Music search and recommendation

Karlheinz Brandenburg, Christian Dittmar, Matthias Gruhne, Jakob Abeßer, Hanna Lukashevich, Peter Dunker, Daniel Gärtner, Kay Wolter, Stefanie Nowak, and Holger Grossmann

Fraunhofer IDMT, Ehrenbergstr. 31, 98693 Ilmenau, Germany
dmr@idmt.fraunhofer.de

1 Introduction

In the last ten years, our ways to listen to music have drastically changed: In earlier times, we went to record stores or had to use low bit-rate audio coding to get some music and to store it on PCs. Nowadays, millions of songs are within reach via online distributors. Some music lovers already got terabytes of music on their hard disc. Users are now no longer desperate to get music, but to select, to find the music they love. A number of technologies has been developed to address these new requirements. There are techniques to identify music and ways to search for music. Recommendation today is a hot topic as well as organizing music into playlists.

For online music shops, recommendation is a way to significantly improve accessibility of music, therefore helping them to sell music. For users, playlist generation helps them to find their favorite track in a much more convenient way than just to search for artists or title. Casual users can broaden their knowledge about music by finding new tracks in the style they are looking for. Expert users can find their way in the millions of tracks available including new releases without having to search through music magazines or listen to thousands of new tracks. Thus, automatic recommendation seems to pose a possible solution for the so called long-tail phenomenon that has been detailed in [4].

Basically, organizing music can be done using Web 2.0 methods, i.e. the knowledge of thousands of other music lovers, or using local, digital signal processing based methods. For well known mainstream music, huge amounts of user generated browsing traces, reviews, playlists and recommendations are available in different online communities. They can be analyzed through collaborative filtering methods in order to reveal relations between artists, songs and genres. For novel or niche content one obvious solution to derive such data is content based similarity search. Since the early days of Music Information Retrieval (MIR) the search for items related to a specific query song or a set of those (Query by Example) has been a consistent focus of scientific interest. Thus, a multitude of different approaches with varying degree of complexity has been proposed [111], [43], [86]. Many publications have addressed suitable modeling methods that represent the musical gist whilst keeping the description blurry enough to account for small but irrelevant differences [8]. Since the human perception of music and their similarities is subjective, context dependent and multi-dimensional, mathematical models can always only be an imprecise estimate of reality.

2 Acoustic Features for Music Modeling

In order to make a computer understand music, it must be translated into numbers. Though partly mathematical in it's original nature, the music as we hear it is a seemingly random series of different air pressure levels. When sampled and quantized, these continuous functions of time become vectors of discrete digits. These vectors still reveal almost no information about their content. Instead, further signal processing operations are necessary to derive the meaning of these data. These operations are called *feature extraction* and their result is called *feature vectors*.

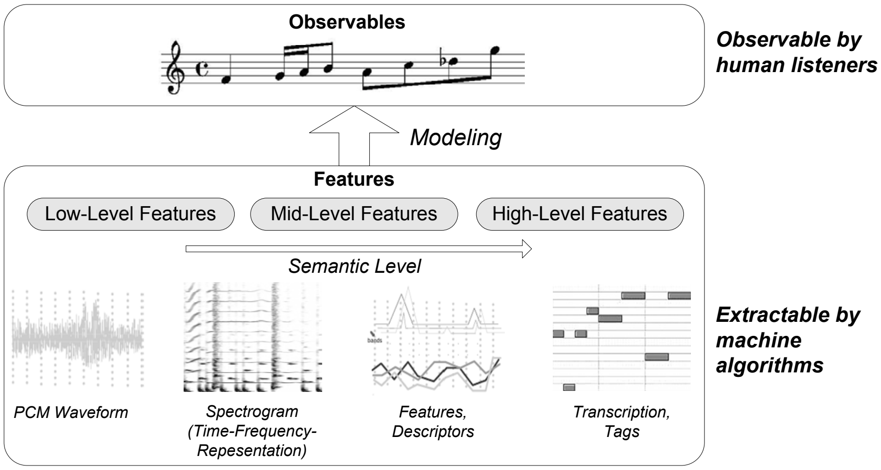

Fig. 1. Different semantic layers of acoustic features for music modeling

As depicted in Fig. 1, acoustic feature vectors can be roughly distinguished into three categories based on their degree of semantic meaning. Low-level features can be extracted from the audio signal via few basic signal processing operations. Mid-level features intend to bridge the gap between low-level features and a full music annotation and transcription. They present an intermediate semantic layer that combines signal processing techniques with a-priori musical knowledge. High-level features carry a high degree of semantic information, since a human listener is able to understand their meaning. They bear a close relation to the musicological vocabulary and can be used as features for special MIR tasks. Prominent examples of these three categories are described in the following subsections.

2.1 Low-level audio features

Low-level acoustic features are directly extracted from the digitized audio signal. Therefore, the signal is divided into adjacent or overlapping frames with a typical size of 10 to 100 milliseconds. Each of the frames is processed independently and the feature extraction results in a feature vector with the length (or dimension) N. Numerous different proposals for computing the various feature vectors have been described in literature. The work of [115], [122] and [79] provide a good overview on low-level features commonly used in MIR-applications. It should be stressed that already slight implementation differences in the signal processing chain for low-level feature extraction can result into huge numerical differences in the feature vectors. Thus, the MPEG-7 standard [53] describes a number of standardized low-level features and an interface that is open to the public. Since it is impossible to detail the vast variety of features that has been introduced during the last years, the focus will be on features that have been most commonly and successfully been used in MIR systems.

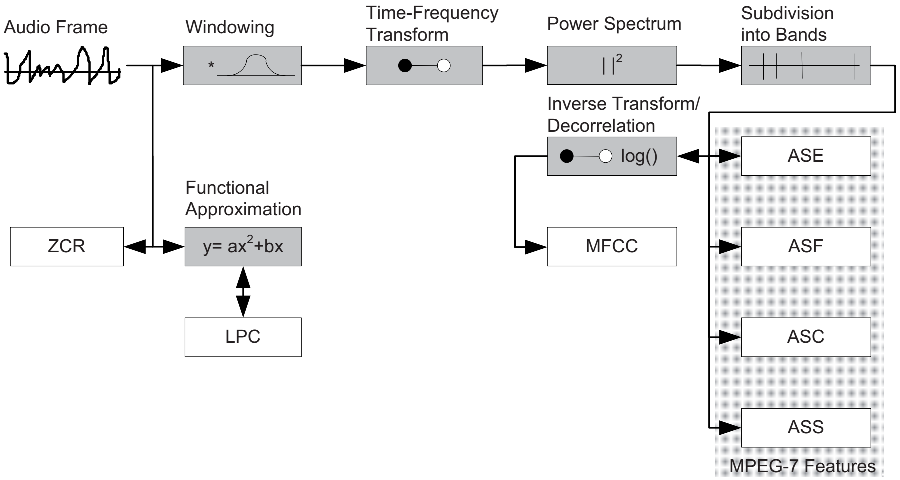

Fig. 2. Processing blocks for the acoustic low-level feature extraction.

Fig. 2 shows the common processing blocks for the acoustic low-level feature extraction. The starting point of this processing chain is always an audio frame. For the different low-level features, one or more of these blocks are applied. The following paragraphs describe each of these blocks in detail:

- *Windowing:* If a time-frequency transformation is applied to the single audio frames, the discontinuities at the edges of each frame lead to erroneous high frequencies. Therefore, a windowing function is usually multiplied with the audio frame before the transformation. Several different windowing functions are used in signal processing, the most common ones are Hamming and Hann windows.

- *Time-Frequency Transformation:* In order to transfer the signal from the time domain into the frequency domain, usually the Discrete Fourier Transform (DFT), respectively the more efficient Fast Fourier Transform (FFT), is applied. Since the human perception of the frequency scale is nonlinear, methods like Constant-Q, Multi-Resolution FFT or Warped FFT are often utilized.
- *Squaring:* As most of the features are based on the spectrum magnitude while the output of the DFT is a complex-valued vector, all elements of vector are squared. The resulting vector is called the *power spectrum*.
- *Subspacing:* In general, the number of coefficients, respectively dimensions N should be kept as low as possible, in order to guarantee for efficient search and retrieval later on. Therefore, the raw spectra are commonly subdivided into adjacent or overlapping frequency bands and a single value is calculated per band depending on the underlying feature extraction algorithm. Commonly, a logarithmic frequency spacing is used. Mel scale is another option.
- *Decorrelation:* The extraction procedure for some features incorporates a decorrelation in order to reduce the number of coefficients in the feature vector. This is often realized by a Karhunen Loéve (KL)-Transform (see Sec. 3.1), which results in a linear combination of the original feature vector elements, sorted by decreasing significance. Thus, the number of elements in the feature vector can be limited, according to the requirements of the system. Logan [70] compared the usage of DCT instead of the KL Transform in the area of music information retrieval and came to the conclusion, that a DCT is adequate for decorrelation.
- *Functional Approximation:* Some features are based on a functional approximation of the original signal. The resulting functional parameters can either be used for parametric coding or as a feature vector. An efficient implementation method is given by Levinson-Durbin algorithm.

The following itemization describes the most common low-level features. Their necessary extraction blocks are depicted in Fig. 1 and their specific purpose is described below:

Mel-Frequency Cepstral Coefficients

The *Mel-Frequency Cepstral Coefficients* (MFCC) low-level feature is one of the most commonly used acoustic low-level feature in the area of speech recognition, where it has already been used almost 30 years ago [51] and later on in MIR [70], [8], [109]. Although speech signals are quite different to audio signals, [70] showed that MFCC are appropriate for music classification, and numerous other publications have proven that. The most noticeable blocks in the extraction process are the logarithm and the inverse DFT after the summation of the frequency bands. This process transfers any multiplication into additions and decorrelates the audio signal in order to reduce the dimensions (see Sec. 3.1) and to facilitate statistical modeling later on (see Sec. 3.2).

Audio Spectrum Envelope

The *Audio Spectrum Envelope* (ASE) low-level feature is a very basic feature that has been applied, e.g., in [13] for audio classification. In order to allow interoperability, ASE is described in the MPEG-7 standard (ISO/IEC 15938-4) [53]. The necessary extraction blocks are described in Fig. 2. The final step is the summation of the subspaced power-spectrum magnitudes inside the logarithmically spaced frequency bands. The disadvantage of this feature is it's dependence on the global level of the signal. If the feature is extracted from exactly the same song with two different amplification factors, the resulting feature vectors will differ significantly. Thus, this feature is often post-processed in MIR systems.

Audio Spectrum Flatness

The feature *Audio Spectrum Flatness* (ASF) is also a standardized low-level feature in the MPEG-7 audio standard [53]. It has originally been used for calculating the noise masking threshold in audio coding [55] and later for several MIR tasks, such as audio identification [3] and audio similarity search [2]. For each subspaced power-spectrum, the ASF factor is estimated by dividing the geometric mean by the arithmetic mean of the corresponding frequency band. This factor indicates the tonality of the signal. The factor ranges from 0 to 1. A value of ”1” indicates, that the spectrum is completely noise-like and a value of ”0” indicates, that the energy in the spectrum is concentrated in narrow-band peaks.

Linear Predictive Coding

One of the fundamentals for digital speech communication is the compression of the speech signal in order to reduce the amount of transmitted data and hence, to save bandwidth. The digital speech compression often bases on the principle of *Linear Predictive Coding* (LPC). This principle optimizes a set of filter coefficients in a manner, that the residual error is minimized when applied on the speech signal. For speech coding, only the filter coefficients and the quantized residual error signal are transmitted from the sender to the receiver, where the original signal is reconstructed by a synthesis filter. These coefficients turned out to be suitable as low-level features for automatic speech recognition. LPC coefficients have also been successfully used in MIR especially when dealing with vocals in music. Kim and Whitman [56] described an algorithm, where conventional LPC have been utilized as features for identifying the performing artist of a popular song. Furthermore, Shao et al. published a paper [100] on automatic music genre classification based on LPC.

Zero Crossing Rate

The *Zero Crossing Rate* (ZCR) simply counts the number of changes of the signum in audio frames. Since the number of crossings depends on the size of the examined window, the final value has to be normalized by dividing by the actual window size. One of the first evaluations of the ZCR in the area of speech recognition have been described by Licklider and Pollack in 1948 [63]. They described the feature extraction process and resulted with the conclusion, that the ZCR is useful for digital speech signal processing because it is loudness invariant and speaker independent. Among the variety of publications using the ZCR for MIR are the fundamental genre identification paper from Tzanetakis et al. [111] and a paper dedicated to the classification of percussive sounds by Gouyon [39].

Audio Spectrum Centroid

The *Audio Spectrum Centroid* (ASC) is another MPEG-7 standardized low-level feature in MIR [89]. As depicted in [53], it describes the center of gravity of the spectrum. It is used to describe the timbre of an audio signal. For each subspaced power-spectrum, the ASC measure is estimated by computing the spectral center of gravity (1st statistical moment) inside the corresponding frequency band.

Audio Spectrum Spread

Audio Spectrum Spread (ASS) is another feature described in the MPEG-7 standard. It is a descriptor of the shape of the power spectrum that indicates whether it is concentrated in the vicinity of its centroid, or else spread out over the spectrum. For each subspaced power-spectrum, the ASS measure is estimated by computing the spectral spread around the centroid (2nd statistical moment) inside the corresponding frequency band, as described in [53]. The spectrum spread allows a good differentiation between tone-like and noise-like sounds.

2.2 Mid-level audio features

Mid-level features ([11]) present an intermediate semantic layer between well-established low-level features and advanced high-level information that can be directly understood by a human individual. Basically, mid-level features can be computed by combining advanced signal processing techniques with a-priori musical knowledge while omitting the error-prone step of deriving final statements about semantics of the musical content. It is reasonable to either compute mid-level features on the entire length of previously identified coherent segments (see Sec. 3.2) or in dedicated mid-level windows that virtually sub-sample the original slope of the low-level features and squeeze their most important properties into a small set of numbers. For example, a window-size of of approximately 5 seconds could be used in conjunction with an overlap of 2.5 seconds. These numbers may seem somewhat arbitrarily chosen, but they should be interpreted as the most suitable region of interest for capturing the temporal structure of low-level descriptors in a wide variety of musical signals, ranging from slow atmospheric pieces to up-tempo Rock music.

Rhythmic mid-level features

An important aspect of contemporary music is constituted by its rhythmic content. The sensation of rhythm is a complex phenomenon of the human perception which is illustrated by the large corpus of objective and subjective musical terms, such as tempo, beat, bar or shuffle used to describe rhythmic gist. The underlying principles to understanding rhythm in all its peculiarities are even more diverse. Nevertheless, it can be assumed, that the degree of self-similarity respectively periodicity inherent to the music signal contains valuable information to describe the rhythmic quality of a music piece. The extensive prior work on automatic rhythm analysis can (according to [112]) be distinguished into *Note Onset Detection*, *Beat Tracking and Tempo Estimation*, *Rhythmic Intensity and Complexity* and *Drum Transcription*. A fundamental approach for rhythm analysis in MIR is onset detection, i.e. detection of those time points in a musical signal which exhibit a percussive or transient event indicating the beginning of a new note or sound [22]. Active research has been going on over the last years in the field of beat and tempo induction [38], [97], where a variety of methods emerged that aim at intelligently estimating the perceptual tempo from measurable periodicities. All previously described areas result more or less into a set of high-level attributes. These attributes are not always suited as features in music retrieval and recommendation scenarios. Thus, a variety of different methods for extraction of rhythmic mid-level features is described either frame-wise [99], event-wise[12] or beat-wise [37]. One important aspect of rhythm are rhythmic patterns, which can be effectively captured by means of an auto-correlation function (ACF). In [111], this is exploited by auto-correlating and accumulating a number of successive bands derived from a Wavelet transform of the music signal. An alternative method is given in [19]. A weighted sum of the ASE-feature constitutes a so called *detection function* and is auto-correlated. The challenge is to find suitable distance measures or features, that can further abstract from the raw ACF-functions, since they are not invariant to tempo changes.

Harmonic mid-level features

It can safely be assumed that the melodic and harmonic structures in music are a very important and intuitive concept to the majority of human listeners. Even non-musicians are able to spot differences and similarities of two given tunes. Several authors have addressed chroma vectors, also referred to as harmonic pitch class profiles [42] as a suitable tool for describing the harmonic and melodic content of music pieces. In general, chroma vectors can be obtained by accumulating all spectral energy that belongs to a certain note of the chromatic scale over all octaves into a single coefficient. This octave agnostic representation of note probabilities can be used for estimation of the musical key, chord structure detection [42] and harmonic complexity measurements. Chroma vectors are somewhat difficult to categorize, since the techniques for extraction are typical low-level operations. But the fact that they already take into account the 12-tone scale of western tonal music places them half-way between low-level and mid-level. Very sophisticated post-processing can be performed on the raw chroma-vectors. One area of interest is the detection and alignment of cover-songs respectively classical pieces performed by different conductors and orchestras. Recent approaches are described in [98] and [83], both works are dedicated to matching and retrieval of songs that are not necessarily identical in terms of the progression of their harmonic content.

A straightforward approach to use chroma features is the computation of different histograms of the most probable notes, intervals and chords that occur throughout a song ([19]). Such simple post-processing already reveals a lot of information contained in the songs. As an illustration, Fig. 3 shows the comparison of chroma-based histograms between the well known song ”I will survive” by ”Gloria Gaynor” and three different renditions of the same piece by the artists ”Cake”, ”Nils Landgren” and ”Hermes House Band” respectively. The shades of gray in the background indicate the areas of the distinct histograms. Some interesting phenomena can be observed when examining the different types of histograms. First, it can be seen from the chord histogram (right-most), that all four songs are played in the same key. The interval histograms (2nd and 3rd from the left) are most similar between the first and the last song, because the last version stays comparatively close to the original. The second and the third song are somewhat sloppy and free interpretations of the original piece. Therefore, their interval statistics are more akin.

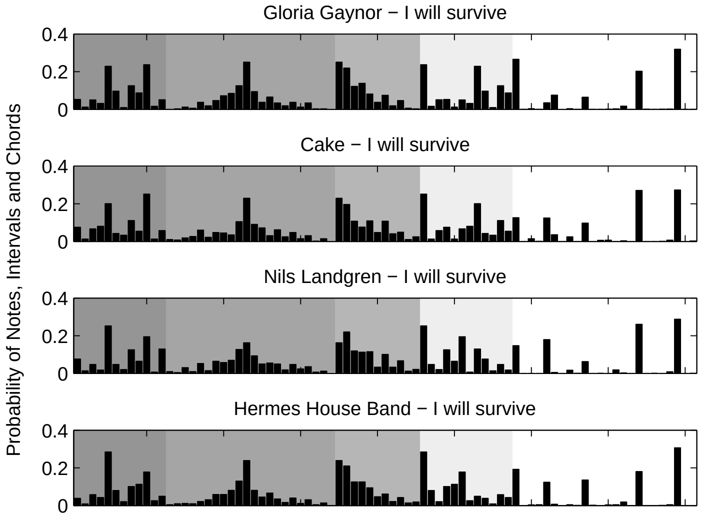

Fig. 3. Comparison of chroma-based histograms between cover songs. From left to right: note, interval, absolute interval, shifted note, and chord histograms.

2.3 High-level music features

High-level features represent a wide range of musical characteristics, bearing a close relation to musicological vocabulary. Their main design purpose is the development of computable features being capable to model the music parameters that are observable by musicologists (see Fig. 1) and that do not require any prior knowledge about signal-processing methods. Some high-level features are abstracted from features on a lower semantic level by applying various statistical pattern recognition methods (see Sec. 4). In contrast, transcription-based high-level features are directly extracted from score parameters like onset, duration and pitch of the notes within a song, whose precise extraction itself is a crucial task within MIR. Many different algorithms for drum [121], [21], bass [93], [40], melody [33], [90] and harmony [42] transcription have been proposed in the literature, achieving imperfect but remarkable detection rates so far. Recently, the combination of transcription methods for different instrument domains has been reported in [20] and [94]. However, modeling the ability of musically skilled people to accurately recognize, segregate and transcribe single instruments within dense polyphonic mixtures still bears a big challenge.

In general, high-level features can be categorized according to different musical domains like rhythm, harmony, melody or instrumentation. Different approaches for the extraction of rhythm-related high-level features have been reported. For instance, they were derived from genre-specific temporal note deviations [36] (the so-called *swing ratio*), from the percussion-related instrumentation of a song [44] or from various statistical spectrum descriptors based on periodic rhythm patters [64]. Properties related to the notes of single instrument tracks like the dominant grid (e.g. 32th notes), the dominant feeling (down- or offbeat), the dominant characteristic (binary or ternary) as well as a measure of syncopation related to different rhythmical grids can be deduced from the *Rhythmical Structure Profile* ([1]). It provides a temporal representation of all notes that is invariant to tempo and the bar measure of a song. In general, a well-performing estimation of the temporal positions of the beat-grid points is a vital pre-processing step for a subsequent mapping of the transcribed notes onto the rhythmic bar structure of a song and thereby for a proper calculation of the related features.

Melodic and harmonic high-level features are commonly deduced from the progression of pitches and their corresponding intervals within an instrument track. Basic statistical attributes like mean, standard deviation, entropy as well as complexity-based descriptors are therefore applied ([25], [78], [74] and [64]).

Retrieval of rhythmic and melodic repetitions is usually achieved by utilizing algorithms to detect repeating patterns within character strings [49]. Subsequently, each pattern can be characterized by its length, incidence rate and mean temporal distance ([1]). These properties allow the computation of the pattern's relevance as a measure for the recall value to the listener by means of derived statistical descriptors. The instrumentation of a song represents another main musical characteristic which immediately affects the timbre of a song ([78]). Hence, corresponding high-level features can be derived from it.

With all these high-level features providing a big amount of musical information, different classification tasks have been described in the literature concerning metadata like the genre of a song or its artist. Most commonly, genre classification is based on low- and mid-level features (see Sec. 4). Only a few publications have so far addressed this problem solely based on high-level features. Examples are [78], [59] and [1], hybrid approaches are presented in [64]. Apart from different classification methods (see Sec. 4.1), some major differences are the applied genre taxonomies as well as the overall number of genres.

Further tasks that have been reported to be feasible with the use of high-level features are artist classification ([26], [1]) and expressive performance analysis ([77], [95]). Nowadays, songs are mostly created by a blending of various musical styles and genres. Referring to a proper genre classification, music has to be seen and evaluated segment-wise. Furthermore, the results of an automatic song segmentation can be the source of additional high-level features characterizing repetitions and the overall structure of a song.

3 Statistical Modeling and Similarity Measures

Nearly all state-of-the-art MIR systems use low-level acoustic features calculated in short time frames as described in Sec. 2.1. Using these raw features results in a $K \times N$ dimension feature matrix $\mathbf{X}$ per song, where K is the number of the time frames in the song, and N is the number of feature dimensions. Dealing with this amount of raw data is computationally very inefficient. Additionally, the different elements of the feature vectors could appear strongly correlated and cause information redundancy.

3.1 Dimension Reduction

One of the usual ways to suppress redundant information in the feature matrix is utilization of dimension reduction techniques. Their purpose is to decrease the feature dimension N while keeping or even revealing the most characteristic data properties. Generally, all dimension reduction methods can be divided into supervised and unsupervised ones. Among the unsupervised approaches the one most often used is *Principal Component Analysis* (PCA). The other well-established unsupervised dimension reduction method is *Self-Organizing Maps* (SOM), which is often used for visualizing the original high-dimensional feature space by mapping it into a two dimensional plane. The most often used supervised dimension reduction method is *Linear Discriminant Analysis* (LDA), it is successfully applied as a pre-processing for audio signal classification.

Principal Component Analysis

The key idea of PCA [31] is to find a subspace whose basis vectors correspond to the maximum-variance directions in the original feature space. PCA involves an expansion of the feature matrix into the eigenvectors and eigenvalues of its covariance matrix, this procedure is called the *Karhunen Loéve expansion*. If $\mathbf{X}$ is the original feature matrix, then the solution is obtained by solving the eigensystem decomposition $\lambda_i v_i = \mathbf{C} v_i$, where $\mathbf{C}$ is a covariance matrix of $\mathbf{X}$, and λ_i and v_i are the eigenvalues and eigenvectors of $\mathbf{C}$. The column vectors v_i form the PCA transformation matrix $\mathbf{W}$. The mapping of original feature matrix into new feature space is obtained by the matrix multiplication $\mathbf{Y} = \mathbf{X} \cdot \mathbf{W}$. The amount of information of each feature dimension (in the new feature space) is determined by the corresponding eigenvalue. The larger the eigenvalue the more effective the feature dimension. Dimension reduction is obtained by simply discarding the column vectors v_i with small eigenvalues λ_i.

Self-Organizing Maps

SOM are special types of artificial neural networks that can be used to generate a low-dimensional, discrete representation of a high-dimensional input feature space by means of unsupervised clustering. SOM differ from conventional artificial neural networks because they use a neighborhood function to preserve the topological properties of the input space. This makes SOM very useful for creating low-dimensional views of high-dimensional data, akin to *multidimensional scaling* (MDS). Like most artificial neural networks, SOM need training using input examples. This process can be viewed as vector quantization. As will be detailed later (see Sec. 6.2), SOM are suitable for displaying music collections. If the size of the maps (the number of neurons) is small compared to the number of items in the feature space, then the process essentially equals k-means clustering. For the emergence of higher level structure, a larger so-called *Emergent SOM* (ESOM) is needed. With larger maps a single neuron does not represent a cluster anymore. It is rather an element in a highly detailed non-linear projection of the high dimensional feature space to the low dimensional map space. Thus, clusters are formed by connected regions of neurons with similar properties.

Linear Discriminant Analysis

LDA [114] is a widely used method to improve the separability among classes while reducing the feature dimension. This linear transformation maximizes the ratio of between-class variance to the within-class variance guaranteeing a maximal separability. The resultant $N \times N$ matrix $\mathbf{T}$ is used to map an N-dimensional feature row vector $\mathbf{x}$ into the subspace $\mathbf{y}$ by a multiplication. Reducing the dimension of the transformed feature vector $\mathbf{y}$ from N to D is achieved by considering only the first D column vectors of $\mathbf{T}$ (now $N \times D$) for multiplication.

3.2 Statistical models of the song

Defining a similarity measure between two music signals which consist of multiple feature frames still remains a challenging task. The feature matrices of different songs can be hardly compared directly. One of the first works on music similarity analysis [30] used MFCC as a feature, and then applied a supervised tree-structured quantization to map the feature matrices of every song to the histograms. Logan and Salomon [71] used a song signature based on histograms derived by unsupervised k-means clustering of low-level features. Thus, the specific song characteristics in the compressed form can be derived by clustering or quantization in the feature space. An alternative approach is to treat each frame (row) of the feature matrix as a point in the N-dimensional feature space. The characteristic attributes of a particular song can be encapsulated by the estimation of the *Probability Density Function* (PDF) of these points in the feature space. The distribution of these points is a-priori unknown, thus the modeling of the PDF has to be flexible and adjustable to different levels of generalization. The resulting distribution of the feature frames is often influenced by the various underlying random processes. According to the central limit theorem, the vast class of acoustic features tends to be normally distributed. The constellation of these factors leads to the fact, that already in the early years of MIR the *Gaussian Mixture Model* (GMM) became the commonly used statistical model for representing a feature matrix of a song [69], [6]. Feature frames are thought of as generated from various sources and each source is modeled by a single Gaussian. The PDF $p(\mathbf{x} \mid \lambda)$ of the feature frames is estimated as a weighted sum of the multivariate normal distributions:

$$p(\mathbf{x} \mid \lambda) = \sum_{i=1}^{M} \omega_i \, \frac{1}{(2\pi)^{N/2} \, |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right) \tag{1}$$

The generalization properties of the model can be adjusted by choosing the number of Gaussian mixtures M. Each single i-th mixture is characterized by its mean vector μ_i and covariance matrix Σ_i. Thus, a GMM is parametrized in $\lambda = \{\omega_i, \mu_i, \Sigma_i\}$, $i = \overline{1,M}$, where ω_i is the weight of the i-th mixtures and $\sum_i \omega_i = 1$. A schematic representation of a GMM is shown in Fig. 4. The parameters of the GMM can be estimated using the Expectation-Maximization algorithm [18]. A good overview of applying various statistical models (ex. GMM or k-means) for music similarity search is given in [7].

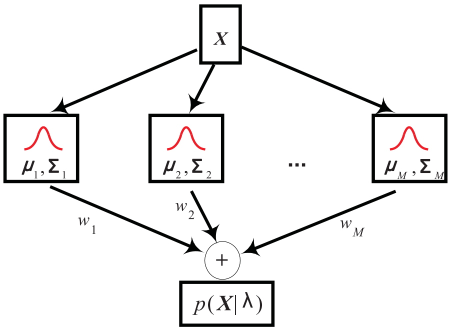

Fig. 4. Schematic representation of Gaussian Mixture Model

The approach of modeling all frames of a song with a GMM is often referred as a ”bag-of-frames” approach [5]. It encompasses the overall distribution, but the long-term structure and correlation between single frames within a song is not taken into account. As a result, important information is lost. To overcome this issue, Tzanetakis [110] proposed a set of audio features capturing the changes in the music “texture”. For details on mid-level and high-level audio features the reader is referred to the Sec. 2.

Alternative ways to express the temporal changes in the PDF are proposed in [28]. They compared the effectiveness of GMM to *Gaussian Observation Hidden Markov Models* (HMM). The results of the experiment showed that HMM better describe the spectral similarity of songs than the standard technique of GMM. The drawback of this approach is a necessity to calculate the similarity measure via log-likelihood of the models (see also Sec. 3.3).

Recently, another approach using semantic information about song segmentation for song modeling has been proposed in [73]. Song segmentation implies a time-domain segmentation and clustering of the musical piece in possibly repeatable semantically meaningful segments. For example, the typical western pop song can be segmented into “intro”, “verse”, “chorus”, “bridge”, and “outro” parts. For similar songs not all segments might be similar. For the human perception, the songs with similar “chorus” are similar. In [73], application of a song segmentation algorithm based on the Bayesian Information Criterion (BIC) has been described. BIC has been successfully applied for speaker segmentation [82]. Each segment state (ex. all repeated “chorus” segments form one segment state) are modeled with one Gaussian. Thus, these Gaussians can been weighted in a mixture depending on the durations of the segment states. Frequently repeated and long segments achieve higher weights.

3.3 Distance Measures

The particular distance measure between two songs is calculated as a distance between two song models and therefore depends on the models used. In [30] the distance between histograms was calculated via *Euclidean distance* or *Cosine distance* between two vectors. Logan and Salomon [71] adopted the *Earth mover’s distance* (EMD) to calculate the distance between k-means clustering models.

The straight forward approach to estimate the distance between the song modeled by GMM or HMM is to rate the log-likelihood of feature frames of one song by the models of the others. Distance measures based on log-likelihoods have been successfully used in [6] and [28]. The disadvantage of this method is an overwhelming computational effort as well as storage requirements. The system does not scale well and is hardly usable in real-world applications dealing with huge music archives. Some details to its computation times can be found in [86].

If a song is modeled by parametric statistical models, such as GMM, a more appropriate distance measure between the models can be defined based on the parameters of the models. A good example of such parametric distance measure is a *Kullback-Leibler divergence* (KL-divergence) [58], corresponding to a distance between two single Gaussians:

$$D(f\|g) = \frac{1}{2}\left(log\frac{|\Sigma_g|}{|\Sigma_f|} + Tr\left[\Sigma_g^{-1}\Sigma_f\right] + \left(\mu_f - \mu_g\right)^T \Sigma_g^{-1}\left(\mu_f - \mu_g\right) - N\right) \tag{2}$$

where f and g are single Gaussians with the means μ_f and μ_g and covariance matrices Σ_f and Σ_g correspondingly, and N is the dimensionality of the feature space. Initially, KL-divergence is not symmetric and needs to be symmetrized

$$D_2(f_a\|g_b) = \frac{1}{2}\left[D(f_a\|g_b) + D(g_b\|f_a)\right]. \tag{3}$$

Unfortunately, the KL-divergence for two GMM is not analytically tractable. Parametric distance measures between two GMM can be expressed by several approximations, see [73] for an overview and comparison.

4 “In the Mood” - Towards Capturing Music Semantics

Automatic semantic tagging comprises methods for automatically deriving meaningful and human understandable information from the combination of signal processing and machine learning methods. Semantic information could be a description of the musical style, performing instruments or the singer’s gender. There are different approaches to generate semantic annotations. Knowledge based approaches focus on highly specific algorithms which implement a concrete knowledge about a specific musical property (see Sec. 2.3). In contrast, supervised machine learning approaches use a large amount of audio features (see Sec. 2) from representative training examples in order to implicitly learn the characteristics of concrete categories. Once trained, the model for a semantic category can be used to classify and thus to annotate unknown music content.

4.1 Classification Models

There are two general classification approaches, a generative and a discriminative one. Both allow to classify unlabeled music data into different semantic categories with a certain probability, that depends on the training parameters and the underlying audio features. Generative probabilistic models describe how likely a song belongs to a certain pre-defined class of songs. These models form a probability distribution over the classes’ features. In contrast, discriminative models try to predict the most likely class directly instead of modeling the class’ conditional probability densities. Therefore, the model learns boundaries between different classes during the training process and uses the distance to the boundaries as an indicator for the most probable class. Only two classifiers that are most often used in MIR will be detailed here, interested readers are referred to [80] for an extensive overview of alternative machine learning techniques.

Classification based on Gaussian Mixture Models

Apart from song modeling described in Sec. 3.2, GMM are successfully used for probabilistic classification because they are well suited to model large amounts of training data per class. One interprets the single feature vectors of a music item as random samples generated by a mixture of multivariate Gaussian sources. The actual classification is conducted by estimating which pre-trained mixture of Gaussians has most likely generated the frames. Thereby, the likelihood estimate serves as some kind of confidence measure for the classification.

Classification based on Support Vector Machines

A support vector machine (SVM) attempts to generate an optimal decision margin between feature vectors of the training classes in an N-dimensional space ([15]). Therefore, only a part of the training samples is taken into account called *support vectors*. A hyperplane is placed in the feature space in a manner that the distance to the support vectors is maximized. SVM have the ability to well generalize data actually in the case of few training samples. Since the quality of SVM training is depending on a set of parameters, it is common to perform a cross validation and grid search in order to optimize them ([48]). This can be a very time-consuming process, depending on the number of training samples.

In most cases classification problems are not linearly separable in the actual feature space. Transformed into a high-dimensional space, non-linear classification problems can become linearly separable. However, higher dimensionality leads to an increase of the computational effort. To overcome this problem, the so called *kernel trick* is used to make non-linear problems separable, although the computation can be performed in the original feature space ([15]). The key idea of the kernel trick is to replace the dot product in a high-dimensional space with a kernel function in the original feature space.

4.2 Mood Semantics

Mood as an illustrative example for semantic properties describes a more subjective information which correlates not only to the music impression but also to individual memories and different music preferences. Furthermore, a distinction between mood and emotion is necessary. Emotion describes an affective perception in a short time frame, whereas mood describes a deeper perception and feeling. In the MIR community sometimes both terms are used for the same meaning. In this article the term mood is used to describe the human oriented perception of music expression.

To overcome the subjective impact, generative descriptions of mood are needed to describe the commonality of different user’s perception. Therefore, mood characteristics are formalized in mood models which describe different peculiarities of the property “mood”.

Mood Models

Mood models can be categorized into category-based and dimension-based descriptions. Furthermore, combinations of both descriptions are defined to combine the advantages of both approaches. The early work on music expression concentrates on category based formalization e.g., Hevner's adjective circle [45] as depicted in Fig. 5(a). Eight groups of adjectives are formulated whereas each group describes a category or cluster of mood. All groups are arranged on a circle and neighbored groups are consisting of related expressions. The variety of adjectives in each group gives a better representation of the meaning of each group and depicts the different user perceptions. Category based approaches allow the assignment of music items into one or multiple groups which results in a single- or multi-label classification problem.

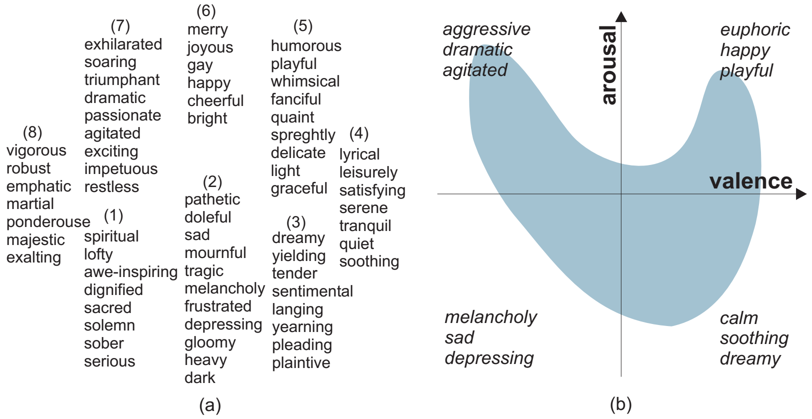

Fig. 5. Category and Dimension based Mood Models based on [45]

The dimension based mood models focus on the description of mood as a point within a multi-dimensional mood space. Different models based on dimensions such as valence, arousal, stress, energy or sleepiness are defined. Thayers model [104] describes mood as a product of the dimensions energy and tension. Russels circumplex model [92] arrange the dimensions pleasantness, excitement, activation and distress in a mood space with 45° dimension steps. As base of its model, Russel defines the dimensions pleasantness and activation. The commonality of different theories on dimension based mood descriptions is the base on moods between positive and negative (valence) and intensity (arousal) as depicted in Fig. 5(b). The labeled area in Fig. 5(b) shows the affect area which was evaluated in psychophysiological experiments as the region that equates a human emotion [41]. Mood models that combine categories and dimensions, typically place mood adjectives in a region of the mood space, e.g. the Tellegen-Watson-Clark model [103]. In [23] the valence and arousal model is extend with mood adjectives for each quadrant, to give a textual annotation and dimensional assignment of music items.

Mood Classification

Scientific publications on mood classification use different acoustic features to model different mood aspects, e.g. timbre based features for valence and tempo and rhythmic features for high activation.

Feng et al. [27] utilize an average silence ratio, whereas Yang et al. [118] use a beats per minute value for the tempo description. Lu et al. [72] incorporate various rhythmic features such as rhythm strength, average correlation peak, average tempo and average onset frequency. Beyond others Li [62] and Tolos [106] use frequency spectrum based features (e.g. MFCC, ASC, spectral flux or spectral roll-off) to describe the timbre and therewith the valence aspect of music expression. Furthermore, Wu and Jeng [117] setup a complex mixture of a wide range of acoustical features for valence and arousal expression: rhythmic content, pitch content, power spectrum centroid, inter-channel cross correlation, tonality, spectral contrast and Daubechies wavelet coefficient histograms.

Next to the feature extraction process the introduced machine learning algorithms GMM and SVM are often utilized to train and classify music expression. Examples for GMM based classification approaches are Lu [72] and Liu [68]. Publications that focus on the discriminative SVM approach are [62, 113, 61, 118]. In [23] GMM and SVM classifiers are compared with a slightly better result of the SVM approach. Liu et al. [67] utilize a nearest-mean classifier. Trohidis et al. [108] compare different multi-label classification approaches based on SVM and k-nearest neighbor.

One major problem of the comparison of different results for mood and other semantic annotations is the lack on a golden standard for test data and evaluation method. Most publications use an individual test set or ground-truth. A specialty of Wu and Jeng's approach [117] is based on the use of mood histograms in the ground truth and the results being compared by a quadratic-cross-similarity, which leads to a complete different evaluation method then a single label annotation.

A first international comparison of mood classification algorithms was performed on the MIREX 2007 in the Audio Music Mood Classification Task. Hu et al.[50] presented the results and lessons learned from the first benchmark. Five mood clusters of music were defined as ground truth with a single label approach. The best algorithm reaches an average accuracy of about 61 %.

5 Music Recommendation

There are several sources to find new music. Record sales are summarized in music charts, the local record dealers are always informed about new releases, and radio stations keep playing music all day long (and might once in a while focus on a certain style of music which is of interest for somebody). Furthermore, everybody knows friends who share the same musical taste. These are some of the typical ways how people acquire recommendations about new music. Recommendation is recommending items (e.g., songs) to users. How is this performed or (at least) assisted by computing power?

There are different types of music related recommendations, and all of them use some kind of similarity. People that are searching for albums might profit from artist recommendations (artists who are similar to those these people like). In song recommendation the system is supposed to suggest new songs. Playlist generation is some kind of song recommendation on the local database. Nowadays, in times of the ”social web”, neighbor recommendation is another important issue, in which the system proposes other users of a social web platform to the querying person - users with a similar taste of music.

Automated systems follow different strategies to find similar items [14].

- Collaborative Filtering. In collaborative filtering (CF), systems try to gain information about similarity of items by learning past user-item relationships. One possible way to do this is to collect lots of playlists of different users and then suggesting songs to be similar, if they appear together in many of these playlists. A major drawback is the cold start for items. Songs that are newly added to a database do not appear in playlists, so no information about them can be collected. Popular examples for CF recommendation are last.fm[1] and amazon.com[2].
- Content-Based Techniques. In the content-based approach (CB), the content of musical pieces is analyzed, and similarity is calculated from the descriptions as result of the content analysis. Songs can be similar if they have the same timbre or rhythm. This analysis can be done by experts (e.g., Pandora[3]) , which leads to high quality but expensive descriptions, or automatically, using signal processing and machine learning algorithms (e.g., mufin[4]), as explained in sections 2,3, and 4. Automatic content-based descriptors cannot yet compete with manually derived descriptions, but can be easily process large databases.
- Context-Based Techniques. By analyzing the context of songs or artists, similarities can also be derived. For example, contextual information can be acquired as a result of web-mining (e.g., analyzing hyperlinks between artist homepages)[66], or collaborative tagging [101].
- Demographic Filtering Techniques. Recommendations are made based on clusters that are derived from demographic information, e.g. “males at your age from your town, who are also interested in soccer, listen to...”.

By combining different techniques to hybrid systems, drawbacks can be compensated, as described in [96], where CB similarity is used to solve the item cold start problem of a CF system.

A very important issue within recommendation is the user. In order to make personalized recommendations, the system has to collect information about the musical taste of the user and contextual information about the user himself. Two questions arise: How are new user profiles initialized (user cold start), and how are they maintained? The user cold start can be handled in different ways. Besides starting with a

[1] http://www.last.fm
[2] http://www.amazon.com
[3] http://www.pandora.com
[4] http://www.mufin.com

blank profile, users could enter descriptions of their taste by providing their favorite artists or songs, or rating some exemplary songs. Profile maintenance can be performed by giving feedback about recommendations in an explicit or implicit way. Explicit feedback includes rating of recommended songs, whereas implicit feedback includes information of which song was skipped or how much time a user spent on visiting the homepage of a recommended artist.

In CB systems, recommendations can be made by simply returning the most similar songs (according to computed similarity as described in Sec. 3.3) to a reference song. This song, often called "seed song" represents the initial user profile. If we just use equal weights of all features, the same seed song will always result in the same recommendations. However, perceived similarity between items may vary from person to person and situation to situation. Some of the acoustic features may be more important than others, therefore the weighting of the features should be adjusted according to the user, leading to a user-specific similarity function.

Analyzing user interaction can provide useful information about the user's preferences and needs. It can be given in a number of ways. In any case, usability issues should be taken into account. An initialization of the user profile by manually labeling dozens of songs is in general not reasonable. In [10], the music signal is analyzed with respect to semantically meaningful aspects (e.g., timbre, rhythm, instrumentation, genre etc.). The user can now weight or disable single aspects or domains to adapt the recommendation process to his own needs. For instance, similarities between songs can be computed by considering only rhythmic aspects.

The settings of weights can also be accomplished by collecting implicit or explicit user feedback. Implicit user interaction can be easily gathered by, e.g., tracing the user's skipping behavior ([87], [116]). The recommendation system categorizes already recommended songs as disliked songs, not listened to, or liked songs. By this means, one gets three classes of songs: songs the user likes, songs the user dislikes and songs, that have not yet been rated and therefore lack a label. Explicit feedback is normally collected in form of ratings. Further information can be collected explicitly by providing a user interface, in which the user can arrange already recommended songs in clusters, following his perception of similarity. Machine learning algorithms can be used to learn the "meaning" behind these clusters and classify unrated songs following the same way. This is analogous to the approach described in Sec. 4, where semantic properties are learned from exemplary songs clustered in classes. In [76], explicit feedback is used to refine the training data for an SVM classifier.

The user model, including seed songs, domain weighting or feedback information, can be interpreted as a reflection of the user's musical taste. The primary use is to improve the recommendations. Now songs are not further recommended solely based on a user-defined song, instead the user model is additionally incorporated into the recommendation process. Besides, the user model can also serve as a base for neighbor recommendation in a social web platform.

Recommendation algorithms should be evaluated according to their usefulness for an individual, but user-based evaluations are rarely conducted since they require a lot of user input. Therefore, large scale evaluations are usually based on similarity analysis (derived from genre similarities) or the analysis of song similarity graphs.

In one of the few user-based evaluations [14] it is shown that CF recommendations score better in terms of relevance, while CB recommendations have advantages regarding to novelty. The results of another user-based evaluation [75] supports the assumption that automatic recommendations are yet behind the quality of human recommendations.

The acceptance of a certain technique further depends on the type of user. People who listen to music, but are far from being music fanatics (about 3/4 of the 16-45 year old, the so called “Casuals” and “Indifferents”, see [54]) will be fine with popular recommendations from CF systems. By contrast the “Savants”, for which “Everything in life seems to be tied up with music” ([54]) might be bored when they want to discover new music.

Apart from that, hybrid recommender systems, which combine different techniques and therefore are able to compensate for some of the drawbacks of a standalone approach, have the largest potential to provide good recommendations.

6 Visualizing Music for Navigation and Exploration

With more and more recommendation systems available, there is a need to visualize the similarity information and to let the user explore large music collections. Often an intuitively understandable metaphor is used for exploration. As already illustrated in Sec. 5, there are several ways to obtain similarities between songs. The visualization of a music archive is independent from the way the similarity information was gathered from the recommenders. There exist visualization interfaces that illustrate CB, CF or web-based similarity information or that combine different sources for visualization.

This section deals with approaches and issues for music visualization. First, a brief overview of visualizing musical work is given. The next subsection deals with visualizing items in music archives followed by a description of browsing capabilities in music collections.

6.1 Visualization of Songs

Early work on visualizing songs was performed by [29]. Self-similarity matrices are used to visualize the time structure in music. Therefore, the acoustic similarity between any two instances of a musical piece is computed and plotted as a two-dimensional graph. In [65], Lillie proposes a visualization technique based on acoustic features for the visualization of song structure. The acoustic features are computed based on the application programmers interface (API) of EchoNest[5]. In the 2-dimensional plot, the x-axis represents the time of the song and the y-axis the chroma indices. Additionally, the color encodes the timbre of the sound. An example is given in Fig. 6. The acoustic features for the *Moonlight Sonata* of Beethoven are displayed on the left and the song *Cross the Breeze* from Sonic Youth is displayed on the right.

[5] http://developer.echonest.com/pages/overview

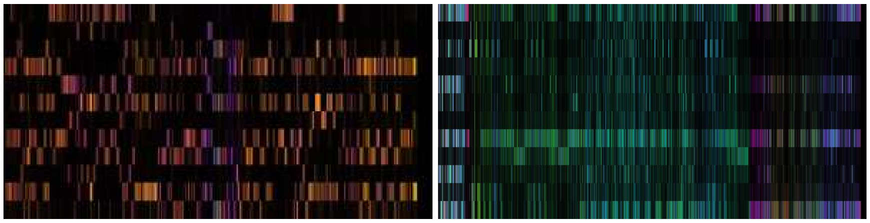

Fig. 6. Visualizing the structure of songs. Left: Visualization of the *Moonlight Sonata* of Beethoven, Right: Visualization of the song *Cross the Breeze* from Sonic Youth (http://www.flyingpudding.com/projects/viz_music/)

In [119], Yoshii et al. propose the visualization of acoustic features through image thumbnails to let the user guess the music content through the appearance of the thumbnail and decide if he wants to listen to it. The mapping between the acoustical space and the visual space is performed via an optimization method, additionally taking some constraints into account. Hiraga et al. [47] propose a 3-D visualization technique for MIDI data. They visualize the performance of musical pieces by focusing on the musical expression like articulation, tempo, dynamic change and structure information. For further reading, the interested reader is referred to [52], where an overview of visualization techniques for musical work with MIR methods is given.

Most work done in song visualization is independent of the work performed in visualization of music archives. From the next subsection it becomes apparent, that visualization of music archives mainly concentrates on the arrangement of songs in the visualization space. One main focus is to realize the paradigm of *closeness encodes similarity* rather than a sophisticated visualization of the song itself. Nevertheless one has to keep in mind that music archives consist of songs. Combined visualization techniques that also stress the musical characteristics of each song in a music archive are still an open research issue.

6.2 Visualization of Music Archives

The key point when visualizing music archives is how to map the multidimensional space of music features per song (compare also Sec. 2) to a low dimensional *visualization space*. Usually a 2-D plot or a 3-D space are used as visualization spaces. The placement of a song in the visualization space is depending on the similarity of this song to neighbored songs. Therefore a mapping of the acoustic features to a spatial distance is performed. For the user it is intuitive and easy to understand that closely positioned songs have similar characteristics. Next to the placement of the songs in this visualization space, additional features can be encoded via the color or the shape of the song icon.

Islands of Music [88] is a popular work for visualizing music archives. The similarities are calculated with content-based audio features and organized in a SOM (compare Sec. 3.1). Continents and islands in the geographic map represent genres.

The *MusicMiner* system [81] uses ESOM to project the audio features onto a topographic map. An example is illustrated in Fig. 7.

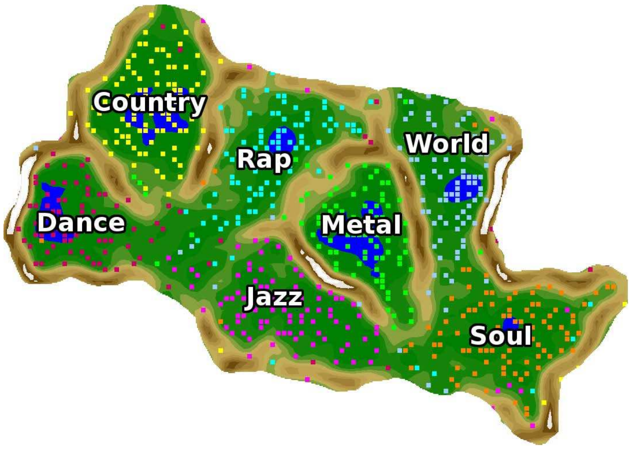

Fig. 7. MusicMiner: 700 songs are represented as colored dots.

Kolhoff et al. [57] use glyphs to represent each song based on its content. The songs are projected into a 2-D space by utilizing a PCA for dimension reduction (see Sec. 3.1) with a special weighting and relaxation for determining the exact position. Also in [85], a PCA is used to determine the three most important principal components and project the feature vectors onto the three resulting eigenvectors. The feature vectors are deskewed and the resulting vectors are reduced to two dimensions via a second PCA. Torrens et al. [107] propose different visualization approaches based on metadata. In their *disc visualization*, each sector of the disc represents a different genre. The songs are mapped to the genres while tracks in the middle are the oldest. They use this visualization technique to visualize playlists.

Requirements for the visualization of music archives are the scalability to large numbers of songs and computational complexity. Even for music archives containing hundreds of thousands of songs, the algorithm has to be able to position every song in the visualization space quickly.

6.3 Navigation and Exploration in Music Archives

Digital music collections are normally organized in folders, sorted corresponding to artists or genres, forcing the user to navigate through the folder hierarchy to find songs. They only allow for a text-based browsing in the music collection. A completely different paradigm for exploring music collections is the comprehensive search for similar music by browsing through a visual space. In this section, a short review about navigation and browsing capabilities is given. There are some overlaps to the section about visualization of music archives since most visualization scenarios also offer a browsing possibility. Here the focus is on approaches that concentrate more on browsing.

A popular method is the use of metaphors as underlying space for visualization. A metaphor provides an intuitive access for the user and an immediate understanding of the dimensions. There were already examples of geographic metaphors in the previous section. In [35] the metaphor of a *world of music* is used. The authors focus on compactly representing similarities rather than on visualization. The similarities are obtained via CF methods, a graph from pairwise similarities is constructed and mapped to Euclidean space while preserving distances. [46] uses a *radar system* to visualize music. Similar songs are located closely to each other and the eight directions from the radial plot denote different oppositional music characteristics like calm vs. turbulent or melodic vs. rhythmic. The actual chosen song is placed in the middle of the radar. *MusicBox* is a music browser that organizes songs in a 2D-space via a PCA on the music features [65]. It combines browsing techniques, visualization of music archives and visualization of the song structure in one application.

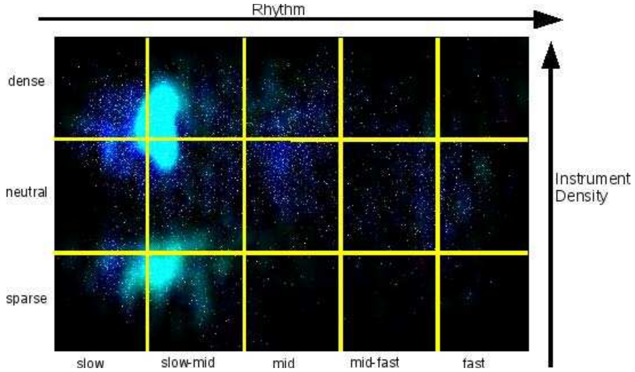

Fig. 8. Semantic browsing in a stars universe. The x-axis encodes the rhythm of the songs and the y-axis the instrument density. For illustration purposes the semantic regions are marked in yellow.

In Fig. 8 we show an example of the metaphor *stars universe*. The 2-D universe is representing the musical space and stars are acting as visual entities for the songs. The user can navigate through this universe finding similar songs arranged closely to each other, sometimes even in star concentrations. The visualization space is subdivided into several semantic regions. On the x-axis there are the rhythmic characteristics from slow to fast subdivided in five gradations and the y-axis contains the instrument density from sparse to full in three gradations. To position a song in the universe, a similarity query on a rhythmic and an instrument density reference set is performed. Each reference set contains the feature vectors of three songs per gradation. For both reference sets the winning song determines the subregion in the visualization space, the rhythmic one for the x-axis and the other for the y-axis. The exact position in the subregion is influenced by locally translating each song in the subspace in dependence from the mean and standard deviations of the song positions belonging to the same region (compare [85]).

A quite different approach is performed in [9]. Here, the *collaging technique*, emerged from the field of digital libraries, is used to visualize music archives and enable browsing based on metadata. Other research focuses on visualizing music archives on mobile devices, e.g., [84]. In [17] a music organizer and browser for children is proposed. The authors stress the needs from children for music browsing and provide a navigation software.

6.4 Summary and Open Issues

All the presented approaches offer the user a different insight into his music collection and allow for a discovery of new, unknown songs, that match to the preferences of the user. The main drawback of visualization and browsing methods that project the high-dimensional feature space of acoustic features into a low (2-D or 3-D) visualization space with dimension reduction methods, is the lack of *semantic browsing*. For the user it is not apparent which semantic entity changes by navigating along one axis. Although nearly located songs are most similar to each other, it is not intuitive which musical characteristic changes when browsing through the visualization space. As a solution many approaches introduce semantic entities like genre mountains. These can serve as a landmark for the user and describe which musical characteristics are typical for a specific direction. Another possibility is the use of high-level features. One example from Sec. 6.3 is the radar system, where each radial direction refers to a change in a special semantic characteristic. Another example is the stars universe, also presented in Sec. 6.3. Problems with these approaches are due to the fact that music is not eight-dimensional or two-dimensional, but multidimensional. So it is not possible to define the holistic impression of music along a few semantic dimensions. One has to abstract that the songs are similar in the mentioned dimensions but regarding other musical aspects, neighbored songs can sound very differently.

7 Applications

Today both physical products (records and CDs) as well as virtual goods (music tracks) are sold via Internet. To find the products, there is an increasing need for search functionalities. This need has been addressed by a number of search paradigms. Some just work, even without scientific foundation, others use elaborated models like the ones described in this book.

During the last years, a large amount of MIR-based applications and services appeared. Some of them generated quite some attention in online communities. Some of the underlying techniques are still subject to basic research and not yet understood to the utmost extent. However, the competition for unique features incited many small start-up companies as well as some innovation-oriented big players to push immature technologies to the market. Below we list some applications, that integrate automatic CB methods to enable retrieval and recommendation of music. The focus is clearly on CB based systems. Beyond the applications below, there are a large number of strictly CF-based systems around. Applications that are merely scientific showcases without significant commercial ambitions will not be mentioned here. Furthermore, a distinction is made between projects that make their applications publicly available and companies that approach other entities and offer them their services. In the latter case, it is difficult to assess whether the capabilities of the real product can live up to their marketing promises. It should be noted, that this section does not claim to be absolutely comprehensive. There are probably some more projects and companies on the Asian market which we do not know due to language-barriers. Furthermore, the market for MIR-applications is quite volatile, so the examples in the following sections can only provide a snapshot of the current situation.

7.1 Business to Business Applications

The American company Gracenote[6] is probably best known for providing the CDDB CD identification service. Today, they have added different solutions for music identification and recommendation to their portfolio. Their recommendation service ”Discover” is based on editorial recommendations, CB and CF methods.

The Canadian company Double V3[7] provides audio identification services to the music and entertainment industry.

The US-based company One Llama[8] can rely on a core team with long experience in academic MIR research. One Llama’s flagship is called ’Artificial Ear’ and is said to have extracted hundreds of music features from millions of songs. Their music discovery tools are based on a combination of CB and CF techniques.

The Echo Nest’s[9] APIs are based on the so-called ”Musical Brain”. Following their description, the MIR-platform combines CB-recommendation with webcrawling and knowledge extraction. The founders of the company have a history with the MIT Media Lab.

The San Francisco based company Music Intelligence Solutions[10] and its Barcelona based predecessor Polyphonic Human Media Interface (PHMI) are offering diverse solutions for music discovery. They are especially well known for the “Hit Song Science” tool that claims to reliably measure the hit potential of novel songs.

The New York based company Music Xray[11] has a common history with aforementioned Music Intelligence Solutions. Their portfolio comprises a web service that allows artists and music industry professionals to measure, monitor, stimulate the demand for novel artists and their songs. They have teamed up with Queen Mary University’s Centre For Digital Music.

[6] http://www.gracenote.com/business_solutions/discover/

[7] http://www.doublev3.com/

[8] http://onellama.com/

[9] http://the.echonest.com/

[10] http://www.uplaya.com/company.html

[11] http://www.musicxray.com/music-xray

The Spanish company BMAT[12] is a commercial spin-off of the Music Technology Group, the music and audio research lab of the Universitat Pompeu Fabra in Barcelona. BMAT generated quite some public attention when they powered the casting of a Spanish idol show with a web application that automatically evaluated the singing.

The Norwegian company Bach Technology[13] benefits from a long tradition with related projects in the digital content domain. Bach Technology develops and distributes audio search and annotation technology to stimulate sales in the ”Long Tail” of music catalogs.

7.2 Business to Consumer Applications

The goal of the aforementioned German company mufin is to foster music consumption and sales by helping end-users to discover new music that is relevant to their personal preferences. Their products enable discovery and management in large-scale music collections. In addition, mufin delivers several applications for free download, such as a recommender Plug-In for Apple iTunes and a stand-alone media player.

The California based company MusicIP[14] was one of the pioneers that made MIR-applications accessible to end users. Their flagship application is called “MyDJ Desktop”. It allows the creation of CB based similarity playlists. Additionally, their music identification service “MusicDNS” has an extensive database of reference music fingerprints available. It is the basis for the community music metadatabase MusicBrainz[15].

Midomi[16] is a melody search engine combined with a community portal. Midomi circumvents the typical problems of how to acquire melody information in a clever way. They let their end-users maintain and update the melody database. The input can be either singing, humming or whistling. The company behind the service is MELODIS, based in Silicon Valley. Their goal is the development of next generation of search and sound technologies for global distribution on a wide range of mobile platforms and devices.

The U.K. based company Shazam[17] started their business in audio identification in 2002 and emerged as the leading mobile music identification service provider. They claim to have a fingerprint database of over 6 million tracks. The integration of their service into the Apple iPhone made the service very well known among technology-affine communities.

The Berlin based company AUPEO![18] is one of the first that combine a music-lovers social network with a mood-based personalized internet radio. The mood an-

[12] http://bmat.com/

[13] http://www.bachtechnology.com/

[14] http://www.musicip.com/

[15] http://musicbrainz.org/

[16] http://www.midomi.com/

[17] http://www.shazam.com/music/web/home.html

[18] http://www.aupeo.com/

notations of their music catalogs are computed by CB methods. Their unique business idea is the integration of their service into hardware devices, such as inexpensive internet-radios.

8 Future Directions and Challenges

The former sections of this article presented important aspects and first results of state-of-the-art MIR research. However, it seems that many available technologies are just in their infancy as it was summarized in a mentionable survey by Lew [60] for the whole multimedia information retrieval sector. Despite considerable research progress and the astonishing amount of different projects and applications already on the market, there is no final solution to be seen that would solve the aforementioned problems related to music recommendation sufficiently. Even worse, there is a lack of adequate business models to make MIR-technologies an indispensable helper for modern and technology-oriented lifestyle. The very first MIR-based applications to become publicly available seemed like toys. This situation is changing slowly with the integration of recommendation technologies into mobile devices and other consumer electronics hardware. This takes originally strictly web-based applications directly into everybodies living room or car.

Context-Sensitivity

Purely CB methods for music recommendation and retrieval know very little about the real world. Learning systems can be taught some semantics about music, styles and moods. But there are complex inter-dependencies between sociocultural aspects, the users' current condition as well as environmental factors. In [16] some future directions are proposed that are promising with respect to mobile applications. In fact, new interactive devices (positional tracking, health, inclination sensors) may provide new possibilities, such as human emotional state detection and tracking. Both, content of a song and context of the user are important to understand why a user likes or dislikes a song. The decoding of this relation will indeed require lots of further research. And once it is done, it remains to be shown that knowledge about the "why" will help finding other songs that satisfy these conditions. This will be a step towards high-quality individual recommendations that are independent from what other users feel.

Semantic Web and Music Ontologies

A conceptualization is an abstract, simplified view of the world. Every knowledge-based system is committed to some conceptualization, explicitly or implicitly. An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of existence. For MIR systems, the musicological knowledge could be described in an ontology. As an example, the Music Ontology Specification[19] provides main concepts and properties for describing music (i.e. artists, albums, tracks, but also performances, arrangements, etc.) on the Semantic Web. This initiative shall enable the interlinking of music related databases on the semantic level [91].

Folksonomies

The majority of the existing search engines are using simple keyword based approaches. Large heterogeneous collections of music can probably not be sufficiently described using a rigid, pre-defined taxonomy. By freely assigning tags, unstructured files can be made searchable since there exists a multitude of stable and robust solutions for text search. Tagging is strongly connected to concepts like wisdom of the crowds or crowd-sourcing. It is based on the assumption that tags assigned by a number of human listeners will result in ”wise tags”. Although the individual can only provide weak labeling, the crowd is assumed to provide more reliable metadata in case consistent labeling can be observed. This idea is appealing and made last.fm and MP3.com[20] useful and popular. In [102], it can be found that taxonomies created by experts are useful for cataloging and hierarchical browsing while the flat view of folksonomies allows better organization and access of a personal collection. Thus, it can be assumed that a combination of taxonomy and folksonomy are a promising future direction. Furthermore, adaptive MIR models can be trained using the music examples labeled with tags in order to assign such tags automatically afterwards [24].

Hybrid Systems

According to first published work ([120],[105]) the combination of automatic content based and CF methods might be beneficial for the further developments of music retrieval and recommendation systems. One intuitive advantage is the aforementioned (see Sec. 5) possibility to avoid the cold-start problem inherent to CF based systems, by recommending novel or unknown songs based on their acoustic properties. More advantages are to be expected from merging social and CB music metadata with musicological knowledge as described before. Such systems should then be able to derive the importance of given or computed information for a certain task in a certain context in order to optimize the decision process or to assess the precision of data sources in order to suppress uncertain information in an autonomous manner. As another example CB similarity measures can probably be utilized to automatically correlate the meaning of different tags given by users.

[19] http://www.musicontology.com/

[20] http://www.mp3.com

Scalability

There are different approaches to deal with large amounts of music content in identification scenarios which have proven to work reliably in real-world applications. However, it is still an open problem how to deal with millions of songs in more fuzzy retrieval and recommendation tasks. As an example, currently music similarity lists in catalogs of several million songs have to be pre-computed. This however collides with the demand for personalized music recommendations tuned to the listeners very own preferences. It is an interesting question whether the consideration of musical knowledge in hybrid recommenders will be able to improve the scalability problem.

Diversity

While most CB MIR tasks today are dealing with popular western music content, the diversity of all the music available on a global scale is much wider. More and more musicians and listeners from currently underrepresented regions of the world will be joining the global stage in the near future facing the MIR community with new challenges regarding musical and cultural diversity. An ongoing project that is investigating these topics is GlobalMusic2One[21].

Scientific exchange

For the future development of the MIR research scientific exchange is an essential issue. In that regard the Music Information Retrieval Evaluation eXchange (MIREX)[22] is a very commendable initiative of the University of Illinois at Urbana-Champaign. The big problem for such contests is the struggle with the limited availability of common music test beds. In the past some independent labels have released content for certain competitions, but most researchers have originally started with their own test sets, often ripped from commercial CDs. These sets are annotated, but may not be shared due to copyright issues. There exists some databases (e.g., [34], [32]) that are intended to be shared among researchers. Unfortunately, their usage is not as widespread as it could be.

Conclusion

As a conclusion it can be clearly stated, that most problems in CB MIR are still far from being finally solved. The only task that has matured to real-world applicability is probably the audio identification task, as the very successful examples in Sec. 7 show. Generally speaking, all of the tasks described in this chapter need significant further research.

[21] http://www.globalmusic2one.net

[22] http://www.music-ir.org/mirex/2009/index.php/Main_Page

References

1. Abeßer, J., Dittmar, C., Großmann, H.: Automatic genre and artist classification by analyzing improvised solo parts from musical recordings. In: Proceedings of the Audio Mostly Conference (AMC). Piteå, Sweden (2008)
2. Allamanche, E., Herre, J., Hellmuth, O., Kastner, T., Ertel, C.: A multiple feature model for music similarity retrieval. In: Proceedings of the 4th International Symposium of Music Information Retrieval (ISMIR). Baltimore, Maryland, USA (2003)
3. Allamanche, E., Herre, J., Helmuth, O., Froba, B., Kastner, T., Cremer, M.: Content-based identification of audio material using MPEG-7 low level description. In: Proceedings of the 2nd International Symposium of Music Information Retrieval (ISMIR). Bloomington, Indiana, USA (2001)
4. Anderson, C.: The Long Tail: Why the Future of Business is Selling Less of More. Hyperion, New York, NY, USA (2006)
5. Aucouturier, J.J., Defreville, B., Pachet, F.: The bag-of-frame approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. Journal of the Acoustical Society of America **122**(2), 881–891 (2007)
6. Aucouturier, J.J., Pachet, F.: Music similarity measures: What's the use? In: Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR). Paris, France (2002)
7. Aucouturier, J.J., Pachet, F.: Improving timbre similarity: How high is the sky? Journal of Negative Results in Speech and Audio Sciences **1**(1), 1–13 (2004)
8. Aucouturier, J.J., Pachet, F., Sandler, M.: The way it sounds: timbre models for analysis and retrieval of music signals. IEEE Transactions on Multimedia **7**(6), 1028–1035 (2005)
9. Bainbridge, D., Cunningham, S., Downie, J.: Visual collaging of music in a digital library. In: Proceedings of the International Conference on Music Information Retrieval (ISMIR). Barcelona, Spain (2004)
10. Bastuck, C., Dittmar, C.: An integrative framework for content-based music similarity retrieval. In: Proceedings of the 35th German Annual Conference on Acoustics (DAGA). Dresden, Germany (2008)
11. Bello, J.P., Pickens, J.: A robust mid-level representation for harmonic content in music signals. In: Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR). London, UK (2005)
12. Brown, J.: Determination of the meter of musical scores by autocorrelation. Journal of the Acoustical Society of America **94**(4), 1953–1957 (1993)
13. Casey, M.: MPEG-7 sound recognition. IEEE Transactions on Circuits and Systems Video Technology, special issue on MPEG-7 **11**, 737–747 (2001)
14. Celma, O.: Music recommendation and discovery in the long tail. Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain (2008)
15. Chen, P.H., Cheh-Jen, L., Schölkopf, B.: A turorial on ν-support vector machines. Tech. rep., Department of Computer Science and Information Engineering, Taipei, Max Planck Institute for Biological Cybernetics, Tübingen (2005)
16. Cunningham, S., Caulder, S., Grout, V.: Saturday night or fever? Context aware music playlists. In: Proceeding of the Audio Mostly Conference (AMC). Piteå, Sweden (2008)
17. Cunningham, S., Zhang, Y.: Development of a music organizer for children. In: Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR). Philadelphia, Pennsylvania (2008)
18. Dempster, A.P., Laird, N.M., Rdin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, Series B **39**, 1–38 (1977)

19. Dittmar, C., Bastuck, C., Gruhne, M.: Novel mid-level audio features for music similarity. In: Proc. of the Intern. Conference on Music Communication Science (ICOMCS). Sydney, Australia (2007)
20. Dittmar, C., Dressler, K., Rosenbauer, K.: A toolbox for automatic transcription of polyphonic music. In: Proceedings of the Audio Mostly Conference (AMC). Ilmenau, Germany (2007)
21. Dittmar, C., Uhle, C.: Further steps towards drum transcription of polyphonic music. In: Proceedings of the AES 116th Convention (2004)
22. Dixon, S.: Onset detection revisited. In: Proceedings of the 9th International Conference on Digital Audio Effects (DAFx06). Montréal, Québec, Canada (2006)
23. Dunker, P., Nowak, S., Begau, A., Lanz, C.: Content-based mood classification for photos and music: A generic multi-modal classification framework and evaluation approach. In: Proceedings of the International Conference on Multimedia Information Retrieval (ACM MIR). Vancouver, Canada (2008)
24. Eck, D., Bertin-Mahieux, T., Lamere, P.: Autotagging music using supervised machine learning. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR). Vienna, Austria (2007)
25. Eerola, T., North, A.C.: Expectancy-based model of melodic complexity. In: Proceedings of the 6th International Conference of Music Perception and Cognition (ICMPC). Keele, Staffordshire, England (2000)
26. Ellis, D.: Classifying music audio with timbral and chroma features. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR). Vienna, Austria (2007)
27. Feng, Y., Zhuang, Y., Pan, Y.: Music information retrieval by detecting mood via computational media aesthetics. International Conference onWeb Intelligence (IEEE/WIC) pp. 235–241 (2003)
28. Flexer, A., Pampalk, E., Widmer, G.: Hidden markov models for spectral similarity of songs. In: Proceedings of the 8th International Conference on Digital Audio Effects (DAFX'05). Madrid, Spain (2008)
29. Foote, J.: Visualizing music and audio using self-similarity. In: Proceedings of the seventh ACM international conference on Multimedia (Part 1). New York, NY, USA (1999)
30. Foote, J.T.: Content-based retrieval of music and audio. In: Proceeding of SPIE Conference on Multimedia Storage and Archiving Systems II. Dallas, TX, USA (1997)
31. Fukunaga, K.: Introduction to Statistical Pattern Recognition, Second Edition (Computer Science and Scientific Computing Series). Academic Press (1990)
32. Gillet, O., Richard, G.: Enst-drums: an extensive audio-visual database for drum signals processing. In: Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR). Victoria, BC, Canada (2006)
33. Goto, M.: A real-time music-scene-description system - predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. Speech Communication **43**, 311–329 (2004)
34. Goto, M.: AIST annotation for the RWC music database. In: Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR). Victoria, BC, Canada (2006)
35. Goussevskaia, O., Kuhn, M., Lorenzi, M., Wattenhofer, R.: From Web to Map: Exploring the World of Music. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). Sydney, Australia (2008)
36. Gouyon, F., Fabig, L., Bonada, J.: Rhythmic expressiveness transformations of audio recordings - swing modifications. In: Proceedings of the 60th International Conference on Digital Audio Effects (DAFx). London,UK (2003)

37. Gouyon, F., Herrera, P.: Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors. In: Proceedings of the 114th AES Convention. Amsterdam, Netherlands (2003)
38. Gouyon, F., Klapuri, A., Dixon, S., Alonso, M., Tzanetakis, G., Uhle, C., Cano, P.: An experimental comparison of audio tempo induction algorithms. IEEE Transactions on Speech and Audio Processing **14**, 1832–1844 (2006)
39. Gouyon, F., Pachet, F., Delerue, O.: The use of zerocrossing rate for an application of classification of percussive sounds. In: COST G-6 Conference on Digital Audio Effects (DAFx). Verona, Italy (2000)
40. Hainsworth, S.W., Macleod, M.D.: Automatic bass line transcription from polyphonic music. In: Proceedings of the International Computer Music Conference (ICMC). Havana, Cuba (2001)
41. Hanjalic, A.: Extracting moods from pictures and sounds. IEEE Signal Processing Magazine **23**(2), 90–100 (2006)
42. Harte, C.A., Sandler, M.B.: Automatic chord identification using a quantised chromagram. In: Proceedings of the 118th AES Convention. Barcelona, Spain (2005)
43. Herre, J., Allamanche, E., Ertel, C.: How similar do songs sound? In: Proceedings of the IEEE Workshop on Applications of Singal Processing to Audio and Acoustics (WASPAA). Mohonk, New York, USA (2003)
44. Herrera, P., Sandvold, V., Gouyon, F.: Percussion-related semantic descriptors of music audio files. In: Proceedings of the 25th International AES Conference. London, UK (2004)
45. Hevner, K.: Experimental studies of the elements of expression in music. American Journal of Psychology **48**(2), 246–268 (1936)
46. Hilliges, O., Holzer, P., Kluber, R., Butz, A.: AudioRadar: A metaphorical visualization for the navigation of large music collections. Lecture Notes in Computer Science **4073**, 82 (2006)
47. Hiraga, R., Mizaki, R., Fujishiro, I.: Performance visualization: a new challenge to music through visualization. In: Proceedings of the 10th ACM international conference on Multimedia. New York, NY, USA (2002)
48. Hsu, C., Chang, C., Lin, C., et al.: A practical guide to support vector classification. Tech. rep., National Taiwan University, Taiwan (2003)
49. Hsu, J.L., Liu, C.C., Chen, A.L.P.: Discovering nontrivial repeating patterns in music data. IEEE Transactions on Multimedia **3**(3), 311–325 (2001)
50. Hu, X., Downie1, J.S., Laurier, C., Bay, M., Ehmann, A.F.: The 2007 MIREX Audio Mood Classification Task: Lessons Learned. In: Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR). Philadelphia, Pennsylvania, USA (2008)
51. Hunt, M.J., Lennig, M., Mermelstein, P.: Experiments in syllable-based recognition of continuous speech. In: Proceedings of the International Conference on Acoustics and Signal Processing (ICASSP). Denver, Colorado, USA (1980)
52. Isaacson, E.: What you see is what you get: on visualizing music. In: Proceedings of the International Conference on Music Information Retrieval. London, UK (2005)
53. ISO/IEC: ISO/IEC 15938-4 (MPEG-7 Audio). ISO (2002)
54. Jennings, D.: Net, Blogs and Rock ’n’ Roll: How Digital Discovery Works and What it Means for Consumers. Nicholas Brealey Publishing (2007)
55. Johnston, J.: Transform coding of audio signals using perceptual noise criteria. IEEE Journal on Selected Areas in Communications **6**(2), 314–322 (1988)

56. Kim, Y., Whitman, B.: Singer identification in popular music recordings using voice coding features. In: Proceedings of 3rd International Symposium on Music Information Retrieval (ISMIR). Paris, France (2002)
57. Kolhoff, P., Preuß, J., Loviscach, J.: Content-based icons for music files. Computers & Graphics **32**(5), 550–560 (2008)
58. Kullback, S.: Information Theory and Statistics (Dover Books on Mathematics). Dover Publications (1997)
59. de Léon, P.J.P., Inesta, J.M.: Pattern recognition approach for music style identification using shallow statistical descriptors. IEEE Transactions on System, Man and Cybernetics - Part C : Applications and Reviews **37**(2), 248–257 (2007)
60. Lew, M.S., Sebe, N., Lifl, C.D., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications (2006)
61. Li, T., Ogihara, M.: Detecting emotion in music. Proceedings of the Fifth International Symposium on Music Information Retrieval pp. 239–240 (2003)
62. Li, T., Ogihara, M.: Content-based music similarity search and emotion detection. Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). **5** (2004)
63. Licklider, J., Pollack, I.: Effects of differentiation, integration, and infinite peak clipping on the intelligibility of speech. Journal Acoustical Society of America **20**, 42–51 (1948)
64. Lidy, T., Rauber, A., Pertusa, A., Iesta, J.M.: Improving genre classification by combination of audio and symbolic descriptors using a transcription system. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR). Vienna, Austria (2007)
65. Lillie, A.S.: Musicbox: Navigating the space of your music. Master's thesis, Massachusetts Institute of Technology, USA (2008)
66. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer, New York, NY, USA (2008)
67. Liu, C., Yang, Y., Wu, P., Chen, H.: Detecting and classifying emotion in popular music. In: 9th Joint International Conference on Information Sciences (2006)
68. Liu, D., Lu, L., Zhang, H.: Automatic mood detection from acoustic music data. In: Proceedings International Symposium Music Information Retrieval (ISMIR), pp. 81–87 (2003)
69. Liu, Z., Huang, Q.: Content-based indexing and retrieval-by-example in audio. In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME). New York City, NY, USA (2000)
70. Logan, B.: Mel frequency cepstral coefficients for music modeling. In: Proceedings of 1st International Symposium on Music Information Retrieval (ISMIR). Plymouth, Massachusetts, USA (2000)
71. Logan, B., Salomon, A.: A music similarity function based on signal analysis. In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME). Tokyo, Japan (2001)
72. Lu, L., Liu, D., Zhang, H.: Automatic mood detection and tracking of music audio signals. IEEE Transactions on Audio, Speech & Language Processing **14**(1), 5–18 (2006)
73. Lukashevich, H., Dittmar, C.: Applying statistical models and parametric distance measures for music similarity search. In: Proceedings of the 32nd Annual Conference of German Classification Society. Hamburg, Germany (2008)
74. Madsen, S.T., Widmer, G.: A complexity-based approach to melody track identification in midi files. In: Proceedings of the International Workshop on Artificial Intelligence and Music (MUSIC-AI). Hyderabad, India (2007)

75. Magno, T., Sable, C.: A comparison of signal-based music recommendation to genre labels, collaborative filtering, musicological analysis, human recommendation, and random baseline. In: Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR). Philadelphia, USA (2008)
76. Mandel, M.I., Poliner, G.E., Ellis, D.P.: Support vector machine active learning for music retrieval. Multimedia Systems **12**, 1–11 (2006)
77. de Mántaras, R.L., Arcos, J.L.: AI and music: From composition to expressive performances. AI Magazine **23**, 43–57 (2002)
78. McKay, C., Fujinaga, I.: Automatic genre classification using large high-level musical feature sets. In: Proceedings of the International Conference in Music Information Retrieval (ISMIR). Barcelona, Spain (2004)
79. Mierswa, I., Morik, K.: Automatic feature extraction for classifying audio data. Machine Learning Journal **58**, 127–149 (2005)
80. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York, USA (1997)
81. Mörchen, F., Ultsch, A., Nöcker, M., Stamm, C.: Databionic visualization of music collections according to perceptual distance. In: Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR). London, UK (2005)
82. Moschou, V., Kotti, M., Benetos, E., Kotropoulos, C.: Systematic comparison of BIC-based speaker segmentation systems. In: Proceedings of IEEE 9th Workshop on Multimedia Signal Processing (MMSP). Crete, Greece (2007)
83. Müller, M., Appelt, D.: Path-constrained partial music synchronization. In: Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Las Vegas, USA (2008)
84. Neumayer, R., Dittenbach, M., Rauber, A.: PlaySOM and PocketSOMPlayer, alternative interfaces to large music collections. In: Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR). London, UK (2005)
85. Nowak, S., Bastuck, C., Dittmar, C.: Exploring music collections through automatic similarity visualization. In: Tagungsband der DAGA Fortschritte der Akustik. Dresden, Germany (2008)
86. Pampalk, E.: Computational models of music similarity and their application in music information retrieval. Ph.D. thesis, Vienna University of Technology, Vienna, Austria (2006)
87. Pampalk, E., Pohle, T., Widmer, G.: Dynamic playlist generation based on skipping behaviour. In: Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR). London, UK (2005)
88. Pampalk, E., Rauber, A., Merkl, D.: Content-based organization and visualization of music archives. In: Proceedings of the 10th ACM international conference on Multimedia. New York, NY, USA (2002)
89. Peeters, G.: A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Tech. Rep. CUIDADO I.S.T. Project, Institut de Recherche et Coordination Acoustique/Musique (IRCAM), Paris, France (2004)
90. Poliner, G.E., Ellis, D.P.W., Ehmann, A.F., Gómez, E., Streich, S., Ong, B.: Melody transcription from music audio: Approaches and evaluation. IEEE Transactions on Audio, Speech, and Language Processing **15**, 1247–1256 (2007)
91. Raimond, Y.: A distributed music information system. Ph.D. thesis, Queen Mary, University of London, London, UK (2008)
92. Russell, J.: A circumplex model of affect. Journal of Personality and Social Psychology **39**(6), 1161–1178 (1980)

93. Ryyänen, M., Klapuri, A.: Automatic bass line transcription from streaming polyphonic audio. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Honolulu, Hawaii, USA (2007)
94. Ryynnen, M.P., Klapuri, A.P.: Automatic transcription of melody, bass line, and chords in polyphonic music. Computer Music Journal **32**, 72–86 (2008)
95. Saunders, C., Hardoon, D.R., Shawe-Taylor, J., Widmer, G.: Using string kernels to identify famous performers from ther playing style. In: Proceedings of the 15th European Conference on Machine Learning (ECML). Pisa, Italy (2004)
96. Schein, A.I., Popescul, R., Ungar, L.H., Pennock, D.M.: Methods and metrics for cold-start recommendations. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland (2002)
97. Schuller, B., Eyben, F., Rigoll, G.: Tango or waltz?: Putting ballroom dance style into tempo detection. EURASIP Journal on Audio, Speech, and Music Processing (JASMP) **2008**(6), 1–12 (2008)
98. Serra, J., Gomez, E.: Audio cover song identification based on tonal sequence alignment. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP). Las Vegas, USA (2008)
99. Sethares, W., Staley, T.: Meter and periodicity in musical performance. Journal of New Music Research **30**(2), 149–158 (2001)
100. Shao, X., Xu, C., Kankanhalli, M.: Unsupervised classification of music genre using hidden markov model. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Edinburgh, Scotland,United Kingdom (2004)
101. Smith, G.: Tagging: People-Powered Metadata for the Social Web. New Riders, Berkeley, CA, USA (2008)
102. Sordo, M., Celma, Ó., Blech, M., Guaus, E.: The quest for musical genres: Do the experts and the wisdom of crowds agree? In: Proceedings of the Ninth International Conference on Music Information Retrieval (ISMIR). Philadelphia, Pennsylvania, USA (2008)
103. Tellegen, A., Watson, D., Clark, L.: On the dimensional and hierarchical structure of affect. Psychological Science **10**, 297–303 (1999)
104. Thayer, R.: The Biopsychology of Mood and Arousal. Oxford University Press (1989)
105. Tiemann, M., Pauws, S., Vignoli, F.: Ensemble learning for hybrid music recommendation. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR). Vienna, Austria (2007)
106. Tolos, M., Tato, R., Kemp, T.: Mood-based navigation through large collections of musical data. In: 2nd IEEE Consumer Communications and Networking Conference. Las Vegas, Nevada, USA (2005)
107. Torrens, M., Hertzog, P., Arcos, J.: Visualizing and exploring personal music libraries. In: Proceedings of the International Conference on Music Information Retrieval (IS-MIR). Barcelona, Spain (2004)
108. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multilabel classification of music into emotions. In: Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR). Philadelphia, Pennsylvania, USA (2008)
109. Tsai, W., Wang, H.: Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. IEEE Transactions on Audio, Speech, and Language Processing **14**(1), 330–431 (2006)
110. Tzanetakis, G.: Manipulation, analysis and retrieval systems for audio signals. Ph.D. thesis, Princeton University, NJ, USA (2002)
111. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE transactions on Speech and Audio Processing **10**(5), 293–302 (2002)

112. Uhle, C.: Automatisierte extraktion rhythmischer merkmale zur anwendung in music information retrieval-systemen. Ph.D. thesis, Ilmenau University, Ilmenau, Germany (2008)
113. Wang, M., Zhang, N., Zhu, H.: User-adaptive music emotion recognition. In: 7th International Conference on Signal Processing, vol. 2, pp. 1352–1355 (2004)
114. Webb, A.: Statistical Pattern Recognition, 2nd edn. John Wiley and Sons Ltd. (2002)
115. West, K., Cox, S.: Features and classifiers for the automatic classification of musical audio signals. In: Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR). Barcelona, Spain (2004)
116. Wolter, K., Bastuck, C., Gärtner, D.: Adaptive user modeling for content-based music retrieval. In: Proceedings of the 6th Workshop on Adaptive Multimedia Retrieval (AMR). Paris, France (2008)
117. Wu, T., Jeng, S.: Probabilistic estimation of a novel music emotion model. In: 14th International Multimedia Modeling Conference. Springer (2008)
118. Yang, D., Lee, W.: Disambiguating music emotion using software agents. In: Proc. of the International Conference on Music Information Retrieval (ISMIR). Barcelona, Spain (2004)
119. Yoshii, K., Goto, M.: Music thumbnailer: Visualizing musical pieces in thumbnail images based on acoustic features. In: Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR). Philadelphia, Pennsylvania, USA (2008)
120. Yoshii, K., Goto, M., Komatani, K., Ogata, T., Okuno, H.G.: Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. In: Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR). Victoria, BC, Canada (2006)
121. Yoshii, K., Goto, M., Okuno, H.G.: Automatic drum sound description for real-world music using template adaption and matching methods. In: Proceedings of the 5th International Music Information Retrieval Conference (ISMIR). Barcelona, Spain (2004)
122. Zils, A., Pachet, F.: Features and classifiers for the automatic classification of musical audio signals. In: Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR). Barcelona, Spain (2004)