1
Introduction

1.1 The human auditory system

The human auditory system serves several important purposes in daily life. One of the most prominent features is to understand spoken words, which allows people to communicate in an efficient and interactive manner. In case of potential danger, the auditory system may provide means to detect dangerous events, such as an approaching car, at an early stage and react accordingly. In such cases, the great advantage of the auditory system compared with the visual system is that it allows us to monitor all directions simultaneously, including positions behind, above and below. In fact, besides a 360-degree 'view' in terms of both elevation and azimuth, the auditory system also provides an estimate of the distance of sound sources. This capability is remarkable, given the fact that humans have only two ears and yet are capable of analyzing an auditory scene in multiple dimensions: elevation, azimuth, and distance, while recognition of a sound source might be considered as a fourth dimension.

But besides being a necessary means for communication and to provide warning signals, the human hearing system also provides a lot of excitement and fun. Listening to music is a very common activity for relaxation and entertainment. Movies rely on a dedicated sound track to be exciting and thrilling. Computer games become more lifelike with the inclusion of dedicated sound tracks and effects.

In order to enjoy music or other audio material, a sound scene has to be recorded, processed, stored, transmitted, and reproduced by dedicated equipment and algorithms. During the last decade, the field of processing, storing, and transmitting audio has shifted from the traditional analog domain to the *digital* domain, where all information, such as audio and video material, is represented by series of bits. This shift in representation method has several advantages. It provides new methods and algorithms to process audio. Furthermore, for many applications, it can provide higher quality than traditional analog systems. Moreover, the quality of the material does not degrade over time, nor does making copies have any negative influence on the quality. And finally, it allows for a more compact representation in terms of information quantity, which makes transmission and storage more efficient and cheaper, and allows devices for storage and reception to

Spatial Audio Processing: MPEG Surround and Other Applications Jeroen Breebaart and Christof Faller

be of very small form factor, such as CDs, mobile phones and portable music players (e.g. MP3 music players).

1.2 Spatial audio reproduction

One area where audio systems have recently gained the potential of delivering higher quality is their *spatial* realism. By increasing the number of audio channels from two (stereophonic reproduction) to five or six (as in many home cinema setups), the spatial properties of a sound scene can be captured and reproduced more accurately. Initially, multi-channel signal representations were almost the exclusive domain of cinemas, but the advent of DVDs and SACDs have made multi-channel audio available in living rooms as well. Interestingly, although multi-channel audio has now been widely adopted on such storage media, broadcast systems for radio and television are still predominantly operating in stereo. The fact that broadcast chains still operate in the two-channel domain has several reasons. One important aspect is that potential 'upgrades' of broadcast systems to multi-channel audio should ensure backward compatibility with existing devices that expect (and are often limited to) stereo content only. Secondly, an increase in the number of audio channels from two to five will result in an increase in the amount of information that has to be transmitted by a factor of about 2.5. In many cases, this increase is undesirable or in some cases simply unavailable. With the technology that is currently being used in broadcast environments it is very difficult to overcome these two major limitations.

But besides the home cinema, high-quality multi-channel audio has made its way to mobile applications as well. Music, movie material, and television broadcasts are received, stored, and reproduced by mobile phones or mobile audio/video players. On such devices, an upgrade from stereo to multi-channel audio faces two additional challenges on top of those mentioned above. The first is that the audio content is often reproduced over headphones, making multi-channel reproduction more cumbersome. Secondly, these devices are often operating on batteries. Decoding and reproduction of five audio channels requires more processing and hence battery power than two audio channels, which has a negative effect on a very important aspect of virtually all mobile devices: their battery life. Furthermore, especially in the field of mobile communication, every transmitted bit has a relatively high price tag and hence high efficiency of the applied compression algorithm is a must.

1.3 Spatial audio coding

Thus, the trend towards high-quality, multi-channel audio for solid-state and mobile applications imposes several challenges on audio compression algorithms. New developments in this field should aim at unsurpassed compression efficiency, backward compatibility with existing systems, have a low complexity, and preferably support additional capabilities to optimize playback on mobile devices. To meet these challenges, the field of spatial audio coding has developed rapidly during the last 5 years. Spatial audio coding (SAC), also referred to as binaural cue coding (BCC), breaks with the traditional view that the amount of information that has to be transmitted grows linearly with the

number of audio channels. Instead, spatial audio coders, or BCC coders, represent two or more audio channels by a certain *down-mix* of these audio channels, accompanied by additional information (spatial parameters or binaural cues) that describe the loss of spatial information caused by the down-mix process.

Conventional coders are based on waveform representations attempting to minimize the error induced by the lossy coding process using a certain (perceptual) error measure. Such perceptual audio coders, for example MP3, weight the error such that it is largely masked, i.e. not audible. In technical terms, it is said that 'perceptual irrelevancies' present in the audio signals are exploited to reduce the amount of information. The errors that are introduced result from *removal* of those signal components that are perceptually irrelevant.

Spatial audio coding, on the other hand, represents a multi-channel audio signal as a down-mix (which is coded with a conventional audio coder) and the before mentioned spatial parameters. For decoding, the down-mix is 'expanded' to the original number of audio channels by restoring the inter-channel cues which are relevant for the auditory system to perceive the correct auditory spatial image. Thus, instead of achieving compression gain by removal of irrelevant information, spatial audio coding employs *modeling* of perceptually *relevant* information only. As a result, the bitrate is significantly lower than that of conventional audio coders because the spatial parameters contain much less information than the (compressed) waveforms of the original audio channels. As will also be explained in this book, the representation of a multi-channel audio signal as a down-mix plus spatial parameters not only provides a significant compression gain, it also enables new functionality such as efficient binaural rendering, re-rendering of multi-channel signals on different reproduction systems, forward and backward format conversion, and may provide means for interactivity, where end-users can modify various properties of individual objects within a single audio stream.

1.4 Book outline

Briefly summarized, the contents of the chapters are as follows:

Chapter 2 provides an overview of common audio reproduction, processing, and compression techniques. This includes discussion of various loudspeaker and headphone audio playback techniques, conventional audio coding, and matrix surround.

Chapter 3 reviews the literature on important aspects of the human spatial hearing system. The focus is on the known limitations of the hearing system to perceive and detect spatial characteristics. These limitations form the fundamental basis of spatial audio coding and processing techniques.

Chapter 4 explains the basic concepts of spatial audio coding, and describes the inter-channel parameters that are extracted, the required signal decompositions, and the spatial reconstruction process.

Chapter 5 describes the structure of the MPEG 'enhanced aacPlus' codec and how spatial audio coding technology is embedded in this stereo coder.

Chapter 6 describes the structure of MPEG Surround, a multi-channel audio codec that was finalized very recently. Virtually all components of MPEG Surround are based

on spatial audio coding technology and insights. The most important concepts and processing stages of this standard is be outlined.

Chapter 7 describes the process of generating a virtual sound source (for headphone playback) by applying spatial audio coding concepts.

Chapter 8 expands the spatial audio coding approach to complex auditory scenes and describes how parameter-based virtual sound source generation processes are incorporated in the MPEG Surround standard.

Chapter 9 reviews methods to incorporate user interactivity and flexibility in terms of spatial rendering and mixing. By applying parameterization on individual objects rather than individual channels, several modifications to the auditory scene can be applied at the audio decoder side, such as re-panning, level adjustments, equalization or effects processing of individual objects present within a down-mix.

Chapter 10 describes algorithms to optimize the reproduction of stereo audio on different reproduction systems than the audio material was designed for, such as 5.1 home cinema setups, or wavefield synthesis systems.