Algorithmic Information Theory

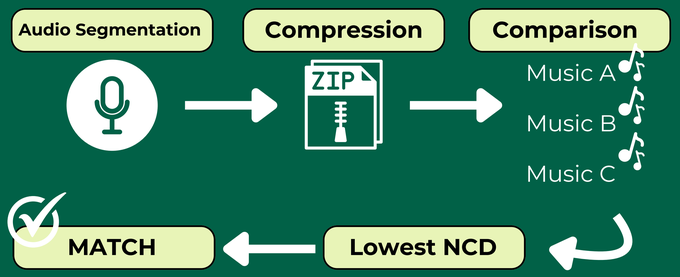
Music Identification with information distances

Introduction

This project explores a method for **automatic music identification** using the **Normalized Compression Distance** a practical, compression-based approximation of the theoretical **Normalized Information Distance** (NID), which is based on **Kolmogorov complexity**

Objective & Aim

Identify music tracks using **compressed string similarity (NCD)**



- Use **NCD** to match query x with full songs mi.
- **Test multiple compressors**: gzip, bzip2, lzma, zstd.
- Evaluate robustness with noisy query segments.
- **Dataset**: ≥25 music tracks.
- **Segment length**: ~30 seconds.

Technologies

Python

Used for orchestration, distance calculation (**NCD**), evaluation, and result analysis

Audio Preprocessing

SoX: used to trim segments, convert formats, and add noise.
Ensures uniform input for comparison

Compressors Tested

gzip, bzip2, lzma, zstd
Each tested to evaluate impact on NCD-based identification

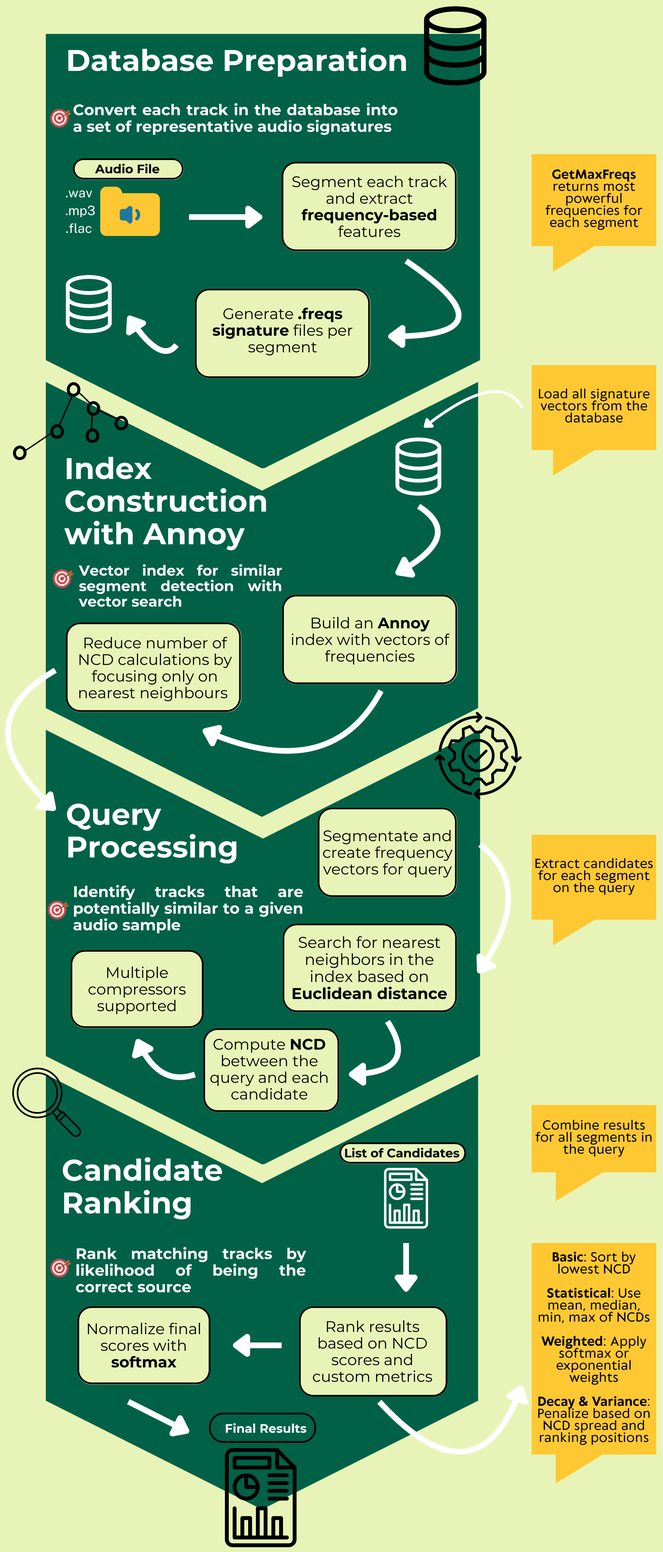
GetMaxFreqs

C++ tool to extract dominant frequency components from audio
Produces a "**frequency signature**" useful for enhanced comparison or noise robustness.

Annoy (Spotify)

Reference implementation for **Approximate Nearest Neighbor Search** (LSH)
Inspired by efficient matching strategies used for audio identification

Methodology

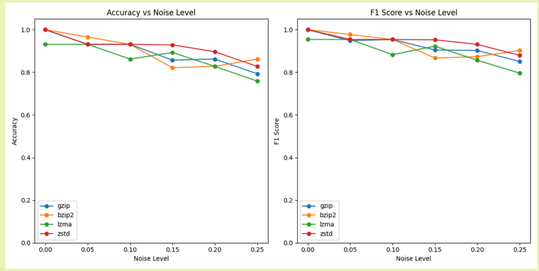


Results

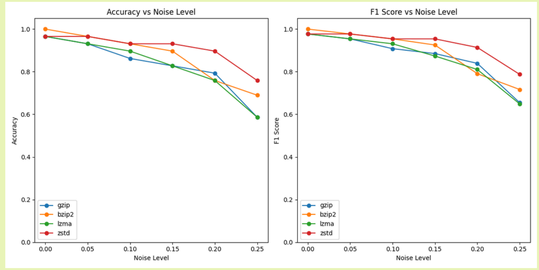
To evaluate robustness, we tested the system on queries corrupted by **white**, **pink,** and **brown** noise of increasing intensity (0.1 to 1.0). For each noise level, we measured both **accuracy** and weighted **F1-score** across all compressors.

We choose the parameters of our solution to optimize the performance on brown noise, since that type of noise has the worst effect on results.

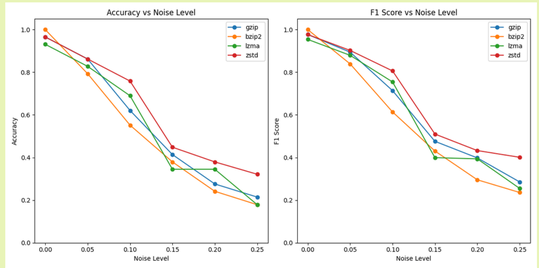
White Noise



Pink Noise



Brown Noise



Conclusion & Future Work

This work demonstrates that data **compression-based similarity**, when combined with robust ranking strategies and indexing structures like **Annoy**, can **effectively identify** music even under noisy conditions.

Future Work

- Experiment with **different compression standards**, especially on small files
- Better tuning of **LSH** i.e distance function
- Different segmenting function
- Ranking function based on **RNN**

Authors:
Diogo Marto - Nº Mec 108 298,
Diogo Pinto - Nº Mec 110 341
Ilker Atik - Nº Mec 123 947
Miguel Vieira - Nº Mec 85095

Acknowledgments:
António Neves
Diogo Pratas