The Europeana Sounds Music Information Retrieval Pilot

Alexander Schindler[1], Sergiu Gordea[1], and Harry van Biessum[2]

[1] Digital Insight Lab, Digital Safety and Security Department
Austrian Institute of Technology
{alexander.schindler,sergio.gordea}@ait.ac.at
[2] Research and Development
Netherlands Institute for Sound and Vision
hvbiessum@beeldengeluid.nl

Abstract. This paper describes the realization of a Music Information Retrieval (MIR) pilot for a huge audio corpora of European cultural sound heritage, which was developed as part of the Europeana Sounds project. The demonstrator aimed at evaluating the applicability of technologies deriving from the MIR domain to content provided by various European digital libraries and audio archives. To approach this aim, a query-by-example functionality was implemented using audio-content based similarity search. The development was preceded by an elaborated evaluation of the Europeana Sounds collection to assess appropriate combinations of music content descriptors that are capable to effectively discriminate the various types of audio-content provided within the dataset. The MIR-pilot was evaluated both by using an automatic and a user based evaluation. The results showed that the quality of the implemented query-by-example algorithm is comparable to state-of-the-art music similarity approaches reported in literature.

1 Introduction

The Europeana Sounds project aims at emphasizing on Europe's cultural audio heritage by aggregating content provided by 20 partner institutions including digital libraries and audio archives. The descriptions of the contributed audio content is made accessible to the public through the Europeana portal. Moreover, the object descriptions and Web Links to the media files are also available through a public API, making these data-sets reusable for 3rd party applications and for research purposes. The aggregated sound content ranges from music to interviews, animal or ambient sounds, broadcasts, news, etc. This high variety of content, the large number of audio items - more than 350.000 items - and the various languages used to describe it (i.e. there are 28 languages used in Europeana) states a problem concerning the retrieve-ability of the provided content. Although the items are rich in descriptive meta-data, these descriptions are often not sufficient to support sophisticated search scenarios or (musicological) research. Simple queries like finding recordings by artist name or a certain year are well supported by the prevalent retrieval system. More complex scenarios, like searching for contemporary music that was inspired by a classical composer or music style would certainly be problematic, as this information is not available in the meta-data. Especially in the case of music recordings, it is not feasible to describe in details the quality and the emotions generated by particular tunes. Within this paper we present the Music Information Retrieval (MIR) Pilot developed

within the scope of the Europeana Sounds project with the aim to develop alternative search and exploration functionality for the sound content available in Europeana. The content based search algorithms are aiming at helping end users to overcome various barriers like the language, the domain expertise such as knowing in advance the name of specific music genres like *Tarantella*, and the lack of extensive content description. This demonstrator has the goal of evaluating the feasibility of implementing effective audio retrieval services for this large and heterogeneous sound data-set, while the main target is to provide a reliable service powering the content based retrieval in Europeana Music Collection[3]. A preliminary evaluation of the demonstrator was performed to quantify the accuracy of the proposed algorithm. As presented in Section 4 the experimental results are comparable with state of the art solutions applied on large music data-sets [6]. Furthermore, an user evaluation was carried out with the goal of measuring the user satisfaction when employing the system for completing special music retrieval tasks.

2 Europeana Sounds Data

For creating the development and evaluation data-set, meta-data descriptions of 400,615 items were collected via the Europeana API[4]. Out of these, 389,120 items included Web URLs pointing to the corresponding audio data. A part of these URLs were outdated, pointed to corrupt audio files, or they couldn't be processable by the audio feature extractors. The final dataset size of 312,096 records makes the relevance of this evaluation comparable to large scale experiments on the Million Song Data-set [6]. The statistical analysis by type of content shows that **Music** is by far the biggest category of the collection varying by style, instrumentation and recording quality. **Spoken Word** in form of interviews, radio news broadcast, public speeches, etc. **Animal Sounds** are field recordings of a wide range of animals. Recordings of **Radio Broadcasts**. These audio items are long mixed-content files. They consist of spoken content, music and radio commercials.

3 Implementation

In order to implement an effective retrieval algorithm, capable of providing a good accuracy over different types of audio content, appropriate feature set have been selected. Effective retrieval of different music styles needs to take in consideration various music properties such as timbre, rhythm and harmony, as well as their progression and variety over the complete performance. Different feature sets are known to work better on certain music genres, but to be inferior when applied on other genres. A further obstacle is the presence of old historic recordings for which, scratches and noise resulting from decaying media distort the feature values. *Spoken word* shows completely different spectral properties than *music* and thus require different audio features to distinguish them from music content. For *animal sounds* it was considered to be sufficient to match animals by the same family. A more detailed discrimination of animal sound was not required for this demonstration.

[3] `http://europeana.eu/portal/collections/music`
[4] `http://labs.europeana.eu/api`

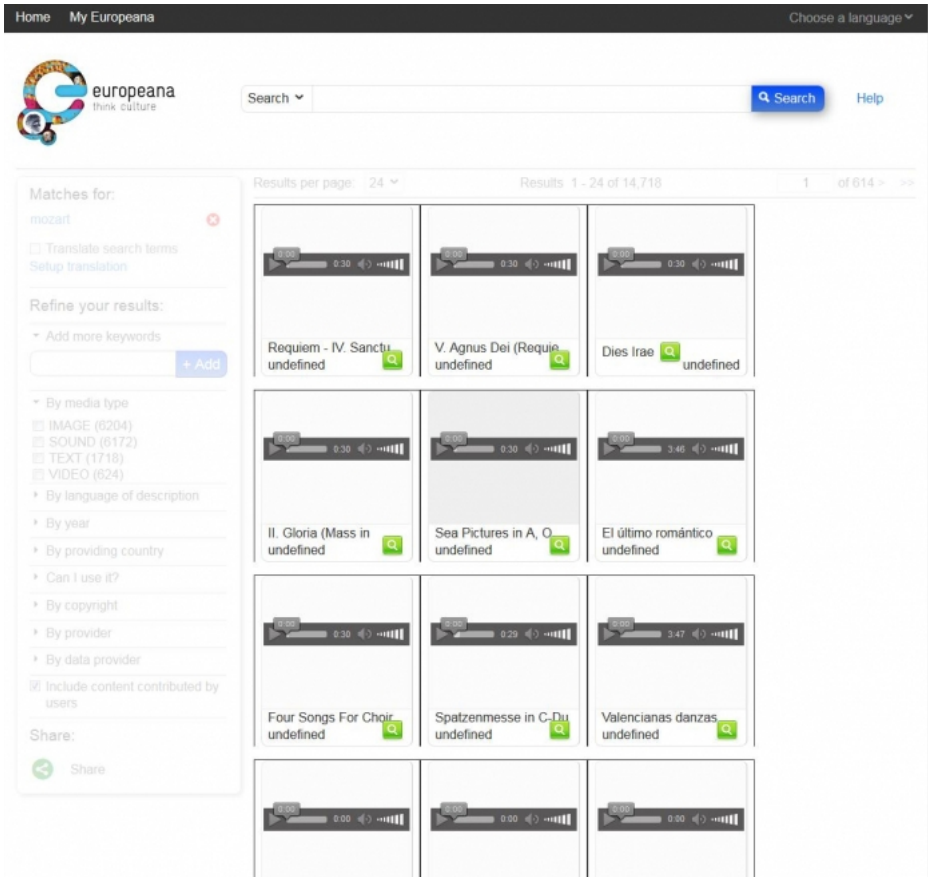

Fig. 1: User interface of the MIR pilot demonstration developed for Europeana Sounds.

3.1 Audio-Content Descriptors

The following audio-content descriptors were evaluated in preceding experiments:

- **Statistical Spectrum Descriptors (SSD)** subsequently computes seven statistical measures for the 24 critical bands of hearing. Mean, median, variance, skewness, kurtosis, min- and max-values, for different segments of a song are aggregated by calculating the median of the descriptors of all segments. SSDs are part of the *Psycho-acoustic Music Descriptors* as proposed by [5] and are based on a psycho-acoustically modified Sonogram representation that reflects human loudness sensation.
- **Rhythm Patterns (RP)** describe rhythmical characteristics by applying a discrete Fourier transform to the transformed Sonogram, resulting in a (time-invariant) spectrum of loudness amplitude modulation per modulation frequency for each individual critical band which provides a rough interpretation of the rhythmic energy of a song. For feature extraction we employed a Python-based implementation[5]

[5] `https://github.com/tuwien-musicir/rp_extract`

Table 1: Overview of the audio-content *descriptors*, their corresponding acoustic *categories*, their assigned feature weight *(f. W.)* as well as the cumulative category weight *(c. W.)*.

Category	Feature	Description	f. W.	c. W.
Timbre	**MFCC**	Timbre description	23%	39%
	SSD	General spectral description	8%	
	SPEC CENT	Pitch description	8%	
Rhythm	**RP**	Rhythmic patterns	18%	25%
	BPM	Tempo	7%	
Harmony	**CHROMA**	Harmonic Scale	12%	24%
	TONNETZ	Traditional harmonic description	12%	
Loudness	**RMSE**	Loudness description	9%	9%
Noise Behaviour	**ZCR**	Noisiness description	3%	3%

- **Mel Frequency Cepstral Coefficients (MFCC)** [8] are derived from speech recognition and also apply log-scale transformations to anneal the feature response to the human auditory systems. MFCCs are good descriptors of timbre.
- **Chroma** [8] features project the entire spectrum onto 12 bins representing the 12 distinct semitones of the musical octave. Both MFCC and Chroma were extracted using the well known MARSYAS toolset [8].
- **Root Mean Square (RMSE)** is a way of comparing arbitrary waveforms based upon their equivalent energy. The RMS method takes the square of the instantaneous voltage, before averaging, then takes the square root of the average.
- The **Spectral Centroid (SPEC CENT)** [8] is the frequency-weighted sum of the power spectrum, normalized by its unweighted sum. It determines the frequency area around which most of the signal energy concentrates and gives an indication of how *dark* or *bright* a sound is.
- **Tempo** measured in Beats per Minute (BPM) [2] is calculated from audio events which are detected in the audio signal.
- **TONNETZ** features [3] are able to detect changes in the harmonic content of musical audio signals based on a model for Equal Tempered Pitch Class Space using 12-bin Chroma vectors. Close harmonic relations such as fifths and thirds appear as small Euclidean distances. Peaks in the detection function denote transitions from one harmonically stable region to another.
- **Zero Crossing Rate (ZCR)** [8] measures the noise behavior of an audio-signal.

The summary of these feature sets together with their weighting for similarity computation is presented in Table 1.

3.2 Composite Feature-Sets

Content based audio and music features attempt to capture certain aspects of music. To provide an ensemble description of a recorded music track it is required to make use of multiple features. The introduced audio features were grouped into the following five music properties which have been chosen to describe music similarity upon:

- **Timbre** is a fundamental property of music and generally reflects the instrumentation used during the performance. Timbre is often a good discriminator for music styles as well as moods expressed by a song.
- **Rhythm** is a similarly strong intrinsic music property.
- **Harmony** describes the tonality of a composition. In terms of an analytic perspective, it analyses how the spectral energy is distributed among a certain (usually western) scale.
- **Loudness** is actually not relevant for music similarity, it was considered referring to recent observations in contemporary music which tends to steadily increase on loudness [7]. By reducing the dynamic range the resulting sound is subjectively more attractive.
- **Noise Behavior** analysis refers to the different recording qualities of audio content. This captures the degradation of the original carriers such as shellac or wax tapes. Adding these features to the stack prefers performance over composition, and thus groups the records with similar sound quality.

3.3 Similarity Calculations

Exhaustive experimentation was applied using the audio features introduced in Section 3.1 and a selection of 18 distance measures discussed in [1]. No general pattern could be identified on which distance measure works best for all features. A general observation was that L1 based metrics usually rank high for the presented feature combinations. Among them the Canberra distance [4] includes an implicit normalization step. The Canberra distance was mostly top-ranked and provided stable results with increasing result list length. Thus, it was decided to use this distance measure for the MIR-pilot. A *late fusion* approach was used to combine the different feature-sets. The similarities are calculated for each feature separately and the distinct similarity values for each song are combined arithmetically. *Feature weighting* was applied to reduce overrated influence of distinct audio-descriptors. Feature weight estimation and optimization was approached empirically through a predefined set of similar records. During an iterative process the weights of the different features were adapted. The final feature weights used for the implementation are provided in Table 1.

3.4 User Interface:

The user-interface was aligned to the design of the Europeana portal. The MIR pilot supports the following use cases:

- **Term-based queries** accept text-based input to query the meta-data to facilitate elementary means to explore the Europeana Sounds collection, or to search for content based on certain terms such as “blues”, “love” or “piano”.
- **Query by Example** through supplying an example song the system searches for similar ones based on their acoustic properties.
- **Usage of External Content to Query for Europeana Content**. To demonstrate further possibilities the query by example approach has been extended to accept also content which is not contained in the Europeana Sounds collection. The Soundcloud API[6] was used which facilitates computational access to the Soundcloud

[6] https://developers.soundcloud.com

music streaming service. By supplying a Soundcloud URL the corresponding audio data is downloaded, processed and its calculated features are analyzed for similarity within the Europeana Sounds data-set.

4 Evaluation

The evaluation of the system was subdivided into a computational part which facilitated the automated evaluation of a large number of queries on a pre-defined ground truth, and a user-questionnaire part which focused on the overall user-perception of the system.

4.1 Automatic Evaluation

For the automatic evaluation the rich meta-data of the data-set has been analyzed to identify a set of semantically descriptive audio categories. The advantage of the data provided by Europeana is that all data items, including their corresponding meta-data, have been curated and edited by domain experts working for national libraries and audiovisual archives. The selected categories provide an overview of various, representative and well known music and sound genres available in the data-set. For each category the corresponding data-set items have been selected. Similar items have been calculated for each of them. The precision was measured by the number of items of the same category at different cut-off points. For very large categories the number of queries was randomly sub-sampled to 1000 items. Results presented in Table 2 describe precision values for queries of the five major categories (Jazz, Classic, Folk, Sounds, Spoken word) at different granularity. Generally it can be observed that spectral homogeneous tracks such as animal sounds and spoken word are better discriminated than polyphonic music. The calculated average precision of 28.7% for all performed queries (including queries not listed in Table 2 is slightly above the top result of 27.4% presented in [6] where k-nearest neighbors classification results on data-set only 12.2% bigger than the Europeana data-set was reported. The results for $k = 1$ are equivalent to the similarity retrieval result at cut-off 1.

4.2 User Evaluation

In order to get an end-user perspective on the results of the MIR pilot, an user evaluation was performed. In sessions of 90 minutes 13 participants provided feedback on their experience with the MIR pilot. A Likert-scale was used to quantify the perception of the calculated music similarity experienced between audio tracks selected from several different categories as well as the overall experience of the provided system. The participants were asked to specify their perception of similarity according the overall similarity of audio tacks as well as their specific music properties *tempo*, *rhythm*, *harmony*, *timbre*, *instrumentation* and *quality of the recording*. The participants were selected from three different types of users: *music lovers* as regular music listeners; *hobby musicians* which play music themselves and have a certain level knowledge with regard to musical concepts; and *music professionals* for whom performing, and or recording music is part of their regular work. Each participant evaluated nine reference tracks and the top tree results (27 tracks in total) and provided a narrative feedback about the applied concept of music information retrieval as well. After playing the reference track, the users were asked to listen and evaluate the top three result tracks provided by the system. Apart from the

Table 2: Precision values for the computational evaluation at cut-off points 1,2,3,5,10. Abbreviations: #: number of class items; *Classic q.a.m.*: Classic quartet allegro major; *Flamenco Guit*: Flamenco + Guitarra; *A.S. Crickets*: Animal Sounds - Crickets

Query	#	1	2	3	5	10
Jazz	31801	38.0	35.0	31.4	31.7	28.6
Smooth Jazz	2419	49.1	45.9	43.8	25.8	20.8
Ragtime	57	24.6	15.8	12.3	7.3	3.6
Classical	28569	44.3	42.1	40.5	38.3	35.1
Classic G maj.	304	17.1	14.8	14.0	12.6	9.3
Classic q.a.m.	191	9.4	6.3	7.3	8.1	5.6
Piano Concerto	510	38.6	32.0	28.0	23.9	17.6
Requiem	463	32.6	26.9	22.0	16.2	10.7
Opera	8278	26.8	24.7	22.7	21.1	18.9
Operette	1081	27.7	22.9	20.8	17.3	14.6
Flamenco	1827	40.7	33.0	29.2	24.3	18.2
Flamenco Guit	287	22.3	17.1	15.3	13.5	10.0
Tarantella	152	33.6	28.0	22.4	16.1	8.5
Tango	3716	30.2	24.9	22.3	19.5	16.0
Animal Sounds	1097	89.7	87.0	85.1	82.8	78.7
Animal Sounds Crickets	113	59.3	55.3	56.6	53.0	48.1
Interview	484	77.5	74.3	72.0	68.6	60.8

nominal scaled ratings the explanations for the experienced similarities and differences were being noted as well during the evaluation. When analyzing the evaluation results no noticeable differences between the user groups were observed. Participants generally agreed upon the similarity of *timbre* related music properties such as *instrumentation* as well as *harmony* of the similar tracks calculated by the MIR pilot (see Figure 2). *Tempo* and *rhythm* earned not as high ratings. From the narrative feedback it was understood that the rhythm dimension was not evaluated in the sense of rhythmic patterns, but as the overall rhythm of the interpretation. Similarly, it was observed very high correlation between the feedback for the *harmony* and *timbre* dimension, which indeed are interdependent musical concepts. The feedback on the *tempo* and *rhythm* dimensions are least correlated with the overall similarity, meaning that their influence on the similarity perception is lower in the case of music content. However, they are good discriminators between music and other types of sound content (i.e. like speech or environmental sounds). While the user evaluation was carried out with a small number of users on a small number of music items, the evaluation results cannot be perceived as an overall evaluation of the systems. However they can be used to validate the weighting of different categories and feature sets in the similarity computation (i.e. which were derived from the experience of the past music information retrieval research and experiments).

5 Conclusion and Future Work

We presented our audio-content similarity estimation based query-by-example implementation on a very large dataset which has been aggregated by the Europeana

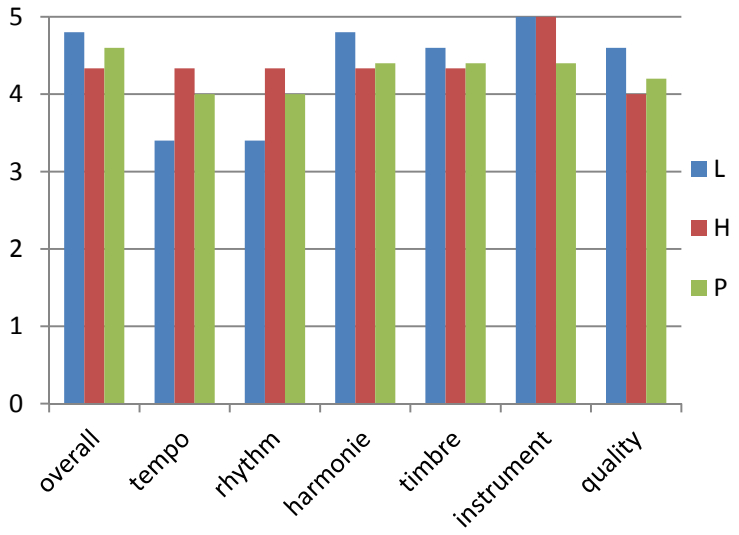

Fig. 2: Quantitative evaluation of the similarity perception over different music properties, tailored by the user groups music lovers (L), hobby musicians (H) and music professionals (P).

Sounds project. The presented approach based on weighted combinations of different audio-content descriptors facilitates similarity estimations of the highly heterogeneous data. The evaluation showed that the presented audio descriptor combinations, as well as the evaluated distance measure and feature space fusion methods are appropriate and results are comparable to results reported in literature. Based on these results it was decided to incorporate these results into the core Europeana search system and to extend the audio search functions by audio-content analysis based approaches.

References

1. S.-H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1, 2007.
2. S. Dixon. Evaluation of the audio beat tracking system beatroot. *Journal of New Music Research*, 2007.
3. C. Harte, M. Sandler, and M. Gasser. Detecting harmonic change in musical audio. In *Proc. 1st ACM WS on Audio and music computing multimedia*, 2006.
4. G. Jurman, S. Riccadonna, R. Visintainer, and C. Furlanello. Canberra distance on ranked lists. In *Proc of Advances in Ranking NIPS WS*, 2009.
5. T. Lidy and A. Rauber. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *ISMIR*, 2005.
6. A. Schindler, R. Mayer, and A. Rauber. Facilitating comprehensive benchmarking experiments on the million song dataset. In *ISMIR*, 2012.
7. J. Serrà, Á. Corral, M. Boguñá, M. Haro, and J. L. Arcos. Measuring the evolution of contemporary western popular music. *Scientific reports*, 2, 2012.
8. G. Tzanetakis and P. Cook. Marsyas: A framework for audio analysis. *Organised sound*, 4(3):169–175, 2000.