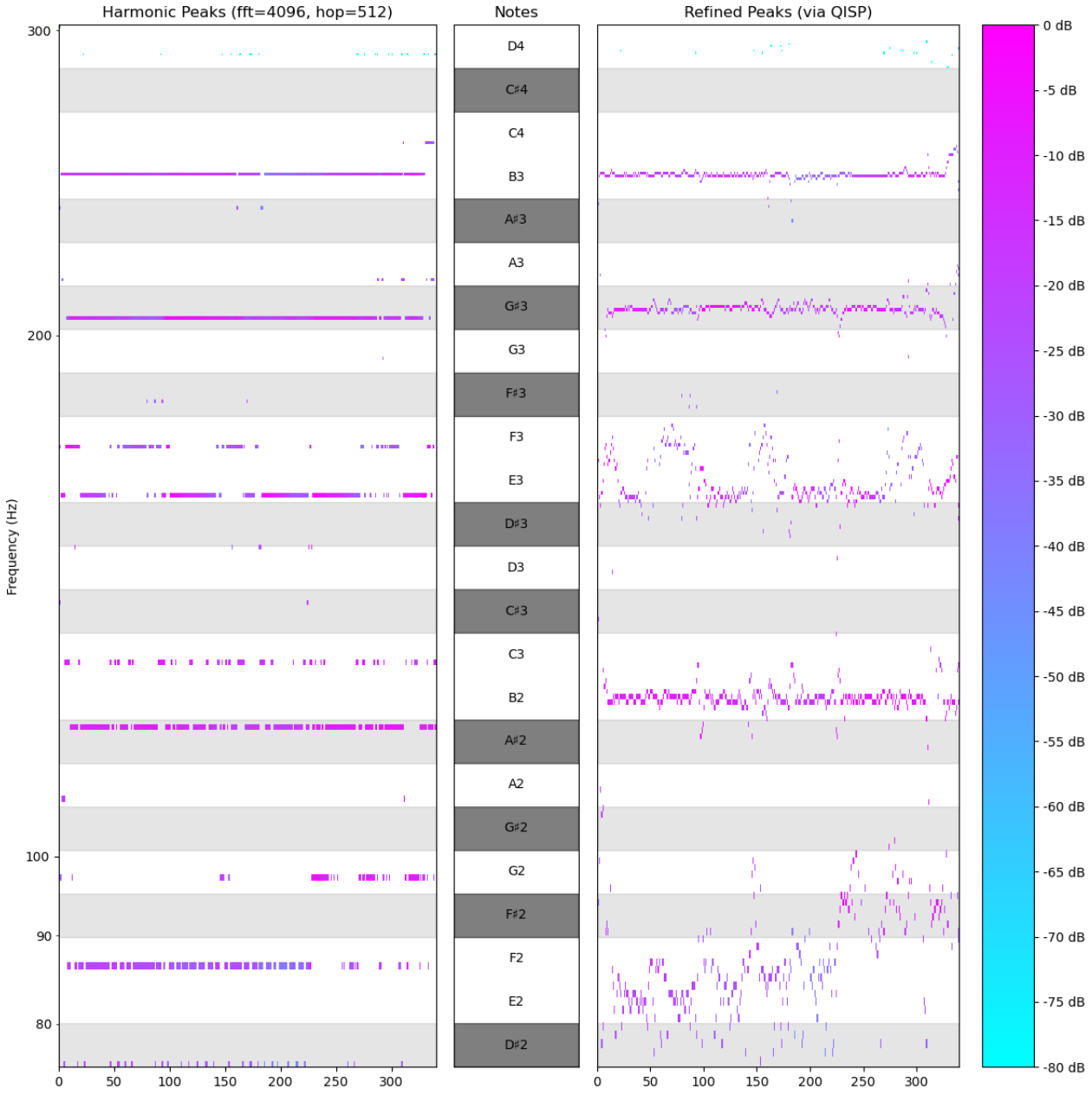


Figure 2: Lower notes (e.g. B2) are particularly susceptible to quantization error

### 3.2 Similarity Metrics

To compare different pieces of music and identify similarities, we will employ various metrics such as Cosine Similarity, TFIDF, and market basket analysis metrics (e.g. Jaccard Index).

### 3.3 Applications

The extracted features and similarity metrics can be utilized in several applications:

- Audio playback and visualization of extracted features.
- Implementation of a search by similarity/dissimilarity feature.
- Automated playlist creation based on the similarity of songs.

### 3.4 Challenges

The primary challenge anticipated is the ambiguity prevalent in various stages, from fundamental detection to chord assignment. This work will focus on functional harmony in the context of a twelve-tone equal-tempered tuning system, aiming for extensibility to a broader analytical framework in future developments.

## 4 Data Collection Plan

Assuming we can obtain access to the correct raw audio data, we would prefer to utilize community datasets such as the Million Song Dataset [4], MagnaTagATune [12], and MTG-Jamendo [15]. These datasets offer a rich collection of songs and associated metadata, which could be useful for training and evaluating the proposed models. If we are unable to obtain a coherent corpus from community sources, our backup plan is to use audio data from personal collections, consisting of original recordings in WAV format and backup copies of commercial audio CDs in MP3 format.

- The Million Song Dataset [4] is a rich collection of audio features and metadata for a substantial number of songs. Obtaining the corresponding raw audio data may prove to be a significant challenge.
- MagnaTagATune provides a collection of music and annotations that should be amenable to the aims of this project, being offered by City University of London, under the Creative Commons Attribution – Noncommercial-Share Alike 3.0 license.
- MTG-Jamendo also provides a large collection of labeled music under various Creative Commons licenses.

## 5 Evaluation Plan

We will implement various forms of ad hoc, manual validation to qualitatively assess the relevance of final results. For example, in the context of note and chord extraction we can use our ears, musical instruments, and any extant musical transcriptions to assess accuracy, but to a large extent this development plan is in the domain of unsupervised learning. Where we have ground truth, we can compare our model against random, so-called "null" models, to assess whether it produces results that are better than guessing.

# References


[1] Jayme Garcia Arnal Barbedo, Amauri Lopes, and Patrick J Wolfe. "High Time-Resolution Estimation of Multiple Fundamental Frequencies." In: *ISMIR*. 2007, pp. 399–402.

[2] Mert Bay, Andreas F Ehmann, and J Stephen Downie. "Evaluation of multiple-f0 estimation and tracking systems." In: *ISMIR*. 2009, pp. 315–320.

[3] Y. Bayle. *Deep learning for music*. 2018. URL: https://github.com/ybayle/awesome-deep-learning-music.

[4] Thierry Bertin-Mahieux et al. *The Million Song Dataset*. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011). 2011. URL: https://labrosa.ee.columbia.edu/millionsong/ (visited on 02/05/2024).

[5] Rachel M Bittner et al. "Deep Salience Representations for F0 Estimation in Polyphonic Music." In: *ISMIR*. 2017, pp. 63–70.

[6] W Bas De Haas et al. "Harmtrace: Improving harmonic similarity estimation using functional harmony analysis". In: *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*. 2011.

[7] Emilia Gómez et al. "Predominant Fundamental Frequency Estimation vs Singing Voice Separation for the Automatic Transcription of Accompanied Flamenco Singing." In: *ISMIR*. 2012, pp. 601–606.

[8] Toni Heittola, Anssi Klapuri, and Tuomas Virtanen. "Musical instrument recognition in polyphonic audio using source-filter model for sound separation." In: *ISMIR*. 2009, pp. 327–332.

[9] Nori Jacoby, Naftali Tishby, and Dmitri Tymoczko. "An information theoretic approach to chord categorization and functional harmony". In: *Journal of New Music Research* 44.3 (2015), pp. 219–244.

[10] Anssi Klapuri. "Multiple fundamental frequency estimation by summing harmonic amplitudes." In: *ISMIR*. 2006, pp. 216–221.

[11] José Pedro Magalhaes and W Bas de Haas. "Functional modelling of musical harmony: an experience report". In: *ACM SIGPLAN Notices* 46.9 (2011), pp. 156–162.

[12] *MagnaTagATune Dataset*. URL: https://mirg.city.ac.uk/codeapps/the-magnatagatune-dataset (visited on 02/05/2024).

[13] Théo Mariotte et al. "An Explainable Proxy Model for Multiabel Audio Segmentation". In: *arXiv preprint arXiv:2401.08268* (2024).

[14] Matthias Mauch and Simon Dixon. "Approximate Note Transcription for the Improved Identification of Difficult Chords." In: *ISMIR*. 2010, pp. 135–140.

[15] *MTG-Jamendo Dataset*. URL: https://mtg.github.io/mtg-jamendo-dataset/ (visited on 02/05/2024).

[16] Jayneel Parekh et al. *Listen to Interpret: Post-hoc Interpretability for Audio Networks with NMF*. 2022. arXiv: 2202.11479 [cs.SD].

[17] H. Purwins et al. "Deep Learning for Audio Signal Processing". In: *IEEE Journal of Selected Topics in Signal Processing* 13.2 (2019), pp. 206–219. DOI: 10.1109/JSTSP.2019.2908700.

[18] *Reference Annotations*. Accessed on: Feb. 6 2024. URL: http://isophonics.net/content/reference-annotations (visited on 02/06/2024).

[19] *Reference Annotations: The Beatles*. Accessed on: Feb. 5, 2024. 2019. URL: http://isophonics.net/content/reference-annotations-beatles (visited on 02/05/2024).

[20] J.O. Smith. "Quadratic Interpolation of Spectral Peaks". In: *Spectral Audio Signal Processing*. 2011.

[21] Tero Tolonen and Matti Karjalainen. "A computationally efficient multipitch analysis model". In: *IEEE transactions on speech and audio processing* 8.6 (2000), pp. 708–716.

[22] Minz Won et al. "Data-driven harmonic filters for audio representation learning". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 536–540.