SPATIAL AUDIO PROCESSING

MPEG Surround and Other Applications

Jeroen Breebaart and Christof Faller

WILEY

TN912
B832

Spatial Audio Processing

MPEG Surround and Other Applications

Jeroen Breebaart
Philips Research, the Netherlands

Christof Faller
EPFL, Switzerland

E2008000524

BICENTENNIAL 1807 WILEY 2007 BICENTENNIAL

John Wiley & Sons, Ltd

Copyright © 2007

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770620.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 42 McDougall Street, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Blvd, Mississauga, Ontario, L5R 4J3, Canada

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Anniversary Logo Design: Richard J. Pacifico

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN-13 978-0-470-03350-0

Typeset in 10/12 Times by Laserwords Private Limited, Chennai, India
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Spatial Audio Processing

Foreword

It is a pleasure and great honor for me to contribute a foreword to this fascinating book on the state-of-the-art in stereo and multi-channel audio processing. Given my own research interest in binaural hearing, it is exciting to follow the detailed description of how scientific insights into human spatial perception in combination with digital signal processing techniques have enabled a major step forward in audio coding. I also feel honored to have a close relationship with the two authors. Both are young scientists with already an impressive output of scientific publications and also patents, and they both have made significant contributions to international standards. The book that they present here documents their deep insights in auditory perception and their ability to translate these into real-time algorithms for the digital representation of multi-channel sounds. I was lucky to follow many of the described developments from close by.

A remarkable aspect of the authors' careers is that both are or have been related to research environments with a long history in using perception insights to steer technological developments. Christof Faller was for many years affiliated with the Bell Laboratories of Lucent Technologies and later with Agere, a spin-off of Bell Laboratories and Lucent. The history of psychoacoustics at Bell Labs started around 1914, when Harvey Fletcher initiated a research program on speech and hearing with the clear goal to improve the design of telephone systems. Since then, many important contributions to audio and speech signal processing applications have come out of this laboratory. Jeroen Breebaart got his academic training at the former Institute for Perception Research (IPO), for many years a joint research institution of Philips Research and the Technische Universiteit Eindhoven. The transfer of his psychoacoustic modeling knowledge into algorithmic applications started after he had joined the Philips Research Laboratories in Eindhoven. Hearing research has been a topic at Philips since the 1930s when the technical possibility of stereophonic sound reproduction required a deeper understanding of spatial hearing. Among the many studies done in that period by, among others, K. de Boer were early applications of a dummy head to study and improve interpersonal communication and support hard-of-hearing subjects.

The book nicely demonstrates how close coding applications have come to present-day psychoacoustic research. When perceptual audio coders were first realized in the 1980s, the bit-rate reduction mainly exploited the concept of spectral masking. This concept was included in the encoder by so-called spectral spreading functions, an approximation which had been known in basic research for at least 30 years. Although using only a very crude description of perceptual processes, these early coders became essential in enabling

试读结束，需要全本PDF请购买 www.ertongbook.com

the internet to be used for music distribution. Up to very recently, audio coding did not take much advantage of the redundancy *between* channels. The problem here is that, despite the *perceptual* similarity between different channels in a recording, the similarity in terms of interchannel correlation is often very low. Using a signal-based analysis thus does not give much room for redundancy removal. In order to capture and remove this redundancy, a time–frequency analysis and parametrization of the *perceptually relevant* spatial parameters is needed. The insights into these perceptual relations is of very recent origin, and the authors were able to apply them so quickly in audio processing because they had, in part, been involved in generating this perceptual knowledge.

Thus, the book is a remarkable document of what can be achieved by combining such complementary knowledge areas as psychoacoustics and digital signal processing. Given the economic and societal impact of audio compression, I hope that this example will help to attract future students to delve into this, certainly not simple, but highly rewarding research domain. Reading this book will certainly help readers to come to such a career choice.

Armin Kohlrausch
Research Fellow,
Philips Research, Eindhoven
Professor for Auditory and Multisensory Perception,
TU Eindhoven, the Netherlands

Eindhoven, May 2007

Preface

The physical and practical limitations that have to be considered in the field of audio engineering include limited directionality of microphones, limited number of audio channels, the need for backwards compatibility, storage and transmission channel constraints, loudspeaker positioning, and cost, which may dictate other restrictions.

There is a long history of interaction between the fields of audio engineering and psychoacoustics. In many cases, the reason for this interaction is to achieve a certain goal given the imposed limitations. For example, in the 1970s the first widely used cinema surround systems applied psychoacoustic knowledge to improve their perceptual performance. Specifically, due to technical limitations the method of representing multi-channel audio often caused a lot of crosstalk between various loudspeaker signals, resulting in a risk that the front dialogue was occasionally perceived from behind. To prevent this, delays were applied motivated by the psychoacoustic 'law of the first wavefront' (sound is often perceived from the direction from which the 'first wavefront' arrives and delayed reflections arriving from different directions are not perceived explicitly). Another example is that the spectral content of the center dialogue loudspeaker in cinemas is modified such that the dialogue is perceived from above the center loudspeaker, i.e. from the center of the screen as is desired.

A more recent example of how psychoacoustic knowledge benefits audio engineering is perceptual audio coding. Invented in the late 1980s at Bell Laboratories, perceptual audio coders reduce the precision of audio waveforms to the minimum such that the error is just not perceived. Due to the psychoacoustic phenomenon of (monaural) masking, i.e. that one sound can render other sounds inaudible, the precision of an audio signal can be reduced in a signal-adaptive manner with hardly any audible impairment. The most prominent example of such a perceptual audio coder is MP3.

Spatial audio coding and processing, the focus of this book, comprises processing of audio signals considering spatial aspects of sound in relation with the abilities and limitations of the human hearing system. As explained in this book, the amount of information required to represent stereo or multi-channel audio can be significantly reduced by considering how humans perceive the spatial aspect of sound. More generally, this book also gives many examples of audio signal processing, considering spatial hearing, for achieving desired results, such as binaural audio processing and two to N-channel audio upmix.

The authors would gratefully like to acknowledge the help, support, valuable insights, comments, suggestions and observations by the following people (in alphabetical order):

Frank Baumgarte, Frans de Bont, Bert den Brinker, Thomas Eisele, Jürgen Herre, Gerard Hotho, Armin Kohlrausch, Jeroen Koppens, Peter Kroon, Juha Merimaa, Francois Myburg, Fabian Nater, Werner Oomen, Mykola Ostrovskyy, Erik Schuijers, Michel van Loon, Leon van de Kerkhof, Steven van de Par, and Martin Vetterli. Furthermore, the authors would like to thank their colleagues from Agere Systems, Coding Technologies, Fraunhofer IIS and Philips for their support in developing and exploiting various audio coders based on the technology explained in this book.

Jeroen Breebaart

Christof Faller

May 2007

Contents

试读结束，需要全本PDF请购买 www.ertongbook.com

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
© 2007 John Wiley & Sons, Ltd

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.

2
Background

2.1 Introduction

Spatial audio processing and coding are not topics that can be treated in an isolated fashion. Spatial audio signals have properties which are related to the specific audio playback system over which the signals are intended to be reproduced. Further, specific properties also result from the microphone setup used if spatial audio signals are directly recorded. In addition to discussing spatial audio playback systems, recording is very briefly discussed. Additionally, conventional audio coding and matrix surround are reviewed.

2.2 Spatial audio playback systems

There has been an ongoing debate about the aesthetic aim of recording and reproducing sound. In recording of classical music or other events implying a natural environment, the goal of recording and reproduction can be to re-create as realistically as possible the illusion of 'being there' live. ('Being there' refers to the recreation of the sound scene at the place and time of the performance. The term 'there and then' is often used to describe the same concept, in contrast to 'here and now', which describes the sound scene at the place and time during playback.) In many other cases, such as movie sound tracks and pop music, sound is an entirely artificial creation and so is the corresponding spatial illusion, which is designed by the recording engineer. In such a case, the goal of recording and reproduction can be to create the illusion of the event 'being here', i.e. the event being in the room where playback takes place.

In any case, the requirement of a spatial audio playback system is to reproduce sound perceived as realistically as possible, either as 'being there' or 'being here'. Note that in 'being there' one would like to create the spatial impression of the concert hall 'there', whereas in 'being here' the acoustical properties of the playback room 'here' are to play a more important role. But these aesthetic issues are to be addressed by the performing artists and recording engineers, given the limits of a specific target spatial audio playback system. In the following, we describe three of the most commonly used consumer spatial audio playback systems: stereo loudspeaker playback, headphone playback, and

Spatial Audio Processing: MPEG Surround and Other Applications Jeroen Breebaart and Christof Faller
© 2007 John Wiley & Sons, Ltd

试读结束，需要全本PDF请购买 www.ertongbook.com