1
Introduction

Today, digital audio applications are part of our everyday lives. Popular examples include audio CDs, MP3 audio players, radio broadcasts, TV or video DVDs, video games, digital cameras with sound track, digital camcorders, telephones, telephone answering machines and telephone enquiries using speech or word recognition.

Various new and advanced audiovisual applications and services become possible based on audio content analysis and description. Search engines or specific filters can use the extracted description to help users navigate or browse through large collections of data. Digital analysis may discriminate whether an audio file contains speech, music or other audio entities, how many speakers are contained in a speech segment, what gender they are and even which persons are speaking. Spoken content may be identified and converted to text. Music may be classified into categories, such as jazz, rock, classics, etc. Often it is possible to identify a piece of music even when performed by different artists – or an identical audio track also when distorted by coding artefacts. Finally, it may be possible to identify particular sounds, such as explosions, gunshots, etc.

We use the term audio to indicate all kinds of audio signals, such as speech, music as well as more general sound signals and their combinations. Our primary goal is to understand how meaningful information can be extracted from digital audio waveforms in order to compare and classify the data efficiently. When such information is extracted it can also often be stored as content description in a compact way. These compact *descriptors* are of great use not only in audio storage and retrieval applications, but also for efficient content-based classification, recognition, browsing or filtering of data. A data descriptor is often called a *feature vector* or *fingerprint* and the process for extracting such feature vectors or fingerprints from audio is called *audio feature extraction* or *audio fingerprinting*.

Usually a variety of more or less complex descriptions can be extracted to fingerprint one piece of audio data. The efficiency of a particular fingerprint

MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval H.-G. Kim, N. Moreau and T. Sikora

used for comparison and classification depends greatly on the application, the extraction process and the richness of the description itself. This book will provide an overview of various strategies and algorithms for automatic extraction and description. We will provide various examples to illustrate how trade-offs between size and performance of the descriptions can be achieved.

1.1 AUDIO CONTENT DESCRIPTION

Audio content analysis and description has been a very active research and development topic since the early 1970s. During the early 1990s – with the advent of digital audio and video – research on audio and video retrieval became equally important. A very popular means of audio, image or video retrieval is to annotate the media with text, and use text-based database management systems to perform the retrieval. However, text-based annotation has significant drawbacks when confronted with large volumes of media data. Annotation can then become significantly labour intensive. Furthermore, since audiovisual data is rich in content, text may not be rich enough in many applications to describe the data. To overcome these difficulties, in the early 1990s content-based retrieval emerged as a promising means of describing and retrieving audiovisual media. Content-based retrieval systems describe media data by their audio or visual content rather than text. That is, based on audio analysis, it is possible to describe sound or music by its spectral energy distribution, harmonic ratio or fundamental frequency. This allows a comparison with other sound events based on these features and in some cases even a classification of sound into general sound categories. Analysis of speech tracks may result in the recognition of spoken content.

In the late 1990s – with the large-scale introduction of digital audio, images and video to the market – the necessity for interworking between retrieval systems of different vendors arose. For this purpose the ISO Motion Picture Experts Group initiated the MPEG-7 "Multimedia Content Description Interface" work item in 1997. The target of this activity was to develop an international MPEG-7 standard that would define standardized descriptions and description systems. The primary purpose is to allow users or agents to search, identify, filter and browse audiovisual content. MPEG-7 became an international standard in September 2001. Besides support for metadata and text descriptions of the audiovisual content, much focus in the development of MPEG-7 was on the definition of efficient content-based description and retrieval specifications.

This book will discuss techniques for analysis, description and classification of digital audio waveforms. Since MPEG-7 plays a major role in this domain, we will provide a detailed overview of MPEG-7-compliant techniques and algorithms as a starting point. Many state-of-the-art analysis and description

algorithms beyond MPEG-7 are introduced and compared with MPEG-7 in terms of computational complexity and retrieval capabilities.

1.2 MPEG-7 AUDIO CONTENT DESCRIPTION – AN OVERVIEW

The MPEG-7 standard provides a rich set of standardized tools to describe multimedia content. Both human users and automatic systems that process audiovisual information are within the scope of MPEG-7. In general MPEG-7 provides such tools for audio as well as images and video data.[1] In this book we will focus on the audio part of MPEG-7 only.

MPEG-7 offers a large set of audio tools to create descriptions. MPEG-7 descriptions, however, do not depend on the ways the described content is coded or stored. It is possible to create an MPEG-7 description of analogue audio in the same way as of digitized content.

The main elements of the MPEG-7 standard related to audio are:

- Descriptors (D) that define the syntax and the semantics of audio feature vectors and their elements. Descriptors bind a feature to a set of values.
- Description schemes (DSs) that specify the structure and semantics of the relationships between the components of descriptors (and sometimes between description schemes).
- A description definition language (DDL) to define the syntax of existing or new MPEG-7 description tools. This allows the extension and modification of description schemes and descriptors and the definition of new ones.
- Binary-coded representation of descriptors or description schemes. This enables efficient storage, transmission, multiplexing of descriptors and description schemes, synchronization of descriptors with content, etc.

The MPEG-7 content descriptions may include:

- Information describing the creation and production processes of the content (director, author, title, etc.).
- Information related to the usage of the content (copyright pointers, usage history, broadcast schedule).
- Information on the storage features of the content (storage format, encoding).
- Structural information on temporal components of the content.
- Information about low-level features in the content (spectral energy distribution, sound timbres, melody description, etc.).

[1] An overview of the general goals and scope of MPEG-7 can be found in: Manjunath M., Salembier P. and Sikora T. (2001) *MPEG-7 Multimedia Content Description Interface*, John Wiley & Sons, Ltd.

- Conceptual information on the reality captured by the content (objects and events, interactions among objects).
- Information about how to browse the content in an efficient way.
- Information about collections of objects.
- Information about the interaction of the user with the content (user preferences, usage history).

Figure 1.1 illustrates a possible MPEG-7 application scenario. Audio features are extracted on-line or off-line, manually or automatically, and stored as MPEG-7 descriptions next to the media in a database. Such descriptions may be low-level audio descriptors, high-level descriptors, text, or even speech that serves as spoken annotation.

Consider an audio broadcast or audio-on-demand scenario. A user, or an agent, may only want to listen to specific audio content, such as news. A specific filter will process the MPEG-7 descriptions of various audio channels and only provide the user with content that matches his or her preference. Notice that the processing is performed on the already extracted MPEG-7 descriptions, not on the audio content itself. In many cases processing the descriptions instead of the media is far less computationally complex, usually in an order of magnitude.

Alternatively a user may be interested in retrieving a particular piece of audio. A request is submitted to a search engine, which again queries the MPEG-7 descriptions stored in the database. In a browsing application the user is interested in retrieving similar audio content.

Efficiency and accuracy of filtering, browsing and querying depend greatly on the richness of the descriptions. In the application scenario above, it is of great help if the MPEG-7 descriptors contain information about the category of

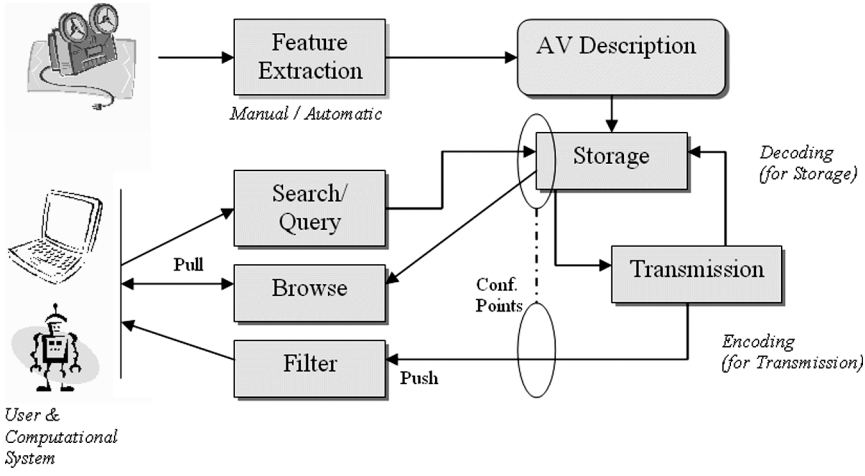

Figure 1.1 MPEG-7 application scenario

the audio files (i.e. whether the broadcast files are news, music, etc.). Even if this is not the case, it is often possible to categorize the audio files based on the low-level MPEG-7 descriptors stored in the database.

1.2.1 MPEG-7 Low-Level Descriptors

The MPEG-7 low-level audio descriptors are of general importance in describing audio. There are 17 temporal and spectral descriptors that may be used in a variety of applications. These descriptors can be extracted from audio automatically and depict the variation of properties of audio over time or frequency. Based on these descriptors it is often feasible to analyse the similarity between different audio files. Thus it is possible to identify identical, similar or dissimilar audio content. This also provides the basis for classification of audio content.

Basic Descriptors

Figure 1.2 depicts instantiations of the two MPEG-7 audio basic descriptors for illustration purposes, namely the audio waveform descriptor and the audio power descriptor. These are time domain descriptions of the audio content. The temporal variation of the descriptors' values provides much insight into the characteristics of the original music signal.

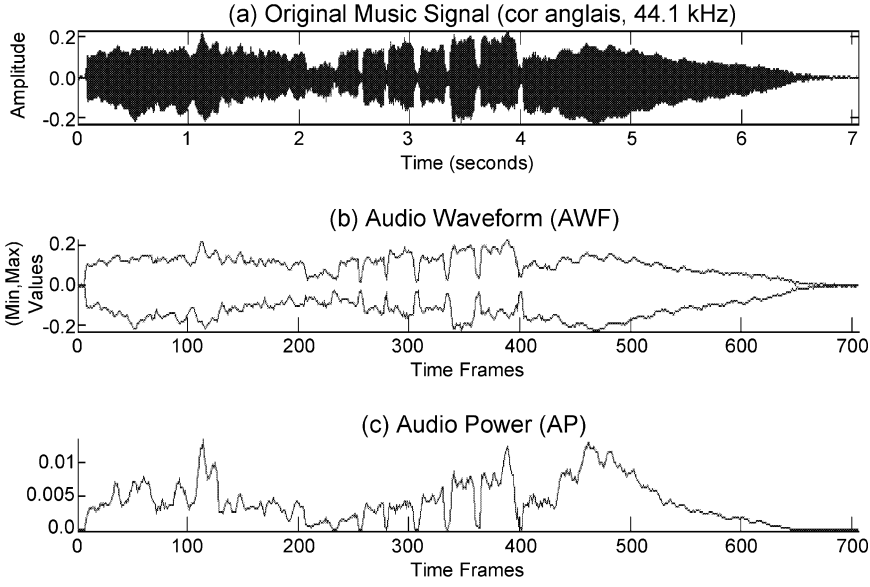

Figure 1.2 MPEG-7 basic descriptors extracted from a music signal (cor anglais, 44.1 kHz)

Basic Spectral Descriptors
The four basic spectral audio descriptors are all derived from a single time–frequency analysis of an audio signal. They describe the audio spectrum in terms of its envelope, centroid, spread and flatness.

Signal Parameter Descriptors
The two signal parameter descriptors apply only to periodic or quasi-periodic signals. They describe the fundamental frequency of an audio signal as well as the harmonicity of a signal.

Timbral Temporal Descriptors
Timbral temporal descriptors can be used to describe temporal characteristics of segments of sounds. They are especially useful for the description of musical timbre (characteristic tone quality independent of pitch and loudness).

Timbral Spectral Descriptors
Timbral spectral descriptors are spectral features in a linear frequency space, especially applicable to the perception of musical timbre.

Spectral Basis Descriptors
The two spectral basis descriptors represent low-dimensional projections of a high-dimensional spectral space to aid compactness and recognition. These descriptors are used primarily with the sound classification and indexing description tools, but may be of use with other types of applications as well.

1.2.2 MPEG-7 Description Schemes

MPEG-7 DSs specify the types of descriptors that can be used in a given description, and the relationships between these descriptors or between other DSs. The MPEG-7 DSs are written in XML. They are defined using the MPEG-7 description definition language (DDL), which is based on the XML Schema Language, and are instantiated as documents or streams. The resulting descriptions can be expressed in a textual form (i.e. human-readable XML for editing, searching, filtering) or in a compressed binary form (i.e. for storage or transmission).

Five sets of audio description tools that roughly correspond to application areas are integrated in the standard: audio signature, musical instrument timbre, melody description, general sound recognition and indexing, and spoken content. They are good examples of how the MPEG-7 audio framework may be integrated to support other applications.

Musical Instrument Timbre Tool

The aim of the timbre description tool is to specify the perceptual features of instruments with a reduced set of descriptors. The descriptors relate to notions such as "attack", "brightness" or "richness" of a sound. Figures 1.3 and 1.4 illustrate the XML instantiations of these descriptors using the MPEG-7 audio description scheme for a harmonic and a percussive instrument type. Notice that the description of the instruments also includes temporal and spectral features of the sound, such as spectral and temporal centroids. The particular values fingerprint the instruments and can be used to distinguish them from other instruments of their class.

Audio Signature Description Scheme

Low-level audio descriptors in general can serve many conceivable applications. The spectral flatness descriptor in particular achieves very robust matching of

```
<AudioDescriptionScheme xsi:type="PercussiveInstrumentTimbreType">
  <LogAttackTime>
    <Scalar>-1.683017</Scalar>
  </LogAttackTime>
  <SpectralCentroid>
    <Scalar>1217.341518</Scalar>
  </SpectralCentroid>
  <TemporalCentroid>
    <Scalar>0.081574</Scalar>
  </TemporalCentroid>
</AudioDescriptionScheme>
```

Figure 1.3 MPEG-7 audio description for a percussion instrument

```
<AudioDescriptionScheme xsi:type="HarmonicInstrumentTimbreType">
  <LogAttackTime>
    <Scalar>-0.150702</Scalar>
  </LogAttackTime>
  <HarmonicSpectralCentroid>
    <Scalar>1586.892383</Scalar>
  </HarmonicSpectralCentroid>
  <HarmonicSpectralDeviation>
    <Scalar>-0.027864</Scalar>
  </HarmonicSpectralDeviation>
  <HarmonicSpectralSpread>
    <Scalar>0.550866</Scalar>
  </HarmonicSpectralSpread>
  <HarmonicSpectralVariation>
    <Scalar>0.001877</Scalar>
  </HarmonicSpectralVariation>
</AudioDescriptionScheme>
```

Figure 1.4 MPEG-7 audio description for a violin instrument

audio signals, well tuned to be used as a unique content identifier for robust automatic identification of audio signals. The descriptor is statistically summarized in the audio signature description scheme. An important application is audio fingerprinting for identification of audio based on a database of known works. This is relevant for locating metadata for legacy audio content without metadata annotation.

Melody Description Tools

The melody description tools include a rich representation for monophonic melodic information to facilitate efficient, robust and expressive melodic similarity matching. The melody description scheme includes a melody contour description scheme for extremely terse, efficient, melody contour representation, and a melody sequence description scheme for a more verbose, complete, expressive melody representation. Both tools support matching between melodies, and can support optional information about the melody that may further aid content-based search, including query-by-humming.

General Sound Recognition and Indexing Description Tools

The general sound recognition and indexing description tools are a collection of tools for indexing and categorizing general sounds, with immediate application to sound effects. The tools enable automatic sound identification and indexing, and the specification of a classification scheme of sound classes and tools for specifying hierarchies of sound recognizers. Such recognizers may be used automatically to index and segment sound tracks. Thus, the description tools address recognition and representation all the way from low-level signal-based analyses, through mid-level statistical models, to highly semantic labels for sound classes.

Spoken Content Description Tools

Audio streams of multimedia documents often contain spoken parts that enclose a lot of semantic information. This information, called *spoken content*, consists of the actual words spoken in the speech segments of an audio stream. As speech represents the primary means of human communication, a significant amount of the usable information enclosed in audiovisual documents may reside in the spoken content. A transcription of the spoken content to text can provide a powerful description of media. Transcription by means of automatic speech recognition (ASR) systems has the potential to change dramatically the way we create, store and manage knowledge in the future. Progress in the ASR field promises new applications able to treat speech as easily and efficiently as we currently treat text.

The audio part of MPEG-7 contains a *SpokenContent* high-level tool targeted for spoken data management applications. The MPEG-7 *SpokenContent* tool provides a standardized representation of an ASR output, i.e. of the semantic information (the *spoken content*) extracted by an ASR system from a spoken signal. It consists of a compact representation of multiple word and/or sub-word

hypotheses produced by an ASR engine. How the *SpokenContent* description should be extracted and used is not part of the standard.

The MPEG-7 *SpokenContent* tool defines a standardized description of either a word or a phone type of lattice delivered by a recognizer. Figure 1.5 illustrates what an MPEG-7 *SpokenContent* description of the speech excerpt "film on Berlin" could look like. A lattice can thus be a word-only graph, a phone-only graph or combine word and phone hypotheses in the same graph as depicted in the example of Figure 1.5.

1.2.3 MPEG-7 Description Definition Language (DDL)

The DDL defines the syntactic rules to express and combine DSs and descriptors. It allows users to create their own DSs and descriptors. The DDL is not a modelling language such as the Unified Modeling Language (UML) but a schema language. It is able to express spatial, temporal, structural and conceptual relationships between the elements of a DS, and between DSs. It provides a rich model for links and references between one or more descriptions and the data that it describes. In addition, it is platform and application independent and human and machine readable.

The purpose of a schema is to define a class of XML documents. This is achieved by specifying particular constructs that constrain the structure and content of the documents. Possible constraints include: elements and their content, attributes and their values, cardinalities and data types.

1.2.4 BiM (Binary Format for MPEG-7)

BiM defines a generic framework to facilitate the carriage and processing of MPEG-7 descriptions in a compressed binary format. It enables the compression,

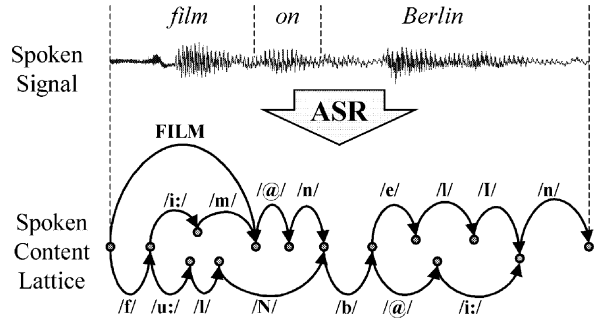

Figure 1.5 MPEG-7 *SpokenContent* description of an input spoken signal "film on Berlin"

multiplexing and streaming of XML documents. BiM coders and decoders can handle any XML language. For this purpose the schema definition (DTD or XML Schema) of the XML document is processed and used to generate a binary format. This binary format has two main properties. First, due to the schema knowledge, structural redundancy (element name, attribute names, etc.) is removed from the document. Therefore the document structure is highly compressed (98% on average). Second, elements and attribute values are encoded using dedicated source coders.

1.3 ORGANIZATION OF THE BOOK

This book focuses primarily on the digital audio signal processing aspects for content analysis, description and retrieval. Our prime goal is to describe how meaningful information can be extracted from digital audio waveforms, and how audio data can be efficiently described, compared and classified. Figure 1.6 provides an overview of the book's chapters.

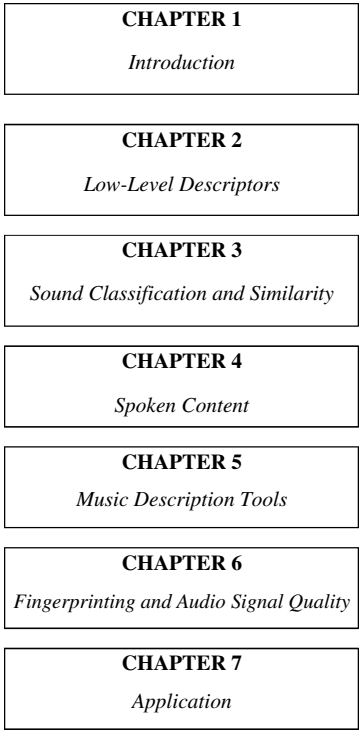

Figure 1.6 Chapter outline of the book

The purpose of Chapter 2 is to provide the reader with a detailed overview of *low-level audio descriptors*. To a large extent this chapter provides the foundations and definitions for most of the remaining chapters of the book. Since MPEG-7 provides an established framework with a large set of descriptors, the standard is used as an example to illustrate the concept. The mathematical definitions of all MPEG-7 low-level audio descriptors are outlined in detail. Other established low-level descriptors beyond MPEG-7 are introduced. To help the reader visualize the kind of information that these descriptors convey, some experimental results are given to illustrate the definitions.

In Chapter 3 the reader is introduced to the concepts of *sound similarity* and *sound classification.* Various classifiers and their properties are discussed. Low-level descriptors introduced in the previous chapter are employed for illustration. The MPEG-7 standard is again used as a starting point to explain the practical implementation of sound classification systems. The performance of MPEG-7 systems is compared with the well-established MFCC feature extraction method. The chapter provides in great detail simulation results of various systems for sound classification.

Chapter 4 focuses on MPEG-7 *SpokenContent* description. It is possible to follow most of the chapter without reading the other parts of the book. The primary goal is to provide the reader with a detailed overview of ASR and its use for MPEG-7 *SpokenContent* description. The structure of the MPEG-7 *SpokenContent* description itself is presented in detail and discussed in the context of the *spoken document retrieval* (SDR) application. The contribution of the MPEG-7 *SpokenContent* tool to the standardization and development of future SDR applications is emphasized. Many application examples and experimental results are provided to illustrate the concept.

Music description tools for specifying the properties of musical signals are discussed in Chapter 5. We focus explicitly on MPEG-7 tools. Concepts for instrument *timbre* description to specify perceptual features of musical sounds are discussed using reduced sets of descriptors. *Melodies* can be described using MPEG-7 description schemes for melodic similarity matching. We will discuss query-by-humming applications to provide the reader with examples of how melody can be extracted from a user's input and matched against melodies contained in a database.

An overview of audio fingerprinting and audio signal quality description is provided in Chapter 6. In general, the MPEG-7 low-level descriptors can be seen as providing a fingerprint for describing audio content. Audio fingerprinting has to a certain extent been described in Chapters 2 and 3. We will focus in Chapter 6 on fingerprinting tools specifically developed for the identification of a piece of audio and for describing its quality.

Chapter 7 finally provides an outline of example applications using the concepts developed in the previous chapters. Various applications and experimental results are provided to help the reader visualize the capabilities of concepts for content analysis and description.