Identifying musics using compressors and NCD

Information distances - Normalized Compression Distance

Guilherme Amorim 107162, José Gameiro 108840, Tomás Victal 109018

Algorithmic Information Theory

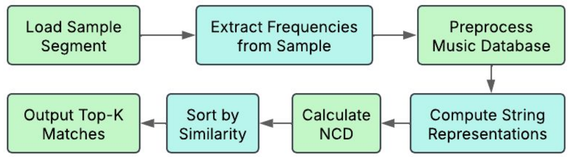
Master's Degree in Informatics Engineering

Introduction

This work aims to identify songs using the NCD (Normalized Compression Distance) metric, based on a music database and a sample provided by the user.

Implementation

- The program starts by reading the sample file an extracting a portion of the sample
- Then the most and least dominant frequencies are obtained
- The most and least dominant frequencies are obtained also for all the songs in the database
- Convert frequency vectors to comparable strings
- Apply compression with one of the one of the following compressors gz, bz2, xz, zstd or lzma
- Calculate the NCD with the sample and all the songs in the DB, sort the results and print them



Data Used

Our group used a total of 30 songs with times ranging between 2min30s and 4min30s, and with the following styles: Rap, Punk-Rock, Grunge, Fado, Pimba, Pop, Jazz, Alternative and Hard-Rock.

In terms of samples, the ones provided by the class professors and also some from other songs with different durations (20s, 40s and 60s).

Used Brown, White, Pink and Green noises

Experimental Analysis

Multiple tests were made including:

- The effect of adding noise to the samples;
- The effect of the size of the sample;
- Identify similar musics from the same artist and album

Results

Tests on musics by the same artist compressor show no clear correlation with higher NCD (used the music hitchin_a_ride_green_day with a sample of 20 seconds)

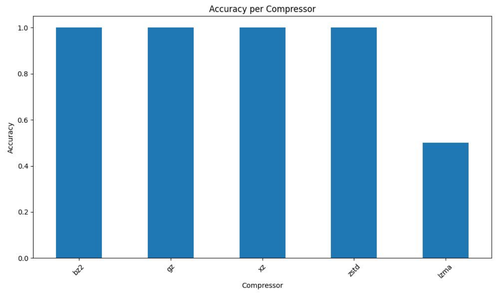
Compressor gz

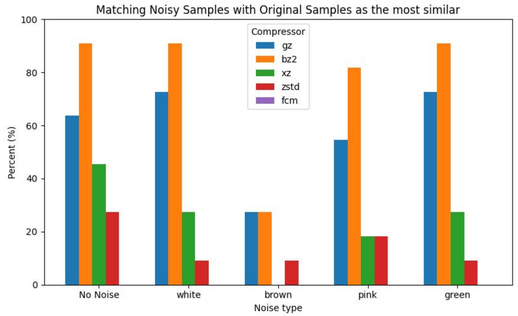
Rank	Most Similar Samples	NCD Score
1	hitchin_a_ride_green_day	0.9201
2	bounce_System_of_a_down	0.9842
3	scattered_green_day	0.9857
4	aint_that_a_kick_in_the_head_dean_martin	0.9859

Compressor bz2

Rank	Most Similar Samples	NCD Score
1	hitchin_a_ride_green_day	0.9281
2	bounce_System_of_a_down	0.9644
3	violent_pornography_systm_of_a_down	0.9688
4	american_idiot_green_day	0.9691

Results changing the sample size with the values 20, 40 and 60 seconds and adding noise to the samples. All the songs were used to create different samples with different sizes and all the compressors were applied to each sample created.





Images Test

As we had already tested images in the previous work, we decided to try again, this time using NCD and the various compressors mentioned in the poster. Below is a portion of the table showing the results for 41 subjects, with 10 images per subject. The last row represents the total, and the ideal total is 410 correctly identified images. For each person, the model needs to correctly identify 10 images.

Person	gz	bz2	xz	zstd	lzma
p1	7	10	3	7	2
p2	7	9	4	7	8
p3	7	4	6	7	5
Total	208	226	126	179	123

Conclusion

This work shows that it is possible to identify music using Normalized Compression Distances (NCD), with good results when matching a sample to the database. However, similarity detection between songs by the same artist or album was not consistently reliable. For images, BZ2 compression showed some ability to identify individuals, but the results were not fully reliable.

Github Source

https://github.com/zegameiro/TAI_Projects

universidade de aveiro

deti
departamento de eletrónica, telecomunicações e informática