



ELSEVIER

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-Western and ethnic music collections

Thomas Lidy^{a,*}, Carlos N. Silla Jr.^b, Olmo Cornelis^c, Fabien Gouyon^d, Andreas Rauber^a, Celso A.A. Kaestner^e, Alessandro L. Koerich^f

^a Vienna University of Technology, Austria

^b University of Kent, Canterbury, UK

^c University College Ghent, Belgium

^d Institute for Systems and Computer Engineering of Porto, Portugal

^e Federal University of Technology of Parana, Brazil

^f Postgraduate Program in Informatics, Pontifical Catholic University of Paraná, Brazil

ARTICLE INFO

Article history:

Received 5 December 2008

Received in revised form

15 September 2009

Accepted 17 September 2009

Keywords:

Ethnic music

Latin music

Non-Western music

Audio analysis

Music information retrieval

Classification

Access

Self-organizing map

ABSTRACT

With increasing amounts of music being available in digital form, research in music information retrieval has turned into a dominant field to support organization of and easy access to large collections of music. Yet, most research is focussed traditionally on Western music, mostly in the form of mastered studio recordings. This leaves the question whether current music information retrieval approaches can also be applied to collections of non-Western and in particular ethnic music with completely different characteristics and requirements.

In this work we analyze the performance of a range of automatic audio description algorithms on three music databases with distinct characteristics, specifically a Western music collection used previously in research benchmarks, a collection of Latin American music with roots in Latin American culture, but following Western tonality principles, as well as a collection of field recordings of ethnic African music. The study quantitatively shows the advantages and shortcomings of different feature representations extracted from music on the basis of classification tasks, and presents an approach to visualize, access and interact with ethnic music collections in a structured way.

© 2009 Published by Elsevier B.V.

1. Introduction

The availability of large volumes of music in digital form has spawned immense research efforts in the field of music information retrieval. A range of music analysis

methods have been devised that are able to extract descriptive features from the audio signal. These are being used to structure large music collections, organize them into different categories, or to identify artists and instrumentation. They also serve as a basis for novel access and retrieval interfaces, allowing users to create personalized playlists, find preferred songs they would like to listen to or to interact with large music collections. Many of these approaches have by now found their way into commercial products.

However, most of this research has been carried out predominantly on Western music. This may be due to the

* Corresponding author.

E-mail addresses: lidy@ifs.tuwien.ac.at (T. Lidy), cns2@kent.ac.uk (C.N. Silla Jr.), olmo.cornelis@ugent.be (O. Cornelis),

fgouyon@inescporto.pt (F. Gouyon), rauber@ifs.tuwien.ac.at (A. Rauber), celsokaestner@utfpr.edu.br (C.A. Kaestner), alekoe@ppgia.pucpr.br (A.L. Koerich).

easier availability of Western music in digital form. It may also reflect the larger familiarity of both the research community as well as the public in general with Western music, making evaluation of new approaches easier and leading to quicker industry take-up.

On the other hand, ethnic audio archives—collections of recordings from oral or tribal cultures—hold huge volumes of valuable music, collected by researchers all over the world over long periods of time. These form the basis of our musical cultural heritage. As a result of large and ongoing digitization and preservation projects, increasing volumes of ethnic music are becoming available in digital form, offering the basis for wider access and greater uptake. In order to fully unlock their value, these collections need to be made accessible with the same ease of use as current commercial music portals.

With ethnic music being in some aspects drastically different from Western music the question arises, in how far the research results stemming from traditional music information retrieval (Music IR) research can be directly applied. This question is not only posed for ethnic music collections but also for other non-Western music, such as Greek folk music, Latin American music, or traditional Indian music. Are the same audio description methods useful, although predominantly tested on music following Western tonality and rhythm principles? Does the optimization of these approaches to Western music benchmark collections lower applicability to non-Western music? Can comparable performance be obtained when trying to categorize ethnic and non-Western music automatically? Can Music IR provide tools and interfaces that allow researchers in ethnic music to access and evaluate their holdings in a sophisticated way, and may these interfaces also serve as an entry point for the general public, thus opening ethnic and cultural music collections to a larger user community?

Three music collections with different characteristics form the basis for detailed evaluations in this paper to address these questions. The first one is a common benchmark collection in Music IR research, consisting of predominantly Western style classical and Rock/Pop music, with some other genres such as World music mixed in. The second one is a collection of Latin American dance music, exhibiting dominating characteristics in terms of instrumentation and rhythm from a particular cultural domain, while still being strongly dominated by Western tonality and having been arranged and mastered using advanced studio technology. The third database consists of a collection of African ethnic music provided by the Ethnomusicological Sound Archive of the Belgian Royal Museum for Central Africa. This collection has drastically different recording standards, uses entirely different instruments and also has drastically different structures corresponding to music functions, geographic information, etc.

The tasks addressed in this article include specifically classification, where music is to be sorted automatically into various categories. These categories differ both in type (genre, instrumentation, geographical region, function) as well as their granularity. While classification of music is only one of many Music IR related tasks, it is also

utilized for evaluation of the audio analysis methods that constitute the fundamental step of many other applications. We thus performed a systematic evaluation of a range of state-of-the-art audio feature extraction algorithms. Support vector machines (SVM) and ensemble classifiers based on time decomposition are used to evaluate performance differences in various settings. Apart from the automatic categorization of music archives we also present an interface to access music collections, based on self-organizing maps (SOM), that facilitates visual exploration and intuitive interaction with music collections and evaluated it regarding its suitability to help in the analysis and usage of ethnic music collections.

This article is organized as follows. Particular aspects to consider when working with ethnic music collections are described in Section 2. In Section 3, a review of the state-of-the-art in relevant fields of music information retrieval is given alongside previous related work on automatic analysis of ethnic and other non-Western music. Section 4 takes a detailed look on audio signal analysis and feature extraction methods that form the basis for the subsequent tasks. Section 5 then outlines the classification approaches used, describes the three characteristically distinct music databases used in the experiments in detail and presents comprehensive evaluation results on various classification strategies on the three databases. Section 6 presents the SOM-based access principles alongside a qualitative evaluation of music map interfaces based on the same three music collections. Conclusions are presented in Section 7, including remarks on issues to be addressed and an outlook on future work.

2. Peculiarities of ethnic music

Preparing an audio data set for Music IR-based research is not just about gathering available audio to build a collection. It needs a well-thought-out scheme of actions and intentions—considering both musical content and formal aspects—especially in the context of non-Western music.

While Western studio recordings are produced by specialists in an idealized environment with a clean song as a result and almost always no direct link between the producer and the consumer, ethnic music recordings are almost always made in the field and not in studio. They reflect a unique moment full of serendipity. Ethnic music is performed with a specific, often social, function serving its community, for instance with court songs, songs for rituals, songs for hunting, praise songs, work songs, etc. Western music is mainly produced for entertaining purposes. Behind the distribution of Western music generally a commercial motive is hidden [1], while for ethnic music it is passed through orally from generation to generation. Orally because there is no written culture, resulting in a musical framework that has neither defined rules nor concepts, an immense contrast with the Western music that relies on a very well-defined musical system. Because of these numerous differences, correct interpretation of ethnic music is not so evident, and researchers must always be aware not to pinpoint ethnic music on

Western musical theory or fall back on the existing musical concepts. Tzanetakis et al. [2] notice, for example, the opposition of Western music with its notion of a composition as a well-defined work to other music cultures where the boundaries between composition, variation, improvisation and performance are more blurred and factors of cultural context have to be taken into account. Since ethnic music comes forward from an oral culture, a popular song can be brought by several, even dozens of local musicians, resulting in plural versions of one song, sometimes of varying quality, but often with personal influences affecting semantics, musical interpretation, instrumentation, and duration.

2.1. Musical content considerations

Audio from archival institutions can display diverging characteristics on pitch, temporal and timbral level if compared with commercial recordings. For instance, we can identify tendencies to differ in rhythmic aspects. Part of modern commercialized Western music (that is recorded, produced and mastered in studio) tends to show rhythmic aspects that are more controlled than in most ethnic music: deviations from perfect tempo and timing are most likely to be cautiously and systematically designed (or even avoided entirely), as opposed to ethnic music, more prone to emerging (bottom-up) tempo and timing deviations as well as to timing errors: for instance at the level of rubato or micro-timing [3,4], but also at a higher level such as the, probably intentional, constant speeding up during an entire song, or the, probably not intentional, slowing down of a group of singers if no percussion instruments are present.

Concerning meter, two main considerations have to be noticed: Western music handles a top-down approach starting with the major unit (a bar, or even a sentence), which is then rhythmically divided into smaller units, usually binary and ternary. Ethnic music tends to be organized additively: a bottom-up approach where the smallest unit can be seen as the starting point for extensions.

In African music timbre is an important aspect: since the variety of pitch and harmony is sober and the melodic lines are repetitive, other ways for enriching the sound are explored. For example, looking at African organology, the *lamellaephone* (also called thumb piano, *sanza*, *ikembe*) has some specific construction details by which its timbre range extends: every instrument is provided with a sound hole on the ventral side of the instrument. The performer can open and close this hole with his or her abdominal wall, generating a timbral and dynamic change of the sound. The lamellaephone is also often equipped with small metal rings that can vibrate when played. Another timbre-related aspect is the choice of material for building the lamellae, the soundboard and the sound box. A wide range of materials is being used to build musical instruments: specific materials can be wood, metal, turtle, reed, seeds, and in a few cases even human skulls have been used for the construction of the resonator, resulting in very different timbre, in spite of being the same

instrument class. Concerning the representation of timbre related research, Western music has only few semantic labels dealing directly with timbre. Timbre is often referred to by descriptive terms, such as dark or brilliant, opaque or transparent, or in terms of metaphors, such as colors or moods. Frequently, Music IR applications avoid the semantic allocation of timbre by the use of similarity retrieval, recommendation or representations such as the SOM. For ethnic music that might also be the best way to handle timbre-related features.

The final remark is about the parameter pitch: when analyzed very precisely, the annotated frequencies that build the musical scale are deviating from the Western scales that are usually tuned according to the well-tempered 100 cents based 12 tone system. Representation of ethnic music scales are often referring to these Western note names, in the best case with their specific individual deviation mentioned. But it is conceptually wrong to try to relate the musical profiles onto the Western pitch classes [5].

2.2. Formal aspects

Ethnic music is usually the product of a field recording implying that its creation was in no case an optimal environment for achieving a perfect recording. The noise level can be very high, depending on the age of the recording, the recording and playback equipment, the deterioration of the original analogue carriers and the amount of time spent during the time-consuming digitalization process. Diverging levels of loudness can occur over separate collections or even within one collection. Some tracks even show unstable speed of the recording resulting in a pitch and tempo shift that is very

Table 1

African music database: functions and number of instances per function.

Festive song	48	
Entertainment	178	101
Dance song	112	
Work song	13	103
Narrative song	68	
Evening song	8	
Court music	45	105
War song	22	
Religious song	14	107
Praise song	54	
Historical song	25	
Hunting	68	109
Cattle	44	
Messages	21	111
Ritual music	79	
Birth	19	
Funeral	21	112
Lullaby	21	
Mourning	26	113
Wedding	22	
Narrative	4	
Satire	1	114
Complaint	2	
Riddle	1	115
Fable	1	
Love song	4	116
Song of grace	4	

hard to correct. Browsing an audio archive reveals the very diverging duration of audio tracks at first sight. While the shortest items in an archive are only a few seconds long, for example when presenting a scale or a very short message of a slit drum, the longest tracks overrun more than 1 h, dealing for example with a ceremony or a dance [6]. For some older collections, the beginning of audio tracks contains spoken information provided by the ethnomusicologist itself. A valuable attempt to provide audio of a physical attachment to its own meta-data, this unfortunately confuses common Music IR algorithms. An important remark concerning the meta-data of ethnic music compared to Western music is the availability and relevance of very different fields of information. Music from an oral culture will attribute no importance to composer, some importance to performer, but then again meta-data such as date and place of recording are more relevant. There is no genre label in a similar sense as the genres of Western music, rather the function of a song is an important attribute. Exemplary functions of an African ethnic music collection are given in Table 1, Section 5.2.3 contains a more specific description of the attributes of this collection.

3. Related work

In Western music, as opposed to what has been said about ethnic music in the previous section, the meta-data fields most frequently used (and searched for) are song title, name of artist, performer or band, composer, album, etc.—and a very popular additional one: the “genre” [7]. However, the concept of a genre is quite subjective in nature and there is no clear standard on how to assign a musical genre [8,9]. Nevertheless, its popularity has led to its usage not only in traditional music stores, but also in the digital world, where large music catalogues are currently labeled manually by genres. However, assigning (possibly multiple) genre labels by hand to thousands of songs is very time-consuming and, moreover, to a certain degree, dependent on the annotating person. Research in Music IR has therefore tackled this problem already in a variety of ways.

A brief analysis of the state of the art shows that there are different approaches in Music IR for the (semi-)automatic description of the content of music. In *content-based* approaches, the content of music files is analyzed and descriptive features are extracted from it. In case of audio files, representative features are extracted from the digital audio signal [10]. In case of symbolic data formats (e.g. MIDI or MusicXML), features are derived from notation-based representations [11]. Additionally, semantic analyses of the lyrics can help in the categorization of music pieces into categories that are not predominantly related to acoustic characteristics [12]. Community meta-data have also been used for such tasks, for instance, collaborative filtering [13], co-occurrence analysis (e.g. on blogs and other music related texts in the web [14,15]), or analysis of meta-information provided by users on dedicated third-party sources (e.g. social tags on last.fm [16]). In cases where manpower is available, expert

analyses are an alternative and can provide powerful representations of music collections extremely useful for automatic categorizations (as in the case of Pandora and the Music Genome Project,¹ or AMG Tapestry²). Hybrid alternatives also exist, they combine several of the previous approaches, e.g. combining audio and symbolic analyses [17], audio features, symbolic features and community meta-data [18] or combining audio content features and lyrics [19]. Although hybrid approaches have proved to be usually better than using a single approach, there are some implications on their use beyond traditional Western music. First of all, naturally, there is a lack of publicly available meta-data for non-Western and ethnic music, which could be used as a resource for hybrid approaches. Moreover, both community meta-data and lyrics-based approaches are dependent on natural language processing (NLP) tools, which are usually not in the same stage of development for English as opposed to other languages. Moreover, as seen in [20] the adaptation of an NLP method from one language to another is far from trivial. This is especially true for ethnic music where the NLP resources might not even exist.

While Music IR research has resulted into a wide range of methods and (also commercial) applications, non-Western music was rarely the scope of this research, and only little research has been performed with focus on ethnic music. Although ethnomusicology is a very traditional field of study with many institutions, both archival and academic, involved, research on the signal level has rarely been performed. Charles Seeger was one of the first researchers to objectively measure, analyze and transcribe sound, using his Melograph [21]. Later, pitch analysis on monophonic audio to score has also been performed by Nesbit et al. onto Aboriginal music [22]. Krishnaswamy focused on pitch contour enhancing annotations by assigning typologies of melodic atoms to musical motives from Carnatic music [23], a technique that is also employed by Chordia et al. on Indian music [24]. Moelants et al. point out the problems and opportunities of pitch analysis of ethnic music concerning the specific tuning systems differing from the Western well-tempered system [25]. Duggan et al. [26] analyzed pitch extraction results achieving segregation of several parts of Irish songs. Pikrakis et al. and Antonopoulos et al. performed meter annotation and tempo tracking on Greek music, and later also on African music [27,28]. Wright focuses on micro-timing of Rumba music, visualizing the smallest deviations of performance opposed to the transcription by the traditional theoretical musical framework [3]. A similar work on Samba music is done in [4]. Only very few authors presented work related to timbre and its usefulness in genre classification of ethnic music [29,30]. The term Computational Ethnomusicology was emphasized by Tzanetakis, capturing some historical, but mostly recent research that refers to the design, development and usage of computer tools within the context of ethnic music [2].

¹ <http://www.pandora.com/>, http://en.wikipedia.org/wiki/Music_Genome_Project

² <http://www.amgtapestry.com/>

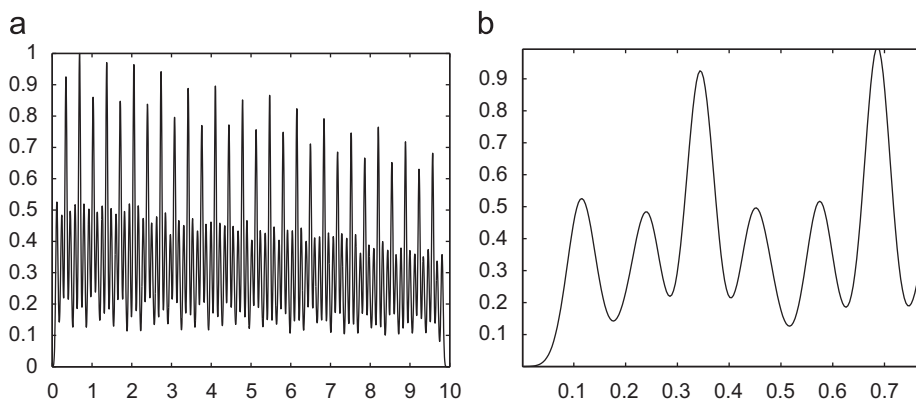


Fig. 1. Inter-onset interval histogram (IOIH): (a) IOIH [0,10] s, (b) IOIH detail view [0.0–0.78] s.

4. Audio analysis and feature extraction

A wealth of audio analysis and feature extraction methods has been devised in the Music IR research area for automated description of music [31]. Major approaches have been reviewed in Section 3 on related work. These feature extraction algorithms are employed in tasks such as automatic music classification, retrieval by (acoustic) similarity, or organization of music archives.

The set of algorithms used in our study comprises features from the MARSYAS framework developed by Tzanetakis et al. [10], inter-onset interval histogram coefficients (IOIHC) by Gouyon et al. [32], rhythm patterns (RP), by Rauber et al. [33] and its derivatives statistical spectrum descriptors (SSD) and rhythm histograms (RH) by Lidy et al. [34]. Additionally, two novel feature sets, based on SSD and RP features, are introduced in this article: temporal SSD and modulation variance descriptors (MVD). Following a brief description of all these feature extraction algorithms in Section 4.1 the creation of hybrid feature sets based on them is detailed in Section 4.2.

4.1. Feature extraction algorithms

4.1.1. MARSYAS features

The MARSYAS framework implements the original feature sets proposed by Tzanetakis and Cook [10]. The features can be divided into three groups: features describing the timbral texture (STFT and MFCC features), features for the rhythmic content (BEAT) and features related to pitch content (PITCH). The features for timbral texture are based on the short-time Fourier transform (STFT) and computed by the mean and variance of framewise spectral centroid, rolloff, flux, the time domain zero crossings, as well as the first five Mel-frequency cepstral coefficients (MFCCs) and low energy. Rhythm-related features aim at representing the regularity of the rhythm and the relative saliences and periods of diverse levels of the metrical hierarchy. They are based on a particular rhythm periodicity function: the so-called “beat histogram” (representing beat strength) and include statistics of the histogram (relative amplitudes, periods, ratios of salient peaks, as well as the overall sum of the

histogram as an indication of beat strength). Pitch related features include the maximum periods of the pitch peak in the pitch histograms. The conjoint features form a 30-dimensional feature vector (STFT: 9, MFCC: 10, PITCH: 5, BEAT: 6).

4.1.2. Inter-onset interval histogram coefficients (IOIHC)

This pool of features tap into rhythmic properties of sound signals. They are computed from a particular rhythm periodicity function, the inter-onset interval histogram (IOIH) [35], that represents (normalized) salience with respect to period of inter-onset intervals present in the signal (in the range 0–10 s, cf. Fig. 1). The IOIH is further parameterized by the following steps: (1) projection of the IOIH period axis from linear scale to the Mel scale (by means of a filterbank), (2) IOIH magnitude logarithm computation, and (3) inverse Fourier transform, keeping the first 40 coefficients.

These steps intend to be an analogy of the Mel-frequency cepstral coefficients (MFCCs), but in the domain of rhythmic periodicities rather than signal frequencies. The resulting coefficients provide a compact representation of the IOIH envelope [32]. Roughly, lower coefficients represent the slowly varying trends of the envelope. It is our understanding that they encode aspects of the metrical hierarchy providing a high level view on the metrical richness, independently of the tempo. Higher coefficients, on the other hand, represent finer details of the IOIH. They provide a closer look at the periodic nature of this periodicity representation and are related to the pace of the piece at hand (its tempo, subdivisions and multiples), as well as to the rhythmical salience (i.e. whether the pulse is clearly established). This is reflected in the shape of the IOIH peaks: relatively high and thin peaks reflect a clear, stable pulse.

4.1.3. Rhythm pattern (RP)

A rhythm pattern is a set of features based on psycho-acoustical models, capturing fluctuations on frequency bands critical to the human auditory system [33,36]. In the first part, the spectrogram of audio segments of approximately 6 s (2^{18} samples) in length is computed using the short time fast Fourier transform (STFT) with a

6

T. Lidy et al. / Signal Processing ■ (■■■■) ■■■–■■■

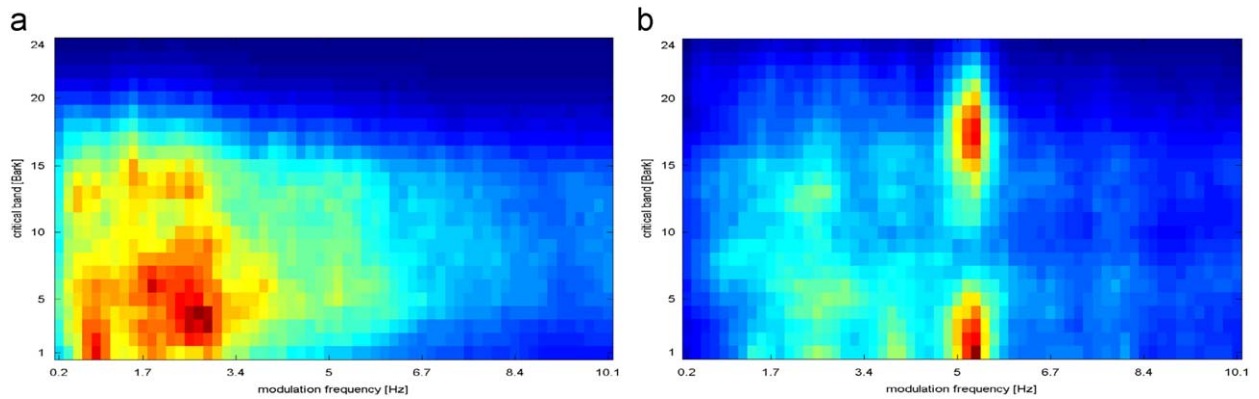


Fig. 2. Rhythm pattern: (a) Classical (b) Rock.

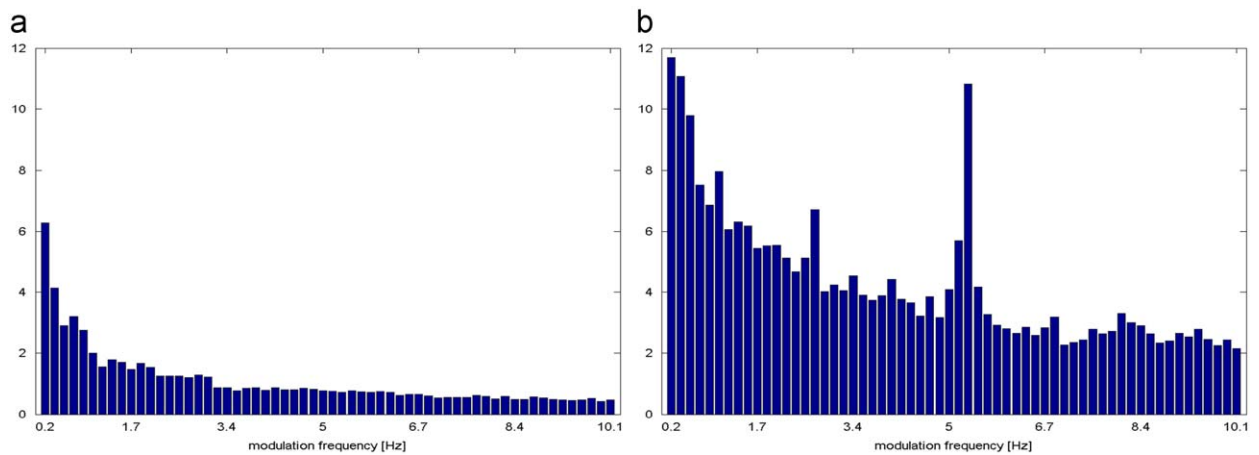


Fig. 3. Rhythm histograms: (a) Classical (b) Rock.

1024 samples³ large Hanning window and 50% overlap. The Bark scale, a perceptual scale which groups frequencies to critical bands according to perceptive pitch regions, is applied to the spectrogram, aggregating it to 24 frequency bands [37]. The Bark-scale spectrogram is then transformed into the Decibel scale. Further psycho-acoustic transformations are applied: computation of the Phon scale to incorporate equal loudness curves which account for the different perception of loudness at different frequencies and transformation into the Sone scale [37] to account for perceived relative loudness. The resulting Bark-scale Sonogram reflects the specific loudness sensation of an audio segment by the human ear.

In the second part, the varying energy on the critical bands of the Bark scale Sonogram is regarded as a modulation of the amplitude over time and its so-called “cepstrum” is retrieved by applying the Fourier transform. The result is a time-invariant signal that contains magnitudes of modulation per modulation frequency per critical band. This matrix represents a rhythm pattern, indicating occurrence of rhythm as vertical bars, but also

describing smaller fluctuations on all frequency bands of the human auditory range. Subsequently, modulation amplitudes are weighted according to a function of human sensation of modulation frequency, accentuating values around 4 Hz, and cutting off frequencies >10 Hz. The application of a gradient filter and Gaussian smoothing improves similarity between rhythm patterns. The final 24 × 60 feature matrix is computed by the median of segmentwise rhythm patterns. Fig. 2 shows examples of a rhythm pattern for a classical piece and a rock piece. While the rock piece shows a prominent rhythm at a modulation frequency of 5.34 Hz, both in the lower critical bands (bass) as well as in higher regions (percussion, e-guitars), the classical piece does not exhibit such a distinctive rhythm but focuses on mid/low critical bands and low modulation frequencies.

4.1.4. Rhythm histogram (RH)

A rhythm histogram aggregates the modulation amplitude values of the 24 individual critical bands computed in a rhythm pattern (before weighting and smoothing), exhibiting the magnitude of modulation for 60 modulation frequencies between 0.17 and 10 Hz [34]. It is a lower-dimensional descriptor for general rhythmic

³ For a sampling rate of 44,100 Hz; adjusted proportionally for lower rates.

characteristics in a piece of audio ($N = 60$, as compared to the 1440 dimensions of an RP). A rhythm histogram is computed for each 6s segment in a piece of audio and the feature vector is then averaged by taking the median of the feature values of the individual segments (cf. Section 4.1.3).

Fig. 3 compares the rhythm histograms of a classical piece and a rock piece (the same example songs as for illustrating rhythm patterns have been used). The rock piece indicates a clear peak at a modulation frequency of 5.34Hz while the classical piece generally contains less energy, having most of it at low modulation frequencies.

4.1.5. Statistical spectrum descriptor (SSD)

In the first part of the algorithm for computation of a statistical spectrum descriptor (SSD) the specific loudness sensation is computed on 24 Bark-scale bands (i.e. a Bark-scale Sonogram), analogously to a rhythm pattern. Subsequently, statistical measures are computed from each of these critical bands: mean, median, variance, skewness, kurtosis, min- and max-value, describing variations on each of the bands statistically. The SSD thus describes fluctuations on the critical bands and captures additional timbral information not covered by other feature sets, such as a rhythm pattern. At the lower dimension of 168 features this feature set is able to capture and describe acoustic content very well [34].

4.1.6. Temporal statistical spectrum descriptor (TSSD)

Feature sets are frequently computed on a per segment basis and do not incorporate time series aspects. As a consequence, TSSD features describe variations over time by including a temporal dimension. Statistical measures (mean, median, variance, skewness, kurtosis, min and max) are computed over the individual statistical spectrum descriptors extracted from segments at different time positions within a piece of audio. This captures timbral variations and changes over time in the audio spectrum, for all the critical Bark-bands. Thus, a change of rhythmic, instruments, voices, etc. over time is reflected by this feature set. The dimension is 7 times the dimension of an SSD (i.e. 1176).

4.1.7. Modulation frequency variance descriptor (MVD)

This descriptor measures variations over the critical frequency bands for a specific modulation frequency (derived from a rhythm pattern, cf. Section 4.1.3). Considering a rhythm pattern, i.e. a matrix representing the amplitudes of 60 modulation frequencies on 24 critical bands, an MVD vector is derived by computing statistical measures (mean, median, variance, skewness, kurtosis, min and max) for each modulation frequency over the 24 bands. A vector is computed for each of the 60 modulation frequencies. Then, an MVD descriptor for an audio file is computed by the mean of multiple MVDs from the audio file's segments, leading to a 420-dimensional vector.

4.2. Hybrid features

We make the hypothesis that a hybrid feature set combining multiple feature sets capturing, as much as possible, complementary characteristics of the music will achieve a better performance in retrieval and classification tasks.

A preliminary evaluation of the previously described individual feature sets on music databases with different characteristics showed also that some feature sets—to be more specific: certain feature attributes—are more discriminative on particular music collections than on others, depending on the musical content. This is a good incentive to try out diverse feature set combinations when dealing with Western vs. non-Western and ethnic audio collections.

Tzanetakis and Cook already proposed a hybrid feature set within the MARSYAS framework, i.e. the combination of STFT, MFCC, PITCH and BEAT features [10], called “MARSYAS-All” in this paper. These features represent multiple aspects of musical characteristics (namely, timbral, tonal and rhythmic). In this paper we propose to extend the hybrid approach by replacing the low-dimensional BEAT features in MARSYAS by the higher-dimensional ones described in Section 4.1, which are assumed to achieve more precise results because they capture a larger number of rhythmical and, for some of them, timbral aspects in the music. On the other hand, some of the feature sets have a strong focus on specific musical facets (e.g. rhythm) and might benefit vice versa from the conjoint feature sets. A number of hybrid feature sets is created, each based on Marsyas STFT, MFCC and PITCH + another feature set and the assumptions stated above are examined experimentally in Section 5.

5. Automatic music classification

A frequent scenario for the organization of audio archives is the categorization into a pre-defined list of categories (or, a related task, the assignment of class labels or tags). It is assumed that such a categorization, or classification, aids in managing an audio library. Based on audio feature extraction and a machine learning algorithm, classification of audio documents can be performed automatically. The machine learning research domain has developed a large range of classifier algorithms that can be employed. These algorithms are intended to find a separation of classes within the feature space. The approaches' premise is the availability of training data from which the learning algorithm induces a model to classify unseen audio documents.

In Section 5.1 we will briefly explain the classification approaches used in our study. Section 5.2 presents the data sets we used, containing Western, Latin American and African music. We will then present our experimental results on these databases in Section 5.3. We are investigating specific aspects of the Latin American and ethnic collections with regard to differences to classification of Western music, which is most frequently categorized into “genres”.

5.1. Classification methods

5.1.1. Support vector machines

A support vector machine (SVM) [38] is a classifier that constructs an optimal separating hyperplane between two classes. The hyperplane is computed by solving a quadratic programming optimization problem, maximizing the distance of the hyperplane from its closest data vector. A “soft margin” allows a number of points to violate these boundaries. Except for linear SVMs the hyperplane is not constructed in the feature space but a kernel is used to project the feature vectors to a higher-dimensional space, in which the problem becomes linearly separable. Polynomial or radial basis function (RBF) kernels are common, however, for high-dimensional problems frequently a linear SVM performs equally or even better.

The sequential minimal optimization (SMO) algorithm is used in our approach, which breaks the large quadratic programming optimization problem of an SVM into a series of smallest possible problems, reducing both memory consumption and computation time, especially for linear SVMs [39].

5.1.2. Time decomposition

Combining multiple classifiers has been shown to improve efficiency and accuracy. Kittler et al. [40] distinguish from two different scenarios for classifier combination. In the first scenario, all the classifiers use the same representation of the input pattern. Although each classifier uses the same feature vector, each classifier will deal with it in different ways. In the second scenario, each classifier uses its own representation of the input pattern.

In this work, we employ the time decomposition (TD) approach [41,42], which is an ensemble-based approach tailored for the task of music classification that is related to the second scenario described above. TD can be seen as a meta-learning approach for the task of music classification as it is not dependent on any particular feature set or classifier. Feature vectors are frequently computed for individual segments of an audio document (cf. Section 4). When using this segmentation strategy it is possible to train a specific classifier for each one of the segments and to compute the final class decision from the ensemble of the results provided by each classifier. There are different ways to combine this information. In this paper we use majority voting (MAJ), the MAX rule (i.e. the output of the classifier with the highest confidence is chosen), the SUM and the PROD rules, where the probabilities for each class from each classifier are summed or multiplied, respectively, and the highest one is chosen.

5.2. Test collections

5.2.1. Western music database

As a reference we use a popular benchmark database of Western music. The collection was compiled for the genre classification task of the ISMIR 2004 Audio Description contest [43,44] and used frequently thereafter by Music IR researchers. The set of 1458 songs is categorized into six popular Western music genres: classical (640 pieces),

electronic (229), jazz and blues (52), metal and punk (90), rock and pop (203), world (244). While the “world” music genre partially covers non-Western music as well, this coarse genre subdivision is typical for average users of Western music collections.

5.2.2. Latin music database

The Latin music database (LMD) [45] contains 3227 songs, which were manually labeled by two human experts who have over 10 years of experience in teaching Latin American dances. The data set is categorized into 10 Latin music genres (Axé, Bachata, Bolero, Forró, Gaúcha, Merengue, Pagode, Salsa, Sertaneja, Tango). Contrary to popular Western music genres, each of these genres has a very specific cultural background and is associated with a different region and/or ethnic and/or social group. Nevertheless, it is important to note that in some aspects the Latin music database is musically similar to the Western music database as it makes use of modern recording and post-processing techniques. By contrast to the Western music database, the LMD contains at least 300 songs per music genre, which allows for balanced experiments.

5.2.3. African music database

The collection of African music used in this study is a subset of 1024 instances of the audio archive of the Royal Museum of Central-Africa (RMCA)⁴ in Belgium, kindly provided by the museum. It is one of the largest museums in the world for the region of Central-Africa, with an audio archive that holds 50,000 sound recordings from the early 20th century until now. This unique collection of cultural heritage is being digitized in the course of the DEKKMMA project [46], one goal being to provide enhanced access through the use of Music IR methods [47]. There is a lot of meta-data available for the collection, related to identification (number/id, original carrier, reproduction right, collector, date of recording, duration), geographic information (country, province, region, people, language), and musical content (function, participants, instrumentation). Unfortunately, not for every recording all fields are available, as often these data cannot be traced.

A number of these meta-data can be used to investigate the methods of Music IR for classification and access. One important meta-data field investigated in this study is the “function”, describing specific purposes for individual pieces of music. Table 1 shows the number of instances for the 27 different functions available in the collection. The database is partially annotated by instrumentation, with a 3-level hierarchy and a single or multiple instruments per song. Level 1 is a categorization by instrument family, on the second level there were 28 different instruments in the database, with an optional subtype on the third level. Instrument families and instruments on level 2 are given in Table 2.

Another category investigated was the country. The list of countries can be seen in Table 9. The database contains also a field with the name of the people (ethnic group) who played the music. Six hundred and ninety three

⁴ <http://music.africamuseum.be>

Table 2

African music database: instrument families and instruments.

Aerophone	Flute, flute (European), horn, pan pipe, whistle, whistling
Chordophone	Fiddle, guitar, harp, lute, musical bow, zither
Idiophone	Bell, handclapping, lamellaphone, percussion pot, pestle, rattle, rhythm stick, scraper, sistrum, slit-drum, struck object, xylophone
Membranophone	Drum, friction drum, single-skin drum, double-skin drum

instances have been annotated with an ethnic group, in total 40 different ethnic groups are known in the database.

5.3. Experimental results

We investigated the classification of audio documents measuring accuracy on a multi-class classification task. The Weka Machine Learning tool [48] was employed in all experiments, using the SMO algorithm, and, in a subset of experiments, the time decomposition approach on top of it. Linear SVMs were trained, with the complexity parameter c set to 1. All experiments were run using stratified 10-fold cross-validation. Potential improvements of the time decomposition ensemble approach over a single SVM were investigated, as well as a comparison of analyzing different audio segments. Apart from the latter experiment, all experiments are based on a feature analysis of the center 30s of the pieces in the music collections. Results are presented as accuracy values in percent and the standard deviation. Though the numerical results cannot be directly compared between the three databases, due to different organization schemes and semantics (e.g. genre vs. function) as well as different sizes of the collections and different numbers of classes, these results allow an assessment of how well the approaches are also applicable to non-Western music archives.

5.3.1. Results on the Western music database

Feature set comparison: The first experiment includes a comparison of the dependence of the results on the particular segment taken as excerpt for analysis from the audio signal. Three different 30-s audio segments were analyzed: the beginning, the center and the end part of each piece (Seg_{beg} , Seg_{mid} , Seg_{end}). Table 3 shows the results of this segmentwise analysis for all the different feature sets described in Section 4. The results indicate a rather moderate performance of rhythm and beat related features (e.g. RH, MARSYAS-BEAT, IOIHC) while other feature sets that capture more timbral information achieve higher results, with a classification accuracy of 76.12% using SSD features.

The MARSYAS-BEAT features have been replaced successively by other feature sets. The lower part of Table 3 presents the results for these hybrid feature sets, where the combination of MARSYAS (excl. BEAT) features with SSD achieved 79% accuracy on Seg_{mid} . In general, the hybrid approaches performed always better than the individual approaches, with all results based on the middle audio segment being higher than 71%. The

Table 3

Western music database: segment comparison (SVM classification by genre).

Feature set	Seg_{beg}	Seg_{mid}	Seg_{end}
MARSYAS-STFT	56.36 ± 1.42	61.72 ± 2.28	59.54 ± 1.84
MARSYAS-PITCH	45.66 ± 1.89	52.49 ± 2.42	49.70 ± 2.22
MARSYAS-MFCC	58.19 ± 2.34	65.47 ± 2.47	60.90 ± 2.80
MARSYAS-BEAT	52.06 ± 2.34	54.87 ± 2.15	53.55 ± 2.57
IOIHC	45.00 ± 1.56	49.71 ± 1.31	42.61 ± 1.10
RH	57.55 ± 1.86	62.84 ± 2.53	59.11 ± 2.86
RP	64.94 ± 3.95	69.78 ± 3.30	64.81 ± 3.91
SSD	71.20 ± 2.52	76.12 ± 3.76	72.71 ± 2.42
TSSD	64.02 ± 4.20	70.14 ± 3.39	65.76 ± 2.28
MVD	62.88 ± 2.51	68.47 ± 1.75	62.73 ± 2.40
MARSYAS-All	66.57 ± 3.03	71.85 ± 2.62	67.54 ± 2.29
HYBRID-IOIHC	64.08 ± 3.14	71.50 ± 2.07	64.48 ± 3.20
HYBRID-RH	66.87 ± 2.63	71.69 ± 2.12	68.94 ± 1.94
HYBRID-RP	71.03 ± 4.47	75.13 ± 2.98	71.65 ± 2.53
HYBRID-SSD	75.40 ± 3.03	79.00 ± 3.13	76.07 ± 2.94
HYBRID-TSSD	68.64 ± 4.39	74.85 ± 3.25	69.74 ± 2.30
HYBRID-MVD	68.76 ± 2.11	73.26 ± 2.34	68.02 ± 1.27

MARSYAS-All combination was improved in all but two cases (HYBRID-IOIHC, HYBRID-RH). The results imply that the additional feature sets capture musical (both rhythmic and a timbral) aspects better than the MARSYAS-BEAT features.

Segment comparison: The highest accuracy for Seg_{beg} is 75.40% with HYBRID-SSD features. For the Seg_{end} the highest accuracy is 76.07% with HYBRID-SSD features. The comparison among Seg_{beg} , Seg_{mid} and Seg_{end} shows that extraction of features from the center segment (Seg_{mid}) performs better in all cases. This comparison was performed also for the Latin music collection with quasi the same outcome, we therefore omit presenting the segment-analysis results for the Latin music database in the next section.

By contrast to the rather clear conclusion on the Western and Latin music databases, where there is a difference of up to 5 percentage points using Seg_{beg} or Seg_{end} instead of Seg_{mid} , the situation is different with the African music database, as will be shown in Section 5.3.3.

Time decomposition: The results presented in Table 4 are based on the ensemble of the features extracted from Seg_{beg} , Seg_{mid} and Seg_{end} . The overall highest result on individual feature sets was 77.20% using SSD features and the majority vote (MAJ) rule (the result using a single SVM was 76.12%). Overall, time decomposition (TD) improved the results for five of the feature sets, but in three cases the results were marginally worse than using linear SVM only. However, the best results were generated using different ensemble voting rules, with the MAX rule being most frequently the best rule, although SUM and PROD seem to be better when using hybrid feature sets.

The highest overall result is 80.37% using HYBRID-SSD features and the majority vote rule. The TD approach generally improved results for hybrid features only very moderately, except for H-RP and H-TSSD features, where the SUM and PROD rules achieved improvements by about 3.4 percentage points.

Table 4

Western music database: classification using the time decomposition approach.

Feature set	Ensemble rule			
	MAJ	MAX	SUM	PROD
MARSYAS-STFT	60.91 ± 1.27	61.58 ± 2.18	60.78 ± 1.33	61.21 ± 1.65
MARSYAS-PITCH	50.34 ± 1.45	52.49 ± 2.30	49.45 ± 1.40	49.45 ± 1.36
MARSYAS-MFCC	63.26 ± 2.50	65.12 ± 2.59	63.92 ± 2.59	63.84 ± 2.83
MARSYAS-BEAT	54.73 ± 2.64	54.65 ± 2.02	54.60 ± 3.34	53.86 ± 1.72
IOIHC	45.00 ± 1.60	49.71 ± 1.31	45.07 ± 1.52	45.07 ± 1.52
RH	60.81 ± 2.06	62.84 ± 2.44	61.58 ± 2.51	61.18 ± 2.11
RP	70.99 ± 3.23	70.39 ± 3.55	72.40 ± 2.61	72.00 ± 2.31
SSD	77.20 ± 3.28	76.19 ± 3.80	76.52 ± 2.57	76.73 ± 2.62
TSSD	72.39 ± 2.38	70.34 ± 3.58	73.42 ± 2.17	72.86 ± 2.99
MVD	67.63 ± 2.01	69.01 ± 2.29	68.65 ± 2.69	68.52 ± 2.72
MARSYAS-All	71.44 ± 2.45	72.21 ± 2.59	71.42 ± 2.26	71.14 ± 1.69
HYBRID-IOIHC	69.51 ± 2.83	71.64 ± 1.96	69.66 ± 1.47	69.65 ± 1.19
HYBRID-RH	70.80 ± 2.73	71.90 ± 2.07	71.45 ± 2.31	72.38 ± 1.93
HYBRID-RP	77.12 ± 3.49	75.54 ± 3.40	78.34 ± 2.54	78.57 ± 2.46
HYBRID-SSD	80.37 ± 2.65	79.02 ± 2.81	79.75 ± 2.65	79.54 ± 2.61
HYBRID-TSSD	78.18 ± 2.22	75.23 ± 3.24	78.25 ± 3.35	77.74 ± 3.46
HYBRID-MVD	73.49 ± 1.05	73.60 ± 1.98	73.80 ± 2.32	73.65 ± 2.72

Table 5

Latin music database: comparison of SVM and the time decomposition approach.

Feature set	SVM <i>Seg_{mid}</i>	Time decomposition	
		MAJ	SUM
MARSYAS-STFT	56.40 ± 2.13	57.73 ± 2.58	56.93 ± 2.32
MARSYAS-PITCH	25.83 ± 1.63	27.33 ± 0.96	29.53 ± 2.17
MARSYAS-MFCC	58.83 ± 2.31	60.20 ± 2.02	60.26 ± 2.86
MARSYAS-BEAT	31.86 ± 1.69	33.40 ± 1.48	34.56 ± 2.43
IOIHC	53.26 ± 2.63	52.53 ± 2.74	47.73 ± 2.76
RH	54.63 ± 1.94	56.96 ± 2.03	57.80 ± 2.27
RP	81.40 ± 1.45	84.76 ± 1.23	84.70 ± 1.25
SSD	82.33 ± 1.36	84.70 ± 1.50	84.06 ± 1.33
TSSD	73.80 ± 1.75	79.40 ± 2.23	81.70 ± 1.11
MVD	67.70 ± 2.75	71.66 ± 2.03	73.00 ± 1.96
MARSYAS-All	68.46 ± 2.03	70.40 ± 2.23	70.40 ± 1.99
HYBRID-IOIHC	77.63 ± 1.74	78.33 ± 1.82	77.13 ± 1.67
HYBRID-RH	74.50 ± 2.47	76.73 ± 2.19	77.16 ± 2.19
HYBRID-RP	84.06 ± 1.42	87.46 ± 1.66	88.06 ± 1.60
HYBRID-SSD	85.30 ± 1.39	87.53 ± 1.20	87.40 ± 1.01
HYBRID-TSSD	75.80 ± 1.93	81.93 ± 2.33	83.96 ± 1.58
HYBRID-MVD	77.06 ± 2.71	81.50 ± 1.56	81.50 ± 1.19

5.3.2. Results on the Latin music database

Feature set comparison: Classification with a single SVM compared by using different audio segments was always better using the center 30s of a song (*Seg_{mid}*). Therefore, results on the beginning and end segments are omitted in Table 5. It seems that both rhythm and timbre play a major role in discriminating Latin music genres, with rhythm patterns (RP) and SSD giving the best results (81.4% and 82.33%, respectively). It is especially noteworthy that pitch seems to play a subordinate role noticeable by the low performance of MARSYAS-PITCH features (25.83%). Pure rhythmic features deliver intermediate results (BEAT, IOIHC, RH). The hybrid approaches bring a major boost to them. This is explainable having a look at the Latin American genres

Table 6

Confusion matrix for Latin music genres, using RH features and SVM.

	Ta	Sa	Fo	Ax	Ba	Bo	Me	Ga	Se	Pa
Tango	263	0	2	0	0	26	0	7	1	1
Salsa	15	158	7	17	4	22	3	29	19	26
Forró	16	15	130	19	1	6	9	59	28	17
Axé	15	34	45	89	4	6	49	35	14	9
Bachata	5	1	1	4	237	4	28	7	9	4
Bolero	66	5	0	1	2	174	1	3	38	10
Merengue	1	2	13	33	17	2	211	8	13	0
Gaúcha	26	15	34	26	9	21	7	129	18	15
Sertaneja	6	16	29	18	28	59	9	15	104	16
Pagode	15	30	13	18	0	31	2	29	30	132

in the database where some genres have a similar rhythm. For example, Forró, Pagode, Sertaneja and Gaúcha are rhythmically similar and for that reason the other features help to distinguish between them. There are also similarities between Bolero and Tango. Additional evidence for these similarities is presented in the confusion matrix in Table 6 where a large portion of Bolero is misclassified as Tango, Sertaneja is confused with Bolero, numerous Pagode songs are misclassified as Salsa, Bolero, Gaúcha or Sertaneja, and many Forró songs are confused with Gaúcha.

The hybrid sets are significantly better than the MARSYAS-All approach in all cases. The addition of more complex features to the MARSYAS set instead of the BEAT features achieved a major increase in classification accuracy. The major trends are similar to the Western music database, although the specific combination of rhythmic and timbral characteristics in the RP features seems to be of particular use for Latin American music.

Time decomposition: The ensemble approach could increase the performance by several percent. The MAX and PROD rules were generally inferior to the MAJ and SUM rules and are therefore not presented. An interesting

aspect is that RP features surpassed SSD using time decomposition, a hint to more individual rhythmic characteristics extracted from individual segments. The same effect appears with HYBRID-RP features on the SUM rule, where the accuracy is as high as 88.06% (a 4% improvement over the single SVM).

5.3.3. Results on the African music database

The many meta-data fields available for the African music database allow classification by multiple facets. The results give an assessment about what kind of information can be detected by current feature analysis and classification approaches and potential challenges specific for classification of ethnic music.

Segment comparison: Before performing classification by different categories, we have carried out a pre-analysis of the performance of different audio segments, as we have done for the Western and Latin music databases. We

Table 7

African music database: segment comparison (SVM classification by function).

Feature set	Seg_{beg}	Seg_{mid}	Seg_{end}
MARSYAS-STFT	21.86 ± 2.63	23.14 ± 2.88	21.92 ± 2.50
MARSYAS-PITCH	21.09 ± 2.22	21.09 ± 2.22	21.09 ± 2.22
MARSYAS-MFCC	30.22 ± 3.91	27.19 ± 4.44	26.36 ± 3.53
MARSYAS-BEAT	21.09 ± 2.22	21.22 ± 2.30	21.09 ± 2.30
IOIHC	21.07 ± 2.28	21.07 ± 2.28	20.81 ± 2.33
RH	21.39 ± 3.07	22.10 ± 2.49	21.84 ± 2.77
RP	36.83 ± 3.42	37.85 ± 5.49	35.29 ± 4.03
SSD	44.27 ± 6.63	45.12 ± 5.98	44.34 ± 4.34
TSSD	37.16 ± 5.23	35.75 ± 3.23	37.14 ± 6.43
MVD	27.76 ± 3.12	28.28 ± 5.66	28.48 ± 4.50
MARSYAS-All	35.38 ± 4.57	32.39 ± 5.56	30.18 ± 4.47
HYBRID-IOIHC	36.72 ± 4.47	30.34 ± 5.45	29.90 ± 5.03
HYBRID-RH	34.21 ± 8.06	34.93 ± 5.33	33.83 ± 4.96
HYBRID-RP	39.63 ± 5.88	41.12 ± 5.81	38.82 ± 3.02
HYBRID-SSD	46.46 ± 6.33	48.25 ± 6.63	47.24 ± 4.27
HYBRID-TSSD	38.81 ± 4.93	38.10 ± 2.66	39.37 ± 7.03
HYBRID-MVD	33.69 ± 7.37	35.08 ± 4.63	35.65 ± 6.84

Table 8

African music database: classification by different meta-data using SVM.

Feature set	Function	Instrument	Country	Ethnic group
MARSYAS-STFT	23.14 ± 2.88	38.71 ± 4.77	58.61 ± 6.62	61.59 ± 9.34
MARSYAS-PITCH	21.09 ± 2.22	36.85 ± 3.92	47.20 ± 3.85	60.53 ± 9.24
MARSYAS-MFCC	27.19 ± 4.44	46.87 ± 8.31	54.64 ± 4.44	70.96 ± 10.84
MARSYAS-BEAT	21.22 ± 2.30	37.64 ± 4.46	50.25 ± 6.16	60.53 ± 9.24
IOIHC	21.07 ± 2.28	39.14 ± 6.88	55.19 ± 6.95	60.53 ± 9.24
RH	22.10 ± 2.49	44.34 ± 5.23	62.17 ± 6.73	63.41 ± 9.70
RP	37.85 ± 5.49	57.42 ± 5.48	72.24 ± 5.31	80.57 ± 5.31
SSD	45.12 ± 5.98	67.61 ± 7.87	81.74 ± 4.70	85.07 ± 5.07
TSSD	35.75 ± 3.23	69.06 ± 6.66	81.21 ± 3.76	84.88 ± 5.12
MVD	28.28 ± 5.66	47.90 ± 7.83	65.22 ± 3.63	68.41 ± 8.08
MARSYAS-All	32.39 ± 5.56	55.44 ± 7.72	64.24 ± 5.00	76.62 ± 6.93
HYBRID-IOIHC	30.34 ± 5.45	59.85 ± 7.04	68.59 ± 5.82	79.96 ± 7.08
HYBRID-RH	34.93 ± 5.33	59.75 ± 10.23	72.61 ± 5.02	81.01 ± 7.14
HYBRID-RP	41.12 ± 5.81	64.11 ± 5.37	77.00 ± 5.25	84.88 ± 5.43
HYBRID-SSD	48.25 ± 6.63	68.79 ± 7.77	82.21 ± 3.34	88.10 ± 3.75
HYBRID-TSSD	38.10 ± 2.66	68.82 ± 4.88	80.71 ± 4.16	88.57 ± 3.89
HYBRID-MVD	35.08 ± 4.63	57.59 ± 8.52	75.03 ± 3.41	77.55 ± 6.29

have carried out classification with SVM considering the function as classes. From Table 7 it is visible that, in general, there is much less difference between the different 30-s segments used for analysis, with deviations below 2 percentage points; for some feature sets, the performance is even equal. However, some other feature sets seem to perform particularly well on the initial segment (Seg_{beg}). This might be a hint that the beginning of ethnic music pieces may be important for characterizing its function, a circumstance that is the contrary with Western music and its frequent lead-in effects. The end segment provides mainly worse results than the center segment, which is similar to Western and Latin music, yet not to the same extent, pointing at a lower presence of fade-out effects in ethnic music. Generally, there is yet again the conclusion that usually the inner content of a piece of music contains more characteristics useful for classification. In the subsequent classification experiments, Seg_{mid} is used for evaluation.

Classification by function: The “function” field in the African music database was of specific interest to us, because it may be considered as the counterpart of the genre in Western music. From the originally 27 different functions available in the set (cf. Table 1) functions with less than 10 instances were ignored for the following experiments, in order to permit a proper cross-validation, keeping 19 functions. The second column of Table 8 provides the results of classification by function using the different feature sets and hybrid approaches (all based on Seg_{mid}). The accuracies achieved were rather low (the baseline considering the largest class is 19.7%); it seems that the concept of a “function” is not captured very well by the audio analysis methods used. The only prominent results are delivered by the SSD features, with 45.12% accuracy, followed by RP features with 37.85% and temporal SSD with 35.75%.

The use of the hybrid feature sets shows an improvement in accuracy well over the baseline for all feature sets compared to the original individual feature sets. They also show an improvement over the MARSYAS-All combination

in all except the HYBRID-IOIHC case. However, the improvement of the best result (SSD) was moderate, with HYBRID-SSD achieving 48.25%. Evaluation of the time decomposition approach showed that the highest result on HYBRID-SSD could be further improved to 50.05% using the SUM rule.

Classification by instrumentation: We have conducted an experiment on classification of the instrument family. Seven hundred and eleven instances were labeled by instrumentation. A mixture of multiple instruments per song was considered as a separate class (see class list in Fig. 7). Instrument recognition naturally is not a task done with rhythmic features as can be also seen from column 3 in Table 8. Better results are accomplished clearly by the timbral features, with MFCC achieving 46.87%, SSD 67.61% and temporal SSD 69.06%. The hybrid set-up could improve the performance of the low-performing rhythm-based feature sets, but not the one of TSSD features and only slightly the one of SSD.

Classification by country: Our subset of the collection contained music from 11 African countries. The countries Eritrea and Niger were, however, represented by one piece each only and were thus ignored for this experiment as it would be impossible to create a training and test set for these.

The results of the classification by country (column 4 of Table 8) indicate that the audio features under investigation allow a proper classification into the originating countries of the African audio recordings. The results indicate that the rhythmic properties differ between the countries (with RH and RP features giving quite high results) but also that timbral aspects play a major role, probably due to the use of different instruments: SSD achieved an accuracy of 81.74% and TSSD 81.21%. Hybrid feature sets could improve these results only marginally.

The confusion matrix in Table 9 shows that the major confusion happens only between Congo DRC and Rwanda. Being geographical neighbors, these countries' cultures are very related and so is their music. Even with the dominance of audio instances available from the Congo DRC and Rwanda, classification of pieces of audio from the other less represented countries performed very well, with 100% recall and precision on classifying music from the Ivory Coast. Overall, precision and recall were 72.8% and 67.5%, respectively.

Table 9

Confusion matrix for African music by country, using TSSD features and SVM.

	Rw	Bu	Co	Ga	RC	Et	Se	Gh	Iv
Rwanda	324	5	65	1	2	0	1	0	0
Burundi	11	6	0	0	0	0	0	0	0
Congo DRC	74	0	384	0	0	0	0	2	0
Gabon	1	0	0	6	4	0	0	0	0
Republic of the Congo	1	0	0	4	7	1	1	0	0
Ethiopia	4	0	5	1	1	12	2	0	0
Senegal	1	0	0	0	1	1	7	0	0
Ghana	2	0	3	0	0	0	0	27	0
Ivory Coast	0	0	0	0	0	0	0	0	15

Classification by ethnic group: In this experiment, we have investigated whether our classification approach is able to separate the music according to the ethnic group that performed it. The baseline for this experiment was 50.22% as there were 348 instances from the “Luba” people. Column 5 of Table 8 shows that the feature sets based mainly on rhythm could not distinguish the music very well by ethnic group, but feature sets incorporating timbral aspects achieved remarkable results on classifying 40 ethnic groups: RP achieved 80.57% classification accuracy, SSD features accomplished remarkable 85.07%, and the TSSD features reached 84.88%. The hybrid approaches could further improve these results to an astounding classification accuracy of more than 88%.

Impressed by these high results, we tried to investigate whether there may be a direct correlation between the recording quality of the pieces of a specific ethnic group. Unfortunately, there was no reference to the recording equipment that was used at the time of the recording of the pieces in the database. From the meta-data we had available we could, however, investigate (1) whether there is an influence introduced by the bitrate used for encoding the recordings and (2) whether there is any correlation between the year of the recording and the ethnic group. Over all the different recordings, only three different MP3 bitrates were used: 128, 192 and 256 kbit/s, so the effect of the bitrate should be negligible. The recordings for 40 different ethnic groups were made in 13 different years between 1954 and 1975, and some in 2005 and 2006. From listening to the recordings we could not perceive major quality differences between the recordings. More importantly, the recordings of a specific ethnic group were not made in a single year, e.g. the music of the Twa people was recorded in 1954, 1971 and 1973–1975. It is plausible that not the same equipment was used in all these years. On the other hand, in several individual years, multiple ethnic groups have been recorded, potentially with the same equipment. Thus, there does not seem to be evidence for any correlation between recognition of ethnic group and recording equipment.

On the other hand, in general, it is hardly possible to avoid that potential recording effects influence the classification results. However, exactly the same is true for Western music, where the instrumentation, voice, etc. of a specific performer and/or the mastering of a certain producer can have an effect on the classification results.

6. Alternative methods of access to music collections

Classification into pre-defined categories faces a particular issue: the definition of the categories. Although classification seems like an objective task, the definition of categories, no matter if done by experts or users of a private music collection, is subjective in its nature. As a consequence, the defined categories are overlapping—a fact that can frequently be observed particularly with Western musical genres—and there are no clear boundaries between the categories, neither for humans, nor for a machine classifier.

This problem is especially prevalent in collections of ethnic audio documents where the concept of a “genre” is

frequently inexistent. The African music database described in Section 5.2.3, for example, contains a “function” category which describes a situation where a song is played rather than a genre in the sense of Western music. Especially for an automatic classification system it is therefore more difficult to determine the “function” of a song by acoustic content than a genre, which is supposed to be, to a certain extent, distinctive by sound. Commonly, a function is also related to the lyrics of a song.

With the lack of the concept of a genre defined by similar musical and sound characteristics, the question arises of how to structure and access ethnic music collections. When the ethnic music collection is thoroughly labeled and entered into a database system it is possible to retrieve music by searching or ordering the available meta-data fields. However, even with meta-data such as function, country or people retrieval by acoustically similar groups is difficult.

Using the concept of self-organizing maps, which organize music automatically according to acoustic content, access by acoustic similarity can be provided to ethnic music which would otherwise not be possible. In the following sections we will describe the underlying principles and a software application that provides this kind of access to music collections by sound similarity.

6.1. The self-organizing map

There are numerous clustering algorithms that can be employed to organize an audio collection by sound similarity based on different acoustic features. An approach that is particularly suitable is the self-organizing map (SOM), an unsupervised neural network that provides a mapping from a high-dimensional input (feature) space to a (usually) two-dimensional output space [49].

A SOM is initialized with an appropriate number of units, arranged on a two-dimensional grid. A weight vector is attached to each unit. The input space is formed by feature vectors extracted from the music. The vectors from the (high-dimensional) input space are randomly presented to the SOM and the activation of each unit for the input vector is calculated using, e.g. the Euclidean distance between the weight vector of the unit and the input vector. Next, the weight vector of the activated unit is adapted towards the input vector. Consequently, the next time the same input signal is presented, the unit's activation will be even higher. The weight vectors of neighboring units are also modified accordingly, yet to a smaller amount. The magnitude of modification of the weight vectors is controlled by a time-decreasing learning rate and a neighborhood function.

This process is repeated for a large number of iterations, presenting each input vector multiple times to the SOM. The result of this training procedure is a topologically ordered mapping of the presented input data in the two-dimensional space. Similarities present in the input data are reflected as faithfully as possible on the map. Hence, similar sounding music is located close to each other, building clusters, while pieces with more distinct acoustic content are located farther away. Clearly

distinguishable musical styles (e.g. distinctive genres) will be reflected by cluster boundaries, otherwise the map will reflect smooth transitions among the variety of different pieces of music.

6.2. The application

Based on the SOMEjB system [36] that extended the purely analytical SOM algorithm by advanced visualizations, the PlaySOM application enhances the principle to a rich application platform that provides direct access to the underlying music database enriched by browsing, interaction and retrieval [50]. The application's main interface visualizes the self-organized music map in one of many different visualization metaphors (acoustic attributes, “Weather Charts”, “Islands of Music”, among various others). It provides a semantic zooming facility which displays different information dependent on the zooming level. The outer zooming level provides a complete overview of the music collection, with numbers indicating the quantity of audio documents mapped at each location. The default visualization, smoothed-data-histograms [51], indicate clusters of music that have coherent acoustic features. Zooming into the map, more information is shown about the individual audio titles.

A search window allows querying the map for specific titles. The benefit of organizing the music collection on a SOM is that similar sounding pieces of music can be retrieved directly by exploring the surroundings of the unit where a searched item has been retrieved from. With the same ease of clicking into the map, a playlist is created on-the-fly. Marking a rectangle selects an entire “cluster” of music that is perceived as acoustically similar, or a subset of the audio-collection matching a particular musical style. A path selection mode allows drawing trajectories through the musical “landscape” and selects all pieces belonging to units beneath that trajectory. This allows creating ad hoc music playlists with (smooth) transitions between various musical styles. These immediate selection and playback modes are particularly useful for a quick evaluation of the clustered content of such a music map. Variants of the PlaySOM application have been created for a range of mobile devices [52], a platform, which is in particular need for enhanced access methods not based on traditional genre-album-artist lists.

6.3. Experimental results

Although the SOM principle and the PlaySOM application are not based on external meta-data at all, an overlay visualization of genre or class meta-information on top of the music map is available. This form of visualization helps in analyzing the experimental results by showing class assignments as pie-charts on top of each SOM unit, using different colors to depict different classes. This will thus provide an implicit kind of evaluation of the automatic organization of a music map by acoustic similarity: the more coherent the colors are on top of

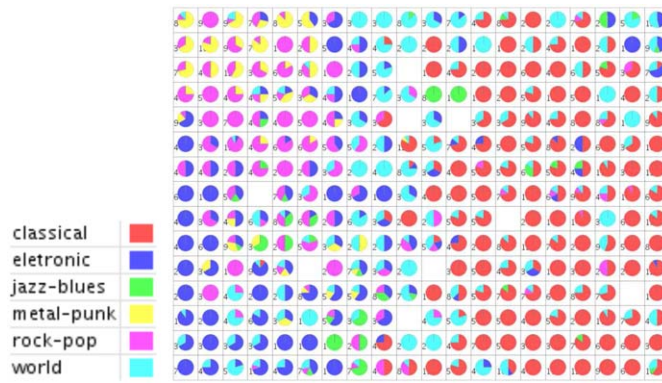


Fig. 4. Map of Western music database, visualized by genre.

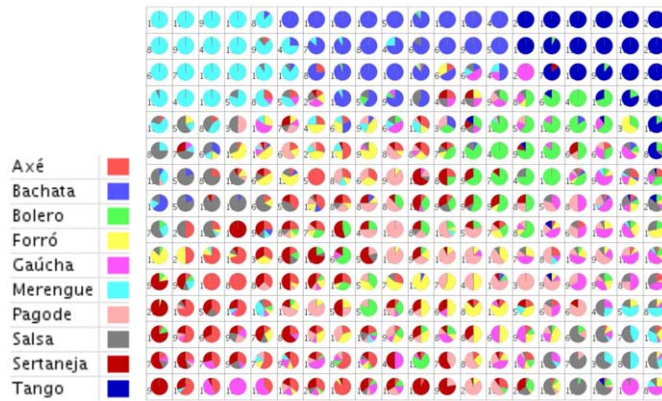


Fig. 5. Map of Latin music database, visualized by genre.

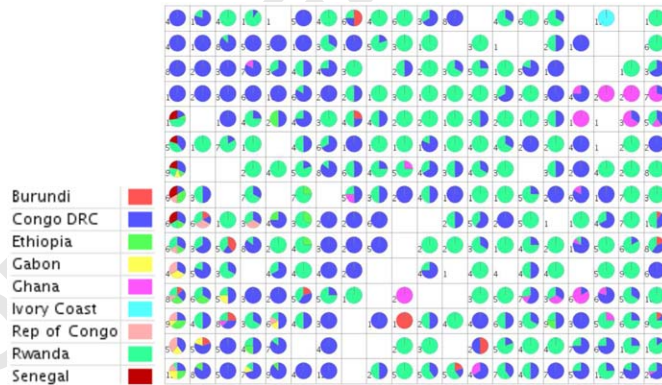


Fig. 6. Map of African music database, visualized by country.

the map, the more it agrees with manual human annotation.⁵

For each of the three music collections studied already in Section 5, a 20×15 SOM was created using SSD features (cf. Section 4.1.5) extracted from the audio as

⁵ White units with numbers represent songs with no class label available. Empty units were not populated by the SOM.

input. Fig. 4 shows the Western music collection automatically aligned by audio content on the units of a SOM. Each unit shows a pie-chart class diagram indicating in various colors the portions of each of the six classes listed in Fig. 4. The number of songs per unit is also given. The semantic classes have been separated quite well by acoustic content, with classical music concentrated on the right part of the map, the most quiet pieces located in the lower right corner. The musical opposite in terms of

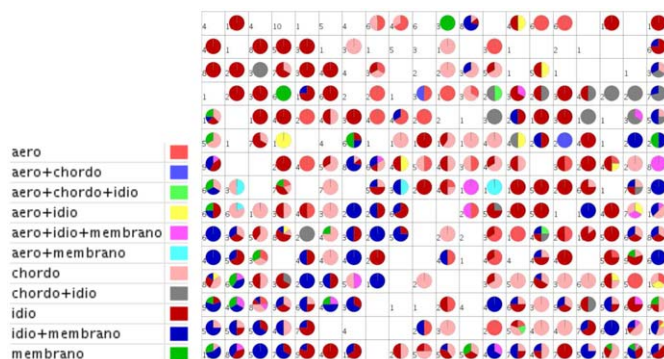


Fig. 7. Map of African music database, visualized by instrument family.

aggressiveness is found on the upper left corner: metal and punk music, followed by rock and pop beneath it and electronic music in the lower left. Both world and jazz music are located in-between classical music and the more energetic musical genres.

The SOM trained for the Latin music database (Fig. 5) was able to separate the genres even better, supposedly due to very distinct (and rather defined) characteristics in the different Latin American dances. Especially Axé, Bachata and Tango are grouped into almost pure clusters, but also the remaining genres are recognizable as cluster structures, although slightly interweaved.

We can produce multiple views for the SOM trained on the African music database, as different meta-data labels that are available. The visualization in Fig. 6 shows the arrangement by country, where we see that the music from Congo DRC and Rwanda is separated on a coarse level (also with partial interleaving), Ghana forms a small cluster, and the less represented countries are aligned on the left edge, with Senegal placed above the Republic of Congo.

Fig. 7 shows the same alignment with the view of instrument families (where pieces with multiple instruments are indicated as separate classes). Although the map seems quite unstructured at first sight, there are some clusters of idiophone instruments, or pieces with chordophone+idiophone instruments. For a better clustering by instruments, however, a dedicated instrument detector should be used as the underlying feature extractor.

Generally speaking, a SOM can give insight into the inherent structure of music depending on the features extracted, and provides multiple views on a collection of music with different visualization metaphors. Especially these multiple views and variable forms of visualization make the SOM (and the PlaySOM application) such a valuable tool for exploring ethnic music collections.

7. Conclusions

Improving means of access is essential to unlock the value that the holdings in ethnic, folk and other non-Western music collections represent. This includes tools

to assist in analyzing, structuring and comparing the audio content of archives. These may help researchers in understanding complex relationships between the various pieces and assist in research work. They may also prove an invaluable asset when it comes to managing the increasing amounts of audio being digitized. Such tools, while not primarily geared towards research use, may also enable a broader public to get in contact with the massive volumes of valuable and rich cultural heritage recordings, such as of Irish or Greek folk music, Indian classical music or ethnic African music, familiarizing a larger public with this music. Yet, while a range of technical solutions are being developed in the field of Music IR the majority of these are designed, optimized and evaluated predominantly on Western music. Considering the peculiarities of non-Western and in particular ethnic music, both in terms of musical content and recording characteristics, the generality of the techniques developed needs to be considered. We conducted an in-depth analysis of the performance of a number of state-of-the-art and novel music analysis and audio feature extraction techniques on both Western and non-Western music. Their performance was evaluated on a range of classification tasks using machine learning techniques to structure music into pre-defined categories. Results were presented for three different music collections, specifically a benchmark collection with predominantly Western music, a database of studio recordings of Latin American music, as well as archival holdings of African music. Overall, the approaches proved to work surprisingly well in all these different settings. Major performance differences can rather be related to different musical characteristics (i.e. dominance in rhythm or timbre) rather than recording settings. It has been shown that state-of-the-art Music IR methods are capable to categorize an ethnic music collection also by meta-data such as the country or ethnic group, while the function of songs, an important attribute for ethnic music—by contrast to the genre used commonly for Western music—could not be recognized accordingly. Another finding is that ethnic music seems to be less susceptible to fade-out effects and feature analysis delivers comparable results also from the beginning of a piece, reducing the effort for segmentation prior to audio feature extraction.

The second major contribution is the evaluation of a SOM-based interface to access recordings of non-Western and ethnic music. While this interface was predominantly developed as an access interface for playlist creation and the intuitive exploration of music collections, we demonstrated in this paper that it may also serve as a useful tool for more structural exploration of audio content. Relationships between various musical characteristics can be visualized and set in relation to each other, while at the same time serving as a convenient interface for the general public, who may lack the necessary in-depth knowledge to understand more traditional musicological organization schemes.

While the audio feature extraction and categorization approaches as well as structuring and access support have proven to work as well for non-Western and ethnic music recordings, the experiments suggest there is room for improvements. Similar to fine-tuning approaches specifically to Western music, dedicated modules considering musical structure in different types of non-Western recordings as well as pre-processing modules to account for different means of recording the audio seem necessary if evaluation across collections should be supported. Furthermore, adaptation of music analysis techniques to the specific characteristics of such music collections from the musicological point of view should be considered. In addition, efforts to analyze the textual content of non-Western and ethnic recordings in the vast range of languages found in such collections are essential to further close the semantic gap between the pure acoustic impression and the intention or function of specific pieces of music.

Acknowledgments

The authors would like to thank the Capes Brazilian Research Agency (process number 4871-06-5) and Mr. Breno Moiana for his invaluable help with the experiment infrastructure. We would also like to thank the reviewers for helping to improve this article through their thorough review.

References

- [1] N. Bernardini, X. Serra, M. Leman, G. Widmer, G. De Poli (Eds.), *A Roadmap for Sound and Music Computing*, The S2S2 Consortium, 2007.
- [2] G. Tzanetakis, A. Kapur, A. Schloss, M. Wright, *Computational ethnomusicology*, *Journal of Interdisciplinary Music Studies* 1 (2) (2006) 1–24.
- [3] M. Wright, A. Schloss, G. Tzanetakis, *Analyzing Afro-Cuban rhythm using rotation-aware Clave template matching with dynamic programming*, in: *Proceedings of the International Conference on Music Information Retrieval*, Philadelphia, PA, USA, 2008, pp. 647–652.
- [4] F. Gouyon, *Microtiming in “Samba de Roda”—preliminary experiments with polyphonic audio*, in: *Proceedings of the Brazilian Symposium on Computer Music*, 2007, pp. 197–203.
- [5] D. Moelants, O. Cornelis, M. Leman, J. Gansemans, R. De Caluwe, G. De Tré, T. Matthé, A. Hallez, *The problems and opportunities of content-based analysis and description of ethnic music*, *International Journal of Intangible Heritage* 2 (2007) 57–68.
- [6] L. Auber, *The Music of the Other*, Ashgate Publishing, 2007.
- [7] J.S. Downie, S.J. Cunningham, *Toward a theory of music information retrieval queries: system design implications*, in: *Proceedings of the International Conference on Music Information Retrieval*, Paris, France, 2002, pp. 299–300.
- [8] J.-J. Aucouturier, F. Pachet, *Representing musical genre: a state of the art*, *Journal of New Music Research* 32 (1) (2003) 83–93.
- [9] C. McKay, I. Fujinaga, *Musical genre classification: Is it worth pursuing and how can it be*, in: *Proceedings of the International Conference on Music Information Retrieval*, Victoria, Canada, 2006, pp. 101–106.
- [10] G. Tzanetakis, P. Cook, *Musical genre classification of audio signals*, *IEEE Transactions on Speech and Audio Processing* 10 (5) (2002) 293–302.
- [11] C. McKay, I. Fujinaga, *Automatic genre classification using large high-level musical feature sets*, in: *Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain, 2004, pp. 525–530.
- [12] R. Neumayer, A. Rauber, *Multimodal analysis of text and audio features for music information retrieval*, in: *Multimodal Processing and Interaction: Audio, Video, Text*, Springer, Berlin, Heidelberg, 2008.
- [13] J.L. Herlocker, J.A. Konstan, L.G. Terveen, J.T. Riedl, *Evaluating collaborative filtering recommender systems*, *ACM Transactions on Information Systems* 22 (1) (2004) 5–53.
- [14] X. Hu, J.S. Downie, K. West, A. Ehmann, *Mining music reviews: promising preliminary results*, in: *Proceedings of the International Conference on Music Information Retrieval*, London, UK, 2005, pp. 536–539.
- [15] M. Schedl, P. Knees, T. Pohle, G. Widmer, *Towards an automatically generated music information system via web content mining*, in: *European Conference on Information Retrieval*, Glasgow, Scotland, 2008, pp. 585–590.
- [16] P. Lamere, *Social tagging and music information retrieval*, *Journal of New Music Research* 37 (2) (2008) 101–114.
- [17] T. Lidy, A. Rauber, A. Pertusa, J.M. Inesta, *Improving genre classification by combination of audio and symbolic descriptors using a transcription system*, in: *Proceedings of the International Conference on Music Information Retrieval*, Vienna, Austria, 2007, pp. 23–27.
- [18] C. McKay, I. Fujinaga, *Combining features extracted from audio, symbolic and cultural sources*, in: *Proceedings of the International Conference on Music Information Retrieval*, Philadelphia, PA, USA, 2008, pp. 597–602.
- [19] A.R. Rudolf Mayer, Robert Neumayer, *Combination of audio and lyrics features for genre classification in digital audio collections*, in: *Proceedings of the ACM International Conference on Multimedia*, Vancouver, Canada, 2008, pp. 159–168.
- [20] I.A. Bolshakov, A. Gelbukh, *Computational linguistics: models, resources, applications*, IPN-UNAM-FCE, 2004.
- [21] C. Seeger, *An instantaneous music notator*, *Journal of the International Folk Music Council* 3 (1951) 103–106.
- [22] A. Nesbit, L. Hollenberg, A. Senyard, *Towards automatic transcription of Australian aboriginal music*, in: *Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.
- [23] A. Krishnaswamy, *Melodic atoms for transcribing carnatic music*, in: *Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.
- [24] P. Chordia, A. Rae, *Raag recognition using pitch-class and pitch-class dyad distributions*, in: *Proceedings of the International Conference on Music Information Retrieval*, Vienna, Austria, 2007, pp. 431–436.
- [25] D. Moelants, O. Cornelis, M. Leman, J. Gansemans, R. De Caluwe, G. De Tré, T. Matthé, A. Hallez, *Problems and opportunities of applying data- and audio-mining techniques to ethnic music*, in: *Proceedings of the International Conference on Music Information Retrieval*, Victoria, Canada, 2006.
- [26] B. Duggan, B. O’Shea, M. Gainza, P. Cunningham, *Machine annotation of sets of traditional Irish dance tunes*, in: *Proceedings of the International Conference on Music Information Retrieval*, Philadelphia, PA, USA, 2008, pp. 401–406.
- [27] A. Pikrakis, I. Antonopoulos, S. Theodoridis, *Music meter and tempo tracking from raw polyphonic audio*, in: *Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.
- [28] I. Antonopoulos, A. Pikrakis, S. Theodoridis, O. Cornelis, D. Moelants, M. Leman, *Music retrieval by rhythmic similarity applied on Greek and African traditional music*, in: *Proceedings of the International Conference on Music Information Retrieval*, Vienna, Austria, 2007, pp. 297–300.

- 1 [29] N.M. Norowi, S. Doraisamy, R. Wirza, Factors affecting auto- 35
 3 matic genre classification: an investigation incorporating
 non-Western musical forms, in: Proceedings of the International
 5 Conference on Music Information Retrieval, London, UK, 2005,
 pp. 13–20.
- 7 [30] S. Doraisamy, S. Golzari, N.M. Norowi, M.N.B. Sulaiman, N.I. Udzir, A
 study on feature selection and classification techniques for
 9 automatic genre classification of traditional malay music, in:
 Proceedings of the International Conference on Music Information
 11 Retrieval, Philadelphia, PA, USA, 2008, pp. 331–336.
- 13 [31] J.S. Downie, Music information retrieval, Annual Review of
 Information Science and Technology, Information Today, vol. 37,
 15 Medford, NJ, USA, 2003, pp. 295–340.
- 17 [32] F. Gouyon, S. Dixon, E. Pampalk, G. Widmer, Evaluating rhythmic
 19 descriptors for musical genre classification, in: Proceedings of the
 25th International AES Conference, London, UK, 2004.
- 21 [33] A. Rauber, E. Pampalk, D. Merkl, Using psycho-acoustic
 models and self-organizing maps to create a hierarchical structur-
 23 ing of music by musical styles, in: Proceedings of the International
 Conference on Music Information Retrieval, Paris, France, 2002,
 25 pp. 71–80.
- 27 [34] T. Lidy, A. Rauber, Evaluation of feature extractors and psycho-
 acoustic transformations for music genre classification, in: Pro-
 ceedings of the 6th International Conference on Music Information
 29 Retrieval, London, UK, 2005, pp. 34–41.
- 31 [35] F. Gouyon, P. Herrera, P. Cano, Pulse-dependent analyses of
 percussive music, in: Proceedings of the 22nd International AES
 33 Conference on Virtual, Synthetic and Entertainment Audio, Espoo,
 Finland, 2002.
- [36] A. Rauber, E. Pampalk, D. Merkl, The SOM-enhanced JukeBox:
 organization and visualization of music collections based on
 perceptual models, *Journal of New Music Research* 32 (2) (2003)
 193–210.
- [37] E. Zwicker, H. Fastl, *Psychoacoustics—Facts and Models*, Springer
 Series of Information Sciences, vol. 22, Springer, Berlin, 1999.
- [38] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer,
 New York, 1995.
- [39] J.C. Platt, *Fast Training of Support Vector Machines using Sequential
 Minimal Optimization*, MIT Press, Cambridge, MA, USA, 1999.
- [40] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers,
IEEE Transactions on Pattern Analysis and Machine Intelligence 20
 (3) (1998) 226–239.
- [41] C.N. Silla Jr., A.L. Koerich, C.A.A. Kaestner, A machine learning
 approach to automatic music genre classification, *Journal of the
 Brazilian Computer Society* 14 (3) (2008) 7–18.
- [42] C.N. Silla Jr., C.A.A. Kaestner, A.L. Koerich, Automatic music genre
 classification using ensemble of classifiers, in: Proceedings of the
 IEEE International Conference on Systems, Man and Cybernetics,
 Montreal, Canada, 2007, pp. 1687–1692.
- [43] P. Cano, E. Gómez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X.
 Serra, S. Streich, N. Wack, ISMIR 2004 audio description contest,
 Technical Report MTG-TR-2006-02, Music Technology Group,
 Pompeu Fabra University, 2006.
- [44] ISMIR 2004 Audio Description Contest, Website, URL <http://
 ismir2004.ismir.net/ISMIR_Contest.html>, 2004.
- [45] C.N. Silla Jr., A.L. Koerich, C.A.A. Kaestner, The Latin music database,
 in: Proceedings of the International Conference on Music Informa-
 tion Retrieval, Philadelphia, PA, USA, 2008, pp. 451–456.
- [46] O. Cornelis, R. De Caluwe, G. De Tré, A. Hallez, M. Leman, T. Matthé,
 D. Moelants, J. Gansemans, Digitisation of the ethnomusicological
 sound archive of the Royal Museum for Central Africa (Belgium),
 International Association of Sound and Audiovisual Archives
 Journal 26 (2005) 35–43.
- [47] T. Matthé, G. De Tré, A. Hallez, R. De Caluwe, M. Leman, O. Cornelis,
 D. Moelants, J. Gansemans, A framework for flexible querying and
 mining of musical audio archives, in: Proceedings of the 16th
 International Conference on Database and Expert Systems Applica-
 tions, 2005, pp. 1041–1045.
- [48] I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning
 Tools and Techniques*, second ed., Morgan Kaufmann, San
 Francisco, 2005.
- [49] T. Kohonen, *Self-organizing Maps*, in: Springer Series in Information
 Sciences, vol. 30, third ed., Springer, Berlin, 2001.
- [50] T. Lidy, A. Rauber, Classification and clustering of music for novel
 music access applications, cognitive technologies, in: *Machine
 Learning Techniques for Multimedia*, Springer, Berlin, Heidelberg,
 2008, pp. 249–285.
- [51] E. Pampalk, A. Rauber, D. Merkl, Using smoothed data histograms
 for cluster visualization in self-organizing maps, in: Proceedings
 of the International Conference on Neural Networks, Springer, Madrid,
 Spain, 2002, pp. 871–876.
- [52] J. Frank, T. Lidy, P. Hlavac, A. Rauber, Map-based music
 interfaces for mobile devices, in: Proceedings of the ACM Interna-
 tional Conference on Multimedia, Vancouver, Canada, 2008.