The original publication is available at www.springerlink.com
http://dx.doi.org/10.1007/s11634-007-0016-x

Classification in Music Research

Claus Weihs · Uwe Ligges · Fabian Mörchen · Daniel Müllensiefen

Accepted for publication in **Advances in Data Analysis and Classification**, 2007, Springer

Abstract Since some few years, classification in music research is a very broad and quickly growing field. Most important for adequate classification is the knowledge of adequate observable or deduced features on the basis of which meaningful groups or classes can be distinguished. Unsupervised classification additionally needs an adequate similarity or distance measure grouping is to be based upon. Evaluation of supervised learning is typically based on the error rates of the classification rules. In this paper we first discuss typical problems and possible influential features derived from signal analysis, mental mechanisms or concepts, and compositional structure. Then, we present typical solutions of such tasks related to music research, namely for organization of music collections, transcription of music signals, cognitive psychology of music, and compositional structure analysis.

Keywords Classification in Musicology · Automatic Transcription · Music Psychology · Organization of Music Collections · Compositional Structure Analysis

1 Introduction

Statistics in music research is since recently a very broad and quickly growing field applying statistical methods to various problems in music information retrieval. In this paper we mainly concentrate on the application of classification methods. Classification can be based on unsupervised learning (clustering) or supervised learning. Unsupervised learning, on the one hand, does not utilize any information about classes. It finds groups that emerge from the properties and the notion of similarity used to describe

Claus Weihs (corresponding author) · Uwe Ligges
Fachbereich Statistik, Universität Dortmund, 44221 Dortmund, Germany,
E-mail: weihs@statistik.uni-dortmund.de, E-mail: ligges@statistik.uni-dortmund.de

Fabian Mörchen
Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540, USA,
E-mail: fabian.moerchen@siemens.com

Daniel Müllensiefen
Department of Computing, Goldsmiths College, University of London, London, UK,
E-mail: d.mullensiefen@gold.ac.uk

music. Supervised learning, on the other hand, utilizes a correct grouping on a certain learning set and is aiming at a classification rule for objects with data to be observed in the future.

Most important for adequate classification is the knowledge of observable or deduced features on the basis of which meaningful groups or classes can be distinguished. Additionally, unsupervised classification needs an adequate similarity or distance measure the grouping is to be based upon. Supervised learning needs relevant classes observable in a learning data set. Evaluation of supervised learning is typically based on the error rates of the classification rules.

In this paper we first discuss typical problem types and corresponding influential features derived from signal analysis, mental mechanisms or concepts, and compositional structure (see section 3). Then, we present typical solutions for classification tasks related to music research, as

- organization of music collections,
- transcription of music signals,
- cognitive psychology of music, and
- compositional structure analysis.

Let us start with some motivation for the usefulness of these four tasks and an overview on the sections of this survey.

Organization of music collections: Collections of recorded musical pieces are often organized by categories such as genre, artist, and album. This information is often manually assigned to pieces, e.g. in the detailed analysis of music by the Music Genome Project (`http://www.pandora.com`). Supervised classification can be used to determine rules for automatical categorization of unlabelled pieces of music. A new artist or album can be assigned to a genre in order to list it in the appropriate section of an online music shop. Unsupervised classification can be used to group (or cluster) pieces of music together, for example by similarity of their timbre. The distribution of features like timbre over a collection of music can be used to categorize (Tzanetakis and Cook 2002), visualize (Pampalk et al 2002; Mörchen et al 2005a), and recommend (Stenzel and Kamps 2005) music. Audio features suggested to characterize the content of music pieces are discussed in section 3, proposed similarity measures are discussed together with solutions to the task in section 4.

Transcription of music signals: Transcription from some recorded or online performed music into notes is a widely desirable application. Example applications are: Transcription of folk songs, laymen who try to transcribe music for themselves, and correcting errors when teaching or practicing to sing or play instruments. The currently most popular application in the field of transcription is the game *SingStar* (`http://www.singstargame.com`) for the Sony Playstation that allows for Karaoke singing and shows whether the singer intonated correctly or not.

More formally, the transcription task is the transformation of music audio input into sheet music. We discuss transcription as a supervised classification problem, where the correct notes are known in a learning data set (see section 5). In such a problem, classes correspond to the note pitches and durations as well as the kinds of rests possibly occurring in the analyzed piece of music. The task is to reproduce the classes in sheet music by means of audio features of a music time series. The derived classification rule can be used for future production of sheet music when only the audio signals are available.

Cognitive psychology of music: The task of cognitive psychology of music, as discussed in section 6, is the classification of music-related behavior by means of variables related either to musical features or to mental mechanisms, skills or experience, such as memory, emotions, mental skills, or mental representations as introduced in section 3.3. Such tasks can be unsupervised, as with finding groupings in a psychological perception space, or supervised, as with finding the classification rules for observed abilities or properties.

Compositional structure analysis: Grouping music by their compositional structure has become increasingly popular in recent times, largely due to the increasing amount of music that is digitally available in a symbolic encoding form (e.g. MIDI, EsAC, kern or other codes that encode notes as the basic musical events).

One task within the field of compositional structure analysis is to group music pieces together according to stylistic criteria. This might be done with the aim to identify the composer for pieces where the author is unknown (*author attribution*) or to explore variability of compositional techniques in a large number of pieces from a particular composer (*stylistic analysis*). Classification can be employed in a supervised way or unsupervised from a more exploratory perspective, depending whether reliable author or stylistic information is available. Another task in this area is to identify higher level structures which are not directly observable (e.g. key, meter, form structure or expectation of musical events) from lower level musical events like pitch class or duration distributions or distribution of sequences of such events which can be directly taken from a musical score or other forms of symbolic music encoding (see section 3.2). Here, classification is mainly supervised, since most higher level features are generally provided from human annotations. The aim always is to classify music, this time by their compositional structure. Therefore, the task is similar to the task of organization of music collections, however, not signal analysis is used for classification but analysis of compositional structure based on known notes. Section 7 gives a brief overview.

Even though we concentrate on these four fields for lack of space we will not be able to cover the literature thoroughly because of the huge set of papers published in the meantime. In section 2 we will sketch some history of the field. In section 3 feature pre-processing is discussed. In the following sections on the four main topics of the paper, mainly very recent research is discussed in order to characterize current developments. Also, emphasis is laid more on the underlying problems, corresponding reference, and software, and on the description of basic ideas and concepts, not so much on the technical details of the derived methods.

2 History

Let us now give some historic review on pioneering work for classification in music research. Some milestones are the publications of the German Fucks from the 1950s and 1960s, which inspired very much the research of the first author of this article in this field. Indeed, Fucks was one of the first authors proposing characteristics of music compositions which enable to distinguish between different composers or different musical epoches, respectively. Fucks (1963, 1964, 1968) analyzed the structure of compositions in different music periods by means of statistical methods. For example the frequencies of transition intervals between neighboring or more distant tones for typical composers were analyzed. This way, the highly linear structure of compositions (reduced to the 1st violin) of Beethoven and Bach can be revealed in contrast

to greater independence of consecutive tones with Berg or Webern. Additional to the frequency of tone heights and of transition intervals, the frequency distribution of tone lengths, intervals of parallel tones, interval pairs as well as accords were analyzed (see also Fucks and Lauter 1965). Moreover, in order to generate a more general result, standard deviations of tone heights of many pieces in a music period were computed and compared over the periods. The result was, in a way, to be expected: the variation mainly increased from the past to the present. All these results opened a statistical view on the development of the structure of music compositions over the centuries. A brief summary of some of his results Fucks published also in English (Fucks 1962). For other older approaches to describing the compositional structure of music see also Meyer (1957); Moles (1958, 1971); Steinbeck (1982). Moreover, first visualizations of genre comparisons were presented, e.g., in Fucks and Lauter (1965), where differences between different musical epoches were typically visualized by means of scatterplots with time at the x-axis and various characteristics for musical style at the y-axis, e.g. standard deviation of the pitches or entropy of pitch classes used in a piece. This lead to a first simple classification of compositions into musical epoches by means of characteristics of the compositions.

A very recent book about *Statistics in Musicology* (Beran 2004) also includes chapters about discriminant analysis, cluster analysis and multidimensional scaling related to the topic of this paper.

So much about history. Let us now switch to the very presence, and discuss current classification problems in music research and their solutions.

3 Feature Preprocessing

Several classification problems in music have been introduced and motivated in section 1. All these classification tasks can be tackled by utilizing features suggested to characterize the content of music pieces either from a recorded song or the sheet. Transcription heavily relies on pitch derived from short segments of the audio. Organization of music collections requires features that describe a complete song. Classification of musical behavior additionally relies, e.g., on mental characterizations. Compositional structure analysis mainly relies on musical properties derived from sheet music.

Indeed, preprocessing of the original time series appears to be the most important step in problem solving. In general, direct modelling in time space is not successful for solving the classification problems in this paper. Therefore, preprocessing is dedicated a whole section, and will also re-appear in later sections.

The following sections describe preprocessing methods for deriving numeric music features suitable for the solution of our classification tasks: short-term, long-term and semantic audio-features as well as compositional features and features related to musical skills and mental mechanisms.

3.1 Audio features

Let us first develop characteristics of audio data adequate for building groups of music pieces which are similar in some sense (like for the organization of music classifications in section 4) or for classification rules for the prediction of certain properties of music (like pitch in section 5). The recorded audio data of polyphonic music is not suited

for direct analysis with data mining algorithms. High quality audio data needs a large amount of memory and contains various sound impressions, such as various instruments, percussion, and singing, that are overlayed in a single (or a few correlated) time series.

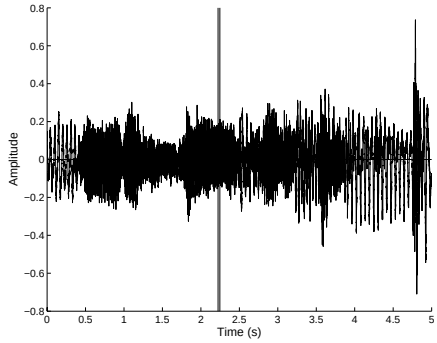
In Figure 1(a) we show the amplitude of the raw audio data from five seconds of a recording of the song *I Shot The Sheriff* by the artist *Bob Marley*. Raw data is given in Wave format (Microsoft Corporation 1991) in CD quality, i.e. with sampling rate 44100 Hertz in 16 bit format (i.e. 2^{16} possible values between -32767 and 32766). The amplitude varies with high frequency around the value zero. At about 2.2 seconds we marked a window of length 23ms that will be used below.

Such time series cannot be compared directly in a meaningful way. A common technique is to describe the sound by extracting audio features at different temporal resolutions. We group the features proposed for recorded audio data in the literature into three categories: *Short-term features* that are extracted from very short time windows during which the sound is assumed to be stationary. These features describe elementary sounds such as a drum beat or a part of a tone played by an instrument for example by measuring the dominant pitch within a few ms of recorded music. Repeating the feature extraction for many consecutive time windows generates a time series of feature values that can be used to detect structure within a musical piece or derive features that describe a longer segment of sound or even a complete song. This builds another type of feature, the so-called *long-term features*. Sometimes it is desirable to understand why a classifier placed a song in a category or what makes songs in a cluster similar to each other. For example when retrieving musical pieces from a large collection, the user might want to emphasize his preference for certain tempo, instruments, or gender of the singing voice. We therefore further distinguish *semantic features*, i.e., features that have an easily understandable interpretation and can thus be utilized directly in end-user applications.

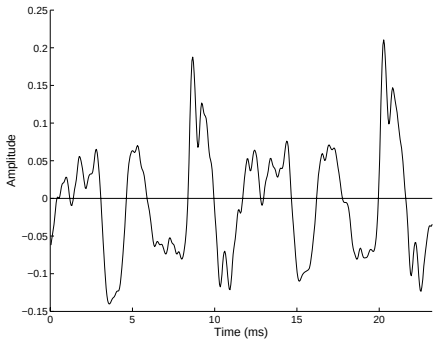
3.1.1 Short-term features.

In this section we survey the plethora of methods that have been proposed to generate numerical features from a short segment of audio data, typically 23-46ms corresponding to a window size of 512 observations at sampling rates of 11025kHz or 22050kHz. An enlarged version of the 23ms window from Figure 1(a) is shown in Figure 1(b). At this scale the time series can be analyzed visually. There seems to be a pattern that is repeating twice. Almost all authors use windows that overlap by 50% as a compromise between temporal resolution and redundancy. If the overlap is large, neighboring frames have very similar sound characteristics leading to redundancy. If the overlap is too small short audio events might not be properly represented if they fall into the border region of neighboring frames.

The most simple short term features are applied in the time domain. The *volume* can be determined by summing up the absolute or squared values of the amplitudes (Li et al 2001). The *low energy* feature is calculated as the percentage of amplitudes whose absolute value is below the root mean square (RMS) of all amplitudes in the window (Tzanetakis and Cook 2002; Mörchen et al 2005b). The *zero-crossing* feature counts the number of times the sign of the amplitudes changes (Li et al 2001). This is correlated with the proportion of high frequencies in the spectrum (Mörchen et al 2005b).

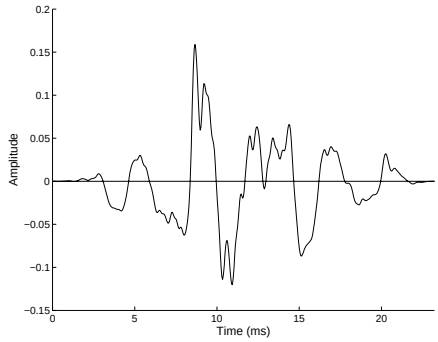


(a) Five seconds of raw audio data. The two neighboring vertical lines mark one window of 23ms used on the right hand side.

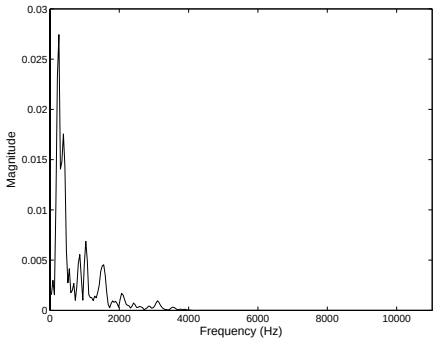


(b) 23ms of raw audio taken from the window marked on the left hand side.

Fig. 1 Raw audio data.



(a) 23ms of raw audio data from Figure 1(b) weighted with a Hann window.

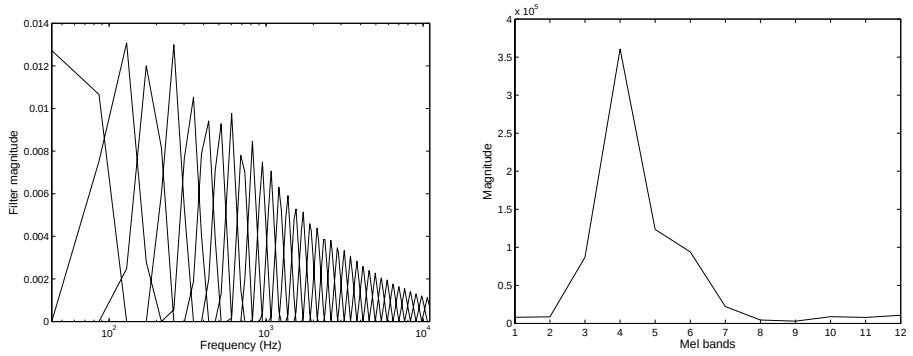


(b) Frequency spectrum calculated with the FFT from the signal in Figure 2(a).

Fig. 2 Hann window and its FFT.

Other frequency related features are based on a preliminary Fast Fourier Transformation (FFT) to obtain a representation of the audio signal in the frequency domain. Such approaches assume that the signal is periodical. To avoid border effects each window is typically weighted with a Hann filter (Oppenheim et al 1999) as shown in Figure 2(a) in order to attain that the values of both borders are zero. The frequency spectrum calculated with the FFT in Figure 2(b) shows the magnitudes of the frequencies that are sampled by the FFT. This can be interpreted as an empirical probability distribution in the frequency space from which other data features are derived.

The frequency spectrum as obtained by the FFT has a rather high resolution of frequency bins and puts equal emphasis on all frequency ranges. Motivated by the success in speech recognition music researchers have used psycho-acoustic transformations of the spectral content (Logan 2000) to summarize frequencies into larger bins (so-called bands) and emphasize frequency ranges that the human ear is most sensitive to. The most popular transformation is the Mel filter bank (Stevens and Volkmann 1940; Rabiner and Juang 1993) consisting of a set of filters with different weights and



(a) Mel filter bands in the frequency space.

(b) Magnitudes of Mel filters applied to spectrum in Figure 2(b).

Fig. 3 Mel filter.

bandwidths each summarizing a different region of the Mel-spectrum

$$\mathrm{Mel}\,(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

where f is the frequency in Hz. Figure 3(a) shows the Mel filters with 50% overlap. Filters in the lower frequency range are much wider and have larger amplitudes. In Figure 3(b) the magnitude of the first 12 Mel bands when applied to the FFT spectrum from Figure 2(b) are displayed. They describe the spectral content of the signal in a more compact way that also takes the human auditory system into account.

Alternative transformation of the frequency spectrum include Bark (Zwicker and Stevens 1957), Equivalent Rectangular Bandwidth (ERB) (Moore and Glasberg 1996), and the well known octave scale, all of which have been used in Mörchen et al (2005b) for audio feature generation.

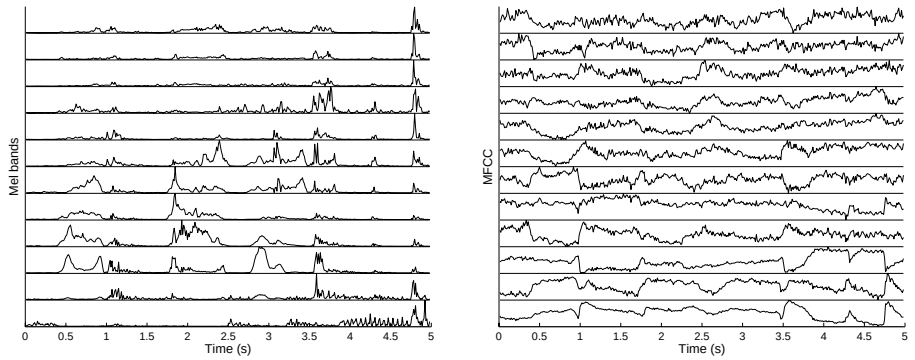
Performing the described steps for many short time windows taken from an audio recording leads to a multivariate time series of Mel magnitudes. Each value of the feature time series describes a short segment of the original audio time series. When windows of length 23ms and 50% overlap of consecutive windows are used the sampling rate of the feature time series is about 85Hz. In Figure 4(a) we show the time series of Mel magnitudes extracted from the five seconds of audio data from Figure 1(a). As can be seen in this example, the amplitudes of neighboring frequency bands are typically highly correlated. Note that Figure 3(b) shows one column of Figure 4(a), i.e. the 12 Mel filters at one time period.

The so-called *Mel Frequency Cepstral Coefficients (MFCC)* are obtained from the Mel spectrum by applying the Discrete Cosine Transform (DCT)

$$y_k = w_k \sum_{n=1}^{N} x_n \cos\frac{\pi(2n-1)(k-1)}{2N},\ k = 1, ..., N, \tag{2}$$

$$w_1 = \frac{1}{\sqrt{N}},\ w_k = \sqrt{\frac{2}{N}},\ k = 2, ..., N,$$

where k is the index of the DCT coefficient and $x = (x_1, ..., x_N)'$ is the N-vector of the logarithms of the amplitudes measured by the Mel band filters.



(a) Magnitudes Mel filters. (b) Magnitudes of MFCC coefficients (cf. (2)).

Fig. 4 Magnitudes of $N = 12$ Mel filters and MFCC coefficients of 23ms windows of data in Figure 1(a).

The resulting series of coefficients is called the *cepstrum* (an anagram of spectrum). Sometimes the Inverse Fourier Transform is used instead of the DCT. The time series of the first 12 MFCC calculated from each spectrum in Figure 4(a) are shown in Figure 4(b), where formula 2 is applied to each column of Figure 4(a). This transformation decorrelates the time series similar to applying principal component analysis but has the advantage to be independent of the data.

The above spectral features are solely based on frequencies and ignore the knowledge about the western musical scale used in many recorded pieces. The *chroma* vector of a spectrum is calculated by summing up the energy of each semitone over all octaves (Wakefield 1999; Goto 2003) indicating the dominance of this pitch. By normalizing the length of this vector to one the volume information can be removed.

3.1.2 Long-term features.

Long-term features describe a long segment of a recorded song or even the complete song. They can directly be used with off-the-shelve supervised and unsupervised algorithms to, e.g., classify music collections. Long-term features can be generated directly from the original audio data or more commonly from a time series of short term features.

This latter approach has been followed by Mörchen et al (2006b, 2005b) who present a systematic study of statistical and time series analysis methods for summarizing short-term feature time series in order to obtain long-term features. First, for each song a multivariate time series with a lot of short-term features was generated. Then, a large set of statistics was combined with each short-term feature to generate a huge number of potentially useful long-term audio features. The features that were best able to separate several manually selected different sounding groups of music were identified in a supervised process and then used for unsupervised classification of other songs. In Mörchen et al (2006a) logistic regression with the lasso method was used to select a small number of features that can collectively describe certain aspects of music well. Many of the statistics used by Mörchen et al had been used before in one way or the other and will be described below.

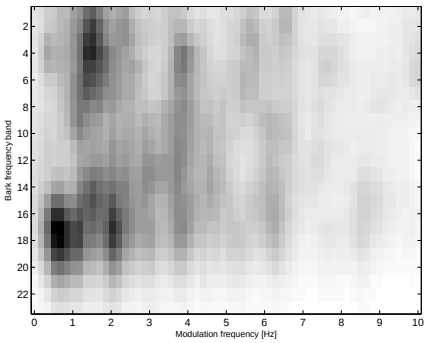
Most authors only use the moments or other descriptive statistics of the empirical probability distribution as a summary of short-term feature time series (e.g. Tzanetakis and Cook 2002; McKinney and Breebaart 2003; Lidy and Rauber 2005). In Mörchen et al (2006b) trimmed versions of the first four moments were used as well to de-emphasize the influence of outlying values that represent only small portions of the song.

A simple way to consider the temporal evolution of the short-term feature values is to use the moments of the 1st and 2nd order differences (e.g. Aucouturier and Pachet 2004). The first use of temporal statistics as long-term features for classification seems to be in Lambrou et al (1998) where entropy and correlation are mentioned but no further details are given. The modulation energy of short-term features was measured in McKinney and Breebaart (2003) by applying the FFT to the feature time series and calculating the energy in three frequency bands motivated by psycho-acoustics: *"1-2Hz (on the order of musical beat rates), 3-15Hz (on the order of speech syllabic rates) and 20-43Hz (in the lower range of modulations contributing to perceptual roughness)"*. In Lidy and Rauber (2005) a histogram with 60 bins for the modulation energy is used. In Mörchen et al (2006b) the autocorrelation function, the cepstrum, and methods from non-linear time series analysis (Kantz and Schreiber 1997) were applied in addition to the modulation spectrum to obtain even more information about the temporal structure of short-term features.

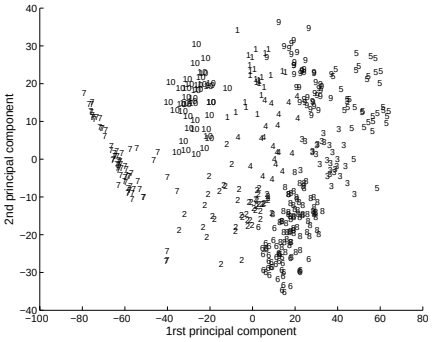
The coefficients of autoregressive (AR) models of univariate short-term feature time series are used in Meng et al (2005) to summarize their temporal structure. This is extended to the case of multivariate autoregressive (MAR) models in Meng (2006) and Meng et al (2006).

The above methods lead to features that individually describe some aspect of the music utilizing the empirical probability distribution of the short-term features or the temporal structure of the time series. In contrast, some authors use 2-dimensional histograms to obtain a vector of features that collectively summarize the short term features. In Pampalk et al (2003a) a histogram indicates how many times one of 50 loudness levels was reached or exceeded in each of 20 frequency bands. The Fluctuation Patterns (FP) (Pampalk et al 2002) measure the magnitude of the modulation energy in 60 frequency bins in time series of 23 frequency bands (see Figure 5(a)). In Li et al (2003) the coefficients of a wavelet transformation of the raw audio signal were summarized with histograms and the moments were used as features.

The composition of short-term and long-term audio features from signal processing and statistical operators was formalized in Mierswa and Morik (2005). Each long-term feature is expressed by an operator tree that is applied to the recorded audio data. Genetic programming (Koza 1992) is used to generate a set of such features that perform well for a specified task. In each iteration only the best performing features are kept. The operator trees are then combined and mutated to generate new, possible even better features. This process is repeated until the improvement from one generation to the next is small. Genre categories and personal taste were used to determine the true class values in this supervised classification study. Some new features based on non-linear time series analysis were found. Genetic programming had previously been used in Pachet and Zils (2003) and Zils and Pachet (2004) to find more general descriptions of acoustic signals. Another supervised approach to generating long-term features is taken in Arenas-Garca et al (2006). Periodograms of short-term MFCC features are summarized with Orthonormalized Partial Least Squares (POPLS) tailored to a specific task, e.g., genre classification.



(a) Fluctuation pattern of audio data in Figure 1(a).



(b) Clustering of MFCC coefficient vectors in Fig. 4(b) by the k-means algorithm with $k = 10$.

Fig. 5 Long-term features.

Often several short-term features are considered collectively leading to a multivariate time series representing a recorded song. Each dimension can be summarized individually with the methods described above, but many authors instead apply methods in the high-dimensional vector space spanned by the individual feature dimensions.

In Logan and Salomon (2001), Berenzweig et al (2004), Pampalk et al (2003b), and Herre et al (2003) vector quantization (VQ) with the k-means clustering algorithm is used to generate a compact summary of the short-term vectors within a single song by few cluster centers. In Figure 5(b) ten clusters obtained by the k-means algorithm are displayed in the 2-dimensional space spanned by the first two principal components of the MFCC data. The audio segment from Figure 1(a) would be represented by the ten cluster centers in the 12-dimensional space of MFCC coefficients.

Self-organizing maps Kohonen (1995) are used for VQ in Vignoli and Pauws (2005). In contrast, Pye (2000), Aucouturier and Pachet (2002), Kulesh et al (2003), Berenzweig et al (2004), and Mandel and Ellis (2005) use a Gaussian Mixture Model (GMM) adapted to short-term data with the Expectation-Maximization algorithm to identify cluster centers.

3.1.3 Semantic features.

We call audio features *semantic* if they describe some more or less commonly and easily understandable properties of music. A Gaussian Mixture Model (GMM) of short term spectra might represent the timbre of a song by a relatively compact model but it is not understandable. The tempo of a song, commonly specified as beats-per-minute (BPM) may be the most widely known semantic music descriptor. It can be estimated from the audio data with signal processing techniques (Scheirer 1998; Tzanetakis et al 2002c; Alonso et al 2003; Gouyon et al 2006). Most semantic features are long-term features as they describe longer segments of recorded sound.

In addition to tempo, rhythm is another important dimension of music that is easily understood by listeners. Many methods for extracting rhythm features are described in Gouyon and Dixon (2005), Gouyon (2005) and the references therein.

The pitch distribution can be summarized with histograms (Tzanetakis et al 2002b) from which the most dominant frequencies can be extracted. The key of the music involves a combination of dominant frequencies, see Gomez (2004, 2006) and references therein. More references on estimating the key, e.g. by analyzing the pitch distribution using histograms, are given in the context of classification of musical structures in section 7 and in the context of transcription algorithms in section 5.4.

Other semantic concepts that have been explored include estimation of danceability (Streich and Herrera 2005), intensity (Sandvold and Herrera 2005), percussiveness (Herrera et al 2004), and complexity (Streich and Herrera 2004). One problem with these approaches is the generation of data for which the true values of such concepts are known.

In Pohle et al (2005) several common audio features are evaluated w.r.t. their ability to express mood, perceived tempo, complexity, emotion, focus, and genre. For each of these concepts supervised classification was performed with different sets of long-term features. This worked well for genre and to some extent also for emotion but the results for the other musical concepts were not satisfying. This suggests that more or different features are needed to classify music according to these semantic aspects.

Characteristics of the register of instruments or voices (such as basso, tenor, alto and soprano) have been analyzed by Weihs et al (2006b) using local probabilistic models for various groups of instruments based on masses and widths of the peaks of so-called pitch-independent periodograms (Weihs and Ligges 2005).

In Berenzweig et al (2003) a supervised approach is taken to model certain aspects of musical sound, e.g., genre or the singer's gender. For each aspect a 2-class feed-forward neural net is trained with a set of short term feature vectors from positive and negative example songs. The output of the model as applied to new songs is interpreted as the strength of this aspect, e.g. a genre, in the music. The resulting feature space is called *Anchor space*. Each song is represented by the high-dimensional distribution of short term feature vectors projected onto this space. The performance for reproducing human judgement of artist similarity is found to be similar to MFCC (Berenzweig et al 2004). Similarly, West et al (2006) map each short sound segment to a vector indicating the likelihood of several genres with regression tree models. A single vector obtained by likelihood smoothing represents a song.

In Mörchen et al (2006a) Bayesian logistic regression (Hastie et al 2001) is directly applied to long term features to obtain a set of semantic descriptors for a complete song in a supervised manner. A very large number of features is generated systematically. A well suited subset of these features is selected automatically via Laplacian priors. Each regression model predicts the likelihood of a genre or some other musical aspect for which true semantic values are available. In Figure 6 the predicted likelihood of belonging to the genre *Metal* is shown for a set of 700 songs (100 from the genre *Metal* and 600 from other genres) that were not used to learn the semantic feature.

Vectors with several such likelihoods can be used to describe a song. Songs from the same genre as used for training of the features will have one dominant entry. Other songs might have a more diverse mixture of likelihoods indicating a musical content that mixes concepts from several different genres. The vectors can be used with off-the-shelve classification and clustering algorithms to build understandable models for music information retrieval.

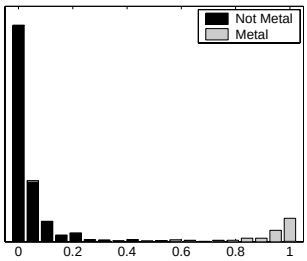


Fig. 6 Values of a semantic feature describing the likelihood of the genre *Metal* in songs not used to train this feature.

3.2 Compositional features

A very different set of features, useful, e.g., for music modelling (see section 7) can be derived from the symbolic representation of music as represented in scores or MIDI files. Such features are expected to characterize compositional structure, i.e. the structures that composers and musicians make conceptual use of when producing music. From the basic information of pitch value and time value (either in absolute time or in metrical time relative to a given meter) a large amount of features can be calculated. These features include the distributions of pitch, pitch interval and duration values over a time window (Müllensiefen and Frieler 2004a). From these distributions relatively simple features can be derived, such as pitch range, extreme pitch values, event density (i.e. number of note events per time unit), leap movement (i.e. the number of all pitch intervals greater than a third divided by the number of all intervals occurring in a monody), or rhythmic variability (defined as the standard deviations of the notated durations), see e.g. Steinbeck (1982); Jürgensen and Knopke (2004); Kranenburg and Backer (2004).

A second class of features are constructed by comparisons between an empirical distribution and a distribution from a representative source like experimentally collected expectation values. E.g., Eerola et al (2002) define *tonal stability* of a melody as the Pearson correlation between a pitch class profile of that melody and the pitch class profile for its major or minor key as derived experimentally by Krumhansl (1990).

Another prominent recent trend in music modelling is to use as the basic musical units not single events like the above mentioned note pitches, intervals, chords or durations, but time related models of longer sequences of such events known as n-grams where n is the length of the sequence (see e.g. Downie 2003). This approach builds on the basic assumption that music is principally produced and perceived in time-ordered sets of events whether these events be notes of a melody or harmonies in a polyphonic piece. The notion that music can be explained, taught, and analyzed as formulae, which specify common or frequent sequences, has been around for several hundred years in music theory, but only the recent availability of large electronic corpora enables to test these hypotheses empirically. n-grams are most helpful for predicting the continuation of a given sequence, be it a melodic sequence of pitches or a sequence of chords. A prerequisite is a model trained on the frequency counts of sequences of a suitable corpus of music. The model is then often used to find the pitch or chord value at position n of a given sequence that has a maximum likelihood given the n-grams of length $n - m$ of the sequence, where m is a model parameter.

3.3 Musical skills and mental mechanisms

Features characterizing musical skills, perceptions, and experience as well as general mental mechanisms and concepts play a major role as predictors in many studies in the field of music cognition and music psychology. They fall broadly into four categories:

- Variables characterizing different aspects of musical activity and experience. Among the common variables in this category are practice hours, years of paid music lessons, time spent listening to music, number of public performances and the like. These variables are often used as predictors in music tasks where performance accuracy or speed is measured.
- Variables characterizing mental skills. These variables are often measured via standardized psychological tests and relate to general cognitive concepts like general intelligence, (working) memory capacity, reaction time in sensorial tasks, handedness (lateralisation), attention span, reading comprehension (see e.g. Kopiez et al 2006).
- Variables characterizing musical skills. In this category fall musicality tests or isolated subtests of these, performance achievements, hearing, singing, and tapping accuracy, perfect pitch, etc. (see e.g. Kopiez et al 2006).
- Variables characterizing perceived musical structure. This category contains e.g. algorithms for estimating the perceived similarity between melodies (see Müllensiefen and Frieler 2004a), or the perceived accent strength of individual notes of a melody (e.g. Thomassen 1982).

4 Classification in Organization of Music Collections

Let us now switch from features to tasks, and start with organization of music collections as motivated and defined in the Introduction.

4.1 Musical similarity

In order to organize music collections with unsupervised classification (clustering) methods (and also for some supervised classification algorithms like k-nearest neighbor) it is necessary not only to have available adequate characteristics (see section 3), but also adequate similarity measures characterizing neighborhood of different pieces of music. Musical similarity is not an objective concept. As noted in Ellis et al (2002) it depends on individual taste, can be described in multiple dimensions, can be asymmetric and context dependent. They propose to use co-occurrence in playlist as one way of obtaining ground truth on similarity. An evaluation of this similarity vs. similarity calculated from the audio content is described in Berenzweig et al (2004). Other authors use artists (Berenzweig et al 2002), genre (Pampalk et al 2003b), and timbre (Aucouturier and Pachet 2004; Mörchen et al 2006b) to group songs into several categories that are considered similar. In (Mörchen et al 2006a) the genre information is obtained from Internet radio stations. The co-occurrence of certain words on the result page of an Internet search is used in Schedl et al (2006) to assign genres to artists.

The recorded version of a song can be represented by a vector of long-term audio features (see section 3.1.2). In this case the so-called *timbre similarity* of different

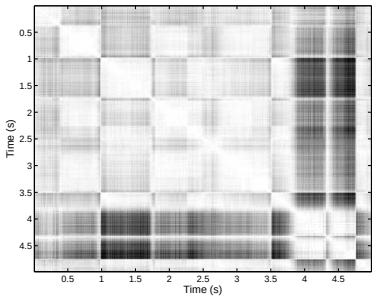


Fig. 7 Similarity matrix of MFCC coefficient vectors in Figure 4(b).

songs can be calculated by vector space distance functions like the L_p distances, the Correlation distance (one minus Pearson correlation), or the Cosine distance

$$d_{\cos}(x, y) = 1 - \frac{x'y}{\sqrt{x'x}\sqrt{y'y}} \tag{3}$$

where $x'y$ is the scalar product of the vectors x and y of long-term audio features.

Care has to be taken to normalize the location and scale of the features, e.g., by subtracting the empirical mean and dividing by the empirical standard deviation. This is in particular important if audio features with different characteristics are combined (Mörchen et al 2006b). If vectors of related features are used like the mean values of all short-term MFCC time series or the histogram as described in section 3.1.1, normalization of individual features is not necessary and may even be harmful, because noise could be amplified. In Pampalk et al (2003b) and Mörchen et al (2006b) distance measures based on vectors of audio features were evaluated on large sets of songs.

Given a multivariate feature time series and a distance function one can calculate the matrix of pairwise distances (or similarities) of all vectors of short-term features in a song (see section 3.1.2). The matrix of pairwise cosine similarities of the MFCC coefficient vectors, i.e. the columns, of Figure 4(b) is shown in Figure 7. One can clearly recognize several segments in this matrix. Bright rectangles around the diagonal indicate homogeneous segments of sound. Bright rectangles off the diagonal connect similar segments that are separated by less similar sounds. In Foote and Uchihashi (2001) and Foote (2002) periodicities are searched by calculating the sum of diagonals in a similarity matrix up to a given time lag, i.e., the autocorrelation of the self-similarity. The Periodicity Histogram (Pampalk et al 2003a) also summarizes the strength of certain beat levels without calculating the self-similarity matrix. In Kurth et al (2006) time scale invariant beat features are derived similar to the chroma for frequencies.

If each song is represented by a collection of clusters (see section 3.1.2), e.g., from a GMM of the short term MFCC time series, vector distances are not applicable. In (Logan and Salomon 2001; Berenzweig et al 2004; Baumann 2003) the Earth Movers Distance (Rubner et al 1998) has been used to compare two recorded songs. This kind of timbre distance is obtained by solving the transportation problem that quantifies the amount of work needed to transform one song representation into the other one based on a distance function between two individual clusters. Each cluster within the song models is described by the mean vector, the covariance matrix and a weight. The symmetric Kullback-Leibler (KL) divergence (Kullback and Leibler 1951; Whittaker 1990) is used for comparing two clusters. Aucouturier and Pachet (2002) use sampling

to directly measure the KL divergence between GMM models obtained from different songs. The pairwise likelihood is estimated by sampling from one model and calculating the likelihood under the other model. A similar technique is used in (Shao et al 2004) for Hidden-Markov Models (HMM).

The k-means and GMM approaches are compared for genre classification under many different parameter settings by Aucouturier and Pachet (2004). Levy and Sandler (2006) compare GMM and VQ models (see section 3.1.2) with various distance functions by quality and efficiency. The KL divergence between GMM models of the MFCC features with a single mixture component and diagonal covariance matrix as used in (Mandel and Ellis 2005) can be efficiently calculated due to the existence of a closed form. The quality when used for genre classification was found to be almost identical to more complicated models and distances. VQ methods are also computationally efficient but were not as effective in the classification. Hidden-Markov Models (HMM) (Rabiner 1989) that incorporate the temporal information of the sequence of feature vectors, are not found to be better than GMM (Aucouturier and Pachet 2004; Flexer et al 2005).

A large-scale evaluation of musical similarity measures on datasets with genre knowledge has been performed by Pampalk (2006b). The best performance was achieved by a somewhat complicated combined measure using the KL divergence between GMM models of the MFCC features with a single mixture component and the full covariance matrix and additional features derived from Fluctuation Patterns (e.g. the center of mass in a Fluctuation Pattern). In addition to accuracy, the author recommends to prefer measures that fulfill the triangle equality more often. Musical similarity measures have further been evaluated with human fixed classes in the Music Information Retrieval eXchange (MIREX) competition (`http://www.music-ir.org/mirex2006`). The two best methods were the above mentioned combination (Pampalk 2006a) and the method of Pohle (2006) that post-processed the KL divergences of GMM to obtain a rank-based distance measure.

4.2 Organization of music collections

Important applications of musical similarity assessment include automated categorization of music by genre, see Tzanetakis and Cook (2002); McKinney and Breebaart (2003); Ahrendt (2006) and references therein. Genre is commonly used to organize a music collection at the highest level but not necessarily available along with the musical pieces. More recently the musical similarity itself is used to explore novel ways of presenting a music collection to the user. Unsupervised classification can group recorded musical pieces into coherent groups according to the similarity measure used. Visualization of songs based on categories and/or similarity of songs and groups can help the user to explore a music collection and help in retrieving pieces of interest.

If information on genre, artist, and album of the songs is available, many scientific visualization techniques can be used to present a music collection to the user. In Torrens et al (2004) disc plots, rectangle plots, and tree maps display the structures of a music collection.

Vignoli et al (2004) display artists on a 2-dimensional map where the axes can be any pair of mood, genre, year, and tempo. The artists are placed such that similar artists are close to each other with a graph drawing algorithm. Self-organizing maps (SOM) (Kohonen 1995) are used to group artists into cells of a 2-dimensional grid

in Knees et al (2004); Vembu and Baumann (2005); Lehwark et al (2007) with text based similarity measures. The text that potentially describes each artist is retrieved from music reviews, Internet search queries, or from the community website *Last.FM*. A limited set of music related words is used to generate feature vectors from the texts using the *term divergence inverse document frequency measure* (Salton and Buckley 1988)

$$tfidf = tf \cdot \log \frac{1}{df} \tag{4}$$

where the *term frequency* (tf) is the frequency of a word within in a single document and the *document frequency* (df) is the frequency of a word over all documents. Sometimes additional normalization terms to account for varying document lengths are used.

This feature weighting is widely used in text mining and based on the intuition that a word is more relevant if it appears frequently in the same document as the artist's name and that a word is generally less discriminative if it appears frequently with many different artists.

In Pampalk et al (2005) this is extended to hierarchical SOM. The MusicRainbow (Pampalk and Goto 2006) is a circular representation of artists. The similarity of artists is calculated from the similarity of the corresponding songs. The representation is color coded by musical style and labelled with information retrieved from the Internet.

At the album level some authors consider manual collaging (Bainbridge et al 2004) of albums, i.e., the user manually orders album covers on a computer screen. Similar to the MusicRainbow similarity of albums could also be determined from the similarity of the individual songs. In general a song-based visualization seems to be preferred. In Cano et al (2002) FastMap and multidimensional scaling are used to create a 2D projection of complex descriptions of songs including audio features. Multidimensional scaling iteratively minimizes the stress function

$$stress = \frac{\sum_{i,j}(\hat{d}_{i,j} - d_{i,j})^2}{\sum_{i,j} d_{i,j}^2} \tag{5}$$

where $d_{i,j}$ is the given original dissimilarity between two objects i and j and $\hat{d}_{i,j}$ is the Euclidean distance of the corresponding representative points in 2D. An initial random configuration can be optimized by gradient descent. In order to avoid the quadratic complexity of MDS the linear FastMap (Faloutsos and Lin 1995) was proposed as an alternative. Each high dimensional point is projected onto 2 orthogonal lines connecting the most dissimilar points in the original data space.

PCA is used in Tzanetakis et al (2002a) to compress audio feature vectors to 3D displays. Pampalk et al (2002) use small SOM trained with song-level features and a density visualization in order to search for possible clusters of songs. In Mörchen et al (2005a); Lehwark et al (2007); Risi et al (2007) the larger Emergent SOM (ESOM) (Ultsch 1993; Ultsch and Mörchen 2005) with distance-based visualization is used.

Small SOM provide results very similar to k-means clustering (Ultsch 1996) as each neuron is typically interpreted as a cluster. The topology preservation of the SOM projection is of little use when using small maps. With larger maps a single neuron does not represent a cluster anymore. It is rather a pixel in a high resolution display of the projection from the high dimensional data space to the low dimensional map space. Clusters are now formed by connected regions of neurons with similar properties. The structure emerges from the large scale cooperation of thousands of neurons during the

Fig. 8 Visualization of music collection with Emergent Self-Organizing map.

ESOM training. Not only global cluster structure is visualized, but also local inner cluster relations are preserved.

In Figure 8 an ESOM of 100 songs from each of seven genres is shown. Each song is described with semantic features (see section 3.1.3, Mörchen et al 2006a) learned from a training set with the same genre structure. The edges and darker regions indicate cluster boundaries whereas the lighter regions indicate clusters. Each small square represents one song. The genre labels were assigned based on the known genres of the songs that were not used to train the map. Songs in the center of a cluster are very typical for a genre. Songs in the area between the labels *Dance*, *Rap*, and *Jazz* seem to mix several genres. In particular the absence of dark regions as found between other genres indicates a soft transition between them.

4.3 Software

The following freely available programs and libraries can be used to generate various audio features for studies in musical similarity:

- Marsyas (Tzanetakis and Cook 2000) can generate a small set of audio features used in one of the most cited publications on genre classification (Tzanetakis and Cook 2002) (`http://marsyas.sf.net`).
- The Matlab Audio Toolbox (Pampalk 2004) can generate MFCC and mixture models thereof as well as most features proposed by (Pampalk 2006b) (`http://www.oefai.at/~elias/ma`).
- MusicMiner (Mörchen et al 2005b) can generate a huge amount of short- and long-term and semantical features that were used in Mörchen et al (2006a) (`http://musicminer.sf.net`). It is based on the free machine learning framework RapidMiner (formerly known as Yale) (Mierswa et al 2006) (`http://www.rapid-i.com`).
- JAudio (McEnnis et al 2005) offers an easy to use interface to select and parameterize many short-term and a few long-term features (`http://coltrane.music.mcgill.ca/ACE/features.html`).
- CLAM (Amatriain et al 2002) is a library that can be used for audio feature generation and many other music information retrieval tasks (`http://www.clam.iua.upf.edu`).

5 Classification in Automatic Transcription

In this section we describe methods for automatic transcription based on audio features. Transcription is transforming audio signals into sheet music, it is in some sense the opposite of playing music from sheet music. The statistical kernel of transcription is classification of notes into classes of pitch (e.g. c, d, ...) and lengths (e.g. dotted eight note, quarter note, ...). A typical transcription algorithm includes at least some of the following steps:

1. Separation of the relevant part of music to be transcribed (e.g. human voice) from other sounds (e.g. piano accompaniment)
2. Estimation of fundamental frequencies
3. Classification of notes, silence and noise
4. Estimation of relative length of notes and meter
5. Estimation of the key
6. Final transcription into sheet music

Note that steps 1 and 2 are again related to pre-processing of the original time series of music. In step 2, time series modelling is used to estimate fundamental frequencies (see sections 5.3 and 5.4.2) which are to be classified into notes afterwards.

Section 5.4 below will be organized along this list of steps and will present more details. For additional information see Klapuri (2004), required signal processing methods are also described in a recent book by Klapuri and Davy (2006). In the following three subsections we will comment on the underlying data and describe the musical and statistical challenges in the transcription task.

5.1 Data

Most existing transcription systems have been invented for the transcription of *MIDI* data (MIDI Manufacturers Association, 2001; both onset times and pitch are already exactly encoded in the data) or for instruments such as piano and other plucked string or percussion instruments.

The transcription of MIDI data is not that difficult, because information related to pitch as well as the begin and end of tones is already explicitly available within the data in digital form. Therefore, this information has not to be estimated from the sound signal. Transcription of plucked and stroked instruments (piano, guitar, etc.) is harder than transcription of MIDI data, but still simpler than, e.g., the human voice. Sudden increases of the signal's amplitude point to new tones for these instruments, which may not be the case for other types of instruments like flute, violin or the human voice to which the algorithm should be applicable to as well.

Typically, the sound that has to be described is given in form of a Wave file (Microsoft Corporation 1991), typically in CD quality with sampling rate 44100 Hertz and in 16–bit format (i.e. 2^{16} possible values).

5.2 Musical Challenges: Partials, Vibrato and Noise

If a tone is played or sung, it commonly does not only produce a single (co)sine wave oscillating with the fundamental frequency but also waves oscillating with integer multiples of the fundamental frequency. These waves are called partials (of the whole

tone). One problem for transcription algorithms is the possible (almost) absence of the fundamental while some of the other partials are well observable.

It is particularly interesting to automatically transcribe one of the most complex musical instruments: the human voice. The human voice can adjust loudness and many properties of the sound very easily within one single tone. Indeed, the sound characterization of the human voice has many more facets than for instruments because the sound is varying in dependence of technical and emotional expression (Wapnick and Ekholm 1997; Kleber 2002). Hence robustness against such variations is very important for the design of transcription systems.

Another problem for transcription algorithms is the presence of vibrato, some kind of intended or unintended adornment. The loudness of a singer's vibrato varies about 2-3 Dezibel while the pitch varies around one semitone (Seidner and Wendler 1997) up to two semitones (Meyer 1995) around the desired pitch of the tone. The vibrato frequency is roughly 5-7 Hertz. Models and detection methods for vibrato have been described, for example, by Rossignol et al (1999a) and Pang and Yoon (2005). The strong vibrato of a professional soprano singer performing the German Christmas song 'Tochter Zion' (G.F. Händel) is shown by the nervously changing line of fundamental frequencies (the lower dark curve) in the spectrum given in Figure 9(a).

A third problem is the presence of noise in the signal. Polotti and Evangelista (2000) model phenomena like pink noise which is the most common kind of noise in audio recordings.

5.3 Statistical Challenge: Piecewise Local Stationarity

For most methods in time series analysis, both in time and in frequency domain, at least some weak stationarity assumptions of the underlying process have to be valid. Unfortunately, even if processes of musical time series might be stationary in the mean, they are not stationary w.r.t. covariance, because the tones (and hence the covariances) are changing quite frequently.

Dahlhaus (1997) defines *locally stationary processes*, which implies stationarity in some ϵ-region around a point in time. This definition is fine for minor changes in pitch, such as for vibrato or other decorations of the tone. It is not sufficient for abrupt changes between one note and another one, because amplitude and pitch might change suddenly. Adak (1998) developed an algorithm for the segmentation of time series and defines *piecewise local stationary processes* as finite series of locally stationary processes by generalizing the definition of Dahlhaus (1997). The new definition is very useful for music time series: for n tones (corresponding to a series of n locally stationary processes), we expect to find at least $n-1$ change points (changes from vowels to consonants within the same tone might lead to change points as well). Most algorithms used in transcription apply Short Time Fourier Transformation (STFT), i.e. calculate periodograms of very small pieces (e.g. 23-46ms, see section 3.1.1) corresponding to windows (mostly overlapping by 50%) of the time series in order to detect those change points and estimate fundamental frequencies.

The SLEX (Smooth Localized Complex Exponential) transformation by Ombao et al (2001) can segment bivariate non-stationary time series into almost stationary segments and it can be flexibly adapted to different time and frequency resolutions. For other related time series methods in frequency domain see also Bloomfield (2000), Brillinger (1975), and particularly for signal analysis, see Van Trees (2001).

5.4 Transcription Algorithm

A sequence of steps for a transcription process was listed at the beginning of section 5 and can be understood as steps from local to global analysis of a music time series. This aspect is further discussed in Weihs and Ligges (2005).

Some complete transcription program has been described by Pressing and Lawrence (1993). Unfortunately, their method seems to be unavailable.

In other projects, researchers try to trace online what performers are playing or singing compared to given notes. For this purpose, Cano et al (1999) use Hidden Markov Models. Raphael (2001) developed an expert system based on Bayes Belief Networks that tries to accompany a singing or instrument playing person based on known sheet music by tracing the notes. These methods have to be extremely fast, because online calculation is mandatory, but they can rely on valuable a-priori knowledge: already known notes.

Let us now go through the steps of automatic transcription as indicated in the beginning of section 5.

5.4.1 Separation of the relevant part of music.

As a first step of the transcription algorithm, the relevant part of music to be transcribed (e.g. human voice) has to be separated from other sounds (e.g. piano accompaniment). The outcome of a separation is a time series of one relevant part of the music. To achieve this Sound Source Separation task, the commonly used standard method is Independent Component Analysis (ICA) as proposed by Hyvärinen et al (2001). Some disadvantages of ICA have been shown by von Ameln (2001). Klapuri (2001) uses the Spectral Smoothness method for both separation and polyphonic fundamental frequency estimation. Another method for Sound Source Separation has been proposed by Viste and Evangelista (2001, 2002). They aim at audio coding and compression for formats like MPEG 3 (Brandenburg and Popp 2000), or integration into hearing aids.

5.4.2 Estimation of Fundamental Frequency.

After having separated the relevant part of music, we have to determine the fundamental frequency f_0 (see section 5.2). Many approaches for the estimation of fundamental frequency, also known as f_0 estimation, for both monophonic and polyphonic sound have been published. Goto (2004) proposes a method called *PreFEst* for the 'predominant f_0 estimation' of melody and bass lines without requiring assumptions about the number of sound sources. Dixon (1996) describes a heuristic method for the identification of notes and Klapuri (2001) describes some method for polyphonic estimation of fundamental frequencies. Smaragdis and Brown (2003) are extending the Fast Fourier Transformation (FFT) by 'Non-Negative Matrix Factorization' for polyphonic transcription. Bayes methods for the f_0 estimation of monophonic and polyphonic sound have been proposed by Walmsley et al (1999), Davy and Godsill (2002), and again Godsill and Davy (2003). A rather theoretical work by Wolfe et al (2004) introduces Bayesian variable selection for spectrum estimation. In the MAMI project (Musical Audio-Mining, see Lesaffre et al (2003)), software for the fundamental frequency estimation has been developed.

Plumbley (2003) proposes 'Algorithms for Nonnegative Independent Component Analysis' (N-ICA) in order to extract features of polyphonic sound, but applies it

only to sound generated by MIDI instruments. Moreover, Plumbley (2004) suggests optimization using Fourier expansion for N-ICA and expresses his hope to be able to extend the method to perform well for regular ICA. In another work, Plumbley et al (2006) propose to use dictionaries of sounds, i.e. databases that contain many tones of different instruments played in different pitches. Using such dictionaries might overcome the problem that different tones containing a lot of partials may not be identifiable for polyphonic problems.

Under some circumstances the frequency of partials is slightly shifted from the expected value. This is a problem for the polyphonic case, if a partial's frequency cannot be assigned to a corresponding fundamental frequency. Hence this phenomenon has to be modelled as done in some recent work by Godsill and Davy (2005).

Polotti and Evangelista (2000) modelled phenomena like pink noise (noise decreasing with frequency; also known as $1/f$ noise) using Wavelet techniques. Later on, Polotti and Evangelista (2001) also modelled other special kinds of unwanted noise or the sound of consonants that do not sound with a well defined fundamental frequency. A more general article about Wavelet analyses of music time series has been written by Evangelista (2001).

Ligges (2006) and Weihs et al (2006a) propose to use a model for fundamental frequency estimation that combines the models of Davy and Godsill (2002) and Rossignol et al (1999a). The first model (Davy and Godsill 2002) includes parameters for phase displacement, frequency displacement of partials, and trigonometric basis functions that model changes in amplitude. The second model (Rossignol et al 1999a) covers vibrato using a sine wave around the 'average audible' frequencies and their partials.

The aim is to model well known physical characteristics of the sound in order to estimate f_0 independently of other relevant factors that might influence estimation. Proposed methods to estimate the model are non-linear optimization of an error criterion such as the MSE distance between the real signal and the signal generated from the model after a transformation of the signals to the frequency domain, and Markov Chain Monte Carlo methods (MCMC, being computationally very intensive).

The fundamental frequencies can be estimated much faster by using a heuristical approach as proposed in, e.g., Weihs and Ligges (2006). In this approach several thresholds are applied to values of the periodograms derived by STFT from the original time series of the music so that the peak representing the fundamental frequency is estimated. The fundamental frequency λ can be estimated by weighting the frequencies λ^* and λ^{**} of the two strongest Fourier frequencies' values $P(\lambda^*)$ (strongest) and $P(\lambda^{**})$ (second strongest) of that peak:

$$\hat{\lambda} := \lambda^* + \frac{\lambda^{**} - \lambda^*}{2} \cdot \sqrt{\frac{P(\lambda^{**})}{P(\lambda^*)}}.$$

5.4.3 Classification of notes, silence and noise.

While it seems to be plausible to segment tones at first and to assign them to notes afterwards, this was found to be less useful in real applications with singing performances and a joint procedure has been proposed by (Ligges 2006), where the classification into notes takes place at first by classifying a tone to the note with minimal (Euclidean) distance in cents of halftones. Afterwards a running median step is applied to the time series of notes in order to smooth it. Finally, the segments are the constant parts of

the time series of smoothed notes, i.e. each change in the pitch of the smoothed notes implies a new segment.

In section 5.3, we already mentioned the SLEX (Ombao et al 2001) procedure and a segmentation algorithm for speech by Adak (1998). The segmentation of sound has also been examined by Rossignol et al (1999b).

5.4.4 Estimation of relative length of notes and meter.

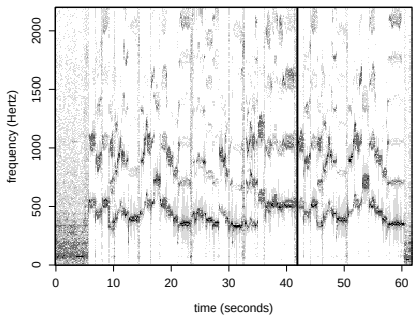
After the segmentation of notes we have to quantize the notes, i.e. to estimate relative lengths of notes. Müllensiefen and Frieler (2004b) define quantized melodies as '[...] melodies where the durations are integer multiples of a smallest time unit T'. For now, we assume that the tempo is fixed throughout a song. An obvious idea is to look for the least common multiple of the divisors of all note lengths to get the smallest T. For example, if there are quavers (length $\frac{1}{8}$), punctuated quavers $\left(\frac{1}{8}+\frac{1}{16}\right)$, quarters and half notes, the searched divisor is $\frac{1}{16}$.

Unfortunately, quite large inaccuracies have to be expected in real data, because humans tend to (it is unessential if intentionally or not) start with notes too late and finish the notes too early (e.g. in order to breathe when singing). Hence quantization has to be very robust against such inaccuracies. Beran (2004) has analyzed (intentional) variations of the tempo by famous pianists. He explains that besides inexact length of notes changes of tempo have to be expected as well.

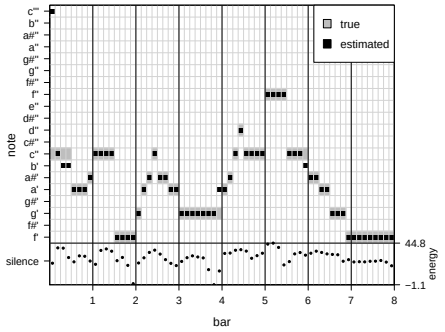
Most published methods are using sudden changes of the amplitude in order to track the tempo, segment the music and perform the quantization. One of these methods has been described by Cemgil et al (2000) and was extended later by (Cemgil et al 2001) in order to take care of dynamic changes of the tempo during time. Alternatively, Cemgil and Kappen (2003) propose some Monte Carlo methods for tempo tracking and Whiteley et al (2006) are using Bayesian models of temporal structures. Davies and Plumbley (2004) try to adapt the quantization to dynamic tempo changes. The perceptual smoothness of tempo in expressively performed music is analyzed by Dixon et al (2006). For more general findings on extracting tempo and other semantic features from the audio data with signal processing techniques, see section 3.1.3.

After pitch estimation, note classification and quantization, the information that has been derived can be presented in some picture like the one given in Figure 9(b), which shows the outcome of analyzing the last 8 bars of the German Christmas song 'Tochter Zion' (G.F. Händel) performed by a professional soprano singer. The 'real' sheet music has been translated to the grey shading (each segment corresponds to an eighth note), black squares indicate the estimated note, and at the bottom an energy bar indicates the loudness.

After a successful quantization, the meter has to be estimated. This is one of the most difficult tasks, and we do not know any method that is capable of it in a general manner. This is not a big surprise, because even humans cannot always distinguish between, for example, $\frac{2}{4}$, $\frac{4}{4}$ and $\frac{4}{8}$ meters. Most of the time, it is, thus, assumed that the meter is given by the algorithms supervisor. A rough distinction between $\frac{4}{4}$, and $\frac{3}{4}$ meters was proposed, e.g., in Weihs and Ligges (2005) by means of the number of quarters between so-called accentuation events.



(a) Spectrum, strong vibrato in sound performed by a professional singer. The vertical line indicates start of last 8 bars as shown in Fig. 9(b).



(b) Collected information during a transcription procedure.

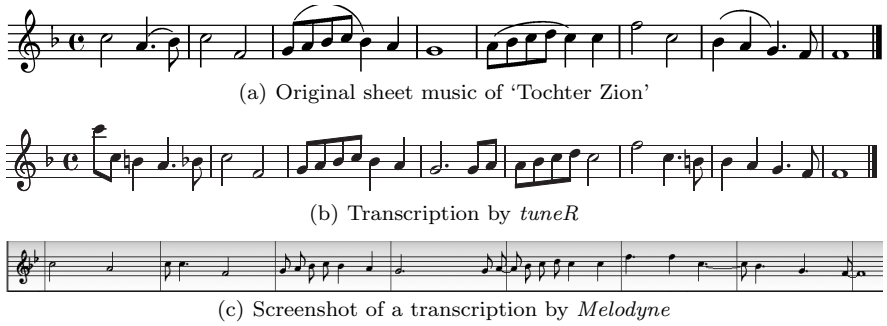
Fig. 9 Representation of a song.

5.4.5 Estimation of the key.

The basic idea for key estimation (Brown et al 1994) is as follows. All notes from some piece of music can be tabulated within some table. Depending on the frequencies of the twelve different notes (including halftones), the most probable key can be estimated. Bayesian modelling can also be used for key estimation. David Temperley (e.g. Temperley 2004, 2006) proposes such a model for a given piece or segment of music (cp. section 7). Here, the probability computation is based not only on the relative frequency with which the twelve scale degrees appear in a key, but also on the probability of a segment being in the same key as the previous segment in the same piece (probability of modulation). Other more sophisticated approaches would also analyze the sequence of tones and chords. More references on estimating the key, e.g. by analysis of the pitch distribution using histograms, are given in section 3.1.3 and in the context of classification of musical structures in section 7.

5.4.6 Final transcription into sheet music.

In the preceding sections we have described how to estimate properties of the sound that are required for the transcription of sound to sheet music. The final part of producing the sheet music is a matter of music notation and score printing. A free and powerful software for music notation is LilyPond (Nienhuys et al 2005) which uses LaTeX (Lamport 1994), the well known enhancement of TeX (Knuth 1984). Beside sheet music, LilyPond is also capable of generating MIDI files. Therefore it is possible to audit results of transcription both visually and audibly. The R package *tuneR* (see the following paragraph) contains a function which implements an interface from the statistical programming language to LilyPond.



(a) Original sheet music of 'Tochter Zion'

(b) Transcription by *tuneR*

(c) Screenshot of a transcription by *Melodyne*

Fig. 10 Comparison of transcriptions: original sheet music vs. results by *tuneR* and *Melodyne*.

5.4.7 Software for Transcription.

The freely available R (R Development Core Team 2007) package *tuneR* (Ligges 2006) is a framework for statistical analysis and transcription of music time series which provides many tools (e.g. for reading Wave files, estimating fundamental frequencies, etc.) in form of R functions. Therefore, it is highly flexible, extendable and allows experimenting and playing around with various methods and algorithms for the different steps of the transcription procedure. A drawback is that knowledge of the statistical programming language R is required, because it does not provide transcription on a single key press nor any graphical user interface – as opposed to commercial products. The outcome of a transcription of 8 bars of 'Tocher Zion' is given in Figure 10(b). For comparison, the original notes of that part of 'Tocher Zion' are shown in Figure 10(a).

Finally, we present the well known commercial software product *Melodyne* (version 3.2, `http://www.celemony.com`), which currently is the best commercial transcription software we tried out. It performs all the steps required by a full featured transcription software, including key and tempo estimation. Its recognition performance is quite good even with default settings on sound that has been produced even by human voices. Some parameters can be tuned in order to improve recognition performance. Figure 10(c) shows a screenshot of Melodyne. This is of the same quality of performance as the outcome of *tuneR* in Figure 10(b).

For both *Melodyne* and *tuneR* we have optimized the quantization by specifying the number of bars and the speed. The quality of the final transcriptions is quite comparable. The software *tuneR* produces more 'nervous' results. At some places additional notes have been inserted where the singer slides smoothly from one note to another. The first note is estimated one octave too high due to an immensely strong second partial in almost absence of any other partials. *Melodyne* omits some notes. Here we guess that *Melodyne* smooths the results too much and even detects smooth transitions of the singer even if the singer intended to sing a separate note.

6 Classification of Musical Behavior

In the next two sections we will discuss recent trendsetting papers that serve as application examples for supervised and unsupervised classification in cognitive psychology of music and compositional structure analysis.

Cognitive psychology of music traditionally aims at explaining music perception or music-related behavior in terms of mental mechanisms or concepts, such as memory, emotions, mental skills, or mental representations, as discussed in section 3.3. We concentrate here on the description of models from the music psychology literature which use classification models to relate measurements of music related human behavior to musical features or to other psychological factors.

Within the recent literature, we find studies that adopt a data mining approach in order to arrive at the best prediction or classification result. For example, Kopiez et al (2006) predict achievements in a sight-reading task by various methods (linear discriminant analysis, linear regression, regression and classification trees). Sight-reading is the ability to play music unrehearsed from score notation. Depending on the difficulty of the music and the ability of the sight-reader renditions of the music are almost always imperfect and, hence, the relative number of notes matching between the score and the rendition served as the dependent variable indicating the performance accuracy. The predictors used by Kopiez et al (2006) included general and elementary cognitive skills, mostly measured by standard psychological tests, as well as practice-related skills and time invested in various musical activities. Motivated by the high amount of 35% unexplained variance in a linear regression model (Kopiez et al 2006) tried classification methods to distinguish low and high performers among their 52 participants. In this application linear discriminant analysis achieved better classification results than classification trees constructed by the CART algorithm (Classification and Regression Trees, Breiman et al 1984) and yielded an acceptable cross-validated classification error of 15% . Among the important predictors that characterize good sight-reading performers are their manual speed at playing trills, mental speed as indicated by a number connection test, and the practice time for sight-reading before the age of 15.

Similar to (Kopiez et al 2006), (Müllensiefen and Hennig 2006) attempt to explain participants' noisy responses in a music memory task by taking a data mining approach and applying a number of different techniques, including classification and regression trees, random forests, linear and ordinal regression, and k-nearest neighbor. In this study the participants' task consisted in spotting differences between a target melody in a musical context and an isolated comparison melody. The study aims at identifying the factors that determine recognition memory for new melodies and tunes. The overall similarity of the melodies and the similarity of their accent structures as well as the musical activity of the participants are identified as the most important predictors in relation to the participants' judgements.

Another important task in music perception is to model the recognition of different instruments playing simultaneously in order to identify (classify) them. To this end, Röver et al (2005) extracted certain instrument-specific long term sound features from the time series representing a given recorded sound. A specific Hough Transform (derived from Shapiro 1978) to detect so-called signal edges, i.e. ascending sections of the music time series, is applied to sounds played by different musical instruments. Other investigated instrument-specific features include characteristics of the amplitude and frequency distributions. Several classification methods are tried out to distinguish between the instruments and it turns out that Regularized Discriminant Analysis (RDA, a hybrid method combining Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), Friedman 1989) performs best. The resulting error rate is better than those achieved by humans (Bruderer 2003).

Questions concerning the variability of intonation perception and intonation judgements are generally a topic of research in the field of psycho-acoustics, but as Weihs

et al (2001) showed they also relate to the psychology of music. Weihs et al (2001) investigate how judgements of a human expert correspond to an objective criterion derived from audio in the case of singing. Taking the expert's judgements as the true class values the following classification rule was found: the maximal permissible error which is accepted in order to classify a tone as 'correct' is about 0.4 semitones below and above the target tone. Note that this is much higher than the 0.03 semitones being the minimal distinguishable distance for the trained ear for isolated tones (Pierce 1992). This, at least at first sight, irritating result is probably caused by the fact that the expert judged the correctness of the intonation of the notes being part of a sung melody, i.e. in the context of other tones and under time constraints.

In contrast to the supervised classification approach, unsupervised classification (clustering) has been widely used in cognitive music psychology for some time. In many cases, data reduction and scaling techniques have been used for exploratory purposes and graphical display. Multidimensional Scaling (see Section 4.2) in particular has been often employed in music psychology for making distances in cognitive judgement space visible in low-dimensional Euclidean space and allow for post-hoc grouping and classification of musical objects or experimental participants. For example, Gromko (1993) scales the rated differences between short audio examples from different classical composers onto a three dimensional space to investigate the criteria of the stylistic judgements of classical music novices and experts. She interprets the three dimensions as representing the secondary musical dimensions of activity, character, and pleasingness to the listener. The individual reliance upon these three dimensions is then used to characterize the two classes of music listeners.

MDS procedures have also been widely used to group musical objects in an unsupervised way. Eerola et al (2002) collected global pair-wise similarity judgements between folk melodies of different origin (e.g. church hymns, Finish Yoik songs, German folk songs). Displaying the resulting MDS solution graphically in two-dimensional space yielded a grouping of the melodies which to a certain extent reflected their origin, but also with several melodies being placed close to tunes from a different origin. The subsequent multiple regression analysis that Eerola et al perform aimed at explaining the global similarity judgements with the aid of music structural features such as the distribution of note intervals and durations, tonal stability, and rhythmic variability (cp. section 3.2).

Conceptually similar, but dealing with psychoacoustic questions of timbre perception, Markuse and Schneider (1996) start off with an MDS solution and a grouping of renditions of short excerpts from R. Wagner's opera *Tristan und Isolde* with the instrumental timbres differing in form and complexity. Again, global subjective similarity judgements were collected experimentally. The authors find a fair agreement between the grouping of the excerpts in MDS space and the two manipulated timbral parameters and conclude that the number of partials corresponds to perceived complexity and the waveform is reflected by some kind of perceived roughness.

7 Classification of Musical Structures

Analyzing and modelling music data in the form of predicting and classifying compositional structures has become increasingly popular in recent times, due largely to the increasing amount of music that is digitally available in a symbolic encoding form.

One challenging task in this area is often termed as *author attribution* and consists in associating pieces by anonymous composers with the name of the most likely composer according to stylistic criteria. This task is highly relevant for medieval, renaissance, and even baroque music where large collections of pieces exist in manuscripts and printed compilations, but frequently without any indication of the individual composers. A recent example is the work by (Jürgensen and Knopke 2004) where the Buxheim Organ Book serves as a training collection for selecting useful features for the author attribution task for fifteenth century music. In this paper a vast number of features is defined to characterize each piece. To first reduce the feature space, a principal component analysis is performed and the pieces are projected onto a two dimensional plane. Unfortunately, the grouping of the pieces appeared to be little informative but the analysis motivated the elimination of many features with high inter-feature correlations from the initial feature set. As a second step, scatterplots for smaller feature sets were inspected by eye leading to some preliminary results.

As one recent approach for the identification and classification of higher level structures Bayesian modelling has grown in popularity (e.g. Temperley 2004, 2006, 2007). Temperley's Bayesian model for key estimation in a given piece or segment of music (Temperley 2004) (cp. section 5.4), finds for a sequence of musical segments a corresponding sequence of keys with a maximum probability given an empirical music surface (here: a pattern of pitches). Temperley's model takes advantage of the relation between pitch class distributions and keys learned from an empirical corpus and from experimental data. He also introduces a probability for the modulation from one segment to the next. Tested against correlation based procedures, like the widely used Krumhansl-Schmuckler key finding algorithm (Krumhansl 1990) and Temperley's own modification of this algorithm (see Temperley 2001), the Bayesian model achieves about equal success rates. Therefore, it remains to be seen what the original Bayesian contribution to these kinds of musical models will be, considering that frequency counts on musical elements have been successfully used as predictors in non-Bayesian models previously (e.g. Eerola et al 2002; Costa et al 2004). Apart from Temperley's work on key finding and the Krumhansl-Schmuckler algorithm, (Chuan and Chew 2005) propose an algorithm that is based on the so-called spiral array, a helix representation of pitch and harmonic relationships proposed by (Chew 2000). Performing a nearest neighbor search, this algorithm finds the key that is geometrically closest to a set of pitch class strength values represented on the spiral array. Tested on renditions of Mozart's Symphonies Chuan's and Chew's algorithm outperforms the correlations based methods, but future evaluations are needed here.

Another prominent, recent trend in music modelling is the use of n-gram models for the prediction and classification of musical events in a sequence. Here, the aim is often to predict the next element, i.e. a pitch class or duration value, based on knowledge of similar sequences. A sophisticated example of the n-gram approach is the work of Pearce and Wiggins (e.g. Pearce and Wiggins 2004, 2006), that is concerned with the occurrence of melodic n-grams. Their research hypothesis is that many aspects of musical expectation are acquired through the spontaneous induction of sequential regularities in the music we are exposed to. Although the basic idea of their modelling approach is straightforward and consists in counting the occurrences of melodic n-grams in a corpus and returning the most probable continuation of chain of $n-1$ notes as a prediction for a given context, examining several model parameters Pearce and Wiggins (2004) outperform a competing model by Narmour (1990) and Schellenberg (1997) based on principles from Gestalt psychology. Moreover, the reasoning of Pearce

and Wiggins (2006) regarding model selection can serve as a guideline for other studies concerned with the comparison of models for music cognition. They not only consider data fit as a model selection criterion, but also scope (the model's failure to predict random data), and simplicity (the number of prior assumptions and principles the model builds upon).

8 Conclusions

In this paper typical solutions for classification tasks related to music research were presented mainly in the fields (1) organization of music collections, (2) transcription of music signals, (3) cognitive psychology of music, and (4) compositional structure analysis. After some historic reminiscences, we have surveyed the features that underly the classification tasks. Some explanatory features are shared between the discussed application areas. Short term audio features are utilized in both (1) and (2), long term audio features in (1) and (3), semantic features in (1), (2) and (4). Other types of features are special to one application or one field like compositional features for (4), and musical skills and mental mechanisms for (3).

As presented in section 4 the automatic comparison of music based on representations and similarities derived purely or mainly from the content of digital audio files has received a lot of attention in the research community. Important applications include the automated categorization and recommendation of music for online sales. The individual user will also benefit from the achievements if the techniques are incorporated into software that organizes a personal music collection and helps in generating playlists.

As presented in section 5 automatic transcription is transforming audio signals into sheet music, i.e. it is in a way the opposite of playing music from sheet music. It is demonstrated that for monophonic sound data transcription results are quite acceptable. Note that automatic transcription is a field where ideas of both audio features and compositional structure is needed in that, e.g., the key has to be classified.

In section 6 prediction/classification of musical behavior is analyzed. Human behavior related to music is such a very broad field, though, that it could only be exemplarily covered in such an overview. In section 7 statistical classification of musical structure helps to learn rules that describe similarities in music compositions. Here, research is only in the beginnings because only since recently the amount of digitally available symbolically encoded music is large enough.

Overall, we demonstrated that the number of papers in the fields has been exploded in the last years. Nevertheless, the results are far from being perfect so that even now there is urgent need for new ideas.

Acknowledgements The work of Claus Weihs has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475. Daniel Müllensiefen is funded by EPSRC grant EP/D038855/1. We thank three unknown referees and the editors for their valuable comments.

References

Adak S (1998) Time-dependent spectral analysis of nonstationary time series. Journal of the American Statistical Association 93:1488–1501

Ahrendt P (2006) Music genre classification systems - a computational approach. PhD thesis, Technical University of Denmark, DTU

Alonso M, David B, Richard G (2003) A study of tempo tracking algorithms from polyphonic music signals. In: Proceedings of the 4th COST 276 Workshop, Information and Knowledge Management for Integrated Media Communication, Bordeaux, France, pp 1–5

Amatriain X, Arumi P, Ramirez M (2002) CLAM, yet another library for audio and music processing? In: Proceedings of the 17th Annual ACM Conference on Object-Oriented Programming, Systems, Languages and Applications, ACM Press, Seattle, WA, USA, pp 46–47

von Ameln F (2001) Blind source separation in der Praxis. Diplomarbeit, Fachbereich Statistik, Universität Dortmund, Dortmund, Germany

Arenas-Garca J, Larsen J, Hansen LK, Meng A (2006) Optimal filtering of dynamics in short-time features for music organization. In: Proceedings of the 7th International Conference on Music Information Retrieval, Victoria, Canada, pp 290–295

Aucouturier JJ, Pachet F (2002) Finding songs that sound the same. In: Proceedings of the IEEE Benelux Workshop on Model based Processing and Coding of Audio, Leuven, Belgium, pp 1–8

Aucouturier JJ, Pachet F (2004) Improving timbre similarity: How high is the sky? Journal of Negative Results in Speech and Audio Sciences 1(1):1–13

Bainbridge D, Cunningham SJ, Downie JS (2004) Visual collaging of music in a digital library. In: Proceedings of the 5th International Conference on Music Information Retrieval, pp 397–402

Baumann S (2003) Music similarity analysis in a P2P environment. In: Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services, London, UK, pp 314–319

Beran J (2004) Statistics in Musicology. Chapman & Hall/CRC, Boca Raton

Berenzweig A, Ellis D, Lawrence S (2002) Using voice segments to improve artist classification of music. In: Proceedings of the 22nd International AES Conference, Espoo, Finland, pp 119–122

Berenzweig A, Ellis D, Lawrence S (2003) Anchor space for classification and similarity measurement of music. In: Proceedings of the IEEE International Conference on Multimedia and Expo, pp I–29–32

Berenzweig A, Logan B, Ellis D, Whitman B (2004) A large-scale evaluation of acoustic and subjective music-similarity measures. Computer Music Journal 28(2):63–76

Bloomfield P (2000) Fourier Analyis of Time Series – An Introduction, 2nd edn. John Wiley and Sons, New York

Brandenburg K, Popp H (2000) An introduction to MPEG Layer 3. EBU Technical review

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and Regression Trees. Wadsworth, Belmont (CA)

Brillinger D (1975) Time Series: Data Analysis and Theory. Holt, Rinehart & Winston Inc., New York

Brown H, Butler D, Jones M (1994) Musical and temporal influences on key discovery. Music Perception 11(4):371–407

Bruderer M (2003) Automatic recognition of musical instruments. Master thesis, Ecole Polytechnique Fédérale de Lausanne

Cano P, Loscos A, Bonada J (1999) Score-performance matching using HMMs. In: Proceedings of the International Computer Music Conference, Beijing, China, pp 441–444

Cano P, Kaltenbrunner M, Gouyon F, Battle E (2002) On the use of FastMap for audio retrieval and browsing. In: Proceedings of the 3rd International Conference on Music Information Retrieval, Paris, France, pp 275–276

Cemgil A, Kappen B (2003) Monte Carlo methods for tempo tracking and rhythm quantization. Journal of Artificial Intelligence Research 18:45–81

Cemgil A, Kappen B, Desain P, Honing H (2001) On tempo tracking: Tempogram representation and Kalman filtering. Journal of New Music Research 29(4):259–273

Cemgil T, Desain P, Kappen B (2000) Rhythm quantization for transcription. Computer Music Journal 24(2):60–76

Chew E (2000) Towards a mathematical model of tonality. PhD thesis, Department of Operaitons Research, MIT, Cambridge, Massachusetts

Chuan CH, Chew E (2005) Audio key finding: Considerations in system design, and the selecting and evaluating of solutions. In: International Conference on Multimedia and Expo

(ICME), pp 21–24
Costa M, Fine P, Ricci Bitti PE (2004) Interval distribution, mode, and tonal strength of melodies as predictors of perceived emotion. Music Perception 22(1):1–14
Dahlhaus R (1997) Fitting time series models to nonstationary processes. The Annals of Statistics 25:1–37
Davies M, Plumbley M (2004) Causal tempo tracking of audio. In: Proceedings of the 5th International Conference on Music Information Retrieval, Audiovisual Institute, Universitat Pompeu Fabra, Barcelona, Spain, pp 164–169
Davy M, Godsill S (2002) Bayesian harmonic models for musical pitch estimation and analysis. Technical Report 431, Cambridge University Engineering Department, Cambridge
Dixon S (1996) Multiphonic note identification. Australian Computer Science Communications 17(1):318–323
Dixon S, Goebl W, Cambouropoulos E (2006) Perceptual smoothness of tempo in expressively performed music. Music Perception 23(3):195–214
Downie JS (2003) Evaluating a simple approach to music information retrieval. evaluating a simple approach to music information retrieval. conceiving melodic n-grams as text. PhD thesis, Faculty of Information and Media Studies, University of Western Ontario, London (Ontario), Canada, URL `http://people.lis.uiuc.edu/~jdownie/mir_papers/thesis_missing_some_music_figs.pdf`
Eerola T, Järvinen T, Louhivuori J, Toiviainen P (2002) Statistical features and perceived similarity of folk melodies. Music Perception 18(3):275–296
Ellis D, Whitman B, Berenzweig A, Lawrence S (2002) The quest for ground truth in musical artist similarity. In: Proceedings of the 3rd International Conference on Music Information Retrieval, pp 170–177
Evangelista G (2001) Flexible wavelets for music signal processing. Journal of New Music Research 30(1):13–22
Faloutsos C, Lin KI (1995) FastMap: A fast algorithm for indexing, data mining and visualization of traditional and multimedia datasets. In: Carey MJ, Schneider DA (eds) Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, California, pp 163–174
Flexer A, Pampalk E, Widmer G (2005) Hidden Markov models for spectral similarity of songs. In: Proceedings of the 8th International Conference on Digital Audio Effects, Madrid, Spain
Foote J (2002) Audio retrieval by rhythmic similarity. In: Proceedings of the 3rd International Conference on Music Information Retrieval
Foote J, Uchihashi S (2001) The beat spectrum: A new approach to rhythm analysis. In: Proceedings of the IEEE International Conference on Multimedia and Expo, Tokyo, Japan, pp 224–228
Friedman J (1989) Regularized discriminant analysis. Journal of the American Statistical Association 84:165–175
Fucks W (1962) Mathematical analysis of formal structure of music. IEEE Transactions on Information Theory 8(5):225–228
Fucks W (1963) Mathematische Analyse von Formalstrukturen von Werken der Musik (mit Diskussion). In: Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen, Westdeutscher Verlag, Köln und Opladen, pp 39–114
Fucks W (1964) Gibt es mathematische Gesetze in Sprache und Musik? In: Frank Hea (ed) Kybernetik – Brücke zwischen den Wissenschaften, Umschau Verlag, Frankfurt am Main, pp 171–183
Fucks W (1968) Nach allen Regeln der Kunst. DVA, Stuttgart
Fucks W, Lauter J (1965) Exaktwissenschaftliche Musikanalyse. Westdeutscher Verlag, Köln und Opladen
Godsill S, Davy M (2003) Bayesian modelling of music audio signals. In: Bulletin of the International Statistical Institute, 54th Session, Berlin, vol LX, book 2, pp 504–506
Godsill S, Davy M (2005) Bayesian computational models for inharmonicity in musical instruments. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, pp 283–286
Gomez E (2004) Tonal description of polyphonic audio for music content processing. INFORMS Journal on Computing, Special Cluster on Computation in Music 18(3):294–304
Gomez E (2006) Tonal description of music audio signals: Harmonic pitch class profiles, tonality and tonal similarity of polyphonic audio signals. PhD thesis, Departament de Tecnologia, Universitat Pompeu Fabra, Barcelona, Spain

Goto M (2003) A chorus-section detecting method for musical audio signals. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp 437–440

Goto M (2004) A predominant-F0 estimation method for polyphonic musical audio signals. In: Proceedings of the 18th International Congress on Acoustics (ICA'04), Acoustical Society of Japan, Kyoto, Japan, pp 1085–1088

Gouyon F (2005) A computational approach to rhythm description: Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing. PhD thesis, Universitat Pompeu Fabra, Departament de Tecnologia, Barcelona, Spain

Gouyon F, Dixon S (2005) A review of automatic rhythm description systems. Computer Music Journal 29(1):34–54

Gouyon F, Klapuri A, Dixon S, Alonso M, Tzanetakis G, Uhle C, Cano P (2006) An experimental comparison of audio tempo induction algorithms. IEEE Transactions on Speech and Audio Processing 14(5):1832–1844

Gromko JE (1993) Perceptual differences between expert and novice music listeners at multidimensional scaling analysis. Psychology of Music 21:34–47

Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. Springer, New York, URL `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`

Herre J, Allamanche E, Ertel C (2003) How similar do songs sound? Towards modeling human perception of musical similarity. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp 83–86

Herrera P, Sandvold V, Gouyon F (2004) Percussion-related semantic descriptors of music audio files. In: Proceedings of the 25th International AES Conference, London, United Kingdom

Hyvärinen A, Karhunen J, Oja E (2001) Independent Component Analysis. John Wiley and Sons, New York

Jürgensen F, Knopke I (2004) A comparison of automated methods for the analysis of style in fifteenth-century song intabulations. In: Parncutt R, Kessler A, Zimmer F (eds) Proceedings of the Conference on Interdisciplinary Musicology (CIM04), URL `http://www-gewi.uni-graz.at/staff/parncutt/cim04/CIM04_paper_pdf/JurgensenKnopke.pdf`

Kantz H, Schreiber T (1997) Nonlinear Time Series Analysis. Cambridge University Press

Klapuri A (2001) Multipitch estimation and sound separation by the spectral smoothness principle. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol 5, pp 3381–3384

Klapuri A (2004) Automatic music transcription as we know it today. Journal of New Music Research 33(3):269–282

Klapuri A, Davy M (eds) (2006) Signal Processing Methods for Music Transcription. Springer, New York

Kleber B (2002) Evaluation von Stimmqualität in westlichem, klassischen Gesang. Diploma Thesis, Fachbereich Psychologie, Universität Konstanz, Germany

Knees P, Pampalk E, Widmer G (2004) Artist classification with web-based data. In: Proceedings of the 5th International Conference on Music Information Retrieval, Barcelona, Spain, pp 517–524

Knuth D (1984) The TEXbook. Addison-Wesley, Reading, Mass

Kohonen T (1995) Self-Organizing Maps. Springer, Berlin

Kopiez R, Weihs C, Ligges U, Lee JI (2006) Classification of high and low achievers in a music sight-reading task. Psychology of Music 34(1):5–26

Koza JR (1992) Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, Mass

Kranenburg Pv, Backer E (2004) Musical style recognition - a quantitative approach. In: Parncutt R, Kessler A, Zimmer F (eds) Proceedings of the Conference on Interdisciplinary Musicology (CIM04), URL `http://www-gewi.uni-graz.at/staff/parncutt/cim04/CIM04_paper_pdf/Kranenburg_Backer_CIM04_proceedings.pdf`

Krumhansl CL (1990) Cognitive Foundations of Musical Pitch. Oxford Psychology Series 17, Oxford University Press, Oxford

Kulesh V, Sethi I, V P (2003) Indexing and retrieval of music via Gaussian mixture models. In: Proceedings of the 3rd International Workshop on Content Based Multimedia Indexing, Rennes, France, pp 201–205

Kullback S, Leibler RA (1951) On information and sufficiency. Annals of Mathematical Statistics 22:79–86

Kurth F, Gehrmann T, Müller M (2006) The cyclic-beat spectrum: Tempo-related audio features for time-scale invariant audio identification. In: Proceedings of the 7th International Conference on Music Information Retrieval, pp 35–40

Lambrou T, Kudumakis P, Speller R, Sandler M, Linney A (1998) Classification of audio signals using statistical features on time and wavelet transform domains. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol 6, pp 3621–3624

Lamport L (1994) LaTeX, a Document Preparation System, 2nd edn. Addison-Wesley, Reading, Mass

Lehwark P, Risi S, Ultsch A (2007) Visualization and clustering of tagged music data. In: to appear in Proc. GfKl 2007, Freiburg, Germany

Lesaffre M, Tanghe K, Martens G, Moelants D, Leman M, De Baets B, De Meyer H, Martens JP (2003) The MAMI Query-By-Voice Experiment: Collecting and annotating vocal queries for music information retrieval. In: Proceedings of the 4th International Conference on Music Information Retrieval, Baltimore, Maryland, USA and Library of Congress, Washington, DC, USA, pp 65–71

Levy M, Sandler M (2006) Lightweight measures for timbral similarity of musical audio. In: Proceedings of the 1st ACM workshop on audio and music computing multimedia (AMCMM), ACM Press, pp 27–36

Li D, Sethi I, Dimitrova N, McGee T (2001) Classification of general audio data for content-based retrieval. Pattern Recognition Letters 22:533–544

Li T, Ogihara M, Li Q (2003) A comparative study on content-based music genre classification. In: Proceedings of the 26th International ACM SIGIR Conference on Research and development in information retrieval, ACM Press, pp 282–289

Lidy T, Rauber A (2005) Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In: Proceedings of the 6th International Conference on Music Information Retrieval, pp 34–41

Ligges U (2006) Transkription monophoner Gesangszeitreihen. Dissertation, Fachbereich Statistik, Universität Dortmund, Dortmund, Germany, URL `http://hdl.handle.net/2003/22521`

Logan B (2000) Mel frequency cepstral coefficients for music modeling. In: Proceedings of the 1st International Conference on Music Information Retrieval, pp 23–25

Logan B, Salomon A (2001) A music similarity function based on signal analysis. In: Proceedings of the IEEE International Conference on Multimedia and Expo, pp 745–748

Mandel M, Ellis D (2005) Song-level features and SVMs for music classification. In: Proceedings of the 6th International Conference on Music Information Retrieval, pp 594–599

Markuse B, Schneider A (1996) Ähnlichkeit, Nähe, Distanz: zur Anwendung multidimensionaler Skalierung in musik-wissenschaftlichen Untersuchungen. Systematische Musikwissenschaft / Systematic Musicology / Musicologie systématique 4:53–89

McEnnis D, McKay C, Fujinaga I, Depalle P (2005) jAudio: A feature extraction library. In: Proceedings of the 6th International Conference on Music Information Retrieval, pp 600–603

McKinney M, Breebaart J (2003) Features for audio and music classification. In: Proceedings of the 4th International Conference on Music Information Retrieval, pp 151–158

Meng A (2006) Temporal feature integration for music organisation. PhD thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU

Meng A, Ahrendt P, Larsen J (2005) Improving music genre classification by short-time feature integration. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol V, pp 497–500

Meng A, Ahrendt P, Larsen J, Hansen LK (2006) Temporal feature integration for music genre classification. IEEE Transactions on Signal Processing 15:1654–1664

Meyer J (1995) Akustik und musikalische Aufführungspraxis. Bochinsky, Frankfurt am Main

Meyer LB (1957) Meaning in music and information theory. Journal of Aesthetics and Art Criticism 15:412–424

Microsoft Corporation (1991) Multimedia Programming Interface and Data Specification, 1.0. Joint design by IBM Corporation and Microsoft Corporation

MIDI Manufacturers Association (2001) Complete MIDI 1.0 Detailed Specification. 2nd edn, URL `http://www.midi.org`

Mierswa I, Morik K (2005) Automatic feature extraction for classifying audio data. Machine Learning Journal 58:127–149

Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T (2006) YALE: Rapid prototyping for complex data mining tasks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, NY, USA, pp 935–940

Moles A (1958) Théorie de l'information et perception estétique. Flammarion, Paris

Moles A (1971) Informationstheorie und ästhetische Wahrnehmung. DuMont Schauberg, Köln

Moore BCJ, Glasberg BR (1996) A revision of Zwickers loudness model. ACTA Acustica 82:335–345

Mörchen F, Ultsch A, Nöcker M, Stamm C (2005a) Databionic visualization of music collections according to perceptual distance. In: Proceedings of the 6th International Conference on Music Information Retrieval, pp 396–403

Mörchen F, Ultsch A, Thies M, Löhken I, Nöcker M, Stamm C, Efthymiou N, Kümmerer M (2005b) MusicMiner: Visualizing timbre distances of music as topographical maps. Tech. rep., Department of Mathematics and Computer Science, University of Marburg, Germany

Mörchen F, Mierswa I, Ultsch A (2006a) Understandable models of music collections based on exhaustive feature generation with temporal statistics. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, Philadelphia, PA, USA, pp 882–891

Mörchen F, Ultsch A, Thies M, Löhken I (2006b) Modelling timbre distance with temporal statistics from polyphonic music. IEEE Transactions on Speech and Audio Processing 14(1):81–90

Müllensiefen D, Frieler K (2004a) Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgments. Computing in Musicology 13:147–176

Müllensiefen D, Frieler K (2004b) Optimizing measures of melodic similarity for the exploration of a large folk song database. In: 5th International Conference on Music Information Retrieval, Audiovisual Institute, Universitat Pompeu Fabra, Barcelona, Spain, pp 274–280

Müllensiefen D, Hennig C (2006) Modeling memory for melodies. In: Siliopoulou M, Kruse R, Borgelt C, Nürnberger A, Gaul W (eds) From Data and Information Analysis to Knowledge Engineering, Gesellschaft für Klassifikation e.V., Springer, Berlin, pp 732–739

Narmour E (1990) The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model. University of Chicago Press, Chicago

Nienhuys HW, Nieuwenhuizen J, et al (2005) GNU LilyPond – The Music Typesetter. Free Software Foundation, URL `http://www.lilypond.org/`, version 2.6.5

Ombao H, Raz J, von Sachs R, Malow B (2001) Automatic statistical analysis of bivariate nonstationary time series. Journal of the American Statistical Association 96(454):543–560

Oppenheim A, Schafer R, Buck J (1999) Discrete-Time Signal Processing, 2nd edn. Prentice-Hall, New Jersey

Pachet F, Zils A (2003) Evolving automatically high-level music descriptors from acoustic signals. In: Proceedings of the International Symposium on Computer Music Modeling and Retrieval, pp 42–53

Pampalk E (2004) A MATLAB toolbox to compute music similarity from audio. In: Proceedings of the 5th International Conference on Music Information Retrieval, Barcelona, Spain, pp 254–257

Pampalk E (2006a) Audio-based music similarity and retrieval: Combining a spectral similarity model with information extracted from fluctuation patterns. In: 3rd Annual Music Information Retrieval eXchange (MIREX'06), URL `http://pampalk.at/publications/`

Pampalk E (2006b) Computational models of music similarity and their application in music information retrieval. PhD thesis, Computer Science Department, Technical University Vienna, Austria

Pampalk E, Goto M (2006) MusicRainbow: A new user interface to discover artists using audio-rased similarity and web-based labeling. In: Proceedings of the 7th International Conference on Music Information Retrieval, pp 367–370

Pampalk E, Rauber A, Merkl D (2002) Content-based organization and visualization of music archives. In: Proceedings of the 10th ACM International Conference on Multimedia, pp 570–579

Pampalk E, Dixon S, Widmer G (2003a) Exploring music collections by browsing different views. In: Proceedings of the 4th International Conference on Music Information Retrieval,

pp 201–208
Pampalk E, Dixon S, Widmer G (2003b) On the evaluation of perceptual similarity measures for music. In: Proceedings of the International Conference on Digital Audio Effects, pp 6–12
Pampalk E, Flexer A, Widmer G (2005) Hierarchical irganization and description of music collections at the artist level. In: Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries, pp 37–48
Pang H, Yoon D (2005) Automatic detection of vibrato in monophonic music. Pattern Recognition 38(7):1135–1138
Pearce MT, Wiggins GA (2004) Improved methods for statistical modelling of monophonic music. Journal of New Music Research 33(4):367–385
Pearce MT, Wiggins GA (2006) Expectation in melody: The influence of context and learning. Music Perception 23(5):377–405
Pierce JR (1992) The science of musical sound, 2nd ed. W.H. Freeman and Co., New York
Plumbley M (2003) Algorithms for nonnegative independent component analysis. IEEE Transactions on Neural Networks 14(3):534–543
Plumbley M (2004) Optimization using Fourier expansion over a geodesic for non-negative ICA. In: Proceedings of the International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004), Granada, Spain, pp 49–56
Plumbley M, Abdallah S, Blumensath T, Jafari M, Nesbit A, Vincent E, Wang B (2006) Musical audio analysis using sparse representations. In: COMPSTAT 2006 – Proceedings in Computational Statistics, Physica Verlag, Heidelberg, pp 104–117
Pohle T (2006) Post processing music similarity computations. In: The Second Annual Music Information Retrieval Evaluation eXchange (MIREX 2006), pp 16–18, URL `http://www.music-ir.org/evaluation/MIREX/2006_abstracts/AS_pohle.pdf`
Pohle T, Pampalk E, Widmer G (2005) Evaluation of frequently used audio features for classification of music into perceptual categories. In: Proceedings of the 4th International Workshop on Content-Based Multimedia Indexing (CBMI), Riga, Latvia
Polotti P, Evangelista G (2000) Harmonic-band wavelet coefficient modeling for pseudo-periodic sound processing. In: Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy, pp 103–108
Polotti P, Evangelista G (2001) Multiresolution sinusoidal/stochastic model for voiced-sounds. In: Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland, pp 120–124
Pressing J, Lawrence P (1993) Transcribe: A comprehensive autotranscription program. In: Proceedings of the International Computer Music Conference, Tokyo, Japan, pp 343–345
Pye D (2000) Content-based methods for managing electronic music. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pp 2437–2440
R Development Core Team (2007) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org`, ISBN 3-900051-07-0
Rabiner L, Juang BH (1993) Fundamentals of Speech Recognition. Prentice-Hall, New York
Rabiner LR (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 77(2):257–286
Raphael C (2001) A probabilistic expert system for automatic musical accompaniment. Journal of Computational and Graphical Statistics 10(3):487–512
Risi S, Mörchen F, Ultsch A, Lewark P (2007) Visual mining in music collections with emergent SOM. In: to appear in Proceedings Workshop on Self-Organizing Maps (WSOM)
Rossignol S, Depalle P, Soumagne J, Rodet X, Collette JL (1999a) Vibrato: Detection, estimation, extraction, modification. In: Proceedings of the COST-G6 Workshop on Digital Audio Effects (DAFx-99)
Rossignol S, Rodet X, Soumagne J, Collette JL, Depalle P (1999b) Automatic characterisation of musical signals: Feature extraction and temporal segmentation. Journal of New Music Research 28(4):281–295
Röver C, Klefenz F, Weihs C (2005) Identification of musical instruments by means of the Hough-Transformation. In: Weihs C, Gaul W (eds) Classification – the Ubiquitous Challenge, Springer, Berlin, Heidelberg, pp 608–615
Rubner Y, Tomasi C, Guibas LJ (1998) A metric for distributions with applications to image databases. In: Proceedings of the IEEE International Conference on Computer Vision, Bombay, India, pp 59–66

Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Information Processing and Management 24(5):513–523
Sandvold V, Herrera P (2005) Towards a semantic descriptor of subjective intensity in music. In: Proceedings of the International Computer Music Conference
Schedl M, Pohle TP, Knees P, Widmer G (2006) Assigning and visualizing music genres by web-based co-occurance analysis. In: Proceedings of the 7th International Conference on Music Information Retrieval, pp 260–265
Scheirer ED (1998) Tempo and beat analysis of acoustic musical signals. Journal of the Acoustical Society of America 103(1):588–601
Schellenberg EG (1997) Simplifying the implication-realization model of simplifying the implication-realization model of melodic expectancy. Music Perception 14:295–318
Seidner W, Wendler J (1997) Die Sängerstimme. Henschel, Berlin
Shao X, Xu C, Kankanhalli MS (2004) Unsupervised classification of music genre using Hidden Markov Model. In: Proceedings of the IEEE International Conference on Multimedia and Expo, pp 2023–2026
Shapiro S (1978) Feature space transforms for curve detection. Pattern Recognition 10:129–143
Smaragdis P, Brown J (2003) Non-negative matrix factorization for polyphonic music transcription. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp 177–180
Steinbeck W (1982) Struktur und Ähnlichkeit: Methoden automatisierter Melodieanalyse. Bärenreiter, Kassel
Stenzel R, Kamps T (2005) Improving content-based similarity measures by training a collaborative model. In: Proceedings of the 6th International Conference on Music Information Retrieval, pp 264–271
Stevens S, Volkmann J (1940) The relation of pitch to frequency. American Journal of Psychology 53(3):329–353
Streich S, Herrera P (2004) Toward describing perceived complexity of songs: Computational methods and implementation. In: Proceedings of the 25th International AES Conference
Streich S, Herrera P (2005) Detrended fluctuation analysis of music signals: Danceability estimation and further semantic characterization. In: Proceedings of the 118th AES Convention
Temperley D (2001) The Cognition of Basic Musical Structures. MIT Press, Cambridge, MA
Temperley D (2004) Bayesian models of musical structure and cognition. Musicae Scientiae 8(2):175–205
Temperley D (2006) A probabilistic model of melody perception. In: Proceeding of the 7th International Conference on Music Information Retrieval, pp 276–279, URL `http://ismir2006.ismir.net/PAPERS/ISMIR0630_Paper.pdf`
Temperley D (2007) Music and Probability. MIT Press, Cambridge, MA
Thomassen J (1982) Melodic accent: Experiments and a tentative model. Journal of the Acoustical Society of America 71:1596–1605
Torrens M, Hertzog P, Arcos JL (2004) Visualizing and exploring personal music libraries. In: Proceedings of the 5th International Conference on Music Information Retrieval, pp 421–424
Tzanetakis G, Cook P (2000) MARSYAS: A framework for audio analysis. Organised Sound 4(30):169–175
Tzanetakis G, Cook P (2002) Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing 10(5):293–302
Tzanetakis G, Ermolinskyi A, Cook P (2002a) Beyond the query-by-example paradigm: New query interfaces for music. In: Proceedings of the International Computer Music Conference, pp 177–183
Tzanetakis G, Ermolinskyi A, Cook P (2002b) Pitch histograms in audio and symbolic music information retrieval. In: Proceedings of the 3rd International Conference on Music Information Retrieval, pp 31–38
Tzanetakis G, Essl G, Cook P (2002c) Human perception and computer extraction of beat strength. In: Proceedings of the International Conference on Digital Audio Effects (DAFx-02), pp 257–261
Ultsch A (1993) Self-organizing neural networks for visualization and classification. In: Opitz O, Lausen B, Klar R (eds) Information and Classification - Concepts, Methods, and Applications, Springer-Verlag, Berlin, pp 307–313

Ultsch A (1996) Self organizing neural networks perform different from statistical k-means clustering. In: BMBF Statusseminar Künstliche Intelligenz, Neuroinformatik und Intelligente Systeme, München, pp 433–443

Ultsch A, Mörchen F (2005) ESOM-Maps: Tools for clustering, visualization, and classification with emergent SOM. Tech. Rep. 46, Department of Mathematics and Computer Science, University of Marburg, Germany

Van Trees H (2001) Detection, Estimation, and Modulation Theory, Part I, reprint edn. Wiley-Interscience, Melbourne, FL, USA

Vembu S, Baumann S (2005) A self-organizing map based knowledge discovery for music recommendation systems. In: Computer Music Modeling and Retrieval, pp 119–229

Vignoli F, Pauws S (2005) A music retrieval system based on user driven similarity and its evaluation. In: Proceedings of the 6th International Conference on Music Information Retrieval, pp 272–279

Vignoli F, van Gulik R, van de Wetering H (2004) Mapping music in the palm of your hand, explore and discover your collection. In: Proceedings of the 5th International Conference on Music Information Retrieval, pp 409–414

Viste H, Evangelista G (2001) Sounds source separation: Preprocessing for hearing aids and structured audio coding. In: Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland, pp 67–70

Viste H, Evangelista G (2002) An extension for source separation techniques avoiding beats. In: Proceedings of the 5th International Conference on Digital Audio Effects (DAFx-02), Hamburg, Germany, pp 71–75

Wakefield G (1999) Mathematical representation of joint time-chroma distributions. In: Proceedings of the SPIE International Symposium on Optical Science, Engineering and Instrumentation, Denver, Colorado, pp 637–645

Walmsley P, Godsill S, Rayner P (1999) Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, pp 119–122

Wapnick J, Ekholm E (1997) Expert consensus in solo voice performance evaluation. Journal of Voice 11(4):429–436

Weihs C, Ligges U (2005) From local to global analysis of music time series. In: Morik K, Boulicaut JF, Siebes A (eds) Local Pattern Detection, Springer-Verlag, Berlin, Lecture Notes in Artificial Intelligence 3539, pp 217–231

Weihs C, Ligges U (2006) Parameter optimization in automatic transcription of music. In: Spiliopoulou M, Kruse R, Nürnberger A, Borgelt C, Gaul W (eds) From Data and Information Analysis to Knowledge Engineering, Springer-Verlag, Berlin, pp 740–747

Weihs C, Berghoff S, Hasse-Becker P, Ligges U (2001) Assessment of purity of intonation in singing presentations by discriminant analysis. In: Kunert J, Trenkler G (eds) Mathematical Statistics and Biometrical Applications, Josef Eul, Bergisch-Gladbach, pp 395–410

Weihs C, Ligges U, Sommer K (2006a) Analysis of music time series. In: Rizzi A, Vichi M (eds) COMPSTAT 2006 – Proceedings in Computational Statistics, Physica Verlag, Heidelberg, pp 147–159

Weihs C, Szepannek G, Ligges U, Luebke K, Raabe N (2006b) Local models in register classification by timbre. In: Batagelj V, Bock HH, Ferligoj A, Žiberna A (eds) Data Science and Classification, Springer-Verlag, Berlin, pp 315–322

West K, Cox S, Lamere P (2006) Incorporating machine-learning into music similarity estimation. In: Proceedings of the 1st ACM workshop on Audio and music computing multimedia (AMCMM), ACM Press, pp 89–96

Whiteley N, Cemgil A, Godsill S (2006) Bayesian modelling of temporal structure in musical audio. In: 7th International Conference on Music Information Retrieval, Victoria, Canada, pp 29–34

Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, New York

Wolfe P, Godsill S, Ng WJ (2004) Bayesian variable selection and regularization for time-frequency surface estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66(3):575–589

Zils A, Pachet F (2004) Automatic extraction of music descriptors from acoustic signals using EDS. In: Proceedings of the 116th AES Convention

Zwicker E, Stevens S (1957) Critical bandwidths in loudness summation. Journal of the Acoustical Society of America 29(5):548–557