MODIS: an audio motif discovery software

Laurence Catanese, Nathan Souviraà-Labastie, Bingqing Qu, Sébastien Campion, Guillaume Gravier, Emmanuel Vincent, Frédéric Bimbot

MODIS: an audio motif discovery software

Laurence Catanese[1], *Nathan Souviraà-Labastie*[2], *Bingqing Qu*[2], *Sebastien Campion*[1], *Guillaume Gravier*[2], *Emmanuel Vincent*[3], *Frédéric Bimbot*[2]

[1] INRIA Centre de Rennes - Bretagne Atlantique – Campus de Beaulieu, 35042 Rennes, France
[2] IRISA/CNRS UMR 6074 – Campus de Beaulieu, 35042 Rennes, France
[3] INRIA Nancy - Grand Est, 54600 Villers-lès-Nancy, France

{laurence.catanese,sebastien.campion}@inria.fr,
{nathan.souviraa-labastie,bingqing.qu,guillaume.gravier,frederic.bimbot}@irisa.fr

Quotation : *"Everything that happens once can never happen again. But everything that happens twice will surely happen a third time."* Paulo Coelho, *The Alchemist (*after an Arab proverb).

Abstract

MODIS[1] is a free speech and audio motif discovery software developed at IRISA Rennes. Motif discovery is the task of discovering and collecting occurrences of repeating patterns in the absence of prior knowledge, or training material. MODIS is based on a generic approach to mine repeating audio sequences, with tolerance to motif variability. The algorithm implementation allows to process large audio streams at a reasonable speed where motif discovery often requires huge amount of time.

Index Terms: word discovery software, pattern matching, unsupervised learning, dynamic programming

1. Introduction

The task that consists in mining useful information from audio data grew significantly these past years given the expansion and the velocity of digital media creation and diffusion. Recent work have already studied the motif discovery problem. A pioneering work in word discovery is presented in [1] and an algorithm is proposed in [2] to automatically extract repeating patterns in multimedia stream. We have developed a generic approach, introduced in [3] and [4], to mine repeating audio sequences, with tolerance to motif variability.

The open source software MODIS is based on this algorithm from which several implementation improvements have been achieved in order to speed-it up. We first present the software and then discuss current and possible application as well as future work.

2. General overview of MODIS

The general input/output system of MODIS is presented in Figure 1. The software only requires an audio features file input. The feature type choice depends on the targeted task. In particular, the software can handle for example classic mel frequency cepstral coefficient (MFCC) and posteriorgrams.

The result of the motif discovery process is a library of motifs. Each motif is a set of occurrences, each occurrence being represented by a starting and an ending point relative to its location in the input stream. Motifs properties like their identity and their number are unknown as well as the number of occurrences, their length and their locations.

MODIS is developed in C. It has been used under Linux and has never been compiled under other operating systems. As shown on Figure 1, MODIS requires the SPRO library[2] installed on the system. It is an open source library widely used to handle audio feature buffers. OpenCV[3] that provides useful image-video processing tools and Audioseg[4] that provides silence segmentation, are optional libraries.

3. Algorithm

A summary of the global internal architecture is shown in Figure 2. The process is based on a general framework called *seeded* discovery. This technique comes from the ARGOS segmentation framework proposed in [2]. After a fixed length fragment of a motif is found, the final occurrences are grown through a match extension procedure. Motifs discovered are stored and a library of motifs is incrementally built. Instead of looking for a seed match in the entire stream, similarities are searched in this library at first, permitting long term occurrences. The search for local similarities between the seed block and a buffer is performed using variations of the popular dynamic time warping (DTW) procedure. The technique used is called "segmental locally normalized dynamic time warping" (SLNDTW). It aims at detecting local alignments relaxing the boundary constraint of the classical algorithm since starting points of potential matches are unknown.

In order to improve the robustness to the variability of speech patterns, a template matching technique based on the comparison of self-similarity matrices (SSM) is used on speech sequences. This technique allows confirming or invalidating the similarity between speech patterns complementary to a DTW-based matching and thus provides for speech related task, a speaker independent system. This technique has shown some benefits although it requires more computation time.

Among other optional features, two main parameters of the audio motif search are left to be defined by the user:

- The length of the seed and the near-future buffer in which the search is performed into can be modifiable. The length of the seed has to be set with respect to the expected size of the output motifs. The computation time generally grows with the buffer sizes.
- The threshold used for the similarity detection corresponds to the value below which two sequences are considered

[1] MODIS: https://gforge.inria.fr/projects/motifdiscovery/

[2] Spro: http://www.gforge.inria.fr/projects/spro
[3] Opencv: http://opencv.willowgarage.com/wiki/
[4] AudioSeg: https://gforge.inria.fr/projects/audioseg/

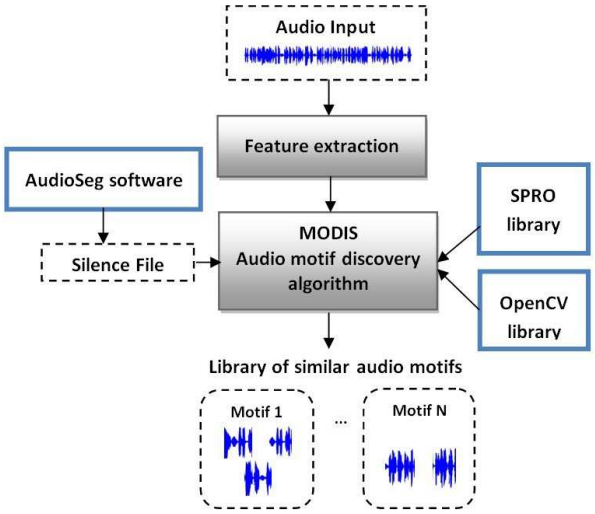

Figure 1 :MODIS software architecture. *It processes an input feature file generated from an audio file. The released output indicates the portions of audio segments that are considered similar by the system. MODIS relies on SPRO and OpenCV (optional) library. A file generated by Audioseg software can optionally be used as an input to process the stream without silences.*

similar. Changing this value affects precision and recall scores.

4. Enhanced implementation

The algorithm is implemented in a modular way, so that the functions, specific for one task, can be replaced or completed for potential future developments like other form of DTW for instance. It also permits to identify its critical stages that make the motif discovery a time-consuming task, and therefore optimize the code. Some functions are set in macro type, and allocation memory management is improved. However, self-similarity matrices use a lot of pre-written functions from OpenCV and so are difficult to optimize.

Some additional functionalities are implemented to improve the software performance such as the down-sampling of the input sequence. The processing of a low-resolution representation of the input stream enables to save a huge amount of computation time. The use of this approximate technique is extremely advantageous to discover patterns with a low variability like near duplicates.

Experiments done in [3] are used as a baseline to compare results of the algorithm before and after the final implementation. Those comparisons show significant benefit in terms of execution time without affecting performances. The best improvement is obtained for French phone posteriorgram with SSM calculation whose time fall from more than 26 hours to 8 hours.

5. Applications

Multimedia contents keep evolving and diversifying in the way public can access to it. Despite this diversity, content redundancy seems to be an inherent feature of multimedia, partly as a consequence of practical and accessible production techniques. In audio, it is widely present at different scales and for different type of sounds, for instance chorus in songs, spoken keywords, or jingles in radio or TV on a one day scale.

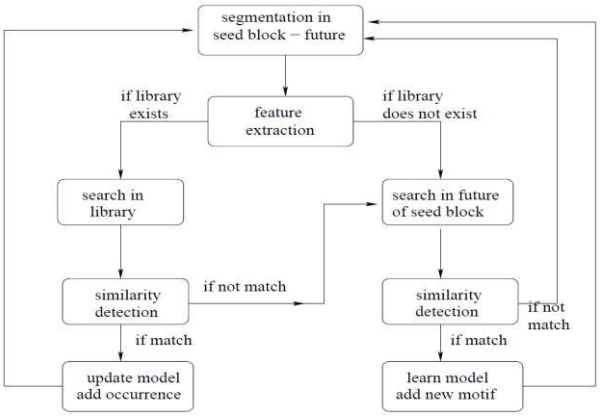

Figure 2: Audio motif discovery algorithm design. *At each step, the seed block is compared with the models in the library and, if not found, is searched in its near future. At the following step, the pair "seed block-future" shifts of a seed block length along the audio stream and the process is iterated.*

As far as audio is present in almost every multimedia data, it opens a large scope of potential application to MODIS. Currently, it is employed for near duplicate discovery in radio stream, TV show structuring, *e.g.* [5], or old movies and series remastering. Other applications can be considered such as multiple spoken-document summarization, *e.g.* [6], topic clustering by keywords extraction, *e.g.* [7], or source separation informed by redundancy, *e.g.* [8]. More innovative applications can be imagined such as data compression, as suggested in [9], or audio source localization, both informed by redundancy.

6. Conclusions and future work

This paper presents the MODIS open source software. An efficient implementation addresses the issue of computation time, and an adjustable framework allows to discover any kind of audio motif. It also permits to the scientific community to contribute to its different modules, various applications are thus targeted. The software is downloadable on https://gforge.inria.fr/projects/motifdiscovery/.

In a new project, we plan to increase the variety of possible distortions to better detect high variability motif. In this way, we expect MODIS to handle more sophisticated contents produced by successive mixing process, such as movies. Targeted motifs will be then sound effects, ambiance or musical leitmotifs. Another track to examine is to modify the similarity detection stage of the algorithm with indexing technics such as in [10].

7. Acknowledgments

The authors would like to acknowledge Armando Muscariello for his thesis works on which this software is based. This development was partly funded by OSEO the French State Agency for Innovation, under the Quaero project.

8. References

[1] A. S. Park, “Unsupervised pattern discovery in speech: applications to word acquisition and speaker segmentation”, Ph.D dissertation, MIT, Cambridge, MA, 2006.

[2] C. Herley. “ARGOS: Automatically extracting repeating objects from multimedia streams”. IEEE Transactions on Multimedia, 2006.

[3] A. Muscariello, et al., “Unsupervised motif acquisition in speech via seeded discovery and template matching combination”. Accepted for publication in IEEE Trans. on Audio, Speech and Language.

[4] A. Muscariello, et al., “An efficient method for the unsupervised discovery of signalling motifs in large audio streams”, International Workshop on Content-Based Multimedia Indexing, Jun 2011, Madrid, Spain.

[5] A. E. Abduraman, et al., “An unsupervised approach for recurrent tv program structuring”, Proceedings of the 9th international interactive conference on Interactive television, June 29-July 01, 2011, Lisbon, Portugal.

[6] X. Zhu, et al., “Summarization multiple spoken documents: finding evidence from untranscribed audio”, 47th annual meeting of the ACL and the 4th IJCNLP and of the AFNLP, 2009, 549-557.

[7] M. Dredze, A. Jansen, G. Coopersmith and K. Church, “NLP on Spoken Documents without ASR”, Empirical Methods in Natural Language Processing, 2010.

[8] A. Liutkus, Z. Rai, R. Badeau, B. Pardo, and G. Richard. “Adaptive filtering for music/voice separation exploiting the repeating musical structure”. In ICASSP, 2012

[9] M. Moussallam,"Représentations redondantes et hiérarchiques pour l'archivage et la compression de scènes sonores" Thèse de Telecom ParisTech (French december 2012).

[10] A. Jansen, et al., “Towards spoken term discovery at scale with zero resources”, in Interspeech, 2010.