Audio Similarity Retrieval Engine

Pavel Jurkas
Masaryk University
Brno, Czech Republic
xjurkas@fi.muni.cz

Milan Štefina
Masaryk University
Brno, Czech Republic
xstefina@fi.muni.cz

David Novak
Masaryk University
Brno, Czech Republic
xnovak8@fi.muni.cz

Michal Batko
Masaryk University
Brno, Czech Republic
batko@fi.muni.cz

ABSTRACT

This paper briefly describes an audio similarity retrieval engine included in the MUFIN project. The engine uses low-level audio descriptors defined by MPEG-7 standard for calculation of similarity measure between audio recordings. The core of the engine is implemented in Java with the use of the MESSIF framework that provides support for metric-based indexing and searching. The presentation layer of the engine is provided by the MUFIN interface.

Keywords

similarity search, metric space, audio, MPEG-7, MESSIF

1. INTRODUCTION

The presented audio similarity retrieval engine is based on the data concept of metric spaces. The specific space for this purpose is defined by descriptors extracted from audio recordings and a similarity measure on these descriptors. Indexing techniques and algorithms for processing similarity queries have already been implemented within the MUFIN project [1] with the aid of the MESSIF framework [2].

2. AUDIO DESCRIPTORS

MPEG-7 Audio is a standard describing audio metadata [5] that can be generally divided into low-level and high-level. Low-level audio descriptors are more general and applicable on any type of input audio signal. They can be used for various tasks and use cases. High-level audio descriptors are more specifically focused and are not used in the engine.

Low-level audio descriptors are typically formed by time series of samples computed from original audio signal. Samples are located in equidistant points in time (typical sampling rate is 100 Hz). The engine uses the following four low-level descriptors from the MPEG-7 Audio standard:

- Audio Power (AP) – one-dimensional sequence capturing the power of original audio signal, see Figure 1b;

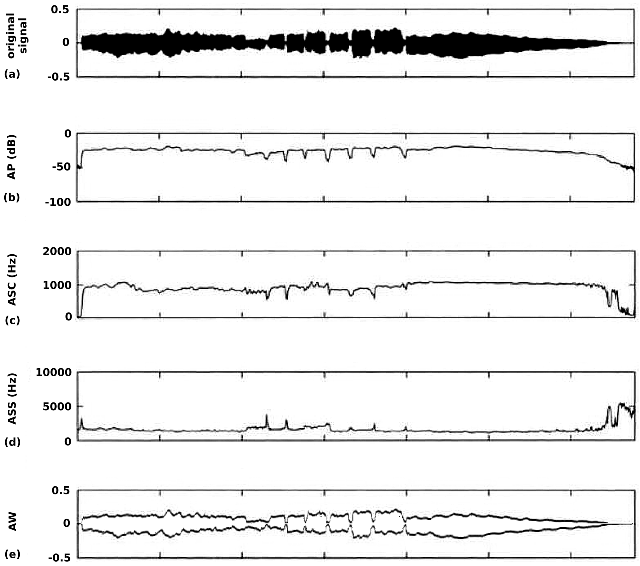


Figure 1: Example of the used MPEG-7 audio descriptors: (a) original 44.1 kHz sampled audio signal; (b) Audio Power; (c) Audio Spectrum Centroid; (d) Audio Spectrum Spread; (e) Audio Waveform.

- Audio Spectrum Centroid (ASC) – one-dimensional sequence which capture the central frequency in original audio signal, see Figure 1c;
- Audio Spectrum Spread (ASS) – one-dimensional sequence which capture the frequency spread in original audio signal, see Figure 1d;
- Audio Waveform (AW) – two-dimensional sequence capturing the minimal and maximal value of amplitude of original audio signal, see Figure 1e.

3. SIMILARITY MEASURE

A metric space $\mathcal{M} = (\mathcal{U}, d)$ is defined by a domain $\mathcal{U}$ of objects and a metric function $d : \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}$ that determines the (dis)similarity between pairs of objects. In our case, the domain consists of particular audio descriptors extracted from the original audio recordings. The Euclidean distance L_2 was used as a metric function for AP, ASC and ASS descriptors. AW descriptor can be viewed as two one-dimensional sequences – one for minimum values and one for maximum values of amplitude, see Figure 1e. L_2 distances

SISAP'10 September 18–19, 2010, Istanbul, Turkey

are computed for both one-dimensional sequences separately and arithmetic mean is used as a final distance.

We have used the following data concept [3] in order to be able to search subsequences in the audio database. Each audio descriptor sequence in the database is split into overlapping subsequences of constant length w using the sliding window approach. Subsequences are associated with an offset k determining their position in the original sequence. M-Index is used for indexing all these subsequences [4].

The query sequence is split into disjoint subsequences of the same length w associated with index l (tail of the query sequence shorter than w is ignored). Separate nearest neighbor queries (*kNN*) are processed for each of these subsequences. The results of these queries consist of matching pairs of query-data subsequences that form a candidate set of possibly similar audio recordings.

This candidate set has to be refined. For each matching pair, we have corresponding offsets k and l determining how the query sequence matches the data sequence. Final similarity is computed as the distance between the longest overlapping subsequences of the query and the matching data sequence and is normalized by the number of pairs of samples in the overlapping subsequences.

Such indexing structure was separately created for all the mentioned audio descriptors using L_2 metric as well as for aggregation 1:1:1:1 of them. Aggregation function is computed as a weighted sum of all audio descriptors.

4. USER INTERFACE

A web application integrated in the MUFIN project is used as a user interface to the engine. The main web page is by default loaded with a random selection of audio recordings from the database. An audio recording can be played directly in the web browser, downloaded, or user can search for similar recordings. The *kNN* search is realized as described above and the results are displayed. User can also uploaded their own query audio recording.

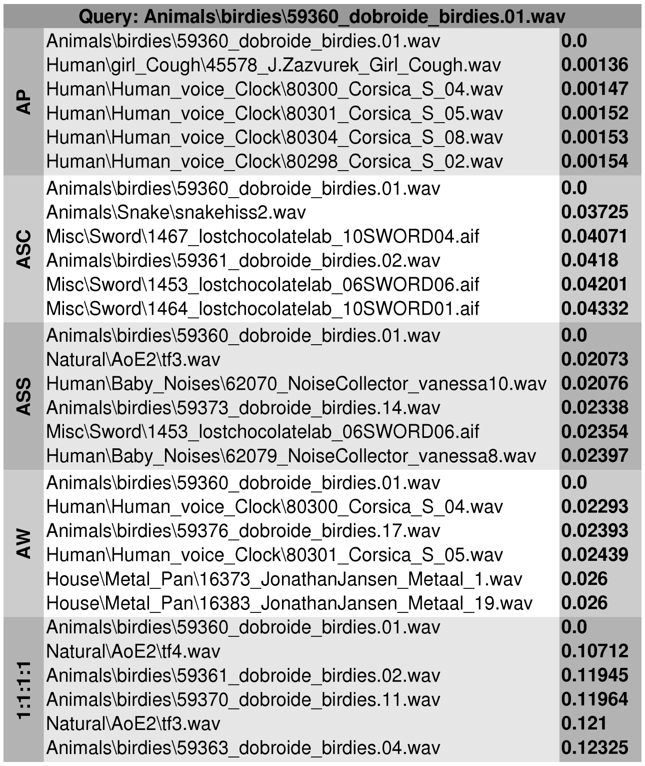
A prototype version of this engine is running on a standard PC and is available at http://mufin.fi.muni.cz/audio.

5. DATA COLLECTIONS

Data collections from servers Partners In Rhytme[1] and The Freesound Project[2] were used. Audio recordings were divided into the following sets according their origin – Animals, House, Human, Instruments, Machines, Misc, and Natural. The whole database contains approximately 1,000 audio recordings. The four mentioned audio descriptors were extracted from each audio recording.

In Figure 2, we can see results of a 6*NN* query from the Animal collection (bird singing) for individual audio descriptors and aggregation 1:1:1:1 of all four descriptors. Using the AP descriptor, engine did not find any relevant result (the evaluation was done by human judges). Using the ASC, one relevant result was found, the ASS found two results, and the AW one result. Using the aggregation of all the descriptors, the results were all relevant.

All performed experiments showed that any audio descriptor is not sufficient itself. Different audio descriptors are suitable for different types of queries but aggregation of all descriptors gives at least as good results as usage of the most suitable audio descriptor separately. In the future, we would like to extend our audio database and add support for melody-based similarity search using high-level descriptors.



Query: Animals\birdies\59360_dobroide_birdies.01.wav		
AP	Animals\birdies\59360_dobroide_birdies.01.wav	0.0
	Human\girl_Cough\45578_J.Zazvurek_Girl_Cough.wav	0.00136
	Human\Human_voice_Clock\80300_Corsica_S_04.wav	0.00147
	Human\Human_voice_Clock\80301_Corsica_S_05.wav	0.00152
	Human\Human_voice_Clock\80304_Corsica_S_08.wav	0.00153
	Human\Human_voice_Clock\80298_Corsica_S_02.wav	0.00154
ASC	Animals\birdies\59360_dobroide_birdies.01.wav	0.0
	Animals\Snake\snakehiss2.wav	0.03725
	Misc\Sword\1467_lostchocolatelab_10SWORD04.aif	0.04071
	Animals\birdies\59361_dobroide_birdies.02.wav	0.0418
	Misc\Sword\1453_lostchocolatelab_06SWORD06.aif	0.04201
	Misc\Sword\1464_lostchocolatelab_10SWORD01.aif	0.04332
ASS	Animals\birdies\59360_dobroide_birdies.01.wav	0.0
	Natural\AoE2\tf3.wav	0.02073
	Human\Baby_Noises\62070_NoiseCollector_vanessa10.wav	0.02076
	Animals\birdies\59373_dobroide_birdies.14.wav	0.02338
	Misc\Sword\1453_lostchocolatelab_06SWORD06.aif	0.02354
	Human\Baby_Noises\62079_NoiseCollector_vanessa8.wav	0.02397
AW	Animals\birdies\59360_dobroide_birdies.01.wav	0.0
	Human\Human_voice_Clock\80300_Corsica_S_04.wav	0.02293
	Animals\birdies\59376_dobroide_birdies.17.wav	0.02393
	Human\Human_voice_Clock\80301_Corsica_S_05.wav	0.02439
	House\Metal_Pan\16373_JonathanJansen_Metaal_1.wav	0.026
	House\Metal_Pan\16383_JonathanJansen_Metaal_19.wav	0.026
1:1:1:1	Animals\birdies\59360_dobroide_birdies.01.wav	0.0
	Natural\AoE2\tf4.wav	0.10712
	Animals\birdies\59361_dobroide_birdies.02.wav	0.11945
	Animals\birdies\59370_dobroide_birdies.11.wav	0.11964
	Natural\AoE2\tf3.wav	0.121
	Animals\birdies\59363_dobroide_birdies.04.wav	0.12325

Figure 2: Example of 6NN query of bird singing for individual audio descriptors and their aggregation 1:1:1:1 together with distances from the query.

6. ACKNOWLEDGMENTS

This work has been supported by national research projects MSMT 1M0545, GACR 201/08/P507, GACR 201/09/0683, GACR 103/10/0886, and GACR P202/10/P220.

7. REFERENCES

[1] M. Batko, V. Dohnal, D. Novak, and J. Sedmidubsky. Mufin: A multi-feature indexing network. In *Proceedings of SISAP '09*, pages 158–159, Washington, DC, USA, 2009. IEEE Computer Society.

[2] M. Batko, D. Novak, and P. Zezula. MESSIF: Metric similarity search implementation framework. In *First International DELOS Conference, Pisa, Italy, Revised Selected Papers*, volume 4877 of *Lecture Notes in Computer Science*, pages 1–10. Springer, 2007.

[3] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. *SIGMOD Rec.*, 23(2):419–429, 1994.

[4] D. Novak and M. Batko. Metric index: An efficient and scalable solution for similarity search. In *Proceedings of SISAP '09*, pages 65–73, Washington, DC, USA, 2009. IEEE Computer Society.

[5] P. Salembier and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface.* John Wiley & Sons, Inc., New York, NY, USA, 2002.

[1] http://www.partnersinrhytme.com/

[2] http://www.freesound.org/