Corinna Mackenow

A benchmark and evaluation of motif discovery on audio books

Bachelorarbeit Exposé

First Supervisor: Dr. rer. nat. Patrick Schäfer
Second Supervisor: tba.

Institute for Computer Science

Humboldt-Universität zu Berlin

Germany

Table of Contents

1 Introduction

When working with time series, motif discovery is one of the most common problems. In simple terms motif discovery is the problem of finding shorter sequences that repeat themselves within the time series. Since motif discovery is an unsupervised learning problem, it is especially practical for exploratory data analysis and is used in a lot of fields like computational biology and biochemical engineering [1] or seismic signals [2]. One of the hurdles when it comes to motif discovery is the lack of ground-truth data. The goal of this thesis will therefore be to establish more ground truth data and evaluate the effectiveness and applicability of motif discovery algorithms to audio books. Applying time series analysis algorithms to data extracted from human language is especially interesting since it is such a widely common and relatively easy to comprehend type of data, therefore opening up the possibilities of providing results that are interesting to a wide range of scientific fields. However, human language can also be very complex and diverse, potentially making it more challenging to work with such recordings.

2 Background

This section provides a brief rundown of important definitions, since there are some inconsistencies throughout literature.

Definition 2.1 Time Series: A time series $T = (t_1, t_2, ..., t_n)$ is an ordered sequence of real-values $t_i \in \mathbb{R}$ with length n.

Since time series can be very large, it is often helpful to look at a subset of values.

Definition 2.2 Subsequence: A subsequence $T_p = (t_p, ..., t_{p+m-1})$ is a subset of m values from T starting from position p.

To compare subsequences with each other, a distance function is necessary.

Definition 2.3 z-normalized Euclidean distance: Given two subsequences T_p, T_q with mean μ_p, μ_q, standard-deviation σ_p, σ_p and length n, their z-normalized Euclidean distance (z-ED) is defined as

$$z - ED(T_p, T_q) = \sqrt{\sum_{m=0}^{n-1} \left(\frac{t_{p+m} - \mu_p}{\sigma_p} - \frac{t_{q+m} - \mu_q}{\sigma_q}\right)^2} \quad [3]$$

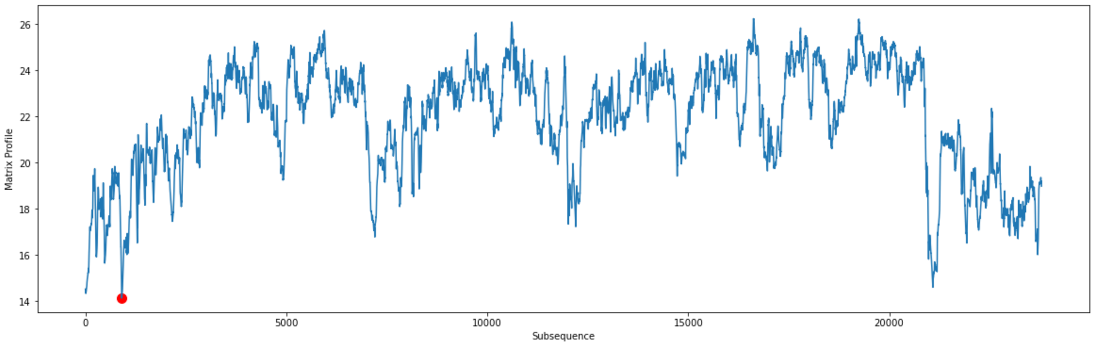

Figure 1: The global minimum of a matrix profile indicates the most conserved pattern (motif). Source: https://stumpy.readthedocs.io/en/latest/Tutorial_AB_Joins.html

Using this distance function, subsequences that are similar to each other can be detected.

Definition 2.4 r-Match: Two subsequences T_p, T_q (both with length n) of T are called an r-match if and only if they share less than $n/2$ common offsets of T and $z-ED(T_p, T_q) \leq r \in \mathbb{R}$.

In order to define k-Motiflets, the extent of a Motif set needs to be defined, first.

Definition 2.5 Extent: Given a time series T and a set S of subsequences of T, each with length n, the extent of S is the maximal pairwise distance of elements from S:

$extent(S) = max_{(S^{(1)},S^{(2)}) \in SxS}(z-ED(S^{(1)},S^{(2)}))$ [3]

With those definitions, the k-Motiflet can be introduced next.

Definition 2.6 Top k-Motiflet: Given a time series T, cardinality $k \in \mathbb{N}$ and length n, the top k-Motiflet of T is the set S with $|S| = k$ subsequences of T of length n for which the following holds: All elements of S are pairwise d-matching, with $d = extent(S)$, and there exists no set S' with $extent(S') < extent(S)$ also fulfilling these constraints. [3]

A Pair Motif can also be represented as the global minimum of a matrix profile
(Figure 1) of a time series.

Definition 2.7 Matrix profile: A matrix profile P of time series T is a vector that stores the z-normalized Euclidean distances between each subsequence of T of length n and its nearest neighbor [4].

3 Related Work

Motiflets [3] is a motif discovery algorithm, that - unlike most motif discovery algorithms - takes the number of occurrences of a certain motif as the central parameter instead of the maximal distance between the motif's occurrences, making it more intuitive and easy to use and providing more accurate results.

In their paper The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing [5] Eyben et al. present a ”basic standard acoustic parameter set for various areas of automatic voice analysis” containing a total of 88 features. Their work is publicly available with the openSMILE toolkit [6], which will be the base to isolate the speakers voice in the audio book time series used in this project.

4 Objectives

This project's goal is to create a benchmark with ground truth data for motif discovery, extracted from audio books. Audio books provide a good source for this, since a single book results in multiple hours of audio files, therefore providing a large amount of data. The correctness of results based on human language tend to be easier to evaluate and analyse than if we were to use other sources (like animal noises or acceleration data). As opposed to songs, audio books do not contain different instruments and effects that could skew the motifs. When reading a text, we assume that a person pronounces reoccurring words in similar ways every time they appear but the same thing can't be said about singers, since they often manipulate the pronunciation to match the melody and add flourishments to their words. This leads to the hypothesis that audio books could produce more accurate results for motif discovery.

A few problems that need to be addressed while creating the benchmark are (a) isolating the speakers voice, (b) dealing with silence to avoid having that be the most common motif, (c) synchronizing the words and audio with each other and potentially (d) filtering out stop words like ”the”, ”a” or ”it”. After creating the benchmark, the Motiflets algorithm will be applied in order to evaluate and plot the results. Afterwards, an opportunity to further extend the benchmark would be using different books to increase variety or using multiple speakers to evaluate whether or not different vocal tones, mother tongues or accents have an effect on the results. Both cases bring up the challenge of detecting motif-sets between two time series, which needs to be dealt with sensibly.

4.1 Data Collection

There is a plethora of audio books online to choose a fitting sample from. Alternatives to that are reading the text and creating new recordings or utilizing AI technologies that are capable of producing human-like audio files from a text. In either case it is important to ensure that the audio and text files match exactly and neither file contains added filler words like ”Chapter x” that aren't present in the other.

4.2 Data Organization

The audio files need to be converted into csv files so that the algorithm can work with them, which can be done in python using Scipy [7]. The corresponding text does not need to be converted since it is already in plain text.

4.3 Data Preparation

One of the possible problems that need to be solved within the course of this thesis is silence. In most audio recording of human language, silence will be the most common motif. This can be dealt with proactively by removing the silent parts of audio beforehand or coping with the silent motifs afterwards. The speaker's voice needs to be isolated within the recording, which can be done by experimenting with different acoustic parameters like the 88 audio features mentioned above. Another problem is accurately identifying and matching the words to the correct timestamp of the audio file to ensure each motif can be labeled correctly. To compare motifs found in audio files of the same book, read by different speakers, it might be necessary to match the lengths of the words by ensuring they are speaking at the same pace in the first place or editing the files afterwards.

4.4 Benchmarking and presentation

Once the data has been prepared for further use, the Motiflets algorithm will be applied and the results can be evaluated. A valid result is the correct identification of the most frequent words or phrases. Filtering out stop words offers a result that is more tailored to the specific book and it's storyline. This might also be achievable by searching long motifs instead, which can be a more efficient approach, since filtering out stop words in recordings is not trivial.

5 Results

The main programming language of this project will be Python. The results will be visualized using the plotting library Seaborn [8]. Everything will be documented in a Jupyter Notebook [9] and published in a GitHub repository.

References

[1] Mark Philip-Walter Styczynski. 2007. Applications of motif discovery in biological data, derived from `https://dspace.mit.edu/handle/1721.1/38976`.

[2] M Ashraf Siddiquee et al. 2019. SeiSMo: Semi-supervised Time Series Motif Discovery for Seismic Signal Detection. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 99–108.

[3] Patrick Schäfer, Ulf Leser. 2022. Motiflets - Simple and Accurate Detection of Motifs in Time Series, `https://www.vldb.org/pvldb/vol16/p725-schafer.pdf`.

[4] The Matrix Profile, `https://stumpy.readthedocs.io/en/latest/Tutorial_The_Matrix_Profile.html`. Last accessed Apr 25th 2023.

[5] Florian Eyben et al. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing, `https://ris.utwente.nl/ws/portalfiles/portal/6721593/07160715.pdf`.

[6] openSMILE 3.0 GitHub `https://github.com/audeering/opensmile/releases/tag/v3.0.0`. Last accessed Apr 22th 2023.

[7] scipy.io.wavfile.read `https://docs.scipy.org/doc/scipy/reference/generated/scipy.io.wavfile.read.html`. Last accessed Apr 26th 2023.

[8] Seaborn: statistical data visualization `https://seaborn.pydata.org/`. Last accessed Apr 26th 2023.

[9] Project Jupyter `https://jupyter.org/`. Last accessed Apr 26th 2023.