Project Acronym: **Europeana Sounds**
Grant Agreement no: **620591**
Project Title: **Europeana Sounds**

D2.6 Music Information Retrieval Pilot delivery report

Revision: Final

Date: 04/12/2015

Authors: Alexander Schindler (AIT)

Harry van Biessum (NISV)

Abstract: The theoretical principles of the Music Information Retrieval (MIR) pilot were developed as part of the Europeana Sounds project and are described in this document. MIR focuses on extracting music relevant information from either the music content itself or from related sources. The MIR-pilot aims to evaluate the applicability of technologies deriving from the MIR domain, to content provided by the Europeana Sounds project. To approach this aim, a query-by-example functionality was chosen to be implemented. The MIR-pilot was evaluated using an automatic approach as well as one including user-questioning. The results show that the quality of the implemented query-by-example algorithm is comparable to state-of-the-art music similarity approaches reported in literature.

Dissemination level	
Public	X
Confidential, only for the members of the Consortium and Commission Services	

Coordinated by the British Library, the Europeana Sounds project is co-funded by the European Union, through the ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme (CIP) http://ec.europa.eu/information_society/activities/ict_psp/

Revision history

Version	Status	Name, organisation	Date	Changes
0.1	ToC	Alexander Schindler, AIT	26/10/2015	
0.2	1st draft	Alexander Schindler, AIT	28/10/2015	Automatic evaluation
0.3	1st draft	Harry van Biessum, NISV	24/11/2015	1st draft user-evaluation
0.4	2nd draft	Alexander Schindler, AIT	26/11/2015	
0.5	Review	Maarten Brinkerink, NISV Sergiu Gordea, AIT	26/11/2015	Introduction, implementation, similarity computation
0.6	Final draft	Alexander Schindler, AIT	28/11/2015	Input from reviews
0.7		Harry van Biessum, NISV	30/11/2015	Finalized user evaluation
0.8	Review	Maarten Brinkerink, NISV	30/11/2015	
0.9	Final draft	Alexander Schindler, AIT	30/11/2015	Input from reviews
1.0	Final	Laura Miles, BL Richard Ranft, BL	01/12/2015	Layout, minor changes

Review and approval

Action	Name, organisation	Date
Reviewed by	David Haskiya Maarten Brinkerink	26/11/2015 27/11/2015
Approved by	Coordinator and PMB	01/12/2015

Distribution

No.	Date	Comment	Partner / WP
1	01/12/2015	Submitted to the European Commission	BL/WP7
2	01/12/2015	Posted on Europeana Pro website	BL/WP7
3	01/12/2015	Distributed to project consortium	BL/WP7

Application area

This document is a formal output for the European Commission, applicable to all members of the Europeana Sounds project and beneficiaries. This document reflects only the author’s views and the European Union is not liable for any use that might be made of information contained therein.

Statement of originality

This document contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Project summary

Europeana Sounds is Europeana's 'missing' fifth domain aggregator, joining APEX (Archives), EUscreen (television), the Europeana film Gateway (film) and TEL (libraries). It will increase the opportunities for access to and creative re-use of Europeana's audio and audio-related content and will build a sustainable best practice network of stakeholders in the content value chain to aggregate, enrich and share a critical mass of audio that meets the needs of public audiences, the creative industries (notably publishers) and researchers. The consortium of 24 partners will:

- Double the number of audio items accessible through Europeana to over 1 million and improve geographical and thematic coverage by aggregating items with widespread popular appeal such as contemporary and classical music, traditional and folk music, the natural world, oral memory and languages and dialects.
- Add meaningful contextual knowledge and medium-specific metadata to 2 million items in Europeana's audio and audio-related collections, developing techniques for cross-media and cross-collection linking.
- Develop and validate audience specific sound channels and a distributed crowd-sourcing infrastructure for end-users that will improve Europeana's search facility, navigation and user experience. These can then be used for other communities and other media.
- Engage music publishers and rights holders in efforts to make more material accessible online through Europeana by resolving domain constraints and lack of access to commercially unviable (i.e. out-of-commerce) content.

These outcomes will be achieved through a network of leading sound archives working with specialists in audiovisual technology, rights issues, and software development. The network will expand to include other data-providers and mainstream distribution platforms (Historypin, Spotify, SoundCloud) to ensure the widest possible availability of their content.

For more information, visit http://pro.europeana.eu/web/europeana-sounds and http://www.europeanasounds.eu

Copyright notice

Contents

Executive summary: D2.6 Music Information Retrieval Pilot delivery report

This document describes the theoretical principles and implementation detail of the Music Information Retrieval (MIR) pilot which was developed as part of the Europeana Sounds project. MIR focuses on extracting music relevant information from either the music content itself or from related sources such the World Wide Web, human estimators or sensoric data. The MIR-pilot aimed to evaluate the applicability of technologies deriving from the MIR domain to content provided by the Europeana Sounds project. To approach this aim, a query-by-example functionality was chosen to be implemented because it is a highly researched topic which facilitates comparability and interpretability of the results of the MIR-pilot evaluation. The implementation was preceded by an evaluation which elaborated on the Europeana Sounds collection and which technologies, such as appropriate music feature extraction methods, are relevant for the implementation. The MIR-pilot was evaluated using an automated approach as well as one including user-questioning. 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 aggregates metadata records corresponding to hundreds of thousands of audio-related objects and makes them available via the Europeana portal, Channels and API. The described content ranges from music to interviews, animal or ambient sounds, broadcasts and news. The huge variety of audio content makes this collection both fascinating and challenging. Although the records are rich in descriptive metadata, these textual descriptions are often not sufficient to support sophisticated search scenarios or (musicological) research. For example, finding recordings from a specific artist is not a problem, nor finding works from a certain year. A search for contemporary music which was inspired by a classical composer or music style would certainly be problematic. This would require references to influencing artists and styles to be recorded somewhere in the metadata. Yet, manually annotating records in digital library collections is cumbersome and error prone. Audio, especially music recordings, is even more complex to describe. This is because, despite of some objective properties (such as composer, performer, instrumentation, structure or meter), music is highly subjective. The perception of properties like genre, timbre or mood is both personal and ambivalent. Even if such information has been recorded in the metadata, the algorithms of conventional search engines require a certain amount of consistency in descriptive terminology. Thus it appears that the state-of-the-art text based approach faces many obstacles concerning the effectiveness of music or audio search.

The information required for implementing an efficient audio search is encoded in audio files, but it still needs to be unlocked. Human listeners are able to identify performing artists, instrumentation, melodies, songs, moods, genres, as well as even subtle similarities between songs within fractions of a second. Based on this premise, the aim is to retrieve this information, to extract it, and to store it in a format that can be machine processed. This process is called feature extraction[1] and transforms music information into a series of descriptive numbers. Based on the applied feature extraction method these numbers express or relate to musical properties such as timbre, rhythm or harmony. Figure 1 shows an example of rhythmic features extracted from two different music tracks. Image a) represents the rhythmic energy from a classical, and image b) from a heavy metal track. The algorithm used to calculate these features is explained in more detail in Chapter 2. At a glance it can be observed, that the two songs obviously differ significantly in their peak values (red coloured pixels refer to high intensity values).

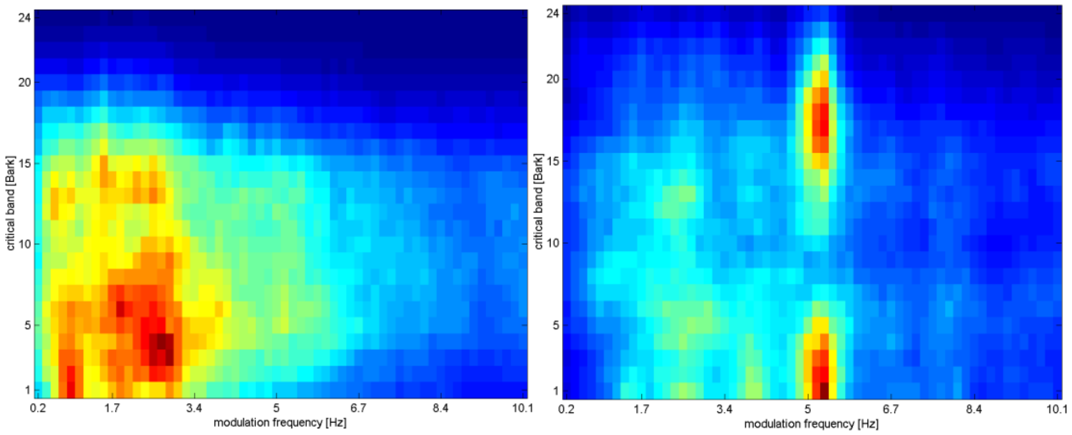


Figure 1: Visualization of the content based music feature *Rhythm Patterns* which capture the rhythmic energy of a recording. Rhythm Patterns of two recorded songs are provided: a) classic music "An der schönen blauen Donau" by Johann Strauss (left). b) Heavy metal music "No one knows" by Queens of the Stone age (right).

Music features calculated from the audio signal, such as the examples provided above, are further used to make songs comparable on a numerical basis. In that sense, songs exhibiting similar patterns will be considered as musically similar, at least in terms of rhythmic expression. This relationship between numerical and music similarity facilitates new approaches to music search. Besides the traditional text-based approach, where music information has to be textually encoded and provided, a content-based approach adds a wide range of advantages and possibilities to search engines.

1.1 Content based search - benefits and difficulties

Content-based search refers to approaches using information available within the audio content and not within the descriptive metadata. As explained in the previous section, this usually requires a pre-processing step where this information is extracted and quantized to make it comparable on a numerical level. This is also referred to as vector representation. Usually a content-based feature - independently of the actual multimedia domain (e.g. music, image, video) is not a single value but a set of values. The Chroma features which will be introduced in Chapter 2 describe the distribution of the spectral energy over the 12 semitones of a musical octave. Thus, the feature consists of 12 values which are represented as one vector with a length of 12. Extracting these features from a collection of audio files thus provides

[1] see https://en.wikipedia.org/wiki/Music_information_retrieval#Statistics_and_machine_learning

n feature vectors of length 12 and thus creates a so-called vector space of the size *n x 12.* Referring to this, search approaches based on content-based features are also often referred to as *Vector Space Approaches* or *Vector Space Models*[2].

Searches in vector space function on different principles from traditional text-based approaches. Contrary to query languages which have strictly defined rule sets, such as SQL[3], vector space models are based on similarity estimations which are generally derived from numerical differences of feature vectors. There are various methods to calculate vector similarities by considering aspects of different probability distributions, but all of them provide an estimate of how similar the values within the vectors are. In the context of a content-based search engine, these estimates are used to facilitate sophisticated search scenarios like the popular use case of Query *by Example*. In this scenario, the information need is expressed by an example song and the search engine is expected to retrieve similar songs. Thus, as a first step, the feature vector of the example song will be used to calculate the pairwise similarities to each other feature vector in the song collection. The result is a list of *n-1* similarity estimations. By sorting this list, top-entries with high similarity values are considered to refer to highly similar songs. The theory of this approach will be further explained and discussed in Chapter 2. Generally, this interpretation of music similarity is theoretical and highly dependent on the models used to calculate the feature vectors. Although this is a considerable weakness of this approach, such approaches are state-of-the-art in multimedia research in general. Research specifically on music feature extraction is progressing rapidly and remarkable results are already made available as open source software [Fu 2011, Tzanetakis 2002, McFee 2015].

An advantage of a vector space approach is that many properties can be compared at once. Usually content-based search engines do not utilise only a single feature, but combine various feature sets which describe different aspects of the content. In music features describing timbre, rhythm, pitch, tempo and harmony are often used together to describe the music content. In image retrieval, features representing colour information are often combined with descriptions of texture or shape. Many of these distinct features are by themselves composed of dozens, to hundreds, of attributes. Comparing such high numbers of attributes is a difficult task to solve in structured search approaches.

The different approach to compute query results using vector space approaches presents the main obstacle of a content-based model. It is difficult to add a vector space model to an existing structured search engine. Thus, it is often required to migrate to different frameworks which support such approaches. Such frameworks, on the other hand, provide a wide range of search options where content-based information can be combined with metadata to create highly sophisticated searches.

1.2 Query by example on the Europeana Sounds collection

Within the scope of the Task 2.4.2 *Music Information Retrieval* it was decided to implement a query by example scenario. The main goals are to demonstrate the pilot approach and to add Music Information Retrieval (MIR) technologies to the Europeana Platform. Based on the feedback and evaluation of the pilot it will be decided if the technology is mature enough for enhancing search functionality within the Music Channel. Query by example scenario is an actively researched method, and the on-going

[2] see https://en.wikipedia.org/wiki/Vector_space_model
[3] https://en.wikipedia.org/wiki/SQL

discussions within the MIR community concerning music similarity make this approach a very good experiment to assess the objectives provided.

One obstacle the MIR research community constantly faces is the absence of adequate datasets (i.e. containing a large number of labelled songs, which are publicly accessible). Such publicly available test and benchmark datasets[4] are required to make experiments and evaluations reproducible and their results comparable. Among others, one obstacle is related to the fact that the majority of music content is still subject to IPR[5]. Consequently, many approaches described by the MIR community are often evaluated against either very small or highly specific datasets. The Europeana Sounds collection on the other hand is a large dataset, consisting of a wide range of music and non-music content. This circumstance makes the facilitation of a query by example search approach both interesting and challenging. Reports of mixed-content evaluations which are not restricted to music but also include different audio content during the evaluation are scarce in literature. Thus, adequate combinations of audio and music features to capture the acoustic properties of the Europeana Sounds music and non-music content are unknown and have to be evaluated in advance.

1.3 Obstacles for the Europeana Sounds collection

The Europeana Sounds Collection will be discussed in more detail in Chapter 2. The main challenge of this dataset is its wide variety. The following main categories have to be considered when evaluating appropriate feature sets for the implementation of a query by example search engine.

- **Music**: Music content is by far the biggest category of the collection. Yet, this category has to be subdivided into subcategories concerning music style, instrumentation and recording quality. Distinguishing different music styles requires a consideration of 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. Scratches and noise resulting from decaying media distort the feature values. Human ears are able to filter such noise and to still recognize the underlying music content. Music features which are robust enough to withstand against such noise have to be considered and evaluated.

- **Spoken Word**: Spoken word in the form of interviews, radio news broadcast or public speeches show completely different spectral properties and thus require different audio features to distinguish them from music content. This introduces additional problems where different features mask properties of others. To avoid this, weighting can be applied to reduce the influence of distinct feature sets on the final similarity estimation.

- **Animal Sounds**: Recordings of animal sounds mostly contain noise and a small percentage of actual sound. Matching or identifying the appropriate animal has not been considered an objective of the MIR-pilot. It should be sufficient to match bird sounds with other bird sounds. A more detailed discrimination is not required.

[4] https://en.wikipedia.org/wiki/Benchmark_%28computing%29
[5] https://en.wikipedia.org/wiki/Music_law

- **Radio Broadcasts**: Some sub-collections consist of live radio recordings. These audio items are long mixed-content files. They consist of spoken content, music and radio commercials.

1.4 Objectives of the MIR-Pilot task

The following objectives are considered for the elaboration of the MIR-pilot task.

- **Europeana Sounds Collection:**
  - *Obtain an overview of the content provided by Europeana-Sounds*. Which content type will be available? Does it only consist of music, or are other types also available? As previously explained, this influences the decision on which feature sets have to be applied to efficiently capture the various properties of the different audio types.
  - *Assess the availability of source audio files*. Direct access to media files is required to extract content based audio features.
  - *Obtain an estimation of the collection size*. Feature extraction is resource demanding. It takes a lot of computation time to calculate the features, and the extracted feature vectors are demanding in terms of disk and memory space. Space considerations have to be known in advance to choose adequate algorithms that are capable of processing such amounts of data in considerable time.
- **Music Information Retrieval**
  - *Which features are suitable?* Identify state-of-the-art feature extractors that capture all types of audio content provided by the Europeana Sounds collection. This is an iterative process. Combinations of different feature set have variant influence on the final similarity estimation. Finding the best performing combinations is often based on exhaustive search and thus time consuming.
  - *What pre-processing is required?* Every content-based feature captures a distinct aspect of music expression. Still the resulting values generally lack of descriptive units. Thus, the values of distinct feature values are not comparable. A common method to overcome this is to normalize the vector spaces. The method of normalization has a substantial influence on the results of similarity calculations and thus a suitable one has to be evaluated.
  - *Which similarity measure is appropriate?* Various similarity measures are available. Each of them captures some aspects of probability distributions better than the others. Proper selection of similarity measures is reported and known to be essential for music similarity estimation. Many similarity or distance measures require specific feature normalization methods. An extensive evaluation of which distance measure and normalization combination is most suitable for the Europeana Sounds collection is required.

- *Assess appropriate feature weights:* Feature weights are used to adjust the influence of a feature on the entire feature space, and thus the resulting similarity estimation. Adjusting feature weights is a cumbersome task that requires a lot of experimentation. Due to the size of the collection, it is not feasible to employ a processing-intensive approach for automatic estimation of feature weights.

- **Pilot Implementation**
  - *Provide a pilot implementation capable of processing the Europeana Sounds Collection.* Depending on the size of the collection, adequate technologies have to be chosen to facilitate similarity computations with reasonable response times.
  - *Provide a user interface for demonstration and evaluation:* A user interface that aligns to the Europeana Framework needs to be implemented. This UI will be used for demonstration, discussions and evaluations, so that user evaluation focussing on user perception and satisfaction has evaluated this interface.

2 Music similarity

Music similarity is an attractive but also challenging research field, due to its inherent ambiguity and subjectivity. Music is part of our culture. It is a prevalent part of our daily life. It sounds from radios, TVs, shops and coffee shops on your way to work. Parents sing songs to make us fall asleep, hymns are taught in school, as are traditional songs and an overview of the rich cultural heritage of classical music. Teenagers use music to express their personality, and at that age the music starts to play a major role in our lives.

Figure 2: Music in everyday life

Because of the subjective perception of music properties, the objective estimation of music similarity becomes a highly challenging task. Everyone has their own interpretation of similarity. In terms of music similarity calculations, objectivity is generally a requirement for generalization. Subjectivity and personalization concerns add further complexity and require the presence of user-related (personal) data.

Furthermore, there are different expectations between professionally-trained and untrained listeners. Especially in the classical domain, notions of similarity are discussed on a different level between musicologists. In that sense, the context plays a more important role. Although contextual information is easier to ingest in a structured search system, it takes huge effort to provide and curate relevant information in an adequate format.

Due to the mentioned obstacles, a common solution is to use acoustic similarity as an estimate for music similarity. This is still controversial because the employed methods derive from digital signal processing, which differs from how humans perceive music. Humans are able to separate and concentrate on distinct properties of music such as instruments, and melodies played by these instruments. This source separation in digital signal processing is still under research, and is not yet developed enough to process polyphonic digital music. As a consequence, music is often interpreted by its distribution of spectral energy; melodic or harmonic attributes need to be carefully reconstructed from the sampled audio. Different models have been developed attempting to capture specific properties of music. The developed pilot uses the following properties:

- **Timbre:** Timbre is a fundamental property of music and generally reflects the instrumentation used during the performance. Timbre is often a good discriminator for music genres as well as moods expressed by a song. Several timbre features have been presented such as the Mel Frequency Cepstral Coefficients (MFCC)[6] which are currently the most widely used features in music research.

- **Rhythm:** Rhythm defines music, as well as timbre does. Other than with traditional or musicological descriptions of rhythm, where rhythms have names assigned to them, computational rhythm features provide a statistical description of rhythmic energy. This facilitates a better comparability, but abstracts from our human understanding of rhythm.

- **Harmony:** Harmony describes the tonality of a composition. In terms of an analytical perspective, it analyses how the spectral energy is distributed among a certain (usually western) scale.

- **Loudness:** Although loudness is not relevant for music similarity, it has to be considered in certain ways. The similarity will make no difference for two identical songs played at different amplitudes. However, on a more global level it has a discriminative notion. It was reported that contemporary music tends to steadily increase on loudness [Serrà 2012]. This was proclaimed as the “loudness war”. It refers to a common habit of modern music production to apply several levels of compression in the post-processing phase. This reduces the dynamic range by pushing silent frequencies up. This results in a sound that is more attractive (subjectively) and demands the attention of the listener (objectively).

- **Noisiness:** Noise behaviour analysis refers to the different recording qualities of the Europeana Sounds collection’s items. Adding these features to the stack would prefer performance over composition, and thus group the records by their quality.

2.1 Audio and music features

Feature extraction is the core of the content-based description of audio files. With feature extraction from audio, a computer is able to recognize the content of a piece of music without the need of annotated labels such as artist, song title or genre. This is the essential basis for information retrieval tasks, such as similarity based searches (query-by-example, query-by-humming), automatic classification into categories, or automatic organization and clustering of music archives. Features extracted from the audio signal are intended to describe the stylistic content of the music such as beat, presence of voice, and timbre.

2.1.1 Overview of content based audio features

The audio features used for the experiments required to evaluate adequate combinations of audio features – distance measures, pre-processing and feature weights – are well-evaluated music content descriptors [Tzanetakis 2002, Lidy 2005, Fu 2011, McFee 2015] widely used in the music information retrieval domain. They provide a good timbral, temporal, rhythmic and harmonic description of the music content.

[6] https://en.wikipedia.org/wiki/Mel-frequency_cepstrum

2.1.2 Psychoacoustic feature set by TU-Wien

Methods from digital signal processing are used, and psycho-acoustic models are considered in order to extract suitable semantic information from music. Various feature sets have been developed, which are appropriate for different tasks.

All features of the Rhythm Patterns family, which will be introduced subsequently, undergo a series of pre-processing steps. Most of them are applied to anneal the sampled audio signal to an audio sensation experienced by human. These psychoacoustically transformed signals provide a better approximation of perceived sound and thus perform better in describing music content.

As the first step, multiple audio channels (e.g. stereo-channels) are averaged to one channel and the audio is split into segments. Short-time Fourier Transform (STFT)[7] is applied to each segment to convert the audio signal into a frequency representation. The Bark scale[8], a perceptual scale which groups frequencies to critical bands of hearing according to perceptive pitch regions, is applied to the spectrogram, aggregating it to 24 frequency bands. Subsequently, the Bark-scaled spectrogram is transformed to the decibel scale and further to the phon scale[9], which incorporates equal loudness curves accounting for the different perception of loudness at different frequencies. Finally, a Sone scale[10] transformation is applied. The Sone scale relates to the Phon scale and describes the perceived loudness in a linear way: doubling on the Sone scale is like a doubling of the loudness to the human ear. Additionally, further psychoacoustic transformation, such as spectral masking and blurring, are applied before the calculation of the features is started. Figure 3 provides an overview of the previously described process.

[7] https://en.wikipedia.org/wiki/Short-time_Fourier_transform
[8] https://en.wikipedia.org/wiki/Bark_scale
[9] https://en.wikipedia.org/wiki/Phon
[10] https://en.wikipedia.org/wiki/Sone

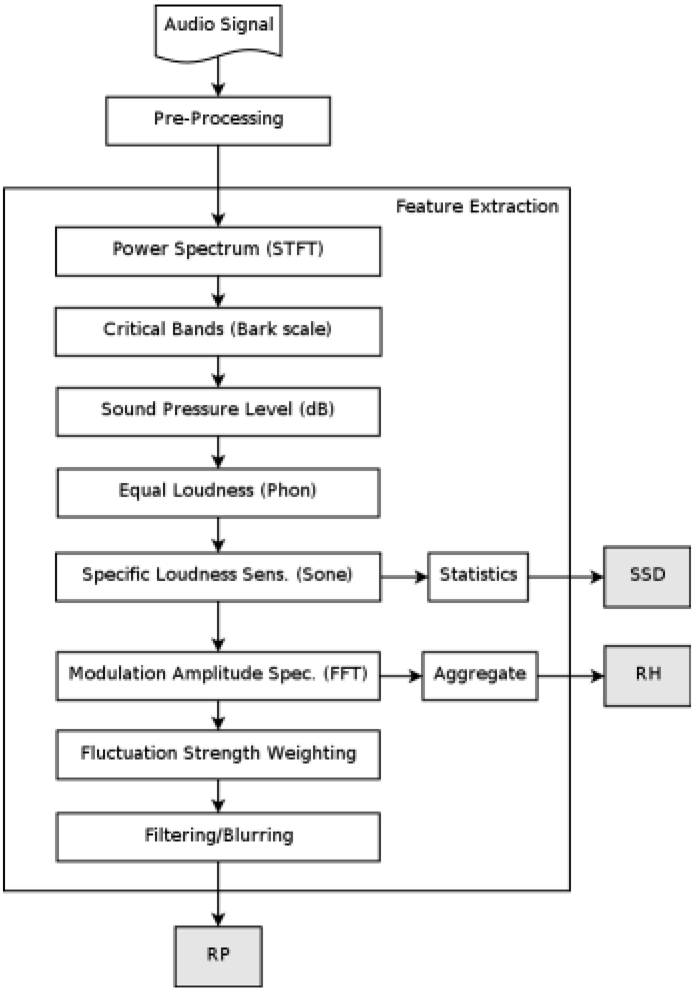

Audio Signal
Pre-Processing
Feature Extraction
Power Spectrum (STFT)
Critical Bands (Bark scale)
Sound Pressure Level (dB)
Equal Loudness (Phon)
Specific Loudness Sens. (Sone)
Statistics
SSD
Modulation Amplitude Spec. (FFT)
Aggregate
RH
Fluctuation Strength Weighting
Filtering/Blurring
RP

Figure 3: Chart depicting the feature extraction process of the Rhythm Patterns feature family.

Rhythm Patterns

Rhythm Patterns (RP) [Lidy 2005] describe modulation amplitudes for a range of modulation frequencies on "critical bands" of the human auditory range, by applying a discrete Fourier transform[11] to the psycho-acoustically transformed Sonogram. This results in a (time-invariant) spectrum of loudness amplitude modulation per modulation frequency, for each individual critical band. These fluctuations in modulation frequency provide a rough interpretation of the rhythmic energy of a song.

Relevance for MIR Pilot: Rhythm is a fundamental property of music. Rhythm Patterns are relevant to distinguish different rhythmic energies in the sound files, and to make similarity calculations more efficient.

Statistical Spectrum Descriptors

The Statistical Spectrum Descriptors (SSD) [Lidy 2005] are based on the previously described pre-processing steps. After the application of the psychoacoustic transformations the mean, median, variance, skewness, kurtosis, minimum and maximum value are calculated subsequently for each

[11] https://en.wikipedia.org/wiki/Fourier_transform

individual critical band of the Bark scale. SSDs are able to capture additional timbral information, compared to Rhythm Patterns, but at a much lower dimension of the feature space.

Relevance for MIR Pilot: Timbre is another fundamental property of music. SSDs are not the best timbre descriptors, but outperform traditional ones due to their overall description of the psychoacoustically transformed audio spectrum.

Further music features of the Rhythm Pattern family - not included in the MIR Pilot

Rhythm Patterns and Statistical Spectrum Descriptors provide an adequate description of spectral and rhythmic properties of music. Thus they were considered to be sufficient for the implementation of the MIR-pilot. The following descriptions of the remaining features of the RP-feature family are provided for reasons of completeness and to have them considered for further improvement.

Rhythm Histogram

Rhythm Histograms (RH) [Lidy 2005] features capture rhythmical characteristics of an audio track by aggregating the modulation values of the critical bands computed in a Rhythm Pattern. Rhythm Histograms provide a much lower-dimensional descriptor for general rhythmic characteristics. The aggregated information is still able to describe the rhythmic energy, although the variance information of the critical bands gets lost.

Discussion for MIR Pilot: Although RH's would require less storage capacity, they are not as descriptive as RP's. The experiments preceding the implementation of the MIR-pilot showed, that the RH's are not able to discriminate satisfactory in this variegated dataset. Thus, it was decided to include the more memory intensive Rhythm Patterns due to their superior precision in describing the rhythmic content.

Temporal Statistical Spectrum Descriptor

Temporal Statistical Spectrum Descriptor (TSSD) [Lidy 2005] features describe variations over time by including a temporal dimension to incorporate time series aspects. Statistical Spectrum Descriptors are extracted from segments of a musical track at different time positions. Thus, TSSDs are able to reflect rhythmical and instrumental changes by capturing variations and changes of the audio spectrum over time.

Discussion for MIR Pilot: TSSDs are usually superior to SSDs due to the additional perspective. The best performances are expected on collections of audio files with similar lengths. In that case the TSSDs are able to describe differences in the structure of the audio file, and in case of music, the structure of the composition. The Europeana Sounds collection is not uniform and contains tracks of various lengths with great variety. Most of the items of the collection consist of 30 seconds samples which are often randomly extracted from the original file. Thus, the advantage of the TSSDs is dampened and the globally calculated SSDs compensate for many problems provided by the 30 second samples.

Temporal Rhythm Histograms

Temporal Rhythm Histograms (TRH) [Lidy 2005] capture change and variation of rhythmic aspects in time. Similar to the Temporal Statistical Spectrum Descriptor, statistical measures of the Rhythm Histograms of individual 6-second segments in a musical track are computed.

Discussion for MIR Pilot: Similar to the TSSD, TRHs are able to capture structural properties of a song. They are similarly able to detect variances in rhythm, such as changes in rhythm or percussive playing styles during the recording. Similarly, they face the same disadvantages concerning the Europeana Sounds collection.

2.1.3 Standard low-level audio features

Mel Frequency Cepstral Coefficients (MFCC)

This feature set[12] was used previously in speech recognition and aims to model human auditory response by transforming it to the Mel scale[13]. The "cepstrum" ("s-p-e-c" reversed) results from taking the Fourier transform (FFT) of the decibel spectrum as if it were a signal. The result shows the rate of change in the different spectrum bands. It is a dominant feature in speech recognition, because of its ability to represent the speech amplitude spectrum in a compact form. It also has proved to be highly efficient in Music Retrieval [Logan 2001]. Representing the rate of change in the different spectrum bands is a good timbre descriptor. The MFCCs are the most commonly used features in music processing [Fu 2011].

Chroma

Chroma Features [Bartsch 2001] represent the 12 distinct semitones (or chroma) of the musical octave. This results in one or a sequence of 12 dimensional vectors where, for example, the bin that corresponds to the pitch class A captures the spectral energy of A0 and all its corresponding subband pitches A1, A2.

Root Mean Square

Root Mean Square (RMS)[14] is a way of comparing arbitrary waveforms based upon their equivalent energy. The RMS method takes the square of the instantaneous voltage, before averaging, and then takes the square root of the average.

Spectral Centroid

The Spectral Centroid (SC) [Tzanetakis 2002] is the frequency-weighted sum of the power spectrum, normalized by its un-weighted sum. It could be described as the centre of gravity or the balancing point of the spectrum. 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.

Spectral Bandwidth

The Spectral Bandwidth (SBW) [Scheirer 1997] represents the weighted spread between minimal and maximal frequency and is calculated similar to the spectral centroid.

Spectral Contrast

[12] https://en.wikipedia.org/wiki/Mel-frequency_cepstrum
[13] https://en.wikipedia.org/wiki/Mel_scale
[14] https://en.wikipedia.org/wiki/Root-mean-square_deviation

The Spectral Contrast (SC) [Scheirer 1997] is calculated from the spectral peaks and valleys and the difference in each sub-band. Strong spectral peaks roughly correspond with harmonic components, while spectral valleys correspond with non-harmonic components such as noise. Thus, Spectral Contrast feature can roughly reflect the relative distribution of harmonic and non-harmonic components in the spectrum.

Spectral Rolloff

The Spectral Rolloff (SRO) [Scheirer 1997] is the frequency below which some fraction k (typically 0.85, 0.9 or 0.95 percentile) of the cumulative spectral power resides. It is a measure of the skewness of the spectral shape and provides an indication of how much energy is in the lower frequencies. It is often used to distinguish voiced from unvoiced speech or music.

Tonnetz

Tonnetz features [Harte 2006] are able to detect changes in the harmonic content of musical audio signals based on a model for Equal Tempered Pitch Class Space[15] using 12-bin chroma vectors. Close harmonic relations such as fifths and thirds appear as small Euclidian distances. Peaks in the detection function denote transitions from one harmonically stable region to another.

Zero Crossing Rate

Zero-crossing rate (ZCR) [Tzanetakis 2002] is a simple, straightforward and inexpensive feature. It measures whether two sets of time series measurements exhibit similar patterns. It is particularly useful to analyse measurements that are corrupted by noise. For example, a measurement with a high zero-crossing rate, such as the number of samples per second that cross the zero reference line, indicates that it is noisy.

Discussion for MIR Pilot: The ZCR is used to group audio files by their recording quality. The Europeana Sounds dataset contains many items that have been digitized from old records and even older formats. Consequently they exhibit a strong noise behaviour resulting from degradations of the original carriers such as shellac or wax tapes. The ZCR groups audio files by their noise behaviour.

Beats per Minute

Beats per Minute (BPM) [Dixon 2007] is a common description of the tempo of a music track. It is calculated from audio events which are detected in the audio signal.

2.2 Audio similarity calculations

This section describes the fundamentals of vector-based audio similarity search algorithms. Audio features are descriptive numbers calculated from the audio spectrum of a track. A good example is the Spectral Centroid which can be interpreted as the centre of gravity of an audio recording. It describes the average frequency weighted by its intensity and distinguishes brighter from darker sounds. Such features are usually calculated for several intervals of a track and finally aggregated into a single vector

[15] https://en.wikipedia.org/wiki/Pitch_class

representation. The latter step, which is a requirement for many machine/statistical learning tasks, is accomplished by calculating statistical measures such as mean or standard deviation.

In the following example, the Spectral Centroids of 10 different tracks are provided using their mean and standard deviation aggregations. Thus, the Spectral Centroid feature (-set) is represented by a two-dimensional feature vector such as the following example:

```
IDX    Mean                Standard Deviation
0      1517.5993814237531  291.1855836731788
```

In this example the centre frequency is 1518 Hz and it deviates by 291 Hz. These numbers already describe the audio content, and can be used to find similar tracks. The common approach to calculate music similarity from audio content is based on vector difference. The assumption is that similar audio feature-values correspond with similar audio content. Thus, feature vectors with smaller vector differences correspond to more similar tracks. The following data represents the extracted Spectral Centroids of our 10-tracks collection:

```
IDX    Mean                Standard Deviation
0      1517.5993814237531  291.1855836731788
1      1659.1988993873124  327.64811981777865
2      1507.4617047141264  340.8830079395701
3      1597.6019371942953  507.1007933367403
4      1498.8531206911534  288.3780838480238
5      535.5910732230583   89.90893994909047
6      2261.4032345595674  353.5971736260454
7      2331.881852844861   406.33517225264194
8      1868.690426450363   342.7489751514078
9      2204.6324484864085  328.94334883095553
```

The tracks are indexed and the content-based search is using the track with index 5 to query for similar items. This step requires a similarity metric which defines how the vector distance has to be calculated. The most common choices are the Manhattan (L1) and Euclidean (L2) distance measures. The Euclidean Distance is the square root of the sum of squared differences of two vectors.

To calculate the Euclidean Distance between track 5 and track 0

```
IDX    Mean                Standard Deviation
0      1517.5993814237531  291.1855836731788
5      535.5910732230583   89.90893994909047
```

The similarity search takes the difference

```
       982.008308          201.276644
```

Squares them to get the absolute magnitude:

```
       964340.317375       40512.287309
```

and takes the sum of these values:

```
       1004852.6046840245
```

Per definition the square root has to be calculated from the sum, but this step is normally skipped for efficiency, because it does not alter the ranking. Calculating the distance from all feature vectors in the collection, the similarity search retrieves a list of distance values where the smaller distances correspond to more similar audio content and the higher values should sound more dissimilar.

```
IDX    Distance
0      1004852.6046840245
1      1319014.4646621975
2      1007520.5071585375
3      1301916.1177259558
4      967263.7731724023
5      0.0
6      3047959.100796666
7      3326786.1254441254
8      1841081.968976167
9      2842836.5609704787
```

To retrieve a ranked list of similar sounding tracks, the list of vector distances has to be in ascending order

```
IDX    Distance
5      0.0
4      967263.7731724023
0      1004852.6046840245
2      1007520.5071585375
3      1301916.1177259558
1      1319014.4646621975
8      1841081.968976167
9      2842836.5609704787
6      3047959.100796666
7      3326786.1254441254
```

This so-called vector space model is predominant in content-based multimedia retrieval. The most crucial and problematic part is feature crafting. If the extracted numbers do not describe the audio well enough, the vector-based similarity will also fail to provide results that are perceived as similar.

The described state-of-the-art approach requires the availability of all feature vectors of all items of a collection. Thus, the feature vectors must be stored. No matter which retrieval approach (pre-calculated / indexed / on demand) is used, all features will be required at a certain time. Because feature extraction is an expensive task, in terms of processing resources, the extracted features should be stored and easily accessible using a common data format.

3 Implementation

The previous chapters introduced the principal objectives of this task, as well as concepts and technologies available to facilitate them. This chapter will focus on the concrete implementation of the MIR-pilot.

3.1 Composite feature-set used in the MIR-Pilot

Chapter one introduced the audio features that were used in the implementation, and discussed their relevance to the MIR pilot for Europeana Sounds. Referring to the five music properties mentioned in the introduction, which have been chosen to describe music similarity upon (timbre, rhythm, harmony, loudness and noisiness)., the previous introduced content-based audio and music features are placed into the five groups as depicted by Figure 4.

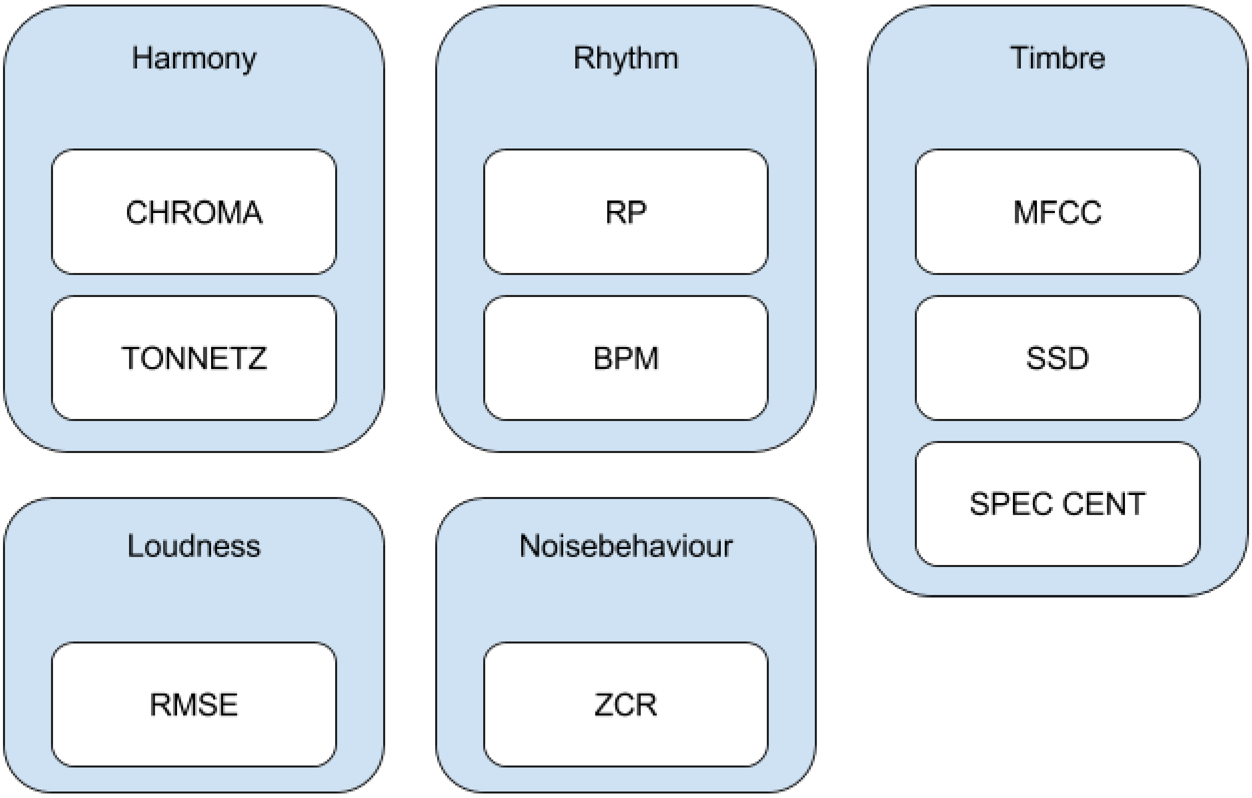


Figure 4: Overview of music property categories and the content based music features used to describe them. Timbre: Mel Frequency Spectrum Coefficients (MFCC), Statistical Spectrum Descriptors (SSD), Spectral Centroid (SPEC CENT), Rhythm: Rhythm Patterns (RP), Tempo / Beats per Minute (BPM), Harmony: Chroma, Tonnetz, Loudness: Root Mean Squared Error (RMSE), Noisebehaviour: Zero Crossing Rate (ZCR)

3.2 Evaluation of appropriate distance measures and normalization methods

Distance measures have a major impact on similarity calculations. Unfortunately, extensive comparative performance evaluations of different distance measures have not been reported, yet. Thus, a preceding evaluation was performed to assess appropriate distance measure and normalisation method combinations for the MIR pilot.

Exhaustive experimentation was applied, using the audio features used in the implementation of the MIR pilot and a selection of 18 distance measures [Cha 2007]:

- Chebyshev Distance[16]
- Sørensen–Dice coefficient[17]
- Canberra Distance[18]
- Wave Hedges [Cha 2007]
- Cosine Distance[19]
- Taneja Distance[20]
- Divergence [Cha 2007]
- Symetric chi square divergence [Taneja 2005]
- Braycurtis Dissimilarity[21]
- Kulbeck/Leibler Divergence[22]
- Czekanowski Distance [Cha 2007]
- Kumar Johnson [Cha 2007]
- Pearson Chi square [Cha 2007]
- Euclidean Distance (L2)[23]
- Jaccard Distance[24]
- Manhattan Distance (L1)[25]
- Lorentzian Distance [Cha 2007]

The evaluation attempted to answer the following questions:

- Which distance measures work best for a certain audio feature?
- Which normalization method is most appropriate for a certain distance measure on a certain audio feature?
- Is the performance of a distance measure constant for different result-list lengths?
- Which method of combining audio-features performs better?

[16] https://en.wikipedia.org/wiki/Chebyshev_distance
[17] https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient
[18] https://en.wikipedia.org/wiki/Canberra_distance
[19] https://en.wikipedia.org/wiki/Cosine_similarity
[20] http://www.mtm.ufsc.br/~taneja/book/node20.html
[21] https://en.wikipedia.org/wiki/Bray%E2%80%93Curtis_dissimilarity
[22] https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence
[23] https://en.wikipedia.org/wiki/Euclidean_distance
[24] https://en.wikipedia.org/wiki/Jaccard_index
[25] https://en.wikipedia.org/wiki/Taxicab_geometry

Figure 5 shows intermediate results of the evaluation for the MFFC feature-set. The cut-off refers to the result list length. It can be observed that the accuracy values drop as expected with increasing result-list length. The left chart displays the performance without normalization applied. Obviously the performance is superior by applying appropriate normalization such as Min-Max or Standardization. Unit length normalization is usually only a good choice for histogram based feature-sets.

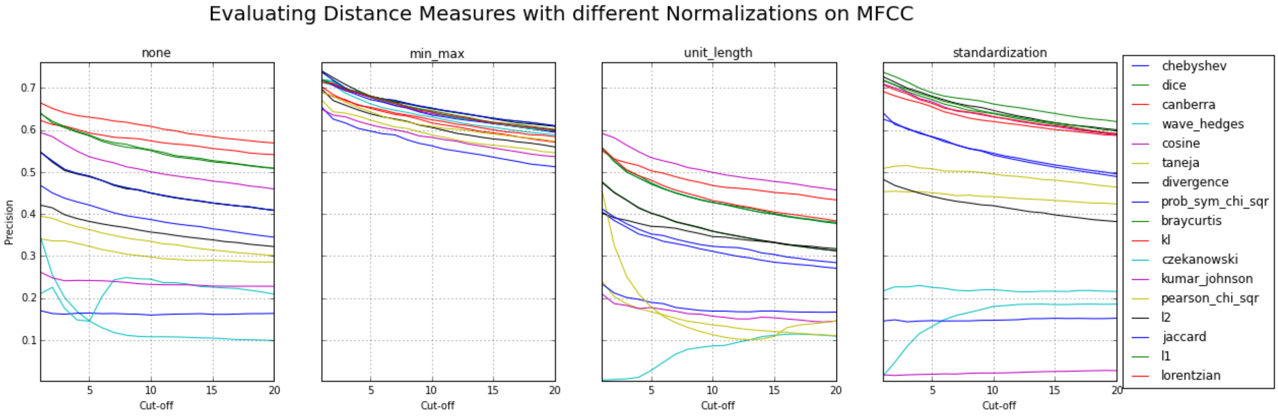


Figure 5: Results of the preliminary evaluation on distance measure and normalisation method combinations. The charts depict similarity estimation performance values measured in average precision at different result list lengths (cut-offs).

Intermediate results of the evaluation were analysed and interpreted to decide which distance measure to use, which normalization method to apply during the pre-processing phase, and how to combine the different audio features into a combined feature space.

Distance measure

No general pattern could be identified on which distance measure works best for all features. Every audio feature had a different best performing distance measure/ normalization method combination. A general observation was that L1 based metrics usually rank highly. Among them the Canberra distance, which is also a derivative of the Manhattan Distance but includes an implicit normalization step. Nevertheless, different normalisation methods showed to have a positive effect on the performance of the Canberra distance. Although, the Canberra distance hardly peaked in terms of performance, it was mostly top-ranked and the difference to the top-performing distance measure was negligible. This distance measure also provides good and stable results with increasing result list length. Thus, it was decided to use this distance measure for the MIR-pilot.

Normalisation method

The experiments showed that choosing the right normalisation method can increase the accuracy of the evaluated system by up to 10% or more. This highly depends on the feature set, the underlying dataset and the distance measure applied. Nevertheless, feature normalisation introduces a certain amount of complexity depending on the applied method. It is often required to reprocess the entire collection on expansion. It was decided that the Canberra distance would be used, which implicitly scales the feature values. Thus, a computationally demanding pre-processing of the feature space is not necessary. Although improvements were observed by explicitly applying normalization to the feature space, the increase in accuracy was generally negligible.

Combination of content based audio features

Two methods are commonly used to combine different feature spaces. One concatenates the feature vectors of each audio feature into a single vector. This results in a single vector for each song which contains all the attributes of all features. This approach is also referred to as *early fusion*. The other common approach is *late fusion*. In this approach the similarities are calculated for each feature separately and the distinct similarity values for each song are combined arithmetically or geometrically. The preceding evaluation showed that late fusion performs better than early fusion. In terms of computation time no significant difference was observed.

3.3 Evaluation of appropriate feature weights

Content-based audio and music features attempt to capture certain aspects of music. To provide an ensemble description of a recorded music track requires the use of multiple features. It is important to consider is that, independent from the approach to combine these features, all similarity results of the distinct audio features have the same weight in estimating the final music similarity. This could lead to inferior performance. The collection of features chosen for the implementation of the MIR-pilot provides a good example. The Zero crossing rate feature has been added to capture effects of audio degradation within old records. This audio property has to be considered in the similarity estimation, but it is not as important as rhythmic, timbral or harmonic similarity. Thus, a common approach to reduce overrated influence of distinct features is to apply weighting to them. Feature weight estimation and optimization is approached empirically. Currently there is no generalized or automatic approach available. For the MIR-pilot the weighting was estimated through a predefined set of similar records. During an iterative process the weights of the different features were adapted. If the changes were reflected by an improvement of the similarity values of the predefined set these weights were kept and further improved. The final feature weighting used for the implementation of the MIR pilot are provided in Table 1:

Table 1: Overview of content based music features and their evaluated weights.

Category	Feature	Description	Weight	Category Weight
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%

The distribution of feature weights is further depicted in Figure6. The final setting favours timbre as the most important music property, followed by rhythm and harmony. Loudness and noisiness are considered in the similarity calculation process but have a minor impact on the overall ranking.

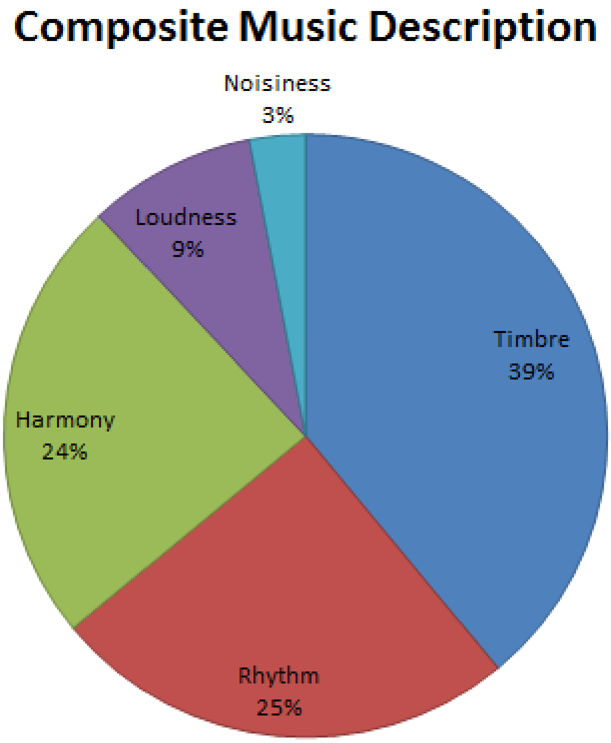


Figure 6 Overview of the distribution of feature weights grouped by categories.

Figure 7 provides an overview of the similarity calculation process. Similarity is estimated using a late fusion approach. Thus, similarities are calculated for each feature distinctively. These estimates are calculated through the application of a distance measure. The resulting distance values are converted into similarity estimates by dividing all distances of a feature by their maximum and subtracting them from one. This results in similarity values for each audio feature that are scaled between 0 and 1, where 1 refers to identical items and 0 to the most dissimilar. In the next step, these similarity estimates are multiplied by the empirically estimated feature weights and summed to retrieve an average estimate of similarity. The division by the number of features as required for calculating the arithmetic mean is omitted due to its indifference on the final ranking of the result list.

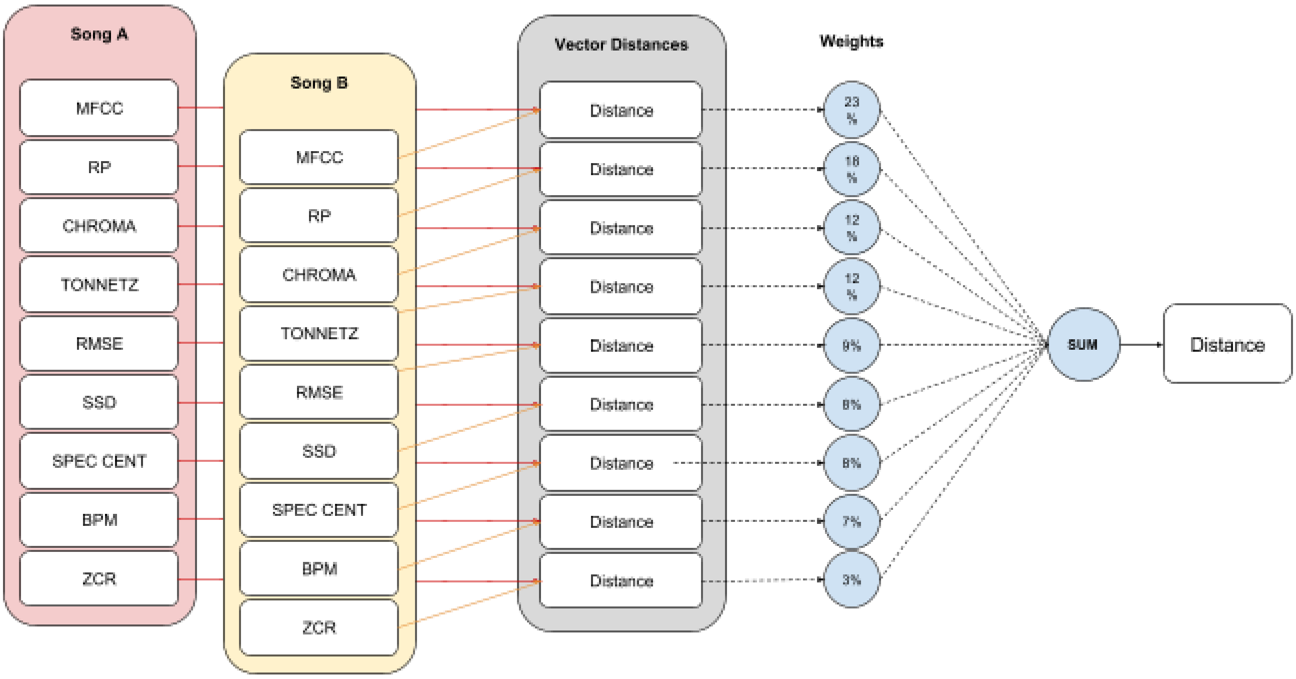


Figure 7: Overview of the music similarity estimation process. Vector distance of two song-feature-vectors is calculated using a late-fusion approach. Weights are applied to the distance values of the distinct features. The sum of all distances is calculated to aggregate final similarity estimation.

3.4 User interface

The user-interface of the MIR pilot is aligned to the current design of the Europeana search interface. The HTML-code for the search result webpage has been used and extended to output the results of the similarity calculations. The pilot only implements the query-by-example feature. All other faceted search functionalities provided by the Europeana webpage are not available and thus greyed-out in the interface. The MIR pilot supports the following use cases:

- **Term-based queries:** The term-based query option accepts text-based input and uses them to query the "title" attribute of the metadata. This attribute seems to be overloaded and commonly contains also information about composer and performer.

- **Query by Example:** By supplying an example song, the system searches for similar ones based on their acoustic properties.

Term-based queries were added 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". The search can be initiated by supplying the query-terms in the search textbox at the top of the web-form. After submitting the query the system responds with a list of 24 matches which corresponds to the current page size of the Europeana search interface. To simplify experimentation and evaluation of the system, a web-based audio player has been added directly into the result page.

The query by example functionality can be triggered by clicking on the green button with the magnifying glass symbol for the desired query song. The system then calculates the similarities to all records of the collection, ranks them in descending order by their similarity and returns the top 24 entries of the result-list.

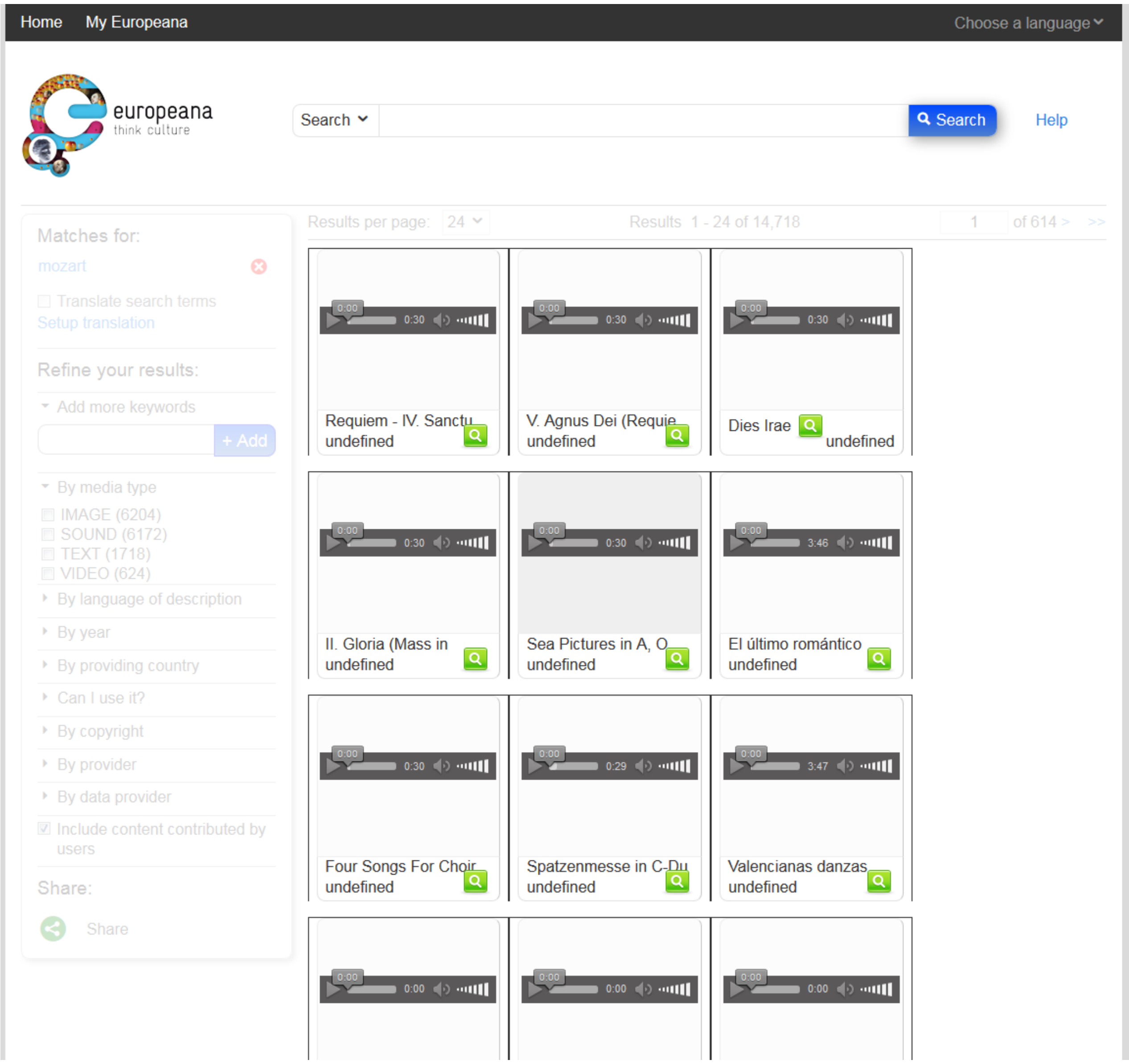


Figure 8: Example of a query result of the MIR-pilot. The web-form features a text-field on top which accepts query terms to query for music metadata. The green magnifying glass triggers the query-by-example search.

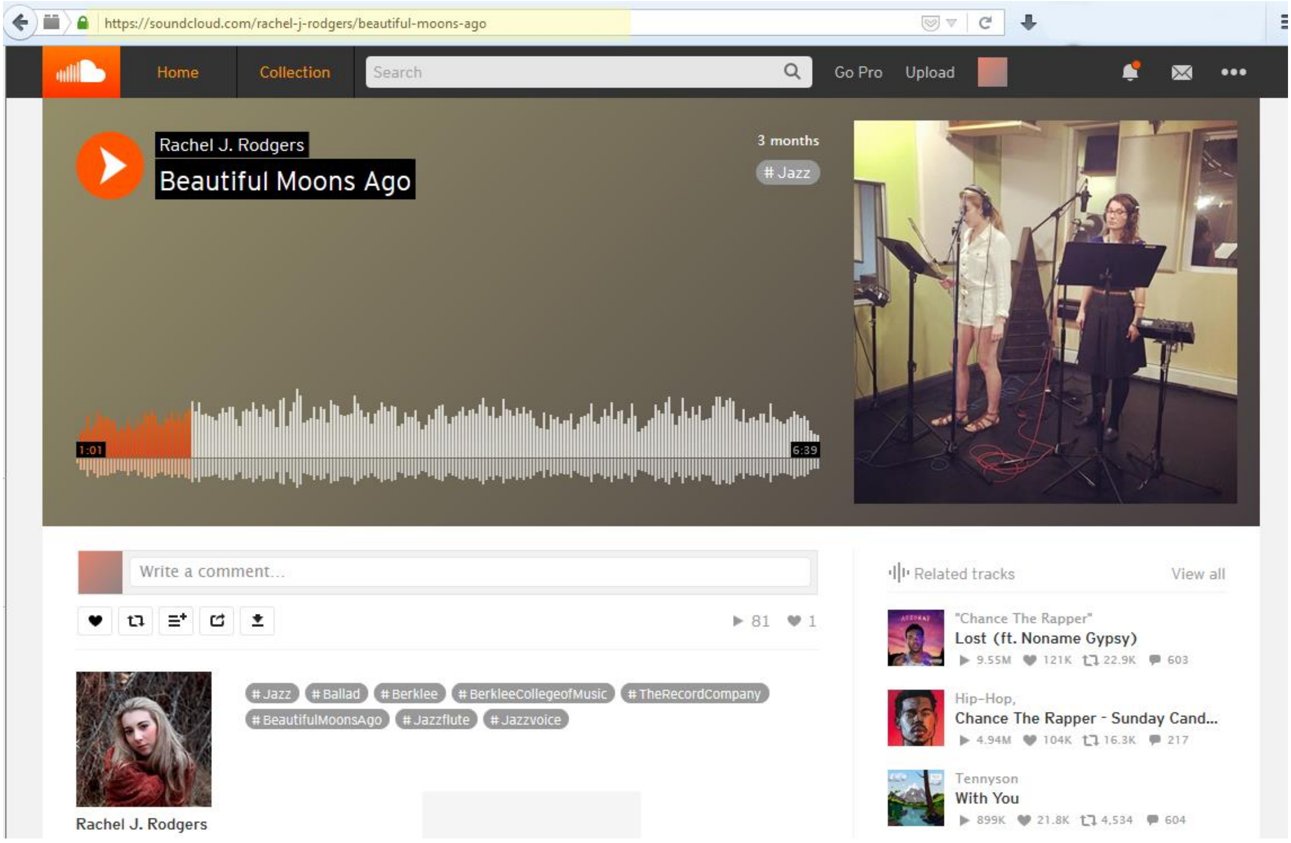
To simplify the evaluation of the query-by-example functionality, the query-song is added on top of the result-list. Thus, the first entry on a result-page is the query-song. This facilitates better comparability since it eliminates unnecessary back- and forth navigation. The result-item right next to the query-song is the first entry of the similarity result-list and has the highest calculated similarity. Thus, according to the system this entry represents the most acoustically similar audio item within the collection. The remaining result items are ordered by their similarity and displayed from left to right and from top to bottom.

3.5 Usage of external content to query for Europeana content

To demonstrate further possibilities of how to apply query-by-example or MIR technologies in general, the query-by-example approach has been extended to accept also content which is not contained in the Europeana Sounds collection.

This demonstration used the SoundCloud API[26] which facilitates computational access to the SoundCloud music streaming service. To trigger the external query by example mechanism the following steps are required:

1. On the SoundCloud webpage search for a desired song which should be used to find similar sounding items within the Europeana Sounds collection.
2. Click on that song to navigate to the song-page.
3. Copy the address of the song-page from the URL-bar of the browser (see Figure 9).

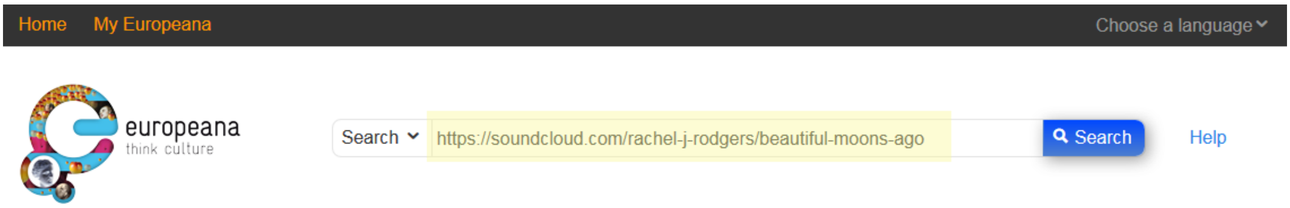


https://soundcloud.com/rachel-j-rodgers/beautiful-moons-ago

Figure 9: Example of a SoundCloud song page. The URL of the page which will be used to query for Europeana content is highlighted on top.

4. Navigate to the MIR-pilot
5. Supply the URL of the SoundCloud song-page to the search text-field and press search (see Image 9).

[26] https://developers.soundcloud.com/docs/api/guide



https://soundcloud.com/rachel-j-rodgers/beautiful-moons-ago

Figure 10: Example of supplying a SoundCloud song-page URL as search term to query for Europeana content.

This initiates the system to login to the SoundCloud via its API and to retrieve the corresponding metadata. In a next step the actual audio content is downloaded and the feature extraction process is initiated. The resulting feature vector is then used to calculate similarities to Europeana Sounds records. The process of estimating these similarities is identical to the one applied to query within the Europeana collection.

4 Evaluation

This section describes the evaluation approach in theory, and how it is applied to the evaluation of the MIR pilot. The applied evaluation method is the standard information retrieval evaluation approach, also called the Cranfield process[27] that dates to the late 1950s. It is based on a document collection, a test suite of expressible queries and a set of relevance judgements, also called gold standard or ground truth. During the evaluation, the results of the queries are compared against the relevance judgements of the ground truth data and a set of expressible metrics are calculated that are used to make different systems comparable. The results from the user evaluation of the MIR pilot and the recommendations derived from these results are described in the last section of this chapter. Next to these results of the user evaluation, the participants' experience with the MIR pilot is compared with the calculated similarity estimation which provides a hypothesis for an online survey that can be distributed in the future.

4.1 MIR Pilot evaluation data and ground truth

For the evaluation of the MIR pilot the following data corpus and ground truth structure were used:

Document Collection: The evaluation dataset will consist of 312,096 audio items which were downloaded from Europeana via the Europeana API[28]. The data consists of mp3- and FLAC-encoded[29] audio data of variable size, sample rate and bitrate. The audio content varies from speech, to recorded radio broadcast, music varying in age, style and quality, nature and ambient sound recording. For every audio file all corresponding metadata-items, as available via the Europeana API, has been downloaded. Unfortunately, great parts of the Europeana audio content is only provided in form of 30 seconds pre-listening samples. Although, 30 seconds exhibit enough information to calculate music similarities, it has to be considered that these estimates refer to only a snippet of an entire recording and thus results may be biased. A more detailed overview of the dataset is provided in Section 3.2.

Test suite of Expressible Queries: These queries define what will be evaluated. In the MIR pilot these corresponds to the query-by-example (QBE) functionality. As explained in the previous chapters, QBE takes an instance as query and searches for similar items. Consequently, queries correspond to example audio files where the expected result-list consists again of audio files that "sound similar".

Ground Truth/Relevance Judgements: To compare the query result against an expected result, a ground truth is required that labels the data according such expectations. Creating a ground truth for large datasets is highly expensive in terms of time and budget constraints. The advantage of the data provided by Europeana is that all data items, including their corresponding metadata, have been curated and edited by often highly trained staff at national libraries and audio-visual archives. The inter-rater agreement ratios used to assess the quality of ground truth assignments consider untrustworthy raters, whereas annotators from national libraries and audio archives can be considered trustworthy. Further, the subset of content used in this pilot is dominated by the DISMARC collection which is known

[27] http://www.informationr.net/ir/8-3/paper152.html
[28] http://labs.europeana.eu/api
[29] Major audio formats (mp3, wav, flac, etc.) are supported by most of state-of-the-art audio feature extractors. To process unconventional audio formats an adequate decoder has to be added to the feature extraction chain.

from previous Europeana projects to have good title information (dc:title), good creator information (dc:creator, dc:contributor) including roles and also provides genre classification (dc:subject). Based on this, it is plausible to use the metadata of the Europeana dataset to create a ground truth without consulting further annotators.

Yet, the automatic generation of the ground truth from the Europeana metadata has to be similarly plausible. Music similarity is a highly subjective concept and metadata entries are required that facilitate objective comparability of music. Generally, music similarity can be defined by a mixture of timbre, rhythm, pitch, key, etc. Based on empirical evaluations of the Europeana metadata it was observable that the data is rich in literal descriptions of music properties. As an example the following query would provide a substantial description of the acoustic properties of a musical track: ["string", "quartet", "C Minor", "allegro"]. These terms describe the used instrumentation and thus the expected timbre, as well as the key. The term “allegro” is historically overloaded and refers to the tempo. It provides a hint to rhythm, as well refers to the mood-related characteristics joyful, lively and fast. Executing a string-based search on the Europeana metadata using these terms results in a result-list of 14 entries:

keywords = ["string", "quartet", "allegro", "C Minor"]

- String Quartet N. 8 In C Minor Op. 110: Allegro Molto;Dmitry Shostakovich (Performer);Classical;00:03:03, 00:00:30 (preview duration)
- "String Quartet In C Minor, D 703, ""Quartettsatz"": Allegro assai";Franz Schubert (Performer), Chamber Orchestra Kremlin (Performer), Misha Rachlevsky (Conductor);Orchestral Music;00:09:34, 00:00:30 (preview duration)
- String Quartet No. 12 in C minor, Op. posth. D. 703 (Quartettsatz) - Allegro assai;Franz Schubert (Performer), Enesco Quartet (Ensemble);Classical;00:09:02, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, I. Allegro;Ludwig Van Beethoven (Performer);Other;00:07:30, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, IV. Allegro;Ludwig Van Beethoven (Performer);Other;00:03:57, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, I. Allegro;Ludwig Van Beethoven (Performer);Orchestral Music;00:07:31, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, I. Allegro;Ludwig Van Beethoven (Performer);Orchestral Music;00:07:30, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, IV. Allegro;Ludwig Van Beethoven (Performer);Orchestral Music;00:03:57, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, I. Allegro;Ludwig Van Beethoven (Performer);Romantic;00:07:30, 00:00:30 (preview duration)
- String Quartet No.4 in C minor, Op.18 No.4, IV. Allegro;Ludwig Van Beethoven (Performer);Romantic;00:03:57, 00:00:30 (preview duration)
- String Quartet No.4 In C Minor, Op.18 No.4, I. Allegro;Ludwig Van Beethoven (Performer);Classical;00:07:31, 00:00:30 (preview duration)
- String Quartet No.4 In C Minor, Op.18 No.4, Iv. Allegro;Ludwig Van Beethoven (Performer);Classical;00:03:55, 00:00:30 (preview duration)
- String Quartet in C minor, Op. 51, No. 1: Allegro;Mandelring Quartett Johannes Brahms (Performer);Classical (Core - Classical);
- String Quartet in C minor, Op. 51, No. 1: Allegro;Mandelring Quartett Johannes Brahms (Performer);Classical (Core - Classical);

Usually the class sizes in standardized information retrieval test-collections are bigger than just 10-20 instances. Though, the advantage of the suggested approach is that it facilitates highly customized evaluations, whereas standard collections are mostly dedicated to a specific task. By identifying diverse relevant music-property-descriptions, it is possible to evaluate the MIR-pilot in multiple facets. It is possible to draw conclusions of its performance on different types of audio, such as classic or contemporary music, spoken word, and animal or ambient sounds, based on recording quality.

The provided example is a highly specific query. To provide a broad overview of the query-by-example algorithm implemented in the MIR pilot, queries at different levels of particularity will be applied.

Evaluation Metrics: The evaluation has been executed in different runs. Each run evaluated the system based on the ground truth provided by a previously defined query. To assess the performance of the system, its precision is calculated as described:

- from the query-generated ground truth data take one song and compute the result list.
- compare the entries of the result list against the ground truth.
- the precision defines how many entries of the ground truth are presented in the result list.

This procedure is repeated for every entry of the query-generated ground truth. If the number of items of a ground truth is very large, the evaluation set is subsampled to 1000 entries. For example the query for traditional Italian “Tarantella” provides 152 metadata entries. Thus, this represents the size of the ground truth. During an evaluation run, every entry is used as input for the query-by-example algorithm to compute the list of similar sounding entries. Each result-list is intersected with the ground truth data to find out how many of the computed similar sounding entries are part of the ground truth. In the given example a precision of 100% would refer to all entries of a result-list being part of the ground truth and thus having the term “Tarantella” applied in their metadata.

The computed result-list is ranked by the calculated similarity. Thus, the items on top of the list are expected to be more similar than those at the bottom. To reflect this behaviour in the evaluation, the precision is calculated using result-lists of different lengths:

- Length 1: Corresponds to the first entry which is according to the computation the most similar song.
- Length 2 and 3: Searching for music requires more time and attention than looking at images, thus the first three entries are the most important entries when browsing or searching for similar songs.
- Length 5 and 10: These precision values are provided to give an overview of the system performance concerning longer result lists. A length of 10 is generally known to be a threshold for user attention.
- Length 24: This represents the number of entries per page of the Europeana Web page’s search result at the time of the evaluation.

These different lengths of result list sizes provide a broad overview of the general performance of this approach to music similarity computations on the Europeana Sounds collection. The choice of distance measures, normalization methods and feature weights has been focussed on optimizing the system performance for the list entries 1 to 10. If future design improvements decide to minimize the page size to fewer entries, no performance degradation is to be expected.

4.2 Dataset overview

For the dataset creation 400,615 entries were downloaded via the Europeana API. The JSON formatted data was stored for each entry and aggregated into a complete dataset. During the aggregation step, empty and erroneous JSON responses were removed. After this cleansing and aggregation step 389,120 entries remained in the set. In parallel the corresponding audio data was downloaded and the previously described audio features were extracted. It was not possible to obtain the audio files to all entries in the dataset, and some of them were corrupt or failed to be processed for the audio feature extractors. 327,261 RP-features and 323,664 features using the librosa[30] library were extracted. The intersection of aggregated metadata and available audio features results in a final dataset size of 312,096 entries.

4.2.1 Dataset statistics

Data Providers:

- Number of Data Providers: 1,002
- Top-10 Data Provides:
  - Preiser Records; Austria (30,738 items)
  - National Library of Spain (10,462)
  - The Orchard (8,860)
  - JSP Records (7,686)
  - Hacienda Records (6,537)
  - Carinco AG (6,534)
  - Ovação; Portugal (6,032)
  - Gesellschaft für Historische Tonträger; Austria (5,985)
  - Duck Records; Italy (5,691)
  - Arts Productions Ltd (5,644)

Data Aggregators:

- Number of Data Aggregators: 20
- Top-10 Data Aggregators:
  - DISMARC (286,189)
  - Judaica Europeana (8,033)
  - Hispana (2,373)
  - Deutsche Digitale Bibliothek (1,661)
  - OpenUp! (1,097)
  - National Library of Finland (183)
  - EuropeanaLocal Romania (136)
  - HOPE - Heritage of the People's Europe (123)
  - The Natural Europe Project (104)
  - Europeana 1914-1918 (47)
- Median number of sound files per aggregator: 39

Collections:

- Number of collections: 34

[30] https://github.com/bmcfee/librosa

4.2.2 Technical overview of the dataset

The download of the dataset was managed via the Europeana API. Audio files were downloaded sequentially to keep the number of requests per minute within an acceptable rate and to avoid stressing the bandwidth and server capacities of the distinct partner sites.

The download of the audio files took approximately six days. This was due to some slow responding servers from some content partners. It has to be considered that a processing time of one second per items results in a total processing time of 4.1 days on the entire collection. The download of the JSON-formatted metadata took about two to three days.

Feature extraction was executed in two separate batches, one for each feature extractor. Optimized performance was not an objective of this task. Running both feature extractor on a file took on average about 6 seconds for 30 seconds audio samples and up to two five minutes for 2 hours or longer recordings. A rough estimation for extracting the features of the entire collection resulted in about 50 days of processing time. Thus, a parallel feature extraction framework was implemented to process the files in up to 16 processes simultaneously. The feature extraction processes were also identified to be demanding in terms of memory usage. While 30 seconds to five minutes require approximately 1GB of memory, 2 hours or more already exceed 16GB of memory usage.

Disk space requirements to store the audio, metadata and feature files are as follows:

- audio files: 42.2GB
- JSON-formatted metadata: 3.2GB
- aggregated metadata stored in a binary format: 270MB
- extracted features in binary format: 9.5GB

4.3 User evaluation

In order to get an end-user perspective on the results of the MIR pilot, a user evaluation was executed by the Netherlands Institute for Sound and Vision. In sessions of 90 minutes 13 participants provided feedback on their experience with the MIR pilot, based on the similarities and differences they experienced between songs from several genres. Snippets of 30 seconds were played from different genres, always starting with one reference track. After playing the reference track, three other tracks from a particular genre (the three that came on top using the content based search) were played. After each of these three tracks the participants needed to indicate if they experienced similarities and/or differences with the initial reference track. To note these indications in an orderly fashion, a scale between 1 (very different from reference track) and 5 (very similar to reference track) was used[31]. These grades were given for the general experience of the tracks, but also for the tempo, rhythm, harmony,

[31] The scale that was being used is as following: 1 stands for experienced as very different from reference song, 2 for different from reference song, 4 for similar to reference track, 5 for very similar to reference track, 3 was put down if participants experienced a balance between elements that were different and elements that were similar to the reference track.

timbre, instrumentation and quality of the recording as a separate element of the music. Note that the general experience of differences and similarities got a separate grade by the participants, so this is not necessarily the average of all the elements. After the participants compared the nine reference tracks with three songs each (27 tracks in total), questions about the concept of music information retrieval were asked.

Most participants did not make a distinction between tempo and rhythm, nor between harmony and timbre in their grading. Apart from the grade on the scale of 5, the explanation for the experienced similarities and differences were being noted as well during the evaluation. There were 3 different types of users involved in the evaluation: music listeners, participants who also played music themselves (and could, for example, differentiate between harmony and timbre more easily) and those who worked professionally with audio by recording music and/or radio (this group observed a lot about the different recording techniques and for example the balance between the instruments from a more technical point of view).

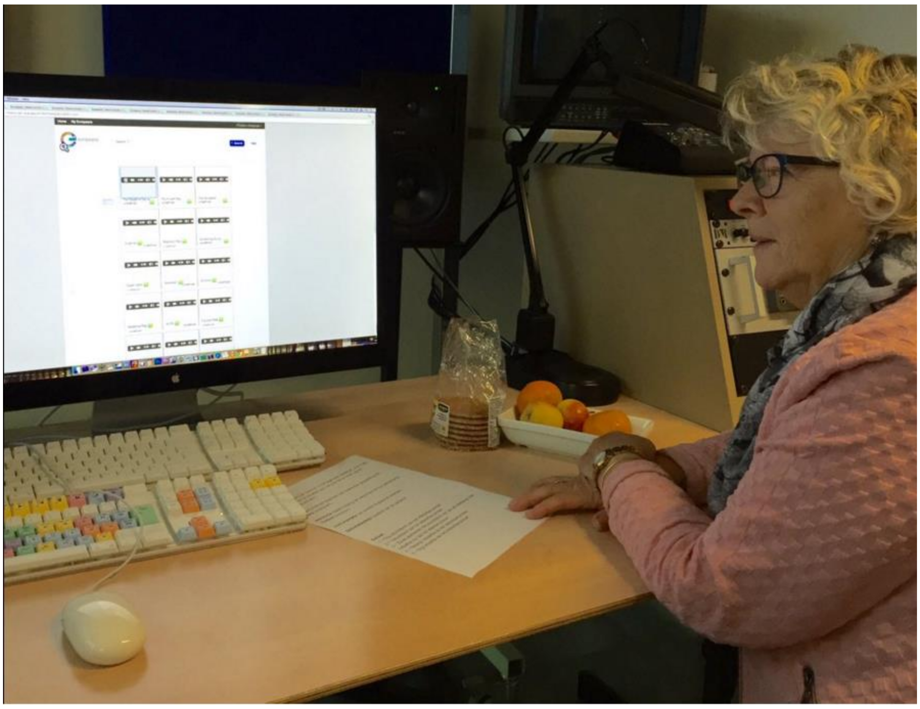


Figure 11: Participant of the MIR user evaluation at the Netherlands Institute for Sound and Vision

Eight different music genres were tested in this user evaluation: ragtime (twice), smooth jazz, classical jazz, requiem, rock, tango, flamenco and classical. The reference tracks were provided by AIT. From each of the categories of the automatic evaluation one or two example songs were chosen randomly and provided as offline webpages for the user-evaluation. During the user-evaluation a sub-sample was chosen out of these examples. Before comparing the user evaluation results with the results from the calculated similarity and summarising the answers to the last questions (about the concept of MIR), the user evaluation results accompanied with comments from the participants will be described underneath first (in the same order as during the evaluation).

4.3.1 User evaluation results

Reference track one: The Ragtime Dance (genre: ragtime)[32]

Track one: Palm Leaf Rag[33]. Average general grade: 4.54. Comments: ‘this track is more serious and dark’, ‘same music category’, ‘I would dance to the reference track more easily’. Lowest scores on tempo and rhythm (both scored 3.69 on those elements); most participants commented that this track was slower than the Ragtime Dance. Highest score on instrumentation (4.62), one participant heard a difference between the types of piano being used in both songs (regular piano versus a fortepiano).

Track two: The Nonpareil[34]. Average general grade: 4.54. Comments: ‘more melody’, ‘sounds a lot like the reference track’. Lowest score on timbre (4), ‘lower tones’, ‘sounds like this track needs to come from further away than the reference track’. Highest score on instrumentation (4.54), ‘almost no difference’, ‘this track is recorded more classical and is recorded with more resonance than the first track’ (last comment by a former audio-technician who recorded music professionally).

Track three: Eugenia[35]. Average general grade: 3.92. Comments: ‘this track should be played later in the night than the reference track’, ‘lighter tone colour’, ‘harmony is more simple’, ‘less clear sound colour’. Lowest score on tempo and rhythm (both got 3.46), ‘slower’. Highest score on instrumentation (4.69).

Reference track two: Oasis of Love (genre: smooth jazz)[36]

Track one: Patterns in the Rain[37]. Average general grade: 2.92. Comments: ‘same category’, ‘totally different’, ‘this track has nothing to do with the reference track’. Lowest score on instrumentation (2.54), ‘with vocals, this causes a different experience’. Highest score on quality, tempo and rhythm (4.54; 3.85; 3.85), ‘this track is recorded drier than the reference track’, ‘tempo gives calmness’.

Track two: Beautiful Moons Ago[38]. Average general grade: 1.85. Comments: ‘different genre’, ‘this is more jazzy’, ‘totally different, this is easy-listening music, so not filmic music (like the reference track)’, ‘this is music that would be played in a bar or restaurant’. Lowest score on timbre (1.85), ‘much lighter’ ‘very relaxed’. Highest score on quality, tempo and rhythm (4.69; 3.77; 3.77), ‘emphasis of recording balance is on the vocals’.

[32] Link to the Ragtime Dance on Europeana: http://www.europeana.eu/portal/record/2023601/oai_eu_dismarc_MACR_USA370964416.html?start=1&query=title%3AThe+Ragtime+Dance&startPage=1&qt=false&rows=24

[33] Link to 30 second snippet of Palm Leaf Rag: http://econnect.ait.co.at/samples/orchard/843/436/053/843436053277/843436053277_1_41_CLIP.mp3

[34] Link to 30 second snippet of the Nonpareil: http://econnect.ait.co.at/samples/orchard/843/436/053/843436053277/843436053277_1_29_CLIP.mp3

[35] Link to 30 second snippet of Eugenia: http://econnect.ait.co.at/samples/orchard/843/436/053/843436053277/843436053277_1_47_CLIP.mp3

[36] Link to Oasis of Love on Europeana: http://www.europeana.eu/portal/record/2023601/oai_eu_dismarc_15BR_USBK41000007.html

[37] Link to 30 second snippet of Patterns in the Rain: http://econnect.ait.co.at/samples/orchard/884/385/251/884385251413/884385251413_1_11_CLIP.mp3

[38] Link to 30 second snippet of Beautiful Moons Ago: http://econnect.ait.co.at/samples/orchard/884/385/611/884385611170/884385611170_1_4_CLIP.mp3

Track three: Just The Way You Are[39]. Average general grade: 1.77. Comments: ‘background music’, ‘this tune sounds happier’. Lowest score on harmony and timbre (1.85 on both) ‘more jazzy’, ‘more resonance’. Highest score on quality and instrumentation (4.31; 2.38), ‘real music instruments being used in this track’.

Reference track three: Alexander’s Ragtime (genre: ragtime)[40]

Track one: Boneyard Shuffle[41]. Average general grade: 4.31. Comments: ‘very similar, both authentic recordings’, ‘a bit older than reference track, but the same music’, ‘harmony is more complicated, giving more variety in the timbre’. Lowest score on instrumentation (3.77), ‘more copper instruments than reference track’, ‘trumpet gets more space in this track’. Highest score on tempo and rhythm (4.46 on both).

Track two: My Daddy Rocks Me[42]. Average general grade: 2.38. Comments: ‘more atmosphere, more space’, ‘more swing, less blues’, ‘tracks are the same age’. Lowest score on tempo (1.77), ‘tempo is slower’. Highest score on quality and instrumentation (4.38; 2.62), ‘better recorded, for this track they used multiple microphones’.

Track three: Ägyptischer Marsch[43]. Average general grade: 1.15. Comments: ‘very different, gives me a whole different feel’, ‘this track is Vienna, the reference track is New Orleans, there is whole world between the two’, ‘this track has nothing to do with the reference track, it’s like saying that a car looks like a bicycle when they are traveling at the same speed’. Lowest score on harmony and timbre (1.31 on both). Highest score on quality, instrumentation, tempo and rhythm (4 on quality; 1.54 on the other elements).

Reference track four: V Lednu Je Maj (genre: classical jazz)[44]

Track one: Someone Stole Gabriel's Horn[45]. Average general grade: 1.85. Comments: ‘same period’, ‘more swing’, ‘same time in history, but this is not background music like the reference track’, ‘christmas track’. Lowest score on rhythm (2.31), ‘slower’. Highest score on quality, harmony and timbre (4.54; 2.46; 2.46), ‘less heavy’.

[39] Link to 30 second snippet of Just The Way You Are: http://econnect.ait.co.at/samples/orchard/829/410/212/829410212267/829410212267_1_9_CLIP.mp3

[40] Link to 30 second snippet of Alexander’s Ragtime: http://econnect.ait.co.at/samples/orchard/843/436/002/843436002169/843436002169_1_13_CLIP.mp3

[41] Link to 30 second snippet of Boneyard Shuffle: http://econnect.ait.co.at/samples/orchard/803/680/768/803680768259/803680768259_1_5_CLIP.mp3

[42] Link to 30 second snippet of My Daddy Rocks Me: http://econnect.ait.co.at/samples/orchard/829/410/234/829410234764/829410234764_1_8_CLIP.mp3

[43] Link to 30 second snippet of Ägyptischer Marsch: http://www.preiserrecords.at/streamMP3Clip.php?mp3link=/mnt/wdc01/imported/717281902915/717281902915_1_03_clip.mp3&fileName=%C3%84gyptischer%20Marsch.mp3

[44] Link to V Lednu Je Maj on Europeana: http://www.europeana.eu/portal/record/09326/233628E7740903E6F0378BA1A7D9FF62204AE6DA.html?start=1&query=V+Lednu+Je+Maj+&startPage=1&qt=false&rows=24

[45] Link to the track Someone Stole Gabriel's Horn: http://bdh-rd.bne.es/high.raw?id=0000013087&name=00000001.original.mp3

Track two: Cielo Azul Fox-Trot[46]. Average general grade: 3.08. Comments; 'same atmosphere, but big band makes it more bombastic', 'this track is less sharp', 'different genre, this is music to dance to'. Lowest score on tempo and rhythm (2.54 on both), 'faster tempo'. Highest score on instrumentation (3.46), 'mainly wind instruments'.

Track three: When It's Sleepy Time In The Old South[47]. Average general grade: 2.54. Comments: 'different genre', 'Dixieland music, this affects me more', 'similar, except the added vocal'. Lowest score on tempo and rhythm (2.46 on both), 'faster tempo'. Highest score on quality and instrumentation (4.15; 3.08), 'recording quality in line with reference track'.

Reference track five: Requiem IV Sanctus (genre: requiem)[48]

Track one: V. Agnus Dei (Requiem|Op. 48)[49]. Average general grade: 2.31. Comments: 'more dramatic, I see images of dead rabbits instead of the happy rabbits I saw when you played the reference track', 'less modern', 'church music, could be on the same compilation album'. Lowest score on harmony and timbre (2.31 on both), 'bombastic'. Highest score on quality and instrumentation (4.38; 3.77) 'better recording'.

Track two: Dies Irae[50]. Average general grade: 2.38. Comments: 'minor, not major', 'more dramatic', 'same atmosphere'. Lowest score on harmony (2.54), 'more harmony'. Highest score on quality and instrumentation (4.23; 3.62), 'sounds like it's mono instead of stereo', 'recorded more directly'.

Track three: II. Gloria (Mass in A Flat Major D678)[51]. Average general grade: 2.69. Comments: 'not the same atmosphere, sounds a bit happier than the reference track'. Lowest score on timbre (2.77), 'warmer sound'. Highest score on quality and instrumentation (4.46; 3.54), 'better recorded'.

Reference track six: Rock 'n Roll Soldier (genre: rock)[52]

Track one: Mladost[53]. Average general grade: 2.54. Comments: 'less authentic', 'friendlier', 'more pop than rock, would not fit the same compilation album', 'completely different'. Lowest score on harmony

[46] Link to the track Cielo Azul Fox-Trot: http://bdh-rd.bne.es/high.raw?id=0000062523&name=00000001.original.mp3

[47] Link to the track When It's Sleepy Time In The Old South: http://bdh-rd.bne.es/high.raw?id=0000134485&name=00000001.original.mp3

[48] Link to Requiem IV Sanctus on Europeana: http://www.europeana.eu/portal/record/2023601/oai_eu_dismarc_ACAD_USA370698745.html?start=2&query=Requiem+IV&startPage=1&qf=TYPE%3ASOUND&qt=false&rows=24

[49] Link to 30 second snippet of V. Agnus Dei (Requiem|Op. 48): http://econnect.ait.co.at/samples/orchard/884/385/520/884385520694/884385520694_1_5_CLIP.mp3

[50] Link to 30 second snippet of Dies Irae: http://econnect.ait.co.at/samples/orchard/884/385/014/884385014322/884385014322_1_3_CLIP.mp3

[51] Link to 30 second snippet of II. Gloria (Mass in A Flat Major D678): http://econnect.ait.co.at/samples/orchard/884/385/055/884385055967/884385055967_1_2_CLIP.mp3

[52] Link to Rock 'n Roll Soldier on Europeana: http://www.europeana.eu/portal/record/2023601/oai_eu_dismarc_NEXT_USA560931966.html?start=1&query=Rock+%E2%80%98n+Roll+Soldier&startPage=1&qt=false&rows=24

[53] Link to 30 second snippet of Mladost: http://econnect.ait.co.at/samples/orchard/843/436/055/843436055394/843436055394_1_7_CLIP.mp3

and timbre (2.62 on both), 'lighter tone colour'. Highest score on quality, tempo and rhythm (4.31; 3.77; 3.77), 'sounds like it comes from far away', 'tempo is slower'.

Track two: Believer[54]. Average general grade: 3.69. Comments: 'sounds more serious', 'same genre'. Lowest score on harmony and timbre (3.62 on both), 'more harmony'. Highest score on quality and instrumentation (4.77; 4.62), 'less quality, not recorded as compact as the reference track'.

Track three: Rock Like Hell[55]. Average general grade: 4.08. Comments: 'repeat of previous track (Believer)?', 'could have been made by the same band, although this track is a little bit more laid-back'. Lowest score on rhythm (3.92). Highest score on quality and instrumentation (5. 4, 62), 'more room for guitar in the recording', 'a richer recording'.

Reference track seven: Tangos de Malaga 300 por minuto (genre: tango)[56]

Track one: Tangos de Malaga 280 por minuto[57]. Average general grade: 4.85. Comments: 'almost identical', 'same atmosphere', 'could have been the same song in my opinion'. Lowest score on rhythm (4.69) 'rhythm is a bit different'. Highest score on instrumentation and quality (4.92 on both), 'recorded with more warmth'.

Track two: Tangos de Malaga Subida Rapida[58]. Average general grade: 4.23. Comments: 'less thrilling', 'same music', 'guitar more prominent'. Lowest score on rhythm (3.92), 'slower'. Highest score on instrumentation (4.54), 'I think I hear just one person clapping hands, in the reference track I heard two persons clapping'.

Track three: Solea Corta 300 por minuto. Average general grade: 4.54. Comments: 'I again get an image of a lady in a red flamenco dress (same as with reference track)', 'I hear more clapping'. Lowest score on rhythm (3.62), 'other rhythm, wavering', 'less constant rhythm'. Highest score on quality and instrumentation (5; 4.69), 'recorded more flat, guitar sounds fiercer'.

Reference track eight: De Pajisas Encarnada[59] (genre: flamenco)

Track one: Que Yo A Ella No La Quería "Fandangos"[60]. Average general grade: 4.69. Comments: 'sounds like the same musician, but another song', 'same kind of music, same period', 'similar in sadness'. Lowest score on harmony (4.54), 'more vividness in the harmony'. Highest score on quality, tempo and

[54] Link to 30 second snippet of Believer:
http://econnect.ait.co.at/samples/orchard/884/385/534/884385534578/884385534578_1_5_CLIP.mp3

[55] Link to 30 second snippet of Rock Like Hell:
http://econnect.ait.co.at/samples/orchard/884/385/534/884385534578/884385534578_1_11_CLIP.mp3

[56] Link to 30 second snippet of Tangos de Malaga 280 por minuto:
http://econnect.ait.co.at/samples/orchard/884/385/675/884385675806/884385675806_1_31_CLIP.mp3

[57] Link to 30 second snippet of Tangos de Malaga 280 por minuto:
http://econnect.ait.co.at/samples/orchard/884/385/675/884385675806/884385675806_1_31_CLIP.mp3

[58] Link to 30 second snippet of Solea Corta 300 por minuto:
http://econnect.ait.co.at/samples/orchard/884/385/675/884385675806/884385675806_1_16_CLIP.mp3

[59] Link to De Pajisas Encarnada on Europeana:
http://www.europeana.eu/portal/record/2023601/oai_eu_dismarc_BLA_USA371065489.html?start=1&query=De+Pajisas+Encarnada&startPage=1&qt=false&rows=24

[60] Link to 30 second snippet of Que Yo A Ella No La Quería "Fandangos":
http://econnect.ait.co.at/samples/orchard/884/385/974/884385974909/884385974909_1_4_CLIP.mp3

rhythm (4.92; 4.75; 4.75)[61], 'a lot of variety in the rhythm of these songs (both this one and the reference track)', 'the recording quality is annoying'.

Track two: Tienen sus Perfumes Finos "Fandangos"[62]. Average general grade: 4.54. Comments: 'singer gets a more prominent role', 'elegy'. Lowest score on harmony (4.31). Highest score on instrumentation (4.77) 'guitar is more soft'.

Track three: Fandangos[63]. Average general grade: 4.08. Comments: 'happier, more guitar, I like this track the most', 'this track is the most deviant'. Lowest score on timbre (3.69), 'the guitars get more room in this track'. Highest score on instrumentation (4.15).

Reference track nine: Entracte, Extrait De (genre: classical)[64]

Track one: Sonata in B-flat Major[65]. Average general grade: 3.15. Comments: 'more powerful', 'there is more dialogue between the instruments in this track (more equally balanced)', 'this track displays the impermanence more, this track gives me a summer feeling instead of a spring feeling (where everything is growing and blooming)'. Lowest score on instrumentation (3), 'more violin', 'flute is replaced by strings and therefore more beautiful'. Highest score on quality, tempo and rhythm (4.69; 3.85; 3.85), 'this track is recorded more directly'.

Track two: Concerto in D Major[66]. General: 1.92. Comments: 'same atmosphere', 'very different, this is a solo plus an orchestra', 'this is still classical, but that is the only comparison with the reference track', 'different, sounds more royal, this is a happy message, very different feeling'. Lowest score on instrumentation (1.77) 'an added trumpet', 'this is a whole orchestra, totally different'. Highest score on quality and tempo (4.69; 2.92) 'again a good recording, but recording setting is different, because this is a baroque song (and not chamber music anymore)'.

Track three: Concerto Pour Hautbois[67]. Average general grade: 2.31. Comments: 'different genre', 'an everything-is-alright track', 'more dynamic', 'baroque instead of chamber music'. Lowest score on instrumentation (1.92), 'very different, more wood instruments', 'other lead instrument'. Highest score on quality, tempo and rhythm (4.69; 3; 3) 'a flat recording in terms of quality'.

[61] One of the participants didn't score the tempo and rhythm for this track in comparison with the reference track, so the avarage of 4,75 is based on 12 participants.
[62] Link to 30 second snippet of Tienen sus Perfumes Finos "Fandangos": http://econnect.ait.co.at/samples/orchard/884/385/974/884385974909/884385974909_1_1_CLIP.mp3
[63] Link to 30 second snippet of Fandangos: http://econnect.ait.co.at/samples/orchard/884/385/423/884385423117/884385423117_1_1_CLIP.mp3
[64] Link to Entracte, Extrait De on Europeana: http://www.europeana.eu/portal/record/2023601/oai_eu_dismarc_7PRO_FRY510779933.html?start=1&query=Entracte%2C+Extrait+De*&startPage=1&qf=TYPE%3ASOUND&qt=false&rows=24
[65] Link to 30 second snippet of Sonata in B-flat Major: http://econnect.ait.co.at/samples/orchard/669/910/667/669910667361/669910667361_1_9_CLIP.mp3
[66] Link to 30 second snippet of Concerto in D Major: http://econnect.ait.co.at/samples/orchard/829/410/496/829410496469/829410496469_1_4_CLIP.mp3
[67] Link to 30 second snippet of Concerto Pour Hautbois: http://econnect.ait.co.at/samples/orchard/884/385/616/884385616540/884385616540_1_2_CLIP.mp3

4.3.2 Experienced similarity and calculated similarity

When comparing the results of the user evaluation with the calculated similarity estimation based on vector distance the conclusion is that there is a very rough correspondence between the two. In total there were 27 tracks measured in the user evaluation (nine reference tracks in total with three tracks per reference track). When a track scores high on the calculated similarity estimation to a reference track it will not necessarily mean that it will be experienced by participants as a similar track based on the user evaluation. For example the track that scored relatively (relative to the scores of the other 26 tracks) lowest on the calculated similarity (88.40%), is ranked in the top 10 highest graded tracks in the user evaluation (4.08 average). However, when ranking the 27 tracks from lowest (88.40%) to highest (92.68%) based on the calculated similarity a rough pattern emerges when compared to the results of the user evaluation. In the list underneath the ranking of the 27 tracks based on the calculated similarity with the corresponding average grades given by the 13 participants in the user evaluation is shown.

Table 2: Tracks ranked based on the calculated similarity based on vector distance from high (#1) to low (#27)

	Track Name	Calculated similarity[68]	Experienced similarity[69]
1	Tangos de Malaga 280 por minuto (tango)	93.68%	4.85
2	Que Yo A Ella No La Quería "Fandangos" (flamenco)	93.23%	4.69
3	Tangos de Malaga Subida Rapida (tango)	93.14%	4.23
4	Solea Corta 300 por minuto (tango)	92.61%	4.54
5	Someone Stole Gabriel's Horn (classical jazz)	92.45%	1.85
6	Cielo Azul Fox-Trot (classical jazz)	92.11%	3.08
7	When It’s Sleepy Time In The Old South (classical jazz)	91.87%	2.54
8	Mladost (rock)	91.86%	2.54
9	Believer (rock)	91.85%	3.69
10	Palm Leaf Rag (ragtime 1)	91.76%	4.54
11	Rock Like Hell (rock)	91.75%	4.08
12	The Nonpareil (ragtime 1)	91.65%	4.54
13	Eugenia (ragtime 1)	91.47%	3.92

[68] See Figure 7 in chapter 3.3 for an overview of the music similarity estimation process. Vector distance of two song-feature-vectors is calculated using a late-fusion approach. Weights are applied to the distance values of the distinct features. The sum of all distances is calculated to aggregate a final similarity estimation.

[69] The average grade given by the 13 participants in the user evaluation.

14	Sonata in B-flat Major (classical)	90.54%	3.15
15	V. Agnus Dei (Requiem|Op. 48) (requiem)	90.44%	2.31
16	Dies Irae (requiem)	90.41%	2.38
17	II. Gloria (Mass in A Flat Major D678) (requiem)	90.26%	2.69
18	Patterns In The Rain (smooth jazz)	89.80%	2.92
19	Beautiful Moons Ago (smooth jazz)	89.65%	1.85
20	Boneyard Shuffle (ragtime 2)	89.52%	4.31
21	Just The Way You Are (smooth jazz)	89.46%	1.77
22	Tienen sus Perfumes Finos "Fandangos" (flamenco)	89.34%	4.54
23	My Daddy Rocks Me (ragtime 2)	89.27%	2.38
24	Concerto in D Major (classical)	89.21%	1.92
25	Concerto Pour Hautbois (classical)	89.20%	2.31
26	Ägyptischer Marsch (ragtime 2)	89.03%	1.15
27	Fandangos (flamenco)	88.40%	4.08

Although the track with the lowest calculated similarity (88.40%) got a relatively high average score in the user evaluation (4.08), the lowest five tracks together got an average (2.37) that is lower than the highest five tracks (4.03). Apparently this rule follows through in the whole list of results in a roughly linear manner. The top 13 tracks (based on calculated similarity) scored an average of 3.78 in the user evaluation and the bottom 13 tracks got an average of 2.66. So there is a difference of more than one point in the grading scale average between the top half of the scores and the bottom half[70]. Underneath a list is shown with the average general grade given by participants for the highest and lowest tracks taken from the calculated similarity. For the highest three tracks an average of 4.59 is written down, this is the average of grades from the user evaluation of the top three tracks in the list above[71].

Table 3: Average general grade given by participants for the tracks that scored highest and lowest on the calculated similarity estimation based on vector distance

Highest/lowest # tracks:	Average grade:
Highest 1	4.85
Highest 2	4.77

[70] 3,78 - 2,66 = 1,12. A 1,12 point difference between the top 13 and bottom 13 tracks (based on calculated similarity). And a (3,73 - 2,7) 1,03 point difference between the top 14 and bottom 14 tracks. There are 27 tracks in total, so the track exactly in the middle (Sonata in B-flat Major, #14 in image 11) is measured in both averages for the top and bottom 14 tracks.

[71] 4,85 (#1) + 4,69 (#2) + 4,23 (#3) = 13,77. 13,77 divided by three is an average grade of 4,59 taken from the user evaluation based on the highest three tracks from the calculated similarity estimation based on vector distance.

Highest 3	4.59
Highest 4	4.58
Highest 5	4.03
Highest 6	3.87
Highest 7	3.68
Highest 8	3.54
Highest 9	3.56
Highest 10	3.66
Highest 11	3.69
Highest 12	3.76
Highest 13	3.78
Highest 14	3.73
Lowest 14	2.70
Lowest 13	2.66
Lowest 12	2.69
Lowest 11	2.72
Lowest 10	2.72
Lowest 9	2.70
Lowest 8	2.81
Lowest 7	2.59
Lowest 6	2.73
Lowest 5	2.37
Lowest 4	2.37
Lowest 3	2.51
Lowest 2	2.61

To make the rough correspondence from the list more visual, the data from Table 3 has been put into the graph below. The sample of participants that this rough correspondence is based on is very low (13 participants), so the rough correspondence still could very well be a coincidence. In order to prove that there is (or is not) a correspondence between the calculated similarity between tracks based on the vector distance and the actual experienced similarity between tracks by end-users a much bigger data sample must be used. An online survey could make this relatively easy to execute. The hypothesis for the online survey (based on this user evaluation) would be that if there is a difference of at least 5.28 points[72] in the percentages distracted from the calculated similarity (if the other conditions will be the same[73]) the difference between the average grade in the top and bottom half of the list (ranked on calculated similarity) would be at least one point (which is a difference of at least 25%[74] in the used scale).

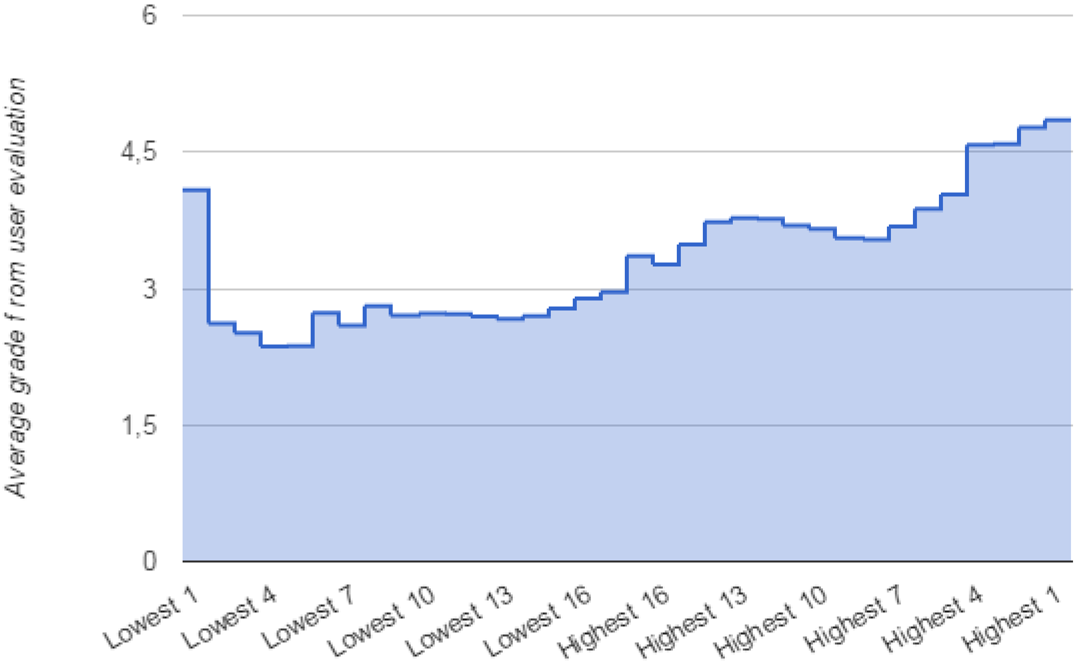


Figure 12: Average general grade given by participants for the tracks that scored highest and lowest on the calculated similarity estimation based on vector distance

4.3.3 User evaluation of the MIR concept

After listening and giving grades and comments about the experienced differences and similarities between the snippets of songs, the functionality of the music information retrieval was explained. The participants were asked some questions in order to gain more feedback on the concept of music information retrieval. Nine participants were positive about developing music information retrieval, ranging from ‘great’ and ‘perfect’ to ‘useful, could be interesting for future generations to keep them interested in old music’. Three participants were negative about the concept ‘I’m against this, you cannot let machines perform human tasks’ and ‘I don’t like it when I get music suggestions’. One participant was more neutral and gave the nuanced reply ‘not everybody wants to use this I think, but some groups could make good use of this of course’.

[72] The top track (#1 in image 11) has a calculated similarity of 93,68%, the bottom track (#27 in image 11) 88,40%, so there is a difference of 5,28 points (93,68 - 88,40 = 5,28).

[73] Measuring only the first three tracks to a reference track, using at least eight different music genres, having 30 second snippets of tracks, using the same scale from one to five (participants were not allowed to give half-grades)

[74] Participants could give five different grades, translated to percentages this comes down to 1 = 0% on the scale, 2 = 25%, 3 = 50%, 4 = 75% and 5 = 100%. One point or more difference in the scale means that there is at least 25% difference in the scale.

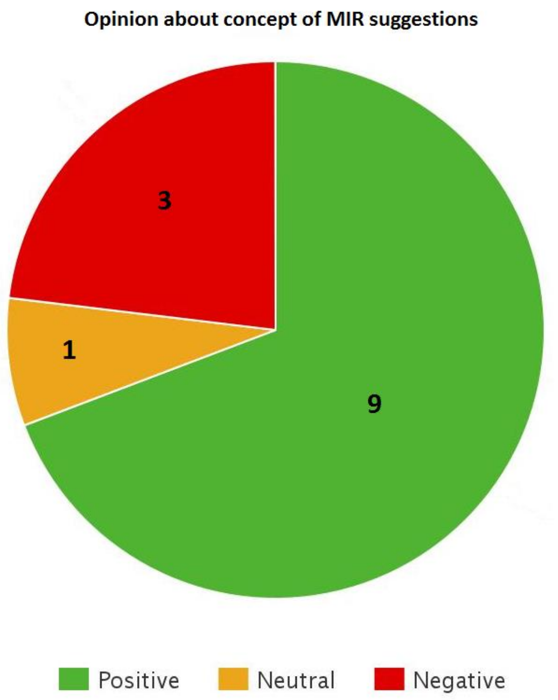


Figure 13: Opinion about the concept of music information retrieval based suggestions from the user evaluation

When asked if they would use music information retrieval based suggestions six participants said yes, one person didn’t know this yet and another six participants think they would not use this themselves. Some interesting answers were given from the people who said ‘no’ to this question. One participant would not use this because he said he wants to explore music himself. This is an indication that it might be good to give the end-user a music information retrieval experience that communicates that it is bringing the end-user a unique experience in which they get the feeling the discovery of new music is also their accomplishment. Another participant that said he would not use the service noted that it would be a very good technique for archives with poor information about their songs and collection. One participant said he would use it, but that for him lyrics and the charisma of a musician is very important and that he would hate it to get a suggestion of a track made by a bad imitator of the musician he loves. So it could be a way to go to give certain end-users music information retrieval based suggestion that weights certain variables (lyrics, favourite artists) more than other types of end-users.

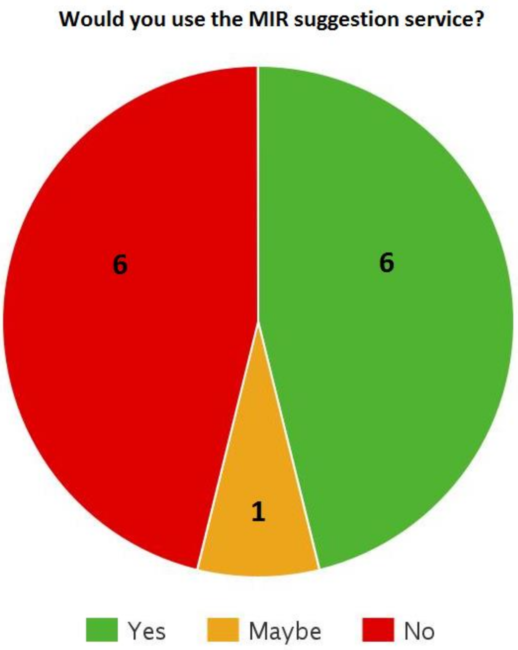


Figure 14: Answers to the question if participants would use the music information retrieval based suggestion service

The third question was focussed on what a participant would experience if they got a music information retrieval based suggestion, but that the suggested track is very different from the music they are listening to. Eight participants would not like this, ranging from ‘I would be pissed off’ to ‘pointless, I would have the feeling that I didn’t ask for that’. This is an indication that the suggestions should not be too different from the reference track. Three participants were neutral about this, one of the three pointed out that he would really like to have the option of giving feedback in such an instance. This is an indication to indeed give end-users at least the possibility of flagging wrong suggestions. Two participants were very positive about receiving a very different suggestion ‘nice to get something else, I like variety’ and ‘amusing, because it brings forth another mind-set’.

4.3.4 Conclusion of the user evaluation

In conclusion, the following results and recommendations can be given based on the user evaluation of the MIR pilot. Based on this user evaluation there is no indication that a high similarity in the calculated similarity means that there will be a high similarity in the experience of two tracks by an end-user[75]. Although there is no indication to make a real estimation about the possible experience of similarity by one end-user between two tracks, there is a correspondence from the user evaluation when analysing all the grades taken from the user evaluation. A rough linear correspondence is measured between the calculated similarity estimation based of vector distance and the grades given by participants in the user evaluation. There is more than one point difference in the average rating between the top and the bottom half of the tracks when the tracks are ranked from high to low based on the calculated similarity

[75] Note that 30 second snippets have been used in this user evaluation and that the experience of similarity might be different if end-users can listen to a whole track.

estimation. This could mean that there is a prediction to be made about the experience of similarity by a group of end-users based on calculated similarity for multiple tracks. To actually prove this a bigger data sample must be used. Since this statement is based on only 13 participants there is still a possibility that this result is a coincidence. An online survey is the most practical way to go in order to check this[76]. The hypothesis for the online survey is that if there is a difference of at least 5.28 points in the percentages distracted from the calculated similarity (if the other conditions will be the same), the difference between the average grade in the top and bottom half of the list (ranked on calculated similarity) would be at least one point. Taken from the answers to the questions at the end of the evaluation some indications for future work were given as well. Overall, it wouldn't be a good idea to give suggestions that are too different from the reference track based on this user evaluation (only two out of thirteen participants were positive about such an occurrence). A point that needs further discussion and research is the way the music information retrieval suggestion service will communicate to end-users; does the target audience want to have a feeling of exploring music themself? If yes, must this be facilitated and how? Another area to give more attention too is the possibility of giving certain types of end-users tailor-made music information retrieval suggestions (for instance by weighing certain variables differently). Another indication derived from the answers to the questions at the end of the user evaluation is to give end-users the possibility of flagging incorrect suggestions and/or giving the possibility of providing descriptive feedback on suggestions.

5 Results

This section provides and discusses the results of the evaluation runs.

5.1 Evaluation results

The table below provides a summary of the evaluation runs. Each run has a dedicated ID to refer to in the succeeding discussion of the results. The quoted query terms refer to actual words that were used to search in the metadata. The search approach was case insensitive and sequentially. This means that for the query “Jazz” + “Traditional” the metadata was first filtered by the term “Jazz” and the resulting subset was filtered by the term “Traditional”. This example resulted in a final ground-truth subset of size 512 which is denoted by the column Num. Tracks. The rightmost six columns represent the calculated precision values for the corresponding query at different result-list lengths. As explained in Section 3.1, refers a result-list length of 1 to the precision of the first and most similar entry of the list whilst a length of 24 corresponds to the precision of a complete search result page of the Europeana Web-page.

[76] A participant in the user evaluation said that he wouldn't follow through with the evaluation after three reference tracks if it would be online. So a good way to go for the online survey could be to lose some of the elements (harmony, tempo and quality are the best candidates to lose).

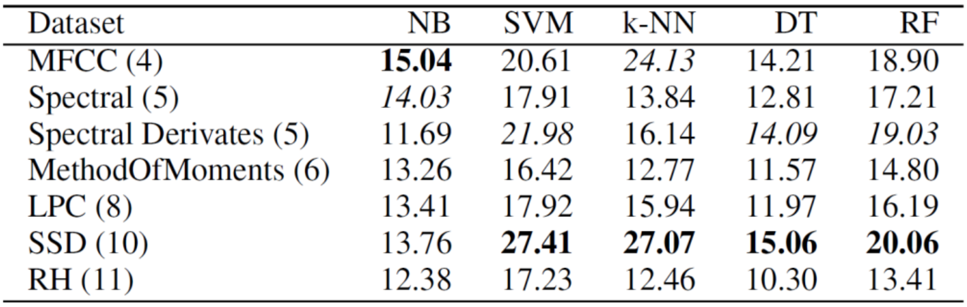
Table 4: Evaluation results for different queries. "Num. Tracks" refers to the size of the query-generated ground truth. The number-labelled columns refer to the calculated precision values at different result-list lengths.

ID	Query	Num. Tracks	1	2	3	5	10	24
Jazz								
1	"Jazz"	31801	38.0	35.0	31.4	31.7	28.6	26.2
2	"Smooth Jazz"	2419	49.1	45.9	43.8	25.8	20.8	16.0
3	"Jazz" + "Traditional"	577	25.3	21.1	17.1	13.0	9.2	6.3
4	"Ragtime"	57	24.6	15.8	12.3	7.3	3.6	1.6
5	"Shuffle"	112	22.3	14.3	11.0	7.0	4.0	1.7
6	"Jazz" + "Blues"	1398	14.6	10.1	7.5	5.6	4.2	2.9
7	"Jazz" + "Bob Crosby, Andy Kirk, June Richmond"	24	12.5	6.3	4.2	2.5	3.8	1.8
8	"Jazz" + "Cuban Jazz"	105	11.2	7.6	5.7	4.2	2.0	1.0
Classical								
9	"Classical"	28569	44.3	42.1	40.5	38.3	35.1	32.3
10	"Piano Concerto"	510	38.6	32.0	28.0	23.9	17.6	10.2
11	"Requiem"	463	32.6	26.9	22.0	16.2	10.7	6.6
12	"operette", "operetta", "opereta", "zarzuela"	1081	27.7	22.9	20.8	17.3	14.6	11.5
13	"Opera"	8278	26.8	24.7	22.7	21.1	18.9	16.1
14	"Classical" + "g major" or "g dur" or "g-dur" or "g majeur"	304	17.1	14.8	14.0	12.6	9.3	5.6
15	"Classical" + "g major" or "g dur" or "g-dur" or "g majeur" + "quartett"	13	15.4	7.7	5.1	3.1	2.3	1.6
16	"Classical" + "major" or "dur" or "majeur" + "quartett" + "allegro"	191	9.4	6.3	7.3	8.1	5.6	3.7
Non-Music								
17	"Interview"	484	77.5	74.3	72.0	68.6	60.8	51.4
18	"Biodiversity Center" (=Animal Sounds)	1097	89.7	87.0	85.1	82.8	78.7	73.9
19	"Biodiversity Center" + "Chorthippus"	113	59.3	55.3	56.6	53.0	48.1	43.0

	(=Crickets)							
European Popular/Traditional Music								
20	“Flamenco”	1827	40.7	33.0	29.2	24.3	18.2	12.8
21	“Tarantella”	152	33.6	28.0	22.4	16.1	8.5	4.0
22	“Tango”	3716	30.2	24.9	22.3	19.5	16.0	12.5
23	“Flamenco” + “Guitarra”	287	22.3	17.1	15.3	13.5	10.0	8.3
24	“Jodler”	61	16.4	8.2	5.5	3.6	2.0	1.4
25	“Serenata”	436	9.7	7.5	5.5	4.4	3.0	2.0
26	“Volksmusik”	13	7.7	3.9	2.6	4.6	4.6	2.9
27	“Fados”	107	6.5	10.8	9.7	7.3	4.9	2.7
Contemporary								
28	“Rock ‘n Roll”	24	16.7	8.3	5.6	3.3	2.0	1.6
29	“Hip Hop”	72	11.1	11.1	12.5	11.9	7.5	5.2

5.2 Discussion

The results of the evaluation runs can generally be observed to correspond with state-of-the-art music similarity retrieval results presented in the literature. The most representative related publication is Schindler (2012). This study presented new genre ground truth assignments for the Million Songs Dataset, including the first baseline results for automatic genre classification experiments. One of the classifiers used in the evaluation was the k-nearest neighbour’s classifier. The principle of this classifier to label an unknown instance is to calculate the vector distances to all items in the dataset and to sort them by their distance. The k top most entries with the smallest distance to the unknown instance - the k nearest neighbours, are used to determine the label for the processed instance by majority voting. An example could be to set k to 3, which means, that the top 3 nearest neighbours will be used for majority voting. If one of them is of “genre 1” and two of them are from “genre 2”, the unknown instance is classified as “genre 2”. A special case is knn with k = 1. In this case only the label of the top nearest neighbour is used to classify the unknown instance. This approach was used in Schindler, 2012 and it is equivalent to the similarity retrieval approach of the MIR pilot with a result list length of 1. In this case, the results of similarity retrieval are comparable to those of classification experiments. The results presented in Schindler, 2012 are the only in literature matching the scale of the Europeana dataset. The subset used in their evaluation contained 273,936 music tracks. This is only 12.2% smaller than the Europeana dataset with 312,096 audio files. The results reported by Schindler, 2012 are presented in Table 2. Best results for the k-NN classifier with k = 1 were reported for the Statistical Spectrum Descriptors (SSD). Due to scalability problems the authors were not able to include higher dimensional feature sets such as Rhythm Patterns (RP) in their evaluation as well as combined feature spaces.

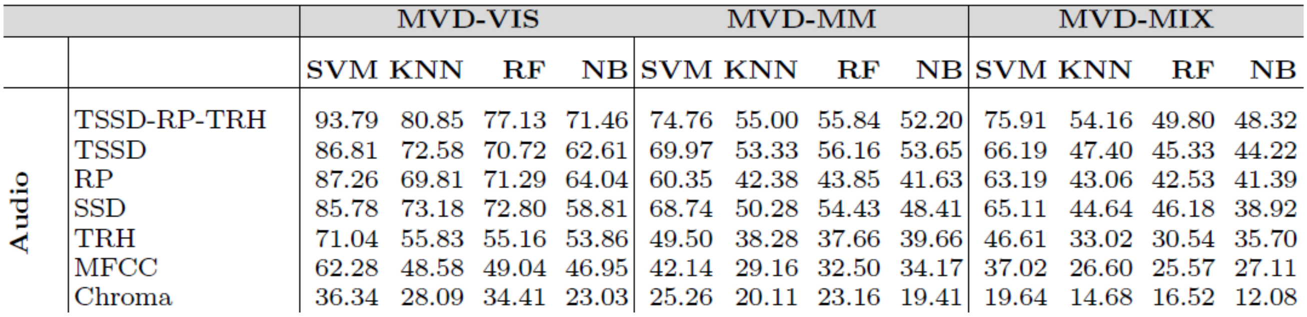


Dataset	NB	SVM	k-NN	DT	RF
MFCC (4)	**15.04**	20.61	*24.13*	14.21	18.90
Spectral (5)	*14.03*	17.91	13.84	12.81	17.21
Spectral Derivates (5)	11.69	*21.98*	16.14	*14.09*	*19.03*
MethodOfMoments (6)	13.26	16.42	12.77	11.57	14.80
LPC (8)	13.41	17.92	15.94	11.97	16.19
SSD (10)	13.76	**27.41**	**27.07**	**15.06**	**20.06**
RH (11)	12.38	17.23	12.46	10.30	13.41

Figure 15: Results of classification experiments with the Million Song Dataset taken from [Schindler 2012]. For the k-NN classifier k = 1 was used. Thus, these results are comparable to the similarity retrieval results presented in Table 1 with a result-list length of 1.

A naive approach to compare the results of Schindler, 2012 with those of the MIR pilot evaluation presented in Figure 15 would be to calculate the mean precision of all evaluation runs. This would result in an average precision of 28.7% and be slightly above the top results presented in literature. Thus, the performance of the implementation corresponds to comparable state-of-the-art similarity retrieval system. However, systems have been reported with much higher precision values[77]. Unfortunately, the datasets used to evaluate these systems have not been made public, or are too small in size to make them reliably comparable. In Schindler, 2015 a similar approach to evaluate the performance of a music video classification system was used. For the evaluation the Music Video Dataset (MVD)[78] has been created. It is a highly specialized dataset to develop visual feature extractors for music video analysis, and consists of two sub-sets of music videos of different genres. All subsets had been assembled by their audio properties - they had been selected because they sounded similar or stereotypic for that specific genre. The tracks of the MVD-VIS's genres sound especially similar. Figure 16 shows the precision results for the genre classification experiments. Combinations of Rhythm Patterns features reach a precision of 93.79% using Support Vector Machine (SVM)[79] classifiers and 80.85% for K-NN classifier with k=1. Consequently this is currently the top result to be expected under optimal conditions such as clearly defined genres without overlaps. The MVD-MM subset was assembled to incorporate such overlaps. Consequently, the precision values are notably smaller than those of the MVD-VIS dataset. The MVD-MIX set is a non-overlapping combination of the MVD-VIS and MVD-MM subset. The combination creates a bigger dataset with a higher number of genres which is an important evaluation scenario for automatic genre classification experiments. In both sets, the MVD-MM and the bigger MVD-MIX set, the precision values for the K-NN classifier with k=1 range between 26% for the standard feature set MFCC and about 55% for the combination of the Rhythm Patterns features. Although the MVD essentially differs in size, having only 1,600 entries, its specialization provides good boundaries of which maximal and average values can be expected from the evaluated features.

[77] http://www.music-ir.org/mirex/wiki/2015:Audio_Music_Similarity_and_Retrieval
[78] http://www.ifs.tuwien.ac.at/mir/mvd/
[79] https://en.wikipedia.org/wiki/Support_vector_machine



		MVD-VIS				MVD-MM				MVD-MIX			
		SVM	KNN	RF	NB	SVM	KNN	RF	NB	SVM	KNN	RF	NB
Audio	TSSD-RP-TRH	93.79	80.85	77.13	71.46	74.76	55.00	55.84	52.20	75.91	54.16	49.80	48.32
	TSSD	86.81	72.58	70.72	62.61	69.97	53.33	56.16	53.65	66.19	47.40	45.33	44.22
	RP	87.26	69.81	71.29	64.04	60.35	42.38	43.85	41.63	63.19	43.06	42.53	41.39
	SSD	85.78	73.18	72.80	58.81	68.74	50.28	54.43	48.41	65.11	44.64	46.18	38.92
	TRH	71.04	55.83	55.16	53.86	49.50	38.28	37.66	39.66	46.61	33.02	30.54	35.70
	MFCC	62.28	48.58	49.04	46.95	42.14	29.16	32.50	34.17	37.02	26.60	25.57	27.11
	Chroma	36.34	28.09	34.41	23.03	25.26	20.11	23.16	19.41	19.64	14.68	16.52	12.08

Figure 16: Results of classification experiments with the Million Song Dataset taken from [Schindler 2015]. For the k-NN classifier k = 1 was used.

The MVD has been used in the development stage to evaluate which features to use for MIR pilot, as well as how to combine them to get an optimal performance. Image 13 shows one of those preliminary evaluations. It depicts the precision values of the Rhythm Patterns feature set with different feature space normalization methods applied and with different distance measures calculated. The chart reproduces the precision values of 69.81 for RPs on the MVD-VIS subset presented in Table 4. It further depicts the performance at different result-list lengths. It is observable that the precision drops by about 20% on average from k=1 to k=20. This behaviour can also be observed in the results of this evaluation as presented in Table 1. This is an artefact that is ascribable to flaws in the ground truth data which was not created for similarity retrieval experiments.

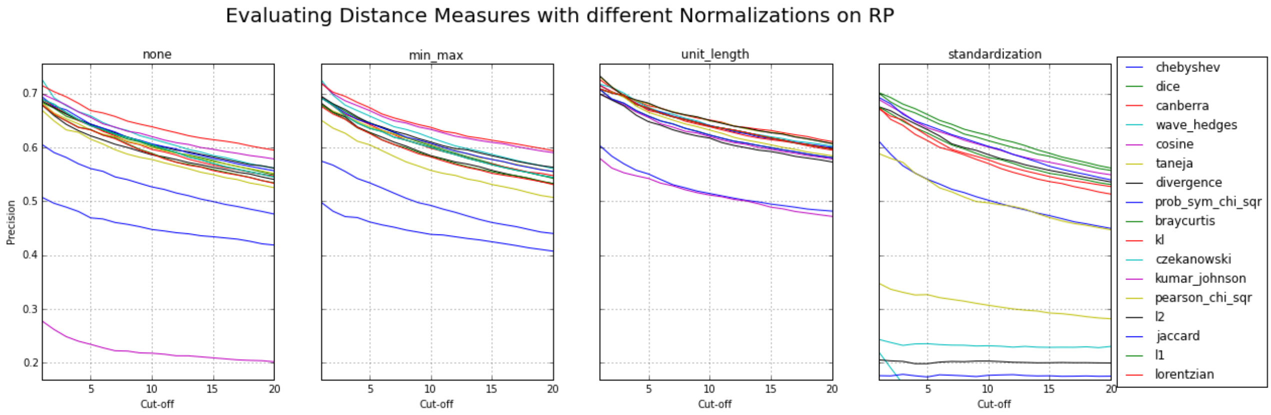


Figure 17: Performance evaluation of the Rhythm Patterns (RP) feature set with different normalization methods in combination with different distance measures on the MVD-VIS subset of the Music Video Dataset. Numbers depict precision values for result-list length ranging from 1 to 20 entries.

Figure 18 shows the same evaluation on the MVD-MM subset of the Music Video Dataset. This subset has more overlaps between the genres and thus results in weaker performance values. This set corresponds more to large mixed datasets such as the Europeana Sounds dataset. The relative performance values and the decline in precision in the case of longer result-lists are comparable to those of the MVD-VIS dataset with its highly similar sounding tracks per genre. These two similar observations on the different subsets indicate that such performance curves can also be expected from similarity retrieval experiments on the Europeana dataset. Referring to the Europeana sounds evaluation results in Table 4 similar behaviour can be observed. Precision values of “Jazz”, “Smooth Jazz”, “Piano Concerto”, etc. drop by 20% on average from result-list lengths from 1 to 24 items. One objective indicator for huge declines of precision towards longer result-lists, are small ground-truth sizes. The possibility that all audio tracks of a 24 items ground-truth are on the same result-list page is very low. Thus, precision values of small ground-truth sets generally drop to values around one or two percent.

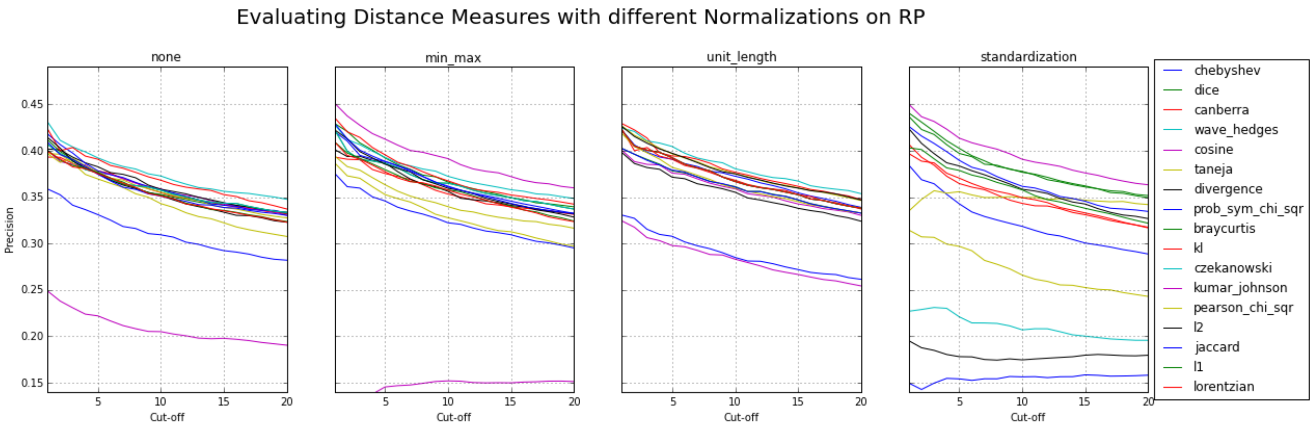


Figure 18: Performance evaluation of the Rhythm Patterns (RP) feature set with different normalization methods in combination with different distance measures on the MVD-MM subset of the Music Video Dataset. Numbers depict precision values for result-list length ranging from 1 to 20 entries.

Besides a direct comparison of the averaged results with results provided in literature, a more detailed examination of the presented results provides a good assessment of the performance of the MIR pilot’s implementation. The results of the high-level query terms “Jazz” and “Classic” are high compared to previously discussed values. Yet, they cannot be considered representative for music similarity, since Jazz and Classic have huge varieties and many sub genres differ in style, rhythm and instrumentation. Query 2 provides a more discrete insight by focussing on “Smooth Jazz”. This subgenre is stylistically more precisely described, and the annotators of the metadata obviously shared a common understanding of this description. This resulted in every second query delivering a correct result, and these high precisions are also noticed for longer result-lists. Similar performance values are observed for the queries 3-5 and 10-13. “Piano Concerto” for example describes classical music performed on pianos. Thus, it gives a clear description of the instrumentation and the expected timbre. Also the variation in timbre is clearly defined. This results in high precision values.

6 Conclusion

It is interesting that precision values on different datasets are comparable, particularly datasets with high varieties in their content (for example; different music genres, spoken content, animal sounds) show highly similar performance values. The chosen evaluation approach provided a good and comprehensive overview of the dataset and the performance of the implemented algorithm of the MIR pilot. Yet, flaws of the ground-truth queries were observed. One of these problems can be deduced by the queries 12 and 14 where the search terms were formulated in four different languages. For query 12 an increase in precision and ground-truth size was observed after adding further languages. This is a general weakness of string based approaches.

For this MIR pilot deliverable a user evaluation was performed to assess how the automatically calculated results are experienced by users. The user evaluation showed a rough linear correspondence between the calculated similarity estimation and the experienced similarity, with the recommendation to analyse this further in an online survey. For future integration of the MIR pilot results in production environments, like the Europeana Music Collection, WP2 that works closely with WP3 and WP4 to examine the IPR aspects of applying MIR technologies to the Europeana aggregation.

7 References

Ref 1	[Schindler 2012] Alexander Schindler, Rudolf Mayer, and Andreas Rauber. Facilitating comprehensive benchmarking experiments on the million song dataset. In Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012), pages 469-474, Porto, Portugal, October 8-12 2012.
Ref 2	[Schindler 2015] Alexander Schindler and Andreas Rauber. An audio-visual approach to music genre classification through affective color features. In Proceedings of the 37th European Conference on Information Retrieval (ECIR'15), Vienna, Austria, March 29 - April 02 2015.
Ref 3	[Lidy 2005] Thomas Lidy, Andreas Rauber.Evaluation of Feature Extractors and Psycho-acoustic Transformations for Music Genre Classification. Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR 2005), pp. 34-41, London, UK, September 11-15, 2005.
Ref 4	[Logan 2001] Logan, Beth, and Ariel Salomon. "A music similarity function based on signal analysis." *null*. IEEE, 2001.
Ref 5	[Fu 2011] Fu, Zhouyu, et al. "A survey of audio-based music classification and annotation." *Multimedia, IEEE Transactions on* 13.2 (2011): 303-319.
Ref 6	[Bartsch 2001] Bartsch, Mark, and Gregory H. Wakefield. "To catch a chorus: Using chroma-based representations for audio thumbnailing." *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*. IEEE, 2001.
Ref 7	[Tzanetakis 2002] Tzanetakis, George, and Perry Cook. "Musical genre classification of audio signals." *Speech and Audio Processing, IEEE transactions on* 10.5 (2002): 293-302.
Ref 8	[Scheirer 1997] Scheirer, Eric, and Malcoh Slaney. "Construction and evaluation of a robust multifeature speech/music discriminator." *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*. Vol. 2. IEEE, 1997.
Ref 9	[Harte 2006] Harte, Christopher, Mark Sandler, and Martin Gasser. "Detecting harmonic change in musical audio." *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. ACM, 2006.
Ref 10	[Dixon 2007] Dixon, Simon. "Evaluation of the audio beat tracking system beatroot." *Journal of New Music Research* 36.1 (2007): 39-50.
Ref 11	[Cha 2007] Cha, Sung-Hyuk “Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions” INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES, 2007
Ref 12	[Taneja 2005] Taneja, Inder Jeet. "Generalized symmetric divergence measures and inequalities." *arXiv preprint math/0501301* (2005).
Ref 13	[McFee 2015] McFee, Brian, et al. "librosa: Audio and Music Signal Analysis in Python." *Proceedings of the 14th Python in Science Conference*. 2015.
Ref 14	[Serrà 2012] Serrà, Joan, et al. "Measuring the evolution of contemporary western popular music." *Scientific reports* 2 (2012).

Appendix A: Terminology

A project glossary is provided at: http://pro.europeana.eu/web/guest/glossary.

Additional terms are defined below:

Term	Definition
AB	Advisory Board
APEX	Archives Portal Europe network of excellence
BPM	Beats per Minute
EC-GA	Grant Agreement (including Annex I, the Description of Work) signed with the European Commission
IPR	Intellectual Property Rights
MFCC	Mel Frequency Cepstral Coefficients
MIR	Music Information Retrieval
PC	Project Coordinator
PI	Performance Indicator
PMB	Project Management Board
QBE	Query-by-example
RP	Rhythm Patterns
STFT	Short-time Fourier Transform
SSD	Statistical Spectrum Descriptors
TSSD	Temporal Statistical Spectrum Descriptor
TRH	Temporal Rhythm Histograms
UAP	User Advisory Panel
WP	Work Package