CHAPTER 1

INTRODUCTION

The objective of *Audio Content Analysis (ACA)* is the extraction of information from audio signals such as music recordings stored on digital media. The information to be extracted is usually referred to as *meta data*: it is data about (audio) data and can essentially cover any information allowing a meaningful description or explanation of the raw audio data. The meta data represents (among other things) the musical content of the recording. Nowadays, attempts have been made to automatically extract practically everything from the music recording including formal, perceptual, musical, and technical meta data. Examples range from tempo and key analysis — ultimately leading to the complete transcription of recordings into a score-like format — over the analysis of artists' performances of specific pieces of music to approaches to modeling the human emotional affection when listening to music.

In addition to the meta data extractable from the signal itself there is also meta data which is neither implicitly nor explicitly included in the music signal itself but represents additional information on the signal, such as the year of the composition or recording, the record label, the song title, information on the artists, etc.

The examples given above already imply that ACA is a multi-disciplinary research field. Since it deals with audio signals, the main emphasis lies on (digital) signal processing. But depending on the task at hand, the researcher may be required to use knowledge from different research fields such as musicology and music theory, (music) psychology, psychoacoustics, audio engineering, library science, and last but not least computer science for pattern recognition and machine learning. If the research is driven by commercial interests, even legal and economical issues may be of importance.

An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics, First Edition. Alexander Lerch.

The term *audio content analysis* is not the only one used for systems analyzing audio signals. Frequently, the research field is also called *Music Information Retrieval (MIR)*. MIR should be understood as a more general, broader field of which ACA is a part. Downie and Orio have both published valuable introductory articles in the field of MIR [1, 2]. In contrast to ACA, MIR also includes the analysis of symbolic non-audio music formats such as musical scores and files or signals compliant to the so-called *Musical Instrument Digital Interface (MIDI)* protocol [3]. Furthermore, MIR may include the analysis and retrieval of information that is music-related but cannot be (easily) extracted from the audio signal such as the song lyrics, user ratings, performance instructions in the score, or bibliographical information such as publisher, publishing date, the work's title, etc. Therefore the term audio content analysis seems to be the most accurate for the description of the approaches to be covered in the following. In the past, other terms have been used more or less synonymously to the term audio content analysis. Examples of such synonyms are *machine listening* and *computer audition*. *Computational Auditory Scene Analysis (CASA)* is closely related to ACA but usually has a strong focus on modeling the human perception of audio.

Historically, the first systems analyzing the content of audio signals appear shortly after technology provided the means of storing and reproducing recordings on media in the 20th century. One early example is Seashore's Tonoscope, which allowed one to analyze the pitch of an audio signal by visualizing the fundamental frequency of the incoming audio signal on a rotating drum [4]. However, the development of digital storage media and digital signal processing during the last decades, along with the growing amount of digital audio data available through broadband connections, has significantly increased both the need and the possibilities of automatic systems for analyzing audio content, resulting in a lively and growing research field. A short introduction to extracting information from audio on different levels has been published by Ellis [5].

Audio content analysis systems can be used on a relatively wide variety of tasks. Obviously, the automatic generation of meta data is of great use for the retrieval of music signals with specific characteristics from large databases or the Internet. Here, the manual annotation of meta data by humans is simply not feasible due to the sheer amount of (audio) data. Therefore, only computerized tags can be used to find files or excerpts of files with, e.g., a specific tempo, instrumentation, chord progression, etc. The same information can be used in end consumer applications such as for the automatic generation of play lists in music players or in automatic music recommendation systems based on the user's music database or listening habits. Another typical area of application is music production software. Here, the aim of ACA is on the one hand to allow the user to interact with a more "musical" software interface — e.g., by displaying score-like information along with the audio data — and thus enabling a more intuitive approach to visualization and editing the audio data. On the other hand, the software can support the user by giving suggestions of how to combine and process different audio signals. For instance, software applications for DJs nowadays include technology allowing the (semi-) automatic alignment of audio loops and complete mixes based on previously extracted information such as the tempo and key of the signals. In summary, ACA can help with

- automatic organization of audio content in large databases as well as search and retrieve audio files with specific characteristics in such databases (including the tasks of song identification and recommendation),
- new approaches and interfaces to search and retrieval of audio data such as query-by-humming systems,

- new ways of sound visualization, user interaction, and musical processing in music software such as an audio editor displaying the current score position or an automatically generated accompaniment,
- intelligent, content-dependent control of audio processing (effect parameters, intelligent cross fades, time stretching, etc.) and audio coding algorithms, and
- automatic play list generation in media players.

1.1 Audio Content

The content or information conveyed by recordings of music is obviously multi-faceted. It originates from three different sources:

- *Score*: The term score will be used broadly as a definition of musical ideas. It can refer to any form of notating music from the *basso continuo* (a historic way of defining the harmonic structure) and the classic western score notation to the lead sheet and other forms of notation used for contemporary and popular music.

  Examples of information originating in the score are the melody or hook line, the key and the harmony progression, rhythmic aspects and specific temporal patterns, the instrumentation, as well as structural information such as repetitions and phrase boundaries.

- *Performance*: Music as a performing art requires a performer or group of performers to generate a unique acoustical rendition of the underlying musical ideas. The performers will use the information provided by the score but may interpret and modify it as well as they may dismiss parts of the contained information or add new information.

  Typical performance aspects include the tempo and its variation as well as the micro-timing, the realization of musical dynamics, accents and instantaneous dynamic modulations such as tremolo (see Sect. 4.2), the usage of specific temperaments (see Sect. 5.2.5.2) and expressive intonation and vibrato (see Sect. 5.2.5.3), and specific playing (e.g., bowing) techniques influencing the sound quality.

- *Production*: The process of recording the performance and the (post-) production process will impact certain characteristics of the recording.

  These are mainly the sound quality of the recording (by microphone positioning, equalization, and by applying effects to the signal) and the dynamics (by applying manual or automatic gain adjustments). Changes in timing and pitch may occur as well by editing the recording and applying software for pitch correction.

There are certain characteristics which cannot easily be assigned to a single category; the timbre of a recording can be determined by the instrumentation indicated by the score, by the specific choice of instruments (e.g., historical instruments, specific guitar amps, etc.), by specific playing techniques, and by sound processing choices made by the sound engineer or producer.

ACA systems may in principle cover the extraction of information from all three categories. In many cases, however, no distinction is being made between those categories by researchers and their systems, respectively. The reason is that popular music in the tradition of western music is one of the main targets of the research for several — last but not

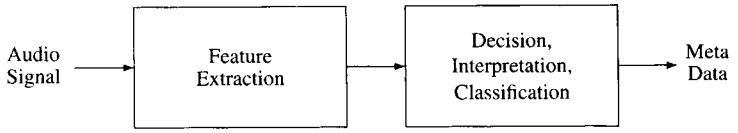

Figure 1.1 General processing stages of a system for audio content analysis

least commercial — reasons and that with popular music a score-like raw representation of musical ideas cannot be distinguished as easily from the performance and production as in "classical" or traditional western music.

From a technical point of view, five general classes can be identified to describe the content of a music recording on a low level:

- *statistical or technical signal characteristics* derived from the audio data such as the amplitude distribution etc. (see Sects. 3.2 and 3.4),
- *timbre or sound quality characteristics* (see Sect. 3.3),
- *intensity-related characteristics* such as envelope-, level-, and loudness-related properties (see Chap. 4),
- *tonal characteristics* which include the pitches and pitch relations in the signal (see Chap. 5), and
- *temporal characteristics* such as rhythmic and timing properties of the signal (see Sect. 6).

The basic information clustered in each individual class can be used and combined to gather a deeper knowledge of the music such as on musical structure, style, performance characteristics, or even transported mood or emotional affection. Especially the last example, however, shows that while many parameters to be extracted from an audio file are objective in a way that they describe music properties independent of the perceptual context (e.g., key, tempo), other properties depend on the individual listener's music experience or way of perceiving music. Not only might this experience vary between the multitude of different listeners, but it might also vary with the listener's individual mood and situation.

1.2 A Generalized Audio Content Analysis System

Most existing systems for the analysis of audio content can be structured into two major processing stages as depicted in Fig. 1.1.

In the first processing stage so-called features are extracted from the audio signal. This extraction process serves two purposes:

- *Dimensionality reduction*: When processing a whole audio file, the raw amount of data in a whole audio file is too large to handle it in a meaningful way. One channel of a digital audio file in Compact Disc (CD) quality (44, 100 samples per second, 16 bits per sample) with a length of 5 minutes contains

$$5\,\text{min} \cdot 60\,^{\text{s}}/_{\text{min}} \cdot 44100\,^{\text{samples}}/_{\text{s}} \cdot 16\,^{\text{bits}}/_{\text{sample}} = 211,680,000\,\text{bits}. \quad (1.1)$$

A feature (or a series of featuress) is used to represent this data with fewer values by suppressing (hopefully) irrelevant information. A typical instantaneous feature will produce one single feature value for each block of audio samples or even for the whole signal from beginning to end.

- More *meaningful representation*: Although all the information that can possibly be extracted is implicitly contained in the raw audio data, it is necessary to focus on the relevant aspects and to transform the audio data into a representation easily interpretable by humans or machines. If, for instance, the variation of brightness over time is of interest, one would have difficulties to extract such information by simply observing the series of audio samples. Instead, a model of the human perception of brightness is required, however simple or sophisticated this model may be. In the case of brightness, we would probably be interested in a measure of spectral distribution (see Sect. 3.3).

A feature is not necessarily required to be meaningful in a perceptual or musical way and does not have to be interpretable by humans. It may also just be designed to provide condensed information to the second processing stage of an ACA system to support the generation of a reliable overall result. Usually a distinction is made between low-level features and high-level features. Low-level features are generally considered to have no direct (humanly interpretable) meaning as opposed to high-level features which represent terms in which humans refer to music such as tempo, structure, etc. Those high-level features are usually extracted in the second processing stage shown in Fig. 1.1.

Obviously, the term feature is not very clearly defined but is used for any lower dimensional representation of the audio signal to be interpreted. Features can be used to compute a result but can also be used to calculate derived, more meaningful "features."

The second processing stage of an ACA system takes the extracted feature data and attempts to map it into a domain both usable and comprehensible. Thus, it turns the low-level feature data into a high-level feature and meaningful meta data, respectively. This process can be accomplished by a classification system (sorting the input into pre-defined and trained categories) or by applying (empirical or musical) knowledge to the task.

Since there is no clear objective distinction between low-level features and high-level features, it is sometimes a context-dependent decision whether the system output is referred to as low-level or high-level description. In fact, we face probably an unlimited number of abstraction levels between the raw audio data and the different (human) ways of referring to music. While one system might be referring to the tempo of a recording as high-level information, another system might use this information just as one feature amongst many others to, for example, automatically recognize the musical style. Ultimately, there can only be the conclusion that an ACA system may either consist only of one instance of the two processing stages or of any arbitrary number of nested instances of such processing stages with the output of one instance forming one of the inputs of the following instance.