EMERGING TECH CONFERENCE – Edge Intelligence

Volume 02, 2023, Page 23 – 24

Proceedings of Emerging Tech Conference:
Edge Intelligence 2023

An Audio Fingerprinting Approach Based on the 2DFT for Byzantine Hymn Recognition

Dimitrios Kampelopoulos[1], Lazaros Moysis[1], Konstantinos Karasavvidis[1], Achilles D. Boursianis[1], Sotirios K. Goudos[1], Spyridon Nikolaidis[1]

[1] *Aristotle University of Thessaloniki, Thessaloniki, GR*
({dkampelo, lmousis, kokarasa, bachi, sgoudo, snikolaid} @physics.auth.gr

Abstract

An audio fingerprinting technique is proposed, in this work, for the challenging task of Byzantine hymn recognition. This approach involves the extraction of speed and scale invariant fingerprints, which is crucial to the characteristics of the application, as well as an evaluation process in order to reliably match a query to the correct hymn of the database.

1 Introduction

Audio fingerprinting is a technique commonly applied in voice recognition, sound event detection, musical information retrieval and cryptography [1]. In the field of music, there are established methods to recognize songs in a timely manner, but the task is mainly to match a distorted recording of a track to the official track in the database. However in Byzantine music the hymns are primarily performed live by different chanters and in varying tempo and scale. So for hymn recognition, it is important to not only eliminate the noise but also to be resistant to these kinds of variations.

In this work, the focus was given on the extraction of time and scale invariant fingerprints through a process involving the 2-dimensional Fourier Transform (2DFT), originally proposed by [2] for cover song identification, that was customized to the Byzantine music's characteristics. Also, a statistical polling technique was introduced as a final step and the whole process was optimized in order to achieve reasonable querying times. The method was tested on a hymn database containing 359 Byzantine hymn executions by multiple chanters.

2 Method

The first part of the process is fingerprint extraction. The Constant-Q Transform (CQT) was applied on the raw audio data in order to extract a spectrogram with a logarithmic frequency axis divided into the musical scales. This logarithmic transformation makes the tone changes linear and are eliminated in a later part of the process. The next step is performing adaptive thresholding on the CQT which is done by applying a median filter. The last step is to apply the 2DFT on the filtered spectrum in order to extract a sequence of Fourier coefficients. This sequence is the actual fingerprint that will be used for the recognition. This process is performed on each track of the hymn database and the resulting fingerprints are precalculated and stored.

To compare two hymn executions the process involves the calculation of a distance metric. This distance is calculated by the similarity matrix of two fingerprint sequences. First, the mean of the similarity matrix is calculated and then a Gaussian filter is applied, in order to create distinct diagonals. The distance is calculated as the sum of the three main diagonals. As a next step, the track is resampled in varying sampling frequencies and the one with the minimum distance is selected. The correlation metric was chosen for the similarity matrix distance calculation which was found to produce better results than the Euclidean distance used in [2].

So, in order to match a query audio to a hymn in the database, the distance is calculated between the query and every hymn execution of the database. The executions exhibiting the minimum distance are the match candidates and usually the minimum distance corresponds to the same hymn. However, an additional step was introduced in which a probability is associated to every match candidate, analogous to the inverse of the distance. As a result, the choice is not always linked to the minimum distance, but the most probable hymn found in the match candidates.

3 Results

For the purpose of this work, a database of 359 executions were recorded on a set of nine different hymns by different chanters in varying tempo and speed. To evaluate the proposed method, each execution of the database was used as a query and was compared to the rest. The process was, also, optimized in terms of speed, so that a single query can be performed in a few seconds depending on its duration. A parametric execution of the algorithm followed which determined the best combination of parameters for this specific application. With this choice of parameters the algorithm exhibited 99.44% accuracy with the statistical polling method versus 99.16% with the original minimum distance one.

4 Conclusion

With the proposed method, it was possible to make a clear distinction between different executions of the same hymn performed by different chanters, in varying tempo and scale. The algorithm exhibits high accuracy and improved results by introducing a statistical polling step.

References

[1] Cano, P., Batlle, E., Kalker, T., Haitsma, J. (2005). A review of audio fingerprinting. Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology, 41(3), 271–284. https://doi.org/10.1007/s11265-005-4151-3

[2] Seetharaman, P., Rafii, Z. (2017). Cover song identification with 2d Fourier transform sequences. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp.2017.7952229