Foundations and Trends® in
Information Retrieval
Vol. 1, No 1 (November 2006) 1–90

DOI: 10.1561/1500000002

now
the essence of knowledge

Music Retrieval: A Tutorial and Review

Nicola Orio

Department of Information Engineering, University of Padova
Via Gradenigo, 6/b, Padova 35131, Italy, orio@dei.unipd.it,

Abstract

The increasing availability of music in digital format needs to be matched by the development of tools for music accessing, filtering, classification, and retrieval. The research area of Music Information Retrieval (MIR) covers many of these aspects. The aim of this paper is to present an overview of this vast and new field. A number of issues, which are peculiar to the music language, are described—including forms, formats, and dimensions of music—together with the typologies of users and their information needs. To fulfil these needs a number of approaches are discussed, from direct search to information filtering and clustering of music documents. An overview of the techniques for music processing, which are commonly exploited in many approaches, is also presented. Evaluation and comparisons of the approaches on a common benchmark are other important issues. To this end, a description of the initial efforts and evaluation campaigns for MIR is provided.

1

Introduction

The amount of available digital music is continuously increasing, promoted by a growing interest of users and by the development of new technology for the ubiquitous enjoyment of music.

There are a number of reasons that may explain this trend, first of all, the characteristics of music language itself. Music is an art form that can be shared by people with different culture because it crosses the barriers of national languages and cultural backgrounds. For example, Western classical music has passionate followers in Japan, and many persons in Europe are keen on classical Indian music: All of them can enjoy music without the need of a translation, which is normally required for accessing foreign textual works. Another reason is that technology for music recording, digitization, and playback allows users for an access that is almost comparable to the listening of a live performance, at least at the level of audio quality, and the signal-to-noise ratio is better for digital formats than for many analog formats. This is not the case with other art forms, like painting or sculpture, for which the digital format is only an approximate representation of the artwork. The access to digitized paintings can be useful for studying the works of a given artist but cannot substitute the direct interaction

with the *real world* works. Moreover, music is an art form that can be both cultivated and popular, and sometimes it is impossible to draw a line between the two, as for jazz or for most of traditional music.

Or maybe the increasing interest toward digital music is motivated by its portability and the possibility to access music while doing another, possibly primary, activity. Perhaps users are not looking for a cultural experience, or the enjoyment of artworks, but for a suitable soundtrack for the many hours spent commuting, traveling, waiting or even working, and studying. Last but not least, the music business is pushing toward the continuous production of new musical works, especially in genres like pop and rock. The continuous decrease of the average age of persons that regularly purchase and consume music has been paralleled by an increasing simplification of the structure of mainstream music genres, requiring the continuous production of new music. The number of items sold daily by Web-based music dealers, or downloaded from services like i-Tune—not mentioning the illegal sharing of music files through peer-to-peer networks—shows how much music is commercially and culturally important.

Music Information Retrieval (MIR) is an emerging research area devoted to fulfill users' *music* information needs. As it will be seen, despite the emphasis on retrieval of its name, MIR encompasses a number of different approaches aimed at music management, easy access, and enjoyment. Most of the research work on MIR, of the proposed techniques, and of the developed systems are content based.

The main idea underlying content-based approaches is that a document can be described by a set of features that are directly computed from its content. Usually, content-based access to multimedia data requires specific methodologies that have to be tailored to each particular medium. Yet, the core information retrieval (IR) techniques, which are based on statistics and probability theory, may be more generally employed outside the textual case, because the underlying models are likely to describe fundamental characteristics being shared by different media, languages, and application domains [60]. For this reason, the research results achieved in the area of IR, in particular in the case of text documents, are a continuous reference for MIR approaches. Already in 1996 McLane stated that a challenging research topic would

have been the application of some standard principles of text IR to music representation [85].

The basic assumption behind content-based approaches is that metadata are either not suitable, or unreliable, or missing. In the case of MIR, all the three assumptions may hold. As it will be described in more detail in the following sections, music metadata may not be suitable because they are either too generic to discriminate different musical works—there are hundreds of slow ballads sung by a female voice with acoustic guitars in the background—or too specific requiring a precise definition of the information need—name of the singer, name of the album, and date of the first release if not the direct name of the song—or requiring a good musical background—the song is in F lydian, with guitars mainly playing suspended chords, at the end the time signature becomes 3/4 with slower tempo.[1] Metadata can be unreliable in the case of music shared with other users, because there is no control on how the information has been added, and of course metadata may not be present at all.

It is interesting to note that both CD and the first MP3 audio standards, which gave rise to the two digital revolutions of music enjoyment, do not take into account the possibility of carrying also structured metadata information. The possibility to include unstructured textual information in MP3 has been introduced in a later version thanks to an external contribution to the standard. Yet, fields are not mandatory, and it is up to the person who creates the MP3 to spend time adding this information.

Despite these problems with metadata, a number of systems that allow users to access and retrieve music based on textual descriptors have been developed, and are available on the Web. Most of the systems are Digital Libraries devoted to the diffusion of Western classical music, such as Cantate [13], Harmonica [43], and Musica [97], whose names already show the primary interest toward this repertoire—other music digital libraries for Western music are described in [45] and in [27]. The discussion of the relationships between the vast research area of

[1] This sentence has been created on purpose using terms that, although very precise, are not very common outside the music terminology; there is probably no song with these characteristics.

digital libraries and MIR would have required a paper by itself, because it regards important issues on digital acquisition of musical works, the creation of an infrastructure for the management of musical documents and for the access to musical content, which are challenging problem by themselves as discussed in [1, 29].

The focus of this paper is on tools, techniques, and approaches for content-based MIR, rather than on systems that implement them. The interested reader can find the descriptions of more than 35 systems for music retrieval in [127], with links to their Web sites. An interesting survey of a selection of 17 different systems has been presented in [135]. Systems can be compared according to the retrieval tasks that can be carried out, the size of the collections, and the techniques that are employed. To this end, a mapping of the analyzed systems on a bidimensional space was proposed [135] in where the two dimensions of interest were the target audience of the systems—e.g., industry, consumer, professional—and the level at which retrieval is needed—e.g., from a particular instance of a work to a music genre.

The paper is structured as follows: This introductory section ends with a short overview of some music concepts. Chapter 2 discusses the peculiarities of the music language, introducing the dimensions that characterize musical documents and that can be used to describe its content. Chapter 3 highlights the main typologies of MIR users, introducing and specifying a number of information needs that have been taken into account by different MIR approaches. The processing of musical documents, aimed at extracting features related to their dimensions of interest, is discussed in Chapter 4, followed by a classification of the different facets of MIR research areas, which are reported in Chapter 5. The efforts carried out for creating a shared evaluation framework and their initial results are presented in Chapter 6. Finally, some concluding considerations are drawn in Chapter 7.

1.1 Review of Music Concepts

Most of the approaches and techniques for music retrieval are based on a number of music concepts that may not be familiar to persons

without musical training. For this reason, this section presents a short introduction of some basic concepts and terms.

1.1.1 Basic music elements

Each musical instrument, with the exception of some percussions, produces almost periodic vibrations. In particular, the sounds produced by musical instruments are the result of the combination of different frequencies, which are all multiple integers of a *fundamental frequency*, usually called $F0$.

The three basic features of a musical sound are

- **Pitch**, which is related to the perception of the fundamental frequency of a sound; pitch is said to range from *low* or *deep* to *high* or *acute* sounds.
- **Intensity**, which is related to the amplitude, and thus to the energy, of the vibration; textual labels for intensity range from *soft* to *loud*; the intensity is also defined *loudness*.
- **Timbre**, which is defined as the sound characteristics that allow listeners to perceive as different two sounds with same pitch and same intensity.

The perception of pitch and intensity is not as straightforward as the above definitions may suggest. The human ear does not have a linear behavior neither in pitch recognition nor in the perception of intensity. Yet, these two perceptually relevant qualities of sounds can be reasonably approximated considering the fundamental frequency and the energy of a sound.

Timbre, on the other hand, is a multidimensional sound quality that is not easily described with simple features. Timbre perception is related to the recognition of the sound source—telling a saxophone from a violin—of the playing technique—telling whether a string has been plucked or played with the bow—of the playing technique nuances—the velocity of the bow and its pressure on the string—of the surrounding acoustics—telling whether the violinist has been in a small room or in a concert hall—and of the recording equipment and digital representation—telling whether the sound has been recorded

from the broadcast of an AM radio or from a live performance. Given all these characteristics, it is not surprising that timbre has been defined for what it is not.

In the case of many percussive musical instruments, there is no fundamental frequency, and the sound is usually called *noise*. Yet, noises are perceived to be in a low, medium, or high register. Intensity and timbre are still relevant descriptors for noises.

When two or more sounds are played together, they form a *chord*. A chord may have different qualities, depending on the pitch of the different sounds and, in particular, on the distances between them. Chords play a fundamental role in many music genres, in particular in pop, rock, and jazz music, where polyphonic musical instruments—e.g., piano, keyboard, guitar—are often dedicated to the accompaniment and basically play chords.

1.1.2 Music terminology

Apart from the basic concepts introduced in the previous section, there are many terms that are currently used to describe music that may not be familiar to persons without a musical education. Part of the terminology commonly used by the MIR community is derived from music theory and practice. The musical concepts that are relevant for this overview are the following[2]:

- The **tempo** is the speed at which a musical work is played, or expected to be played, by performers. The tempo is usually measured in *beats per minute*.
- The **tonality** of a song is related to the role played by the different chords of a musical work; tonality is defined by the name of the chord that plays a central role in a musical work. The concept of tonality may not be applicable to some music genres.

[2] Being very simple, these operative definitions may not be completely agreed by readers with a musical background, because some relevant aspects are not taken into account. Yet, their purpose is just to introduce some terminology that will be used in the next chapters.

- The **time signature**, usually in the form of a fractional number, gives information on the organization of strong and soft beats along the time axis.
- The **key signature**, usually in the form of a number of alterations—symbols ♯ and ♭—is an incomplete representation of the tonality, which is useful for performers because it express which are the notes that have to be consistently played altered.

Figure 1.1 shows four measures of a polyphonic musical score, an excerpt from Claude Debussy's *Première Arabesque*, for piano. In this excerpt no direct information on tempo and tonality is given, the time signature (the **C** sign) indicates that measures have to be divided in four equal beats, and the three sharps (the ♯ signs) indicate that all occurrences of notes *F*, *C*, and *G* have to be raised by a semitone if not otherwise indicated. The presence of three sharps may suggest that the tonality of the excerpt is either A major or F♯ minor, which have the same number of sharps, with the former more likely to be the actual tonality.

Other concepts are more related to the production of sound and to the parameters that describe single or groups of notes. In particular, a note starts being perceived at its *onset time*, lasts for a given *duration*, and stops to be perceived at its *offset time*. Finally, the sounds are produced by musical instruments and by the voice that depending on their conformation have a limited range of pitches that can be produced, which is called instrument—or voice—*register*.

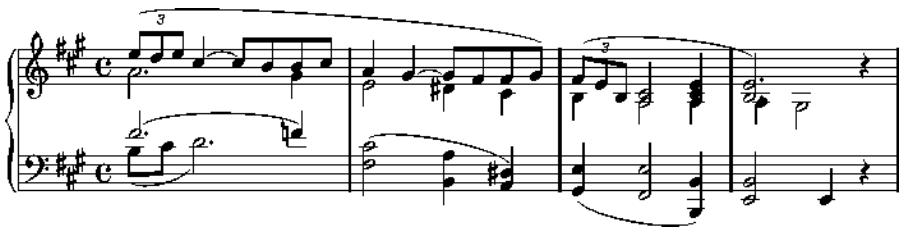

Fig. 1.1 Example of a musical score (excerpt from *Première Arabesque* by Claude Debussy).

2

Characteristics of the Music Language

The effectiveness of tools for music retrieval depends on the way the characteristics of music language, and somehow its peculiarities, are modeled. For example, it is well known that textual IR methodologies are based on a number of shared assumptions about the written text, in particular on the role of words as content descriptors, depending on their frequency and distribution inside the documents and across the collections. Similarly, image retrieval is based on assumptions that images with similar distribution of colors and shapes may be perceived as similar. Video retrieval extends these assumption with the notion of similarity between frames, which are assumed to be related to the concept of movement. All these assumptions are based on scientific results in the field of linguistic, information science, psychology of visual perception, information engineering, and so on.

In the case of music, concepts developed by music theorists, musicologists, psychologists of perception, and information scientists and engineers can be exploited to design and to refine the approaches to

MIR. In particular, the different elements of a musical work and the alternative forms in which it can be instantiated play a main role in the development of methodologies and techniques to music retrieval.

2.1 Which Content Does Music Convey?

Text, images, video, or speech-based documents in general are able to convey information that forms their content. The users may search for these media by giving a description of the kind of content they are looking for. For example, let us consider the concept of *tempest*: A written or spoken text may give its definition; an image may represent most of its visual characteristics; a video may combine this information possibly adding environmental sounds. On the other hand, to convey the concept of tempest through music is a difficult task, and it is still unclear what kind of content is conveyed by musical works, apart from the music itself.

It may be argued that music is an art form, and, like architecture and other structural arts, its goal is not to convey concepts. For example, the first chapter of the play *The Tempest* by Shakespeare or the landscape of the painting *The Tempest* by Giorgione probably are not meant to simply convey the concept of tempest. Yet, users may recognize some common characteristics related to the concept of tempest that they can describe when searching for these media. There are dozens of works of Western classical music whose title, or the title of a passage, is related to tempests; among those the most famous probably are the IV Movement of the Sixth Symphony by Beethoven, the Overture of *William Tell* by Rossini, and the Concerto *La Tempesta di Mare* by Vivaldi. These works differ in most of their features—or *dimensions* as described in Section 2.2—and above all the user may not be able to recognize that the work is about a tempest and not just pure music.

There is a particular kind of music called *musica a programma*, where the title like the symphony *Eroica* by Beethoven or a lyric like in the works *Prélude à l'après-midi d'un faune* by Debussy and *L'apprenti sorcier* by Dukas suggest a meaning to the listener. Nevertheless, it is

a difficult task for a listener to guess the text that inspired a given musical work.

These considerations lead to the conclusion that music characteristics can be described almost only in musical terms. Probably because of this reason, the terminology used in music domain refers to the structure of the musical work or to the overall features, rather than to its content. Hence terms like *B flat minor*, *concerto*, *ballad*, and *remix* refer to some of the possible global characteristics of a musical work but are not useful to discriminate among the hundreds or thousands of different concerti, ballads, and remixes. Clearly, relevant metadata on author's name, work title, and so on are particularly useful to retrieve relevant documents if the user already knew this information. To this end, the effectiveness of metadata for querying a music collection has been studied in [70], where name–title entries, title entries, and contents notes have been compared. From the analysis of the results, keyword retrieval based on a textual description of the content was shown to be less effective than title searching and name–title searching.

All these considerations suggest to focus on music content, in terms of music features, as a way to access and retrieve music.

2.2 Dimensions of the Music Language

Music can be defined as the art of disposing and producing sounds and silences in time; thus the temporal organization of sounds plays a major role in music perception. Music has a *horizontal* dimension—the term is due to the representation used in Western music that associates time to the horizontal axis. In the case of polyphonic music, when more sounds are playing at the same time, the relationships between different tones become a relevant perceptual factor. This is called the *vertical* dimension of music—because tones that are simultaneously active are vertically aligned in the musical score.

Music theory and analysis are based on the definition of a number of additional dimensions that, although correlated, may be used separately to describe a musical work. For the aims of music retrieval, a musical work may be relevant for a user because one or more of these dimensions are of interest for his information needs. The main

dimensions of music that could be effective for music retrieval are the following:

- **Timbre**[1] depends on the perception of the quality of sounds, which is related to the used musical instruments, with possible audio effects, and to the playing techniques. All the musical gestures contribute to the perception of the overall timbre of a performance.
- **Orchestration** is due to the composers' and performers' choices in selecting which musical instruments are to be employed to play the different voices, chords, and percussive sounds of a musical work.
- **Acoustics** can be considered as a specialization on some characteristics of timbre, including the contribution of room acoustics, background noise, audio post-processing, filtering, and equalization. This dimension is conveyed by the sum of the individual timbres, with the goal of fulfilling the users' expectations on sound quality, ambience, and style.
- **Rhythm** is related to the periodic repetition, with possible small variants, of a temporal pattern of onsets alone. Sounds do not need to have a recognizable pitch, because the perception of rhythm is related to the sound onsets; to this end, unpitched and percussive sounds are the most used conveyors of the rhythmic dimension. Different rhythms can be perceived at the same time in the case of polyrhythmic music.
- **Melody** is made of a sequence of tones with a similar timbre that have a recognizable pitch within a small frequency range. The singing voice and monophonic instruments that play in a similar register are normally used to convey the melodic dimension. Melody is a multidimensional feature by itself, because different melodies can be played at the same time by different musical instruments, as in counterpoint.

[1] The dictionary-based definition of timbre given in Section 1.1.1 is too generic for MIR purposes. It has been preferred to redefine it according to its common use in MIR approaches.

- **Harmony** is the organization, along the time axis, of simultaneous sounds with a recognizable pitch. Harmony can be conveyed by polyphonic instruments, by a group of monophonic instrument, or may be indirectly implied by the melody.
- **Structure** is a horizontal dimension whose time scale is different from the previous ones, being related to macro-level features such as repetitions, interleaving of themes and choruses, presence of breaks, changes of time signatures, and so on.

In principle, any combination of these dimensions may be a relevant descriptor of a musical work. Yet, it is likely that the dimensions of interest vary with the level of the user's expertise and the specific user's search task, as described in Section 3.

The dimensions introduced in this section are related to the *facets* of music information retrieval presented in [25], where aspects not directly conveyed by music but strictly related to its content are also taken into account. In particular, editorial, textual, and bibliographic facets are described as relevant for MIR. Another approach for dealing with the complexity of the content conveyed by music is the development of a user-dependent taxonomy of music descriptors [73]. The aim of the taxonomy, which includes the dimensions presented in this section with *loudness* and *subjective qualities* as additional categories, is to help designers and users of MIR systems dealing with the diversities of information needs, background knowledge, and expectations.

Timbre, orchestration, and acoustics are more related to the perception of sounds and are *short-term* features. In fact, their processing is carried out by the human auditory systems using few milliseconds of information. In many musical works they tend to change slowly over time, for instance the same instruments are used for the complete song, with the same playing techniques, and with identical room acoustics and eventual post-processing of the recording. For this reason, even if sound is a function of time, these dimensions can be represented also with a time-independent *snapshot*, which summarizes them for a

complete musical work. Many music genres are characterized by and recognizable through timbre, orchestration, and acoustics.

Rhythm, melody, and harmony are characterized by the way the basic elements of sounds, i.e., tones, percussive sounds, and silence, are organized, disposed, and perceived along the time axis, and are *middle-term* features. They are usually described with time sequences of notes, which are the single sounds produced by the different instruments. Rhythm, melody, and harmony are culturally dependent, because each musical style and genre have different traditions, and sometimes rules, on how to deal with them. Many musical works are based on the introduction, variation, modification, and reprise of some given rhythmic, melodic, and harmonic material, which tends to be characteristic of each music genre.

Structure depends on how short-term and middle-term features are organized on a higher time scale, and is a *long-term* feature. The ability to perceive and recognize a musical structure depends on the musical education, knowledge of the styles of the music genre, exposure to previous musical works, and active listening. A number of different theories have been proposed on musical structure by musicologists, among which the most popular ones are the Generative Theory of Tonal Music [72] and the Implication-Realization Model [102]. Details of these theories are beyond the scope of this overview. What is relevant for the aims of this paper is that in both cases it is stated that listeners perceive music as structured and consisting of different basic elements. The visual representation of musical structure is a difficult task; the interested reader may refer to [67] for a complete overview.

Table 2.1 summarizes the content carried by the different dimensions. The ability to convey all these dimensions depend on the particular music form in which a work is instantiated.

2.3 Music Forms

A simple consideration is that music is both a composing and a performing art. Users may be interested in enjoying either the formal work of a composer without the need of hearing a performance, or the playing

Table 2.1 Time scale and characteristics of the different dimensions of music.

Time scale	Dimension	Content
Short term	Timbre	Quality of the produced sound
	Orchestration	Sources of sound production
	Acoustics	Quality of the recorded sound
Middle term	Rhythm	Patterns of sound onsets
	Melody	Sequences of notes
	Harmony	Sequences of chords
Long term	Structure	Organization of the musical work

technique of a musician without a particular emphasis on the work that is actually played. Clearly, the combination of both aspects is the one users are more likely interested in. What is important for this discussion is that the communication in music is carried out at two levels. A typical scenario is the common praxis of Western classical music, which can be taken as example.

At the first level, the composer communicates to the musicians, which musical gestures have to be performed by way of a symbolic representation called the musical *score*. The symbolic score is the form for representing a musical work as a composition. The information represented in the score and the way gestures are associated with symbols depend on many factors, in particular on the culture, on the music genre, and on the instruments. The representation mostly used in Western countries, which is one of the possible examples of symbolic notation, was introduced by Guido d'Arezzo at the beginning of the 11th century and has been used in its modern form since the 14th century. The same representation has also been applied with minor modifications to other music genres, namely traditional music of many countries, pop, rock, and jazz. For this reason, the Western musical notation is the reference that is almost always assumed in MIR. A definition of score is the following:

> A musical **score** is a structured organization of symbols, which correspond to acoustic events and describe the gestures needed for their production.

At the second level, the musicians communicate to the audience by interpreting the symbols reported in the musical score, carrying out their personal realization of the required gestures on the musical instruments (including the human voice), and producing a sequence of acoustic events called the *performance*. A performance is then the realization of a process that converts symbols into sounds, according to the personal interpretation of the musicians. The recording of the acoustic rendering is the form for representing a musical work as a performance, which can be defined as follows:

> A musical **performance** is made of a sequence of gestures performed by musicians on their musical instruments; the result is a continuous flow of acoustic waves, which correspond to the vibration induced on musical instruments or produced by the human voice.

There are other scenarios that are worth to mention. Traditional and folk songs are usually transmitted orally, and the symbolic scores are the results of transcriptions eventually made decades or centuries after the songs have been composed. Pop and rock musical works are often composed and performed by the same persons, who may not need a symbolic representation of the musical gestures, which may be recollected from memory or with the aid of previous recordings. In this case the symbolic score, which may not be available at all, is the result of a transcription made *a posteriori* from recordings. Moreover, composition and performance are compounded in a single gesture in improvised music. Starting from existing musical works, the jazz musicians may create new melodies, harmonies, and rhythms such that any acoustic recording of a performance may be completely different from the other ones.

Even in the more straightforward situation, where a musical work is first composed and written in a symbolic score and afterwards performed and recorded, the two forms carry significantly different information and are of interest for different typologies of users and information needs.

2.3.1 Information in symbolic scores

As previously defined, a musical score is a symbolic description of a musical work that allows musicians to produce a correct performance. In the following, the discussion will focus on the symbolic notation used in Western countries, which is used to represent most of the music genres addressed by MIR systems; the problem of the bias toward Western music, and thus toward its representation, is common in MIR and is motivated by the large majority of users that live in Western countries together with the spread in the world of many Western music genres.

It requires some musical background to decode the most commonly used symbols of a musical score, but years of study and practice are needed to completely understand the complexity of many scores and recreate a valuable performance. Thus a large majority of users, the ones with little or no musical training, are not interested in symbolic scores, because they cannot exploit this information or even use the score to create their own performance.

The central role played by symbolic representation, which is the main tool used by musicologists in their research, put a particular emphasis on the musical parameters that are easily and directly represented in symbolic notation [91]. These parameters may be classified in two groups:

- **General parameters**: main tonality, modulations, time signatures, tempi, musical form, number of voices and instruments, and repetitions.
- **Local parameters**: tones that have to be played by each instrument, with their relative position in the flow of music events, their duration, and a possible indication of their intensity.

General parameters describe some overall properties of the musical work. Yet, they may not be good discriminants between different musical works. For instance, in tonal Western music there are only

30 different tonalities,[2] 15 major and 15 minor, while probably more than 99% of pop and rock songs have a time signature of 4/4.

From local parameters it is possible to extract information about the melody, the harmony, and the rhythm. The task of computing the melodic information is trivial when different voices are assigned to different musical instruments, each in a single staff, but becomes difficult also for expert musicians when different voices are played by the same polyphonic instrument in a Fugue or in a counterpoint piece. The task of computing the harmony depends on the complexity of the symbolic score, becoming trivial in the case of many pop, rock, and jazz music scores, which already report the chord names written over the melody.

Information about sound intensity, or loudness, is usually vague and expressed in a subjective scale from very soft (***ppp***, *Più che Pianissimo*: softer than the softest sound) to very loud (***fff***, *Più che Fortissimo*: louder than the loudest sound). The symbolic score carries almost no information about other relevant music qualities or dimensions, in particular timbre, articulation, room acoustics, and spatialization of sound sources, which all play a fundamental role in the experience of music listening and are related to music as a performing art. When present, the indications of the musical instruments and the playing techniques may suggest the expected orchestration and timbre of the performance.

Given these considerations, it may be argued that, with perhaps the exception of automatic generated music, where a score codes the algorithms used to produce the performance, the score is only an approximate representation of a musical work, because it is impossible to represent all the possible nuances of musical gestures with a compact symbolic representation. Nevertheless, the musical score has always played an important role in music access, considering sometimes the symbolic representation as the ideal version of a musical work, to which performances are only approximations [91].

[2] The number of tonalities that are perceived as different is slightly smaller, because of enharmonic equivalents—tones that are written differently but actually sounds exactly the same with equal-tempered intonation, such as $C\sharp$ and $D\flat$.

2.3.2 Information in audio performances

Ideally, a performance contains all the information about a musical work that is carried by a symbolic score. The musicians add their personal interpretation, and thus new information, according to the degrees of freedom left by the symbolic score, but usually the directions on dimensions like rhythm, melody, and harmony are maintained. Moreover, as previously mentioned, recordings of performances may be the only form available for many music genres, from jazz to traditional music. Finally, all users may enjoy listening to a music performance regardless of their musical education.

The recordings of performances may be of interest for musicologists, music theorists, and musicians because they may study how one or more performers interpreted a musical work. The choice of tempo, the presence of rallentando and accelerando, the intensity of the different tones, and the timbre are all interesting information carried by the recording of a performance. The room acoustics, the distortions of the audio equipment used for the recording, and the effects of the digitization and of the possible compression are of interest for audio engineers and researchers in digital audio processing.

On the other hand, it may be very difficult, also for trained musicians, to learn how to perform a musical work starting only from the recording of a performance, whether many musicians can play in a reasonably correct way a new musical work by reading a symbolic score for the first time. In the case of polyphonic music, the tracking and recognition of the single musical gestures is a time-consuming and error-prone task. Also the study of the structure of a musical work is more difficult using only the performance. The analysis of the composition itself may be biased by the particular interpretation given by the performers.

Even when a symbolic score exists, performances of traditional, pop and rock, and jazz music may carry additional information also on the orchestration and on the arrangement, because these genres allow performers more freedom on these dimensions than Western classical music does. The scores themselves are less detailed and give more general indications on how the compositions have to be performed.

Score and performance allows users for the extraction to almost complementary information. Melodic lines, harmony, exact timing, and structure are accessible from the former, while timbre, expressive timing, and acoustic parameters are accessible from the latter. For this reason, systems for music retrieval have to deal with both forms.

2.3.3 Versions of musical works

Besides being represented in symbolic or acoustic forms, each musical work may be instantiated in different digital documents that may vary remarkably.

Differences between symbolic scores of the same musical work may be due to different revisions of original material, which correct possible errors in the manuscripts, propose alternate approaches to the performances, or are transcriptions for other musical instruments—with possible changes in the tonality and addition/deletion of the musical material. When symbolic scores are transcriptions of acoustic recordings, the differences may be due to imprecise or simplified notations. The problem of the coexistence of different and alternative versions is common also to other art forms and other media, and it has been faced in the research area on digital libraries. As an example, a digital library approach has been applied to compare and visualize different versions of *Don Quixote de la Mancha* by Cervantes [65]. A similar approach could also be extended to the visualization of music score and to the retrieval of alternative parts.

Performances are the result of the act of interpretation of a musical work—which may be represented in one or more musical scores or recollected from memory—or the results of an act of improvisation. To this end, each performance is different from the others. Moreover, the noise of the environment, the quality of the audio equipment used for the recording, and the post-processing contribute to substantial differences between performances.

The presence of alternative versions of a musical work plays an important role in the concept of relevance to a given user information need. It is expected that all the alternative versions share most of the musical dimensions; yet the difference may be particularly important for the user to judge whether or not a musical document is relevant.

For instance, the original version of *Yesterday* by The Beatles and the orchestral version played by The London Symphony Orchestra—and most of the hundreds of cover versions that have been released—share melody and harmony, but may not all be relevant for a user. On the other hand, also a cover version of *Yesterday* by a garage band available from a weblog, even if the goal was to be as close as possible to the original version, is probably not relevant for many users. This particular aspect of relevance of musical documents has not been addressed in content-based MIR, apart in the case of audio fingerprinting, where particular recordings are tracked for copyright issues. Nevertheless, they will probably become more interesting when MIR systems will be available to a large majority of users.

2.4 Formats of Musical Documents

A number of different digital formats correspond to each of the two forms. Apart from the peculiar characteristics of the forms, different formats are able to capture only a reduced number of dimensions. Therefore, the choice of a representation format has a direct impact on the degree to which a music retrieval system can describe each dimension.

With the aim of taking into account all the variety in which music information can be represented, the Standard Music Description Language (SMDL) [57], as an application of the Standard ISO/IEC Hyper-media/Time-based Structuring Language, has been proposed. In SMDL, a musical work is divided into different domains, each one dealing with different aspects, from visual to gestural, and analytical. SMDL provides a linking mechanism to external, preexisting formats for visual representation or storage of performances. Hence SMDL may be a useful way for music representation standardization, but the solution is just to collect different formats rather than proposing a new one that is able to deal with all the aspects and dimensions of music.

2.4.1 Symbolic formats

The main aim of most symbolic formats is to produce a high quality printout of musical scores. They are thus more oriented toward the visual representation of the information in the score and are not aimed at representing information about musical structures and dimensions.

In a sense, symbolic formats inherit the bias toward the direct representation of a part of the musical content that has been discussed in Section 2.3.1 for symbolic scores. As for textual documents, there are two approaches to music editing: graphical interfaces and markup languages.

The most popular commercial products for music editing are *Finale*, *Sibelius*, *Encore*—whose names reveal a bias toward Western classical music—together with software that takes into account both symbolic and acoustic formats such as *CuBase*. These commercial products provide a graphical user interface and are currently used by music editors—many printed music books have been authored with them. Unfortunately, these valuable digital documents are not available to the public for clear reasons of copyright protection. Although out of the scope of this overview, it is worth mentioning the Wedelmusic project [141], aimed at promoting the distribution of symbolic formats, granting the protection of copyright ownership thanks to a watermarking mechanism on the score printouts.

A visual format for the representation of symbolic scores is called *pianoroll*, which represents notes as horizontal lines, where the vertical position of the line depends on note pitch. The term pianoroll, used in many music editing software packages, derived from the punched rolls that were used in automatic pianos: The beginning of a line, which was actually a hole in the paper, corresponded to a note onset, and the length corresponded to note duration. Figure 2.1 displays the pianoroll view representing the same music excerpt shown in Fig. 1.1; this representation clarifies the horizontal and vertical dimensions of a musical work introduced in Section 2.2.

Portability of digital musical documents authored with commercial software is an issue. If users of text processing software complain about the portability of their documents to different platforms and operative systems, it is probably because they never tried with edited music. Moreover, there have been few attempts to develop software to convert from and to these formats. It has to be mentioned that Finale developed a format, called *Enigma*, whose specifications are public [15] and for which some converters have been already developed.

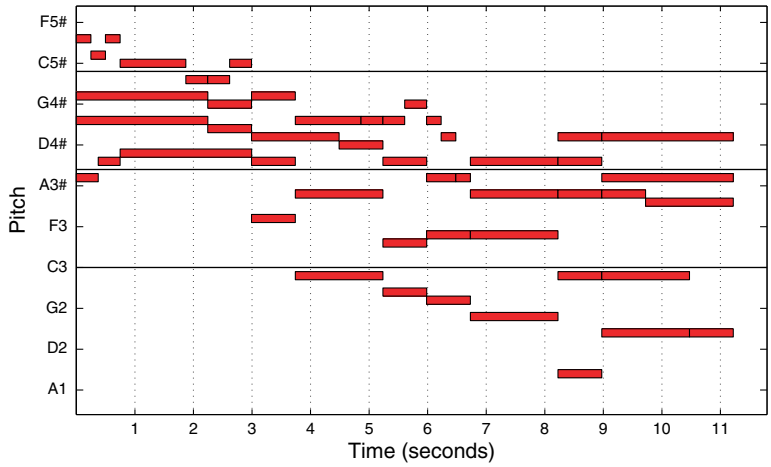

Fig. 2.1 Pianoroll view of the musical score shown in Fig. 1.1.

Markup languages for symbolic music editing are becoming increasingly popular, also because of the existence of converting tools that allows users to create large collections of public available symbolic scores. Moreover, the organizations that promote these emerging formats are usually involved in the creation of music digital libraries based on these formats to store documents.

Examples of well-known markup languages for representing symbolic scores are *ABC*, which is mainly used to code simple monophonic scores of traditional music, *GUIDO* [99] and *MuseData* [96] that have been developed at the University of Darmstadt and Stanford University, respectively, and the emerging format *Lilypond*, which is an open source project.[3] An interesting effort in music representation, which partially builds upon MuseData, is *MusicXML* [98] that exploits the powerful features of XML as a portable format for information exchange. MusicXML is supported both by commercial software and by open source projects. A set of macros and fonts that enable LaTex to typeset music have been developed too [101]. A comprehensive introduction of formats for the representation of symbolic scores can be found in [121], while on-line information can be found in [100].

[3] The scores depicted in this paper have been edited with Lilypond.

It is interesting to note that GUIDO has been developed by a research group that is also involved in MIR research [49], while the group MuseData supports the *Humdrum Toolkit* [51], which has been developed as a set of software tools intended to assist in music research. Hence these formats have also been created within an interest in music retrieval.

2.4.2 Audio formats

Audio formats are aimed at representing the digital recordings of performances. Digital acquisition is based on sampling of the signal and quantization of the sampled values. Most non-compressed audio formats are based on the Pulse Code Modulation (PCM) representation, where each sample is represented as a pulse with a given amplitude. Stereo or multichannel information is normally represented interleaving the samples taken from each channel.

The two relevant parameters of PCM audio formats are the *sampling rate* that is the number of equally spaced samples taken in the time unit, and the *amplitude resolution* that is the number of bits used to represent each sample. Both parameters are related to the perceived audio quality. The values chosen for the audio CD—44.1 kHz for the sampling rate and 16 bits for the amplitude resolution of each channel—become the standard for high-quality audio even if the actual technology allows digital audio for higher sampling rates and amplitude resolutions. Available audio formats—such as AIFF, WAVE, and AU—can be easily converted from one to another. Hence, the choice of the audio format is not a relevant issue. An example of a PCM representation of an audio signal is shown in Fig. 2.2, the audio is a piano performance of the score shown in Fig. 1.1.

Formats based on PCM are not particularly efficient in memory occupation: 10 seconds of complete silence need the same amount of space as 10 seconds of a *tutti*, i.e., all the ensemble play simultaneously, performed by a symphonic orchestra. Compressed formats have been developed to overcome this limitation. Among the different compressed formats, the MPEG 1, Layer 3 (MP3) is the best known [94]. An addi-

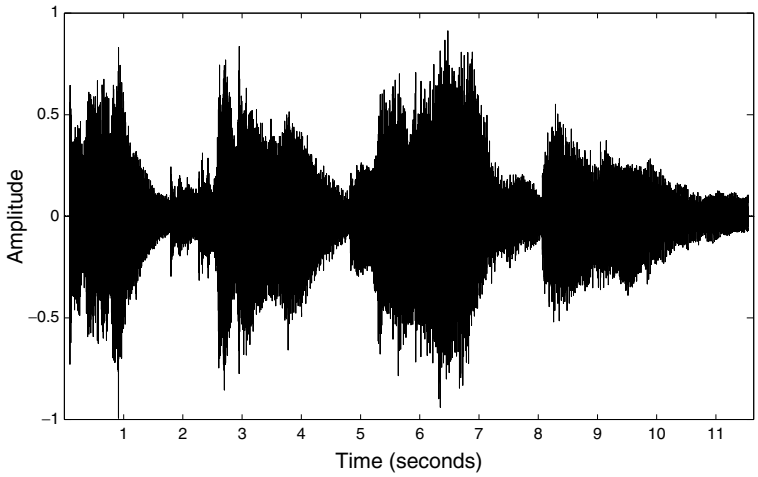

Fig. 2.2 Time domain representation of an audio excerpt corresponding to a performance of the score shown in Fig. 1.1.

tional parameter of compressed formats is the *bitrate*, which is the number of bits needed to encode a second of music. Other proprietary formats have been proposed, in particular the Windows Media Audio that achieve good compression with a good audio quality. Clearly, the lack of public specifications of these formats makes them less suitable for MIR tasks.

There is a number of audio formats that, instead of representing the evolution of the sound, describe the synthesis technique and the synthesis parameters required to generate a given sound, or performance. They can be considered the acoustic parallel of vector graphics and are particularly useful for representing synthetic sounds. These formats achieve a very high compression rate, because the control parameters change much slower than the audio parameters and have been included in the MPEG-4 standard. At the state of the art, these formats did not find an application in MIR, mainly because they are limited to a class of sounds that are not of particular interest for the final users.

Another evolution of the MPEG standard is MPEG-7, which takes into account also the coding of audio descriptors, in particular for the

timbre dimension that can be used for MIR tasks. An approach to audio classification based on MPEG-7 descriptors has been proposed in [2].

2.4.3 The musical instrument digital interface

A format that can be considered as a compromise between symbolic and acoustic formats is the Musical Instrument Digital Interface format [120], or simply Midi. This standard has been proposed in 1982 for data exchange among digital instruments, and thus it is oriented toward the representation of musical gestures performed on digital keyboards. The communication is achieved through the exchange of messages on a sequential interface, with information on which keys have been played or released, the intensity, and the voice (or channel) to which they are assigned.

The use of computers to control external audio devices or software synthesizers motivated the definition of a file format for storing complete performances; a file coded according to this format is called a Standard Midi File (SMF). The specification for SMF defines the structure of Midi messages and a number of global parameters of the musical work, such as key and time signatures or tempo. Structured metadata can be coded with SMF, including information about authorship, title of the musical work, and copyright information, yet the presence of metadata is not mandatory. Most of the SMFs follow the General Midi convention, which defines an association between the 128 Midi channels and the synthetic timbre of musical instruments. SMFs are commonly used to control synthesizers, yet the automatic performances may differ dramatically, depending on the quality of audio devices and on the synthesis techniques.

Because Midi carries information about musical gestures, it is possible to reconstruct an approximate representation of the symbolic score. In a sense, Midi draws a link between the two forms for the representation of musical works. This characteristic, together with the fortune of Midi as an exchange format in the early times of the Internet, can explain why many MIR approaches are based on this format. As for today, there are still hundreds of thousands of SMFs available on the

Web: Downloading them is still a cheap way to create a test collection without running the risk of copyright infringement.[4]

Midi is becoming obsolete because users on the Internet increasingly prefer to exchange digital music stored in compressed audio formats such as MP3. Being a compromise between two different needs—to represent symbols and to be playable—Midi turns out to fit neither the needs of users who want to access a complete digital representation of the symbolic score, nor the needs of users who want to listen to high-quality audio. A number of software tools have been developed to convert SMFs to most of the formats presented in Section 2.4.1, together with tools to create digital recordings of Midi performances. It can be foreseen that the popularity of Midi will slowly fade away at least as a file format. On the other hand, Midi still has a niche of users in interactive music applications and performances, as a standard format for data interchange across different hardware and software music devices.

[4] It could be argued that the author of the SMF, that is the person who sequenced it, owns the same rights of a music editor. In general, it is assumed that the labor needed to obtain a useful printout from Midi files discourages this kind of activity, while authors grant for free the possibility to download and listen to the files. Of course the musical work itself can be copyrighted, and it is probably possible to find these files only because they are not considered interesting for most users.

3

The Role of the User

The concept of *information need* is central in IR. Users interact with a search engine to retrieve information that is relevant to their needs. A survey on music-related information needs [71], carried out on about 500 students at university level, highlighted that information about music content—especially in the form of title, lyrics, artist, and genre—is relevant for a great part of listeners. Yet, searching for music information is usually motivated by a subsequent access to musical documents. According to [71], there are three main reasons why users may want to access digital music:

(1) listening to a particular performance or musical work;
(2) building a collection of music;
(3) verifying or identifying works.

Another crucial aspect is related to how well users can describe their information needs. As it usually happens in IR, the degree of expertise of the user on the particular application domain has a strong impact on retrieval effectiveness, which in the case of MIR may vary dramatically. For instance, the knowledge of the musical theory and practice, of the dimensions conveyed by the two forms and their different formats, may

range from the complete understanding of scholars and professionals to a vague set of notions of casual users. For the sake of simplicity, potential users of MIR systems are divided in three categories: (i) casual users want to enjoy music, listening and collecting the music they like and discovering new *good* music; (ii) professional users need music suitable for particular usages related to their activities, which may be in media production or for advertisements; (iii) music scholars, music theorists, musicologists, and musicians are interested in studying music.

3.1 Casual Users

There is a large number of users that enjoy music even if they do not have a musical training. Users of this kind do not necessarily know about most of the music dimensions and thus are not able to describe precisely and effectively their information needs. In this case, the query-by-example (QBE) paradigm can be useful to help casual users in describing their information needs in a intuitive way. Normally, casual users are interested in the acoustic form, because their goal is listening to or collecting music. Furthermore, they may not be able to interpret the symbols in the symbolic score. Like for most of the media, casual users represent the majority of potential users of a MIR system.

Some typical information needs are reported in the following, as a representative sampling of the needs that have been presented in research work in MIR addressing casual users. It has been chosen to represent the information needs in the form of a direct request that could be made either to an expert in the field or to an automatic system.

Find me a song that sounds like this

This is the classic example of a MIR task, which is often used to introduce music retrieval systems and approaches to people outside the MIR research community. The typical scenario is a user having in mind some music dimensions, usually the melody and/or the rhythm, of an unknown piece of music that he heard recently or that came to his mind from long-term memory; a situation that happened almost to everybody.

In this case there is only one musical work that is relevant to the user's information need—e.g., the song was *Knockin' on Heaven's Door* by Bob Dylan—and sometimes the user is interested in a particular recording of that musical work—e.g., the version was the one performed by Guns n' Roses and included in the *Use Your Illusion II* album. The search task is a particular case of information retrieval, where the relevance of retrieved documents depends on their specific characteristics rather than on the information they convey. An analogous in Web IR can be the *homepage finding* task presented at the TREC [130] Web track, where the only relevant document is the Web page maintained by a given person or organization—e.g., the local Bob Dylan Fun Club—while Web pages describing the person or the organization are not considered relevant.

The use of the QBE paradigm is mandatory, because the only information available to the user is a part of the document content. An example can be provided by the user by singing a melodic/rhythmic excerpt or, when available, by sending a recording of the unknown musical work. In most of the cases, it is assumed that the users want to access a digital recording of the searched musical piece in order to listen eventually to it. Complete metadata information, such as author, title, availability of the recording, may also be interesting for the user.

For this particular information need, the evaluation of the relevance of the retrieved musical documents is straightforward. A person who already knew the musical work corresponding to the user's query—and often the user himself/herself—may easily judge, using a binary scale, whether or not the retrieved musical documents are relevant.

> *Given that I like these songs, find me more songs that I may enjoy*

This is more an information filtering than an information retrieval task, and it can be considered the extension to the music domain of automatic recommendation. Items to be recommended may be either complete CDs, for which this is a typical application of e-commerce [75], or single digital recordings in compressed format in the more recent case of Web services to download music. The evaluation of the relevance of

suggested musical items can be carried out, for music and other media, measuring how well a recommender system predicts the behavior of real users, normally using a manifold cross validation.

The literature on automatic recommendation is generally oriented toward the exploitation of customers' behavior—the choice of buying or rating particular items—as in the case of collaborative filtering. This approach has some drawbacks: the *new item* problem, that is an item that has not been bought or rated cannot be recommended, and the *cold start* problem, that is new customers do not have a reliable profile. For this reason, recommender systems can also be content-based, as it is described in Section 5.2.

Content-based approaches to music recommendations have a motivation in theories on music perception, which are worth to be mentioned. The enjoyment of music listening can be due to the principles of *expectation-confirmation* and *expectation-denial*. When listening to a performance, the listeners create expectations on how the different music dimensions will evolve. When expectations are confirmed too often, the performance sounds uninteresting or even boring. On the other hand, when expectations are denied too often, the performance sounds incomprehensible and eventually boring. The ability of anticipating the evolution of a musical work depends on the user's knowledge of the music genre, besides a general knowledge of music theory and practice.

These considerations suggest that similarity in the musical content can be a viable approach to fulfill this information need, in order to provide the user with previously unknown music that can be enjoyed because the expectation-confirmation and expectation-denial mechanisms may still hold thanks to the similarity with known and appreciated musical works [68].

> *I need to organize my personal collection of digital music (stored in my hard drive, portable device, MP3 player, cell phone, . . .)*

The increasing availability of memory storage at low cost and the ease with which users may download and share music need to be paired

by tools for easy organization of the musical documents. The task, which is more a classification/categorization task, has to be carried out using a content-based approach, because metadata may be incomplete, incorrect, or even totally missing. Moreover, as already mentioned, metadata may be too generic to be useful for a classification task, e.g., to classify all the hundreds of ballads together in the same category may not be particularly useful for the final user.

The *music store* metaphor, where musical documents are organized hierarchically depending on genre, subgenre, period, and author, may not be a suitable representation for large collections of documents. In general, any hierarchical representation, and in particular the directory tree, has the problem that many documents may be left forgotten somewhere especially when they are stored in handheld devices with reduced displaying capability. To this end, other interaction metaphors need to be designed and developed in order to achieve this task. A bidimensional or tridimensional representation of the music collection may be a viable solution, for instance representing music similarity with the distance between musical works or mapping songs depending on their genre and style.

A facet of the need for organizing a personal collection, which is peculiar to music, is the *automatic playlist generation* problem. For the ones that are not used to listen to music through portable devices or Web radios, a playlist can be defined as a list of songs, played in sequential order, that is selected according to a specific goal. The goal may range from enjoying a given artist, genre, or orchestration to propose a path across music to a friend, to listening to songs with a particular mood. The typical situation is that a user wants to have a subset of his/her personal collection organized in a suitable sequence of songs that can be listened to.

The goal is to suggest a list of songs to the user, presenting familiar songs that the user surely likes together with less familiar songs that are similar to the known ones and thus potentially interesting. The ordering of the songs in the playlist is important, because local similarity may be more relevant than global similarity and the novelty has to be spread across the complete playlist. Approaches to playlist generation are presented in Section 5.2.1.

3.2 Professional Users

A number of users may need to access music collections because of their professional activity. Apart from the typical case of radio and television music broadcast, other applications regard the use of music for the soundtrack of movies, commercials, news stories, documentaries, reports, and even Web pages, or the organization of live performances and concerts. Music critics may be interested in retrieving musical works to draw comparisons with newly released albums.

Professional users usually have a good knowledge of the application domain and are familiar with the language of the given media. In the particular case of music language, professional users may be familiar only with a subset of the features and dimensions. For instance, the editor of a radio program of HipHop music may not be familiar with the concepts of tonalities, modulations, and structure. Vice versa, the organizer of concerts of Western classical music may not know, and even not notice, the acoustic and stylistic difference between genres such as hard rock and heavy metal. Moreover, it may be possible that some professional users are not able to interpret notated formats.

Even if it is difficult to generalize the kind of interaction suitable for any professional user, it may be assumed that professional users are able to describe in detail their information need. The retrieval task can be carried out with a QBE approach, but the user may specify which are the most relevant dimensions of the content provided as example.

I am looking for a suitable soundtrack for . . .

In many Western countries, music is ubiquitous: leading theme or leitmotiv in commercials, background in news stories, opening tune for TV serials plus background in restaurants and elevators, ringtone for cellphones, and welcome in answering machines. The ability of a musical work to capture the attention of the audience or to reinforce the message that is conveyed by the video is difficult to define. Consequently it is difficult to describe the information need with metadata. Also in this case, the QBE paradigm may be a viable solution. The professional user gives the system the reference to a number of possible works, which are relevant for some of their musical dimensions but

cannot be directly used for a number of reasons—costs for obtaining permission for commercial utilization, musical works already used in other contexts, and songs that are too familiar to the audience. Moreover, the quest for novelty for previously unknown musical works may promote the use of content-based systems.

It is important to highlight which are the music dimensions involved in this particular task. Timbre, acoustics, style, and genre are relevant dimensions when music is used as a background, because they are more easily perceived also when the attention focuses on something else. Melody and rhythm are important if the musical work has to be used as a signature, for a trademark or a TV serial, and it has to be easily remembered by listeners.

For this kind of information need, the professional users may be interested either in symbolic scores or in audio recordings, depending on the final application. As already mentioned in Section 2.3, the symbolic score allows musicians to produce new performances and carries explicit information on the structure, the melody, and the harmony that may be used to refine the search in a presentation–evaluation loop; the audio recordings carry unique information on the musical gestures of a particular performance. The professional users should be able to describe which are the forms and eventually the formats of interest. In any case, this task can be done using a typical IR approach, where many different documents are potentially relevant to the final user.

> *Retrieve musical works that have a rhythm (melody, harmony, orchestration) similar to this one*

Although it could be considered a subtask of the previous example, the retrieval based on specific dimensions may be motivated by different information needs, affecting the relevance of retrieved musical documents. For instance, a music critic may want to retrieve songs that could have been the source of inspiration for a composer or a performer, or to verify the novelty of a newly released musical work, or to check possible plagiarism. The consistency of a given dimension among musical works classified in the same music genre, e.g. the uptempo rhythm of reggae music, or the discovery of crossgenre dimensions, are other

possible aims of this particular task. Finally, the retrieval of musical works based on given features may be used to create thematic broadcasts and playlists. In particular, professional DJs may need to retrieve songs during their performances, as proposed in [62].

Even if the question implies a QBE approach, being based on a reference musical work, the professional users may be able to describe their information need also at an abstract level. For instance, they can refer to the most peculiar characteristic of a music genre, e.g., harmonic progression of Irish jigs, acoustics of new age music, or to high level descriptors of the musical dimensions, e.g., slow rhythm, distorted timbre. Also in this latter case, content-based retrieval is the most suitable approach.

The concept of music similarity based on its dimensions, which is implicit in many information needs of casual users, is directly involved in this task. A professional user is able to identify which are the dimensions that characterize a musical document, either in symbolic or in acoustic forms, and that can be used to retrieve musical works that are similar according to these dimensions. Clearly, the relevance of retrieved musical documents depends on the dimensions of interest and on the characteristic of the music genre. For example, the harmonic structure is not useful to distinguish two different blues, because the genre is defined because of a particular harmonic structure, while melody is not the most relevant feature in rap music.

3.3 Music Theorists, Musicologists, and Musicians

Differently from casual and professional users, these users are interested mostly in obtaining information from the musical works, which is the subject of their study, rather than obtaining the musical work itself. Music theorists need to access musical scores in order to develop or refine theories on the music language, the composition, and the analysis of musical works. Musicologists are interested in the work of composers, performers, the role of tradition, the cultural interchanges across genres, styles, and geographical areas. Musicians are interested in retrieving scores to be performed and to listen to how renowned musicians interpreted particular passages or complete works.

In general, the information needs of this category of users may be fulfilled either by a single document or by the characteristics of a set of documents, for instance, the consistent use of particular chord progressions by a composer or the use of rallentando by a director when performing musical works of a given period. It can be assumed that music theorists, musicologists, and musicians have a perfect knowledge of the application domain and about the meaning and the role of musical dimensions, at least on the genre they are specialized in. They should be able to express clearly their information needs, because they are aware of the meaning of all the dimensions, the terminology to describe them, and of all the metadata that can be associated with a musical work.

It is difficult to summarize the information needs in simple sentences as it has been done in the previous sections for casual and professional users, because they may regard all of the different aspects of music dimensions. Examples of retrieval tasks that may be useful for these users are reported in [52], which describes a set of music analysis tools aimed at extracting information from a collection of musical works in a symbolic format using regular expressions on the rhythm, melody, and harmony dimensions.

3.4 Interaction with the System

The QBE paradigm applied to the music domain may take advantage by the fact that almost every user is able to provide a music example through a short vocal performance. This is usually known by the MIR community as *query-by-humming*, the term being introduced already in 1995 by [33] (even if the authors cite a previous paper [61] written in Japanese). Of course, humming is only one way in which casual users can perform a melody, together with whistling and singing, while users with a musical education can play an instrument.[1] As MIR is broadening its scope, the term is becoming less popular, as can be seen from the decreasing number of papers on query-by-humming that

[1] As for July 2006, the ACM Digital Library reported 89 papers using the term query-by-humming, 13 query-by-singing, 1 query-by-whistling, and 2 query-by-playing.

are published at the International Conference on Music Information Retrieval (ISMIR) [59].

The idea of allowing the user to provide directly an example without the need of external digital documents has been applied also to other domains, such as image retrieval [63] and 3D models retrieval [32], where users may sketch their information need through a graphical interface. In the case of musical queries, the standard audio equipment of a personal computer is sufficient to allow the user to record his query. On the other hand, for Web-based systems it is not that easy to interact through audio, and systems based on query-by-humming need ad hoc pieces of software for audio acquisition [4, 7].

The interaction through the audio channel is an error-prone process. The query may be dramatically different from the original song that it is intended to represent, and also different from the user's intentions. In fact users may have an approximate recollection of the melody—after all they are looking for musical works they are not supposed to know very well. Psychological studies [22] showed that melodies are remembered by pitch intervals between successive notes; errors in recollection may result in a different sequence of intervals. Moreover, untrained singers can perform several errors due to a difficult control of the voice, and to an imprecise recollection from memory.

A significant contribution on user interaction through music has been given by the *Musical Audio Mining* project [83]. A group of subjects, with different degrees of musical education, was asked to hum, sing, or whistle melodic excerpts from a number of predefined songs. Typical errors are the massive use of glissandi to connect distant notes or to reach difficult notes; an additional vibrato due to imprecise tuning and consequent continuous pitch adjustment; a decrease of the fundamental frequency toward the end of the note; a number of false attacks before the real beginning of a note; and tempo fluctuations, with a tendency to shorten pauses. Another interesting result on user interaction, reported in [74], is that users prefer to sing the lyrics of the songs, which sometimes are the only cues that human listeners can effectively use to recognize the melodies.

The effectiveness of the retrieval may be affected by an error-prone interaction. For this reason, it has been proposed to take into account

users' performing errors either during the pitch tracking by modeling the point of view of the listener [124] or in a post-processing phase [86]. The knowledge of the uncertainty given by a particular query can be exploited also at retrieval time, as discussed in [139]. In general, it is possible to deal with errors in the transcription using quantization, as for some of the approaches described in Section 5.1.1, approximate string matching techniques, as described in Section 5.1.2, or computing the match between the query and the documents through continuous distance measures, as for the approach presented in [133] and described in Section 5.1.3.

4

Music Processing

As for any other medium, the first step toward music retrieval involves the automatic processing of the musical documents to extract relevant content descriptors. There are different approaches to music processing, depending on the form and format in which musical documents are instantiated and on the dimensions of interest. As it can be expected, great part of the research on feature extraction has been devoted to the audio form, from which most of the music dimensions are particularly challenging to extract.

The analysis of the publications on feature extraction aimed at the development of MIR systems shows a clear drift from symbolic toward audio forms. Early works on MIR focused on the symbolic form, because relevant melodic features were easier to extract and also because MP3 was not as popular and pervasive as it is nowadays. SMF was the common format for musical documents, and it has always been considered that Midi is more relevant for its symbolic representation of a score rather than for the information it contains on temporal and intensity performing parameters. Many approaches were based on the

query-by-humming paradigm, described in Section 3.4, that is based almost only on melodic information. For these reasons, great part of the work on feature extraction from the symbolic form is devoted to melodic information.

The growing diffusion of music in digital audio format, together with the advances in the automatic extraction of relevant parameters from audio, motivated the shift toward audio-based MIR systems. The corresponding increase in the number of users that regularly access musical documents, with the corresponding broadening of users' information needs, motivated also the research on other approaches to music access as described in Section 3.1, such as information filtering and clustering. These approaches are based on a wider number of music dimensions.

An overview of problems and methods for the automatic computation of music dimensions from symbolic and audio formats, respectively, is given in the following sections.

4.1 Symbolic Form: Melody

The automatic computation of melodic information from symbolic formats is the cornerstone of query-by-humming systems. The difficulties of this task depend on the typology of musical work and on the format in which the work is represented.

4.1.1 Extraction of the main melodies

The computation of melodic information from a monophonic score is straightforward, because any symbolic format represents directly the melody content that can be immediately processed. The melody is already represented as a unique and clearly identifiable sequence of symbols. A slightly more difficult task is given by a polyphonic score made of several monophonic voices, which is usually addressed as counterpoint. If it is assumed that all the melodic lines are equally relevant content descriptors, the task can be carried out like for a collection of independent monophonic scores, i.e., by extracting as many melodies as the different voices in the score.

Yet, for many genres, and in particular for pop and rock music, it is assumed that there is only a relevant melodic line—usually assigned to the voice—and all the other sounds are part of the accompaniment—assigned to guitars, keyboards, bass, and drums. The task that has to be carried out is the automatic recognition of the main voice. To this end, statistical approaches have been proposed, starting from a training corpus of manually annotated musical scores. The idea is to describe each voice with a number of features that may help discriminating sung melodies from other voices, including mean and variance of the pitch, range, mean and variance of the difference between subsequent notes, and relative length of the voice in relationship with the total length. An approach to the computation of the main voice, or theme, of a polyphonic score is presented in [47].

A more difficult task is to extract the main melody from a polyphonic score when the voices are mixed together and when chords are played together with single notes. The extraction of the main melody can be carried out by exploiting previous knowledge of the perception of melodic lines, and on how composers organize sounds. An example is the system ThemeFinder, presented in [87]. Figure 4.1 shows the main melody, as it is normally perceived by listeners, of the music excerpt shown in Fig. 1.1.

The effectiveness of the automatic computation of the main melody depends also on how users will perform the same task, which is a difficult task at least for listeners without a musical education. In fact, if the query-by-humming approach is used, it is also assumed that the users will recognize the main melody and use it as their query. To this end, an interesting user study is reported in [137]. A number of subjects were asked to assess the effectiveness by which different algorithms extracted the main melody from a polyphonic score. Quite surprisingly, subjects gave a higher score to the algorithm that purely extracted the

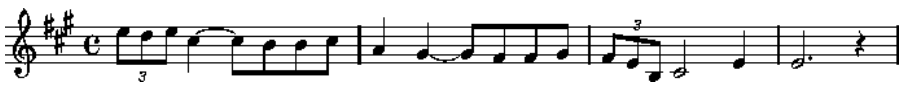

Fig. 4.1 Main melody of the score shown in Fig. 1.1.

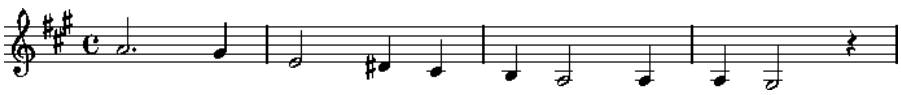

Fig. 4.2 A melodic line extracted from the score shown in Fig. 1.1.

highest note in the polyphony, even when the final extracted melody was clearly a mixing of different voices.[1]

The extraction of the main melody is an error-prone process. For instance, the extracted melody can have notes from other voices or the wrong voice can be picked up as the representative melody. As an example, Fig. 4.2 represents a descending melodic line that, although being extracted from the score shown in Fig. 1.1, is unlikely to be remembered by the users as its most representative melody. On the other hand, the extraction of the main melody is not a necessary step for approaches that are able to process and compare polyphonic scores directly, as proposed in [76] and in [14].

4.1.2 Segmentation of the melody

Once the melody has been extracted from a symbolic score, the further step is to represent its content in a compact and efficient way. This step is not mandatory, and there are approaches to MIR that are based on the representation of the complete melody as a, possibly long, sequence of notes [4]. Alternatively, as in the work presented in [123], melodic information can be represented by the main theme of a given song. A melody can be viewed as a sequence of thousands of symbols; normally a theme is much shorter because repetitions, as in choruses and verses, are not considered.

There is a number of approaches that focus on a higher granularity for representing the melodic content. The basic idea is that short subsequences of the main melody are efficient descriptors of its content. Drawing a parallel with text IR, the basic idea is to find the *lexical*

[1] For example, in the case of the opening bars of *Yesterday* by The Beatles, the algorithm output would have been: the melody on the words *yesterday*, two descending notes of the bass line, the melody on the words *all my troubles seemed so far away*, two more notes of the bass line, and so on.

units of melodic information. Yet, differently from text, melody is a continuous sequence of events without explicit separators. Single notes cannot be considered as effective content descriptors, because relevant melodic information is carried by the combination of notes. Moreover, pauses or rests do not play the same role of blanks in text documents, because they are not necessarily related to a boundary between two subsequent lexical units.

It is therefore necessary to automatically detect the lexical units of a musical document to be used as index terms. Different strategies to melodic segmentation can be applied, each one focusing on particular aspects of melodic information.

A simple segmentation approach consists on the extraction from a melody of all the subsequences of exactly N notes, called N-grams. N-grams may overlap, because no assumption is made on the possible starting point of a theme, neither on the possible repetitions of relevant music passages. The idea underlying this approach is that the effect of musically irrelevant N-grams will be compensated by the presence of all the relevant ones. This approach is quite popular in MIR and can be found in [26] and, with variants to take into the account the polyphony, in [21]. The N-grams approach can be extended by considering that typical passages of a given melody tend to be repeated many times, because of the presence of different choruses in the score or of the use of similar melodic material [113]. Each sequence, of any length, that is repeated at least K times can be used as a content descriptor of the melodic information. Segments can be truncated by applying a given threshold, as suggested in [103].

Alternatively, melodies can be segmented by exploiting *a priori* knowledge of the music domain. According to theories on human perception, listeners have the ability to segment melodies in smaller units, which can be an effective descriptor because they capture melodic information that appears to be relevant for users. Even if the ability of segmenting music may vary depending on the level of musical training, a number of strategies can be generalized for all listeners. Computational approaches have been proposed in the literature for the automatic emulation of listeners' behavior [128]. An alternative approach to segmentation is based on the knowledge of music theory, in particular

for classical music. According to music theorists, music is based on the combination of musical structures [72], which can be inferred by applying a number of rules. A MIR system exploiting this latter approach is presented in [89], where the segmentation algorithm is the one proposed in [10], where differences in melodic intervals and relative note durations are used as clues to automatically detect boundaries between segments.

In order to test the effectiveness of segmentation algorithms, an evaluation study has been carried out on manual segmentation performed by a number of subjects with a musical education, which has been compared to automatic segmentation [88]. Results showed a good consistency across the subjects, although for some music excerpts the segmentation task proved to be particularly difficult. Moreover, automatic segmentation algorithms have the tendency to oversegment the melodies, i.e., the lexical units are in general shorter than the ones highlighted by subjects.

4.2 Symbolic Form: Harmony

Even if harmonic information cannot be directly used in a query-by-humming system, chord sequences are considered as a relevant descriptor of musical works, which can be used to compare or cluster musical documents. The automatic computation of harmonic information is a difficult task also for users, even with symbolic scores. Almost anyone can recognize a melody and sing it, but it takes time, practice, and effort to transcribe the chords of a song. In fact, there are a lot of Web pages dedicated to chord transcription of famous recordings, and musicians (especially guitar players of pop and rock music, which uses guitar tablatures) exchange chord transcription through USENET, mailing lists, and discussion groups.

A chord usually lasts for a longer time than the notes of the melody, and the same chord can be recognized also when the actual notes in the polyphony change. In fact, each chord can be realized by a considerable number of possible combinations of notes. For this reason, recognition and segmentation are usually carried out at the same time. The common assumption in chord transcription is that there are no parallel

harmonic progressions, that is, chord sequences are monodimensional. Even if different chord sequences may sound reasonably good with the same melody, and this is a typical approach to jazz music, where musicians modify the chord sequence according to a particular style,[2] it is normally assumed that a single correct chord sequence exists.

A statistical approach to label contiguous regions of a score with information on the chord—together with additional information on the key signature and the mode—is introduced in [117]. The analysis is performed with a hidden Markov model [116], which is trained from examples of symbolic polyphonic scores and is able to find the globally optimal harmonic labeling.

Another statistical approach to directly model polyphonic information is presented in [69], where horizontal and vertical dimensions of music (see Section 2.2) are handled in single framework using Markov random fields [114]. Although the model is not aimed at annotating chord progressions with textual labels, it can be used to directly model polyphonic information, in a similar fashion as hidden Markov models are used for monophonic information.

A final step in the processing of harmonic information is the transformation of the extracted features in a suitable representation. Music theory offers a variety of different styles to represent the information on chord progressions according to the chord name, e.g., CM7, Dm7, F/C, ..., to its harmonic function, e.g., I7, ii7, IVc, ..., or using a numerical notation called *figured bass*. The work reported in [44] is an interesting proposal on the representation of chords, based on a syntax with a simple tree structure and implemented with a textual markup language.

4.3 Audio Form: Timbre

Among the different dimensions, timbre is probably the most difficult to define and characterize [66]. The standard definition of timbre, reported in Section 1.1, is in negative form: Timbre is defined as the acoustic feature that is neither pitch nor intensity. It has to be mentioned that

[2] Charlie Parker's and John Coltrane's modifications of well-known chord sequences are two typical examples.

in Western classical music the choice of timbre parameters is almost completely left to the performers, because the indication on how to play—if present—is usually generic and open to interpretation. Nevertheless, listeners seem to be very sensitive to this dimension, and in fact the particular timbre of a musician is addressed as his *sound*.

The perception of timbre is related to the structure of the spectrum of a signal, i.e., by its representation in the frequency domain. For this reason, the Fourier transform is one of the most frequently used tools for the analysis of timbre, and audio analysis in general. Temporal evolution of the sound is also important, and for this reason a common way to represent an audio excerpt is the *spectrogram*, which is a bidimensional plot, where the X-axis represents time, the Y-axis represents frequency, and the coloring of each element represents the energy. Figure 4.3 shows the spectrogram of the audio excerpt displayed in Fig. 2.2; some similarities can be found between the spectrogram and the pianoroll representation of the score shown in Fig. 2.1.

The spectrogram and other alternative time-frequency representations are not suitable content descriptors, because of their high

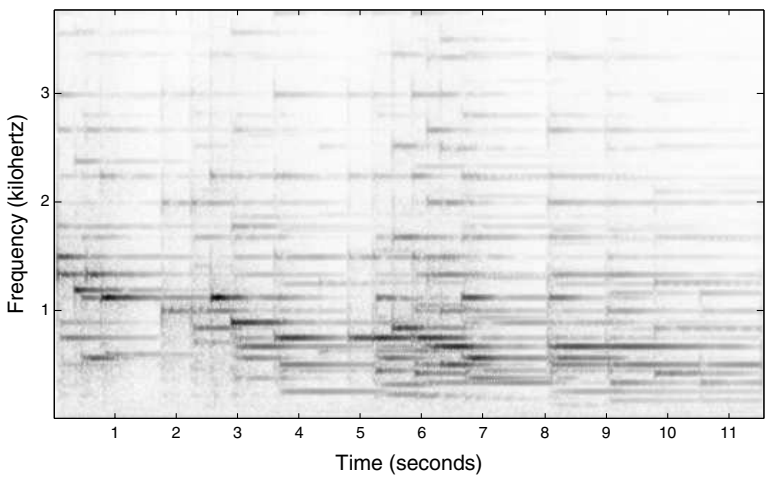

Fig. 4.3 Spectrogram representation of an audio excerpt.

dimensionality, considering also that the amount of data increases with time duration and time-frequency resolution. Moreover, variations in the time-frequency representation may not correspond to variations in the perceived timbre.

Timbre characterization is based on the extraction of low-dimensional parameters, which should correspond to perceptually relevant qualities of the sound. Studies in psychoacoustics highlighted for instance that parameters such as *attack time*, *spectral centroid*, and *roughness* can be considered as relevant content descriptors of the timbre dimension. The interested reader can find a comprehensive discussion of the acoustic parameters and their perceptual correlates in [84]. A set of descriptors that have been extensively used in MIR are the Mel-Frequency Cepstral Coefficients (MFCCs), which have been introduced in [18] for efficient speech recognition. MFCCs have become the state of the art for speech recognition and have been used also in related areas as a content descriptor of musical sounds.

In principle, any transformation of the time-frequency representation of musical sounds may be used to represent timbre information, and probably the analogous ones of MFCCs for musical sounds have not been found yet but will be found in the future.

Timbre features are extensively used in many aspects of MIR, because it is believed that users are particularly sensible to timbre, and less sensible to other middle- and long-term features.

4.4 Audio Form: Orchestration

The recognition of orchestration makes sense only for the audio form, because the symbolic form may either clearly report the names of the musical instruments or not contain this information at all. In this context, orchestration is related to the recognition of the sources, i.e., the musical instruments, involved in the sound production. The style by which they are played and the final sound they contribute are more related to the perception of timbre.

It has to be noted that the recognition of sound sources, though having a long tradition in the computer music research area [39], is not a major topic in MIR. This is probably because the task of recognizing

musical instrument in a polyphonic performance, where the number of instruments is higher than the number of tracks in the recording, is still a very difficult task. Good results in terms of recognition rate have been obtained with monophonic—or polyphonic but with a single sound source—recordings, and with a small amount of different sources. For instance, the results reported in [140] for single instrument identification show that the recognition rate range from 100% for easy recognizable instruments like the flute to 75% for more difficult instruments like the cello. An average recognition rate of 80% is reported in [28] on a larger dataset of musical instruments. Approaches to single instrument recognition are usually based on the extraction of relevant audio features, from MFCCs to peaks in the frequency representation, while classification is carried out using well-known techniques such as Gaussian mixture models or K-nearest neighbors, which are trained on a set of labeled audio examples. Recently, the instrument identification problem has been extended to polyphonic music [64] even if only in the case of two or three instruments playing together. The approach is based on the creation of musical instrument timbre templates, which take into account the variation of timbre due to the performing parameters. Results are comparable to the ones obtained with solo instrument.

Even when the techniques for sound source recognition will be effective enough, it can be argued that the recognition of a particular orchestration does not fit common users' information needs. Yet there are some MIR applications that can be based on orchestration. For instance a violin player may be interested in retrieving audio recordings where his instrument is employed, or a music critic may want to retrieve all the recordings of a given band when a particular set of musical instruments have been employed.

4.5 Audio Form: Rhythm

It is commonly assumed that rhythm is the most easily recognizable dimension for casual users. As usual, this assumption is biased by Western classical music, where the evolution of music language involved more melody and harmony than rhythm. Rhythm information can be

very complex for other music languages, for example the ones developed in Africa or in Eastern Europe, where different time signatures are applied to the same musical work and rhythm is a multidimensional feature.[3]

In the case of pop and rock music, rhythm information is based on variants of the same structure of four equally spaced beats, the first and the third stronger than the second and the fourth. What becomes relevant then is the speed by which the beats are played. To this end, a number of approaches on *tempo tracking* or *foot-tapping*[4] have been proposed. Before its application to MIR, tempo tracking has been applied to interactive music systems, using both Midi [3] and audio [37]. The general approach is to highlight a periodic structure in the amplitude envelope of the signal, for instance using a self-similarity matrix [115] or the autocorrelation function [17]. Tempo tracking systems will give a very general description of the musical work content, which nevertheless may be considered relevant for users interested in creating mixes of songs, where tempo coherence is important, organizing playlists for parties or retrieving music that they intend to dance. In many radio companies, the songs are classified according to their tempo, which is metadata manually added by tapping on the computer keyboard while the song is played.

There is a particular Western music genre where rhythm information, in particular the number and the organization of strong and soft beats, becomes particularly relevant also for casual users: *dance music*. Each rhythm is defined by a style label and is associated with particular steps to dance to it. For example, the approach presented in [38] uses general descriptors, such as tempo and periodicity histograms, to classify dance music. The work reported in [110] proposes to identify patterns in the spectrum of the signal that are considered as the signature of a particular dance. These approaches, which are based on the assumptions that particular music styles are described by temporal patterns,

[3] This kind of music is usually defined *polyrhythmic*.

[4] The term refers to the ability of simulating a user that taps his foot while listening to a driving rhythm.

can be extended also to more general retrieval and classification tasks.

4.6 Audio Form: Melody

The automatic transcription of audio signals is a very active research area, which goes beyond the MIR application. In particular, the task of *pitch tracking*, or *F0-estimation*, consists of computing the fundamental frequency of the different sound sources for each time frame and has many applications ranging from speech recognition and synthesis, computer music, and gestural control of sound. Automatic melody tracking combines the F0-estimation with segmentation, to obtain a representation comparable to the musical score.

As in the cases of symbolic melody processing, Section 4.1, and sound source recognition, Section 4.4, the problem is treatable with monophonic recordings obtaining good results in terms of recognition rate. On the other hand, it is particularly difficult to compute a reliable estimation of the fundamental frequencies from polyphonic recordings, especially when musical instruments with a rich spectrum are employed. For a review on the quite vast research area on pitch tracking, both for monophonic and for polyphonic signals, the interested reader can refer to [19]. Yet, it is important to mention the typical limitations of pitch trackers, because they affect the way an automatically extracted melody can effectively represent an audio recording.

The computed fundamental frequency may be an octave higher or lower than the correct one.[5] A single note played with *vibrato* may be recognized as a sequence of notes going up and down around a central one. Two subsequent notes with the same pitch may be compounded in a single note, or the same note with a steep amplitude envelope may be divided in two or more notes. Fast glissandi, used in the singing voice to reach a particular note, are transcribed as a number of short notes. It is clear that any kind of mismatch results in a decrease of the retrieval effectiveness.

[5] This means that if the real fundamental frequency is an A at 440 Hz, the system may recognize an A at either 220 Hz or 880 Hz.

A subtask of melodic feature extraction, which is particularly suitable for MIR, regards the computation of the main melody from an audio recording [36]. The goal is similar to the one presented in Section 4.1.1 on extracting the main melody from a polyphonic score. Besides the assumptions made for the symbolic case, in the case of audio recordings it is also assumed that the main melody has a higher intensity than the accompaniment and, in case of stereo recordings, is balanced between left and right channels. These assumptions are reasonable in the case of pop and rock music and may hold also for traditional music and for a Western classical music *concerto* [107]. The typical approach to this task is then to perform a monophonic pitch tracking of the polyphonic audio, trying to extract the main voice and considering the background accompaniment as noise.

4.7 Audio Form: Harmony

The automatic recognition of chord sequences is still a difficult task. An alternative solution is to compute descriptors of the harmonic content of the signal. These descriptors can be used to characterize directly an audio recording or as a preprocessing step to compute the chord progressions and the tonality of a musical work.

The basic idea of chord representation is to map the different spectral components of a complex audio signal into a single chromatic scale. The representation should be robust to variations on how the notes that form a given chord are combined and to differences in the energies of the first harmonics of the individual notes.

As an example, Pitch Class Profiles are proposed in [31] as descriptors of the harmonic content of a musical sound. Profiles are computed, mapping the energy on the complete spectrum on the 12 elements of the chromatic scale. Recognition of the actual chords is carried out by comparing the output with a number of chord templates. An extension of the approach called Harmonic Pitch Class Profile has been applied in [35] to the task of recognizing the tonality of a song. A similar technique is the chroma-based approach, presented in [5], where the signal is transformed in 12-dimensional distribution vectors depending on the energy of each frequency component.

4.8 Other Dimensions

The music processing for some of the dimensions have not been included in this discussion. In particular, acoustics and structure have not been related either to symbolic or to audio music processing. This is motivated by the fact that, to the knowledge of the author, these dimensions are not commonly used in existing MIR systems. The choice of including them in the discussion in Section 2.2 is due to a number of consideration of their possible relevance in future research.

Acoustics has been defined as the part of the hearing experience that depends on the way a performance has been recorded and post-processed. This is in general independent from the musical work and from the performers themselves; it is strictly related to the choice of audio engineers, to the available audio equipment at the time of the recording, and to the style of particular recording labels—with this latter characteristic eventually related to music genres. After the introduction of the audio CD, and the increasing availability of high fidelity equipment for playback also to non-professional users, the listeners' expectations of audio quality have been modified. Maybe it cannot be stated that users look for a particular acoustics of the musical works, but it is likely that users' information needs will not be completely fulfilled if the audio quality is below their expectations. Therefore, MIR systems should be able to retrive musical documents also depending on the perceived quality, based for example on objective perceptual measurements of noise [119].

At the state of the art, it is difficult to compute and even to represent in a compact form, information about musical structure. Moreover, structure is the dimension that is more difficult to understand for users without a musical education. Nevertheless, it is likely that short- and middle-term features are intrinsically limited in their ability to model the similarity between musical works, both in symbolic and in audio forms. Structure could be used to achieve further improvements in carrying out many music retrieval tasks. An example of the application of structural information is given in [112], where melodic and harmonic information is combined

with the information on the structure to improve a music clustering task. Experimental results show that the use of structural information may improve the clustering of musical documents stored in symbolic form.

4.9 Alignment of Audio and Symbolic Representations

The two main forms of musical documents, audio and symbolic, are strictly related. A score can be considered as the model from which different performances are created and, at the same time, as an approximate transcription of a given performance. Bridging the gap between the two representations is the focus of a research area that is related both to MIR and to computer music research communities. In the case of MIR, audio–symbolic alignment is not a task by itself, but rather a preprocessing step for retreival and recognition tasks.

One of the main applications of automatic audio alignment is usually called *score following*. The basic idea is that a system aligns in real time an ongoing audio performance with a symbolic representation of the score in order to perform an automatic accompaniment. Early score following systems were based on dynamic programming approaches, such as [16]. It is interesting to notice that, already in 1993, an information retrieval approach was applied to a score following task [125]. More recent approaches are based on statistical models [41], and in particular on the use of hidden Markov models [118, 12].

Similar techniques can be applied to MIR tasks, which do not impose real-time constraints. In this case, alignment can be used to recognize audio recordings, using a database of symbolic scores, as proposed in [106], and in [54] using alternative approaches. Alignment can be carried out also for the automatic recognition of audio recordings. For example, a direct match between an unknown recording and a collection of labeled audio recordings is proposed in [95], where a variant of Dynamic Time Warping has been proposed. A similar approach has been presented in [20], where the goal is mainly to compare the expressive timing of different performances of the same musical work.

5

Systems for Music Retrieval

The techniques presented in the previous chapter are aimed at extracting useful information from musical documents, which is the first step towards the development of a system that fulfills real information needs. It is proposed to classify the different approaches according to three main areas, typical of information management: *searching*, *filtering*, and *browsing*. It can be noted that the three examples of casual users' information needs, reported in Section 3.1, can be mapped easily on these three areas, while the two examples of professional users' information needs—and in general the information needs of musicologists, music theorists, and musicians—are more related to a retrieval task.

5.1 Music Search

Searching for a musical work given an approximate description of one or more of its dimensions is the prototype task for a MIR system, and in fact it is simply addressed as music retrieval. In principle, retrieval can be carried out on any dimension. For instance, the user could provide an example of the timbre—or of the sound—that he is looking for, or describe the particular structure of a song. Systems have been

proposed where the retrieval is based on vocal percussion [62] or on harmony [111]. Yet, most of the approaches are based on melody as the main, and often only, content descriptor.

The research work on melodic retrieval can be grouped depending on the methodologies that have been proposed to compute the similarity between the query and the documents. A classification in three categories is proposed: approaches based on the computation of *index terms*, which play a similar role of words in textual documents, approaches based on *sequence matching* techniques, which consider both the query and the documents as sequences of symbols and model the possible differences between them, and *geometric methods*, which can cope with polyphonic scores and may exploit the properties of continuous distance measures (in particular the triangular inequality) to decrease computational complexity.

5.1.1 Melodic retrieval based on index terms

As it is well known in IR, indexing improves the scalability of a retrieval system, because all the relevant information needed at retrieval time is computed off-line and matching is carried out between query and document indexes. Scalability is the main motivation behind systems based on the computation of index terms. This positive aspect is balanced by a more difficult extraction of document content, with non-trivial problems arising from query errors that may cause a complete mismatch between query and document indexes.

Being based on the automatic extraction of content descriptors from the melody, these particular indexing techniques rely on the automatic extraction of lexical units described in Section 4.1.2. Although reported for the symbolic form, the approach can be applied also to the transcriptions of audio documents. The main difference between the approaches lies in the computation of the lexical units. Indexing and retrieval are then usually carried out using well-known techniques developed for textual IR, such as the Vector Space Model using the $tf \cdot idf$ weighting scheme.

An example of research work in this group has been presented in [26], where melodies were indexed through the use of N-grams.

Experimental results on a collection of folk songs were presented, testing the effects of system parameters such as N-gram length, showing good results in terms of retrieval effectiveness, though the approach did not seem to be robust to decreases in query length. The N-gram approach has been extended in [21] in order to retrieve melodies in a polyphonic score, without prior extraction of the single melodies.

An alternative approach to N-grams has been presented in [89], where indexing was carried out by highlighting musically relevant sequences of notes, called musical phrases. Unlike the previous approaches, the length of indexes was not fixed but depended on the musical context. Phrases could undergo a number of different normalization, from the complete information on pitch intervals and duration to the simple melodic profile. Segmentation approaches can be based also on recurrent melodic patterns, as proposed in [113] and further developed in [103]. In this latter case, patterns were computed using either only rhythm, or only pitch, or the combined information, and the final retrieval was carried out using a data fusion approach.

An extensive evaluation of segmentation techniques aimed at extracting index terms has been presented in [105]. Experimental results on a collection of about 2300 musical documents in Midi format, showed that N-grams are still the best choice for index terms—0.98 of average precision—and that recurrent patterns were almost comparable to them—0.96 of average precision. Segmentation approaches based on *a priori* knowledge of music perception or structure showed to be more sensible to local mismatches between the query and the documents, giving an average precision of about 0.85 in both cases.

5.1.2 Melodic retrieval based on sequence matching

The typical application of these approaches is the retrieval of a precise musical work, given an approximate excerpt provided by the user. To this end, a representation of the query is compared with the representations of the documents in the collection each time a new query is submitted to the system. The main positive aspect of these approaches is that they are able to model the possible mismatches between the query and the documents to be retrieved. As it is well known from

string processing domain, possible sources of mismatches are insertion and deletions of musical notes. The modification of a note can be considered either as a third source of mismatch or the combination of a deletion and an insertion.

Approximate string matching techniques have been applied to melodic retrieval. One of the first examples has been described in [33], where the melodies were represented by three symbols—ascending or descending interval and same note—in order to cope with possible mismatches in pitch between the query and the documents. The work presented in [4] is based on the use of pattern discovery techniques, taken from computational biology, to search for the occurrences of a simplified description of the pitch contour of the query inside the collection of documents. Another approach, reported in [49], applies pattern matching techniques to documents and queries in GUIDO format, exploiting the advantages of this notation in structuring information. Approximate string matching has been used also by [46], adapting the technique to the kind of input provided by the user. The work presented in [55] reports a comparison of different approaches based on a variant of Dynamic Time Warping, with a discussion on computational complexity and scalability of four different techniques. Other examples of sequence matching can be found in [131, 50].

Alternatively to approximate string matching, statistical models have been applied to sequence matching. An application of Markov chains have been proposed in [7] to model a set of themes that have been extracted from musical documents, while an extension to hidden Markov models has been presented in [123] as a tool to model possible errors in sung queries. A mixed methodology has been presented in [53], where the distance function used in a Dynamic Time Warping approach has been computed using a probabilistic model.

Sequence matching techniques are very efficient, with a computational cost for a single comparison that is $\mathbf{O}(m + n)$, where m is the length of the query and n is the size of the document. Yet, the application of sequence matching may require that the sequence representing the query is compared to all the documents in the collections. Thus the computational cost of a single retrieval is linear with the size of the collection. This clearly implies a low scalability of direct sequence match-

ing. To overcome the problem, pruning techniques have been proposed in the literature. In particular, the approach described in [109] is based on the creation of a tree structure over the collection of documents depending on the melodic similarity between them: comparisons are carried out only along the path, from the root to a leaf, that gives the best sequence matches.

5.1.3 Melodic retrieval based on geometric methods

The pianoroll representation of a symbolic score, such as the one shown in Figure 2.1, suggests to compute the matching of the query with documents in a geometric framework. This approach can cope with polyphonic music without requiring prior extraction of the main melody, because the complete score is represented as a set of points, or lines, on a plane: the vertical axis usually corresponds to pitch while the horizontal axis corresponds to time. The same representation applies to queries.

The geometric approach, which has been introduced in [14], is based on the application of a number of translations to the query pattern in order to find the best matches with the geometric representation of each document. Incomplete matches can also be found with a geometric approach, as described in [142], where scores were represented as points on a plane. An extension of the representation of documents is presented in [138], where a polyphonic score is represented as a set of lines on a plane, the position along the time axis and the length of the line are computed from time onset and note duration, respectively. A further improvement has been proposed in [82], where note duration is exploited to create a weight model that penalizes mismatches between long notes.

The computational cost of geometric approaches is $\mathbf{O}(mn \log n)$, where m is the size of the query and n is the size of the score. The increase in computational complexity is compensated by the fact that these approaches can cope with polyphonic scores. As for sequence matching approaches, a retrieval task may require a number of comparisons that is linear with the collection size, if a pruning or indexing technique is not applied.

To this end, an alternative approach to compute the similarity between a bidimensional representation of queries and documents was presented in [133]. The polyphonic scores are represented as weighted points on a plane, where the positions correspond to pitch and onset time of each note, while the weight is computed from note durations. The melodic similarity is computed through two alternative transportation distance, the Earth Mover's Distance and the Proportional Transportation Distance [34]. The Proportional Transportation Distance is a pseudo-metric, for which the triangle inequality holds. This property has been exploited to improve retrieval efficiency, because the query is compared only to a reduced set of documents—called *vantage objects*—exploiting the triangular inequality to rule out all the documents that have a distance from the query higher than a given threshold.

5.2 Music Filtering

Given the number of customers that regularly buy music at on-line stores, it is obvious that a number of automatic recommender systems have been developed also for the music domain. The goal is to provide the user with a substitute of the musical expert of any good CD store; the system can suggest what to purchase depending on users' preferences filtering out all non-relevant items. Usually, recommender systems are based on external information, such as user profiles, purchases, and product ratings, applying techniques known as *collaborative filtering*, and are not necessarily content-based. As mentioned in Section 3.1, being based on customers' behavior, collaborative filtering approaches have the new item and the cold start drawbacks. For this reason, a number of content-based music recommender systems have been proposed.

Content-based approaches are normally based on the idea that, having a root set of items rated by the user, the system recommends a number of new items depending on the similarity—or dissimilarity for negative scores—with the elements in the root set. The choice of the features depends on the dimensions that are considered relevant for the user. For instance, the work reported in [77] uses timbre information, that is low-level features, to group items depending on their similarity.

The final recommendation is based on a classic collaborative filtering approach, where the rating of each item depends also on the content-based grouping of items. The system presented in [80] uses MFCCs as content descriptors, and explores the use of different approaches to compute the distance between the root set and the songs to be recommended. The examples were based on qualitative analyses of the distance between songs in the same album, which is a characteristic peculiar to music that could be explored in more detail.

Regarding the choice of the features, the recommendation and browsing system called *MusicSurfer* is based on high-level features such as rhythm and harmony, which seems as suitable as low-level features for a music recommendation task. Collaborative filtering and content-based recommendation have been compared on a large dataset in [126], showing that collaborative filtering approaches still outperform content-based approaches.

5.2.1 Automatic playlist generation

Once a measure of similarity is given between items, a content-based recommender system could be applied in principle to any media, because the final goal is usually the same—to suggest the user one or more items to purchase. As described in Section 3.1, there is a particular task that can be applied to music only: the automatic *playlist* generation.

It is assumed that songs in a playlist should share some music dimensions, both globally and locally. There are two main differences between automatic playlist generation and the usual recommendation of a set of items: it is likely that users want to listen to songs that they already know and the ordering of the songs is relevant. The approach presented in [79] sees the generation of a playlist as a path across a graph, where nodes are the songs and links are drawn between each couple of similar songs. The focus is hence on local consistence between two subsequent songs, rather than on global coherence of the songs in the playlist.

Automatic playlist generation can also be carried out on an external collection, for instance in the case of users of digital radios. In this case

the approaches for categorization and for retrieval can be combined with collaborative filtering, in order to find the songs that are of interest for a user and then organize them in a playlist.

5.3 Music Browsing, Classification, and Visualization

The direct search proposed in the previous section is only one of the possible approaches to for accessing a music collection. Other metaphors are currently explored for the accessing of text and multimedia documents, including the possibility to browse a collection, to classify documents in a number of categories, or to use visual cues for more effective access to large collections of documents.

Clearly, the same considerations can also be made for music collections. For instance, a user may retrieve, through a query-by-humming system, a number of musical documents that do not completely satisfy his information need. Yet, the user may not be able to refine his query as in common presentation–evaluation loops, because he already gave the best performance he could with a query-by-humming interface: listening to other documents may not necessarily improve his singing ability. Browsing can be exploited to overcome this situation. As another example, a user may be interested only in a given music genre and can be bothered by a single rank list of retrieved documents where traditional music is mixed with HipHop and Hard Rock: classification based on high-level descriptors as the music content can help reorganize the results in a similar fashion as Web search engines based on document clustering. Finally, the users may wish to use visual cues to access music collections. To this end, it has also to be considered that musical documents do not permit an easy representation of their content, and searching for music files, even inside personal collections, may not be an easy task.

5.3.1 Browsing music collections

It is well known that browsing a collection of documents is a viable alternative to direct search. Moreover, as it happens for many similar approaches for other media, navigation may integrate normal content-based search. To match a query, the system produces a list of documents

that are within a given distance of the query, and provides links to retrieve other similar documents.

Music browsing and navigation are based on the concept of similarity between musical documents, which can be applied both to symbolic and to audio forms. All the dimensions that have been presented, and any combination, can be used to create new links between the documents. In principle, similarity is user-dependent, at least because the individual contribution of each single dimension to the similarity score depends on the importance that the user gives to it, and it may vary with time and user expertise. Yet, most of the approaches to music browsing are based on the static computation of similarity, based on a predefined number of dimensions. Browsing can partially overcome the problem of describing the content of a musical document, in particular for casual users. To this end, defining a musical document through a list of links to similar ones may be a useful tool for users in selecting—and eventually purchasing—new musical works.

The first paper on content-based navigation inside a collection of musical documents have been presented in [8]. In that case, similarity was computed using melody as the only relevant dimension, in particular using the pitch contour. An interesting aspect is that an open hypermedia model is adopted, which enables the user to find available links from an arbitrary fragment of music. Another approach to content-based navigation of a music collection is presented in [90]. A collection of musical documents and their lexical units are enriched by a hypertextual structure, called *hypermusic*, that allows the user to navigate inside the document set. An important characteristic is that links are automatically built between documents, between documents and lexical units, and between lexical units. The navigation can be carried out across documents but also across relevant content descriptors, where similarity is computed depending on the co-occurrence of lexical units inside the documents.

5.3.2 Audio classification

The term audio classification has been traditionally used to describe a particular task in the fields of speech and video processing, where the

main goal is to identify and label the audio in three different classes: speech, music, and environmental sound. This first coarse classification can be used to aid video segmentation or decide where to apply automatic speech recognition. The refinement of the classification with a second step, where music signals are labeled with a number of predefined classes, has been presented in [144], which is also worth mentioning because it is one of the first papers that present hidden Markov models as a tool for MIR.

An early work on audio classification, presented in [143], was aimed at retrieving simple music signals using a set of semantic labels, in particular focusing on the musical instruments that are part of the orchestration. The approach is based on the combination of segmentation techniques with automatic separation of different sources and the parameter extraction. The classification based on the particular orchestration is still an open problem with complex polyphonic performances, as described in Section 4.4.

An important issue in audio classification, introduced in [30], is the amount of audio data needed to achieve good classification rates. This problem has many aspects. First, the amount of data needed is strictly related to the computational complexity of the algorithms, which usually are at least linear with the number of audio samples. Second, perceptual studies showed that even untrained listeners are quite good at classifying audio data with very short excerpts (less than 1 sec). Finally, in a query-by-example paradigm, where the examples have to be digitally recorded by users, it is likely that users will not be able to record a significantly large part of audio.

A particular aspect of audio classification is *genre classification*. The problem is to correctly label an unknown recording of a song with a music genre. Labels can be hierarchically organized in genres and subgenres, as shown in Table 5.1. Labeling can be used to enrich the musical document with high-level metadata or to organize a music collection. In this latter case, it is assumed that the organization in genres and subgenres is particularly suitable for a user, because it is followed by almost all the CD sellers, and is one of the preferred access methods for on-line stores. Genre classification is, as other aspects of MIR, still biased by Western music, and thus genres are the ones typically found

Table 5.1 Music classification hierarchy, as proposed in [136].

Genre	Subgenre
Classical	Choir
	Orchestra
	Piano
	String quartet
Country	
Disco	
HipHop	
Jazz	BigBand
	Cool
	Fusion
	Piano
	Quartet
	Swing
Rock	
Blues	
Reggae	
Pop	
Metal	

in Western music stores. Some attempts have been made to extend the approach also to other cultures, for instance in [104] genre classification has been carried for traditional Indian musical forms together with Western genres.

It can be argued that a simple classification based on genres may not be particularly useful for a user, because a coarse categorization will result in hundreds of thousands of musical documents in the same category, and users may not agree on how the classification is carried out with a fine grained categorization. Yet, this part of MIR research is pretty active, because users still base their choices on music genres, and information about genre preferences can be exploited to refine users' profiles, as proposed in [48].

One of the first papers introducing the problem of music classification is [136]. The proposed classification hierarchy is shown in Table 5.1,

from which it can be seen that there is a bias toward classical music and jazz, while some genres—ambient, electronic, and ethnic—are not reported. This is a typical problem of music classification, because the relevance of the different categories is extremely subjective, as well as the categories themselves. These problems are faced also by human classifiers that try to accomplish the same task, and in fact in [136] it is reported that college students achieved no more than about 70% of classification accuracy when listening to three seconds of audio (listening to longer excerpt did not improve the performances). The automatic classification is based on three different feature sets, related to rhythmic, pitch, and timbre features. As also highlighted in subsequent works, rhythm seems to play an important role for the classification.

The features used as content descriptors are normally the ones related to timbre, and described in Section 4.3. This choice depends on the fact that approaches try to classify short excerpts of an audio recording, where middle-term features like melody and harmony are not captured. Common music processing approaches compute the MFCCs, while the use of the wavelet transform is exploited in [40] and in [78]. Systems on genre classification are normally trained with a set of labeled audio excerpts, and classification is carried out using different techniques and models from the classification literature. In particular, k-Nearest Neighbor and Gaussian Mixtures Models are classically used for classifying genres, but Support Vector Machines and Linear Discriminant Analysis have also been successfully applied to this task.

5.3.3 Visualization of music collections

The approaches to the visualization of music collections can be divided into two categories: the ones aimed at a graphical representation of the content of single musical documents and the ones aimed at representing a complete collection. The former is motivated by the difficulties, for a casual user, to retrieve musical documents he has already purchased and downloaded and that are stored in one of his devices: the possibility to have a *musical snapshot* of the content of a song—possibly as a preview given by the operative system like it is common practice for images, presentations, and textual documents—without having to

listen to the songs themselves, will ease the user to browse his own collection. The latter is motivated by the fact that a spatial organization of the music collection will help the users finding particular songs they are interested in, because they can remember their position in the visual representation and they can be aided by the presence of similar songs nearby the searched one.

There is a variety of approaches to music visualization, including the symbolic score, the pianoroll view, the plot of the waveform, and the spectrogram. Any representation has positive aspects and drawbacks, depending on the dimensions carried by the music form it is related to, and on the ability to capture relevant features. Representations can be oriented toward a global representation or local characteristics. The interested reader may refer to [58] for a complete overview on techniques for music visualization.

Visualization of a collection of musical documents is usually based on the concept of similarity. The problem of a graphical representation, normally based on bidimensional plots, is typical of many areas of data analysis. Techniques such as Multidimensional Scaling and Principal Component Analysis are well known for representing a complex and multidimensional set of data when a distance measure—such as the musical similarity—can be computed between the elements or when the elements are mapped to points in a high-dimensional space. The application of bidimensional visualization techniques to music collections has to be carried out considering that the visualization will be given to non-expert users, rather than to data analysts, who need a simple and appealing representation of the data.

One example of system for graphical representation of audio collection is *Marsyas3D*, which includes a variety of alternative 2D and 3D representations of elements in the collection. In particular, Principal Component Analysis is used to reduce the parameter space that describe the timbre in order to obtain either a bidimensional or tridimensional representation. Another example is the *Sonic Browser*, which is an application for browsing audio collections [9] that provides the user with multiple views, including a bidimensional scatterplot of audio objects, where the coordinates of each point depend on attributes of the dataset, and a graphical tree representation, where the tree is depicted

with the root at the center and the leaves over a circle. The *sonic radar*, presented in [81], is based on the idea that only a few objects, called prototype songs, can be presented to the user. Each prototype song is obtained through clustering the collection with k-means algorithm and extracting the song that is closer to the cluster center. Prototype songs are plotted on a circle around a standpoint.

A number of approaches to music visualization are based on Self-Organizing Maps (SOMs) [129]. In particular, the visual metaphor of *Islands of Music* is presented in [108], where musical documents are mapped on a plane and enriched by a third dimension in the form of a geographical map. SOMs give a different visualization of the collection depending on the choice of the audio parameters used for their training. A tool to align the SOMs is proposed to reduce the complexity of alternative representation for non-expert users. Another approach using Emergent SOMs for the bidimensional representation of a music collection is presented in [93], where genre classification was combined with visualization because genre labels are added to the elements in the collection. In this case, instead of allowing the user to choose the combination of dimensions that he prefers, the system is trained with more than 400 different low-level features, which were also aggregated to obtain high-level features, and the selection was made *a posteriori* depending on the ability of each feature to separate a group of similar songs from the others.

6

Evaluation

The evaluation of the effectiveness of a MIR system and the comparison of different approaches are fundamental steps toward continuous improvements of the system's effectiveness. Given the variety of the approaches to MIR, evaluation has to take into account many aspects: the effectiveness of a retrieval engine, which can be computed with commonly agreed measures such as the classic *average precision*, the ability to reliably extract content descriptors, which can be measured using a number of manually labeled examples, and the effectiveness in classifying and clustering musical documents.

Until year 2004, the research results in MIR were evaluated with self-made test collections, each research group using a different set of documents and queries. Collections could be made either of symbolic documents—with Midi the most popular format, thanks to its wide availability—or of audio documents—with MP3 as the most popular format. Differences regarded also the queries, which were either recorded from real users or automatically generated, the size of the collection, which ranged from hundreds to several thousands of documents, and the music genre. Also the measures used to evaluate the systems were different, depending on individual choices.

There is still some research work on MIR that has been evaluated only qualitatively, for instance, all the approaches to the visualization of music collection presented in Section 5.3.3 are difficult to evaluate with classic techniques.

6.1 The Audio Description Contest

The first step toward a common evaluation framework has been carried out by the Music Technology Group of the Audiovisual Institute, Universitat Pompeu Fabra (Barcelona), which in 2004 hosted the International Conference of Music Information Retrieval (ISMIR). Being the research group main expertise on audio analysis and synthesis, the contest has been biased toward the audio form. In fact, the evaluation framework has been called *Audio Description Contest*. It has to be noted that audio recordings are more likely to be of interest for a larger audience than symbolic scores, because they can be accessed also by users without a musical training.

The Audio Description Contest has been divided into six independent tasks. The first three were on classification and identification of artists and genre.

- *Genre Classification*: label an unknown song with one out of six possible music genres.
- *Artist Identification*: identify one artist given three songs of his repertoire, after training the system with seven more songs.
- *Artist Similarity*: mimic the behavior of experts in suggesting an artist similar to a given one, with 53 artists in the training set, and 52 artists in the test set.

For the first task, the participants could have the raw audio data to run their experiments, because they were provided by a Web service that allowed the use of audio content for research. Due to copyright issues, the organizers did not distribute the original recordings for the second and third tasks, but distributed a set of low-level features that they computed from the recording themselves, as proposed in [6].

Three more tasks were devoted to evaluate music processing techniques, with a special focus on rhythm and melody dimensions, described in Sections 4.5 and 4.6, respectively. The tasks were:

- *Rhythm Classification*: label audio excerpts with one out of eight rhythm classes—a training set of 488 instances were available from a third party Web site, and 210 more instances were used by the organizers to test the system.
- *Tempo Induction*: main beat detection from polyphonic audio—no training set was provided, while 3199 instances were used as test set.
- *Melody Extraction*: main melody detection, singing voice or solo instrument, from polyphonic audio—a training set of 10 manually annotated excerpts and a test set of 20 audio excerpts (the training set with 10 additional excerpts) were provided.

An interesting approach, which has been maintained also in subsequent MIR evaluation campaigns, is that the participants had to submit the algorithms they had developed to carry out the proposed tasks rather than submitting directly the results. It was the duty of the organizers to gather, compile, run the algorithms, and compute the results. Different platforms and programming languages were allowed, in order to open the contest to as many participants as possible.

The choice of collecting the algorithms instead of the results has two main drawbacks. From the point of view of the organizers, it is surely demanding to compile and to run software that have been developed for research purposes, and which is likely neither well documented nor tested, and may be the results of last minute choices. From the point of view of the participants, it required to assemble different pieces of software that have been used at different stages of the experimentation, to provide a self-containing set of routines that can be submitted to the contest.

It is interesting to note that the proposed evaluation approach required the participants to completely trust the organizers, who carried out all of the experiments. This was probably possible because of the need for a common evaluation framework, which is hindered by the

problems in sharing musical documents in any form and format: the choice of not distributing the musical documents is a viable solution to avoid any copyright infringement from the beginning.

6.2 The Music Information Retrieval Evaluation eXchange

A large-scale evaluation of MIR systems has been made possible thanks to the many efforts by Dr. J.S. Downie, who organized several workshops on MIR evaluation, collecting ideas and needs of researchers in MIR, and finally starting the International Music Information Retrieval System Evaluation Laboratory (IMIRSEL) project [56]. The main aim of the project is the creation of secure, yet accessible, music collections for MIR evaluation. Researchers may access the documents through a uniform mechanism that allows them to test music processing, accessing, and retrieval techniques without requiring the transfer of the collection at the users' side. A graphical user interface allows researchers and developers the prototyping of new approaches and implementation of new functionalities. An overview of the evaluation framework is given in [23].

The approach to the definition of the tasks is an example of working democracy. Each potential participant is able to propose a particular task, and can involve as many participants as he is able to. It is up to the proposer to define the final goal of the task, to provide the datasets for training and testing the systems, and to define the measures by which the results will be ranked. This is usually carried out with the active collaboration of all the potential participants, who can discuss all the details of the task, contribute musical documents to the datasets, and suggest different approaches for the evaluation. Of course it is up to the organizers to choose which is the ultimate setup of the task they proposed.

6.2.1 MIREX 2005

The first evaluation campaign based on IMIRSEL was organized in the year 2005, and results were presented at the ISMIR of the same year. The campaign has been called *Music Information Retrieval Evaluation eXchange* (MIREX). There have been nine different tasks during

the MIREX 2005 campaign, six on the audio form and three on the symbolic form.

(1) *Audio Genre Classification*: assign a label, from a set of 10 pre-defined genres, to an audio recording of polyphonic music.
(2) *Audio Artist Identification*: recognize the singer or the group that performed a polyphonic audio recording.
(3) *Audio Drum Detection*: detect the onsets of drum sounds in polyphonic pop songs, some synthesized and some recorded, which have been manually annotated for comparison.
(4) *Audio Onset Detection*: detect the onsets of any musical instrument in different kinds of recordings, from polyphonic music to solo drums.
(5) *Audio Tempo Extraction*: compute the perceptual tempo of polyphonic audio recordings.
(6) *Audio Melody Extraction*: extract the main melody, for instance the singing voice in a pop song or the lead instrument in a jazz ballad, from polyphonic audio.
(7) *Symbolic and Audio Key Finding*: extract the main key signature of a musical work, given either a representation of the symbolic score or the recording of a synthetic performance (the dataset was the same for the two tasks).
(8) *Symbolic Genre Classification*: assign a label, from a set of 38 pre-defined genres, to a symbolic representation—e.g., Midi—of a musical work.
(9) *Symbolic Melodic Similarity*: given an incipit, retrieve the most similar documents from a collection of monophonic incipits.

Audio genre classification and artist identification were carried out on similar datasets that were based on two independent collections. The participants had to make independent runs on the two collections, and final results were based on the average performances. In the case of artist identification, the best result obtained was a recognition rate of 72%. The use of alternative collections was done also for other tasks. In particular, onset detection was divided into nine subtasks, depending

on the type of audio—e.g., polyphonic, instruments with sharp attacks, and voice—and the final results were a weighted average of the individual subtasks. The best results, computed using the F-measure, varied dramatically depending on the audio source: from as high as $F = 0.99$ for bars and bells sounds to as low as $F = 0.45$ for the singing voice. These results highlight that the singing voice should be analyzed with ad hoc techniques, in particular for the transcription of users' queries.

Similarly for drum detection, three different collections have been used, each collection provided by one of the participants. In this case the results did not vary sensibly among collections. The average F-measure for the system that gave the highest performance was $F = 0.67$. The only effect of the collection is that in two cases out of three, the participant group that provided the collection was also the one that scored the best.

Both tempo and melody extraction was carried out on a (different) single collection, with a percentage of finding at least one correct tempo of 95% and with an overall accuracy for melody of 71%. It would be interesting to see whether the approximations introduced by approximate matching or indexing already put an upper bound to the performances of a MIR system, which will not improve even if melodic transcription would reach higher accuracy. In general, it would be important that the music processing techniques are evaluated on how well they accomplish a MIR task, rather than purely compared with the behavior of human experts.

Given that the same collection of musical works has been used, it is possible to directly compare results on symbolic and audio key finding. The percentage of correct recognitions were very similar for both forms—maximum recognition rate of 90% for audio and 91% for symbolic form—showing that techniques for chord recognition from audio have reached considerably high precision. Even if with some caution, because the datasets were different—in particular there were 10 genres for audio and 38 for symbolic classification tasks—it is also possible to compare the symbolic and audio classification tasks. In this case, results show that the use of audio cues may give better results than the use of symbolic information in the order of five points in percentage—

classification accuracy reached 84% for audio and 77% for symbolic documents.

The task more similar to an IR task was the one on melodic similarity. The evaluation was carried out using the Cranfield model for information retrieval, with an experimental collection of 582 documents—monophonic incipits—and two sets of queries, to train and test the system respectively. Queries were in the same form of documents, while the relevance judgments have been collected by the proposers of the task [132]. The retrieval effectiveness has been compared using classic IR measures—for instance a non-interpolated average precision of 0.51 has been obtained—together with an ad hoc measure, called *Average Dynamic Recall* and suggested by the proposers [134]; the novel measure took into account the fact that relevance assessments were not binary. The relative scoring of the different systems was not affected by the kind of measure.

Being promoted and organized by different persons, and requiring different kinds of manually annotated labeled data, the nine tasks of MIREX 2005 were carried out with different collections, with very different sizes, sometimes divided into training and test collections. Some information about the type of task, the size of the collections, and the number of participants for each task is reported in Table 6.1. Audio-related tasks were definitely more popular than symbolic ones,

Table 6.1 Main characteristics of the tasks carried out at MIREX 2005 campaign.

Task	Form	Trai—Test	Partic.
Genre classification	A	1005 —510	15
	S	*N/A*—950	4
Artist identification	A	1158 —642	8
Drum detection	A	100 — 20	8
Onset detection	A	*N/A*—85	7
Tempo extraction	A	*N/A*—140	9
Melody extraction	A	*N/A*—25	9
Key finding	A	*N/A*—1252	6
	S	*N/A*—1252	5
Melodic similarity	S	11/582—11/582	6

both on the average number of participants and, of course, in the number of tasks.

The duty of the organizers of MIREX 2005 was also to collect and run the different submissions, which were in the form of algorithms either using the graphical interface provided by IMIRSEL or in different programming languages. This approach, which proved to be successful to overcome copyright issues, did not create any problem even if participants had to trust the organizers and could not test the effectiveness (and sometimes the correctness) of their algorithms on the final data. It has to be noted that the organizers of MIREX 2005 were able, in some cases, to even correct the code and obtain final runs.

The interested reader can refer to [24] and to the MIREX wiki at [56] for a complete description of each task, the collection that have been used for each task, and the results of each participant.

6.2.2 MIREX 2006

The MIREX campaign will also be organized for year 2006 with a similar approach. As of July 2006, nine tasks have been proposed by the participants, namely:

(1) *Audio Beat Tracking*: find the location of each beat in an audio file.
(2) *Audio Melody Extraction*: similar to the same task at MIREX 2005.
(3) *Audio Music Similarity and Retrieval*: compute the distance matrix of a collection of audio files.
(4) *Audio Cover Song*: find alternative performances of the same musical work.
(5) *Audio Onset Detection*: continuation of the same task at MIREX 2005.
(6) *Audio Tempo Extraction*: similar to the same task at MIREX 2005.
(7) *Query by Singing/Humming*: retrieve songs in symbolic format given a set of real users' queries.
(8) *Score Following*: align an audio performance to its score in symbolic format.

(9) *Symbolic Melodic Similarity*: retrieve a rank list of symbolic scores given a melodic query; the task differs from the one at MIREX 2005, which required the computation of similarity between documents in the collection.

A trend can be seen from the comparison between MIREX 2005 and 2006. First of all, the percentage of tasks involving the audio format is increased, because there is only one task based only on the symbolic form. Moreover, all the tasks on the extraction of high-level metadata—genre classification, artist identification, and key finding—are not part of the final MIREX 2006 set of tasks, even though they were part of the initial proposal.

Another relevant difference with MIREX 2005 is that some of the tasks can be grouped together, because they have similar goals, and may be based on the same test collections. For instance, *Audio Beat Tracking* and *Audio Tempo Extraction*, both focus on rhythm as the most relevant music dimension, while *Audio Music Similarity and Retrieval* and *Audio Cover Song*, though different tasks, are based on the same collection. Also *Symbolic Melodic Similarity* and *Query by Singing/Humming*, which are two typical IR tasks, are very similar, with the former focusing on the effects of the collection and the latter on the effects of queries.

The interested reader can find further information on MIREX 2006, and in particular the evaluation results when they will be available, at the official Web site of the contest [92].

7

Conclusions

The term MIR encompasses a number of different research and development activities that have the common denominator of being related to music access. Analyzing the research trends of the last two to three years, maybe the first conclusion that can be drawn is that the term *music information retrieval* is somehow misleading because it does not fit completely with the large variety of research activities that are usually labeled with the MIR acronym. On the one hand, it has been already mentioned that users' information needs and consequent approaches are more aimed at music retrieval rather than music *information* retrieval. On the other hand, *retrieval* is meant in a broader sense, which encompasses tasks such as filtering, classification, and visualization that become increasingly useful for the final users. Anyway, MIR is a nice acronym[1] and the only problem is that it is also used to refer to *Multimedia Information Retrieval*.

One big caveat in music retrieval is that, even when the perfect query-by-humming—or query-by-example—system will be developed,

[1] As almost everybody knows, in particular after the story about the space laboratory with this name, MIR means *peace* in Russian.

it is not granted that it will ever become a real killer application. Newspapers of different countries report, from time to time, that a *new music search engine*, usually content-based, has been released. So far, the existence of these new systems does not seem to have changed the user's behavior when looking for music, because interaction is still based on textual descriptions. It could happen that a query-by-humming system would be used for a while, because of the curiosity that it may stir up, but it has still to be proven that users will switch to this modality of interaction.

The growing interest toward different accessing metaphors may give a new thrust to MIR research. For instance, commercial applications such as content-based music recommender systems may become increasingly important components of e-commerce applications. One of their positive characteristics is that they do not need any effort from the user, who is simply presented with potentially relevant items. One thing that seems clear is that a large majority of users are interested in enjoying music in digital format. New metaphors for music retrieval, for the organization of music collections, and also for sharing the enjoyment with friends, need to be proposed.

MIR can take advantage from, and at the same time contribute to, other emerging research trends. For instance, the research area on disappearing computing may better exploit the audio channel and the existence of techniques for music accessing, browsing and clustering, for new metaphors of human–computer interaction. The same applies to mobile and handheld devices, which have limited displaying and gestural interaction capabilities, but do not have such limitations for the audio channel.

A different kind of consideration has to be made about peer-to-peer networks, because of the ease by which users can illegally share musical documents. On the one hand, peer-to-peer networks are promoting the creation of huge personal music collections, with new users' needs in their management that need to be fulfilled by MIR systems, and on the other hand the continuous infringement of intellectual property rights is bounding the possibility of developing real large-scale systems also for research centers.

There are some other aspects related to MIR that have not been taken into account in this overview. Among these, perhaps the most strictly related to issues of music retrieval is the research area named *audio fingerprinting*. The task is to recognize all the copies of a given audio document also in presence of major distortions, such as additional noise, compression, resampling, or time stretching. Audio fingerprint is more related to automatic music recognition rather than music retrieval, and has significant applications on copyright management. An interesting overview of audio fingerprinting motivation and techniques, presented in a unified framework, can be found in [11]. Audio fingerprinting also found commercial applications, like the one provided by [122]. Another aspect related to MIR, which is out of focus of this overview but is worth at least to be mentioned, is *audio watermarking* aimed at inserting inaudible yet trackable information about the owner and the user of a musical document [42]. Audio watermarks are usually created to protect copyright because the owner of a digital recording can be recognized and also because it is possible to track the final users who, after buying digital recordings, are responsible for their diffusion on file-sharing networks.

Despite its existence as a research area for at least five or six years, MIR is still not well known in the scientific community. It may still happen that, after the presentation of a MIR technique, someone from the audience approaches the speaker asking if it is really possible to retrieve music by singing a melody, while nobody is amazed by a system that can retrieve images. This situation may seem strange because, apart from the successful ISMIR [59], which gathers each year an increasing number of participants, MIR results are more and more published and presented in journals, conferences, and workshops on IR, multimedia, speech and audio processing. The weak perception of MIR as a research discipline can be explained by the peculiarities of the musical language, which make it radically different from other media in particular because of the kind of content that is conveyed and how this content may be related to users' information needs. A side effect of this diversity is that, even though there are a number of national and international projects

on different aspects of MIR, few projects involve, at the same level, music and other, different media.

To this end, MIR systems and techniques need to be integrated with other multimedia approaches, promoting the creation of new approaches to information access in which music could be the added value that makes them particularly interesting for users.

Acknowledgments

The author thanks his colleagues of the *Information Management Systems Group* at the Department of Information Engineering of the University of Padova, and in particular to Prof. Maristella Agosti who leads the research group, for the many ideas and discussions on relevant topics related to information retrieval, digital libraries, and multimedia. A special thank to Prof. Massimo Melucci, who was one of the promoters of the research activities on MIR at the University of Padova and who shared his experience on information retrieval and hypertexts.

Many thanks to the anonymous reviewers, for their valuable suggestions to improve the content and the structure of the paper and for contributing to additional references to relevant research work.

References

[1] M. Agosti, F. Bombi, M. Melucci, and G.A. Mian. Towards a digital library for the venetian music of the eighteenth century. In J. Anderson, M. Deegan, S. Ross, and S. Harold, editors, *DRH 98: Selected Papers from Digital Resources for the Humanities*, pages 1–16. Office for Humanities Communication, 2000.

[2] E. Allamanche, J. Herre, O. Hellmuth, B. Fröba, T. Kastner, and M. Cremer. Content-based identification of audio material using MPEG-7 low level description. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 73–82, 2001.

[3] P.E. Allen and R.B. Dannenberg. Tracking musical beats in real time. In *Proceedings of the International Computer Music Conference*, pages 140–143, 1990.

[4] D. Bainbridge, C.G. Nevill-Manning, I.H. Witten, L.A. Smith, and R.J. McNab. Towards a digital library of popular music. In *Proceedings of the ACM Conference on Digital Libraries*, pages 161–169, 1999.

[5] M.A. Bartsch and G.H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104, 2005.

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

[7] W.P. Birmingham, R.B. Dannenberg, G.H. Wakefield, M. Bartsch, D. Bykowski, D. Mazzoni, C. Meek, M. Mellody, and W. Rand. MUSART: music retrieval via aural queries. In *Proceedings of the International Conference on Music Information Retrieval*, pages 73–82, 2001.

[8] S. Blackburn and D. DeRoure. A tool for content based navigation of music. In *Proceedings of the ACM International Conference on Multimedia*, pages 361–368, 1998.

[9] E. Brazil and M. Fernström. Audio information browsing with the sonic browser. In *Coordinated and Multiple Views in Exploratory Visualization*, pages 26–31, 2003.

[10] E. Cambouropoulos. Musical rhythm: a formal model for determining local boundaries. In E. Leman, editor, *Music, Gestalt and Computing*, pages 277–293. Springer-Verlag, Berlin, DE, 1997.

[11] P. Cano, E. Batlle, T. Kalker, and J. Haitsma. A review of audio fingerprinting. *Journal of VLSI Signal Processing*, 41:271–284, 2005.

[12] P. Cano, A. Loscos, and J. Bonada. Score-performance matching using hmms. In *Proceedings of the International Computer Music Conference*, pages 441–444, 1999.

[13] Cantate. Computer access to notation and text in music libraries, July 2006. http://projects.fnb.nl/cantate/.

[14] M. Clausen, R. Engelbrecht, D. Meyer, and J. Schmitz. PROMS: a web-based tool for searching in polyphonic music. In *Proceedings of the International Symposium of Music Information Retrieval*, 2000.

[15] Coda Music. Enigma transportable file specification. Technical Report, version 98c.0, July 2006. http://www.xs4all.nl/ hanwen/lily-devel/etfspec.pdf.

[16] R.B. Dannenberg and H. Mukaino. New techniques for enhanced quality of computer accompaniment. In *Proceedings of the International Computer Music Conference*, pages 243–249, 1988.

[17] M.E.P. Davies and M.D. Plumbley. Casual tempo tracking of audio. In *Proceedings of the International Conference on Music Information Retrieval*, pages 164–169, 2004.

[18] S.B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustic, Speech, and Signal Processing*, 28(4):357–366, 1980.

[19] A. de Cheveigné and A. Baskind. F0 estimation. In *Proceedings of Eurospeech*, pages 833–836, 2003.

[20] S. Dixon and G. Widmer. Match: a music alignment tool chest. In *Proceedings of the International Conference of Music Information Retrieval*, pages 492–497, 2005.

[21] S. Doraisamy and S. Rüger. A polyphonic music retrieval system using N-grams. In *Proceedings of the International Conference on Music Information Retrieval*, pages 204–209, 2004.

[22] W.J. Dowling. Scale and contour: two components of a theory of memory for melodies. *Psychological Review*, 85(4):341–354, 1978.

[23] J.S. Downie, J. Futrelle, and D. Tcheng. The International Music Information Retrieval Systems Evaluation Laboratory: governance, access and security. In *Proceedings of the International Conference on Music Information Retrieval*, pages 9–14, 2004.

[24] J.S. Downie, K. West, A. Ehmann, and E. Vincent. The 2005 music information retrieval evaluation exchange (mirex 2005): preliminary overview. In

Proceedings of the International Conference on Music Information Retrieval, pages 320–323, 2005.

[25] J.S. Downie. Music information retrieval. *Annual Review of Information Science and Technology*, 37:295–340, 2003.

[26] S. Downie and M. Nelson. Evaluation of a simple and effective music information retrieval method. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 73–80, 2000.

[27] J. Dunn and C. Mayer. VARIATIONS: a digital music library system at indiana university. In *Proceedings of ACM Conference on Digital Libraries*, pages 12–19, 1999.

[28] S. Essid, G. Richard, and B. David. Musical instrument recognition based on class pairwise feature selection. In *Proceedings of the International Conference on Music Information Retrieval*, pages 560–567, 2003.

[29] E. Ferrari and G. Haus. The musical archive information system at Teatro alla Scala. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, volume II, pages 817–821, 1999.

[30] J.T. Foote. A similarity measure for automatic audio classification. In *Proceedings of AAAI Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*, 1997.

[31] T. Fujishima. Realtime chord recognition of musical sound: a system using common lisp Music. In *Proceedings of the International Computer Music Conference*, pages 464–467, 1999.

[32] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, and D. Dobkin. A search engine for 3D models. *ACM Transactions on Graphics*, 22(1):83–105, 2003.

[33] A. Ghias, J. Logan, D. Chamberlin, and B.C. Smith. Query by humming: musical information retrieval in an audio database. In *Proceedings of the ACM Conference on Digital Libraries*, pages 231–236, 1995.

[34] P. Giannopoulos and R.C. Veltkamp. A pseudo-metric for weighted point sets. In *Proceedings of the European Conference on Computer Vision*, pages 715–730, 2002.

[35] E. Gómez and P. Herrera. Estimating the tonality of polyphonic audio files: cognitive versus machine learning modelling strategies. In *Proceedings of the International Conference on Music Information Retrieval*, pages 92–95, 2004.

[36] E. Gómez, A. Klapuri, and B. Meudic. Melody description and extraction in the context of music content processing. *Journal of New Music Research*, 32(1):23–40, 2003.

[37] M. Goto and Y. Muraoka. An audio-based real-time beat tracking system and its applications. In *Proceedings of the International Computer Music Conference*, pages 17–20, 1998.

[38] F. Gouyon and S. Dixon. Dance music classification: a tempo-based approach. In *Proceedings of the International Conference on Music Information Retrieval*, pages 501–504, 2004.

[39] J.M. Grey and J.A. Moorer. Perceptual evaluations of synthesized musical instruments tones. *Journal of Acoustic Society of America*, 62(2):454–462, 1977.

[40] M. Grimaldi, P. Cunningham, and A. Kokaram. A wavelet packet representation of audio signals for music genre classification using different ensemble and feature selection techniques. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, pages 102–108, 2003.

[41] L. Grubb and R.B. Dannenberg. A stochastic method of tracking a vocal performer. In *Proceedings of the International Computer Music Conference*, pages 301–308, 1997.

[42] J. Haitsma, M. van der Veen, T. Kalker, and F. Bruekers. Audio watermarking for monitoring and copy protection. In *Proceedings of the ACM Workshops on Multimedia*, pages 119–122, 2000.

[43] Harmonica. Accompanying action on music information in libraries, July 2006. http://projects.fnb.nl/harmonica/.

[44] C. Harte, M. Sandler, S. Abdallah, and E. Gómez. Symbolic representation of musical chords: a proposed syntax for text annotations. In *Proceedings of the International Conference on Music Information Retrieval*, pages 66–71, 2005.

[45] J. Harvell and C. Clark. Analysis of the quantitative data of system performance. Deliverable 7c, LIB-JUKEBOX/4-1049: Music Across Borders, July 2006. http://www.statsbiblioteket.dk/Jukebox/edit-report-1.html.

[46] G. Haus and E. Pollastri. A multimodal framework for music inputs. In *Proceedings of the ACM Multimedia Conference*, pages 282–284, 2000.

[47] Y. Hijikata, K. Iwahama, K. Takegawa, and S. Nishida. Content-based music filtering system with editable user profile. In *Proceedings of the ACM Symposium on Applied Computing*, pages 1050–1057, 2006.

[48] K. Hoashi, K. Matsumoto, and N. Inoue. Personalization of user profiles for content-based music retrieval based on relevance Feedback. In *Proceedings of the ACM International Conference on Multimedia*, pages 110–119, 2003.

[49] H.H. Hoos, K. Renz, and M. Görg. GUIDO/MIR—an experimental musical information retrieval system based on GUIDO music notation. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 41–50, 2001.

[50] J.-L. Hsu, C.C. Liu, and A.L.P. Chen. Efficient repeating pattern finding in music databases. In *Proceeding of the International Conference on Information and Knowledge Management*, pages 281–288, 1998.

[51] Humdrum. The Humdram Toolkit: software for music research, July 2006. http://www.music-cog.ohio-state.edu/Humdrum/.

[52] D. Huron. *The Humdrum Toolkit: Reference Manual*. Center for Computer Assisted Research in the Humanities, Menlo Park, CA, 1995.

[53] N. Hu, R.B. Dannenberg, and A.L. Lewis. A probabilistic model of melodic similarity. In *Proceedings of the International Computer Music Conference*, pages 509–515, 2002.

[54] N. Hu, R.B. Dannenberg, and G. Tzanetakis. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 185–188, 2003.

[55] N. Hu and R.B. Dannenberg. A comparison of melodic database retrieval techniques using sung queries. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 301–307, 2002.

[56] IMIRSEL. The international music information retrieval system evaluation laboratory project, July 2006. http://www.music-ir.org/evaluation/.

[57] International Organization for Standardization (ISO). Information technology: Standard Music Description Language (SMDL). Technical Report ISO/IECDIS 10743:1995, ISO International Electrotechnical Commission, Geneva, Switzerland, July 2006. ftp://ftp.techno.com/pub/SMDL/.

[58] E. Isaacson. What you see is what you get: on visualizing music. In *Proceedings of the International Conference on Music Information Retrieval*, pages 389–395, 2005.

[59] ISMIR. The international conferences on music information retrieval, July 2006. http://www.ismir.net/.

[60] K. Sparck Jones and P. Willett. *Readings in Information Retrieval*. Morgan Kaufmann, San Francisco, CA, 1997.

[61] T. Kageyama and Y. Takashima. A melody retrieval method with hummed melody (written in japanese). *Transactions of the Institute of Electronics, Information and Communication Engineers*, J77D-II(8):1543–1551, 1994. Cited in Ghias *et al.*, 1995.

[62] A. Kapur, M. Benning, and G. Tzanetakis. Query-by-beat-boxing: music retrieval for the DJ. In *Proceedings of the International Conference on Music Information Retrieval*, pages 170–177, 2004.

[63] M.L. Kherfi, D. Ziou, and A. Bernardi. Image retrieval from the World Wide Web: issues, techniques, and systems. *ACM Computing Surveys*, 36(1):35–67, 2004.

[64] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H.G. Okuno. Instrument identification in polyphonic music: feature weighting with mixed sound, pitch dependent timbre modeling, and use of musical content. In *Proceedings of the International Conference on Music Information Retrieval*, pages 558–563, 2005.

[65] R. Kochumman, C. Monroy, R. Furuta, A. Goenka, E. Urbina, and E. Melgoza. Towards an electronic variorum edition of Cervantes' Don Quixote: visualizations that support preparation. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 199–200, 2002.

[66] C.L. Krumhansl. Why is musical timbre so hard to understand? In S. Nielsen and O. Olsson, editors, *Structure and Perception Electroacoustic Sound and Music*, pages 45–53. Elsevier, Amsterdam, NL, 1989.

[67] C.L. Krumhansl. The geometry of musical structure: a brief introduction and history. *Computer in Entertainment*, 3(4):1–14, 2005.

[68] F.-F. Kuo and M.-K. Shan. Looking for new, not known music only: music retrieval by melody style. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 243–251, 2004.

[69] V. Lavrenko and J. Pickens. Polyphonic music modeling with random fields. In *Proceedings of the ACM International Conference on Multimedia*, pages 120–129, 2003.

[70] G.H. Leazer. The effectiveness of keyword searching in the retrieval of musical works on sound recordings. *Cataloging and Classification Quarterly*, 15(3):15–55, 1992.

[71] J.H. Lee and J.S. Downie. Survey of music information needs, uses, and seeking behaviours: preliminary findings. In *Proceedings of the International Conference on Music Information Retrieval*, pages 441–446, 2004.

[72] F. Lerdhal and R. Jackendoff. *A Generative Theory of Tonal Music*. The MIT Press, Cambridge, MA, 1983.

[73] M. Lesaffre, M. Leman, K. Tanghe, B. De Baets, H. De Meyer, and J.-P. Martens. User-dependent taxonomy of musical features as a conceptual framework for musical audio-mining technology. In *Proceedings of the Stockholm Music Acoustics Conference*, pages 635–638, 2003.

[74] M. Lesaffre, K. Tanghe, G. Martens, D. Moelants, M. Leman, B. De Baets, H. De Meyer, and J.-P. Martens. The MAMI query-by-voice experiment: collecting and annotating vocal queries for music information retrieval. In *Proceedings of the International Conference on Music Information Retrieval*, pages 65–71, 2003.

[75] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[76] N.-H. Liu, Y.-H. Wu, and A.L.P. Chen. Efficient K-NN search in polyphonic music databases using a lower bounding mechanism. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, pages 163–170, 2003.

[77] Q. Li, B.M. Kim, D.H. Guan, and D.W. Oh. A music recommender based on audio features. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 532–533, 2004.

[78] T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 282–289, 2003.

[79] B. Logan. Content-based playlist generation: exploratory experiments. In *Proceedings of the International Conference on Music Information Retrieval*, pages 295–296, 2002.

[80] B. Logan. Music recommendation from song sets. In *Proceedings of the International Conference on Music Information Retrieval*, pages 425–528, 2004.

[81] D. Lübber. Sonixplorer: combining visualization and auralization for content-based exploration of music collections. In *Proceedings of the International Conference on Music Information Retrieval*, pages 590–593, 2005.

[82] A. Lubiw and L. Tanur. Pattern matching in polyphonic music as a weighted geometric translation problem. In *Proceedings of the International Conference of Music Information Retrieval*, pages 289–296, 2004.

[83] MAMI. Musical Audio Mining—"query by humming", July 2006. http://www.ipem.ugent.be/MAMI/.

[84] S. Mcadams. Perspectives on the contribution of timbre to musical structure. *Computer Music Journal*, 23(3):85–102, 1999.

[85] A. McLane. Music as information. In M.E. Williams, editor, *Arist*, volume 31, chapter 6, pages 225–262. American Society for Information Science, 1996.

[86] C. Meek and W. Birmingham. Johnny can't sing: a comprehensive error model for sung music queries. In *Proceedings of the International Conference on Music Information Retrieval*, pages 65–71, 2002.

[87] C. Meek and W. Birmingham. Automatic thematic extractor. *Journal of Intelligent Information Systems*, 21(1):9–33, 2003.

[88] M. Melucci, N. Orio, and M. Gambalunga. An evaluation study on music perception for musical content-based information Retrieval. In *Proceedings of the International Computer Music Conference*, pages 162–165, 2000.

[89] M. Melucci and N. Orio. Musical information retrieval using melodic surface. In *Proceedings of the ACM Conference on Digital Libraries*, pages 152–160, 1999.

[90] M. Melucci and N. Orio. Combining melody processing and information retrieval techniques: methodology, evaluation, and system implementation. *Journal of the American Society for Information Science and Technology*, 55(12):1058–1066, 2004.

[91] R. Middleton. *Studying Popular Music*. Open University Press, Philadelphia, PA, 2002.

[92] Mirex 2006 Wiki. Second annual music information retrieval evaluation exchange, July 2006. http://www.music-ir.org/mirex2006/.

[93] F. Mörchen, A. Ultsch, M. Nöcker, and C. Stamm. Databionic visualization of music collections according to perceptual distance. In *Proceedings of the International Conference on Music Information Retrieval*, pages 396–403, 2005.

[94] MPEG. The MPEG home page, July 2006. http://www.chiariglione.org/mpeg/.

[95] M. Müller, F. Kurth, and M. Clausen. Audio matching via chroma-based statistical features. In *Proceedings of the International Conference of Music Information Retrieval*, pages 288–295, 2005.

[96] MuseData. An electronic library of classical music scores, July 2006. http://www.musedata.org/.

[97] Musica. The international database of choral repertoire, July 2006. http://www.musicanet.org/.

[98] MusicXML. Recordare: Internet music publishing and software, July 2006. http://www.musicxml.org/.

[99] GUIDO Music Notation. The GUIDO NoteServer, July 2006. http://www.noteserver.org/.

[100] Music Notation. Formats, July 2006. http://www.music-notation.info/.

[101] MusiXTeX. MusiXtex and related software, July 2006. http://icking-music-archive.org/software/indexmt6.html.

[102] E. Narmour. *The Analysis and Cognition of Basic Melodic Structures*. University of Chicago Press, Chicago, MI, 1990.

[103] G. Neve and N. Orio. Indexing and retrieval of music documents through pattern analysis and data Fusion Techniques. In *Proceedings of the International Conference on Music Information Retrieval*, pages 216–223, 2004.

[104] N.M. Norowi, S. Doraisamy, and R. Wirza. Factors affecting automatic genre classification: an investigation incorporating non-western musical forms. In *Proceedings of the International Conference on Music Information Retrieval*, pages 13–20, 2005.

[105] N. Orio and G. Neve. Experiments on segmentation techniques for music documents indexing. In *Proceedings of the International Conference on Music Information Retrieval*, pages 104–107, 2005.

[106] N. Orio. Alignment of performances with scores aimed at content-based music access and Retrieval. In *Proceedings of European Conference on Digital Libraries*, pages 479–492, 2002.

[107] R.P. Paiva, T. Mendes, and A. Cardoso. On the detection of melody notes in polyphonic audio. In *Proceedings of the International Conference on Music Information Retrieval*, pages 175–182, 2005.

[108] E. Pampalk, S. Dixon, and G. Widmer. Exploring music collections by browsing different views. In *Proceedings of the International Conference on Music Information Retrieval*, pages 201–208, 2003.

[109] C.L. Parker. A tree-based method for fast melodic retrieval. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 254–255, 2004.

[110] G. Peeters. Rhythm classification using spectral rhythm patterns. In *Proceedings of the International Conference on Music Information Retrieval*, pages 644–647, 2005.

[111] J. Pickens and T. Crawford. Harmonic models for polyphonic music retrieval. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 430–437, 2002.

[112] A. Pienimäki and K. Lemström. Clustering symbolic music using paradigmatic and surface level analyses. In *Proceedings of the International Conference of Music Information Retrieval*, pages 262–265, 2004.

[113] A. Pienimäki. Indexing music database using automatic extraction of frequent phrases. In *Proceedings of the International Conference on Music Information Retrieval*, pages 25–30, 2002.

[114] S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:380–393, 1997.

[115] A. Pikrakis, I. Antonopoulos, and S. Theodoris. Music meter and tempo tracking from raw polyphonic audio. In *Proceedings of the International Conference on Music Information Retrieval*, pages 192–197, 2004.

[116] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

[117] C. Raphael and J. Stoddard. Harmonic analysis with probabilistic graphical models. In *Proceedings of the International Conference on Music Information Retrieval*, pages 177–181, 2003.

[118] C. Raphael. Automatic segmentation of acoustic musical signals using hidden markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4):360–370, 1999.

[119] T. Rohdenburg, V. Hohmann, and B. Kollmeier. Objective perceptual quality measures for the evaluation of noise reduction schemes. In *Proceedings of the*

International Workshop on Acoustic Echo and Noise Control, pages 169–172, 2005.

[120] J. Rothstein. *Midi: A Comprehensive Introduction*. A-R Editions, Madison, WI, 1991.

[121] E. Selfridge-Field. *Beyond MIDI: The Handbook of Musical Codes*. MIT Press, Cambridge, MA, 1997.

[122] Shazam. Home page, July 2006. http://www.shazam.com/.

[123] J. Shifrin, B. Pardo, C. Meek, and W. Birmingham. Hmm-based musical query retrieval. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 295–300, 2002.

[124] H.-H. Shih, S.S. Narayanan, and C.-C.J. Kuo. Multidimensional humming transcription using hidden Markov models for query by humming systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 541–544, 2003.

[125] D.R. Stammen and B. Pennycook. Real-time recognition of melodic fragments using the dynamic timewarp algorithm. In *Proceedings of the International Computer Music Conference*, pages 232–235, 1993.

[126] R. Stenzel and T. Kamps. Improving content-based similarity measures by training a collaborative model. In *Proceedings of the International Conference on Music Information Retrieval*, pages 264–271, 2005.

[127] MIR Systems. A survey of music information retrieval systems, July 2006. http://mirsystems.info/.

[128] J. Tenney and L. Polansky. Temporal gestalt perception in music. *Journal of Music Theory*, 24(2):205–241, 1980.

[129] P. Toiviainen. Visualization of tonal content with self-organizing maps and self-similarity Matrices. *Computers in Entertainment*, 3(4):1–10, 2005.

[130] TREC. Text REtrieval Conference home page, July 2006. http://trec.nist.gov/.

[131] Y. Tseng. Content-based retrieval for music collections. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 176–182, 1999.

[132] R. Typke, M. den Hoed, J. de Nooijer, F. Wiering, and R.C. Veltkamp. A ground truth for half a million musical incipits. *Journal of Digital Information Management*, 3(1):34–39, 2005.

[133] R. Typke, R.C. Veltkamp, and F. Wiering. Searching notated polyphonic music using transportation distances. In *Proceedings of the ACM International Conference on Multimedia*, pages 128–135, 2004.

[134] R. Typke, R.C. Veltkamp, and F. Wiering. Evaluating retrieval techniques based on partially ordered ground truth lists. In *Proceedings of the International Conference of Multimedia and Expo*, 2006.

[135] R. Typke, F. Wiering, and R.C. Veltkamp. A survey on music information retrieval systems. In *Proceedings of the International Conference on Music Information Retrieval*, pages 153–160, 2005.

[136] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[137] A. Uitdenbogerd and J. Zobel. Manipulation of music for melody matching. In *Proceedings of the ACM Conference on Multimedia*, pages 235–240, 1998.

[138] E. Ukkonen, K. Lemström, and V. Mäkinen. Geometric algorithms for transposition invariant content-based music retrieval. In *Proceedings of the International Conference of Music Information Retrieval*, pages 193–199, 2003.

[139] E. Unal, S.S. Narayanan, and E. Chew. A statistical approach to retrieval under user-dependent uncertainty in query-by-humming systems. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, pages 113–118, 2004.

[140] E. Vincent and X. Rodet. Instrument identification in solo and ensemble music using independent subspace analysis. In *Proceedings of the International Conference on Music Information Retrieval*, pages 576–581, 2003.

[141] Wedelmusic. Server of sound and sound-processing, July 2006. http://www.wedelmusic.org.

[142] G.A. Wiggins, K. Lemström, and D. Meredith. SIA(M)ESE: an algorithm for transposition invariant, polyphonic content-based music retrieval. In *Proceedings of the International Conference of Music Information Retrieval*, pages 283–284, 2002.

[143] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 3(3):27–36, 1996.

[144] T. Zhang and C.-C. Jay Kuo. Hierarchical system for content-based audio classification and retrieval. In *Proceedings of International Conference on Speech, Audio, and Signal Processing, Vol. 6*, pages 3001–3004, 1999.