PARAMETRIC CODING OF SPATIAL AUDIO

THÈSE N^{O} 3062 (2004)

PRÉSENTÉE À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

Institut de systèmes de communication

SECTION DES SYSTÈMES DE COMMUNICATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Christof FALLER

ingénieur électricien diplômé EPF
de nationalité suisse et originaire d'Arbon (TG)

acceptée sur proposition du jury:

Prof. M. Vetterli, directeur de thèse
Prof. J. Blauert, rapporteur
Prof. K. Brandenburg, rapporteur
Prof. M. Hasler, rapporteur
Dr F. Rumsey, rapporteur

Lausanne, EPFL
2004

Contents

Abstract v

Zusammenfassung vii

Acknowledgments ix

Frequently Used Terms, Abbreviations, and Notation xi

1 Introduction 1
- 1.1 Thesis motivation 1
- 1.2 Coding of audio signals 2
- 1.3 Contributions and overview 5

2 Background 9
- 2.1 Introduction 9
- 2.2 Spatial hearing 9
  - 2.2.1 Auditory events and auditory spatial image 9
  - 2.2.2 Spatial hearing with one sound source 10
  - 2.2.3 Ear entrance signal properties and lateralization 11
  - 2.2.4 Two sound sources: Summing localization 13
  - 2.2.5 Superposition of signals each evoking one auditory event 15
  - 2.2.6 Source localization in the presence of reflections 17
  - 2.2.7 Other spatial attributes 17
- 2.3 Spatial audio playback systems 20
  - 2.3.1 Stereo audio loudspeaker playback 21
  - 2.3.2 Headphone audio playback 24
  - 2.3.3 Multi-channel audio playback 26
- 2.4 Conclusions 27

3 Parametric Coding of Spatial Audio Using Perceptual Cues 29
- 3.1 Introduction 29
- 3.2 Related techniques 30
  - 3.2.1 Pseudostereophonic processes 30
  - 3.2.2 Intensity stereo coding 31
- 3.3 Binaural cue coding (BCC) 31
  - 3.3.1 Time-frequency processing 31
  - 3.3.2 Downmixing to one channel 32
  - 3.3.3 "Perceptually relevant differences" between audio channels 34
  - 3.3.4 Estimation of spatial cues 35

3.3.5 Synthesis of spatial cues 37
3.3.6 Manipulation of spatial cues at the decoder 40
3.4 Coding of low frequency effects (LFE) audio channels 42
3.5 Subjective evaluations 43
3.5.1 Different time resolutions for ICLD synthesis 43
3.5.2 Stereo audio coding 48
3.6 Conclusions 50

4 Coherence Synthesis Based on Late Reverberation 53
4.1 Introduction 53
4.2 Generating independent late reverberation given one audio channel 54
4.3 Using late reverberation for coherence synthesis 54
4.3.1 Stereo coherence synthesis 55
4.3.2 Multi-channel coherence synthesis 57
4.4 Subjective evaluations 59
4.4.1 Stereo audio signals 59
4.4.2 Multi-channel audio signals 61
4.5 Conclusions 63

5 Variations of Binaural Cue Coding (BCC) 65
5.1 Introduction 65
5.2 Hybrid BCC 66
5.2.1 Encoder and decoder processing 66
5.2.2 Evaluations 68
5.3 C-to-E BCC 71
5.3.1 Encoder processing 73
5.3.2 Decoder processing 73
5.3.3 Specific examples for C-to-E BCC schemes 75
5.3.4 Subjective evaluation 81
5.4 BCC for flexible rendering 84
5.4.1 Encoder processing 86
5.4.2 Decoder processing 87
5.4.3 Applications 87
5.4.4 Subjective evaluation 89
5.5 Conclusions 90

6 Low Complexity Implementation of Binaural Cue Coding (BCC) 93
6.1 Introduction 93
6.2 STFT-based BCC 93
6.2.1 The time-frequency transform 94
6.2.2 Downmixing to one channel 98
6.2.3 Estimation of ICTD, ICLD, and ICC 99
6.2.4 Quantization and coding of ICTD, ICLD, and ICC 100
6.2.5 Synthesis of ICTD, ICLD, and ICC 101
6.2.6 Synthesis of ICC based on late reverberation 103
6.3 STFT-based BCC for flexible rendering 104
6.3.1 Estimation of dominance cues 104
6.3.2 Coding of dominance cues 104
6.3.3 Decoder processing 105
6.4 Complexity and delay of the proposed schemes 105

6.5 Conclusions 106

7 Source Localization in Complex Listening Situations 107
7.1 Introduction 107
7.2 Model description 111
7.2.1 Auditory periphery 111
7.2.2 Binaural processor 111
7.2.3 Higher model stages 113
7.2.4 Discussion 114
7.3 Simulation results 115
7.3.1 Independent sources in free-field 117
7.3.2 Precedence effect 122
7.3.3 Independent sources in reverberant environment 127
7.4 General discussion 129
7.5 Discussion related to binaural cue coding (BCC) 131
7.5.1 Cue selection and stereo audio playback 131
7.5.2 BCC and the cue selection 136
7.5.3 ICC and auditory spatial image width and stability 138
7.5.4 Discussion 139
7.6 Conclusions 139

8 Conclusion 141
8.1 Thesis summary 141
8.2 Impact 142
8.3 Future research 142

Bibliography 145

Curriculum Vitae 159

Abstract

A wide range of techniques for coding a single speech or audio signal channel have been developed over the last few decades. In addition to pure redundancy reduction, sophisticated source and receiver models have been considered for reducing the bitrate. Only a few techniques address joint-coding of the channels of stereo[1] and multi-channel audio signals.

Stereo and multi-channel audio signals evoke an auditory spatial image in a listener. Thus the receiver model may consider properties of spatial hearing of the auditory system for reducing the bitrate. This has been done in previous techniques by considering the importance of interaural level difference cues at high frequencies and by considering the binaural masking level difference for computing the masked threshold for multiple audio channels.

The coding scheme proposed in this thesis aims at being more systematic and parameterized. A stereo or multi-channel audio signal is represented as a single downmixed audio channel plus side information. The side information contains the inter-channel cues inherent in the original audio signal that are relevant for the perception of the properties of the auditory spatial image. At the decoder the stereo or multi-channel audio signal is reconstructed such that its inter-channel cues approximate the corresponding cues of the original audio signal.

This enables coding of stereo or multi-channel audio signals at a bitrate nearly as low as a mono audio coding bitrate because the side information contains about two orders of magnitude less information than the original audio channel waveforms. This corresponds to a significant bitrate reduction compared to conventional state-of-the-art coders. Several subjective tests were conducted, indicating that good audio quality can be achieved by the proposed scheme.

A number of variations of the coding scheme are proposed. These include different combinations of conventional multi-channel audio coders and the proposed coding scheme, and a scheme which provides flexibility at the decoder to manipulate the auditory spatial image.

A model for source localization in the presence of concurrent sound (other sources and reflections) is proposed. The results from a number of previous psychophysical studies are predicted successfully by the model. The model is also applied for comparing audio signals to corresponding signals coded with the proposed scheme.

---

[1] In this thesis, the term "stereo audio signal" always refers to two-channel stereo audio signals.

Zusammenfassung

In den letzten Jahrzehnten wurden eine Vielzahl von Techniken zur Kodierung von monophonen Sprach- und Audio Signal Kanälen entwickelt. Neben Redundanzreduktion wurden auch Sender- und Empfängermodelle berücksichtigt, um die Datenrate weiter zu reduzieren. Nur relativ wenige Techniken berücksichtigen jedoch die gemeinsame Kodierung der Kanäle von Stereo[2]- und Multikanalsignalen.

Stereo- und Multikanalsignale rufen eine räumliche Wahrnehmung bei dem Zuhörer hervor. Deshalb gibt es die Möglichkeit, im Empfängermodell Eigenschaften des räumlichen Hörens zu berücksichtigen um die Kodierungsbitrate zu reduzieren. Existierende Techniken erreichen dies, indem sie die wichtige Rolle der interauralen Intensitätsunterschiede und binaurale Eigenschaften der Verdeckung für die Berechnung der Maskierungsschwellen berücksichtigen.

Das Kodierungsverfahren, das in dieser Dissertation vorgestellt wird, hat das Ziel einer verbesserten Systematik und Parametrierbarkeit. Ein Stereo- oder Multikanalsignal wird als ein heruntergemischter, monophoner Audiokanal plus Seiteninformation represäntiert. Die Seiteninformation repräsentiert Interkanaleigenschaften eines gegebenen Audiosignals, die wichtig für die räumliche Wahrnehmung sind. Das so repräsentierte Audiosignal wird durch Dekodierung in ein Multikanalsignal umgeformt, das die gleichen Interkanaleigenschaften hat wie das ursprüngliche Audiosignal.

Damit wird die Kodierung von Stereo oder Multikanalsignalen mit nahezu einer Monokodierungsbitrate möglich, weil die Seiteninformation um zwei Grössenordnungen weniger Information enthält als die unkodierten Daten des Audiosignals. Das entspricht einer signifikanten Reduktion der Bitrate verglichen mit herkömmlichen Kodierungsverfahren, die dem Stand der Technik entsprechen. Verschiedene subjektive Hörtests zeigen, dass mit dem vorgestellten Kodierungsverfahren eine gute Audioqualität erreicht wird.

Verschiedene Variationen des Kodierungsverfahren werden vorgestellt. Unter anderem verschiedene Kombinationen von herkömmlichen und dem vorgestellten Kodierungsverfahren und eine Variation die es dem Dekoder erlaubt, das räumliche Klangbild zu manipulieren.

Ein Modell für Quellenlokalisation bei Vorhandensein von “Stör-”geräuschquellen (andere Quellen oder Reflektionen) wird vorgestellt. Resultate von verschiedenen publizierten psychoakustischen Studien werden mit dem Modell richtig vorhergesagt. Das Modell wird auch angewandt um Referenzsignale und deren kodierte Version zu vergleichen, die mit dem vorgestellten Kodierungsverfahren erzeugt wurden.

---

[2] Der Ausdruck “Stereosignal” wird in dieser Dissertation immer verwendet für 2-Kanal Stereosignale.

Acknowledgments

I would like to thank my thesis supervisor Martin Vetterli for letting me pursue a Ph.D. degree at EPFL, while being full time employed by Bell Laboratories (research division of Lucent Technologies) and Agere Systems (Lucent Technologies spin-off). Not only did he regularly inspire me and give me a different angle for viewing things, but he also gave me the freedom which enabled efficient contribution to both, my thesis and my company.

Many thanks go also to Peter Kroon, my supervisor at Bell Laboratories and Agere Systems, for his advise and teaching of technical writing and ethics. The outcome of this thesis would have been impossible without the nice and free environment Peter Kroon was creating for our department. Thanks to Frank Baumgarte for the numerous collaborations on projects related to this thesis and other projects related to audio coding.

I am grateful for the fruitful interactions I had with Juha Merimaa on the topic of spatial hearing, which ultimately resulted in the contribution of this thesis in this field. Also thanks to Aki Härmä for the collaboration and discussions we had at Agere Systems and later at Helsinki University. Thanks to my colleagues from EPFL, Harald Viste and Thibaut Ajdler, for the many inspiring discussions. Also thanks to the audio coding team at Fraunhofer IIS for the fruitful collaborations for the “MP3 Surround” project and the “ISO/IEC MPEG Spatial Audio Coding” project. Particularly, I would like to thank Jürgen Herre for the inspiring discussions.

Last but not least, many thanks to my family for their support.

Frequently Used Terms, Abbreviations, and Notation

Terms and abbreviations

Anechoic chamber: Type of room with almost totally sound absorbing walls, frequently used for experimentation under free-field-like conditions.

Auditory event: Perception corresponding to a single sound source. Attributes of auditory events are location and extent.

Auditory spatial image: Illusion of perception of a space with auditory events of specific extent and at specific locations.

BCC: The coding scheme proposed in this thesis is denoted binaural cue coding (BCC), because binaural cues play an important role in it.

Binaural cues: Inter-channel cues between the left and right ear entrance signals (see also ITD, ILD, and IC).

BRIR: Binaural room impulse response, modeling transduction of sound from a source to left and right ear entrances in enclosed spaces. The difference to HTRFs is that BRIRs also consider reflections.

DFT: Discrete Fourier transform.

Direct sound: Sound reaching a listener's ears or microphones through a direct (non-reflected) path.

Enclosed space: A space where sound is reflected from the walls enclosing it.

Externalization: When typical stereo music signals are played back over headphones, the extent of the auditory spatial image is limited to about the size of the head. When playing back headphone signals which realistically mimic ear entrance signals during natural listening, the extent of the auditory spatial image can be very realistic and as large in extent as any auditory spatial image in natural listening. The experience of perceiving an auditory spatial image significantly larger than the head during headphone playback is denoted externalization.

FFT: Fast implementation of the DFT, denoted fast Fourier transform (FFT).

Free-field: An open space with no physical objects from which sound is reflected.

Free-field cues: Binaural cues (ITD and ILD) which occur in a one-source-free-field listening scenario.

HRTF: Head related transfer function, modeling transduction of sound from a source to left and right ear entrances in free-field.

IC: Interaural coherence, i.e. degree of similarity between left and right ear entrance signals. This is sometimes also referred to as “IAC” or “interaural cross-correlation” (IACC).

ICC: Inter-channel coherence. Same as IC but defined more generally between any signal pair (e.g. loudspeaker signal pair, ear entrance signal pair, etc.).

ICLD: Inter-channel level difference. Same as ILD but defined more generally between any signal pair (e.g. loudspeaker signal pair, ear entrance signal pair, etc.).

ICTD: Inter-channel time difference. Same as ITD but defined more generally between any signal pair (e.g. loudspeaker signal pair, ear entrance signal pair, etc.).

ILD: Interaural level difference, i.e. level difference between left and right ear entrance signals. This is sometimes also referred to as “interaural intensity difference” (IID).

ITD: Interaural time difference, i.e. time difference between left and right ear entrance signals. This is sometimes also referred to as “interaural time delay”.

kb/s: Unit for bitrate, kilo-bit per second.

Lateral: From the side, e.g. ”lateral reflections” are reflections arriving at a listener’s ears from the sides.

Lateralization: For headphone playback, a subject’s task is usually restricted to identifying the lateral displacement of the projection of the auditory event to the straight line connecting the ear entrances. The relationship between the lateral displacement of the auditory event and attributes of the ear entrance signals is denoted lateralization.

LFE channel: Low frequency effects channel. Multi-channel surround systems often feature one or more LFE channels for low frequency sound effects requiring higher sound pressure than can be reproduced by the loudspeakers for the regular audio channels. In movie soundtracks, an LFE channel may for example contain low frequency parts of explosion sounds.

Mixing: Given a number of source signals (e.g. separately recorded instruments, multitrack recording), the process of generating stereo or multichannel audio signals intended for spatial audio playback is denoted mixing.

Mixing cues: The cues (level difference, delay difference) with which a source signal was mixed into a stereo signal.

Listener envelopment: A listener's auditory sense of "envelopment" or "spaciousness of the environment".

Localization: The relation between the location of an auditory event and a single or more attributes of a sound event. For example, localization may describe the relation between the direction of a sound source and the direction of the corresponding auditory event.

PDF: Probability density function.

Precedence effect: A number of phenomena related to the auditory system's ability to resolve the direction of a source in the presence of one or more reflections by favoring the "first wave front" over successively arriving reflections.

Reference cues: Binaural cues (ITD and ILD) which occur when playing back a one-auditory-event stereo or multi-channel audio signal.

Reflection: Sound that arrives at a listener's ears or microphones indirectly after being reflected one or more times from the surface of physical objects.

Reverberation: The persistence of sound in an enclosed space as a result of multiple reflections after the sound source has stopped.

Reverberation time: Defined as the time it takes for sound energy to decay by 60 dB. The more reverberant a room is, the larger is its reverberation time. Reverberation time is often denoted "RT" or "RT60".

Sound source: A physical object emitting sound.

STFT: Short time Fourier transform. The term STFT in this thesis is used for short time discrete Fourier transform.

Spatial audio: Audio signals which when played back through an appropriate playback system evoke an auditory spatial image.

Spatial impression: The impression a listener spontaneously gets about type, size, and other properties of an actual or simulated space.

Spatial cues: Cues relevant for spatial perception. In this thesis, this term is used for cues between pairs of channels of a stereo or multi-channel audio signal (see also ICTD, ICLD, and ICC).

Sweet spot: Optimal listening position for a stereo or multi-channel loudspeaker-based audio playback system.

Transparent: An audio signal is transparent when a listener can not distinguish between this signal and a reference signal. For example, transparent audio coding denotes audio coding, where there is no perceptible degradation in the coded audio signals.

Notation and variables

$\star$	Convolution operator;
b	Partition (subband) index;
n	Time index (discrete time);
f_s	Sampling frequency;
T	Time constant of a one-sided exponential estimation window;
m	STFT spectrum frequency index;
c	Audio channel index;
k	Time index of subband signals (also time index of STFT spectra);
C	Number of encoder input channels;
D	Number of decoder output channels (if different than number of encoder input channels C);
E	Number of transmitted channels (if different than number of encoder input channels C);
$x_c(n)$	Encoder input audio channels;
$s(n)$	Transmitted sum signal;
$y_c(n)$	Transmitted audio channels;
$\hat{x}_c(n)$	Decoder output audio channels;
$\tilde{x}_c(k)$	One subband signal of $x_c(n)$ (similarly defined for other signals);
$p_{\tilde{x}_c}(k)$	Short-time estimate of power of $\tilde{x}_c(k)$ (similarly defined for other signals);
$h_c(n)$	Late reverberation filter for channel c;
M	Length of filters $h_c(n)$;
$s_c(n)$	Late reverberation signal channels;
$\tau_{1c}(k)$	ICTD between channel 1 and c;
$\Delta L_{1c}(k)$	ICLD between channel 1 and c;
$c_{c_1 c_2}(k)$	ICC between channel c_1 and c_2;
$\tau(k)$	ITD in specific critical band;
$\Delta L(k)$	ILD in specific critical band;
$c(k)$	IC in specific critical band;
$X_{c,m}(k)$	STFT spectrum of a signal $x_c(n)$ (similarly defined for other signals);

Chapter 1

Introduction

1.1 Thesis motivation

Generally speaking, audio coding is a process for changing the representation of an audio signal to make it more suitable for transmission or storage. Although high capacity channels, networks, and storage systems have become more easily accessible, audio coding has retained its importance. Motivations for reducing the bitrate necessary for representing audio signals are the need to minimize transmission costs or to provide cost-efficient storage, the demand to transmit over channels with limited capacity such as mobile radio channels, and to support variable-rate coding in packet oriented networks.

The audio coding techniques proposed in this thesis enable higher compression ratios for stereo[1] and multi-channel audio coding than what is achieved by previous state-of-the-art techniques. The starting point of this work were experiments testing an "extreme case" scenario, i.e. just downmixing all audio channels of a stereo or multi-channel audio signal to one single channel and trying to reconstruct an audio signal which would sound similar to the original audio signal.

The problem of reconstructing the discrete audio channels of a multi-channel audio signal after downmixing to a single channel seems to be impossible without retaining a substantial amount of information about the original audio channels. Techniques such as blind source separation are not suitable for tackling this problem directly, since usually the channels of a multi-channel audio signal can not be considered to be independent.

But it turned out that it is possible to reconstruct high quality audio signals resembling the original audio signals before downmixing, given the downmixed channel and a small amount of side information. This is possible by considering properties of the human auditory system related to spatial hearing. The side information represents only audio channel properties which are relevant cues for spatial perception. Intuition gained through the work on the proposed coding scheme also lead to the proposal for an auditory model for source localization.

Before further discussing the detailed scope of this thesis, Section 1.2 gives an overview of commonly used techniques for audio coding and describes the

---

[1] In this thesis, the term "stereo audio signal" always refers to two-channel stereo audio signals.

current state-of-the-art in the field. The contributions of this thesis are briefly described in Section 1.3, which also serves as thesis overview with pointers to the chapters of the thesis.

1.2 Coding of audio signals

In this section, various techniques for the coding of audio signals are reviewed. This includes techniques for coding a single audio channel and techniques for coding stereo and multi-channel audio signals.

Audio signals are usually available as discrete time sampled signals. For example, a *compact disc* (CD) stores stereo audio signals as two separate audio channels each sampled with a sampling frequency of 44.1 kHz. Each sample is represented as a 16-bit signed integer value, resulting in a bitrate of $2 \times 44.1 \times 16 = 1411$ kb/s. For multi-channel audio signals or signals sampled at a higher sampling frequency the bitrate scales accordingly with the number of channels and sampling frequency.

Lossless audio coding Ideally, an audio coder reduces the bitrate without degrading the quality of the audio signals. So-called lossless audio coders (Cellier *et al.* 1993, Robinson 1994, Bruekers *et al.* 1996, Purat *et al.* 1997, Gerzon *et al.* 1999) can achieve this by reducing the bitrate while being able to perfectly reconstruct the original audio signal. Bitrate reduction in this case is possible by exploring redundancy present in audio signals, e.g. by applying prediction over time or a time-frequency transform, and controlling the coding process such that during decoding the values are rounded to the same integer sample values as the original samples. For typical CD audio signals, lossless audio coders reduce the bitrate by about a factor of 2. For higher sampling rates the higher degree of redundancy inherent in the corresponding samples result in higher compression ratios.

Perceptual audio coding Most audio coders are “lossy” coders not aiming at reconstructing an audio signal perfectly. The primary motivation for using lossy audio coders is to achieve a higher compression ratio. A *perceptual audio coder* is an audio coder which incorporates a *receiver model*, i.e. it considers the properties of the human auditory system. Usually, a model for computing the *masked threshold* (Zwicker and Fastl 1999) is considered. The masked threshold specifies as a function of time and frequency a signal level below which a signal component is not perceptible, i.e. the maximum level a signal component can have such that it is masked by the audio signal to be coded. By controlling the quantization error as a function of time and frequency such that it is below the masked threshold, the bitrate can be reduced without degrading the perceived quality of the audio signal.

A generic perceptual audio coder, operating in a subband (or transform) domain, is shown in Figure 1.1. The input signal is decomposed into a number of subbands. A perceptual model computes the masked threshold as a function of time and frequency. Each subband signal is quantized and coded. The quantization error of each subband signal is controlled such that it is below the computed masked threshold. This coding principle was introduced for speech coding by Zelinski and Noll (1977). Brandenburg *et al.* (1982) applied this technique to wideband audio signals.

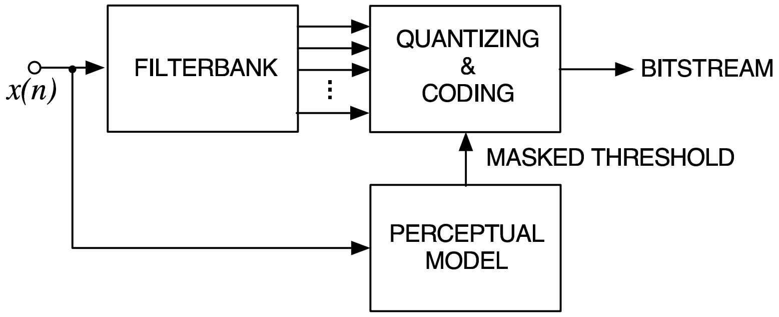


Figure 1.1: Generic perceptual audio encoder. The input signal is decomposed into a number of subbands. The subband signals are quantized and coded. A perceptual model controls the quantization error such that it is just below the masked threshold and thus not perceptible.

When more than one audio channel need to be encoded, redundancy between the audio channels may be explored for reducing the bitrate. *Mid/side* (M/S) coding (Johnston and Ferreira 1992) reduces the redundancy between a correlated channel pair by transforming it to a sum/difference channel pair prior to quantization and coding. The masked threshold depends on the interaural signal properties of the signal (masker) and quantization noise (maskee). When coding stereo or multi-channel audio signals, the perceptual model needs to consider this interaural dependence of the masked threshold. This dependence is often described by means of the *binaural masking level difference* (BMLD) (Blauert 1997).

Today many proprietary perceptual audio coders (Fielder *et al.* 1996, Tsutsui *et al.* 1996, Sinha *et al.* 1997) and international standards-based perceptual audio coders (Brandenburg and Stoll 1994, ISO/IEC 1993, Stoll 1996, Bosi *et al.* 1997, Grill 1999, ISO/IEC 1997) are commercially available. When requiring that the coded audio signal can not be distinguished from the original audio signal, state-of-the-art perceptual audio coders are able to reduce the bitrate of CD audio signals by a factor of about 10. When higher compression ratios are required, the audio bandwidth needs to be reduced or coding distortions will exceed the masked threshold.

Parametric audio coding Perceptual audio coders with coding mechanisms as described above perform only suboptimally when high compression ratios need to be achieved. *Parametric audio coders* are usually based on a *source model*, i.e. they make specific assumptions about the signals to be coded and can thus achieve higher compression ratios. Speech coders are a classical example for parametric coding and can achieve very high compression ratios. However, they perform sub-optimally for signals which can not be modeled effectively with the assumed source model. In the following, a number of parametric coding techniques applicable to audio and speech coding are described.

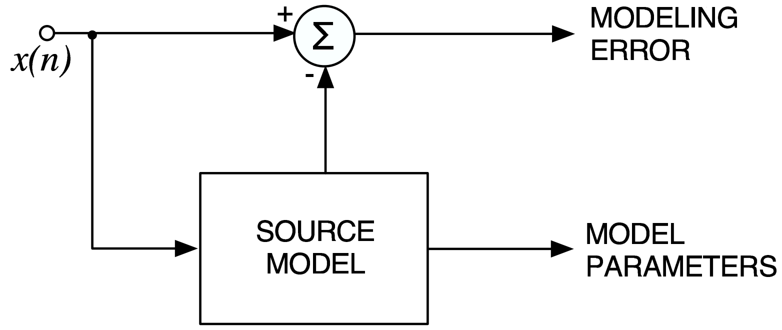


Figure 1.2: A generic parametric encoder. Parameters of a source model are estimated adaptively in time for modeling the input signal. The model parameters and often also the modeling error are transmitted to the decoder.

A parametric encoder is illustrated in Figure 1.2. The parameters of the source model are estimated and transmitted to the decoder. The modeling error is often also transmitted to the decoder.

Linear predictive coding (LPC) is a standard technique often used in speech coders (Atal and Hanauer 1971, Kleijn and Paliwal 1995). In LPC, the source model is an all-pole filter. It has been shown that such an all-pole filter is a good model for the human vocal tract. LPC has also been applied to wide-band audio signals, see e.g. (Singhal 1990, Boland and Deriche 1995, Härmä *et al.* 1997, Härmä and Laine 1999).

Another commonly used parametric coding technique is based on *sinusoidal modeling*, where the source model consists of a number of sinusoids. The parameters to be estimated here are amplitude, frequency, and phase of the sinusoids. Sinusoidal modeling has been applied to speech coding (Hedelin 1981, McAulay and Quatieri 1986) and audio coding (Smith and Serra 1987).

For audio coding, a parametric coder has been proposed which decomposes an audio signal into sinusoids, harmonic components, and noise (Edler *et al.* 1996), denoted *harmonic and individual lines plus noise* (HILN) coder (Purnhagen and Meine 2000, ISO/IEC 1999). Each of these three “objects” is represented with a suitable set of parameters. A similar approach, but decomposing an audio signal into sinusoids, noise, and transients, was proposed by den Brinker *et al.* (2002).

Parametric coders may not only incorporate a source model but also a receiver model. Properties of the human auditory system are often considered for determining the choice or coding precision of the modeling parameters. Frequency selectivity properties of the human auditory system were explored for LPC-based audio coding by Härmä *et al.* (1997). A masking model is incorporated into the object-based parametric coder of Edler *et al.* (1996) and only signal components which are not masked by the remaining parts of the signal are parameterized and coded. A later version of this coder additionally incorporates a loudness model (Purnhagen *et al.* 2002) for giving precedence to signal components with highest loudness.

Most parametric audio and speech coders are only applicable to single-channel audio signals. Recently, the parametric coder proposed by den Brinker *et al.* (2002) was extended for coding of stereo signals (Schuijers *et al.* 2002, Schuijers *et al.* 2003) with techniques very similar to some of the techniques proposed in this thesis (Faller and Baumgarte 2001, Faller and Baumgarte 2002*a*, Faller and Baumgarte 2002*c*).

Combining perceptual and parametric audio coding Usually, parametric audio coding is applied for very low bitrates, e.g. $4 - 32$ kb/s for mono or stereo audio signals, and non-parametric audio coding for higher bitrates up to transparent coding, e.g. $24 - 256$ kb/s for mono or stereo audio signals. The perceptual audio coder as described above is based on the principle of "hiding" quantization noise below the masked threshold. It is by design aimed at transparent coding. For medium bitrates, when the amount of quantization noise has to exceed the masked threshold for achieving the desired bitrate, this coding principle may not be optimal. A number of techniques have been proposed for extending perceptual audio coders with parametric coding techniques for improved quality at such medium bitrates.

One such parametric technique applicable for stereo or multi-channel audio coding is *intensity stereo coding* (ISC) (Herre *et al.* 1994). ISC is a parametric technique for coding of audio channel pairs. It is related to some of the techniques proposed in this thesis and is further discussed in Section 3.2.2. Another parametric technique that has been applied within the framework of perceptual audio coders is denoted *spectral bandwidth replication* (SBR) (Dietz *et al.* 2002). SBR applies perceptual audio coding at frequencies below a certain frequency (e.g. $< 4 - 12$ kHz) and parametric coding above that frequency. Very little information is used to parameterize the upper spectrum and the decoder uses information from the lower spectrum to generate a perceptually meaningful upper spectrum. Ideas related to SBR were already introduced by Makhoul and Berouti (1979) in the context of speech coding.

1.3 Contributions and overview

In terms of coding of the audio waveforms many methods have been tried and optimized. While current state-of-the-art perceptual audio coders incorporate sophisticated auditory models for computation of the masked threshold, only simplistic processing is used to explore phenomena related to spatial hearing. The aim of this thesis is to make further substantial improvements by considering spatial hearing and spatial aspects of audio signals.

Chapter 2 reviews spatial hearing phenomena and explains how commonly used spatial audio playback systems explore these phenomena for reproducing spatial sound. In **Chapters 3 and 4** the proposed coding scheme for stereo and multi-channel audio signals, denoted *binaural cue coding* (BCC) (Faller and Baumgarte 2001, Faller and Baumgarte 2002*a*, Faller and Baumgarte 2002*c*, Faller and Baumgarte 2003, Faller 2003), is described. BCC achieves significant bitrate reduction by transmitting only one single audio channel. This single audio channel contains all signal components (disregarding spatial aspects) which are present in the original stereo or multi-channel audio signal. In the simplest case, this single channel is just the sum of all channels of the audio signal to be coded. In addition, parameters describing "perceptually relevant differences" (in terms of spatial hearing) between the original audio channels are estimated. These parameters contain about two orders of magnitude less information than the waveforms themselves and thus the bitrate is significantly reduced by transmitting them as opposed to transmitting all the audio channels. In the decoder, the transmitted audio channel is processed such that the "perceptually relevant differences" of the synthesized channels approximate those of the original audio channels.

The considered "perceptually relevant differences" between the audio channels are time difference, level difference, and coherence as a function of time and frequency. Synthesis of these cues, given the waveform of the transmitted single audio channel, is discussed. Furthermore, the role of specific music production methods (e.g. coincident pair microphones, amplitude panning) and playback setups (headphone playback, stereo and multi-channel loudspeaker playback) with respect to these "perceptually relevant differences" between channels is discussed.

Subjective tests were carried out for assessing the performance of BCC for audio coding applications. It is shown that BCC, combined with a conventional mono audio coder (for coding of the single transmitted channel), outperforms existing state-of-the-art audio coders for a wide range of bitrates. Furthermore, the upper quality bound of BCC was assessed with subjective tests carried out when the transmitted single audio channel is not coded. Subjective tests indicate that BCC achieves "excellent" audio quality for critical headphone listening.

Chapter 5 describes a number of variations of BCC, i.e. different ways for combining BCC with conventional audio coders (Baumgarte *et al.* 2004). One such scheme offers scalable audio quality up to transparent audio quality. The other scheme represents multi-channel audio signals with more than one audio channel, enabling backwards compatible coding between different stereo and multi-channel surround formats. One application of this scheme is extension of existing stereo coders or systems for multi-channel surround (Herre *et al.* 2004). Another variation of BCC offers flexibility at the decoder for determining the attributes of the auditory spatial image of its output audio signal (Faller and Baumgarte 2001, Faller and Baumgarte 2002*b*).

In **Chapter 6** it is described how BCC schemes can be implemented with low computational complexity using a *short time Fourier transform* (STFT) based filterbank. Furthermore, it is described how to limit, quantize, and code the spatial cues, and computational complexity and algorithmic delay of the algorithms are discussed.

Informal experiments with BCC and coherence synthesis implied that *interaural coherence* (IC) may play an important role not only for attributes related to diffuseness, apparent source width, and envelopment, but that IC is also related to the stability of the auditory spatial image. For example, when two concurrently active independent source signals are amplitude panned to the left and right sides, a listener perceives two clearly localized auditory events in most cases. It turned out that in such cases where no diffuseness is perceived, it is necessary for BCC to synthesize coherence cues in order to maintain auditory image stability and full auditory image width.

This intuition gave rise to motivation for a model for source localization in complex listening scenarios. **Chapter 7** describes this model (Faller and

Merimaa 2004), which is based on processing for modeling the periphery of the auditory system, followed by a mechanism using IC cues to identify time instants when *interaural time difference* (ITD) and *interaural level difference* (ILD) cues correspond to source directions. At other time instants, ITD and ILD cues are ignored. The model is shown to be able to explain various results from the literature related to source localization in the presence of distracting sound and related to the precedence effect. Also, the model was applied for predicting and comparing auditory event localization for stereo signals and corresponding BCC synthesized signals.

Chapter 2

Background

2.1 Introduction

The focus of this thesis is on coding of spatial audio. In other words, coding of audio signals which were generated with the aim to evoke a certain auditory spatial image in a listener when presented over headphones or two or more loudspeakers. Such audio signals have properties related to the intended playback setup (e.g. number and positions of loudspeakers) and the auditory spatial image that is to be perceived by a listener during playback of such signals. The aim of this chapter is to give the reader a brief overview of perceptual phenomena related to spatial hearing and to discuss how these phenomena relate to commonly used spatial audio playback systems. The material presented in this chapter forms the basic knowledge necessary for understanding the context and motivation of the following chapters. A thorough introduction into the field of spatial hearing has been given by Blauert (1997). Introductions into spatial audio and related perceptual phenomena have been given by Streicher and Everest (1998) and Rumsey (2001).

2.2 Spatial hearing

2.2.1 Auditory events and auditory spatial image

Similarly to the way humans perceive a visual image, humans are also able to perceive an *auditory spatial image*. The different objects which are part of the auditory spatial image are denoted *auditory events*. For example, if a listener listens to a musical performance, the auditory events are the different instruments which are playing. In most listening situations, the perceived directions of auditory events correspond well to the directions of the physical *sound sources* emitting the sounds that are associated with the corresponding auditory events. This is a necessity in order that the perceived auditory spatial image corresponds to the physical surroundings of a listener.

In the most commonly used consumer stereo and multi-channel audio playback systems, all loudspeakers are usually placed in the horizontal plane (see also Section 2.3). Therefore, in the following only spatial hearing phenomena for this case are reviewed. The discussion is also limited to phenomena which are relevant to this thesis.

2.2.2 Spatial hearing with one sound source

The simplest listening scenario is when there is one sound source in *free-field*. Free-field denotes an open space with no physical objects from which sound is reflected. *Anechoic chambers* are rooms frequently used for experimentation under free-field-like conditions. Due to their highly sound absorbent walls there are virtually no reflections, similarly as in free-field. *Localization* denotes the relation between the location of an auditory event and one or more attributes of a *sound event*. A sound event denotes sound sources and their corresponding signals responsible for the perception of the auditory event. For example, localization may describe the relation between the direction of a sound source and the direction of the corresponding auditory event. *Localization blur* denotes the smallest change in one or more attributes of a sound event such that a change in location of the auditory event is perceived. For sources in the *horizontal plane*, localization blur with respect to direction is smallest for sources in the forward direction of a listener. It is slightly larger for sources behind the listener and largest for sources to the sides. In other words, the precision with which a listener can discriminate the direction of a source in the horizontal plane is best if the source is in the front and worst if a source is on the side.

In order to understand how the auditory system discriminates the direction of a source, the properties of the signals at the ear entrances have to be considered, i.e. the signals available to the auditory system. Most generally, the ear input signals can be viewed as being filtered versions of the source signal. The filters modeling the path of sound from a source to the left and right ear entrances are commonly referred to as *head related transfer functions* (HRTFs). Figure 2.1(a) illustrates the left and right HRTFs, h_1 and h_2, for a source at angle ϕ. For each source direction different HRTFs need to be used for modeling the ear entrance signals. A more intuitive but only approximately valid view for the relation between the source angle ϕ and the ear entrance signals is illustrated in Figure 2.1(b). The difference in length of the paths to the ear entrances, $d_2 - d_1$, can be expressed as a function of the source angle ϕ. As a result of the different path lengths, there is a difference in arrival time between both ear entrances. The most simple formula for the difference in path length between the left and right ear is the "sine law" for spatial hearing as proposed by von Hornbostel and Wertheimer (1920),

$$\Delta d = \kappa \sin\phi \text{ with } \kappa = 21 \text{ cm}\,, \tag{2.1}$$

where κ is the distance of the two microphones modeling the two ear entrances. Since the effect of the curved path around the head is ignored, κ is chosen larger than the actual distance between the ears. Another limitation of the sine law is that *head shadowing* is ignored, i.e. the effect of the head on the intensities of the ear entrance signals is not considered. Several improved formulas where introduced to account for the curved path of sound around the head. An overview of different path length difference formulas has been given by Blauert (1997).

As a result of the path length difference from the source to the two ear entrances, there is a difference in arrival times of sound at the left and right

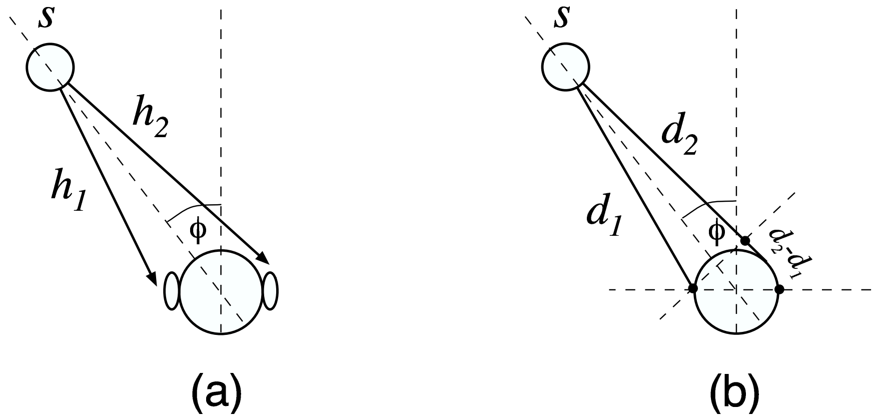


Figure 2.1: (a): Paths of a source to the ear entrances modeled with HRTFs. (b): A more intuitive view relates the source angle to a path distance difference resulting in an arrival time difference at the ear entrances. Additionally, head shadowing results in an intensity difference between the ear entrance signals as a function of the source angle.

ears, denoted *interaural time difference* (ITD). Additionally, the shadowing of the head results in an intensity difference of the left and right ear entrance signals, denoted *interaural level difference* (ILD). For example, a source to the left of a listener results in a higher intensity of the signal at the left ear than at the right ear.

Diffraction, reflection, and resonance effects caused by the head, torso, and the external ears of the listener result in that ITD and ILD not only depend on the source angle ϕ but also on the source signal. Nevertheless, if ITD and ILD are considered as a function of frequency, it is a reasonable approximation to say that the source angle solely determines ITD and ILD as implied by data shown by Gaik (1993). When only considering frontal directions ($-90° \leq \phi \leq 90°$) the source angle ϕ approximately causally determines ITD and ILD. However, for each frontal direction there is a corresponding direction in the back of the listener resulting in a similar ITD-ILD pair. Thus, the auditory system needs to rely on other cues for resolving this front/back ambiguity. Examples of such cues are head movement cues, visual cues, and spectral cues (different frequencies are emphasized or attenuated when a source is in the front or back) (Blauert 1997). The following discussion does not cover these other cues, since these are not considered explicitly in the coding scheme proposed in this thesis. For audio playback systems with loudspeakers these other cues are automatically inherent in the ear entrance signals due to the physical location of the loudspeakers.

2.2.3 Ear entrance signal properties and lateralization

The previous discussion implies that ITD and ILD are ear entrance signal properties which provide to the auditory system information about the direction of a sound source. A specific ITD-ILD pair can be associated with the source direction (when disregarding the front/back ambiguity). With headphones, the ear entrance signals are (ideally) equal to the signals given to the left and right transducers of the headphones. Therefore, it is possible to evaluate the effect of ITD and ILD independently of each other with experiments carried out with headphones. Figure 2.2 shows an experimental setup for generating coherent left and right ear entrance signals, $e_1(n)$ and $e_2(n)$, given a single audio signal $s(n)$. ITD is determined by the delays and equal to $d_2 - d_1$ and ILD is determined by the scale factors a_1 and a_2, and expressed in dB is $20\log_{10}(a_2/a_1)$.

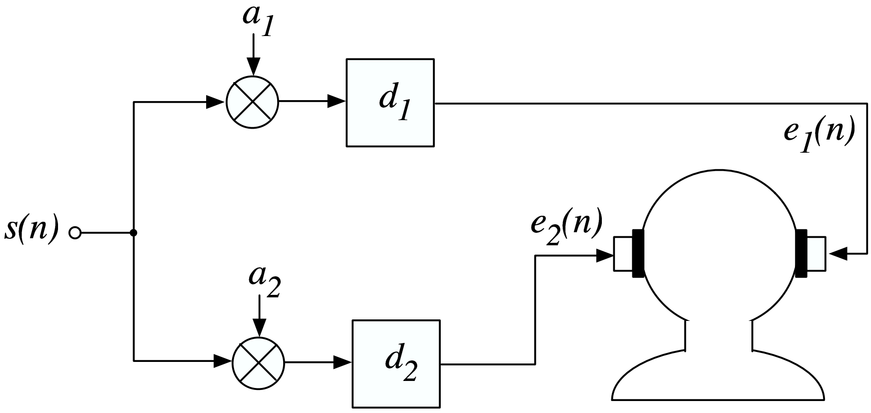


Figure 2.2: Experimental setup for generating coherent left and right ear entrance signals with specific ITD and ILD.

Figure 2.3(a) illustrates perceived auditory events for different ITD and ILD (Blauert 1997) for two coherent left and right headphone signals as generated by the scheme shown in Figure 2.2. When left and right headphone signals are coherent, have the same level (ILD $= 0$), and no delay difference (ITD $= 0$), an auditory event appears in the center between the left and right ears of a listener. More specifically, the auditory event appears in the center of the frontal section of the upper half of the head of a listener, as illustrated by Region 1 in Figure 2.3(a). By increasing the level on one side, e.g. right, the auditory event moves to that side as illustrated by Region 2 in Figure 2.3(a). In the extreme case, when only the signal on the left is active, the auditory event appears at the left side as illustrated by Region 3 in Figure 2.3(a). ITD can be used similarly to control the position of the auditory event. For headphone playback, a subject's task is usually restricted to identifying the lateral displacement of the projection of the auditory event to the straight line connecting the ear entrances. The relationship between the lateral displacement of the auditory event and attributes of the ear entrance signals is denoted *lateralization*.

So far, only the case of coherent left and right ear input signals were considered. Another ear entrance signal property that is considered in this discussion is a measure for the degree of "similarity" between the left and right ear entrance signals, denoted *interaural coherence* (IC). IC here is defined as the

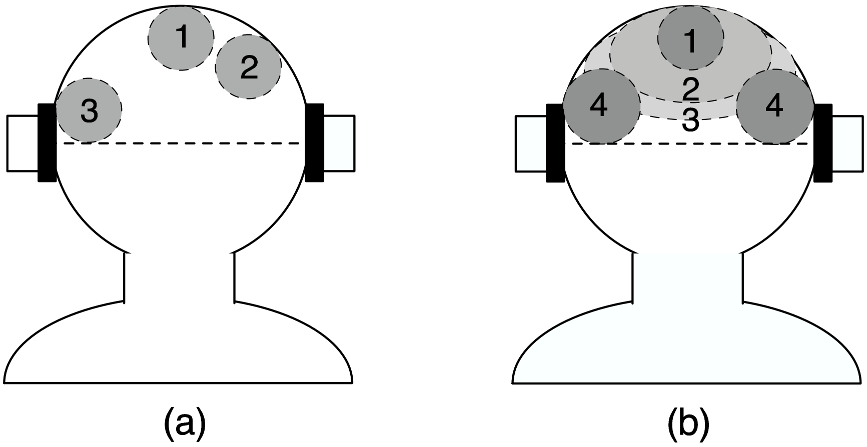


Figure 2.3: (a): ILD and ITD between a pair of headphone signals determine the location of the auditory event which appears in the frontal section of the upper head. (b): The width of the auditory event increases (1-3) as the interaural coherence (IC) between the left and right headphone signals decreases, until two distinct auditory events appear at the sides (4).

maximum absolute value of the normalized cross-correlation function,

$$\mathrm{IC} = \max_{d} \frac{|\sum_{n=-\infty}^{\infty} e_1(n) e_2(n+d)|}{\sqrt{e_1^2(n) e_2^2(n+d)}} \,, \tag{2.2}$$

where delays d corresponding to a range of ±1 ms are considered. IC as defined has a range between zero and one. IC = 1 means that two signals are coherent (signals are equal with possibly a different scaling and delay) and IC = 0 means that the signals are independent. IC may also be defined as the signed value of the normalized cross-correlation function with the largest magnitude, resulting in a range of values between minus one and one. A value of minus one then means that the signals are identical but with a different sign (phase inverted).

When two identical signals (IC = 1) are emitted by the two transducers of the headphones, a relatively compact auditory event is perceived. For noise the width of the auditory event increases as the IC between the headphone signals decreases until two distinct auditory events are perceived at the sides, as illustrated in Figure 2.3(b) (Chernyak and Dubrovsky 1968).

Conclusively, one can say that it is possible to control the lateralization of an auditory event by choosing ITD and ILD. Furthermore, the width of the auditory event is related to IC.

2.2.4 Two sound sources: Summing localization

Spatial hearing with two sound sources has a high practical relevance because stereo loudspeaker listening depends on perceptual phenomena related to two sound sources (the two sources are the two loudspeakers in this case). Also for multi-channel loudspeaker listening the two source case is relevant since it is based on similar phenomena.

Previously, the ear entrance signal properties ITD and ILD were related to source angle. Then the perceptual effect of ITD, ILD, and IC cues was discussed. For two sources at a distance (e.g. loudspeaker pair), ITD, ILD, and IC are determined by the HRTFs of both sources and by the specific source signals. Nevertheless, it is interesting to assess the effect of cues similar to ITD, ILD, and IC, but relative to the source signals and not ear entrance signals. To distinguish between these same properties considered either between the two ear entrance signals or two source signals, respectively, the latter are denoted *inter-channel time difference* (ICTD), *inter-channel level difference* (ICLD), and *inter-channel coherence* (ICC). For headphone playback, ITD, ILD, and IC are (ideally) the same as ICTD, ICLD, and ICC. In the following a few phenomena related to ICTD, ICLD, and ICC are reviewed for two sources located in the front of a listener.

Figure 2.4(a) illustrates the location of the perceived auditory events for different ICLD for two coherent source signals (Blauert 1997). When left and right source signals are coherent (ICC = 1), have the same level (ICLD = 0), and no delay difference (ICTD = 0), an auditory event appears in the center between the two sources as illustrated by Region 1 in Figure 2.4(a). By increasing the level on one side, e.g. right, the auditory event moves to that side as illustrated by Region 2 in Figure 2.4(a). In the extreme case, when only the signal on the left is active, the auditory event appears at the left source position as is illustrated by Region 3 in Figure 2.4(b). ICTD can be used similarly to control the position of the auditory event. This principle of controlling the location of an auditory event between a source pair is also applicable when the source pair is not in the front of the listener. However, some restrictions apply for sources to the sides of a listener (Theile and Plenge 1977, Pulkki 2001*b*). There is an upper limit for the angle between such a source pair beyond which localization of auditory events between the sources degrades.

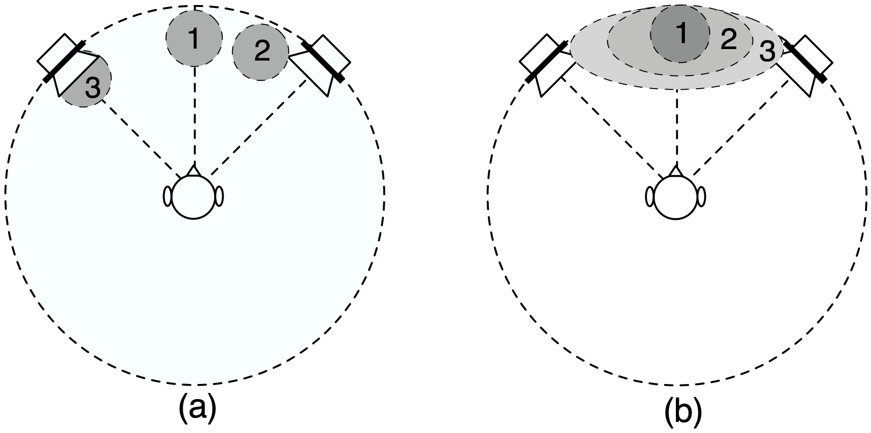


Figure 2.4: (a): ICTD and ICLD between a pair of coherent source signals determine the location of the auditory event which appears between the two sources. (b): The width of the auditory event increases (1-3) as the IC between left and right source signals decreases.

When coherent wideband noise signals (ICC = 1) are simultaneously emitted by a pair of sources, a relatively compact auditory event is perceived. When the ICC is reduced between these signals, the width of the auditory event increases (Blauert 1997), as illustrated in Figure 2.4(b). The phenomenon of summing localization, e.g. an auditory event appearing between a pair of frontal loudspeakers, is based on the fact that ITD and ILD cues evoked at the ears crudely approximate the dominating cues that would appear if a physical source were located at the direction of the auditory event. The mutual role of ITDs and ILDs is often characterized with time-intensity trading ratios (Blauert 1997) or in the form of the classic duplex theory (Rayleigh 1907): ITD cues dominate localization at low frequencies and ILD cues at high frequencies.

The insight that when signals with specific properties are emitted by two sources the direction of the auditory event can be controlled is of high relevance for applications. It is this property, which makes stereo audio playback possible. With two appropriately placed loudspeakers, the illusion of auditory events at any direction between the two loudspeakers can be generated. (More on this topic is discussed in Section 2.3).

Another relevance of the described phenomena is that for loudspeaker playback and headphone playback similar cues can be used for controlling the location of an auditory event. This is the basis, which makes it possible to generate signal pairs which evoke related illusions in terms of relative auditory event location for both loudspeaker and headphone playback. If this were not the case, there would be a need for different signals depending on whether a listener uses loudspeakers or headphones.

2.2.5 Superposition of signals each evoking one auditory event

Above, only phenomena were described where a single auditory event appears. However, an auditory spatial image often consists of a number of auditory events distributed in (perceived) space. Consider the case of two *independent* sound sources. By independent sources we mean sources which generate signals independently, e.g. different speech sources or instruments. Over a short period of time (e.g. 20 ms) such signals may be correlated but when observed over a sufficiently long period of time (e.g. a few seconds) such signals are statistically reasonably independent. When one source emits a signal, a corresponding auditory event is perceived at the direction of that source (left scenario in Figure 2.5). When another independent source at a different location emits a signal, another corresponding auditory event is perceived from the direction of that second source (middle scenario in Figure 2.5). When both sources emit their independent signals simultaneously, usually two auditory events are perceived from the two directions of the two sources. From the linearity property of HRTFs it follows that in the latter scenario the ear entrance signals are equal to the sum of the two ear entrance signals of the two one-active-source scenarios (as indicated in the right scenario in Figure 2.5). More generally speaking, the sum of a number of ear entrance signals associated with independent sources usually results in distinct auditory events for each of the sources. In many cases, the directions of the resulting auditory events correspond to the directions of the sources.

This perceptual feature is not only a remarkable ability of the auditory system, but also a necessity for formation of auditory spatial images corresponding

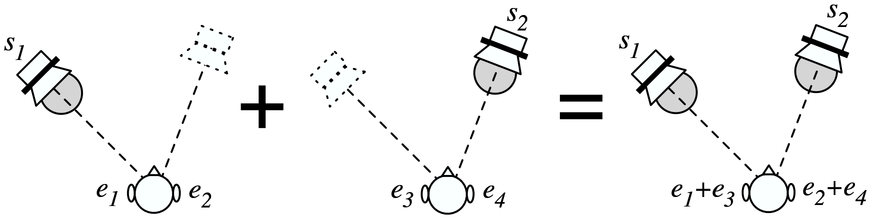
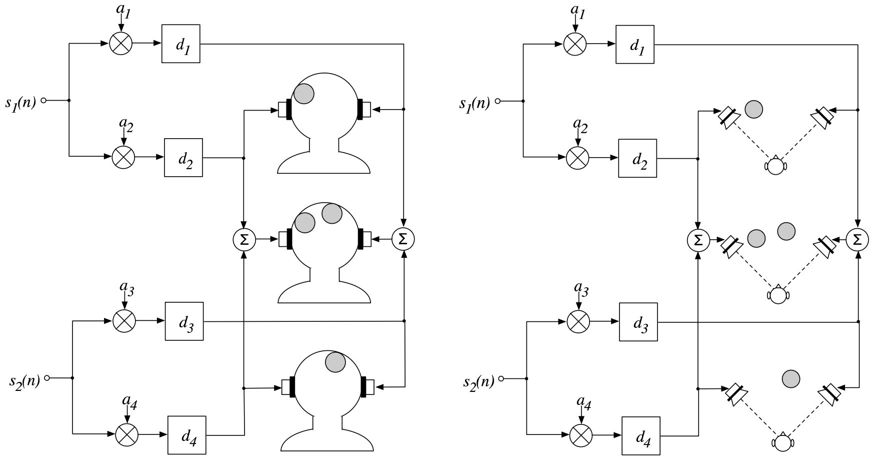


Figure 2.5: When two independent sources are concurrently active (right), usually two auditory events appear at the same directions where auditory events appear for each source emitting a signal individually (left and middle).

to the physical surroundings of a listener. In Chapter 7, a model is proposed aiming at explaining this phenomenon, i.e. how the auditory system is able to localize sources given the complex mix of signals at the ear entrances in multi-source scenarios.



(a) (b)

Figure 2.6: By adding two different pairs of signals corresponding to single auditory events, the composite signal pair usually results in two auditory events at the same locations as the auditory events corresponding to each signal pair for (a) headphone playback and (b) loudspeaker playback.

The described principle for the superposition of ear entrance signals does also hold for signal pairs generated as illustrated in Figure 2.2. For example, if a number of such signal pairs are generated for different independent base signals, then the composite signal pair usually results in perceived auditory events at the locations of the auditory events appearing when the signal pairs are played back separately. From the linearity property of HRTFs this follows not only for headphone playback, but also for two sources emitting these signals. This property is illustrated for headphones and sources in Figure 2.6(a) and (b), respectively. *Mixing* techniques for generating stereo signals given a multitrack recording (separately recorded instruments) are based on the described principle.

2.2.6 Source localization in the presence of reflections

Resolving mixes of multiple concurrently active sources in free-field is not the only challenge the auditory system is facing. In the described free-field scenarios, sound reaches the ears from the sources through only one direct path, modeled by HRTFs for each source. Sound reaching the ears directly from a source is denoted *direct sound*. If a listener is surrounded by physical objects or if a listener is in a room, sound not only reaches the ears directly from each source, but also indirectly as reflections from different directions than the direct sound. Despite of this, in most cases auditory events are perceived at the directions of the sources. Therefore, the auditory system must have the ability of “suppressing” localization associated with reflections.

Usually, the direct sound of a source reaches the ears earlier than the reflections of the same sound because the indirect path associated with reflections is longer than the direct path from the source to the ears. The *precedence effect* describes a number of phenomena related to the auditory system’s ability to resolve the direction of a source in the presence of one or more reflections by favoring the “first wave front” over successively arriving reflections. That is, the directional perception of reflections arriving within a few milliseconds after the direct sound is suppressed and the direct sound and these reflections are “fused” into one single auditory event at the direction of the direct sound. Extensive reviews of the precedence effect have been given by Zurek (1987), Blauert (1997), and Litovsky *et al.* (1999).

HRTFs model the path from a source to the ear entrances in free-field. *Binaural room impulse responses* (BRIRs) are linear filters modeling the multiple paths sound travels through before reaching the ear entrances as direct sound and reflections.

The auditory model proposed in Chapter 7 not only addresses source localization in the case of concurrently active sources, but also source localization in the presence of reflections.

2.2.7 Other spatial attributes

So far the discussion mostly focused on the attribute of perceived direction or lateralization of auditory events. One exception was the discussion of the role IC and ICC play for noise signals in determining the width of the auditory event. In the following, other attributes related to auditory events and the auditory spatial image are briefly discussed. These attributes mostly depend on the properties of reflections relative to the direct sound.

Room reflections contain important information for perception of attributes of auditory events other than only direction. The persistence of sound in an enclosed space as a result of multiple reflections after the sound source has stopped is denoted *reverberation*. A commonly used measure for the degree of reverberation in a room is the *reverberation time* as proposed by Wallace Clement Sabine in the 1890’s (Sabine 1922). It is defined as the time it takes for sound energy to decay by 60 dB. The more reverberant a room is, the larger is its reverberation time.

Spatial impression is defined as the impression a listener spontaneously gets about type, size, and other properties of an actual or simulated space. Note that spatial impression is an attribute of auditory spatial images which is not exclusively associated with auditory event properties. It also includes properties more associated with the auditory spatial image as a whole than with the single auditory events. Spatial impression is largely determined by the relation between direct sounds and reflections, and number, strength, and directions of reflections. In the following, attributes related to spatial impression are briefly reviewed. More complete reviews have been given by Blauert (1997) and Mason (2002).

Coloration

The first early reflections up to about 20 ms later than the direct sound can cause timbral colorization due to a "comb filter" effect which attenuates and amplifies frequency components in a frequency-periodic pattern.

Distance of auditory event

In free-field, the following two ear entrance signal attributes change as a function of source distance: Power of signal reaching the ears and high frequency content (air absorption). Additionally, for sources close to the head a source distance change causes a change in ILD across all frequencies (Brungart and Rabinowitz 1999). There is evidence that the overall level of sound reaching the ears provides potent distance information. For a source for which a listener knows its likely level of emitted sound, such as speech, the overall sound level at the ear entrances provides an absolute distance cue (Mershon *et al.* 1975, Mershon and Bowers 1979). However, in situations when a listener does not expect a source to have a certain emitting level, overall sound level at the ear entrances can not be used for judging absolute distance. In such a situation, overall level does only provide a relative cue (Coleman 1962).

On the other hand, in a reverberant environment there is more information available to the auditory system. The reverberation time and the timing of the first reflections contain information about the size of a space and the distance to the surfaces, thus giving an indication about the expected range of source distances. Thus it is not surprising that many researchers have argued that for relatively distant sources the ratio of the power of direct to reflected sound is a reliable distance cue, see e.g. (Mershon *et al.* 1975, Mershon and Bowers 1979, Bronkhorst and Houtgast 1999).

Distance cues and their importance for generating artificial auditory spatial images have been discussed by Shinn-Cunningham (2000). It is argued that for real-world listening conditions and headphone playback, it is mostly important to consider level and reverberation cues.

Width of auditory events and envelopment

Barron and Marshall (1981) found that lateral reflections from ±90° cause the greatest spatial impression. The closer the direction of the reflections is to the median plane the less is the resulting spatial impression. The spatial impression caused by a pair of early reflections in the range of reflection delays between 8 ms and 90 ms is approximately constant. Based on this, Barron and Marshall proposed a physical measure called *lateral fraction* (LF) for measuring spatial impression. The lateral fraction is the ratio of the lateral sound energy to the total sound energy that arrived within the first 80 ms after the arrival of the direct sound. The lateral fraction measure is mostly associated with the width of the auditory event. More recent studies found that lateral reflections from ±90° are not optimal for creating greatest spatial impression at all frequencies (Ando and Kurihara 1986, Singh *et al.* 1994, Okano *et al.* 1998).

More than 80 ms after the arrival of the direct sound, reflections tend to contribute more to the perception of the environment than to the auditory event itself. This is manifested in a sense of "envelopment" or "spaciousness of the environment", frequently denoted *listener envelopment* (Morimoto and Maekawa 1989). Such a situation occurs for example in a concert hall, where late reverberation arrives at the listener's ears from all directions.

Bradley and Soulodre (1995) extended the research of Barron and Marshall by adding more early reflections and a (late) "reverberation tail". From a number of experiments they concluded that a similar measure as the lateral fraction for early reflections may also be applicable to reverberation. This measure does relate more to listener envelopment than width of auditory events. They termed this measure *late lateral energy fraction* ($\mathrm{LG}_{80}^{\infty}$).

The previously described measures, lateral fraction and late lateral energy fraction, relate properties of rooms (early and late reflections) to the perceptual phenomena of width of auditory events and listener envelopment. Another class of physical measures relates properties of the signals at the ear entrances to such attributes. In the following a few such measures are reviewed.

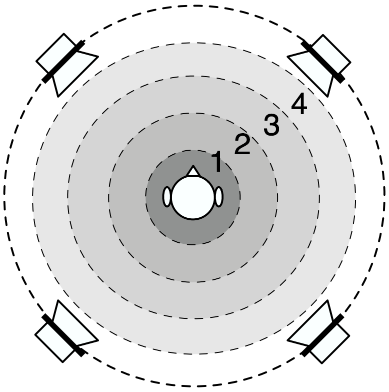


Figure 2.7: For multi-loudspeaker playback the auditory event surrounding the listener increases in width as the ICC between the signals decreases.

As implied by the results presented in Sections 2.2.3 and 2.2.4, IC and ICC are related to the width of auditory events. An artificial experience related to listener envelopment can be evoked by emitting independent noise signals with the same level from loudspeakers distributed all around a listener, as illustrated in Figure 2.7. When the ICC between source signal pairs is increased, the width of the auditory event surrounding the listener decreases (Damaske 1967/68). IC and ICC are in many cases directly related, i.e. lower ICC between a loudspeaker pair results in lower IC between the ear entrance signals (Kurozumi and Ohgushi 1983). Thus both, IC and ICC, seem to be related to auditory event width and listener envelopment. Similarly to the case of relating room properties to auditory event width and envelopment, IC can be related to these two properties by computing it relative to the early and late part of BRIRs. These two measures are often denoted early and late *interaural cross-correlation coefficient*, IACC(E) and IACC(L), respectively (Bradley 1994, Okano *et al.* 1998). A thorough review of IACC and related measures has been given by Mason (2002).

Despite of the fact that IC (or IACC) and measures like lateral fraction and lateral energy fraction seem very different, they are often similarly influenced by lateral reflections (Kuttruff 1991). A time-based division, by considering early and late reflections (e.g. up to 80 ms and later reflections) for measuring auditory event width and envelopment, is not always suitable since both influence both to a certain degree.

The relation between the previously described physical measures and spatial impression evoked by sound reproduced over a spatial audio playback system has been discussed by Rumsey (2002) in the context of spatial quality evaluation.

2.3 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. 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 multi-channel surround loudspeaker playback. A relation to Section 2.2 is established by linking spatial hearing phenomena to the described playback systems. A more thorough overview, covering history and a wide range of playback systems not discussed here can be found in Streicher and Everest (1998) and (Rumsey 2001).

2.3.1 Stereo audio loudspeaker playback

The most commonly used consumer playback system for spatial audio is the stereo loudspeaker setup as shown in Figure 2.8(a). Two loudspeakers are placed in front on the left and right sides of the listener. Usually, these loudspeakers are placed on a circle at angles −30° and 30°. The width of the auditory spatial image that is perceived when listening to such a stereo playback system is limited approximately to the area between and behind the two loudspeakers.

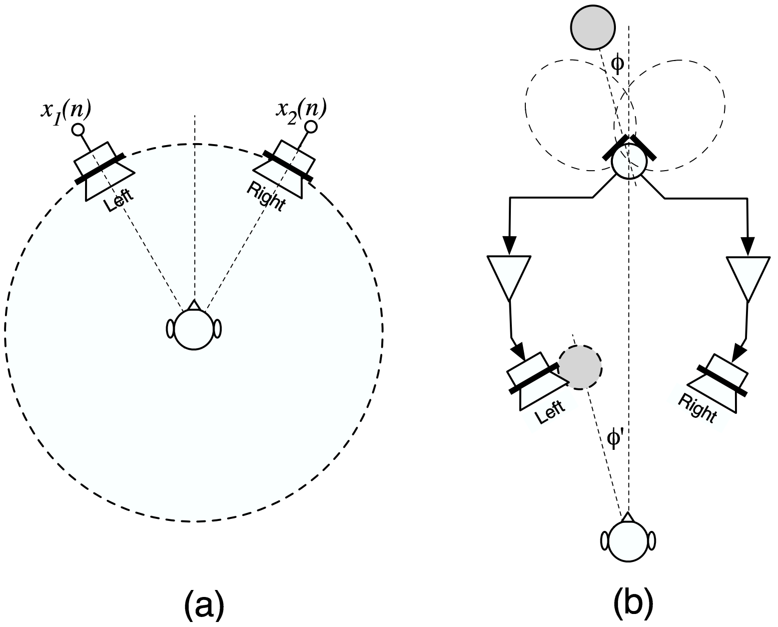


Figure 2.8: (a): Standard stereo loudspeaker setup. (b): Coincident-pair microphone pickup and playback of the resulting stereo signal.

Stereo loudspeaker playback depends on the perceptual phenomenon of summing localization (Section 2.2.4), i.e. an auditory event can be made to appear anywhere between a loudspeaker pair in front of a listener by controlling ICTD and/or ICLD. It was Blumlein (1931) who recognized the power of this principle and filed his now-famous patent. Blumlein showed that when only introducing amplitude differences (ICLD) between a loudspeaker pair, it would be possible to create phase differences between the ears (ITD) similar to those occurring in natural listening. He proposed a number of methods for pickup of sound, leading to the now common technique of *coincident-pair microphones*. However, Blumlein's work remained unimplemented in commercial products for dozens of years. It was not until the late 1950's when stereo vinyl discs became available. These applied methods for cutting two channel stereo signals onto a single disc similar to a technique already proposed in Blumlein's patent.

Capturing natural spatial sound

Figure 2.8(b) illustrates sound pickup and playback with coincident-pair microphones. Two directional microphones at the same location are orientated such that one microphone is headed more to the left and the other more to the right. Since ideally both microphones are at the same location, there is no phase difference (ICTD) between their signals. But due to their directionality there is an intensity difference (ICLD). For example, sources located on the left side result in a stronger signal in the microphone heading towards the left side than in the microphone heading towards the right side. In other words, ICLD between the two microphone signals is a function of the source angle ϕ. When these microphone signals are amplified and played back over a loudspeaker pair, an auditory event will appear at an angle ϕ' which is related to the original source angle ϕ, as illustrated in Figure 2.8(b). If the recording system parameters are properly chosen one can achieve $\phi \approx \phi'$. When there are multiple concurrently active sources to be recorded (e.g. musical instruments playing together) the mentioned recording and playback principle is also applicable and usually results in multiple auditory events, one for each instrument. This is due to the principles described in Section 2.2.5.

Coincident-pair microphones are a commonly used technique for stereo sound pickup. But there are a number of other popular microphone techniques. As mentioned, coincident-pair microphones ideally result in a signal pair without phase differences (ICTD $= 0$). This has the advantage that the resulting signal pair is "mono compatible", i.e. when the signal pair is summed to a single mono signal, no problems will occur due to a comb-filter effect (cancellation and amplification of signal components which are out-of-phase and in-phase, respectively).

Early spatial audio playback experiments based on "spaced microphone configurations" were carried out at Bell Laboratories (Steinberg and Snow 1934). In spaced microphone configurations the different microphones are located at different locations. Therefore, such techniques will result not only in ICLD cues but also in ICTD cues. When the goal is to retain mono compatibility special care has to be taken when mixing the so-obtained microphone signals to the final stereo mix. It goes beyond the scope of this overview to describe such other microphone techniques in more detail. More on this topic can be found in (Streicher and Everest 1998) and (Rumsey 2001).

Artificial generation of spatial sound

As mentioned in Section 2.2.5 and illustrated in Figure 2.6(b), artificial auditory spatial images for stereo loudspeaker playback systems can be generated by mixing a number of separately available source signals (e.g. multitrack recording). In practice, mostly ICLD are used for mixing of sources in this way, denoted *amplitude panning.* The perceived direction of an auditory event appearing when amplitude panning is applied follows approximately the stereophonic law of sines derived by Bauer (1961a),

$$\frac{\sin\phi}{\sin\phi_0} = \frac{a_1 - a_2}{a_1 + a_2}\,, \tag{2.3}$$

where $0^\circ \leq \phi_0 \leq 90^\circ$ is the angle between the forward axis and the two loudspeakers, ϕ is the corresponding angle of the auditory event, and a_1 and

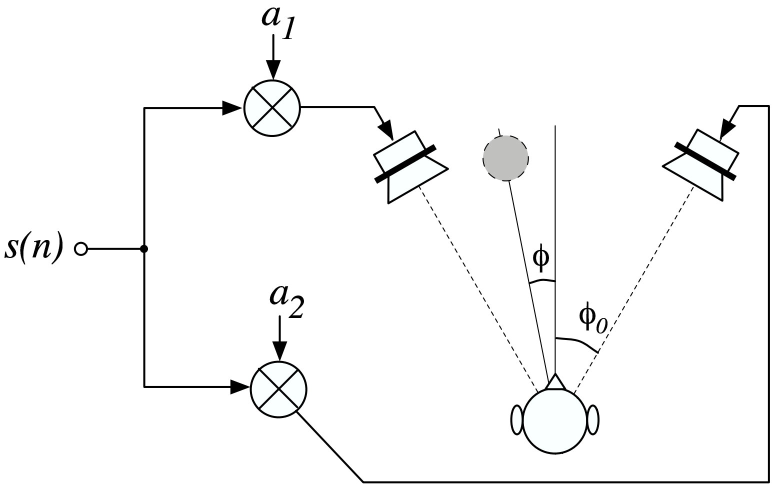


Figure 2.9: Definitions of scale factors and angles for the stereophonic law of sines (2.3).

a_2 are scale factors determining ICLD. Angles and scale factors are illustrated in Figure 2.9. The relation between ICLD, i.e. $20\log_{10}(a_2/a_1)$, and ϕ is shown in Figure 2.10 for a standard stereo listening setup with $\phi_0 = 30°$.

Bennett *et al.* (1985) derived a panning law considering an improved head model compared to the stereophonic law of sines. The result was a "stereophonic law of tangents" which is similar to another earlier proposed law by Bernfeld (1973) but for different listening conditions. Amplitude panning and auditory event direction perception is discussed in more detail in (Pulkki 2001*a*). Note that all the mentioned panning laws are only a crude approximation since the perceived auditory event direction ϕ also depends on signal properties such as frequency and signal bandwidth.

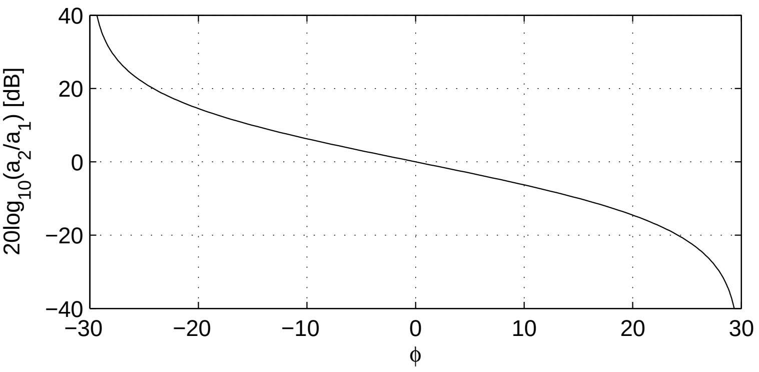


Figure 2.10: The relation between auditory event angle ϕ and ICLD, i.e. $20\log_{10}(a_2/a_1)$, for the stereophonic law of sines.

Implementation of "ICTD panning" in analog mixing consoles would have been much more difficult than implementing amplitude panning. This was surely one reason why ICTD panning was hardly used. But even today, when implementation of ICTD panning would be simple in the digital domain, ICTD panning is not commonly used. This may be due to the fact that ICLD are somewhat more robust than ICTD when a listener is not exactly in the *sweet spot* (optimal listening position). ICLD may be perceived as being more robust because amplitude panning with large-magnitude ICLD results in auditory events at the loudspeaker locations by means of only giving signal to one loudspeaker. In such a case, an auditory event is perceived at the loudspeaker location even in cases when the listener is not in the sweet spot. This is one reason why in movie theaters usually a center loudspeaker is used, i.e. to have auditory events associated with dialogue at the center of the screen for all movie viewers. For pure ICTD panning signal is always given with the same level to more than one loudspeaker and a situation with stable auditory events (i.e. only signal from one loudspeaker) does not occur.

In addition to panning, artificial reverberation may be added to the stereo signal for mimicking the spatial impression of a certain room or hall. Other signal modifications may be carried out for controlling other attributes of auditory events and the auditory spatial image.

2.3.2 Headphone audio playback

Headphone stereo audio playback

Stereo audio signals are mostly produced in an optimized way for loudspeaker playback as described in the previous section. This is reflected by the fact that during the production process the signals are usually monitored with loudspeakers by the recording engineer. As mentioned in Section 2.2.4, the signal cues (ICTD, ICLD, ICC) result in relatively similar phenomena with respect to localization and lateralization of auditory events when a signal is presented over loudspeakers or headphones, respectively. Thus, one single stereo signal can be used for either loudspeaker or headphone playback. A major difference is that headphone listening with such stereo signals is limited to in-head localization, i.e. the width of the auditory spatial image is limited to being inside of the head of the listener (as described in Section 2.2.3).

Headphone binaural audio playback

For regular audio playback (headphones, loudspeakers) the resulting ITD and ILD cues only crudely approximate the cues evoked by sources that are physically placed at the auditory event positions. Furthermore, cues related to other attributes of the auditory events and auditory spatial image are also not entirely realistic but determined largely by the recording engineer as a function of microphone setup parameters, mixing techniques, and sound effects processing.

Binaural audio playback aims at presenting a listener with the same signals at the ear entrances as the listener would receive if he were at the original event to be reproduced. Thus, all the signal cues related to perception of the sound are realistic, enabling a three dimensional sound experience. Note that in this case the width of the auditory spatial image is not limited to inside the head

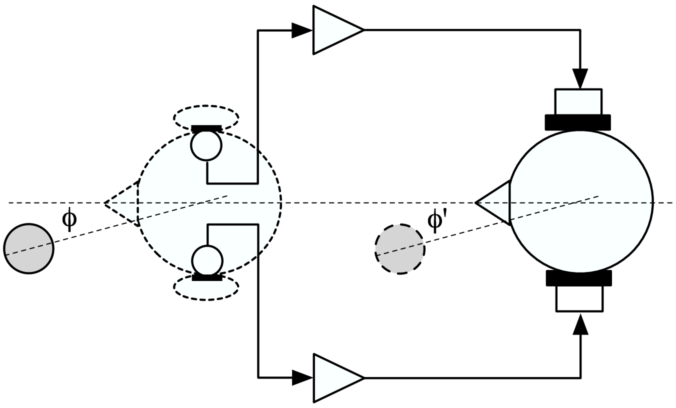


Figure 2.11: A binaural recording is a two-channel audio signal recorded with microphones close to the ear entrances of a listener or dummy head (left). When these signals are played back with headphones (right), a realistic three dimensional auditory spatial image is reproduced mimicking the natural spatial image that occurred during recording.

(this is often called *externalization*).

Figure 2.11 illustrates a system for binaural recording and binaural headphone playback. During a performance, two microphones are placed at or near the ear entrances of a listener or a dummy head and the respective signals are recorded. If these signals are played back with binaural headphones, a listener will experience an auditory spatial image very similar to the image he would perceive if he would be present at the original performance. If the binaural recording and playback are carried out with a single person, the auditory spatial image experienced during playback is very realistic. However, if a different person (or dummy head) is chosen for the binaural recording the sound experience of the listener is often limited (front/back confusion, limited externalization). Front/back confusion can be avoided by modifying the signals as a function of a listener's head movements (Boerger *et al.* 1977).

The dependence on the listener is one reason why binaural recordings are commercially hardly used. Another reason is that binaural recordings do not sound very good when played over a stereo loudspeaker setup. Several approaches have been proposed for playing back binaural recordings over two loudspeakers. Crosstalk cancellation techniques (Bauer 1961*b*, Atal and Schroeder 1966, Huopaniemi 1999) pre-process the loudspeaker signals such that the signals at the ear entrances approximate the binaural recording signals. Disadvantages of this approach are that it only works effectively when a listener's head is located exactly in the sweet spot, and its performance at higher frequencies is limited. A technique where a binaural recording is post-processed with a filter with the goal of being comparable in quality to conventional stereo recordings for loudspeaker playback was proposed in (Theile 1981*b*, Theile 1981*a*, Theile 1981*c*). This can be viewed as post-processing binaural recordings such as to mimic the properties of a good stereo microphone configuration. The idea is to store the so-obtained signals as conventional stereo signals. These signal would play back in good quality on standard stereo loudspeaker setups and when a device is intended for binaural playback it would incorporate a filter which would undo the post-processing that was applied prior to storage of the signal. The result would be a signal similar to the original binaural recording.

Binaural recordings can also be created by artificially mixing a number of audio signals. Each source signal is filtered with HRTFs or BRIRs corresponding to the desired location of its corresponding auditory event. The resulting signal pairs are added resulting in one signal pair.

2.3.3 Multi-channel audio playback

Five-to-one (5.1) surround

Not until in recent years have multi-channel loudspeaker playback systems become widely used in the consumer domain. Such systems are mostly installed as "home theater systems" for playing back audio for movies. This is partially due to the popularity of the digital versatile disc (DVD), which usually stores five or six discrete audio channels designated for such home theater system audio playback. Figure 2.12 illustrates the standard loudspeaker setup for such a system (Rec. ITU-R BS.775 1993), denoted *5.1* surround. For backwards compatibility to stereo in terms of loudspeaker positions, in the front, two loudspeakers are located at angles $-30°$ and $30°$. Additionally, there is a center loudspeaker at $0°$, providing a more stable center of the auditory spatial image when listeners are not exactly in the sweet spot. The two rear loudspeakers, located at $-110°$ and $110°$, are intended to provide the important lateral signal components related to spatial impression. There is one additional channel (the ".1" in 5.1) intended for *low frequency effects* (LFE). The LFE channel has only a bandwidth up to 120 Hz and is for effects for which the other loudspeakers can not provide enough low frequency sound pressure, e.g. explosion sounds in movies.

The angle between the two rear loudspeakers is so large ($140°$) that amplitude panning is problematic. Similarly, it is problematic to apply amplitude panning between the front and rear loudspeakers. Thus, the standard 5.1 system is not optimized for providing a $360°$ general auditory spatial image, but for providing a solid frontal auditory spatial image with lateral sound from the sides for spatial impression. In terms of $360°$ rendering, it is never problematic to place an auditory event at the location of a loudspeaker. In this sense, also the 5.1 system provides some good possibilities for auditory events appearing either at rear left or rear right. With only five main loudspeakers, a system has to be a compromise, as reflected by the 5.1 system.

Capturing and generating sound for 5.1 systems

Different microphone configurations and mixing techniques have been proposed for generating sound for 5.1 systems, see e.g. (Rumsey 2001). Alternatively, techniques applied for recording or mixing two channel stereo can be applied to a specific channel pair of the five main loudspeakers of a 5.1 setup. For example, for obtaining an auditory event at a specific direction, the loudspeaker pair enclosing the desired direction is selected and the corresponding signals are recorded or generated similarly as for the stereo case (resulting in auditory

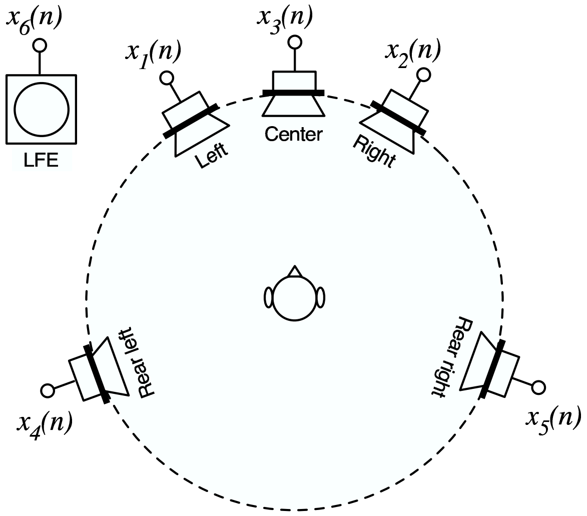


Figure 2.12: Standard 5.1 surround loudspeaker setup with a low frequency effects (LFE) channel.

events between the two selected loudspeakers). Vector base amplitude panning (VBAP) (Pulkki 1997, Pulkki 2001*b*), when applied to two dimensional loudspeaker setups such as 5.1, applies the mentioned principle in terms of amplitude panning (ICLD). But also other techniques have been proposed, feeding signals to more than two loudspeakers simultaneously, e.g. three loudspeakers (Gerzon 1990).

2.4 Conclusions

Aspects of spatial hearing and spatial audio systems were briefly reviewed in this chapter. ITD, ILD, ICTD, and ICLD are primarily related to perceived directions of auditory events. IC and ICC primarily play an important role for spatial impression and need to be properly considered if signals are to be synthesized mimicking a desired spatial impression. The auditory spatial image stereo and multi-channel audio signals evoke depends also on the specific properties of the playback system, e.g. on the number and positioning of loudspeakers. Therefore, spatial audio playback systems were also reviewed and the discussion focused on spatial hearing phenomena that are relevant for audio playback systems.

Chapter 3

Parametric Coding of Spatial Audio Using Perceptual Cues

3.1 Introduction

As mentioned in Chapter 1, we are aiming to design an audio coding scheme for stereo and multi-channel audio signals, which only transmits one single audio channel plus additional parameters describing "perceptually relevant differences" between the audio channels.

Figure 3.1 shows the proposed coding scheme, denoted *binaural cue coding* (BCC) because binaural cues play an important role in it. As indicated in the figure, the input audio channels $x_c(n)$ $(1 \leq c \leq C)$ are downmixed to one single audio channel $s(n)$, denoted *sum signal*. As "perceptually relevant differences" between the audio channels, inter-channel time difference (ICTD), inter-channel level difference (ICLD), and inter-channel coherence (ICC), are estimated as a function of frequency and time and transmitted as *side information* to the decoder. The decoder generates its output channels $\hat{x}_c(n)$ $(1 \leq c \leq C)$ such that ICTD, ICLD, and ICC between the channels approximate those of the original audio signal.

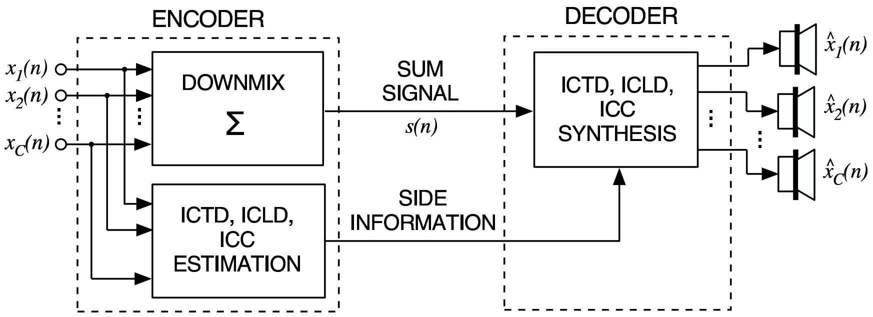


Figure 3.1: Generic scheme for binaural cue coding (BCC).

The described scheme is able to represent multi-channel audio signals at a bitrate only slightly higher than what is required to represent a mono audio signal. This is so, because the estimated ICTD, ICLD, and ICC between a channel pair contain about two orders of magnitude less information than an audio waveform.

Not only the low bitrate but also the backwards compatibility aspect is of interest. The transmitted sum signal corresponds to a mono downmix of the stereo or multi-channel signal. For receivers that do not support stereo or multi-channel sound reproduction, listening to the transmitted sum signal is thus a valid method of presenting the audio material on low-profile mono reproduction setups. BCC can therefore also be used to enhance existing services involving the delivery of mono audio material towards multi-channel audio. For example, existing mono audio radio broadcasting systems can be enhanced for stereo or multi-channel playback if the BCC side information can be embedded into the existing transmission channel.

Section 3.2 reviews previously proposed related techniques. BCC is motivated and explained in detail in Section 3.3. This includes a discussion of how ICTD, ICLD, and ICC relate to properties of auditory events and the auditory spatial image. Multi-channel surround systems often support one or more discrete audio channels for low frequency effects, denoted LFE channel (for more details see Section 2.3.3). Section 3.4 describes how to apply BCC for efficient coding of LFE channels. The results of a number of subjective evaluations are presented in Section 3.5. Conclusions are drawn in Section 3.6.

3.2 Related techniques

3.2.1 Pseudostereophonic processes

BCC relies on a synthesis technique which can generate stereo and multi-channel signals given a mono signal. There is a long history in techniques attempting to “enhance” mono signals to create a spatial impression, i.e. to generate a signal pair or more channels evoking some kind of spatial impression. Such techniques are often called “pseudostereophonic” processes. Janovsky (1948) proposed a scheme where a lowpass filtered version of the mono signal is given to one loudspeaker and a highpass filtered version to the other loudspeaker. Another technique uses complementary comb filters for generating left and right signals (Lauridsen 1954). Schroeder (1961) proposed the use of allpass filters instead of comb filters resulting in a stereo effect with less coloration artifacts. The use of a reverberation chamber with one loudspeaker emitting the mono signal and two microphones generating left and right signals was described by Schroeder (1958) and Lochner and de V. Keet (1960). Another scheme gives the mono signal to both loudspeakers and adds an attenuated and delayed version of the mono signal to one loudspeaker and the same phase inverted attenuated and delayed signal to the other loudspeaker (Lauridsen 1954, Lauridsen and Schlegel 1956). Enkl (1958) proposed the use of time-variant controllable filters controlled by properties of the mono signal. A more thorough review on these pseudosterephonic processes is given in (Blauert 1997).

In all the described techniques, the spatial distribution of the auditory events is independent of where the sound was originally picked up. The fundamental difference between these early techniques for “enhancing” mono signals and the technique applied in BCC is that not an arbitrary auditory spatial image is to be created, but an auditory spatial image similar to the auditory spatial image of the original audio signal. For this purpose information about the auditory spatial image must be available. The before-mentioned “perceptually relevant differences” between the audio channels represent this information.

3.2.2 Intensity stereo coding

Intensity stereo coding (ISC) is a joint-channel coding technique that is part of the ISO/IEC MPEG family of standards (Brandenburg and Stoll 1994, Stoll 1996, Bosi *et al.* 1997, ISO/IEC 1993, ISO/IEC 1997). ISC is applied to reduce “perceptually irrelevant information” of audio channel pairs and originated from (Waal and Veldhuis 1991). In each coder subband that uses ISC, the subband signals are replaced by a sum signal and a direction angle (azimuth). The azimuth controls the intensity stereo position of the auditory event created at the decoder. Only one azimuth is transmitted for a scalefactor band (1024 coder bands are divided into roughly 50 scalefactor bands that are spaced proportionally to auditory critical bands). ISC is capable of significantly reducing the bitrate for stereo and multi-channel audio where it is used for channel pairs. However, its application is limited since intolerable distortions can occur if ISC is used for the full bandwidth or for audio signals with a highly dynamic and wide spatial image (Herre *et al.* 1994). Potential improvements of ISC are constrained since the time-frequency resolution is given by the core audio coder and cannot be modified without adding considerable complexity.

Some of the limitations of ISC are overcome by BCC by using different filterbanks for coding of the audio waveform and for parametric coding of spatial cues (Baumgarte and Faller 2002*c*). Most audio coders use a *modified discrete cosine transform* (MDCT) (Malvar 1992) for coding of audio waveforms. The advantages of using a different filterbank for parametric stereo are reduced aliasing (Baumgarte and Faller 2002*c*) and more flexibility, such as the ability to efficiently synthesize not only intensities (ICLD), but also time delays (ICTD) and coherence (ICC) between the audio channels. Another notable conceptual difference between ISC and BCC is that the latter operates on the full band audio signal and transmits only a mono time domain signal, whereas ISC transmits not a true mono fullband signal, but only spectral portions of a mono signal.

3.3 Binaural cue coding (BCC)

3.3.1 Time-frequency processing

BCC processes audio signal with a certain time and frequency resolution. The frequency resolution used is largely motivated by the frequency resolution of the auditory system. Psychoacoustics suggests that spatial perception is most likely based on a critical band representation of the acoustic input signal (Blauert 1997). This frequency resolution is considered by using an invertible filterbank with subbands with bandwidths equal or proportional to the critical bandwidth of the auditory system (Zwicker and Fastl 1999, Glasberg and Moore 1990).

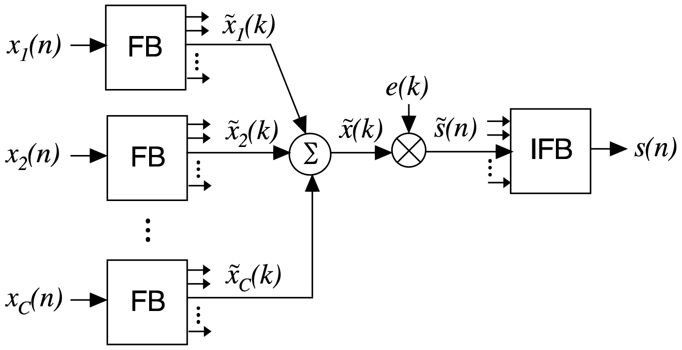


Figure 3.2: The sum signal is generated by adding the input channels in a subband domain and multiplying the sum with a factor in order to preserve signal power. FB denotes filterbank and IFB inverse filterbank. The shown processing is applied independently to each subband.

The specific time and frequency resolution used for BCC is discussed later in Section 3.3.3.

3.3.2 Downmixing to one channel

It is important that the transmitted sum signal contains all signal components of the input audio signal. The goal is that each signal component is fully maintained. Simple summation of the audio input channels often results in amplification or attenuation of signal components. In other words, the power of signal components in the "simple" sum is often larger or smaller than the sum of the power of the corresponding signal component of each channel. Therefore, a downmixing technique is used which *equalizes* the sum signal such that the power of signal components in the sum signal is approximately the same as the corresponding power in all input channels.

Figure 3.2 shows the proposed downmixing scheme. The input audio channels $x_c(n)$ $(1 \leq c \leq C)$ are decomposed into a number of subbands. One such subband is denoted $\tilde{x}_c(k)$ (note that for notational simplicity no subband index is used). Since similar processing is independently applied to all subbands it is enough to describe the processing carried out for one single subband. A different time index k is used since usually the subband signals are downsampled.

The signals of each subband of each input channel are added and then multiplied with a factor $e(k)$,

$$\tilde{s}(k) = e(k) \sum_{c=1}^{C} \tilde{x}_c(k) \,. \tag{3.1}$$

The factor $e(k)$ is computed such that

$$\sum_{c=1}^{C} p_{\tilde{x}_c}(k) = e^2(k) p_{\tilde{x}}(k) \,, \tag{3.2}$$

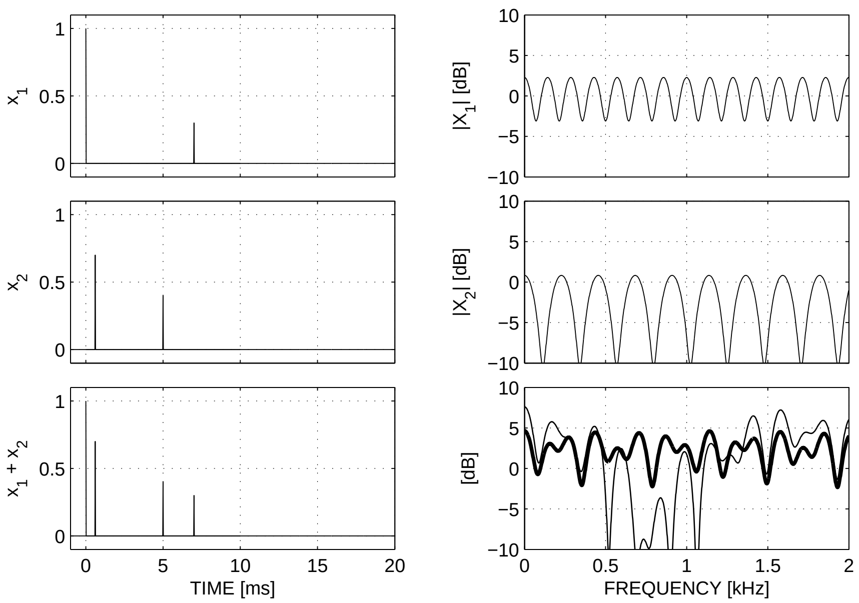


Figure 3.3: Two signals $x_1(n)$ and $x_2(n)$ and their respective magnitude spectra $X_1(j\omega)$ and $X_2(j\omega)$ (top two rows). Sum signal $x_1(n)+x_2(n)$ magnitude spectrum (bottom right, thin) and equalized sum signal magnitude spectrum (bottom right, bold) are shown.

where $p_{\tilde{x}_c}(k)$ is a short-time estimate of the power of $\tilde{x}_c(k)$ at time index k and $p_{\tilde{x}}(k)$ is a short-time estimate of the power of $\sum_{c=1}^{C} \tilde{x}_c(k)$. From (3.2) it follows that

$$e(k) = \sqrt{\frac{\sum_{c=1}^{C} p_{\tilde{x}_c}(k)}{p_{\tilde{x}}(k)}}\,. \tag{3.3}$$

The equalized subbands are transformed back to the time domain resulting in the sum signal $s(n)$ that is transmitted to the BCC decoder.

An example for the effect of the proposed downmixing with equalization is illustrated in Figure 3.3. The top two rows show two signals and their respective magnitude spectra. The bottom row shows the sum of the two signals and the magnitude spectra of the "simple" sum signal and the equalized sum signal as generated with the scheme shown in Figure 3.2. In the range from 500 Hz to about 1 kHz the equalization for this example prevents that the sum signal is significantly attenuated. The shown example is somewhat artificial since the x_1 and x_2 signals look more like impulse responses (direct sound plus one reflection) than real-world signals. Equalization can be viewed as modifying the interactions of impulse responses to prevent attenuation or amplification of signal components.

3.3.3 "Perceptually relevant differences" between audio channels

Given the sum signal, BCC synthesizes a stereo or multi-channel audio signal such that ICTD, ICLD, and ICC approximate the corresponding cues of the original audio signal. In the following, the role of ICTD, ICLD, and ICC in relation to auditory spatial image attributes is discussed.

The discussion in Section 2.2 implies that for one auditory event ICTD and ICLD are related to perceived direction. When considering *binaural room impulse responses* (BRIRs) of one source, there is a relationship between width of the auditory event and listener envelopment and IC estimated for the early and late parts of the BRIRs. However, the relationship between IC (or ICC) and these properties for general signals (and not just the BRIRs) is not straightforward.

Stereo and multi-channel audio signals usually contain a complex mix of concurrently active source signals superimposed by reflected signal components resulting from recording in enclosed spaces or added by the recording engineer for artificially creating a spatial impression. Different source signals and their reflections occupy different regions in the time-frequency plane. This is reflected by ICTD, ICLD, and ICC which vary as a function of time and frequency. In this case, the relation between instantaneous ICTD, ICLD, and ICC and auditory event directions and spatial impression is not obvious. The strategy of BCC is to blindly synthesize these cues such that they approximate the corresponding cues of the original audio signal.

We use filterbanks with subbands of bandwidths equal to two times the *equivalent rectangular bandwidth* (ERB) (Glasberg and Moore 1990). Informal listening revealed that the audio quality of BCC did not notably improve when choosing higher frequency resolution. A lower frequency resolution is favorable since it results in less ICTD, ICLD, and ICC values that need to be transmitted to the decoder and thus in a lower bitrate.

Regarding time-resolution, ICTD, ICLD, and ICC are considered at regular time intervals. Best performance is obtained when ICTD, ICLD, and ICC are considered about every $4 - 16$ ms (Section 3.5.1). Note that unless the cues are considered at very short time intervals, the precedence effect is not directly considered. Assuming a classical lead-lag pair of sound stimuli, when the lead and lag fall into a time interval where only one set of cues is synthesized, localization dominance of the lead is not considered. Despite of this, BCC achieves audio quality reflected in an average MUSHRA score (Rec. ITU-R BS.1534 2003) of about 87 ("excellent" audio quality) on average and up to nearly 100 for certain audio signals (Section 4.4).

The often achieved perceptually small difference between reference signal and synthesized signal implies that cues related to a wide range of auditory spatial image attributes are implicitly considered by synthesizing ICTD, ICLD, and ICC at regular time intervals. In the following, some arguments are given on how ICTD, ICLD, and ICC may relate to a range of auditory spatial image attributes.

Source localization (auditory event direction)

Several years of working with and optimizing BCC gave rise to an intuition that ICC (and thus implicitly IC) may play an important role for source localization independent of the attributes related directly to ICC and IC described in Section 2.2. The model for source localization proposed in Chapter 7 is based on this intuition and speculates about a possibly important role IC may play for source localization. This includes localization of sources in the presence of concurrent sound and reflections. The validity of this model would in many cases justify the use of ICTD, ICLD, and ICC only at regular time intervals without explicitly considering the precedence effect for real-world audio signals, as discussed in Section 7.5.

Attributes related to reflections

Early reflections up to about 20 ms result in coloration of sources' signals. This coloration effect is different for each audio channel determined by the timing of the early reflections contained in the channel. BCC does not attempt to retrieve the corresponding early reflected sound for each audio channel (which is a source separation problem). However, frequency dependent ICLD synthesis imposes on each output channel the spectral envelope of the original audio signal and thus is able to mimic coloration effects caused by early reflections.

Most perceptual phenomena related to spatial impression seem to be related directly to the nature of reflections that occur following the direct sound. This includes the nature of early reflections up to 80 ms and late reflections beyond 80 ms. Thus it is crucial that the effect of these reflections is mimicked by the synthesized signal.

ICTD and ICLD synthesis ideally result in that each channel of the synthesized output signal has the same temporal and spectral envelope as the original signal. This includes the decay of reverberation (the sum of all reflections is preserved in the transmitted sum signal and ICLD synthesis imposes the desired decay for each audio channel individually). ICC synthesis de-correlates signal components that were originally de-correlated by lateral reflections. Also, there is no need of considering reverberation time explicitly. Blindly synthesizing ICC at each time instant to approximate ICC of the original signal has the desired effect of mimicking different reverberation times, since ICLD synthesis imposes the desired rate of decay.

The most important cues for auditory event distance are overall sound level and direct sound to total reflected sound ratio (Shinn-Cunningham 2000). Since BCC generates level information and reverberation such that it approaches that of the original signal, also auditory event distance cues are represented by considering ICTD, ICLD, and ICC cues.

3.3.4 Estimation of spatial cues

In the following, it is described how ICTD, ICLD, and ICC are estimated. The bitrate required for transmission of these spatial cues is just a few kb/s and thus with BCC it is possible to transmit stereo and multi-channel audio signals at bitrates close to what is required for a single audio channel. Quantization and coding of the spatial cues is discussed in Section 6.2.4.

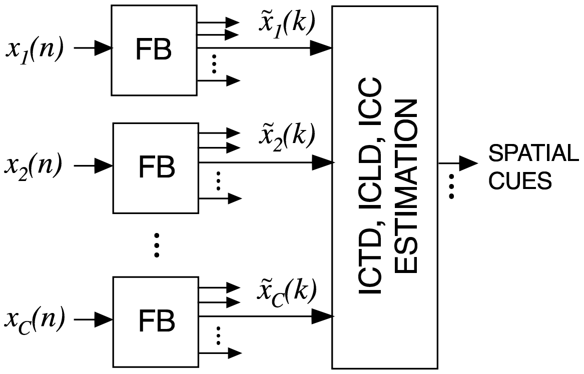


Figure 3.4: The spatial cues, ICTD, ICLD, and ICC are estimated in a subband domain. The spatial cue estimation is applied independently to each subband.

Estimation of ICTD, ICLD, and ICC for stereo signals

The scheme for estimation of ICTD, ICLD, and ICC is shown in Figure 3.4. The following measures are used for ICTD, ICLD, and ICC for corresponding subband signals $\tilde{x}_1(k)$ and $\tilde{x}_2(k)$ of two audio channels:

- ICTD [samples]:

  $$\tau_{12}(k) = \arg\max_d \{\Phi_{12}(d,k)\}\,, \tag{3.4}$$

  with a short-time estimate of the normalized cross-correlation function

  $$\Phi_{12}(d,k) = \frac{p_{\tilde{x}_1\tilde{x}_2}(d,k)}{\sqrt{p_{\tilde{x}_1}(k-d_1)p_{\tilde{x}_2}(k-d_2)}}\,, \tag{3.5}$$

  where

  $$\begin{aligned} d_1 &= \max\{-d,0\} \\ d_2 &= \max\{d,0\}\,, \end{aligned} \tag{3.6}$$

  and $p_{\tilde{x}_1\tilde{x}_2}(d,k)$ is a short-time estimate of the mean of $\tilde{x}_1(k-d_1)\tilde{x}_2(k-d_2)$.

- ICLD [dB]:

  $$\Delta L_{12}(k) = 10\log_{10}\left(\frac{p_{\tilde{x}_2}(k)}{p_{\tilde{x}_1}(k)}\right)\,. \tag{3.7}$$

- ICC:

  $$c_{12}(k) = \max_d |\Phi_{12}(d,k)|\,. \tag{3.8}$$

  Note that the absolute value of the normalized cross-correlation is considered and $c_{12}(k)$ has a range of [0, 1]. Out-of-phase signal pairs can not be represented by these cues as defined. Real-world audio signals only contain phase-inverted signal components in unusual cases and we thus do not consider this case explicitly

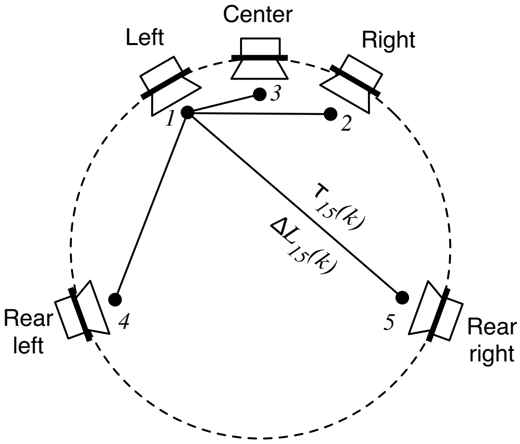


Figure 3.5: ICTD and ICLD are defined between the reference channel 1 and each of the other $C-1$ channels.

Estimation of ICTD, ICLD, and ICC for multi-channel audio signals

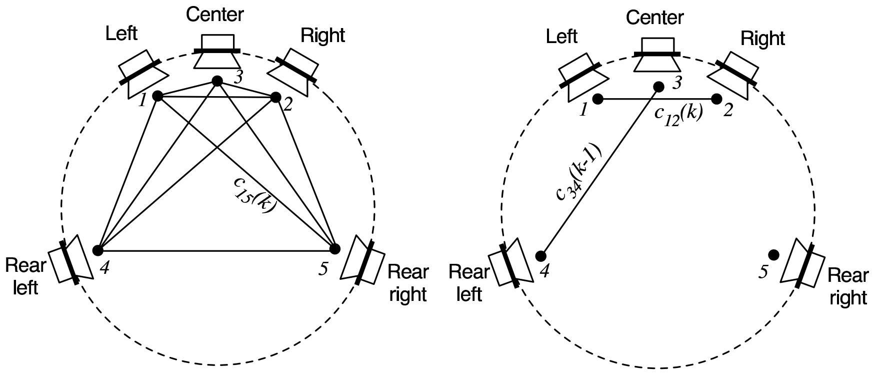
It is enough to define ICTD and ICLD between a reference channel (e.g. channel number 1) and the other channels as illustrated in Figure 3.5 for the case of $C = 5$ channels. $\tau_{1c}(k)$ and $\Delta L_{1c}(k)$ denote the ICTD and ICLD between the reference channel 1 and channel c.

As opposed to ICTD and ICLD, ICC has more degrees of freedom. The ICC as defined can have different values between all possible input channel pairs. For C channels there are $C(C-1)/2$ possible channel pairs, e.g. for 5 channels there are 10 channel pairs as illustrated in Figure 3.6(a). However, such a scheme requires that for each subband at each time index $C(C-1)/2$ ICC are estimated and transmitted, resulting in high computational complexity and high bitrate.

For each subband, ICTD and ICLD determine the direction at which the auditory event of the corresponding signal component in the subband is rendered. One single ICC parameter per subband is used to describe the overall coherence between all audio channels. We obtained good results by estimating and transmitting only ICC cues between the two channels with most energy in each subband at each time index. This is illustrated in Figure 3.6(b), when for time instants $k-1$ and k the channel pairs (3, 4) and (1, 2) are strongest, respectively. A heuristic rule is used for determining ICC between the other channel pairs (see Chapter 4.3.2 for details).

3.3.5 Synthesis of spatial cues

Figure 3.7 shows the scheme which is used in the BCC decoder to generate a stereo or multi-channel audio signal, given the transmitted sum signal plus the spatial cues. The sum signal $s(n)$ is decomposed into subbands, where $\tilde{s}(k)$ denotes one such subband. For generating the corresponding subbands of each



(a) (b)

Figure 3.6: Computation of IC for multi-channel audio signals. (a): In the most general case, ICC is considered for each subband between each possible channel pair. (b): BCC considers for each subband at each time instant k, the ICC between the channel pair with the most power in the considered subband. In the example shown the channel pair is $(3,\,4)$ at time instant $k-1$ and $(1,\,2)$ at time instant k.

of the output channels, delays d_c, scale factors a_c, and filters h_c are applied to the corresponding subband of the sum signal. (For simplicity of notation, the time index k is ignored in the delays, scale factors, and filters).

ICTD synthesis

The delays are determined by the ICTDs,

$$d_c = \begin{cases} -\frac{1}{2}(\max_{2\le l\le C}\tau_{1l}(k) + \min_{2\le l\le C}\tau_{1l}(k)), & c=1 \\ \tau_{1c}(k)+d_1, & 2\le c\le C\,. \end{cases} \qquad (3.9)$$

The delay for the reference channel, d_1, is computed such that the maximum magnitude of the delays d_c is minimized. The less the subband signals are modified, the less there is a danger for artifacts to occur. If the subband sampling rate does not provide high enough time-resolution for ICTD synthesis, delays can be imposed more precisely by using suitable all-pass filters.

ICLD synthesis

In order that the output subband signals have desired ICLDs (3.7) between channel c and the reference channel 1, $\Delta L_{1c}(k)$, the gain factors a_c must satisfy

$$\frac{a_c}{a_1} = 10^{\frac{\Delta L_{1c}(k)}{20}}\,. \qquad (3.10)$$

Additionally, the output subbands are normalized such that the sum of the power of all output channels is equal to the power of the input sum signal. Since the total original signal power in each subband is preserved in the sum signal (Section 3.3.2), this normalization results in that the absolute subband power for each output channel approximates the corresponding power of the

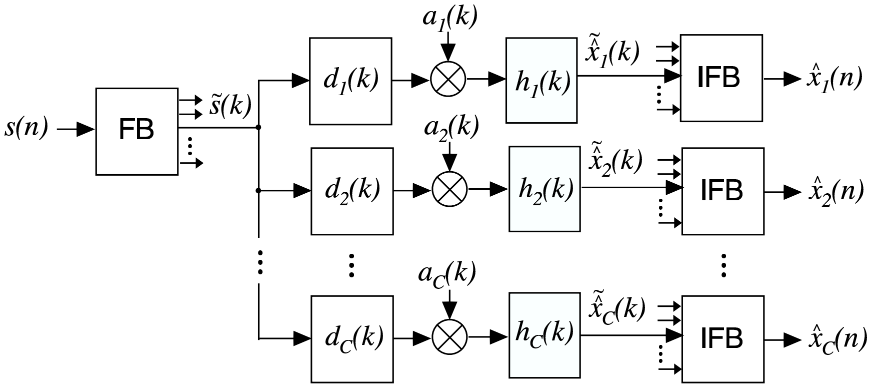


Figure 3.7: ICTD are synthesized by imposing delays, ICLD by scaling, and ICC by applying de-correlation filters. The shown processing is applied independently to each subband.

original encoder input audio signal. Given these constraints, the scale factors are

$$a_c = \begin{cases} 1/\sqrt{1+\sum_{i=2}^{C} 10^{\Delta L_{1i}/10}}, & \text{for } c = 1 \\ 10^{\Delta L_{1c}/20} a_1, & \text{otherwise}. \end{cases} \tag{3.11}$$

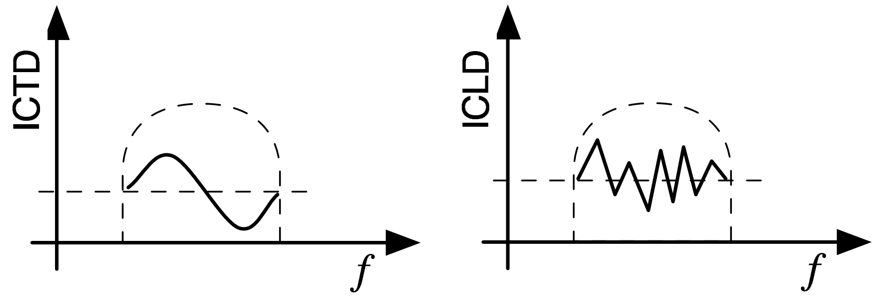


Figure 3.8: ICC is synthesized in subbands by varying ICTD and ICLD as a function of frequency.

ICC synthesis

The aim is to reduce correlation between the subbands after delays and scaling have been applied, without affecting ICTD and ICLD. This is achieved by designing the filters h_c in Figure 3.7 such that ICTD and ICLD are effectively varied as a function of frequency such that the average variation is zero in each subband (auditory critical band). Figure 3.8 illustrates how ICTD and ICLD are varied within a subband as a function of frequency. The amplitude of ICTD and ICLD variation determines the degree of de-correlation and is controlled as a function of ICC. Note that ICTD are varied smoothly while ICLD are varied randomly. One could vary ICLD as smoothly as ICTD, but this would result in more coloration of the resulting audio signals. Detailed processing for ICTD and ICLD variation as a function of ICC is described in Chapter 6.2.5. Another method for synthesizing ICC, particularly suitable for multi-channel ICC synthesis, is described in Chapter 4. As a function of time and frequency, specific amounts of artificial late reverberation is added to each of the output channels for achieving a desired ICC. Additionally, spectral modification is applied such that the spectral envelope of the resulting signal approaches the spectral envelope of the original audio signal.

3.3.6 Manipulation of spatial cues at the decoder

The parametric nature of BCC allows modification of auditory spatial image attributes at the decoder by modifying the transmitted ICTD, ICLD, and ICC before BCC synthesis is applied. By linear scaling of ICTD and ICLD the overall width of the auditory spatial image can be controlled. Manipulation of ICC can be used to modify overall diffuseness of the auditory spatial image. In the following, two other examples are described for modification of the spatial cues at the decoder.

Generation of ICTD cues at the decoder

As mentioned in Chapter 2.3, often only ICLD are used as directional cues when stereo signals are recorded or artificially mixed (e.g. coincident-pair microphones, amplitude panning). Therefore, one may decide not to transmit ICTD, resulting in a lower bitrate. However, according to the duplex theory (Rayleigh 1907), for headphone playback ICTDs are important at frequencies below about $1 - 1.5$ kHz. Thus for improved headphone performance without transmitting ICTD cues possibly present in the signal, ICTDs can alternatively be generated at the decoder with values proportional to the corresponding transmitted ICLDs,

$$\tau_{1c,b} = -q f_s \Delta L_{1c,b}\,, \tag{3.12}$$

with a scaling factor of q. Positive values of q result in ICTDs shifting the auditory events in the same direction as the given ICLDs. (3.12) is applied only at frequencies below $1 - 1.5$ kHz, where ICTDs are most important according to the duplex theory.

The criterion for determining q may be based on wideband empiric psychoacoustic lateralization data ((Blauert 1997) Figs. 2.68 and 2.80). The factor q is determined by fitting the linear frequency-independent function (3.12) to the data and determining ICTD such that the resulting lateralization is equal to the lateralization caused by the corresponding ICLD, resulting in $q = 80\ \mu\text{s/dB}$ (Baumgarte and Faller 2002*a*). This can be motivated in the following way: The ICLD resulting from amplitude panning or coincident-pair microphones correspond to the “intended” lateralization. Thus for low frequencies ICTD are computed such that they correspond to that intended lateralization.

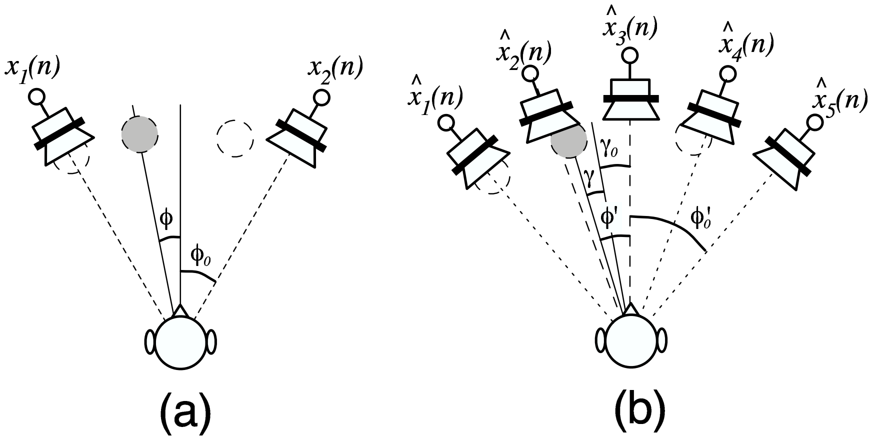


Figure 3.9: Playback setups. (a): Standard stereo playback setup. (b): Signal intended for standard stereo setup mapped to 5 loudspeakers for extended auditory spatial image width and improved stability.

Mapping from stereo to multi-channel

Suppose BCC is used for coding of a stereo signal. The default playback setup for such a stereo signal is shown in Figure 3.9(a). The goal is to generate more than two audio channels at the decoder for a playback setup with more than two front loudspeakers as shown in Figure 3.9(b). The intention is to generate more signal channels with a similar spatial image. For example, by using five as opposed to two loudspeakers as illustrated in Figure 3.9, the auditory spatial image can be made wider (by increasing the angle enclosed by the loudspeakers) and stability can be improved.

Avendano and Jot (2002) proposed a technique for upmixing stereo signals for multi-channel surround loudspeaker playback. Ambience signal components are extracted from the stereo signal and given to the rear loudspeakers. The technique proposed here has a different aim since it only attempts to reproduce the same frontal spatial image but with more front loudspeakers.

The loudspeaker angles of a standard stereo setup are $\phi_0 = \pm 30^\circ$. In the example shown in Figure 3.9(b) the outer loudspeakers are placed at $\phi_0' = \pm 40^\circ$. The mapping from stereo to five channels works as follows. ICLD which coincide with ICC larger than a certain threshold are treated as strong directional cues. Thus in this case the subband signal is amplitude panned between the two loudspeakers closest to the intended auditory event direction. Figure 3.10(a) shows the relation between the ICLD of the stereo signal, ΔL_{stereo}, and auditory event angle ϕ according to the stereophonic law of sines (2.3). This angle is scaled to a range covering the multi-loudspeaker setup. In our example, $\phi' = \frac{4}{3}\phi$. Then amplitude panning is applied to the two loudspeakers closest to the auditory event direction ϕ'. In the example shown in Figure 3.9(b) these are loudspeaker number 2 and 3. The angles corresponding to this selected loudspeaker pair, γ_0 and γ, are illustrated in Figure 3.9(b). The stereophonic law of sines (2.3) is applied for computing the gain factors for the loudspeaker

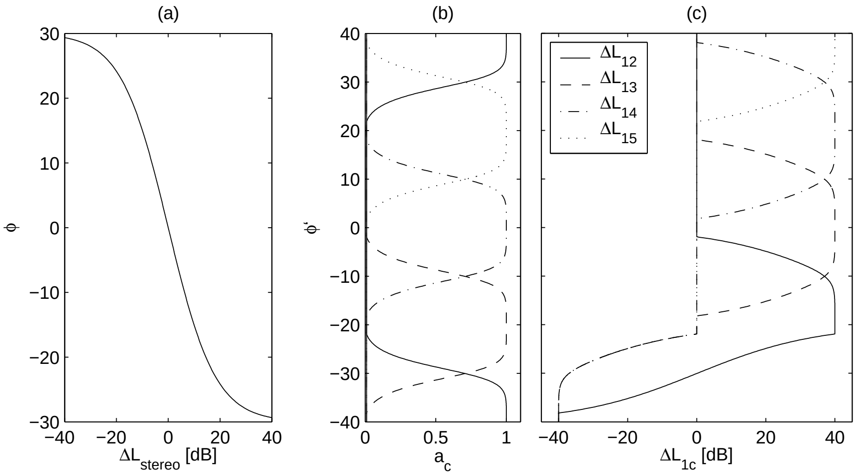


Figure 3.10: (a): Relation between ICLD of a stereo signal, ΔL_{stereo}, and the angle of the auditory event ϕ. (b): Gain factors for the five loudspeakers, a_c $(1 \leq c \leq C)$, as a function of auditory event angle ϕ'. (c): ICLD between channel pairs, ΔL_{1c}, as a function of auditory event angle ϕ'.

pair resulting in a relative auditory event angle γ. The gain factors a_c for each loudspeaker $(1 \leq c \leq C)$ as a function of ϕ' are shown in Figure 3.10(b). Given these gain factors ΔL_{1c} can be computed. These are shown in Figure 3.10(c). Note that ΔL_{1c} were limited to a range of ±40 dB. Most ICLDs in Figure 3.10(c) are positive since ICLD are defined relative to a reference channel (channel 1) and negative ICLDs only occur when for amplitude panning the loudspeaker pair with the reference channel is selected.

When the ICC is smaller than the chosen threshold, ΔL_{1c} are set such that signal is emitted from the two side loudspeakers (in the given example the loudspeakers at $\phi' = \pm 40°$) such that IC is still effectively lowered by signals emitted by relatively widely spaced loudspeakers. Note that while ICLD are re-computed for the multi-channel case, ICC are directly used as given from the stereo signal.

3.4 Coding of low frequency effects (LFE) audio channels

Commonly used multi-channel surround formats, such as 5.1 surround (Section 2.3.3), use LFE channels. An LFE channel, as defined for the 5.1 standard (Rec. ITU-R BS.775 1993), contains only frequencies up to 120 Hz. In the following we describe how the LFE channel in 5.1 is coded with BCC. The same principles are applicable to other surround formats.

At frequencies below 120 Hz, 6−channel BCC is applied, i.e. all six channels including the LFE channel are coded. At frequencies above 120 Hz, 5−channel BCC is applied, i.e. all channels except the LFE channel. The LFE channel is not considered at higher frequencies since it does not contain any signal energy there.

This is implemented specifically by using a filterbank with a lowest subband covering $0-120$ Hz. For this lowest subband the LFE channel is considered and for all other subbands the LFE channel is ignored. BCC only considering ICLD is applied to this lowest frequency subband since at such low frequencies spatial hearing is limited and the purpose of applying BCC is merely for providing to each loudspeaker the same power as was present in the original audio signal.

Since LFE channels are considered only in the lowest frequency BCC subband, at each time instant only a single ICLD parameter per LFE channels is used. Thus the amount of side information does not notably increase by including one or two LFE channels.

3.5 Subjective evaluations

At the time when these tests were carried out (October 2001), only algorithms for ICTD and ICLD synthesis were available. Thus the results presented in the following consider BCC without ICC synthesis. Subjective evaluations for BCC with ICC synthesis are presented in Section 4.4.

Thus, the results presented in this section do not have the aim of demonstrating the best possible quality BCC can achieve. The aim of the subjective evaluations presented here was two-fold. The first evaluation was aimed at assessing the impact of different time-resolutions used for spatial cue synthesis. The second evaluation compares a BCC-based scheme to a conventional audio coder in terms of bitrate and audio quality.

BCC was implemented using a *short time Fourier transform* (STFT) as described in Chapter 6. The STFT is implemented using a *fast Fourier transform* (FFT) and Hann or sine windows with 50% overlap. The perceptually motivated subband widths in which BCC operates are mimicked by grouping spectral coefficients together such that each group of spectral coefficients conceptually represent one BCC subband (for details see Section 6.2.1).

3.5.1 Different time resolutions for ICLD synthesis

A subjective test was designed to assess the effect of different time resolutions for ICLD synthesis on the audio quality. The time resolution of ICLD synthesis was varied by using various STFT window sizes. The used FFT size always corresponded to the STFT window size (no zero padding was used). In the following the term “window size” is used to refer to the different STFT window sizes.

Subjects and playback setup

Five experienced subjects participated in the tests. For audio playback an *Apple PowerBook G4* laptop computer was used with an external digital audio-out device (*Emagic EMI A26*) connected to a *Mark Levinson N360 Audio Processor* (D/A), a *Mark Levinson N380S Pre-amp*, and a *Mark Levinson N335 Power amp* connected to *B&W Nautilus 801* loudspeakers. The test was performed by each subject sitting at the standard listening position for conventional stereo playback. The test was carried out in a quiet (but not sound proof) listening room.

Stimuli

The ICLD cues are estimated and synthesized at different rates. ICLD are updated every 4, 8, 16, and 32 ms. This corresponds to window sizes of 256, 512, 1024, and 2048 samples. Note that sine windows are used with 50% window overlap. Thus ICLD are updated two times during the time span of one window, resulting in the above mentioned update rates. The choice of ICLD update rates is motivated by experimental psychoacoustic data (Holube *et al.* 1998) that shows a binaural time/frequency resolution in the same range. Moreover, preliminary experiments suggested that FFT lengths below 256 should be excluded since the average overall audio quality did not improve while the side information rate increased.

The STFT-based BCC implementation was only used for ICLD synthesis and one common algorithm was used for ICLD estimation, i.e. a cochlear filterbank was used and ICLD were estimated in critical bands (Baumgarte and Faller 2002*b*). These estimates were downsampled and/or interpolated to match the time/frequency resolution of the synthesis schemes with the various window sizes. Due to the auditory filter properties of the cochlear filterbank, the effective frequency resolution of the estimated ICLD corresponds to BCC with subbands of width 1 ERB.

Four different stereo audio excerpts, each with a duration of approximately 10 s sampled at 32 kHz, were used in the test. Table 3.1 summarizes their contents. The fourth excerpt is a stereophonic recording of applause. It is known as a critical signal for stereo audio coding since the spatial image is very dynamic.

Table 3.1: List of the audio excerpts used. The last two columns contain the sources of excerpt 1, 2, and 3, that are placed to the left or right side of the auditory spatial image by amplitude panning with ±10 dB.

excerpt	*category*	*left*	*right*
1	speech	male	female
2	singing	tenor	soprano
3	percussions	castanets	drums
4	applause	(stereo recording)	

Test method

The subjects were asked to grade different specific degradations and the overall audio quality of the processed excerpts with respect to the known reference, i.e. the original excerpt. The four different grading tasks of this test are summarized in Table 3.2. Tasks 1 and 2 assess two properties that are thought to be important factors for auditory spatial image quality, i.e. width and stability. Task 3 evaluates distortions introduced by BCC that do not result in auditory spatial image artifacts. For example, aliasing and blocking artifacts should be detected here. Task 4 is an important measure for global optimization of BCC.

Table 3.2: Tasks and scales of the subjective test in experiment 3.

task		*scale*
1	image width	stereo...mono
2	image stability	stable...unstable
3	audio quality ignoring spatial image distortions	ITU-R 5-grade impairment
4	overall audio quality	ITU-R 5-grade impairment

During the test, each subject was able to randomly access each test item processed with the four different window sizes (corresponding to different ICLD update rates) and the reference. Each item could be accessed with a corresponding “Play” button of a graphical user interface. This Play function stops a possibly active audio output at any time, such that the subject can do quick initial listening through all items before proceeding with a more thorough evaluation. The gradings were entered via graphical sliders that are permanently visible for all test items and can be adjusted at any time to reflect the proper grading and ranking. It is important to note that subjects were specifically asked to pay attention to the rank order of the test items. The feature of being able to play the items according to their rank order greatly facilitates this task as opposed to other testing schemes that allow to listen only once to each item in a pre-defined order. The ordering of the different window sizes was randomly chosen for each subject and each excerpt but not changed during the four different tasks performed for each excerpt. The philosophy of this test method corresponds closely to MUSHRA (Rec. ITU-R BS.1534 2003).

Results

The experimental results are shown for the individual excerpts only. Averaging over the gradings of different excerpts cannot be justified due to substantially deviating ratings. The gradings of each task is discussed in the following.

Image width The gradings for image width are shown in Figure 3.11 for each excerpt and each window size with respect to the reference. Apparently, all window sizes reduce the image width for all test items. For excerpt 2 there is a trend toward a smaller image width with reduced window size. This trend is reverse for excerpt 3. This result can be explained by the more stationary character of excerpt 2 (singing) requiring higher frequency resolution in contrast to the non-stationary excerpt 3 (percussions) which requires a higher time resolution. The overall performance of all considered window sizes is comparable.

Image stability Gradings for image stability with respect to the reference are given in Figure 3.12. The image stability is best if the auditory event locations are perceived as stationary. These locations are well defined for the reference excerpts 1, 2, and 3. For excerpt 4 (applause) each source is only active for a short time so that a moving auditory event cannot be detected. That is why excerpt 4 appears relatively close to “stable” for all window sizes. From the remaining excerpts, 1 and 3 are more critical

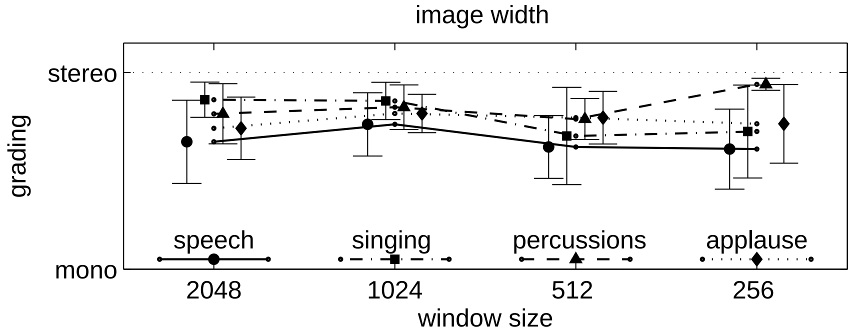


Figure 3.11: Auditory image width gradings and 95% confidence intervals. The grading "stereo" corresponds to the auditory image width of the reference item.

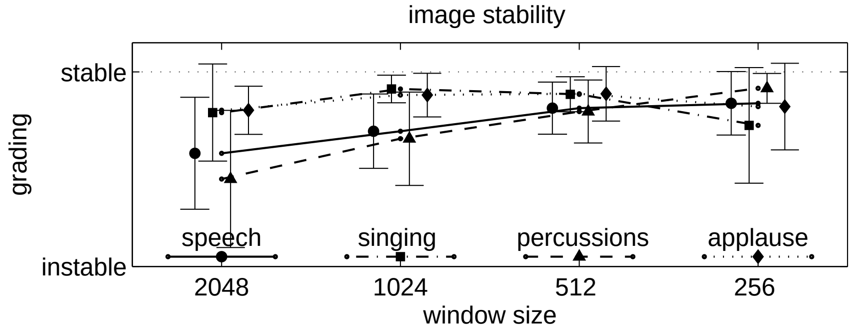


Figure 3.12: Auditory image stability gradings and 95% confidence intervals. The grading "stable" corresponds to the auditory image stability of the reference item.

than 2. For excerpt 1 and 3 the stability increases consistently as the window size becomes shorter. For excerpt 2 a medium window size (512 and 1024) shows best gradings. Window sizes of 256 and 512 perform best followed closely by the window size of 1024. The window size of 2048 is clearly worse than the shorter window sizes.

Quality, ignoring image distortions In task 3 the audio quality is assessed with respect to the reference without considering auditory spatial image degradations. The results in Figure 3.13 show no significant degradations except for excerpt 3 which appears critical for the window sizes 2048 and 1024. The time resolution is apparently insufficient for this excerpt (percussions) containing many transients.

Overall quality The overall quality gradings in Figure 3.14 show the integral impact of all noticeable degradations on audio quality to facilitate the selection of the window size with best overall performance. Obviously, the overall quality reflects the influence of the degradations assessed in task 1, 2, and 3 and it combines these individual components into a perceptually meaningful global measure. From visual inspection it is concluded that

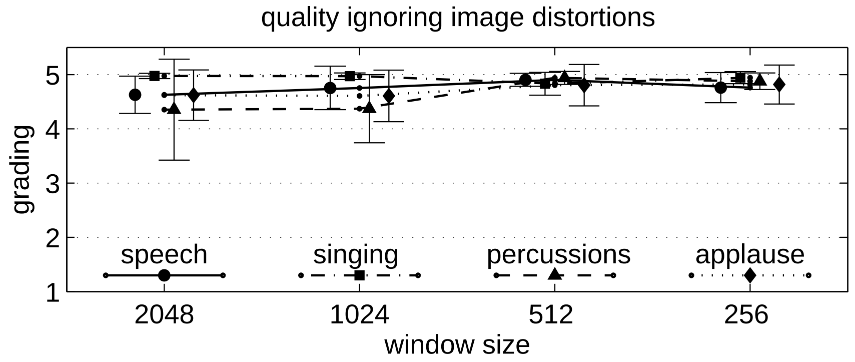


Figure 3.13: Perceived audio quality ignoring spatial image distortions gradings and 95% confidence intervals.

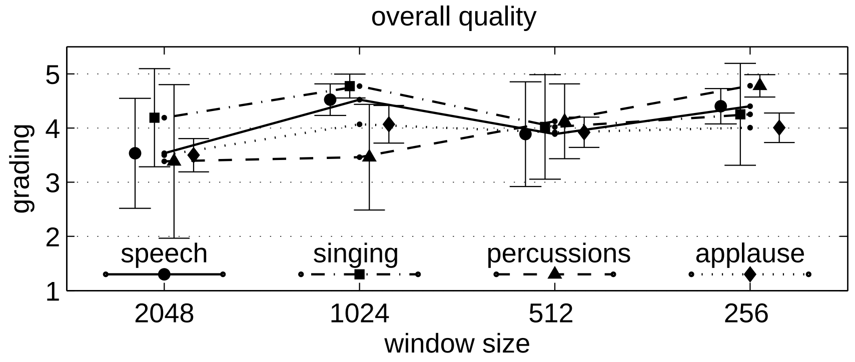


Figure 3.14: Perceived audio quality gradings and 95% confidence intervals.

a window size of 256 has best performance for the test excerpts followed by the window sizes 512 and 1024. Window size 2048 shows significantly reduced quality for three excerpts. The window size 256 has a clear advantage over longer window sizes for excerpt 3 (percussions) which requires a high time resolution. For the more stationary excerpts 1 and 2 a window size of 1024 seems to be best.

Discussion

It is interesting to note that the time resolution corresponding to window size 256 is higher than the measured binaural time resolution summarized in (Holube *et al.* 1998). This may be related to phenomena related to the precedence effect, i.e. a lower time resolution results in that localization dominance is more often incorrectly synthesized by averaging the cues of lead and lag within the same window. The frequency resolution of the window size 256 (256-point FFT) is slightly lower than experimental data of (Holube *et al.* 1998), which may explain why this window size performs worse for tonal signals. The overall results imply that a window size of 256 is best (4 ms ICLD update rate). The performance of the window sizes 512 and 1024 is slightly worse. Window size 2048 is clearly worse compared to the shorter window sizes. For reasons of algorithm simplicity and bitrate, we often use longer window sizes than 256. Actually, we use ICLD update rates between $4 - 16$ ms, corresponding to window sizes in the range of $256 - 1024$ for 32 kHz sampling frequency.

3.5.2 Stereo audio coding

We compared audio coding schemes based on BCC with conventional stereo audio coders at various bitrates. Each conventional stereo audio coder was compared with a BCC-based scheme using the same audio coder for encoding the mono sum signal. For that purpose we used PAC (Sinha *et al.* 1997) and MPEG-1 Layer 3 ("MP3") (Brandenburg and Stoll 1994, ISO/IEC 1993). The MP3 encoder incorporated into Apple's iTunes program was used. The goal of this test was to assess in which range of bitrate BCC-based coders outperform conventional stereo audio coders.

For this test BCC was used with a Hann window of size 896 with 50% overlap. This results in that ICLD and ICTD are updated every 14 ms.

Subjects and playback setup

Ten experienced subjects participated in the tests. The playback setup was the same as in the previous test and is described in Section 3.5.1.

Stimuli

The rows in Table 3.3 show the bitrates of the audio coders for encoding the stereo signals and the sum signals when used with BCC. The bitrate of the BCC schemes is shown as the sum of the mono audio coder bitrate and the BCC side information bitrate. The mono audio coder bitrate is chosen lower than the bitrate for stereo because for the same level of distortion and audio bandwidth less bits are needed to encode the mono signal. For PAC we chose a sampling rate of 32 kHz and an audio bandwidth of 13.5 kHz. The parameters for the MP3 encoder were a bitrate of 40 kb/s for stereo and 32 kb/s for mono, joint-stereo coding enabled, and a sampling rate of 24 kHz. For both, PAC and MP3, we chose the fixed bitrate encoding mode.

To achieve a lower bitrate of the BCC side information, only ICLDs were used as inter-channel cues. We used an ICLD-to-ICTD scaling factor of $q = 25 \cdot 10^{-6}$ seconds/dB (3.12), with an ICLD range of ±18 dB. This results in an ICTD range of ±180 μs. Note that we have chosen q smaller than proposed in Section 3.3.6, to prevent some artifacts which occur during ICTD synthesis for large ICTD values. The ICLD cues were quantized and coded as described in Section 6.2.4.

For each of the tests we chose the same 14 music clips. Each of these clips has a pronounced wide spatial image. BCC is challenged by a wide spatial image in the sense that it needs to spatially separate auditory events. Also, for the conventional stereo audio coder a wide spatial image is challenging because the redundancy between the channels is small in that case resulting in a high bit demand. Different kinds of music signals such as jazz, rock, and

Table 3.3: The coders and bitrates for the three subjective tests conducted. The numbers (1-5) denote the five different coding configurations used.

Coder	*Bitrate for Stereo*	*Bitrate for BCC-based Coder*
PAC	(1) 64 kb/s	(2) $52 + 2$ kb/s
PAC	(3) 56 kb/s	(2) $52 + 2$ kb/s
MP3	(4) 40 kb/s	(5) $32 + 2$ kb/s

percussive music were selected. We did not consider any classical recordings because without ICC synthesis very reverberant items can not be reproduced well.

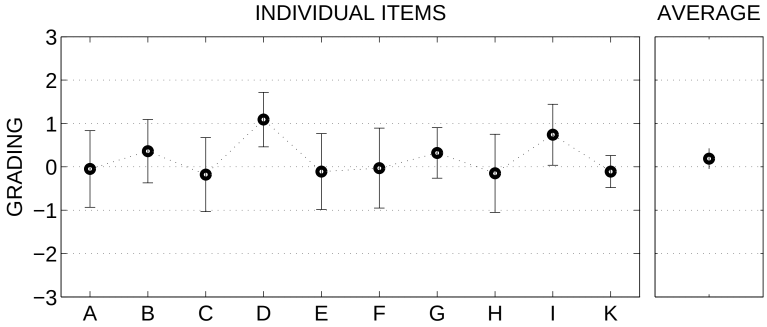


Figure 3.15: Relative grading and 95% confidence intervals of the BCC-based coder at $52 + 2$ kb/s compared to stereo PAC at 64 kb/s. The results are shown for the items individually (left panel) and averaged over all items (right panel). BCC is better than PAC for positive gradings (1: slightly better, 2: better, 3: much better).

Test method

Four of the clips were used as training items and 10 as test items. The type of test was a blind triple-stimulus test (Rec. ITU-R BS.562.3 1990) to grade the quality difference of two processed versions with respect to a reference using a seven-grade comparison scale. The subjects were presented with triples of signals, each of 12 s length for each trial. The uncoded source signal (reference) was presented first followed by the coded clips of the conventional stereo audio coders and BCC-based coders in random order. After initial presentation of the items, the subject could selectively listen to the items as many times as he wished. Switching between items was possible at any time.

Results

Figures 3.15, 3.16, and 3.17 show the results of the three subjective tests. Positive gradings correspond to preference for the BCC-based schemes. For

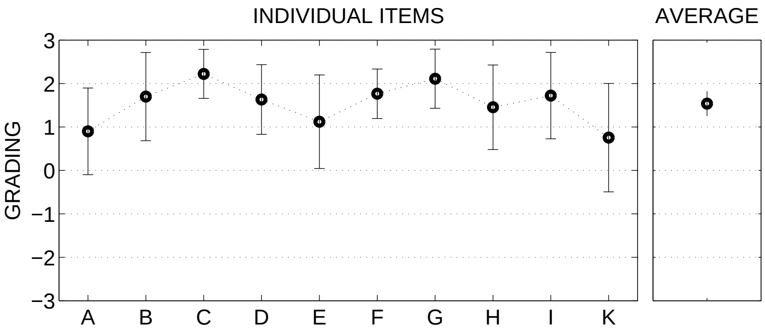


Figure 3.16: Relative grading and 95% confidence intervals of the BCC-based coder at $52 + 2$ kb/s compared to stereo PAC at 56 kb/s. The results are shown for the items individually (left panel) and averaged over all items (right panel). BCC is better than PAC for positive gradings (1: slightly better, 2: better, 3: much better).

every test the total bitrate of the BCC-based scheme was lower than the bitrate of the stereo audio coder. For the average of each test, the BCC-based scheme outperforms the stereo audio coder despite of its lower bitrate. It has to be noted that the artifacts of the two coding schemes are quite different. The BCC-based coder generally modifies the auditory spatial image more while the conventional stereo audio coders introduce more quantization distortions. From the test results one can conclude that the subjects preferred auditory spatial image modification (as introduced primarily by BCC) over quantization distortions (introduced primarily by the conventional coders).

Derived from the test results (Figures 3.15-3.17), Figure 3.18 shows qualitatively the subjective quality of each coding configuration that was used for the tests (the same numbering as in Table 3.3 is used): (1) Stereo PAC 64 kb/s, (2) Stereo PAC 56 kb/s, (3) BCC with mono PAC $52 + 2$ kb/s, (4) Stereo MP3 40 kb/s, and (5) BCC with mono MP3 $32 + 2$ kb/s. At bitrates high enough for transparent or nearly transparent coding the conventional coder is better since BCC does not achieve transparent quality. The test results give an indication that for bitrates in the range of about $24 - 70$ kb/s the BCC-based coding scheme has better quality than conventional perceptual transform audio coders for stereo.

3.6 Conclusions

BCC was motivated and described. A given stereo or multi-channel audio signal is downmixed to a single channel containing all signal components inherent in the input channels. ICTD, ICLD, and ICC are estimated in subbands at regular time intervals and transmitted to the decoder. The use of ICTD, ICLD, and ICC for representing the auditory spatial image of the encoder input signal was motivated. The BCC decoder generates an output signal such that ICTD, ICLD, and ICC approximate the corresponding cues of the original audio signal.

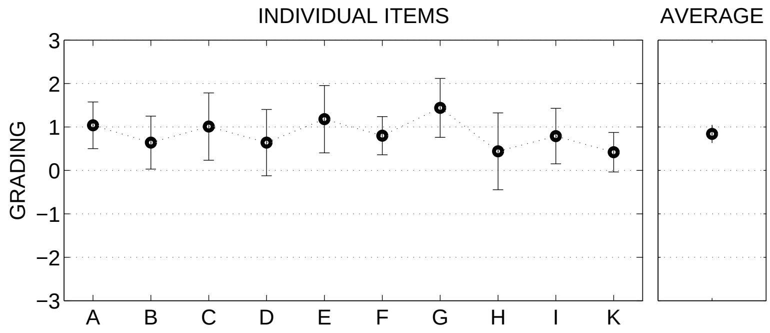


Figure 3.17: Relative grading and 95% confidence intervals of the BCC-based coder at $32 + 2$ kb/s compared to stereo MP3 at 40 kb/s. The results are shown for the items individually (left panel) and averaged over all items (right panel). BCC is better than MP3 for positive gradings (1: slightly better, 2: better, 3: much better).

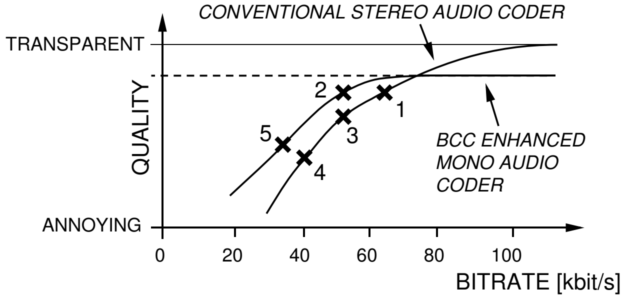


Figure 3.18: For bitrates at which conventional perceptual transform audio coders operate near transparent quality these are better than BCC-based schemes. At lower bitrates BCC-based schemes are better on average.

It was also described how a low frequency effects channel can be encoded with minor additional expense in terms of bitrate. A subjective test was carried out to get an indication of suitable time resolution for ICLD synthesis. A second subjective test implies that a BCC-based coder using only ICTD and ICLD outperforms a conventional stereo audio coder at bitrates below about 70 kb/s.

Chapter 4

Coherence Synthesis Based on Late Reverberation

4.1 Introduction

Binaural cue coding (BCC) was motivated and described in Chapter 3. BCC downmixes a stereo or multi-channel audio signal to a single sum signal and transmits that to the decoder. Additionally, side information is transmitted containing estimated values of inter-channel cues of the original audio signal. The considered cues are inter-channel time difference (ICTD), inter-channel level difference (ICLD), and inter-channel coherence (ICC). Given the transmitted sum signal, BCC synthesizes ICTD, ICLD, and ICC as a function of time and frequency such that the corresponding cues of the original audio signal are approximated.

In Section 3.3.5, ICC synthesis based on de-correlation by means of ICTD and ICLD variation in auditory critical bands was described. This technique is computationally very efficient (when for example implemented as shown in Section 6.2.5) and performs reasonably well for stereo audio signals. However, for more audio channels ICTD and ICLD variation does not yield enough de-correlation capability to generate independent signal components for each audio channel (unless the degree of ICTD and ICLD variation is chosen so large that the resulting audio quality is impaired). This problem is more pronounced at low frequencies (e.g. up to 1 kHz).

Both of these drawbacks are addressed by the ICC synthesis technique described in this chapter. The method is scalable to any number of audio channels and effectively de-correlates even very low frequency signal components if necessary. The drawback of the method presented in this chapter is its high computational complexity, which is partially addressed in Section 6.2.6 by describing how to implement the technique entirely in the frequency domain.

The chapter is organized as follows. Section 4.2 motivates and describes how a number of de-correlated audio channels are generated given the transmitted sum signal. Schemes for stereo and multi-channel ICTD, ICLD, and ICC synthesis making use of such de-correlated audio channels are described in Section 4.3. The results of subjective audio quality evaluations of the proposed schemes are presented in Section 4.4. Conclusions are drawn in Section 4.5.

4.2 Generating independent late reverberation given one audio channel

A natural listening scenario where low interaural coherence (IC) occurs is in a concert hall. Late reverberation sound arrives at the ears from random angles with random strength, such that the IC is low. Particularly lateral reflections, known to be important for good concert hall acoustics, result in low IC (Kuttruff 1991). The property of late reverberation to result in low IC gives the motivation for generating a number of de-correlated audio channels by applying late reverberation models to the given transmitted BCC sum signal $s(n)$.

C late reverberation audio channels $s_c(n)$ $(1 \leq c \leq C)$ with low ICC between pairs of channels are obtained by,

$$s_c(n) = h_c(n) \star s(n)\,, \tag{4.1}$$

where $\star$ denotes convolution and $h_i(n)$ are the filters modeling late reverberation. Late reverberation is modeled by impulse responses given by

$$h_i(n) = \begin{cases} r_i(n)(1 - \frac{1}{f_s T_h})^n, & 0 \leq n < M \\ 0, & \text{otherwise}\,, \end{cases} \tag{4.2}$$

where $r_i(n)$ $(1 \leq i \leq C)$ are independent stationary white Gaussian noise signals, T_h is the time constant in seconds of the exponential decay of the impulse response, f_s is the sampling frequency, and M is the length of the impulse response in samples. An exponential decay is chosen, because the strength of late reverberation is decaying exponentially in time.

The reverberation time of many concert halls is in the range of 1.5 to 3.5 seconds (Beranek 1962). In order that the late reverberation signals $s_c(n)$ are enough independent for generating as low IC or ICC as occur in concert halls, T_h is chosen such that reverberation times of $h_c(n)$ are in the same range. Given an impulse response, the reverberation time is computed as described in (Schroeder 1965). We choose $T_h = 0.4$ seconds, corresponding to a reverberation time of 2.8 seconds. M is chosen such that the impulse responses are 0.3 s long. (Note that the reverberation time considerations were only used to determine the rate of decay and not the length of the filters. The length of the filters was determined heuristically to be as short as possible without compromising the quality of the BCC synthesis scheme).

4.3 Using late reverberation for coherence synthesis

In this section, it is described how the late reverberation audio channels $s_c(n)$ are used to synthesize ICC. Each subband of each output signal channel is computed as a weighted sum of the corresponding subbands of $s(n)$ and $s_c(n)$ $(1 \leq c \leq C)$. The factors of the weighted sum are determined such that the ICC cues between the output subbands approximate those of the original audio signal.

As mentioned in the previous Chapter, ICC are synthesized every $4-16$ ms. The length and decay of the filters h_c determines the upper bound of de-correlation capability of the system. In Section 3.3.3 it was discussed how by spatial cue synthesis different degrees of reverberation are synthesized. Since the spatial cues are updated every 4−16 ms, the de-correlation and time spreading effect of the long tail of h_c is effectively suppressed if necessary. Thus, the long h_c's are suitable not only for synthesizing ICC mimicking very reverberant environments, but also ICC mimicking less reverberant environments.

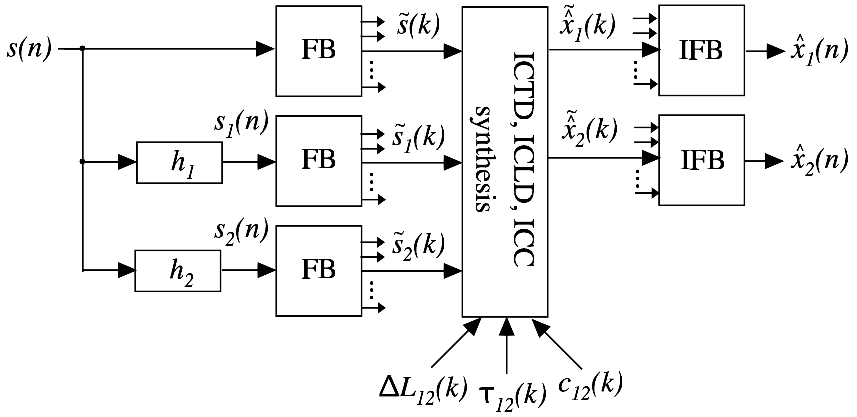


Figure 4.1: Scheme for generating stereo signals. ICTD, ICLD, and ICC synthesis is applied independently to each subband.

4.3.1 Stereo coherence synthesis

The proposed scheme is shown in Figure 4.1. The late reverberation channels $s_1(n)$ and $s_2(n)$ are generated by filtering the sum signal $s(n)$ as described in the previous section. The signals $s(n)$, $s_1(n)$, and $s_2(n)$ are decomposed into a number of spectral components by non-uniform filterbanks with subbands reflecting the frequency resolution of the auditory system (as described in Section 3.3.1). One subband signal of the decomposed three input signals is denoted $\tilde{s}(k)$, $\tilde{s}_1(k)$, or $\tilde{s}_2(k)$, respectively. As illustrated in Figure 4.2, the subbands of the two output channels are computed as a weighted sum of the given subband signals,

$$\begin{aligned} \tilde{\tilde{x}}_1(k) &= a_1\tilde{s}(k-d_1) + b_1\tilde{s}_1(k) \\ \tilde{\tilde{x}}_2(k) &= a_2\tilde{s}(k-d_2) + b_2\tilde{s}_2(k)\,, \end{aligned} \tag{4.3}$$

where the scale factors (a_1, a_2, b_1, b_2) and delays (d_1, d_2) are determined as a function of the desired ICTD $\tau_{12}(k)$, ICLD $\Delta L_{12}(k)$, and ICC $c_{12}(k)$. (The time index of the gain factors and delays is neglected for a simpler notation). The subband signals $\tilde{\tilde{x}}_1(k)$ and $\tilde{\tilde{x}}_2(k)$ are computed for all subbands and the output signals $x_1(n)$ and $x_2(n)$ are generated by applying the inverse of the filterbank used.

The ICTD, $\tau_{12}(k)$, is synthesized by imposing two different delays, d_1 and d_2 (3.9), on $\tilde{s}(k)$. If the subband sampling rate does not provide high enough time-resolution for ICTD synthesis, delays can be imposed more precisely by

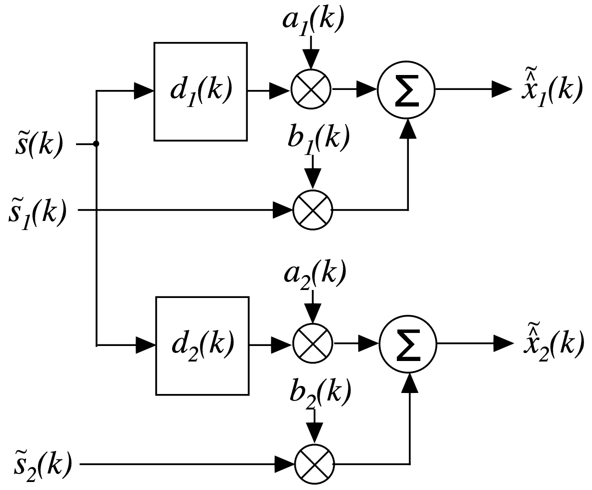


Figure 4.2: Processing of one subband for ICTD, ICLD, and ICC synthesis for stereo.

using suitable all-pass filters. In order that the output subband signals have an ICLD (3.7) equal to $\Delta L_{12}(k)$, the gain factors (a_1, a_2, b_1, b_2) must satisfy

$$\frac{a_1^2 p_{\tilde{s}}(k) + b_1^2 p_{\tilde{s}_1}(k)}{a_2^2 p_{\tilde{s}}(k) + b_2^2 p_{\tilde{s}_2}(k)} = 10^{\frac{\Delta L_{12}(k)}{10}}\,, \tag{4.4}$$

where $p_{\tilde{s}}(k)$, $p_{\tilde{s}_1}(k)$, and $p_{\tilde{s}_2}(k)$ are short-time power estimates of the subband signals $\tilde{s}(k)$, $\tilde{s}_1(k)$, and $\tilde{s}_2(k)$, respectively.

For the output subband signals to have a certain ICC (3.8), $c_{12}(k)$, the gain factors must satisfy

$$\frac{(a_1^2 + a_2^2) p_{\tilde{s}}(k)}{\sqrt{(a_1^2 p_{\tilde{s}}(k) + b_1^2 p_{\tilde{s}_1}(k))(a_2^2 p_{\tilde{s}}(k) + b_2^2 p_{\tilde{s}_2}(k))}} = c_{12}(k)\,, \tag{4.5}$$

as is easily shown by applying (3.5) and (3.8) to the signals given in (4.3) and assuming that $\tilde{s}(k)$, $\tilde{s}_1(k)$, and $\tilde{s}_2(k)$ are independent.

BCC normalizes its output signals, such that the sum of the power of all output channels is equal to the power of the input sum signal (as described in Section 3.3.5). This yields another equation for the gain factors,

$$(a_1^2 + a_2^2) p_{\tilde{s}}(k) + b_1^2 p_{\tilde{s}_1}(k) + b_2^2 p_{\tilde{s}_2}(k) = p_{\tilde{s}}(k). \tag{4.6}$$

Since there are four gain factors and three equations, there is one degree of freedom in the choice of the gain factors. Thus an additional condition can be formulated,

$$b_1^2 p_{\tilde{s}_1}(k) = b_2^2 p_{\tilde{s}_2}(k)\,. \tag{4.7}$$

Condition (4.7) forces the amount of late reverberation to be the same in the left and right channel. There are several motivations for doing this:

- Late reverberation as appearing in concert halls has a level which is nearly independent of position (for relatively small displacements). Thus, the level difference of the late reverberation between left and right is always about 0 dB.
- The sound of the stronger channel is modified less, reducing negative effects of the long convolutions (4.1), such as time spreading of transients.

The gain factors are the non-negative solutions of the equation system given by (4.4)-(4.7),

$$\begin{aligned}
a_1 &= \sqrt{\frac{10^{\frac{\Delta L_{12}(k)}{10}} + c_{12}(k)10^{\frac{\Delta L_{12}(k)}{20}} - 1}{2(10^{\frac{\Delta L_{12}(k)}{10}} + 1)}} \\
a_2 &= \sqrt{\frac{-10^{\frac{\Delta L_{12}(k)}{10}} + c_{12}(k)10^{\frac{\Delta L_{12}(k)}{20}} + 1}{2(10^{\frac{\Delta L_{12}(k)}{10}} + 1)}} \\
b_1 &= \sqrt{\frac{(10^{\frac{\Delta L_{12}(k)}{10}} - c_{12}(k)10^{\frac{\Delta L_{12}(k)}{20}} + 1)p_{\tilde{s}}(k)}{2(10^{\frac{\Delta L_{12}(k)}{10}} + 1)p_{\tilde{s}_1}(k)}} \\
b_2 &= \sqrt{\frac{(10^{\frac{\Delta L_{12}(k)}{10}} - c_{12}(k)10^{\frac{\Delta L_{12}(k)}{20}} + 1)p_{\tilde{s}}(k)}{2(10^{\frac{\Delta L_{12}(k)}{10}} + 1)p_{\tilde{s}_2}(k)}} \; .
\end{aligned} \tag{4.8}$$

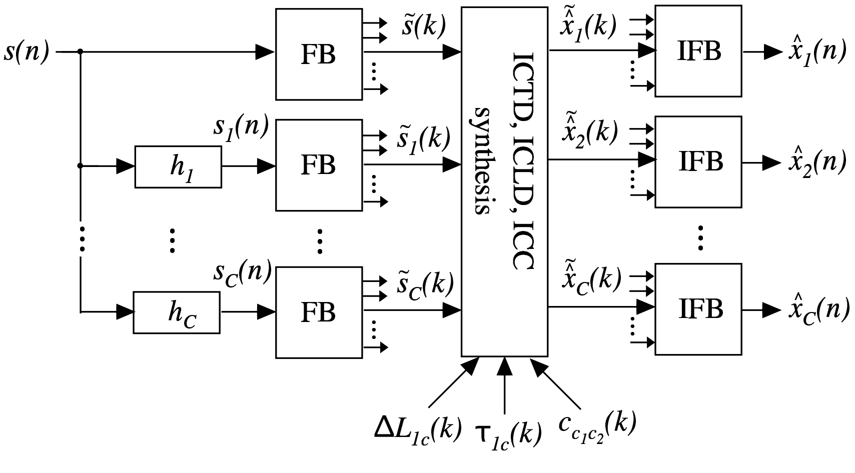


Figure 4.3: BCC synthesis scheme for generating multi-channel audio signals. ICTD, ICLD, and ICC synthesis is applied independently to each subband.

4.3.2 Multi-channel coherence synthesis

Similar to the stereo case, the output subband signals are computed as weighted sums of the subband signals of the sum signal and the late reverberation audio

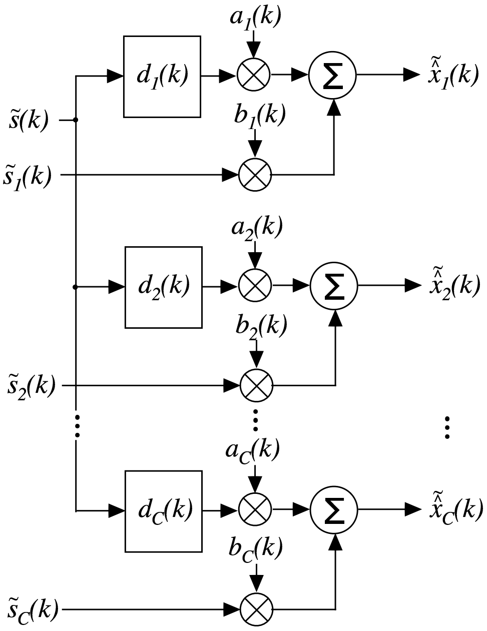


Figure 4.4: Processing of one subband for ICTD, ICLD, and ICC synthesis for multi-channel audio signals.

channels,

$$\begin{aligned}
\tilde{\hat{x}}_1(k) &= a_1\tilde{s}(k-d_1)+b_1\tilde{s}_1(k) \\
\tilde{\hat{x}}_2(k) &= a_2\tilde{s}(k-d_2)+b_2\tilde{s}_2(k) \\
&\vdots \\
\tilde{\hat{x}}_C(k) &= a_C\tilde{s}(k-d_C)+b_C\tilde{s}_C(k)\,,
\end{aligned} \tag{4.9}$$

as illustrated in Figure 4.4. The delays are determined by the ICTDs (3.9).

$2C$ equations are needed to determine the $2C$ scale factors in (4.9). In the following the conditions leading to these equations are briefly described.

ICLD: $C-1$ equations similar to (4.4) are formulated between the channel pairs such that the output subband signals have the desired ICLD cues.

ICC for the two strongest channels: Two equations similar to (4.5) and (4.7) between the two strongest audio channels, with indices c_1 and c_2, are formulated such that the ICC between these channels is the same as estimated in the encoder and the amount of late reverberation in both channels is the same, respectively.

Normalization: One equation is obtained by extending (4.6) to C channels,

$$\sum_{c=1}^{C} a_c^2 p_{\tilde{s}}(k) + \sum_{c=1}^{C} b_c^2 p_{\tilde{s}_c}(k) = p_{\tilde{s}}(k). \tag{4.10}$$

ICC for $C-2$ weakest channels: The ratio between the power of late reverberation to sum signal for the weakest $C-2$ channels ($c \neq c_1 \wedge c \neq c_2$) is chosen to be the same as for the second strongest channel c_2,

$$\frac{b_c^2 p_{\tilde{s}_c}(k)}{a_c^2 p_{\tilde{s}}(k)} = \frac{b_{c_2}^2 p_{\tilde{s}_{c_2}}(k)}{a_{c_2}^2 p_{\tilde{s}}(k)}\,, \tag{4.11}$$

resulting in another $C-2$ equations, for a total of $2C$ equations. The gain factors are the non-negative solutions of the equation system consisting of the described $2C$ equations.

4.4 Subjective evaluations

BCC and the spatial cue synthesis scheme described in this chapter was implemented with a short time Fourier transform (STFT) based filterbank as described in detail in Chapter 6. Spatial cue estimation and synthesis was carried out with a spectral resolution corresponding to subbands with bandwidths equal to two times the equivalent rectangular bandwidth (ERB) (Glasberg and Moore 1990). A Hann window of size 680 samples with 50 % overlap was used for the STFT, resulting in that for the used sampling frequency of 32 kHz ICTD, ICLD, and ICC are updated every 11 ms.

4.4.1 Stereo audio signals

Subjects and playback setup

Seven experienced subjects participated in the test, all with an age between 22 and 37 years. The test was conducted with *Stax SR-404 Signature* electrostatic headphones and a *Stax SRM Monitor* driver unit. A *Lake People DAC A-54* D/A converter was connected to an SGI workstation through its digital audio interface. The test was conducted in a sound insulated room.

Stimuli

The subjective evaluation was carried out using 14 music clips with a length of 12 s each. The clips were sampled at 32 kHz. Different kinds of music clips such as classical recordings, jazz, rock, and percussive music were selected. Each of these clips has a pronounced wide auditory spatial image and different degrees of ambience and reverberance.

Two signal types were generated by BCC: BCC without ICC synthesis and BCC with ICC synthesis. BCC without ICC synthesis was considered to assess the improvement of using ICC synthesis compared to not using ICC synthesis as was assessed in Section 3.5. The sum signal was not coded with an audio coder. The aim was to assess BCC without the influence of distortions resulting from coding of the sum signal.

Test method

A double-blind MUSHRA test (Rec. ITU-R BS.1534 2003) was used, comparing a number of degraded items to the reference. Besides the reference, the items with and without ICC synthesis were presented to the subjects in one panel.

Additionally, two anchor signals were used plus a hidden reference signal. The anchors were 3.5 kHz and 7 kHz lowpass filtered versions of the reference signal. Anchor signals are used in MUSHRA to obtain an indication how a coder compares to well-known audio quality levels. The subjects could listen to the reference and the other items as many times as they desired. Switching between the items was possible at any time. The subjects were instructed not only to pay attention to the absolute grading but also to the rank order of their judgments.

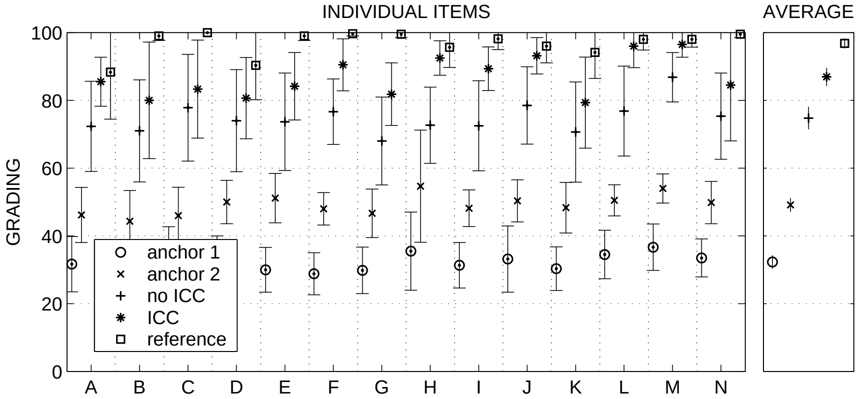


Figure 4.5: Subjective test results of the 14 test items. The gradings and 95 % confidence intervals of the two band-limited anchor signals, BCC without ICC synthesis, BCC with ICC synthesis, and hidden reference are shown for each item. The rightmost panel contains the gradings averaged over all items. The absolute quality grading scale used is: 80–100: “excellent”; 60–80: “good”; 40–60: “fair”; 20–40: “poor”; 0–20: “bad”.

Results

Figure 4.5 shows the results of the subjective test. The grading of each of the 14 items is shown in the left panel for each signal type (anchor signals, BCC without ICC synthesis, BCC with ICC synthesis, hidden reference). The gradings averaged over the 14 music items are shown in the right panel. BCC with ICC synthesis is better than BCC without ICC synthesis for each of the 14 items. BCC with ICC synthesis reaches on average a score of about 87 which corresponds to “excellent” quality, whereas BCC without ICC synthesis reaches on average about a score of 75 corresponding to “good” quality.

Discussion

ICC synthesis clearly improves the quality of BCC. The average quality that is reached by the proposed scheme is a MUSHRA score of about 87 corresponding to “excellent”. For some items (A, H, J, L, M) BCC reaches nearly transparent quality, i.e. the reference signal and BCC synthesized signal can hardly be distinguished. For other more critical items for BCC there is still a significant difference to the reference.

4.4.2 Multi-channel audio signals

Subjects and playback setup

The test was conducted in two different listening rooms with different equipment and subjects at different locations (EPFL Lausanne, Fraunhofer IIS):

1. Four adults with an age range of 22-29 participated as subjects in the listening tests at EPFL. Two subjects are experienced listeners and two are non-experienced. During the test, the subjects were sitting on a chair that was placed in the sweetspot of a standard 5.1 listening setup (Rec. ITU-R BS.775 1993) in a sound insulated room. The loudspeakers were placed on a circle with a radius of 2 m. Informal listening revealed that a relatively small scale loudspeaker setup is more critical than a larger scale loudspeaker setup. For audio playback an *Apple PowerBook G4* laptop computer was used with an external multi-channel D/A converter (*Emagic EMI A26*) directly connected to active loudspeakers (*Genelec 1031A*).

2. Five experienced adult listeners participated in the listening test at Fraunhofer IIS. During the test, the subjects were sitting on a chair that was placed in the sweetspot of a standard 5.1 listening setup in a sound insulated room. A personal computer with an *RME Hammerfall* digital sound output interface connected to *Lake People DAC F20* D/A converters was used. The D/A converters were directly connected to active loudspeakers (*Geithain RL 901*).

Stimuli

Different kinds of reference $5-$channel audio material was selected: Classical recordings mimicking a concert hall experience and movie soundtrack style items with auditory events occurring in all directions. We chose audio material that we consider critical for multi-channel BCC coding (e.g. applause). The reference items (R) were compared to two kinds of BCC synthesized items: BCC with ICC synthesis (A) and BCC without ICC synthesis (B). We included items B in the evaluation for assessing the improvement achieved when considering ICC over the case of not considering ICC. The sum signal was not coded to avoid affecting the test results due to coding artifacts.

Test methods

Test 1: A test was conducted to assess the quality of items A (BCC with ICC synthesis) relative to the reference items R. The test method used was the hidden reference method, used according to (Rec. ITU-R BS.1116.1 1997). The reference item is played, followed by the reference item and the degraded item in random order. A $5-$grade impairment scale was used for comparing the degraded item to the reference. After the three items were initially played, the listener could selectively listen to the items again while switching between the items at any time. This method is suitable for subjective assessment of small impairments. We decided to use this method, after informal listening revealed that for the considered items the degree of impairment is fairly small.

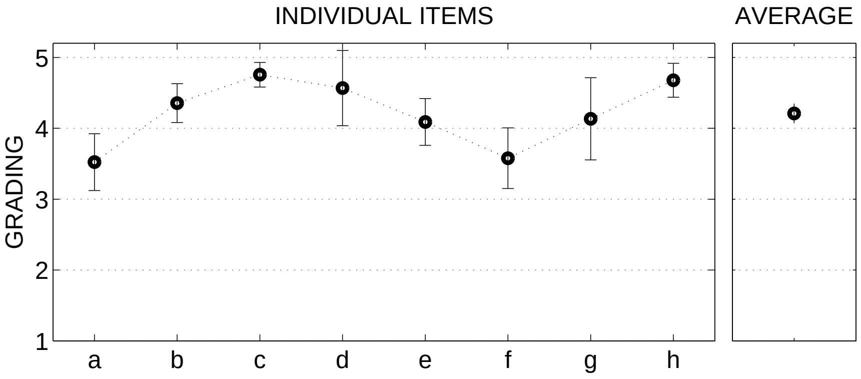


Figure 4.6: Test 1: Hidden reference test results. The test results averaged over the subjects and 95 % confidence intervals are shown for each item (left panel) and averaged for all items (right panel). (Grading scale judges difference between BCC with ICC synthesis and reference: 5: "not perceptible", 4: "perceptible but not annoying", 3: "slightly annoying", 2: "annoying", 1: "very annoying").

Test 2: Informal listening revealed that items B (BCC without ICC synthesis) were impaired significantly more than items A (BCC with ICC synthesis). Therefore, we decided not to compare these items with an absolute scale to the reference items, but to use a relative 7-grade comparison scale to compare items A and B, according to (Rec. ITU-R BS.562.3 1990). In this case, the reference item is played, followed by items A and B in random order. Similarly as in test 1, after the three items were initially played, the listener could selectively listen to the items again while switching between the items at any time.

Results

The results for both tests obtained from the two different locations (EPFL Lausanne, Fraunhofer IIS) were fairly similar and thus we averaged the results over all subjects at both locations.

Test 1: Figure 4.6 shows the results for the individual items averaged for all subjects and the overall average. The proposed scheme for ICTD, ICLD, and ICC synthesis has an overall grading between "perceptible but not annoying" and "imperceptible". The items with the best quality in Figure 4.6 (*b*, *c*, *d*, *h*) are a classical recording, movie soundtracks, and a scene with auditory events all around the subject. The most critical item is the applause signal (*a*). Item *e* also contains critical applause and a talker at the side. Item *f* is a classical recording with very tonal components, where the ICC synthesis introduces some distortions. Item *g* is a movie soundtrack signal.

Test 2: The results are shown in Figure 4.7 for the individual items averaged for all subjects and the overall average. The items with ICC synthesis are

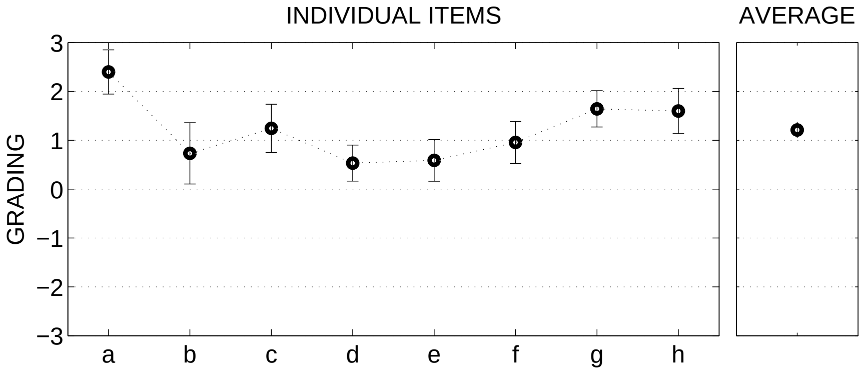


Figure 4.7: Test 2: 7-grade relative comparison scale test results. The test results averaged over the subjects and 95 % confidence intervals are shown for each item (left panel) and averaged for all items (right panel). (Positive values correspond to better performance of BCC with ICC compared to BCC without ICC synthesis: 1: “slightly better”, 2: “better”, 3: “much better”).

significantly better than the items without ICC synthesis. The proposed ICC synthesis gives an improvement for all items compared to no ICC synthesis. Interestingly, the worst item in Test 1 (item a, applause) is most improved by ICC synthesis.

4.5 Conclusions

In this chapter an ICC synthesis method was described based on generating a number of de-correlated audio channels by applying a late reverberation model. For each channel and subband of the output signal specific amounts of late reverberation are used for controlling ICC as a function of time and frequency. Furthermore, the spectral envelope of the original audio signal is imposed on the synthesized signal by scaling each subband such that at each time its power approximates the power of the original audio signal.

Two subjective tests were carried out applying the synthesis scheme proposed in this chapter for generating stereo and multi-channel audio signals. One of these tests was carried out with stereo signals and headphone playback and the other with five-channel surround signals and loudspeaker playback. The transmitted sum signal was not coded because the goal was to assess the quality of BCC disregarding possible distortions added by an external mono audio coder. The results of the subjective tests indicate that the proposed scheme achieves good quality for both headphone and multi-channel loudspeaker playback.

Chapter 5

Variations of Binaural Cue Coding (BCC)

5.1 Introduction

In this chapter, three variations of binaural cue coding (BCC) are described. Summaries and motivations of these schemes are:

Hybrid BCC: The results of the subjective tests presented in the previous chapter imply that the quality that can be achieved by only transmitting a single audio channel plus inter-channel time difference (ICTD), inter-channel level difference (ICLD), and inter-channel coherence (ICC) cues is good, but the original and synthesized signals are usually distinguishable. In Section 5.2, we are describing a hybrid scheme (Baumgarte *et al.* 2004) which applies at lower frequencies a conventional audio coder and at higher frequencies BCC with a mono signal transmission. It will be shown that at the expense of additional bitrate (compared to fullband BCC) the proposed hybrid scheme can achieve quality up to transparent quality.

C-to-E BCC: In Section 5.3 a variation of BCC is described which represents C audio channels not as one single (transmitted) channel, but as E channels. There are two motivations for C-to-E BCC:

1. Regular BCC provides a backwards compatible path for upgrading existing mono systems for stereo or multi-channel audio playback. The upgraded systems transmit the BCC downmixed sum signal through the existing mono infrastructure, while additionally transmitting the BCC side information. C-to-E BCC is applicable to E-channel backwards compatible coding of C-channel audio.
2. While Hybrid BCC introduces scalability in terms of applying BCC not at all frequencies, C-to-E BCC introduces scalability in terms of different degrees of reduction of the number of transmitted channels. It is expected that the more audio channels are transmitted the better the audio quality will be.

BCC for flexible rendering: This scheme is not applied for coding of stereo and multi-channel audio signals. A number of independent source signals (e.g. separately recorded instruments) are represented as a single audio channel plus side information different from regular BCC. The decoder can freely generate binaural, stereo, or multi-channel audio signals with an auditory spatial image as desired. In other words, a number of audio signals are transmitted jointly as one audio channel, while still being able to mix them at the decoder as if these signals were transmitted separately.

5.2 Hybrid BCC

The experimental results of Section 3.5.2, comparing a BCC-based coder with conventional stereo audio coders such as PAC (Sinha *et al.* 1997) and MP3 (Brandenburg and Stoll 1994, ISO/IEC 1993) in terms of subjective quality versus bitrate, are illustrated in Figure 5.1. These results suggest that a BCC-based coder yields higher quality than a conventional stereo audio coder for bitrates below roughly 70 kb/s. Above that bitrate a conventional audio coder can achieve better quality because a BCC-based coder is limited by some auditory spatial image distortions that can not be avoided by increasing the bitrate. (Note that the estimation of the threshold of 70 kb/s is based on BCC without ICC synthesis, BCC with ICC synthesis is likely to result in a higher threshold).

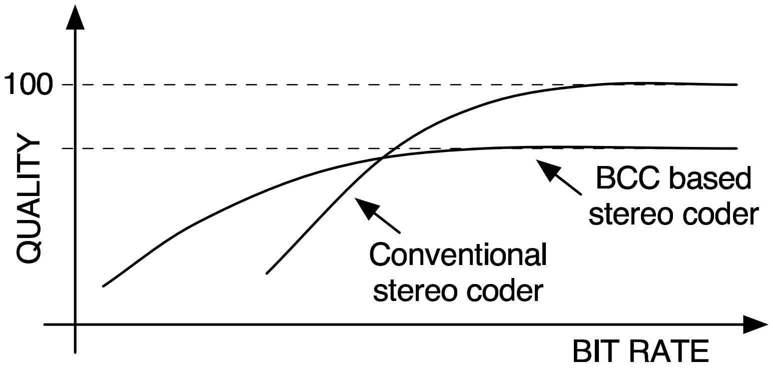


Figure 5.1: Schematic performance characteristics of a conventional stereo coder and a BCC-based coder (derived from data in Section 3.5.2). A quality value of 100 corresponds to transparent quality (no degradations).

Given the performance curves of both coders, it is desirable to combine them in such a way that a smooth transition occurs from a BCC-based coder to a conventional audio coder when the bitrate is increased beyond the crossover point in Figure 5.1. Such a hybrid coder would provide highly efficient multi-channel coding in a large bitrate range, i.e. from very low bitrate coding to transparent coding.

5.2.1 Encoder and decoder processing

Figure 5.2 shows the described hybrid BCC scheme. Each audio input channel is decomposed into two subbands using a filterbank with two bands with a cut-off frequency denoted f_0. The low frequency subband is not processed.

The high frequency subband is processed with a BCC encoder. The sum signal output of the BCC encoder and the non-processed low frequency subbands are combined resulting in the C-channel output signal $y_c(n)$ $(1 \leq c \leq C)$ which is transmitted, denoted *hybrid signal*. The hybrid signal has the property that at high frequencies it contains effectively only one audio channel.

As can be seen in Figure 5.2, the decoding process is as follows. Each channel of the transmitted hybrid signal $y_c(n)$ $(1 \leq c \leq C)$ is decomposed by the same filterbank as used in the encoder into a low and high frequency subband. The low frequency subband corresponds to the original low frequency part of the audio channels and thus does not need to be further processed. The high frequency subband of channel one is processed with a BCC decoder to generate multi-channel high frequency subbands. Then, the low and high frequency subbands are combined with the inverse filterbank resulting in the hybrid BCC decoder output multi-channel signal.

By means of choosing the cut-off frequency f_0 between the two subbands one can choose the ratio between discrete multi-channel and BCC multi-channel. As f_0 approaches the Nyquist frequency, the scheme shown in Figure 5.2 converges to transparent audio quality (when the transmitted audio channels are not coded or transparently coded). For $f_0 = 0$ the scheme corresponds to a fullband BCC scheme. Thus this scheme provides scalable audio quality between the quality of a fullband BCC scheme and transparent audio quality.

Now it will be explained that when a conventional multi-channel audio coder is used for coding of the hybrid signal the bitrate will scale from the bitrate of coding one audio channel to the bitrate of coding C audio channels.

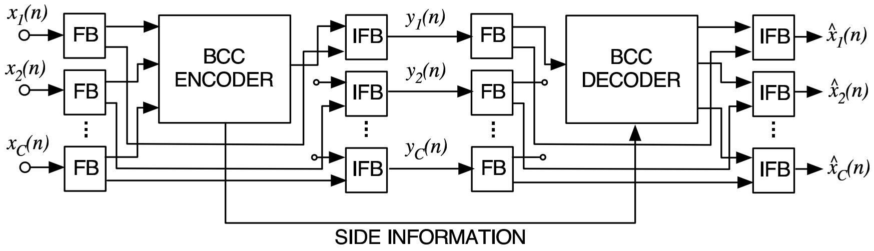


Figure 5.2: Hybrid BCC coder. Two band filterbanks (FB) and corresponding inverse filterbanks (IFB) are used such that only frequencies higher than a certain threshold f_0 are processed by BCC. For lower frequencies the discrete audio channels are transmitted.

A conventional stereo or multi-channel audio coder such as PAC (Sinha *et al.* 1997) or MPEG-2 AAC (Stoll 1996, Bosi *et al.* 1997, ISO/IEC 1997) can be used for coding the hybrid signal without any modifications. Such a coder takes advantage of the downmixed part of the upper spectrum by effectively only coding the upper spectrum of channel one. The other channels have no power in the upper spectrum and therefore effectively no spectral coefficients are coded. That is, because usually audio coders feature a mechanism such that a multitude of coefficients that are zero are not coded and transmitted (Sinha *et al.* 1997, Stoll 1996, Bosi *et al.* 1997, ISO/IEC 1997). Therefore, the larger f_0 is chosen the less spectral coefficients are effectively coded and transmitted, and the lower is the bitrate. Additional bitrate savings of a few

Table 5.1: The bitrate required for transparent coding of the hybrid signals with different cut-off frequencies.

cut-off frequency	*bitrate*	*relative bitrate*
16000 Hz	100.8 kb/s	100%
6000 Hz	84.3 kb/s	84%
2000 Hz	74.2 kb/s	74%
1000 Hz	67.8 kb/s	68%
0 Hz	58.9 kb/s	59%

kb/s are possible by making explicit use of the knowledge that only one upper spectrum needs to be coded.

5.2.2 Evaluations

Hybrid BCC was assessed in different ways. First, the bitrate necessary for transparent waveform quality was estimated for different cut-off frequencies. Then, a subjective test was conducted for assessing the quality of hybrid BCC for different cut-off frequencies.

Both, bitrate analysis and subjective evaluation, were carried out using the same 14 music clips with a length of 12 s each. The clips were sampled at 32 kHz. Each of these clips has a pronounced wide spatial image and different degrees of ambience and reverberance. BCC is challenged by a wide spatial image in the sense that it needs to perceptually separate audio sources. Also for a conventional stereo audio coder a wide spatial image is challenging because the redundancy between the channels is small in that case resulting in a high bit demand. Different kinds of music signals such as classical recordings, jazz, rock, and percussive music were selected.

Bitrate analysis

We implemented the "hybrid coder" by combining hybrid BCC with PAC. PAC was configured such that the quantization noise would always be just below the masked threshold, i.e. the coder was granted as many bits as it needed for transparent coding (this is often called "variable bitrate coding"). Furthermore, the audio bandwidth was always set to 16 kHz. The described coding configuration is suitable for estimating the bitrate reduction of the hybrid coder as a function of the cut-off frequency. The so-obtained bitrate is related to the perceptual entropy (Johnston 1988) of the hybrid signal (stereo at lower frequencies and mono at higher frequencies) and thus gives the fundamental degree of bitrate saving achieved in principle by any similar hybrid system. Table 5.1 shows the bitrates for different cut-off frequencies in kb/s and as percentage of the bitrate of fullband stereo. Note that a cut-off frequency of 16 kHz corresponds to fullband stereo.

Subjective evaluation

Subjects and playback setup Seven experienced subjects participated in the test, all with an age between 22 and 37 years. The test was conducted with *Stax*

SR-404 Signature electrostatic headphones and a *Stax SRM Monitor* driver unit. A *Lake People DAC A-54* D/A converter was connected to an SGI workstation through its digital audio interface.

Stimuli BCC as described in Chapter 3 and the spatial cue synthesis scheme described in Chapter 4 were used for the test. The algorithms were implemented with a short time Fourier transform (STFT) based filterbank as described in detail in Chapter 6. Spatial cue estimation and synthesis was carried out with a spectral resolution corresponding to subbands with bandwidths equal to two times the equivalent rectangular bandwidth (ERB) (Glasberg and Moore 1990). A Hann window of size 680 samples with 50 % overlap was used for the STFT, resulting in that ICTD, ICLD, and ICC are updated every 11 ms.

The hybrid signal was not coded by an audio coder. Hybrid cut-off frequencies as shown in Table 5.1 were used. Note that 0 Hz corresponds to regular fullband BCC and 16000 Hz corresponds to fullband stereo (= reference signal with no degradations).

Test method A double-blind MUSHRA test (Rec. ITU-R BS.1534 2003) was used, comparing a number of degraded items to the reference. Besides the reference, the items with the 5 different hybrid cut-off frequencies f_0 were presented to the subjects in one panel (note that one of these is equal to the reference). Additionally, two anchor signals were used. These were 3.5 kHz and 7 kHz lowpass filtered versions of the reference signal. Anchor signals are used in MUSHRA to obtain an indication how a coder compares to well-known audio quality levels. The subjects could listen to the reference and the other items as many times as they desired. Switching between the items was possible at any time. The subjects were instructed not only to pay attention to the absolute grading but also to the rank order of their judgments.

Results Figure 5.3 shows the results of the subjective test. The grading and 95% confidence interval of each of the 14 items is shown in each panel. The different panels show the results for the two anchor signals and the hybrid items with different cut-off frequencies (f_0 = 0, 1000, 2000, 6000, 16000 Hz). The gradings averaged over the 14 music items are shown in the right panel. Note that as expected, the average quality increases as the hybrid cut-off frequency increases. The fullband BCC item ($f_0 = 0$ Hz) has an average grading of about 87. The item with $f_0 = 6000$ Hz is nearly as good as the hidden reference item ($f_0 = 16000$Hz).

Discussion

The subjective test results of Figure 5.3 show that the average quality of hybrid BCC improves with increasing cut-off frequency, as expected. The quality increase is considerably larger at a low cut-off frequency in comparison with higher ones for the same frequency increment. For instance, the quality increase obtained from changing f_0 from 0 to 1 kHz is larger than that obtained by changing the cut-off frequency from 1 to 16 kHz or for any step in-between.

The bitrate analysis results of Table 5.1 show that there is no proportional relationship between the amount of bitrate increase and the quality increase in

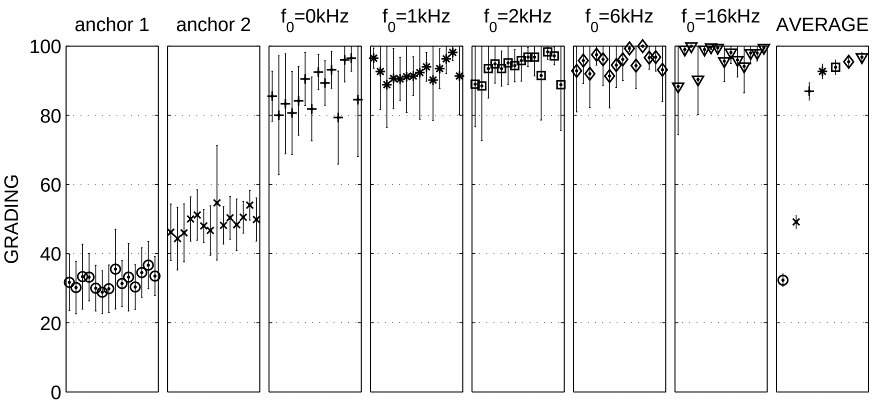


Figure 5.3: Subjective test results of the 14 test items. The two left panels show the gradings and 95% confidence intervals of the two band-limited anchor signals. The following 5 panels show the gradings and 95% confidence intervals for the BCC-coded items for different cut-off frequencies f_0. The rightmost panel contains the gradings averaged over all items. The absolute quality grading scale used is: 80–100: "excellent"; 60–80: "good"; 40–60: "fair"; 20–40: "poor"; 0–20: "bad".

Figure 5.3. For instance, the bitrate increases from 59% to 68% when changing f_0 from 0 to 1 kHz. When f_0 is changed from 6 to 16 kHz, the bitrate increases from 84% to 100%. However, the corresponding quality increase is much smaller in the latter case than in the first case. This relation indicates that hybrid BCC with a low cut-off frequency, e.g. 1 kHz, provides a reasonable trade-off between bitrate and quality which enhances the quality of fullband BCC. Further increased cut-off frequencies provide a continuous transition toward conventional transparent stereo audio coding ($f_0 = 16$ kHz).

The results confirm that the "hard" upper quality limit of fullband BCC as illustrated in Figure 5.1 can be overcome by hybrid BCC. This follows from the observed gradual quality increase with increasing cut-off frequency f_0 beyond the quality of fullband BCC. As illustrated in Figure 5.4, the hybrid coder operates at bitrates in the range where neither fullband BCC nor a conventional audio coder operates optimally.

Although not specifically tested, we observed in experiments with hybrid BCC and PAC that the hybrid coder provides better quality than fullband BCC or PAC in a bitrate range around the cross-over point in Figure 5.1. Therefore, we are confident that the audio quality achieved by the hybrid coder is considerably better than the quality of each of the single coders as illustrated in Figure 5.4. A subjective test with items including an audio coder was not performed in this study to avoid any dependencies on the specific audio coder selected. Moreover, the quality optimization is a complex trade-off between bitrate, audio bandwidth, auditory spatial image quality, and audio coding artifacts. It is hard to justify the different parameter settings for the different coders without conducting extensive subjective tests.

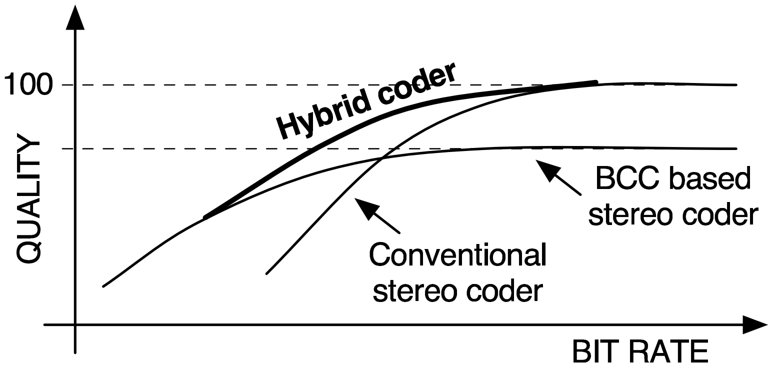


Figure 5.4: The hybrid coder is used in the range of bitrates, where neither BCC nor conventional audio coders operate optimally.

5.3 C-to-E BCC

A BCC scheme with multiple audio transmission channels is shown in Figure 5.5. In the encoder, the C input channels are downmixed to the E transmitted audio channels. ICTD, ICLD, and ICC between certain pairs of input channels are estimated as a function of time and frequency. The estimated cues are transmitted to the decoder as side information. A BCC scheme with C input channels and E transmission channels is denoted C-to-E BCC.

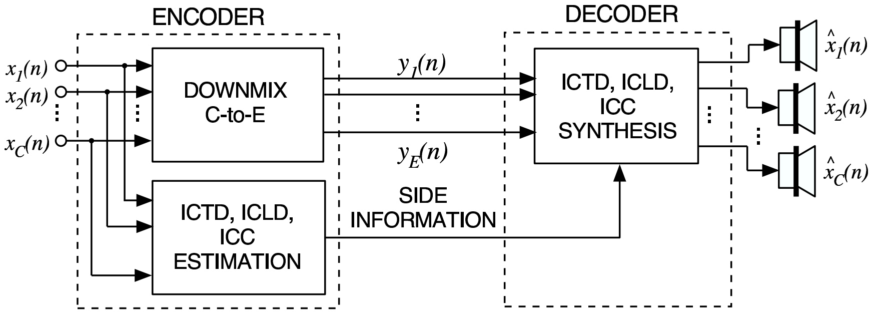


Figure 5.5: Generic BCC scheme with multiple transmission channels. The C input channels are downmixed to E channels and transmitted to the decoder together with side information.

One application for regular BCC with its single transmitted audio channel is backwards compatible extension of existing mono systems for stereo or multichannel audio playback. Since the transmitted single audio channel is a valid mono signal, it is suitable for playback by the legacy receivers. Similarly, C-to-E BCC can be used for backwards compatible extension of existing E-channel systems to C-channel systems.

Most of the installed audio broadcasting infrastructure (analog and digital radio, television, etc.) and audio storage systems (vinyl discs, compact cassette, compact disc, VHS video, MP3 sound storage, etc.) are based on two-channel stereo. On the other hand, "home theater systems" conforming to the 5.1 standard (Rec. ITU-R BS.775 1993) (Section 2.3.3) are becoming more popular. Thus, BCC with two transmission channels (C-to-2 BCC) is particularly interesting for extending existing stereo systems for multi-channel surround.

Another application for C-to-E BCC, interesting in the longer term, may be to extend the 5.1 surround standard (e.g. audio on DVD video) or surround on movie theater media to support more audio channels. Legacy home theater systems or legacy movie theaters would still be able to play back the audio while a new generation of systems may support more independent loudspeakers.

In the analog domain, matrixing algorithms such as "Dolby Surround", "Dolby Pro Logic", and "Dolby Pro Logic II" (Hull 1999, Dressler 2000) have been popular for years. Such algorithms apply "matrixing" for mapping the 5.1 audio channels to a stereo compatible channel pair. However, matrixing algorithms only provide significantly reduced flexibility and quality compared to discrete audio channels (Herre *et al.* 2004). If limitations of matrixing algorithms are already considered when mixing audio signals for 5.1 surround, some improvements can be achieved (Hilson 2004) (compared to the case when such limitations are not considered).

C-to-E BCC can be viewed as a scheme with similar functionality as a matrixing algorithm. But it is more general since it supports mapping from any number of channels to any number of channels. C-to-E BCC is intended for the digital domain and its low bitrate additional side information usually can be included into the existing data transmission in a backwards compatible way (i.e. legacy receivers will ignore the additional side information and play back the E transmitted channels directly). The goal is to achieve audio quality similar to discrete channels, i.e. significantly better quality than what can be expected from a conventional matrixing algorithm.

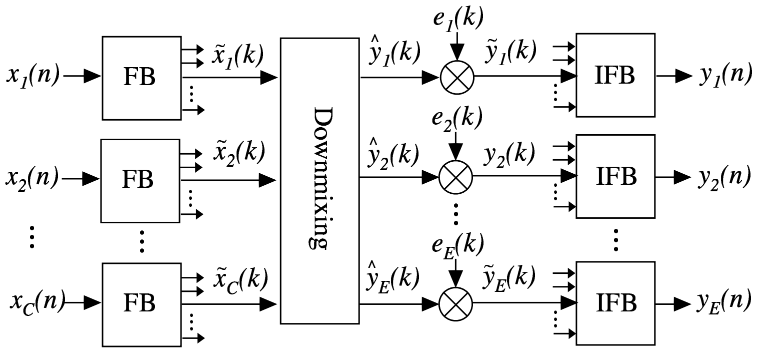


Figure 5.6: Downmixing with equalization. The downmixing is applied in subbands and equalization is carried out by scaling the subbands. The shown processing is carried out independently for each subband.

5.3.1 Encoder processing

Similarly as downmixing in regular BCC (Section 3.3.2), downmixing for C-to-E BCC is also carried out in the subband domain as illustrated in Figure 5.6. The E downmixed subbands are generated by

$$\begin{bmatrix} \hat{y}_1(n) \\ \hat{y}_2(n) \\ \vdots \\ \hat{y}_E(n) \end{bmatrix} = \mathbf{D}_{CE} \begin{bmatrix} \tilde{x}_1(n) \\ \tilde{x}_2(n) \\ \vdots \\ \tilde{x}_C(n) \end{bmatrix} , \tag{5.1}$$

where the real-valued C-by-E matrix $\mathbf{D}_{CE}$ is denoted downmixing matrix.

As illustrated in Figure 5.6, the downmixing is followed by scaling. The motivation of this is the same as for the equalization for computing the regular BCC sum signal (Section 3.3.2), but generalized for downmixing with arbitrary weighting factors for each channel. If the input channels are independent, then the power of the downmixed signal in each subband $p_{\tilde{y}_i}(k)$ is equal to

$$\begin{bmatrix} p_{\tilde{y}_1}(k) \\ p_{\tilde{y}_2}(k) \\ \vdots \\ p_{\tilde{y}_E}(k) \end{bmatrix} = \bar{\mathbf{D}}_{CD} \begin{bmatrix} p_{\tilde{x}_1}(k) \\ p_{\tilde{x}_2}(k) \\ \vdots \\ p_{\tilde{x}_C}(k) \end{bmatrix} , \tag{5.2}$$

where $p_{\tilde{x}_i}(k)$ is the power in a subband of the input signal $x_i(n)$ and $\bar{\mathbf{D}}_{CD}$ is the downmixing matrix with each matrix element squared. If the subbands are not independent, the power values of the downmixed signal $p_{\tilde{y}_i}(k)$ will be larger or smaller than as computed by (5.2), due to signal cancellations and amplifications when signal components are in-phase and out-of-phase, respectively. To prevent this, the downmixing matrix is applied in subbands followed by a scaling operation as illustrated in Figure 5.6. The scaling factors $e_i(k)$ $(1 \leq i \leq E)$ are chosen to be

$$e_i(k) = \sqrt{\frac{p_{\tilde{y}_i}(k)}{p_{\hat{y}_i}(k)}} , \tag{5.3}$$

where $p_{\tilde{y}_i}(k)$ is the subband power as computed by (5.2) and $p_{\hat{y}_i}(k)$ is the power of the corresponding downmixed subband signal $\hat{y}_i(k)$.

ICTD, ICLD, and ICC are estimated similarly as in regular BCC (Section 3.3.4), however not necessarily between all signal channels. Specific examples between which channels to estimate the cues are given in Section 5.3.3.

5.3.2 Decoder processing

The decoder processes the transmitted E audio channels to generate its C output channels, considering how the encoder downmix was carried out and the transmitted cues. Figure 5.7 illustrates how the C audio output channels are generated given the E transmitted channels. The input channels are converted to the subband domain. Upmixing is applied to generate C subband signals given the E subband signals. The upmixed C subband signals are scaled and delayed such that the desired ICTD and ICLD appear between pairs of channels. Block A in Figure 5.7 is a generic scheme for ICC synthesis, e.g. as described in Section 3.3.5 or Chapter 4. Note that C-to-E BCC synthesis is very similar to regular BCC synthesis (Figure 3.7). The difference is that for each output channel a different *base channel*, as generated by the upmixing, is used prior to applying processing for ICTD, ICLD, and ICC synthesis. These base channels are linear combinations of the transmitted channels,

$$\begin{bmatrix} \tilde{s}_1(n) \\ \tilde{s}_2(n) \\ \vdots \\ \tilde{s}_E(n) \end{bmatrix} = \mathbf{U}_{EC} \begin{bmatrix} \tilde{y}_1(n) \\ \tilde{y}_2(n) \\ \vdots \\ \tilde{y}_C(n) \end{bmatrix}, \tag{5.4}$$

where the real-valued E-by-C matrix $\mathbf{U}_{EC}$ is denoted upmixing matrix. Note that the upmixing (5.4) is applied in subbands. This has the advantage that as opposed to C filterbanks only E filterbanks have to be used. Additionally, one may apply “dynamic upmixing” individually in each subband as is discussed later.

The synthesis of ICLD is relatively unproblematic compared to synthesis of ICTD and ICC, since it involves merely scaling of subband signals. Furthermore, ICLD cues are the most commonly used directional cues (amplitude panning, coincident-pair microphones) and thus it is important that ICLD cues approximate those of the original signal. Thus unless some audio channels are transmitted unmodified, ICLD are estimated between all channel pairs (Section 3.3.4). Similarly as in ICLD synthesis for regular BCC (Section 3.3.5), the scaling factors $a_c(k)$ $(1 \leq c \leq C)$ for each subband are chosen such that the subband power of each output channel approximates the corresponding power of the original audio signal.

Additionally, the goal is to apply less signal modifications for synthesizing ICTD and ICC than would be required in regular BCC. For this purpose the scheme considers ICTD and ICC which are present between the transmitted channels and synthesizes ICTD and ICC cues only between certain output channel pairs.

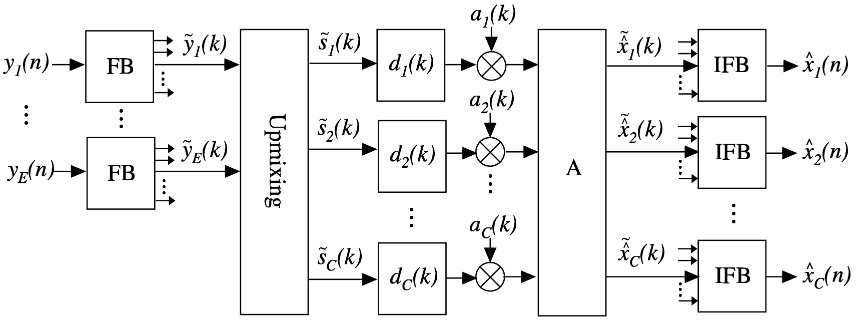


Figure 5.7: BCC synthesis applied to the E transmitted audio channels. The transmitted channels are converted to subbands. The given E subbands are upmixed to C subbands, followed by delays, scaling, and other processing (block A) for ICTD, ICLD, and ICC synthesis, respectively. The shown processing is carried out independently for each subband.

5.3.3 Specific examples for C-to-E BCC schemes

5-to-2 BCC

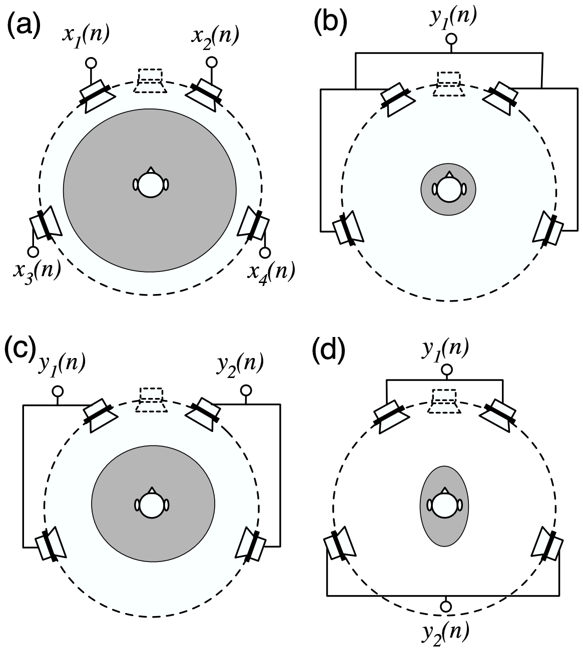


Figure 5.8: Perception of wideband noise signals: (a): Four independent signals for all four speakers. (b): Same signal for all four speakers. (c): Two independent signals for left two speakers and right two speakers. (d): Two independent signals for front two speakers and rear two speakers. The gray area illustrates the perceived auditory events in the scenarios (a)-(d) in a reverberant room.

A simple experiment is described for motivating the choice of the specific downmixing and upmixing matrices. A four loudspeaker setup is considered with the front loudspeakers at ±30° and the rear loudspeakers at ±110° (standard 5.1 setup without center loudspeaker and without subwoofer for low frequency effects).

In the following, "scenario (a), (b), (c), and (d)" denote the four parts of Figure 5.8. In scenario (a), four independent Gaussian noise signals are played back from the left, right, rear left, and rear right loudspeakers. In this scenario, the auditory event is surrounding the listener. This is the reference scenario with a maximum degree of listener envelopment and compared to the other scenarios resulting in the smallest interaural coherence (IC) values. A single Gaussian noise signal is played back from all loudspeakers in scenario (b), resulting in a minimum degree of listener envelopment and the largest IC values. Assuming free-field and left/right symmetry of the loudspeaker setup and listener's head, the ear input signals for this scenario are identical, i.e. IC $= 1$.

Scenarios (c) and (d) correspond to two ways of reducing the four independent channels of scenario (a) to two independent channels given to the four loudspeakers. Scenarios (c) and (d) play back two independent Gaussian noise signals through the left two and right two loudspeakers and through the front two and back two loudspeakers, respectively. Again, free-field and left/right symmetry of loudspeaker setup and listener's head is assumed. In scenario (c), the resulting ear input signals are not identical and IC < 1. For scenario (d), the ear input signals are identical as in scenario (b), i.e. IC $= 1$. Thus, scenario (c) is mimicking the reference scenario better than scenario (b).

It is expected, that also in reverberant rooms scenario (c) performs better than scenario (d). Informal listening experiments indicate that this is indeed the case. The gray areas in Figures 5.8(a)-(d) conceptually illustrate the extent of the corresponding auditory events.

The previous discussion implies the following rules for reducing the number of independent channels:

- Independence between signals of loudspeakers with different left/right positions should be maintained, i.e. ICC and ICTD cues are important in this case.
- Signals of loudspeakers with different front/rear positions can be coherent while IC cues are still low as long as left/right ICC is low.

5-to-2 BCC for stereo backwards compatible coding of 5-channel surround transmits different audio channels for different left/right positions such that independence of audio channels with different left/right positions is maintained.

One transmitted channel is computed from right, center, and rear right and the other from left, center, and rear left. Given the channel assignment indicated in Figure 5.9(a), this corresponds to a downmixing matrix of

$$\mathbf{D}_{52} = \begin{bmatrix} 1 & 0 & \frac{1}{\sqrt{2}} & 1 & 0 \\ 0 & 1 & \frac{1}{\sqrt{2}} & 0 & 1 \end{bmatrix}, \tag{5.5}$$

where the scale factors are chosen such that the sum of the square of the values in each column is one, resulting in that the power of each input signal contributes equally to the downmixed signals. The corresponding upmixing matrix copies each transmitted channel to the channels which were used for the corresponding downmixes,

$$\mathbf{U}_{25} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{5.6}$$

The scaling of the rows in upmixing matrices is not relevant, since the upmixed signals are normalized and re-scaled during ICLD synthesis.

Figure 5.9(a) illustrates the downmixing of the five input channels to the two transmitted channels. The upmixing is illustrated in Figure 5.9(b), where the left transmitted channel is used as base channel for left and rear left, the right transmitted channel as base channel for right and rear right, and the sum of both transmitted channels as base channel for the center channel. The

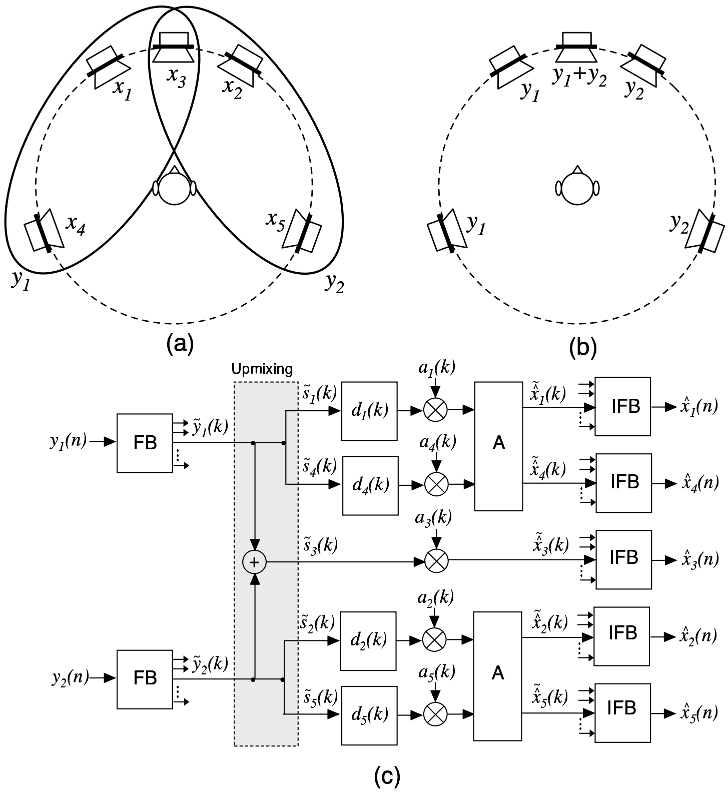


Figure 5.9: 5-to-2 BCC: (a) Downmixing to two channels, (b) computation of base channels for each output channel (upmixing), (c) synthesis of 5 channels given the 2 transmitted channels.

process of generating the 5-channel output signal, given the two transmitted channels, is shown in Figure 5.9(c). Note that ICTD and ICC synthesis is applied between the channel pairs for which the same base channel is used, i.e. between left and rear left, and right and rear right. The two blocks A in Figure 5.9(c) are schemes for 2-channel ICC synthesis.

The side information, estimated at the encoder, which is necessary for computing all parameters for the decoder output signal synthesis are the following cues: ΔL_{12}, ΔL_{13}, ΔL_{14}, ΔL_{15}, τ_{14}, τ_{25}, c_{14}, and c_{25}. (Different level differences could be used. The condition is just that enough information is available at the decoder for computing the scale factors, delays, and parameters for ICC synthesis).

6-to-5 BCC

Figure 5.10 illustrates the different processing steps in a 6-to-5 BCC scheme, i.e. a scheme that can be used for 5-channel backwards compatible coding of 6-channel surround. A 6-channel surround system with an additional rear center channel is considered. Such a loudspeaker setup is used in "Dolby Digital - Surround EX" (Rumsey 2001). Downmixing as illustrated in Figure 5.10(a) is used, corresponding to a downmixing matrix of

$$\mathbf{D}_{65} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & 0 & 0 & 1 & \frac{1}{\sqrt{2}} \end{bmatrix} , \tag{5.7}$$

where the front channels are transmitted non-modified and the rear three channels are downmixed to two channels for a total of 5 transmitted channels.

The upmixing, i.e. choice of base channels for BCC synthesis, is illustrated in Figure 5.10(b). In this case, all base channels are different audio channels and no ICTD and ICC synthesis is applied as is illustrated in the 6-to-5 BCC synthesis scheme in Figure 5.10(c). This choice of base channels corresponds to an upmixing matrix of

$$\mathbf{U}_{56} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} . \tag{5.8}$$

Note that channels 1, 2, and 3 are used unmodified. Thus no filterbank is used and these channels are just delayed for compensating the filterbank delay of the other channels, as indicated in Figure 5.10(c). The necessary side information for this case is: ΔL_{46} and ΔL_{47}.

Another possibility to downmix the 6 input audio channels would be to add left and rear left and right and rear right and leave the other channels unmodified. In this case, the synthesis scheme would apply ICTD and ICC synthesis between left and rear left and right and rear right. This way of downmixing and upmixing corresponds to giving more emphasis to left/right independence, whereas (5.7) and (5.8) give more emphasis to front/back independence.

7-to-5 BCC

This scheme transmits a 7-channel surround signal over 5 audio channels. For brevity, the downmixing and upmixing matrices are not explicitly written down. The "Lexicon Logic 7" surround matrix process uses 7 main loudspeakers approximately placed as illustrated in Figure 5.11(a) (Rumsey 2001). 7-to-5 BCC is applied for providing 7 independent channels with such a loudspeaker setup backwards compatibly to 5-channel surround. Figure 5.11(a) illustrates the downmixing that is used for this purpose. The two rear left and the two rear right channels are downmixed, while transmitting the other channels unmodified, for a total of 5 transmitted audio channels. The upmixing is illustrated in Figure 5.11(b). For synthesis of the 7 output channels, ICTD, ICLD, and ICC synthesis is applied only between the two rear side audio channel pairs, as illustrated in Figure 5.11. In this case, the transmitted side information is: ΔL_{46}, ΔL_{56}, τ_{46}, τ_{56}, c_{46}, and c_{56}.

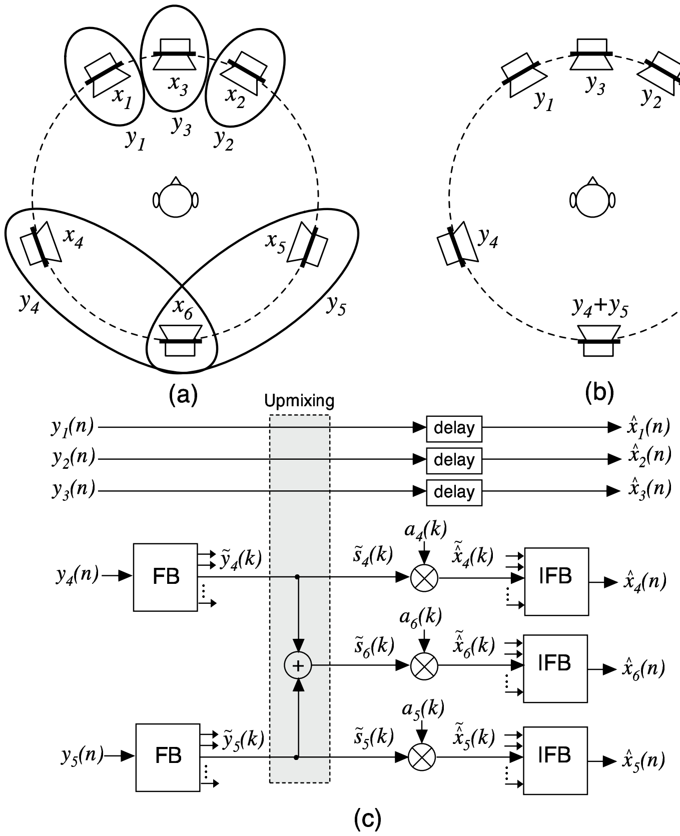


Figure 5.10: 6-to-5 BCC: (a) Downmixing to six channels, (b) computation of base channels for each output channel (upmixing), (c) synthesis of 6 channels given the 5 transmitted channels.

Low frequency effects (LFE) channels

In the 5-to-2 BCC example given, the low frequency effects (LFE) channel is treated similarly as the center channel, i.e. it is attenuated by 3 dB and added to both transmitted channels. The base channel for LFE channel synthesis at the decoder is the sum of both transmitted channels, also similarly as for the center channel. Processing for the LFE channel may only be applied at low frequencies as is described in Section 3.4.

In the other examples given, 6-to-5 and 7-to-5 BCC, the LFE channel is transmitted as the 5.1 LFE channel (assuming 5.1 backwards compatible coding) and no additional processing is necessary.

For 5.1 backwards compatible coding of surround formats with more than one LFE channel, e.g. 10.2 surround (Rumsey 2001), all the LFE channels are added and transmitted as the 5.1 LFE channel. At the decoder processing is applied for generating the multiple LFE channels, similarly as the generation of $\hat{x}_4$ and $\hat{x}_6$ (or $\hat{x}_5$ and $\hat{x}_7$) in the 7-to-5 BCC example (Figure 5.11).

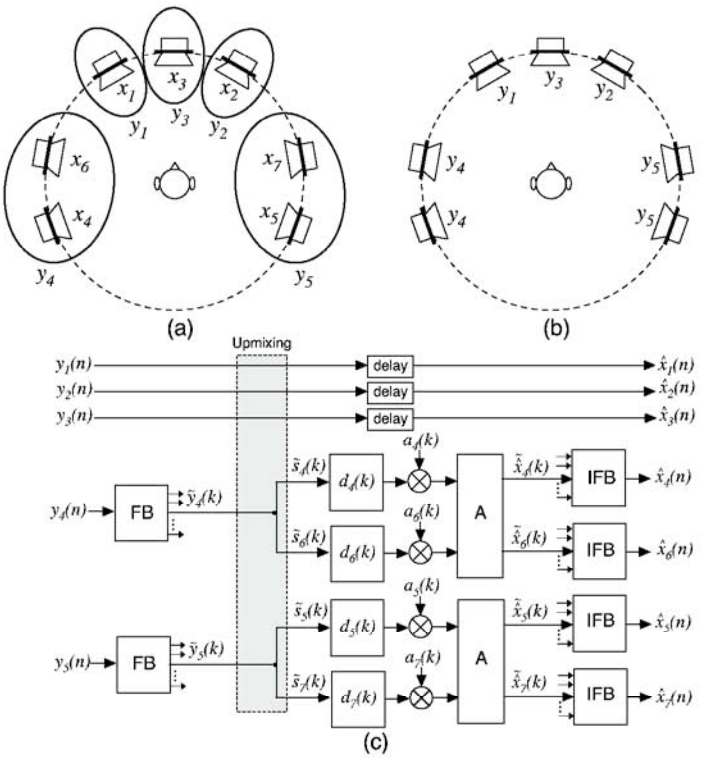


Figure 5.11: 7-to-5 BCC: (a) Downmixing to five channels, (b) computation of base channels for each output channel (upmixing), (c) synthesis of 7 channels given the 5 transmitted channels.

Dynamic upmixing

Informal listening revealed that 5-to-2 BCC with processing as illustrated in Figure 5.9 suffers from a certain reduction of overall width of the auditory spatial image for certain signals (e.g. applause signal). In the following, we are describing a technique for improving the auditory spatial image width for such signals. The technique is applicable to all cases when an input channel is mixed into more than one of the transmitted channels (e.g. the previous examples of 5-to-2 and 6-to-5 BCC). For simplicity of the discussion, the technique is described only for the specific case of 5-to-2 BCC.

The before mentioned problem of auditory spatial image width reduction occurs mostly for audio signals which contain independent fast repeating transients from different directions (e.g. applause signal). The image width reduction may be caused by insufficient time resolution of ICLD synthesis (for more discussion on the topic of BCC synthesis and time resolution refer to Section 7.5.2). As opposed to using a higher time resolution in this case, we aim at removing the center channel signal component from the side channels.

According to Figure 5.9 and Equations (5.5) and (5.6), the base channels for the 5 output channels of 5-to-2 BCC are:

$$\begin{aligned}
\tilde{s}_1(k) &= \tilde{y}_1(k) = \tilde{x}_1(k) + \frac{\tilde{x}_3(k)}{\sqrt{2}} + \tilde{x}_4(k) \\
\tilde{s}_2(k) &= \tilde{y}_2(k) = \tilde{x}_2(k) + \frac{\tilde{x}_3(k)}{\sqrt{2}} + \tilde{x}_5(k) \\
\tilde{s}_3(k) &= \tilde{y}_1(k) + \tilde{y}_2(k) = \tilde{x}_1(k) + \tilde{x}_2(k) + \sqrt{2}\tilde{x}_3(k) + \tilde{x}_4(k) + \tilde{x}_5(k) \\
\tilde{s}_4(k) &= \tilde{s}_1(k) \\
\tilde{s}_5(k) &= \tilde{s}_2(k) \,. \qquad (5.9)
\end{aligned}$$

Note that the original center channel signal component $\tilde{x}_3$ appears 3 dB amplified in the center base channel subband $\tilde{s}_3$ (factor $\sqrt{2}$) and 3 dB attenuated in the remaining (side channel) base channel subbands. For further attenuating the center channel signal component in the side base channel subbands, the center channel subband estimate, $\hat{\tilde{x}}_3$, is attenuated by 3 dB and subtracted from the side base channels as illustrated in Figure 5.12. The resulting base channel subbands are

$$\begin{aligned}
\tilde{s}_1(k) &= \tilde{y}_1(k) - \frac{a_3(k)}{\sqrt{2}}(\tilde{y}_1(k) + \tilde{y}_2(k)) \\
\tilde{s}_2(k) &= \tilde{y}_2(k) - \frac{a_3(k)}{\sqrt{2}}(\tilde{y}_1(k) + \tilde{y}_2(k)) \\
\tilde{s}_3(k) &= \tilde{y}_1(k) + \tilde{y}_2(k) \\
\tilde{s}_4(k) &= \tilde{s}_1(k) \\
\tilde{s}_5(k) &= \tilde{s}_2(k) \,. \qquad (5.10)
\end{aligned}$$

The described technique can also be viewed as using dynamic upmixing, i.e. using a different upmixing matrix for each subband at each time k,

$$\mathbf{U}_{25} = \frac{1}{\sqrt{2}} \begin{bmatrix} \sqrt{2} - a_3(k) & -a_3(k) \\ -a_3(k) & \sqrt{2} - a_3(k) \\ \sqrt{2} & \sqrt{2} \\ \sqrt{2} - a_3(k) & -a_3(k) \\ -a_3(k) & \sqrt{2} - a_3(k) \end{bmatrix} . \qquad (5.11)$$

More generally, one could also use different factors for computation of the output center channel subbands and the factors for "dynamic upmixing", as opposed to the same $a_3(k)$ for both.

5.3.4 Subjective evaluation

In the following, a short description of the test presented in (Herre *et al.* 2004) is given. Refer to this paper for a more detailed description of the subjective test. Here, more detail than in (Herre *et al.* 2004) is given about the specific 5-to-2 BCC algorithm parameters that were used for the test.

Subjects and playback setup Ten subjects participated in the test. Eight of the ten subjects were expert listeners with years of experience in audio

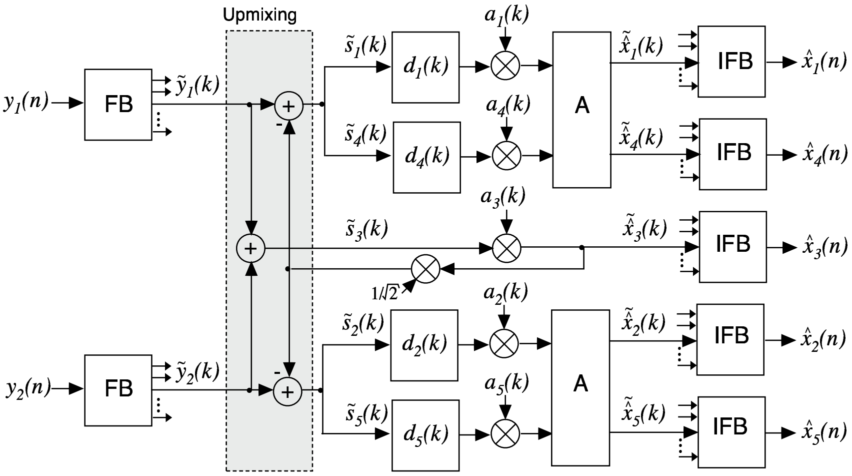


Figure 5.12: 5-to-2 BCC: A center channel subband estimate is subtracted from the base channels for the side channels for improving independence between the channels.

coding, while the other two were less experienced. During the test, the subjects were sitting on a chair that was placed in the sweetspot of a standard 5.1 listening setup in a sound insulated room. A personal computer with an *RME Hammerfall* digital sound output interface connected to *Lake People DAC F20* D/A converters was used. The D/A converters were directly connected to active loudspeakers (*Geithain RL 901*).

Stimuli Eleven items were selected for the listening test, ten of them being commercial music of different styles (3 pop music, 3 jazz music, 4 classical music). The remaining item was created artificially and features a high dissimilarity between different audio channels (item "fountain": the sound of a fountain on the center channel, a piano on the front side channels and singing birds on the surround channels).

Three different coded versions of the items were compared to the reference. The goal was to compare a 5-to-2 BCC based scheme to two other established schemes for providing multi-channel surround (matrixing, and discrete multichannel audio coding).

- 5-to-2 BCC was implemented with a short time Fourier transform (STFT) based filterbank as described in detail in Chapter 6. Spatial cue estimation and synthesis was carried out with a spectral resolution corresponding to subbands with bandwidths equal to 2 ERB. A sampling rate of 44.1 kHz was used and a Hann window of size 2 × 574 samples with 50% overlap, resulting in that ICLD are updated every 13 ms. No explicit ICTD and ICC synthesis and no center channel cancellation was applied, since the test was carried out before these techniques were implemented and optimized.

ICLD were quantized as described in Section 6.2.4, resulting in a BCC side information bitrate of about 14 kb/s. The 5-to-2 BCC encoder output stereo signal was coded using an MP3 audio coder (Brandenburg and Stoll 1994, ISO/IEC 1993). The total bitrate (MP3 and 5-to-2 BCC) was 192 kb/s.

- Dolby Prologic II (Dressler 2000) was chosen to compare the proposed scheme to an established technology with similar functionality. The downmixed stereo signal was not coded, giving Prologic II an advantage for the test.

- As a discrete multi-channel codec, MPEG-2 AAC (Stoll 1996, Bosi *et al.* 1997, ISO/IEC 1997) was used at a bitrate of 320 kb/s where it operates near transparent quality.

Test method In order to obtain both an absolute grade for each of the coders as well as a consistent relative rating between them, the chosen listening test method closely resembles MUSHRA (Rec. ITU-R BS.1534 2003). Several time-aligned audio signals were presented to the listener who performed on-the-fly switching between these signals using a keyboard and a screen. The signals included the original signal, which was labeled as “Reference”, and several anonymized items, arranged in random order. Using a graphical user interface based software, the subjects had to grade the basic audio quality of the anonymized items on a scale with five equally sized regions labeled “Excellent”, “Good”, “Fair”, “Poor” and “Bad”. To check the listener’s reliability and enable relating the ratings to results of other tests, a hidden reference (original) and an anchor were included in addition to the three coded items. The anchor was a 3.5 kHz lowpass filtered version of the reference.

Results Figure 6 shows the results of the listening test and 95% confidence interval for individual test items together with their overall mean. As can be seen, the ratings of AAC coded items overlap with those of the hidden reference in their confidence interval for all items. This shows that the listeners were not able to distinguish between the coded/decoded items and the reference in a statistical sense.

The quality of the Prologic II coded signals was rated mostly in the “good” region of the grading scale. The listeners were able to distinguish these signals from the reference. Listeners frequently reported a change in both, spatial impression and location of auditory events.

The quality of the 5-to-2 BCC+MP3 encoded/decoded signals was mostly rated within the “excellent” range of the grading scale, and their confidence intervals overlap with those of the hidden reference for two out of the 11 test items. For the other nine cases, the listeners were able to distinguish statistically between the 5-to-2 BCC+MP3 coded version and the reference. The subjects reported no significant alterations of the auditory spatial image.

Discussion

Considering that the basic coding efficiency of MP3 is significantly lower than that of AAC and that 5-to-2 BCC+MP3 uses a far lower bitrate than AAC (192

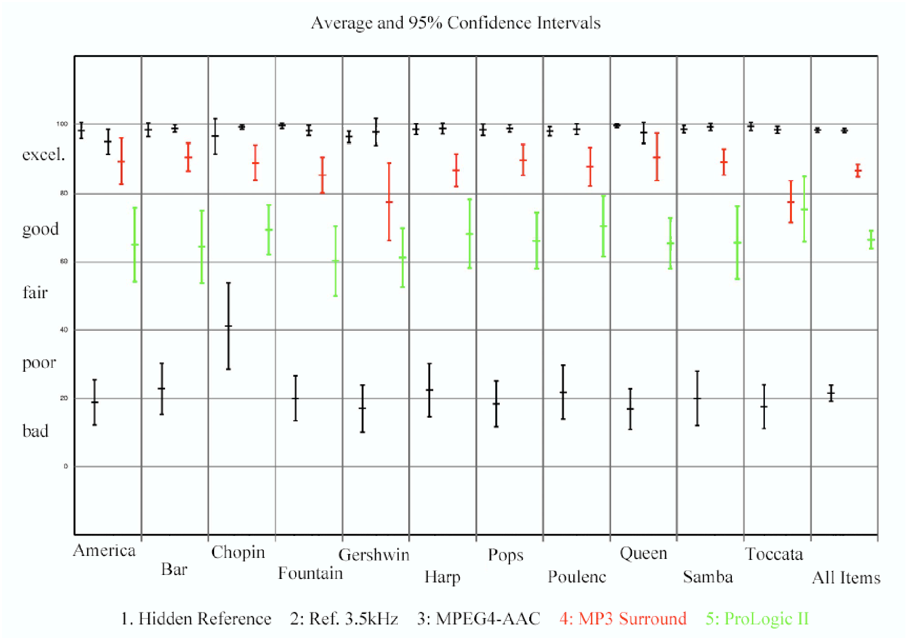


Figure 5.13: The gradings and 95% confidence intervals of the reference, anchor signal, 5-to-2 BCC+MP3 192 kb/s, AAC 320 kb/s, and Dolby Prologic II (ordering from left to right for each item). The gradings averaged over all items are shown on the right. The absolute quality grading scale used is: 80–100: "excellent"; 60–80: "good"; 40–60: "fair"; 20–40: "poor"; 0–20: "bad".

kb/s versus 320 kb/s), the test results imply that 5-to-2 BCC is an efficient technique for coding multi-channel surround. The results indicate that the overall subjective sound quality provided by 5-to-2 BCC+MP3 is much closer to the quality of a discrete multi-channel coder than the quality of a matrixing algorithm (Prologic II) even though the 5-to-2 BCC-based scheme spends only a small fraction of its overall bitrate on encoding of the spatial information.

5.4 BCC for flexible rendering

In this Section, a scheme for audio compression with a *flexible rendering* capability at the decoder is presented. Flexible rendering means that the decoder can determine the auditory spatial image of its output signal. A number of discrete source signals (e.g. separately recorded instruments) are encoded and transmitted jointly. The decoder generates stereo or multi-channel audio signals with an artificial auditory spatial image determined by the user at the decoder. Note that this includes not only determining the auditory spatial image at the decoder, but also the number of playback channels and the rendering method (rendering with ICTD and ICLD, rendering with HRTFs or BRIRs).

A generic audio compression scheme with flexible rendering capability at the decoder is shown in Figure 5.14(a). The C input sound sources are encoded and transmitted as one bitstream to the decoder. A straight forward way of implementing this is to encode each of the C input source signals independently with mono audio coders and combining the resulting C bitstreams to one combined bitstream. The decoder then parses the combined bitstream and decodes each of the C sound sources. The sound sources are then processed to generate a stereo or multi-channel audio signal intended for headphone or loudspeaker playback. Parameters locally determined at the decoder are the playback setup (e.g. loudspeakers/headphones, number and positioning of loudspeakers) and the mixing parameters (e.g. rendering method and desired direction of the auditory event of each source). The drawback of this approach is that the bitrate scales with the number of sound sources C.

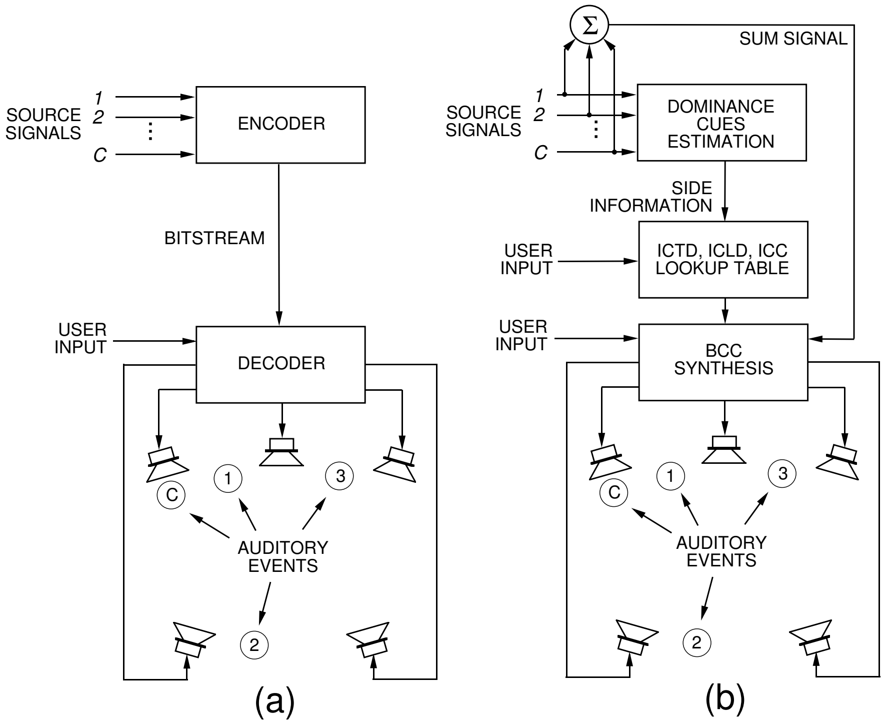


Figure 5.14: (a): Generic scheme for audio compression with flexible rendering capability at the decoder. (b): BCC for flexible rendering.

BCC for flexible rendering only transmits the sum signal of all source signals and side information to the decoder. The sum signal can be coded with any mono audio or speech coder. Given the sum signal plus the side information, the decoder can freely render a custom auditory spatial image of the sound sources for any specific playback setup as if the sound sources were transmitted separately. The described scheme is shown in Figure 5.14(b).

Regular BCC relies on a perceptually motivated synthesis technique for generating stereo or multi-channel audio signals at the decoder given the sum signal. BCC for flexible rendering relies on the same synthesis technique. The difference lies in how the sum signal is computed and the nature of side information that is transmitted.

The chapter is organized as follows. In Sections 5.4.1 and 5.4.2, the encoder and decoder processing is described, respectively. A number of applications for BCC for flexible rendering are described in Section 5.4.3. The response performance of a subject in multitalker communication scenarios is usually improved if the various talkers are perceived as spatially separated. Section 5.4.4 describes a subjective test carried out to assess if the same degree of improvement is observed with signals generated with BCC for flexible rendering.

5.4.1 Encoder processing

As opposed to generating the sum signal from a stereo or multi-channel audio signal, the sum signal is generated by simply adding the discrete source signals (e.g. separately recorded instruments) which are to be part of the auditory spatial image of the decoder output signal which is a priori undefined. BCC for flexible rendering determines the auditory spatial image, composed of an auditory event for each source signal in the sum signal, at the decoder. In the following, we motivate what kind of side information is needed for this purpose.

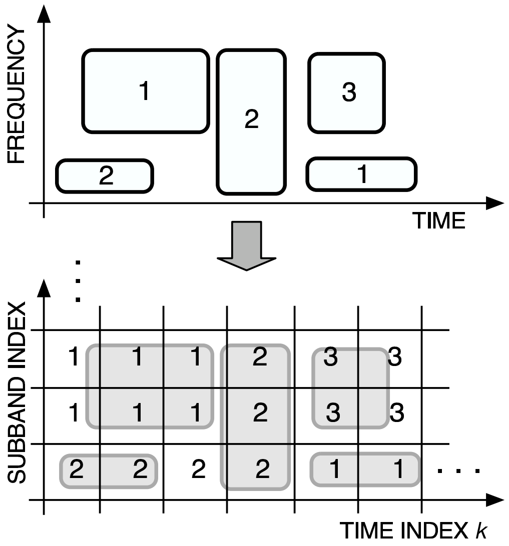


Figure 5.15: Different source signals dominate in different regions of the time-frequency plane of the sum signal (top). For each subband at each time k the source index of the strongest source (bottom) is transmitted to the decoder.

The top of Figure 5.15 schematically shows a time-frequency representation of the sum signal. In this example, there are three source signals mixed into the sum signal. As indicated, these sources dominate in different regions of the time-frequency plane. In the area between the regions where one source dominates, there is either vanishing signal power or a mix of power of various sources. BCC for flexible rendering transmits the structure of such regions to the decoder. This is done by transmitting, for each subband at regular time intervals, the source index of the source with most power at the corresponding time instant,

$$I(k) = \arg \max_{1 \leq c \leq C} p_{\tilde{x}_c}(k) \,, \tag{5.12}$$

where $p_{\tilde{x}_c}(k)$ is a short-time power estimate at time k of the considered subband of source c. Such an assignment of source indices is illustrated in the bottom of Figure 5.15. An entropy coder for reducing the bitrate of the side information is described in Chapter 6.3.2. As with regular BCC, the side information bitrate is only a few kb/s.

5.4.2 Decoder processing

At the decoder, the sum signal plus the source indices of the dominating source in each subband at each time is given. Whereas regular BCC transmits the spatial cues (ICTD, ICLD, and ICC), BCC for flexible rendering obtains the spatial cues from a local table which stores one set of spatial cues for each source c $(1 \leq c \leq C)$. (This corresponds to the "User Input" in Figure 5.14(b)). For each subband, the spatial cues are chosen according to the transmitted source index $I(k)$. Then the multi-channel output signal is generated by applying ICTD/ICLD/ICC synthesis (as described in Section 3.3.5) given these spatial cues. The ICTD and ICLD stored in the table for each source determine the direction, whereas the ICC determines the width of the auditory event. Time adaptive flexible rendering is implemented by (smoothly) modifying the spatial cues in the table in real-time.

For multi-channel playback, the question arises what ICLD values between the pairs of channels correspond to which auditory event direction. A multi-channel amplitude panning law, such as vector base amplitude panning (VBAP) (Pulkki 1997), can be used for this purpose. This is implemented by applying the VBAP algorithm for computing gain factors for each output channel for a desired auditory event direction. Given these gain factors, the corresponding ICLD are computed.

BCC for flexible rendering can also support other rendering methods for generating its output audio signal. For example, HRTFs or BRIRs can be used to generate signals for binaural audio playback. An example how to implement this is given in Chapter 6.3.3.

5.4.3 Applications

Tele-conferencing

Figure 5.16 shows a scheme for a stereo tele-conferencing client incorporating BCC for flexible rendering. It consists of a speech decoder (A), a BCC decoder (B), a speech encoder (C), and a stereo echo canceler (D) (Sondhi *et al.* 1995, Benesty *et al.* 2001). In the server, each client is connected to the other clients with a scheme as shown in Figure 5.17. (A) and (C) are speech decoders and encoders, respectively. There are several reasons why BCC is interesting for applications in tele-conferencing:

Flexibility at the decoder: Each client (decoder) can not only determine the direction at which each other conference participants appears but also the number of playback channels. The same side information supports different types of client systems (e.g. mono, stereo, or multi-channel).

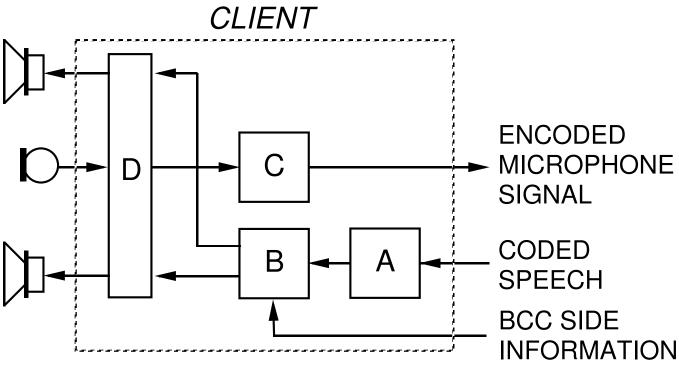


Figure 5.16: Scheme for a stereo tele-conferencing client with one microphone based on BCC. A: speech decoder, B: BCC decoder, C: speech encoder, D: stereo echo canceler.

Low bitrate: The bitrate is as low as in a mono system with a small overhead for the BCC side information.

Backwards compatibility: If the BCC side information can be embedded into the transmission channel between the server and the clients of an existing mono system, this system can be upgraded for stereo or multi-channel tele-conferencing while maintaining backwards compatibility. For example, one could use LSB "bit-flipping" in μLaw (Rec. ITU-T G.711 1993) to embed the BCC side information.

We tried two speech coders, G.722 (Rec. ITU-R G.722.1 1997) and G.729 (Rec. ITU-T G.729 1996), which operate at 24 kb/s and 8 kB/s, respectively. Since these speech coders model a single human vocal tract, it is not obvious that they perform well for multiple simultaneously active speech signals. We found that the speech quality degrades gradually as the number of simultaneous speech signals increases. When rendered to different directions with BCC, speech can be well understood with as many as five simultaneous talkers. In contrast, the speech intelligibility is poor when the speech signals are not spatially separated. In Section 5.4.4 a subjective test is described that was carried out for assessing the degree of improvement achieved using BCC for spatially separating the auditory events corresponding to the talkers.

Virtual reality

In a virtual reality system, users interact with the visual and auditory scene which is presented to them, e.g. they can move around within the virtual scene. It is desirable that in such a system the auditory spatial image adapts to the visual scene. For example, depending on the position of the person in the virtual scene, sound sources are located in different directions. The flexible rendering capability of BCC for flexible rendering can be used for rendering different auditory events corresponding to the sound sources to different directions within the virtual scene. For a flexible rendering capability without BCC, each source signal would need to be stored or transmitted separately. Therefore, with BCC,

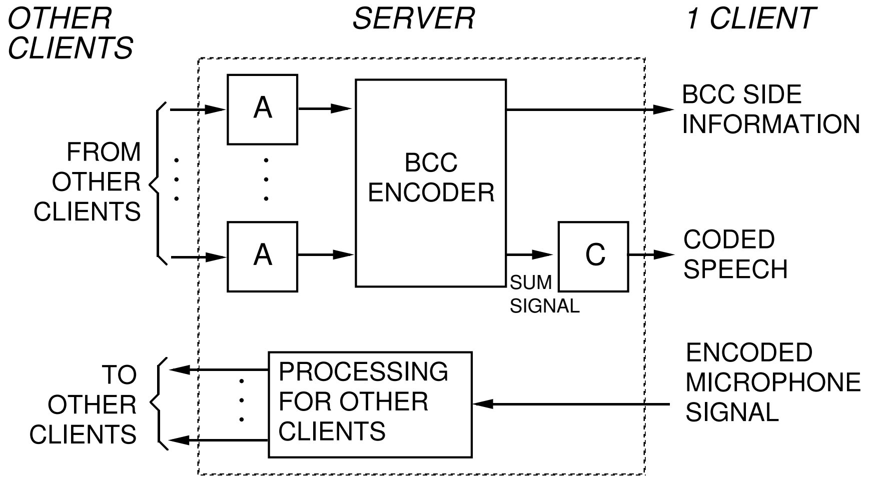


Figure 5.17: Each client is connected to the other clients with a scheme as shown. A: speech decoder, C: speech encoder.

such systems can be implemented with a lower bitrate for the audio. Similarly, BCC for flexible rendering can be used for interactive computer games. Especially games which are played over a network (e.g. Internet) benefit from a low bitrate for the audio transmission.

5.4.4 Subjective evaluation

It is a well known fact that a listener's performance of responding to simultaneous talkers is better when these are spatially separated than without spatial separation (Bolia *et al.* 1999, Spieth *et al.* 1954). This improvement may be explained by informational masking that is reduced by differences in perceived locations of the talkers (Freyman *et al.* 1999). The aim of this subjective test was to assess whether this is also the case for talkers rendered with BCC for flexible rendering. Also, we were examining how close the performance of the BCC-based scheme comes to the performance of separation by regular amplitude and time-delay panning.

Subjects and playback setup

Twelve experienced subjects participated in the test. The signals were presented to the subjects with high quality headphones (*Sennheiser HD 600*) in an acoustically isolated room.

Stimuli

Five tests with five different signal types were conducted to assess the ability of a listener to respond to one of two simultaneous messages:

- diotic: sum signal of the talkers to both ears

- $\mathrm{ICLD}_{\mathrm{ref}}$: separate talker signals rendered with ICLDs (regular amplitude panning)
- $\mathrm{ICTD}_{\mathrm{ref}}$: separate talker signals rendered with ICTDs (regular time-delay panning)
- $\mathrm{ICLD}_{\mathrm{BCC}}$: signals generated with BCC and ICLDs
- $\mathrm{ICTD}_{\mathrm{BCC}}$: signals generated with BCC and ICTDs

For the diotic signals, both talkers are localized in the center. For the other signals, the talkers are localized at the sides (left and right) with ICLDs of ± 16 dB and ICTDs of ± 500 μs. ICC was chosen equal to one for all BCC items.

Test method

The subjects were given a task which required responding to one of two simultaneous voice messages. This is a variation of the "cocktail party effect" of attending to one voice in the presence of others. Each of the subjects took all of the tests in a randomized order. For the test we used the speech corpus by Bolia *et al.* (2000). A typical sentence of the corpus is "READY LAKER, GO TO BLUE FIVE NOW", where LAKER is the call sign and BLUE FIVE is a color-number combination. There are eight different call signs, four colors, and eight numbers. All these sentences are synchronized, i.e. the call sign and color-number combination occur at approximately the same time. One out of four female voices was randomly chosen for each of the two talkers in each item of each test. We chose the call sign, color, and number randomly with the restriction that the call sign assigned to the subject occurred in 50 % of the cases. The subjects were instructed to respond when their call sign (always LAKER) was called by indicating the color-number combination of the sentence in which the call sign appeared. Each of the five tests consisted of 10 training items followed by 20 test items.

Results

Table 5.2 shows the results for the case when the subject was called (50 % of the items for each signal type). As expected, these results suggest that the percentage of correct identification of the call sign and of the color-number combination significantly improve for the signals generated with separated source signals ($\mathrm{ICLD}_{\mathrm{ref}}$, $\mathrm{ICTD}_{\mathrm{ref}}$). The signals generated with BCC for flexible rendering ($\mathrm{ICLD}_{\mathrm{BCC}}$ and $\mathrm{ICTD}_{\mathrm{BCC}}$) are almost as good. For the case when the subject was not called, the percentages of the subjects responding was below two percent for all tests.

5.5 Conclusions

Three variations of BCC were described in this chapter:

Hybrid BCC: Above a certain bitrate, the quality of a BCC-based audio coder saturates. That is, because even if the transmitted mono signal is coded

Table 5.2: Results for the case when the subjects were called by their call sign. The middle column shows the percentage of correct identification of the call sign and the right column shows the conditional percentage of the correct color-number combination given that the subject's call sign was correctly identified.

	call sign	*color-number*
diotic	70 %	64 %
ICLD_{ref}	78 %	98 %
ICTD_{ref}	85 %	88 %
ICLD_{BCC}	77 %	96 %
ICTD_{BCC}	78 %	91 %

without distortions auditory spatial image artifacts remain. For better quality at higher bitrates, BCC is only applied above a certain frequency f_0. By choosing f_0 between 0 Hz and the Nyquist frequency, the resulting audio quality ranges from fullband BCC quality to transparent audio quality. The bitrate of hybrid BCC combined with a conventional audio coder scales from a bitrate slightly larger than a mono audio coding bitrate ($f_0 = 0$, fullband BCC) to the bitrate necessary for conventional multi-channel audio coding ($f_0 =$ Nyquist frequency). The choice of f_0 offers a trade-off between low bitrate and audio quality above fullband BCC quality. An audio coder based on hybrid BCC offers not only better quality at lower bitrates but also better quality at higher bitrates (below transparent coding) where regular BCC performs worse than a conventional audio coder.

C-to-E BCC: This scheme transmits more than one audio channel, but less than the number of input audio channels. There are two motivations for transmitting more than one audio channel. Firstly, most of the existing audio infrastructure is based on two-channel stereo. For upgrading such systems for multi-channel playback in a backwards compatible way it would be desirable to have BCC with two transmission channels. More generally speaking, C-to-E BCC can upgrade E-channel audio systems in a backwards compatible way to C-channel systems. Secondly, one can take advantage of the fact that more than one audio channel is transmitted resulting in a higher audio quality. C-to-E BCC was described in detail for the general case of any number of transmission channels. Special considerations were discussed for practical application of 5-to-2, 6-to-5, and 7-to-5 BCC.

BCC for flexible rendering: This scheme transmits the sum of a number of independent source signals to the decoder plus low bitrate side information. With the help of the side information, the decoder can generate stereo or multi-channel audio signals as if the source signals were given separately. That is, the decoder determines the signal format (e.g. stereo or multi-channel) and auditory spatial image properties. Optionally, also binaural signals rendered with HRTFs or BRIRs can be generated. When combined with a conventional mono audio coder or speech coder, BCC for flexible rendering allows low bitrate joint transmission of independent source signals for the purpose of rendering at the decoder.

Chapter 6

Low Complexity Implementation of Binaural Cue Coding (BCC)

6.1 Introduction

In this chapter, it is described how binaural cue coding (BCC) schemes are implemented with low computational complexity, using a short time Fourier transform (STFT) based filterbank. The presented implementations have low computational complexity and low delay, making them suitable for affordable implementation on microprocessors or digital signal processors for real-time applications[1].

The BCC schemes were thoroughly motivated previously and here less motivation is presented in favor of more signal processing details for the specific STFT-based implementations. The STFT-based implementation of BCC, including schemes for quantization and coding of the BCC side information, is presented in Section 6.2. Section 6.3 describes the STFT-based implementation of BCC for flexible rendering, including coding of the side information. Section 6.4 discusses computational complexity and algorithmic delay of the presented schemes. Conclusions are drawn in Section 6.5.

6.2 STFT-based BCC

Figure 6.1 shows the various processing stages of a BCC scheme. This scheme is functionally the same as the corresponding scheme introduced in Chapter 3, except that quantization and coding of the spatial cues has been added. Note that usually the sum signal is coded with a conventional mono audio coder, not explicitly shown in the figure. Before describing the details of each processing stage in detail, the STFT-based time-frequency transform is described that is used instead of a more complex non-uniform filterbank.

---

[1] A reader who is mostly interested in concepts and not implementation details may skip this chapter.

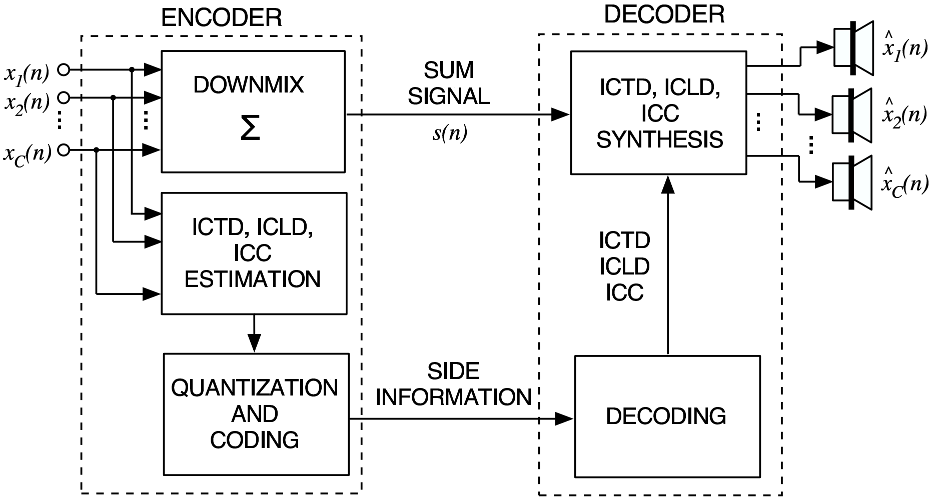


Figure 6.1: BCC scheme, including quantization and coding of the spatial cues.

6.2.1 The time-frequency transform

Short-time Fourier transform (STFT)

BCC encoder and decoder, based on an auditory filterbank closely mimicking the spectral decomposition carried out in the auditory system, is described in (Baumgarte and Faller 2003). Baumgarte and Faller (2003) found that a low complexity STFT-based filterbank achieves the same performance as the significantly more complex auditory filterbank (the assessment was carried out only for inter-channel level difference synthesis). In the following, the STFT that is applied in BCC implementations is described in detail.

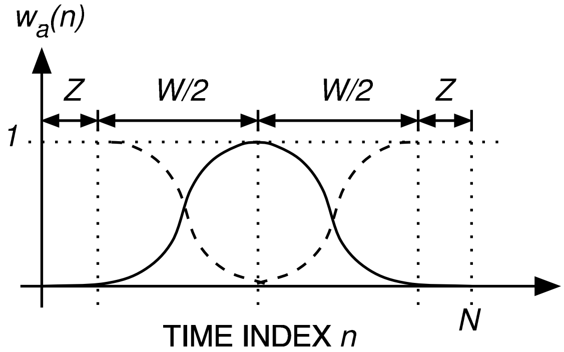


Figure 6.2: Analysis window. The time-span of the window W is shorter than the DFT length N, such that non-circular time-shifts and linear filtering can be implemented in the STFT domain.

BCC synthesis needs to be able able to introduce delays and modify the level of audio signals adaptively in time and frequency to generate inter-channel time differences (ICTDs) and inter-channel level differences (ICLDs) between pairs of channels. Furthermore, it is desirable that time adaptive filtering can be carried out efficiently for inter-channel coherence (ICC) synthesis. The STFT used provides these desired properties.

The STFT applies a *discrete Fourier transform* (DFT) to windowed portions of a signal $x(n)$. A signal frame of N samples is multiplied with a window before an N-point DFT is applied. We use a Hann window with zero padding at both sides,

$$w_a(n) = \begin{cases} 0 & \text{for} \quad 0 \leq n < Z \text{ or} \\ & \phantom{\text{for}} \quad N - Z \leq n < N \\ \sin^2(\frac{(n-Z)\pi}{W}) & \text{for} \quad Z \leq n < Z + W \,, \end{cases} \tag{6.1}$$

where Z is the width of the zero region before and after the non-zero part of the window. Figure 6.2 shows the described window schematically. The non-zero window span is W. Adjacent windows are overlapping and are shifted by $W/2$ samples (hop size). The window was chosen such that the overlapping windows add up to a constant value of 1. Therefore, for the inverse transform there is no need for additional windowing. A plain inverse DFT of size N with time advance of successive frames of $W/2$ samples is used. If the spectrum is not modified, perfect reconstruction is achieved by overlap/add.

Implementation of short convolutions in the STFT domain

Given a generic signal $x(n)$ and an impulse response $h(n)$, their convolution is written as

$$y(n) = h(n) \star x(n) \,, \tag{6.2}$$

where $\star$ is the convolution operator. In the following, it is shown how to carry out this convolution in the STFT frequency domain.

At time index k, the windowed signal is

$$x_k(n) = \begin{cases} w_a(n - k\frac{W}{2} + Z)x(n), & k\frac{W}{2} \leq n < k\frac{W}{2} + W \\ 0, & \text{otherwise}\,. \end{cases} \tag{6.3}$$

Windows other than (6.2) can be used which fulfill the (in the following assumed) condition,

$$x(n) = \sum_{k=-\infty}^{\infty} x_k(n) \,. \tag{6.4}$$

First, the simple case of implementing a convolution of the windowed signal $x_k(n)$ in the frequency domain is considered. Figure 6.3(a) illustrates the non-zero span of an impulse response $h(n)$ of length M. Similarly, the non-zero span of $x_k(n)$ is illustrated in Figure 6.3(b). It is easy to verify that $h(n) \star x_k(n)$ has a non-zero span of $W + M - 1$ samples as illustrated in Figure 6.3(c).

Figure 6.4 illustrates at which time indices DFTs of length $W + M - 1$ are applied to the signals $h(n)$, $x_k(n)$, and $h(n) \star x_k(k)$, respectively. Figure 6.4(a) illustrates that H_m denotes the spectrum with frequency index m obtained by applying the DFT starting at time index $n = 0$ to $h(n)$. Figures 6.4(b) and (c) illustrate the computation of $X_m(k)$ and $Y_m(k)$ from $x_k(n)$ and $h(n) \star x_k(n)$,

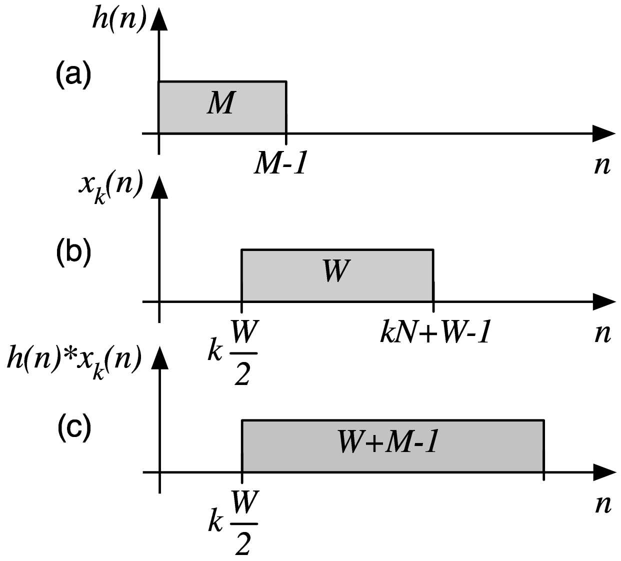


Figure 6.3: Illustration of the non-zero span of $h(n)$, $x_k(n)$, and $h(n) \star x_k(n)$.

respectively, by applying the DFTs starting at time index $n = k\frac{W}{2}$. It can easily be shown that $Y_m(k) = H_m X_m(k)$, if $M \leq Z$ or for non-causal filters

$$h(n) = 0 \text{ for } |n| > Z\,. \tag{6.5}$$

That is, because the zeros at the start and end of the signals $h(n)$ and $x_k(n)$ result in that the circular convolution imposed on the signals by the spectrum product is equal to a linear convolution.

From the linearity property of convolution and (6.4) it follows that

$$h(n) \star x(n) = \sum_{k=-\infty}^{\infty} h(n) \star x_k(n)\,. \tag{6.6}$$

Thus, it is possible to implement a convolution in the domain of the STFT, by computing at each time k the product $H_m X_m(k)$ and applying the inverse STFT (inverse DFT plus overlap/add). Condition (6.5) implies that delays in the range $[-Z, Z]$ can be imposed to the underlying signal by spectral modification. The described technique is similar to overlap/add convolution (Oppenheim and Schaefer 1989) with the generalization that overlapping windows can be used (any window fulfilling condition (6.4)).

Implementation of long convolutions in the STFT domain

Previously, it was explained how to implement a convolution in the STFT frequency domain for filters satisfying condition (6.5). That method is not practical for long impulse responses (e.g. $M \gg W$) since then a DFT of a much larger size than W needs to be used (because at least $Z = M - 1$ zeroes need to be padded at the end of the transform window). In the following, we

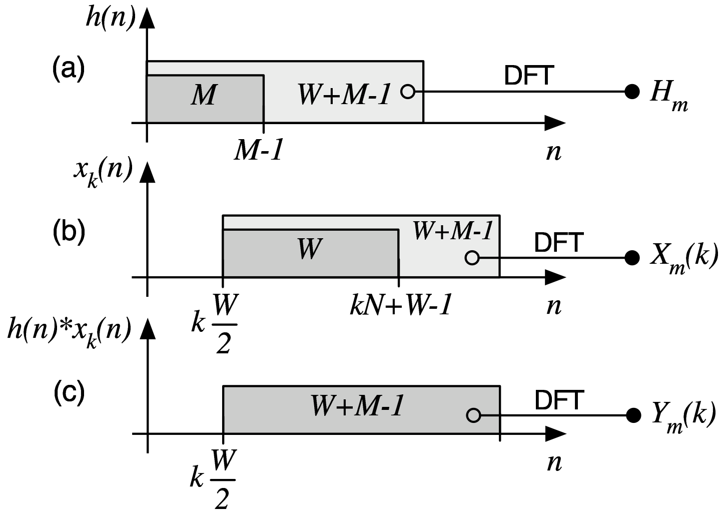


Figure 6.4: Illustration of the DFTs of size $W + M - 1$ that are applied to $h(n)$, $x_k(n)$, and $h(n) \star x_k(n)$.

are extending that method such that only a DFT of size $W + \frac{W}{2} - 1$ needs to be used.

A long impulse response $h(n)$ of length $M = L\frac{W}{2}$ is partitioned into L shorter impulse responses ($0 \leq l < L$),

$$h_l(n) = \begin{cases} h(n + l\frac{W}{2}), & 0 \leq n < \frac{W}{2} \\ 0, & \text{otherwise}. \end{cases} \tag{6.7}$$

(If the length of $h(n)$ is not a multiple of $\frac{W}{2}$, zeroes are added to the tail of $h(n)$). The convolution with $h(n)$ can then be written as a sum of shorter convolutions,

$$h(n) \star x(n) = \sum_{l=0}^{L-1} h_l(n) \star x(n - l\frac{W}{2}). \tag{6.8}$$

Applying (6.6) and (6.8) at the same time, yields

$$h(n) \star x(n) = \sum_{k=-\infty}^{\infty} \sum_{l=0}^{L-1} h_l(n) \star x_k(n - l\frac{W}{2}). \tag{6.9}$$

The non-zero time span of one convolution in (6.9), $h_l(n) \star x_k(n - l\frac{W}{2})$, as a function of k and l is $(k+l)\frac{W}{2} \leq n < (k+l+1)\frac{W}{2} + W$. Thus, for obtaining its spectrum, $\tilde{Y}_m(k + l)$, the DFT is applied to this interval (corresponds to DFT position index $k+l$). It can easily be shown that $\tilde{Y}_m(k+l) = H_{l,m}X_m(k)$, where $X_k(j\omega)$ is defined as previously with $M = \frac{W}{2}$ and $H_{l,m}$ is defined similarly as previously H_m but for the impulse response $h_l(n)$.

The sum of all spectra $\tilde{Y}_m(i)$ with the same DFT position index, $i = k + l$,

Table 6.1: The partition boundaries A_b $(0 \leq b \leq B = 20)$ for the case of partition bandwidths of 2 ERB, $N = 1024$, and a sampling rate of $f_s = 32$ kHz.

A_0 0	A_1 2	A_2 4	A_3 7	A_4 11	A_5 15	A_6 20
A_7 26	A_8 34	A_9 44	A_{10} 56	A_{11} 71	A_{12} 90	A_{13} 113
A_{14} 142	A_{15} 178	A_{16} 222	A_{17} 277	A_{18} 345	A_{19} 430	A_{20} 513

is

$$Y_m(i) = \sum_{k+l=i} \tilde{Y}_m(k+l) \tag{6.10}$$

$$= \sum_{l=0}^{L-1} H_{l,m} X_m(i-l)\,. \tag{6.11}$$

Thus, the convolution $h(n) \star x_k(n)$ is implemented in the STFT domain by applying (6.11) at each spectrum index i to obtain $Y_m(i)$. The inverse STFT applied to $Y_m(i)$ is equal to the convolution $h(n) \star x(n)$, as desired.

Note that, independently of the length of $h(n)$, the required amount of zero padding is upper bounded by $\frac{W}{2} - 1$ (one sample less than the STFT window hop size). DFTs larger than $W + \frac{W}{2} - 1$ can be used if desired (e.g. using an fast Fourier transform (FFT) with a length equal to a power of two).

Perceptually motivated frequency resolution

The uniform spectral resolution of the STFT is not well adapted to human perception. Therefore, the uniformly spaced spectral coefficients $X_m(k)$ $(0 \leq m \leq N/2)$ are grouped into B non-overlapping partitions with bandwidths better adapted to perception. One partition conceptually corresponds to one "subband" of the previous descriptions of BCC. Only the first $N/2+1$ spectral coefficients of the spectrum are considered, because the spectrum is symmetric. The indices of the DFT coefficients which belong to the partition with index b $(1 \leq b \leq B)$ are $m \in \{A_{b-1}, A_{b-1}+1, \ldots, A_b - 1\}$ with $A_0 = 0$. Since each partition corresponds conceptually to one BCC subband, only one set of inter-channel cues (ICTD, ICLD, ICC) are synthesized within each partition.

Unless otherwise noted, we used $W = 896$, $Z = 64$, and $N = 1024$ for a sampling rate of $f_s = 32$ kHz. $B = 20$ partitions are used, each having a bandwidth of approximately two times the equivalent rectangular bandwidth (ERB) (Glasberg and Moore 1990). The resulting partition boundaries A_b are shown in Table 6.1. N was chosen to be a power of two, such that an efficient FFT algorithm can be applied for implementing the DFT.

6.2.2 Downmixing to one channel

Implementation of the downmixing of the input audio channel to one single channel, as proposed in Chapter 3.3.2, is described in the following. The spectral coefficients of one partition b (indices: $A_{b-1} \leq m < A_b$) of the equalized sum signal $S_m(k)$ are computed by

$$S_m(k) = e_b(k) \sum_{c=1}^{C} X_{c,m}(k)\,, \tag{6.12}$$

where $X_{c,m}(k)$ are the spectra of the input audio channels and $e_b(k)$ is a gain factor computed as

$$e_b(k) = \sqrt{\frac{\sum_{c=1}^{C} p_{\tilde{x}_{c,b}}(k)}{p_{\tilde{x}_b}(k)}}\,, \tag{6.13}$$

with partition power estimates,

$$\begin{aligned} p_{\tilde{x}_{c,b}}(k) &= \sum_{m=A_{b-1}}^{A_b-1} |X_{c,m}(k)|^2 \\ p_{\tilde{x}_b}(k) &= \sum_{m=A_{b-1}}^{A_b-1} |\sum_{c=1}^{C} X_{c,m}(k)|^2\,. \end{aligned} \tag{6.14}$$

To prevent artifacts resulting from large gain factors when attenuation of the sum of the subband signals is significant, the gain factors $e_b(k)$ are limited to 6 dB, i.e. $e_b(k) \leq 2$.

6.2.3 Estimation of ICTD, ICLD, and ICC

The estimation of the spatial cues is carried out as described in the following.

ICTD estimation: At frequencies below about 1.5 kHz the phase delay between a pair of channels is relevant for spatial perception (Blauert 1997) and the ICTDs are estimated by averaging the phase delay within each partition between a pair of channels. At frequencies above about 1.5 kHz the group delays ("envelope delay") is relevant for spatial perception (Blauert 1997). To estimate the group delay, the phase difference between a pair of channels is computed. Then, for each partition b linear regression is used to compute the slope of the phase difference of the spectral coefficients within the partition (indices: $A_{b-1} \leq m < A_b$). The group delay between a pair of channels is proportional to the slope of the regression line.

ICLD estimation: First, for each of the audio input channels $1 \leq c \leq C$ the power within each partition $1 \leq b \leq B$, $p_{\tilde{x}_{c,b}}(k)$, is estimated similarly to (6.14). The estimated ICLD in dB between channel c and the reference channel 1 for partition b is,

$$\Delta L_{1c,b}(k) = 10 \log_{10} \left(\frac{p_{\tilde{x}_{c,b}}(k)}{p_{\tilde{x}_{1,b}}(k)} \right)\,. \tag{6.15}$$

ICC estimation: At each time for each partition the two strongest input channels are selected for the ICC estimation as described in Section 3.3.4. If the spectra of the selected two input channels are denoted $X_{i,m}(k)$ and $X_{j,m}(k)$ (in this paragraph the letters i and j are used independently of the other parts of the chapter). The coherence between these two channels is

$$c_m(k) = \sqrt{\frac{|\Phi_{ij}(n,k)\Phi^*_{ji}(m,k)|}{\Phi_{ii}(m,k)\Phi_{jj}(m,k)}}\,, \tag{6.16}$$

with

$$\Phi_{ij}(m,k) = E\{X_{i,m}(k)X^*_{j,m}(k)\}\,, \tag{6.17}$$

where $E\{\ldots\}$ denotes mathematical expectation and * is the complex conjugate of a complex number. A short-time estimate of $c_m(k)$ is obtained by computing $\Phi_{ij}(n,k)$ according to

$$\Phi_{ij}(m,k) = \alpha X_{i,m}(k)X^*_{j,m}(k) + (1-\alpha)\Phi_{ij}(m,k-1)\,. \tag{6.18}$$

The factor α determines the degree of smoothing of the estimation over time. The time constant of the exponential decay in seconds is

$$T_{\text{ICC}} = \frac{1}{\alpha f_s}\,, \tag{6.19}$$

where f_s denotes the sampling frequency. We use $T_{\text{ICC}} = 50$ ms.

The coherence for the partition is computed as a weighted average of the coherence estimates of the spectral coefficients within the partition,

$$c_b(k) = \frac{\sum_{m=A_{b-1}}^{A_b-1} c_m(k)(\Phi_{ii}(m,k)+\Phi_{jj}(m,k))}{\sum_{m=A_{b-1}}^{A_b-1} \Phi_{ii}(m,k)+\Phi_{jj}(m,k)}\,. \tag{6.20}$$

6.2.4 Quantization and coding of ICTD, ICLD, and ICC

The estimated ICTDs and ICLDs, $\tau_{1c,b}(k)$ and $\Delta L_{1c,b}(k)$), are limited to a range of $\pm\tau_{max}$ and $\pm\Delta L_{max}$ with $\tau_{\max} = 800$ μs and $\Delta L_{max} = 18$ dB. For stereo audio playback, these limits correspond to auditory events at the far right and far left. The ICC is by definition limited to $0 \leq c_b(k) \leq 1$. After limiting, uniform quantizers are used for quantizing the estimated inter-channel cues. Uniform quantizers may not be optimal but we use them for simplicity. For the experiments previously reported in this thesis that use quantization 7 quantizer values were used for ICTDs and ICLDs and 8 quantizer values for ICCs (Section 3.5.2 and 5.3.4). More details about the quantization of ICTDs and ICLDs are given in (Faller and Baumgarte 2002*b*).

For a lower bitrate, the quantizer indices are coded as described in the following. For each frame, two sets of quantizer index differences are computed. Firstly, for each partition b the difference between the current (time index k) and the previous (time index $k-1$) quantizer index is computed. Secondly, the difference of the quantizer index of each partition b with respect to the index of the partition with the next higher index, $b+1$, is computed. Both of these sets of index differences are coded with a Huffman coder. We use one single Huffman codebook for all cases. For transmission, the set is chosen which uses less bits. One additional bit is transmitted indicating which set is transmitted.

The resulting average bitrate for ICTDs or ICLDs for one channel pair for the entropy coded quantizer indices is approximately 2 kb/s. The average bitrate for the ICC is also about 2 kb/s.

6.2.5 Synthesis of ICTD, ICLD, and ICC

The given sum signal $s(n)$ is converted to the STFT domain. As a function of the given ICTDs, ICLDs, and ICCs, the spectral coefficients are modified for generating the spectra of the multiple output channels. These spectra are converted back to the time domain resulting in the multi-channel output signal.

Given the spectral coefficients $S_m(k)$ of the sum signal, the spectral coefficients $\hat{X}_{c,m}(k)$ for each channel c are obtained by

$$\hat{X}_{c,m}(k) = F_{c,m}(k) G_{c,m}(k) S_m(k)\,, \tag{6.21}$$

where $F_{c,m}(k)$ is a positive real number determining a level modification for each spectral coefficient. $G_{c,m}(k)$ is a complex number of magnitude one determining a phase modification for each spectral coefficient. The following two paragraphs describe how $F_{c,m}(k)$ and $G_{c,m}(k)$ are obtained given $\{\Delta L_{1c,b}(k),\ \tau_{1c,b}(k), c_b(k)\}$. For notational simplicity, the time index k is often ignored in the following.

Determining the level modification for each channel

The factors $F_{c,m}$ for channel $c > 1$ are computed for each spectral coefficient within a partition b (indices: $A_{b-1} \leq m < A_b$) given $\Delta L_{1c,b}$,

$$F_{c,m} = 10^{(\Delta L_{1c,b} + r_{1c,m})/20} F_{1,m}\,, \tag{6.22}$$

where $r_{1c,m}$ are random numbers for controlling the ICC between the channel pairs. The factors $F_{1,m}$ for the reference channel ($c = 1$) are computed such that for each partition the sum of the power of all channels is the same as the power of the sum signal,

$$F_{1,m} = 1 \Bigg/ \sqrt{1 + \sum_{i=2}^{C} 10^{(\Delta L_{1i,b} + r_{1i,m})/10}}\,. \tag{6.23}$$

For de-correlation (ICC synthesis), ICLDs are varied within partitions by adding a random sequence to the ICLDs along the frequency axis. The random sequence $\bar{r}_{1c,m}$ for each channel pair is chosen such that the variance is approximately constant for all partitions in all channels and the average is zero in each partition. Moreover, clustering of large or small values in the sequence should be avoided. For each channel pair a different independent random sequence is used. The random sequences are not varied in time, i.e. the same sequences are applied to each frame. ICC is controlled by modifying the variance of the random sequence. The variance modification is done in each partition b (indices: $A_{b-1} \leq m < A_b$) as a function of $c_b(k)$,

$$r_{1c,m}(k) = [1 - c_b(k)]\bar{r}_{1c,m}\,. \tag{6.24}$$

A suitable amplitude distribution for the random sequence is a uniform distribution. The range of the distribution of $\bar{r}_{1c,m}$ determines the strength of the correlation reduction. We use a range of ±3 dB or smaller, i.e. $\bar{r}_{1c,m} \in [1/\sqrt{2}, \sqrt{2}]$.

Figure 6.5 shows an example of $F_{c,m}$ for stereo signals, i.e. $C = 2$. In the example shown, specific ICLD are synthesized in partitions $b - 1$, b, and $b + 1$.

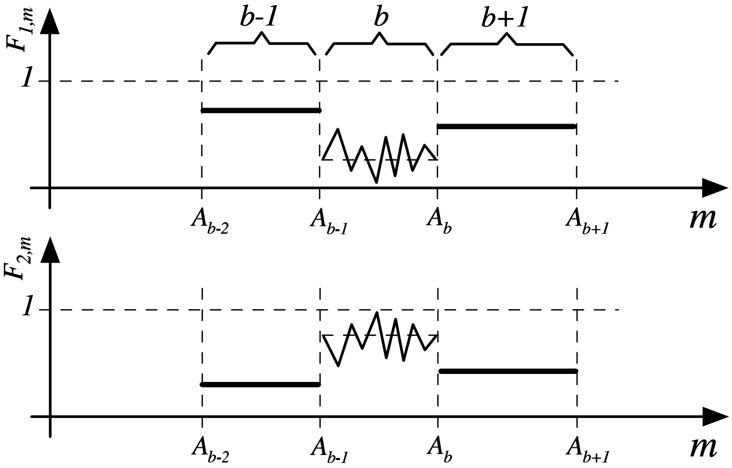


Figure 6.5: Example for $F_{c,m}$ for three partitions $b-1$, b, and $b+1$ for two output channels.

For partitions $b-1$ and $b+1$, $c_b(k)$ is equal to one and no level fluctuations are introduced within the partition (6.24). For partition b, ICLD fluctuations are introduced for synthesis of a specific $c_b(k) < 1$.

Before applying (6.21), $F_{c,m}$ is smoothed at the partition boundaries by interpolating between the different $\Delta L_{c,b}$ to reduce aliasing artifacts.

Determining the phase modification for each channel

The complex factors $G_{c,n}$ for channel $c > 1$ are computed for each spectral coefficient within a partition b (indices: $A_{b-1} \leq m < A_b$) given $\tau_{1c,b}$ in sampling intervals,

$$G_{c,m} = \exp\left(-j\frac{2\pi[m(\tau_{1c,b} + \tau_b) + g_{c,b}]}{N}\right) , \tag{6.25}$$

where τ_b is the delay which is introduced into reference channel 1,

$$\begin{aligned} \tau_b &= -\frac{1}{2}(\max_{2\leq c\leq C} \tau_{1c,b} + \min_{2\leq c\leq C} \tau_{1c,b}) \\ G_{1,m} &= \exp\left(-j\frac{2\pi(m\tau_b + g_{1,b})}{N}\right) . \end{aligned} \tag{6.26}$$

The delay for the reference channel τ_b is computed in (6.26) such that the maximum absolute delay introduced to any channel in a specific partition is minimal. By minimizing the maximum phase modification, the maximum audio signal distortion resulting from (6.25) is minimized.

For reducing correlation (synthesizing ICC), not only ICLDs are varied within partitions but also ICTDs. ICTDs are varied by imposing a frequency variant group delay between channel pairs. The average group delay imposed within each partition is zero such as not to affect directional perception (ICTD) of the corresponding signal component. We are varying the group delay sinu-

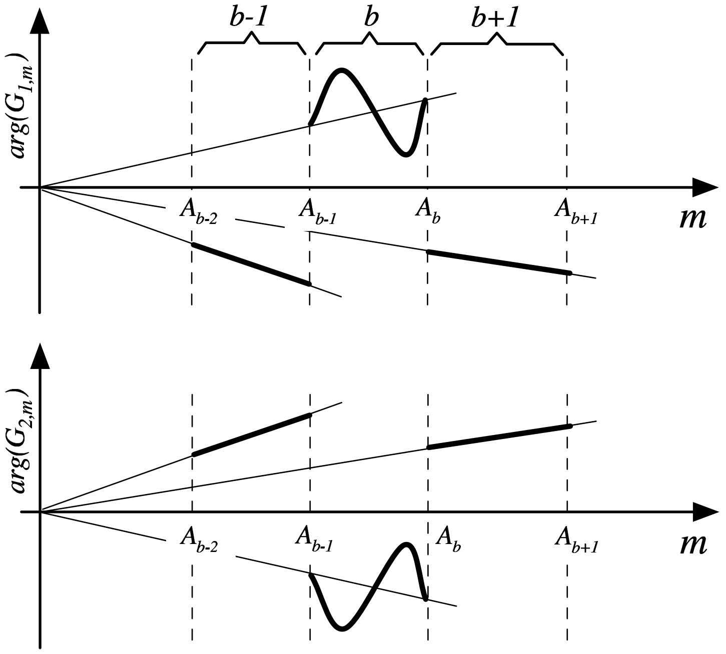


Figure 6.6: Example for the phase of $G_{c,m}$ for three partitions $b-1$, b, and $b+1$ for two output channels.

soidally with one period per partition,

$$g_{c,b}(k) = \frac{g_0}{2}(1 - c_b(k)) \sin\left(\frac{2\pi(m - A_{b-1} + 0.5)}{A_b - A_{b-1}} + \frac{2\pi(c-1)}{C}\right) , \qquad (6.27)$$

where g_0 determines the range of imposed group delays between the channels. For stereo signals we use $g_0 = 40$.

An example of the phase of $G_{c,m}$ is shown in Figure 6.6. In the example, specific ICTD are synthesized in partitions $b-1$, b, and $b+1$. For partitions $b-1$ and $b+1$, $c_b(k)$ is equal to one and no group delay fluctuations are introduced within the partition (6.27). For partition b, ICTD fluctuations are introduced for synthesis of a specific $c_b(k) < 1$.

6.2.6 Synthesis of ICC based on late reverberation

This section outlines how to implement the ICC synthesis scheme described in Chapter 4 with low complexity. For reproducing naturally sounding late reverberation (Karjalainen and Järveläinen 2001), the impulse responses $h_c(n)$ (4.2) need to be as long as several hundred milliseconds, resulting in high computational complexity. Furthermore, for each $h_c(n)$ $(1 \leq c \leq C)$ an additional filterbank is required (as shown in Figures 4.1 and 4.3).

The computational complexity could be reduced by using artificial reverberation algorithms (Schroeder 1962, Gardner 1998) for generating late reverberation and using that for $s_c(n)$. Another possibility is to still carry out the convolutions (4.1) but applying an algorithm based on the FFT for reduced computational complexity (Gardner 1995). We chose a related method. The difference to (Gardner 1995) is that we are operating in the STFT domain, i.e. overlapping windows are used. The motivation is to use the same STFT for carrying out the convolutions and the BCC processing. This results in lower computational complexity of the convolution computation and no need in using an additional filterbank for each $h_c(n)$.

The convolutions (4.1) are implemented in the STFT domain as described in Section 6.2.1 by applying (6.11). The STFT spectra resulting from (6.11) are directly used as late reverberation in the frequency domain without going back to the time domain.

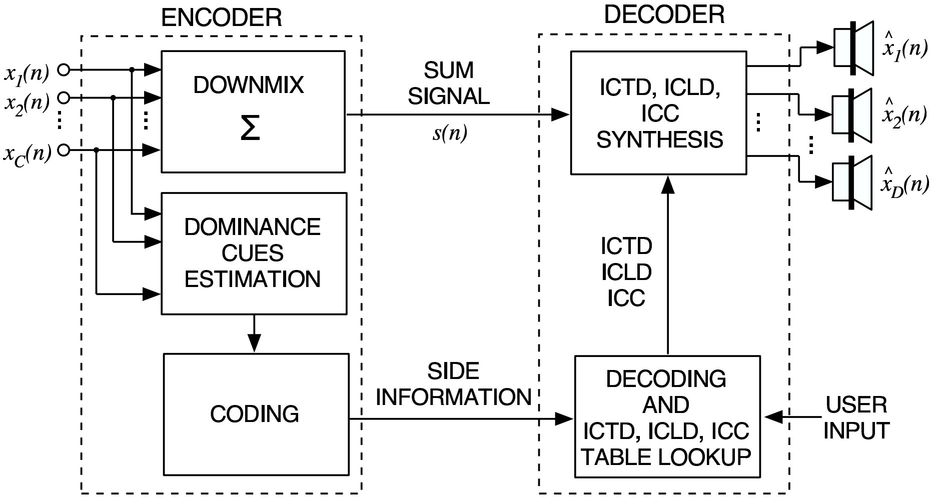


Figure 6.7: BCC for flexible rendering scheme including coding of the dominance cues.

6.3 STFT-based BCC for flexible rendering

Figure 6.7 shows a detailed scheme for BCC for flexible rendering. In the following, the dominance cues estimation, coding, and decoder processing are described.

6.3.1 Estimation of dominance cues

To each partition b a source index $I_b(k)$ $(1 \leq I_b(k) \leq C)$,

$$I_b(k) = \arg \max_{1 \leq c \leq C} p_{\tilde{x}_{c,b}}(k) \,, \tag{6.28}$$

is assigned, where C is the number of source input signals and the partition power estimates are computed similarly as in 6.14.

6.3.2 Coding of dominance cues

A run-length coding algorithm (Jayant and Noll 1984) is applied to the source indices $I_b(k)$ over frequency (a more sophisticated algorithm may also consider dependencies between frames). For the chosen parameters (Section 6.2.1), this results in a bitrate of about 2 kb/s for the case of having three simultaneously active speech signals as source signals. If not all speech signals are active simultaneously, the run-length coding is more efficient and results in lower bitrates.

6.3.3 Decoder processing

As mentioned in Chapter 3, BCC for flexible rendering obtains its ICTD, ICLD, and ICC cues at the decoder from a lookup table. These cues are then applied similarly as described in Section 6.2.5 for generating the stereo or multi-channel audio signal. In the following, the detailed processing is described when BCC for flexible rendering is used with head related transfer functions (HRTFs) or binaural room impulse responses (BRIRs) as opposed to ICTD, ICLD, and ICC.

In this case, the local table in the BCC decoder stores for each source c a left and right HRTF or BRIR frequency response, $H_{\mathrm{L},c,m}$ and $H_{\mathrm{R},c,m}$. For obtaining the left and right binaural signals, each partition b is multiplied with the coefficients corresponding to the left and right HRTF or BRIR associated with source $I_b(k)$,

$$\hat{X}_{1,m}(k) = H_{\mathrm{L},I_b(k),m} S_m(k) \text{ and } \hat{X}_{2,m}(k) = H_{\mathrm{R},I_b(k),m} S_m(k)\,, \qquad (6.29)$$

where $n \in \{A_{b-1}, A_{b-1}+1, \ldots, A_b - 1\}$ are the indices of spectral coefficients within partition b. It must be made sure that the impulse response of each HRTF or BRIR satisfies the condition of (6.5), otherwise more than one STFT spectrum in time needs to be considered for implementing the convolution (6.11). An example of this process of "mixing" HRTFs or BRIRs is shown for the left channel in Figure 6.8. To prevent aliasing artifacts, the transitions between partitions need to be smoothed. This is done by using overlapping spectral windows for the different portions of the HRTFs or BRIRs between the partitions.

6.4 Complexity and delay of the proposed schemes

The complexity of the presented BCC implementation is reasonably low. The most demanding operations are the FFT and inverse FFT (used for implementing the STFT-based filterbank) and, if used, late reverberation generation (Section 6.2.6). The FFTs are of size 1024 for the specific parameters chosen (Section 6.2.1). Implementations of BCC encoder and decoder (without late reverberation generation) run both in real-time on a laptop computer (500 MHz PowerPC G4 processor), each process showing about 5 % processor load for stereo audio input and output. Also a fixed-point version of a BCC decoder was implemented, running in real-time together with a PAC (Sinha *et al.* 1997) decoder on a general purpose digital signal processor.

The algorithmic delay is defined as the delay for encoding and decoding of a signal excluding the time needed for calculations and data transmission. When the sum signal is computed by simple addition of the input signals in the time domain, the sum signal can be fed to the BCC decoder without any delay. For this case, the STFTs of encoder and decoder operate synchronously and

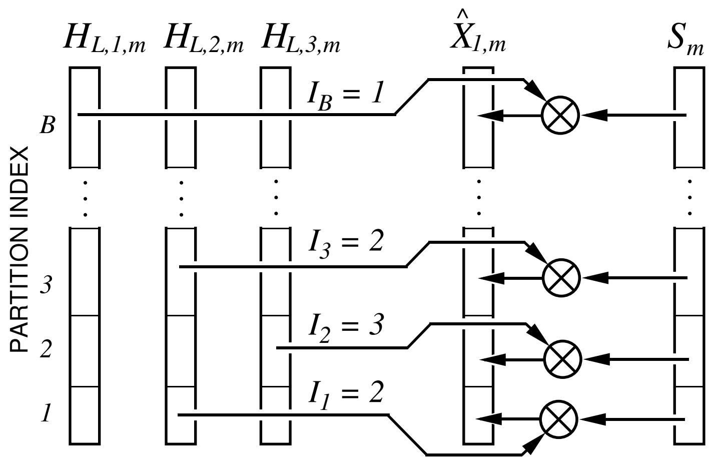


Figure 6.8: The process of generating the left channel of a binaural signal with HRTFs or BRIRs. As a function of the source index $I_b(k)$, portions of different HRTFs or BRIRs are applied in the different partitions. For simplicity, the time index k is ignored in the figure.

no delay compensation for the side information needs to be considered. Thus, the algorithmic delay is determined by just the decoder window size $W = 896$, which corresponds to 28 ms.

If the sum signal is computed in the frequency domain (Section 6.2.2), then the additional delay of the transmitted sum signal is $W/2$ samples. The resulting total algorithmic delay is $3W/2 = 1344$ samples or 42 ms.

It should be noted that the value of W was chosen to be a compromise between quality, delay, and bitrate. Choosing smaller values of W will reduce the delay and will potentially improve quality (Section 3.5.1). However, smaller values of W will also increase the bitrate, assuming the coding schemes described in Sections 6.2.4 and 6.3.2.

6.5 Conclusions

In this chapter, it was described how to implement BCC and BCC for flexible rendering, using an STFT-based time-frequency transform as opposed to a non-uniform filterbank. It was described how to carry out ICTD, ICLD, and ICC synthesis in the STFT domain. Furthermore, it was shown how to implement convolutions with filters of any length in the STFT domain for the purpose of efficient ICC synthesis. A variation of the scheme for BCC for flexible rendering was described, using HRTFs or BRIRs (as opposed to ICTD, ICLD, and ICC) for generating binaural signals. Simple schemes for quantization and coding of the BCC side information were described. The presented schemes have all low complexity and relatively low delay.

Chapter 7

Source Localization in Complex Listening Situations

7.1 Introduction

In Chapter 2, perceptual phenomena related to perception of the auditory spatial image were reviewed. The relation between interaural time difference (ITD) and interaural level difference (ILD) and source direction in free-field is obvious and a conclusion that the auditory system discriminates the source direction as a function of ITD and ILD is in this case rather plausible. However, some phenomena were also described for which it is not obvious how the auditory system processes ear entrance signal properties for localization of sound sources. For example, when a number of sources are concurrently active (as described in Chapter 2.2.5), ITD and ILD are likely to be time varying and in many cases their values do not correspond directly to source directions. In an enclosed space, when sound from sources not only reaches the ears of a listener directly, but also indirectly from different directions (as described in Chapter 2.2.6), the matter of localization of the sources becomes even more complicated. Playback of "real-world" stereo and multi-channel audio signals usually mimics listening to multiple concurrently active sources in rooms. For the task of designing a coding scheme for spatial audio it is helpful to understand which signal properties are important to the auditory system for auditory event localization. These properties need to be maintained when coding stereo and multi-channel audio signals.

In the following, an auditory model for source localization in complex listening scenarios is described. This model is then also discussed related to spatial audio playback and binaural cue coding (BCC). Before describing the proposed model in more detail, related psychophysical localization experiments and psychoacoustic models are reviewed.

Localization accuracy in the presence of concurrent sound from different directions has been investigated by several authors. A detailed review is given by Blauert (1997). The effect of independent distracters on the localization of a target sound has been recently studied by Good and Gilkey (1996), Good *et al.* (1997), Lorenzi *et al.* (1999), Hawley *et al.* (1999), Drullman and Bronkhorst (2000), Langendijk *et al.* (2001), Braasch and Hartung (2002), and Braasch

(2002). The results of these studies generally imply that the localization of the target is either not affected or only slightly degraded by introducing one or two simultaneous distracters at the same overall level as the target. When the number of distracters is increased or the target-to-distracter ratio (T/D) is reduced, the localization performance begins to degrade. However, for most configurations of a target and a single distracter in the frontal horizontal plane, the accuracy stays very good down to a target level only a few dB above the threshold of detection (Good and Gilkey 1996, Good *et al.* 1997, Lorenzi *et al.* 1999). An exception to these results is the outcome of the experiment of Braasch (2002), where two incoherent noises with exactly the same envelope were most of the time not individually localizable.

In order to understand the localization of a source in the presence of reflections from different directions, the precedence effect needs to be considered. Extensive reviews have been given by Zurek (1987), Blauert (1997), and Litovsky *et al.* (1999). The operation of the precedence effect manifests itself in a number of perceptual phenomena: fusion of subsequent sound events into a single perceived entity, suppression of directional discrimination of the later events, as well as localization dominance by the first event. The directional perception of a pair of stimuli with an interstimulus delay shorter than 1 ms is called summing localization. The weight of the lagging stimulus reduces with increasing delay up to approximately 1 ms, and for delays greater than that the leading sound dominates the localization judgment, although the lag might never be completely ignored. Echo threshold refers to the delay where the fusion breaks apart. Depending on stimulus properties and individual listeners, thresholds between 2–50 ms have been reported in the literature (Litovsky *et al.* 1999).

Localization accuracy within rooms has been studied by Hartmann (1983), Rakerd and Hartmann (1985, 1986), and Hartmann and Rakerd (1989) (see also a review by Hartmann 1997). Overall, in these experiments the localization performance was slightly degraded by the presence of reflections. Interestingly, using slow-onset sinusoidal tones and a single reflecting surface, Rakerd and Hartmann (1985) found that the precedence effect sometimes failed completely. In a follow-up study, the relative contribution of the direct sound and the steady state interaural cues to the localization judgment was found to depend on the onset rate of the tones (Rakerd and Hartmann 1986). Nevertheless, absence of an attack transient did not prevent the correct localization of a broadband noise stimulus (Hartmann 1983). Giguère and Abel (1993) reported similar findings for noise with the bandwidth reduced to one-third octave. Rise/decay time had little effect on localization performance except for the lowest center frequency (500 Hz), while increasing the reverberation time decreased the localization accuracy. Braasch *et al.* (2003) investigated the bandwidth dependence further, finding that the precedence effect started to fail when the bandwidth of noise centered at 500 Hz was reduced to 100 Hz.

The auditory system features a number of physical, physiological, and psychological processing stages for accomplishing the task of source direction discrimination and ultimately the formation of the auditory spatial image. The structure of a generic model for spatial hearing is illustrated in Figure 7.1. There is little doubt about the first stages of the auditory system, i.e. the physical and physiological functioning of the outer, middle, and inner ear are

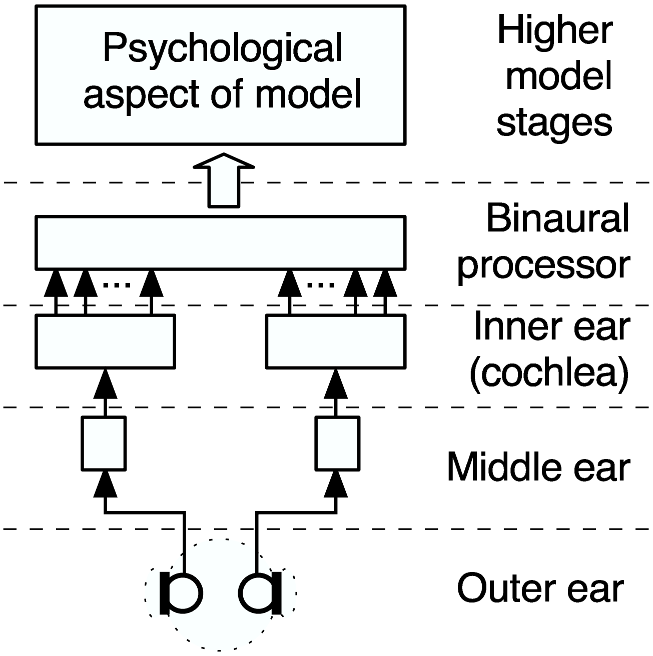


Figure 7.1: A model of spatial hearing covering the physical, physiological, and psychological aspects of the auditory system.

known and understood to a high degree. However, already the stage of the binaural processor is less well known. Different models have used different approaches to explain various aspects of binaural perception. The majority of proposed localization models are based on analysis of ITD cues using a coincidence structure (Jeffress 1948), or a cross-correlation implementation that can be seen as a special case of the coincidence structure. Evidence for cross-correlation-like neural processing has also been found in physiological studies (Yin and Chan 1990). However, such excitation-excitation (EE) type cells are but one kind of neural units potentially useful for obtaining binaural information (see e.g. the introduction and references of Breebaart *et al.* 2001). With current knowledge, the interaction between the binaural processor and higher level cognitive processes can only be addressed through indirect psychophysical evidence.

For a single source in free-field, sound from only one direction arrives at the ears of a listener and thus causally determines the ITD and ILD cues (Gaik 1993), which appear in the auditory system as a result of reflections, diffraction, and resonance effects caused by the head, torso, and the external ears of the listener. However, in complex listening situations, i.e., in the presence of several sound sources and/or room reflections, it often occurs that sound from several different directions concurrently reaches the position of the listener. Furthermore, the superposition of sound emanating from several directions results in instantaneous ITD and ILD cues that most of the time do not correspond to any of the source directions. Nevertheless, humans have a remarkable ability to resolve such complex composites of sound into separate localizable auditory events at directions corresponding to the sound sources.

Few binaural models have specifically considered localization in complex listening situations. To begin with, Blauert and Cobben (1978) investigated a model with the essential features of most current models, including a simulation of the auditory periphery and cross-correlation analysis. In a precedence effect experiment they concluded that the correct cross-correlation peaks were available but the model could not explain how to identify them. Later, Lindemann (1986*a*) extended the model with contralateral and temporal inhibition, combining the analysis of both ITD and ILD cues within a single structure that was shown to be able to simulate several precedence effect phenomena (Lindemann 1986*b*). The model of Lindemann was further extended by Gaik (1993) to take into account naturally occurring combinations of ITD and ILD cues in free-field. A different phenomenological model, using localization inhibition controlled by an onset detector, was proposed by Zurek (1987), and developed into a cross-correlation implementation by Martin (1997). Hartung and Trahiotis (2001) were able to simulate the precedence effect for pairs of clicks without any inhibition, just taking into account the properties of the peripheral hearing. However, this model was not able to predict the localization of continuous narrowband noises in a comparison of several models by Braasch and Blauert (2003). The best results were achieved with a combined analysis of ITD cues with the model of Lindemann (1986*a*) and ILD cues using a modified excitation-inhibition (EI) model (Breebaart *et al.* 2001) extended with temporal inhibition. For independent localization of concurrent sources with non-simultaneous onsets, Braasch (2002) has proposed a cross-correlation difference model.

In this chapter, we propose a single modeling mechanism to explain various aspects of source localization in complex listening situations. The basic approach is very straightforward: only ITD and ILD cues occurring at time instants when they represent the direction of one of the sources are selected, while other cues are ignored. It will be shown that the interaural coherence (IC) can be used as an indicator for these time instants. More specifically, in many cases by selecting ITD and ILD cues coinciding with IC cues larger than a certain threshold, one can obtain a subset of ITD and ILD cues similar to the corresponding cues of each source presented separately in free-field. The proposed cue selection method is implemented in the framework of a model that considers a physically and physiologically motivated peripheral stage, whereas the remaining parts are analytically motivated. Fairly standard binaural analysis is used to calculate the instantaneous ITD, ILD, and IC cues. The presented simulation results reflect psychophysical data from a number of localization experiments cited earlier, involving independent distracters and precedence effect conditions.

The chapter is organized as follows. The binaural model, including the proposed cue selection mechanism, is described in Section 7.2. The simulation results are presented in Section 7.3 with a short discussion of each case related to similar psychophysical studies. Section 7.4 includes a general discussion of the model and results, followed by a discussion in Section 7.5 about the cue selection model in relation to spatial audio playback and BCC. Conclusions are presented in Section 7.6.

7.2 Model description

The model can be divided into three parts: auditory periphery, binaural processor, and higher model stages. In this section, each of the model stages is described in detail, followed by a discussion of the features of the model.

7.2.1 Auditory periphery

Transduction of sound from a source to the ears of a listener is realized by filtering the source signals either with head-related transfer functions (HRTFs) or with measured binaural room impulse responses (BRIRs). HRTF filtering simulates the direction dependent influence of the head and outer ears on the ear input signals. BRIRs additionally include the effect of room reflections in an enclosed space. In multi-source scenarios, each source signal is first filtered with a pair of HRTFs or BRIRs corresponding to the simulated location of the source, and the resulting ear input signals are summed before the next processing stage.

The effect of the middle ear is typically described as a bandpass filter. However, since this chapter is only considering simulations at single critical bands, the frequency weighting effect of the middle ear has been discarded in the model. The frequency analysis of the basilar membrane is simulated by passing the left and right ear signals through a gammatone filterbank (Patterson *et al.* 1995). Each resulting critical band signal is processed using a model of neural transduction as proposed by Bernstein *et al.* (1999). The envelopes of the signals are first compressed by raising them to the power of 0.23. The compressed signals are subjected to half-wave rectification followed by squaring and a 4^{th} order low-pass filtering with a cutoff frequency of 425 Hz. The resulting nerve firing densities at the corresponding left and right ear critical bands are denoted x_1 and x_2. These parts of the model are implemented using the freely available Matlab toolboxes from Slaney (1998) and Akeroyd (2001).

Internal noise is introduced into the model in order to describe the limited accuracy of the auditory system. For this purpose independent Gaussian noise, filtered with the same gammatone filters as the considered critical band signals, is added to each critical band signal before applying the model of neural transduction. The noise is statistically independent for each critical band, as well as for the left and right ears. For the critical band centered at 2 kHz, a sound pressure level (SPL) of 9.4 dB has been chosen according to Breebaart *et al.* (2001) who fitted the level of the noise to describe detection performance near the threshold of hearing. For other critical bands the level is scaled according to the hearing threshold curves (ISO 389 1975). For the 500 Hz band, an SPL of 14.2 dB is used.

7.2.2 Binaural processor

As mentioned in Section 7.1, the present study does not make a specific physiological assumption about the binaural processor. The only assumption is that its output signals (e.g. binaural activity patterns) yield information which can be used by the upper stages of the auditory system for discriminating ITD, ILD, and IC. Given this assumption, the proposed model computes the ITD, ILD, and IC directly. Note that here ITD, ILD, and IC are defined with respect to critical band signals after applying the neural transduction.

The ITD and IC are estimated from the normalized cross-correlation function. Given x_1 and x_2 for a specific center frequency f_c, at the index of each sample n, a running normalized cross-correlation function is computed according to

$$\gamma(n,m) = \frac{a_{12}(n,m)}{\sqrt{a_{11}(n,m)a_{22}(n,m)}} , \tag{7.1}$$

where

$$\begin{aligned} a_{12}(n,m) &= \alpha x_1(n-\max\{m,0\})x_2(n-\max\{-m,0\}) + (1-\alpha)a_{12}(n-1,m) , \\ a_{11}(n,m) &= \alpha x_1(n-\max\{m,0\})x_1(n-\max\{m,0\}) + (1-\alpha)a_{11}(n-1,m) , \\ a_{22}(n,m) &= \alpha x_2(n-\max\{-m,0\})x_2(n-\max\{-m,0\}) + (1-\alpha)a_{22}(n-1,m) , \end{aligned}$$

and $\alpha \in [0,1]$ determines the time-constant of the exponentially decaying estimation window

$$T = \frac{1}{\alpha f_s} , \tag{7.2}$$

where f_s denotes the sampling frequency. $\gamma(n,m)$ is evaluated over time lags in the range of $[-1,1]$ ms, i.e., $\frac{m}{f_s} \in [-1,1]$ ms. The ITD (in samples) is estimated as the lag of the maximum of the normalized cross-correlation function,

$$\tau(n) = \arg\max_m \gamma(n,m) . \tag{7.3}$$

Note that the time resolution of the computed ITD is limited by the sampling interval.

The normalization of the cross-correlation function is introduced in order to get an estimate of the IC, defined as the maximum value of the instantaneous normalized cross-correlation function,

$$c(n) = \max_m \gamma(n,m) . \tag{7.4}$$

This estimate describes the coherence of the left and right ear input signals. In principle, it has a range of $[0,1]$, where 1 occurs for perfectly coherent $x_1(n)$ and $x_2(n)$. However, due to the DC offset of the halfwave rectified signals, the values of $c(n)$ are typically higher than 0 even for independent (non-zero) $x_1(n)$ and $x_2(n)$. Thus, the effective range of the interaural coherence $c(n)$ is compressed from $[0,1]$ to $[a,1]$ by the neural transduction. The compression is more pronounced (larger a) at high frequencies, where the lowpass filtering of the halfwave rectified critical band signals yields signal envelopes with a higher DC offset than in the signal waveforms (Bernstein and Trahiotis 1996).

The ILD is computed as

$$\Delta L(n) = 10\log_{10}\left(\frac{L_2(n,\tau(n))}{L_1(n,\tau(n))}\right) , \tag{7.5}$$

where

$$\begin{aligned} L_1(n,m) &= \alpha x_1^2(n-\max\{m,0\}) + (1-\alpha)L_1(n-1,m) , \\ L_2(n,m) &= \alpha x_2^2(n-\max\{-m,0\}) + (1-\alpha)L_2(n-1,m) . \end{aligned}$$

Note that due to the envelope compression the resulting ILD estimates will be smaller than the level differences between the ear input signals. For coherent ear input signals with a constant level difference, the estimated ILD (in dB) will be 0.23 times that of the physical signals.

The sum of the signal power of x_1 and x_2 that contributes to the estimated ITD, ILD, and IC cues at time index n is

$$p(n) = L_1(n, \tau(n)) + L_2(n, \tau(n)) \,. \tag{7.6}$$

Choosing the time constant T is a difficult task. Studies of binaural detection actually suggest that the auditory system integrates binaural data using a double-sided window with time constants of both sides in the order of $20 - 40$ ms (e.g. Kollmeier and Gilkey 1990). However, a double sided window with this large time constant will not be able to simulate the precedence effect, where the localization of a lead sound should not be influenced by a lagging sound after only a few milliseconds. The difference could be explained by assuming that the auditory system responsible for binaural detection further integrates the binaural data originally derived with a better time resolution. We have chosen to use a single-sided exponential time window with a time constant of 10 ms, in accordance with the time constant of the temporal inhibition of the model of Lindemann (1986*a*).

7.2.3 Higher model stages

A vast amount of information is available to the upper stages of the auditory system through the signals from the auditory periphery. The focus of this study lies only in the analysis of the three inter-channel properties between left and right critical band signals that were defined in the previous section: ITD, ILD, and IC. It is assumed that at each time instant n the information about the values of these three signal properties, $\{\Delta L(n), \tau(n), c(n)\}$, is available for further processing in the upper stages of the auditory system.

Consider the simple case of a single source in free-field. Whenever there is sufficient signal power, the source direction determines the nearly constant ITD and ILD which appear between each left and right critical band signal. The (average) ITDs and ILDs occurring in this scenario are denoted *free-field cues* in the following. The free-field cues of a source with an azimuthal angle ϕ are denoted τ_ϕ and ΔL_ϕ. It is assumed that this kind of a one-source-free-field scenario is the reference for the auditory system. That is, in order for the auditory system to perceive auditory events at the directions of the sources, it must obtain ITD and/or ILD cues similar to the free-field cues corresponding to each source that is being discriminated. The most straightforward way to achieve this is to select the ITD and ILD cues at time instants when they are similar to the free-field cues. In the following it is shown how this can be done with the help of IC.

When several independent sources are concurrently active in free-field, the resulting cue triplets $\{\Delta L(n),\ \tau(n),\ c(n)\}$ can be classified into two groups: (1) Cues arising at time instants when only one of the sources has power in that critical band. These cues are similar to the free-field cues (direction is represented in $\{\Delta L(n), \tau(n)\}$, and $c(n) \approx 1$). (2) Cues arising when multiple sources have non-negligible power in a critical band. In such a case, the pair

$\{\Delta L(n), \tau(n)\}$ does not represent the direction of any single source, unless the superposition of the source signals at the ears of the listener incidentally produces similar cues. Furthermore, when the two sources are assumed to be independent, the cues are fluctuating and $c(n) < 1$. These considerations motivate the following method for selecting ITD and ILD cues. Given the set of all cue pairs, $\{\Delta L(n), \tau(n)\}$, only the subset of pairs is considered which occurs simultaneously with an IC larger than a certain threshold, $c(n) > c_0$. This subset is denoted

$$\{\Delta L(n), \tau(n) | c(n) > c_0\} . \tag{7.7}$$

The same cue selection method is applicable for deriving the direction of a source while suppressing the directions of one or more reflections. When the "first wave front" arrives at the ears of a listener, the evoked ITD and ILD cues are similar to the free-field cues of the source, and $c(n) \approx 1$. As soon as the first reflection from a different direction arrives, the superposition of the source signal and the reflection results in cues that do not resemble the free-field cues of either the source or the reflection. At the same time IC reduces to $c(n) < 1$, since the direct sound and the reflection superimpose as two signal pairs with different ITD and ILD. Thus, IC can be used as an indicator for whether ITD and ILD cues are similar to free-field cues of sources or not (while ignoring cues related to reflections).

For a given c_0 there are several factors determining how frequently $c(n) > c_0$. In addition to the number, strengths, and directions of the sound sources and room reflections, $c(n)$ depends on the specific source signals and on the critical band being analyzed. In many cases, the larger c_0 the more similar the selected cues are to the free-field cues. However, there is a strong motivation to choose c_0 as small as possible while still getting accurate enough ITD and/or ILD cues, because this will lead to the cues being selected more often, and consequently to a larger proportion of the ear input signals contributing to the localization.

It is assumed that the auditory system adapts c_0 for each specific listening situation, i.e., for each scenario with a constant number of active sources at specific locations in a constant acoustical environment. Since the listening situations do not usually change very quickly, it is assumed that c_0 is adapted relatively slowly in time. In Section 7.3.2, it is also argued that such an adaptive process may be related to the buildup of the precedence effect. All simulations reported in this chapter consider only one specific listening situation at a time. Therefore, for each simulation a single constant c_0 is used.

7.2.4 Discussion

The physiological feasibility of the cue selection depends on the human sensitivity to changes in interaural correlation. The topic has been investigated by Pollack and Trittipoe (1959*a*, 1959*b*), Gabriel and Colburn (1981), Grantham (1982), Koehnke *et al.* (1986), Jain *et al.* (1991), Culling *et al.* (2001), and Boehnke *et al.* (2002). These investigations agree in that the sensitivity is highest for changes from full correlation, whereas the estimates of the corresponding just noticeable differences (JNDs) have a very large variance. For narrowband noise stimuli centered at 500 Hz, the reported JNDs range from

0.0007 (Jain *et al.* 1991, fringed condition) to 0.13 (Culling *et al.* 2001) for different listeners and different stimulus conditions. The sensitivity has been generally found to be lower at higher frequencies. However, all the cited studies have measured sensitivity to correlation of the ear input waveforms instead of correlation computed after applying a model of neural transduction. As discussed in Section 7.2.2, the model of Bernstein *et al.* (1999) reduces the range of IC, indicating overall lower JNDs of IC as defined in this chapter. Furthermore, the model has been specifically fitted to yield constant thresholds at different critical bands when applied to prediction of binaural detection based on changes in IC (Bernstein and Trahiotis 1996). With these considerations it can be concluded that at least the JNDs reported by Gabriel and Colburn (1981), Koehnke *et al.* (1986), and Jain *et al.* (1991) are within the range of precision needed for the simulations in Section 7.3.

The auditory system may not actually use a hard IC threshold for selecting or discarding binaural cues. Instead of pure selection, similar processing could be implemented as an IC based weighting of ITD and ILD cues with a slightly smoother transition. However, the simple selection criterion suffices to illustrate the potential of the proposed method, as will be shown in Section 7.3. Interestingly, van de Par *et al.* (2001) have argued that the precision needed for normalization of the cross-correlation function is so high that it is unlikely that the auditory system is performing the normalization *per se*. Since normalized cross-correlation, nevertheless, describes the perception of IC well, it will be utilized in this chapter.

The cue selection can also be seen as a multiple looks approach for localization. Multiple looks have been previously proposed to explain monaural detection and discrimination performance with increasing signal duration (Viemeister and Wakefield 1991). The idea is that the auditory system has a short term memory of “looks” at the signal, which can be accessed and processed selectively. In the case of localization, the looks would consist of momentary ITD, ILD, and IC cues. With an overview of a set of recent cues, ITDs and ILDs corresponding to high IC values could be adaptively selected.

7.3 Simulation results

As mentioned earlier, it is assumed that in order to perceive an auditory event at a certain direction, the auditory system needs to obtain cues similar to the free-field cues corresponding to a source at that direction. In the following, the proposed cue selection is applied to several stimuli that have been used in previously published psychophysical studies. In all cases both the selected cues as well as all cues prior to the selection are illustrated, and the implied directions are discussed in relation to the literature.

The effectiveness of the proposed cue selection is assessed using a number of statistical measures. The biases of the ITD and ILD cues with respect to the free-field cues τ_ϕ and ΔL_ϕ are defined as

$$\begin{aligned} b_\tau &= \left|E\{\tau(n)\} - \tau_\phi\right| \\ b_{\Delta L} &= \left|E\{\Delta L(n)\} - \Delta L_\phi\right| , \end{aligned} \tag{7.8}$$

respectively, and the corresponding standard deviations are given by

$$\begin{aligned}\sigma_\tau &= \sqrt{E\{(\tau(n) - E\{\tau(n)\})^2\}}\\ \sigma_{\Delta L} &= \sqrt{E\{(\Delta L(n) - E\{\Delta L(n)\})^2\}}\,.\end{aligned} \tag{7.9}$$

The biases and standard deviations are computed considering only the selected cues (7.7). When there is more than one source to be discriminated, these measures are estimated separately for each source by grouping the selected cues at each time instant with the source known to have free-field cues closest to their current values.

For many cases, the larger the cue selection threshold c_0, the smaller the bias and standard deviation. The choice of c_0 is a compromise between the similarity of the selected cues to the free-field cues and the proportion of the ear input signals contributing to the resulting localization. The proportion of the signals contributing to the localization is characterized with the fraction of power represented by the selected parts of the signals, given by

$$p_0 = \frac{E\{p(n)w(n)\}}{E\{p(n)\}}\,, \tag{7.10}$$

where $p(n)$ is defined in (7.6) and the weighting function $w(n)$ is

$$w(n) = \begin{cases} 1, & \text{if } c(n) > c_0 \\ 0, & \text{otherwise} \end{cases}\,. \tag{7.11}$$

In this chapter, the cue selection is only considered independently at single critical bands. Except for different values of c_0, the typical behavior appears to be fairly similar at critical bands with different center frequencies. For most simulations, we have chosen to use the critical bands centered at 500 Hz and/or 2 kHz. At 500 Hz the binaural processor operates on the input waveforms, whereas at 2 kHz the model of auditory periphery extracts the envelopes of the input signals and feeds them to the binaural processor. Where appropriate, results for other critical bands are also shown or briefly discussed. However, considering the way the auditory system eventually combines information from different critical bands is beyond the scope of this thesis. As mentioned earlier, the simulations are carried out with a single constant cue selection threshold c_0 for each case. It is assumed that the auditory system has already adapted c_0 to be effective for the specific listening situation. Unless otherwise noted, the specific c_0 was chosen such that a visual inspection of the simulation results implies an effective cue selection.

Two kinds of plots are used to illustrate the cue selection. In some cases the instantaneous ITD and ILD values are plotted as a function of time, marking the values which are selected. For other examples, the effect of the cue selection is visualized by plotting short-time estimates of *probability density functions* (PDFs) of the selected ITD and ILD cues. Unless otherwise noted, the PDFs are estimated by computing histograms of ITD and ILD cues for a time span of 1.6 s. The height of the maximum peak is normalized to one in all PDFs. In both types of plots, free-field cues resulting from simulations of the same source signals without concurrent sound sources or reflections, are also indicated.

Listening situations in free-field are simulated using HRTFs measured with the KEMAR dummy head with large pinnae, taken from the CIPIC HRTF

Database (Algazi *et al.* 2001). All simulated sound sources are located in the frontal horizontal plane, and, unless otherwise noted, all the stimuli are aligned to 60 dB SPL averaged over the whole stimulus length.

7.3.1 Independent sources in free-field

In this section, the cue selection method is applied to independent stimuli in an anechoic environment. As the first example, the operation of the selection procedure is illustrated in detail for the case of independent speech sources at different directions. Subsequently, simulation results of the effect of target-to-distracter ratio (T/D) on localization of the target stimulus are presented.

Concurrent speech

Localization of a speech target in the presence of one or more competing speech sources has been investigated by Hawley *et al.* (1999) and Drullman and Bronkhorst (2000). Drullman and Bronkhorst (2000) utilized an anechoic virtual environment using both individualized and non-individualized HRTFs for binaural reproduction of the stimuli. They reported slight but statistically significant degradation in localization performance when the number of competing talkers was increased beyond 2. The experiment of Hawley *et al.* (1999), on the other hand, was conducted in a "sound-field room" (reverberation time of approximately 200 ms), as well as using headphone reproduction of the stimuli recorded binaurally in the same room. While not strictly anechoic, their results are also useful for evaluating our anechoic simulation results. Hawley *et al.* (1999) found that apart from occasional confusions between the target and the distracters, increasing the number of competitors from 1 to 3 had no significant effect on localization accuracy. As discussed in Section 7.1, room reflections generally make the localization task more difficult, so a similar or a better result would be expected to occur in an anechoic situation. Note that the overall localization performance reported by Drullman and Bronkhorst (2000) was fairly poor, and the results may have been affected by a relatively complex task requiring listeners to recognize the target talker prior to judging its location.

Based on the previous discussion, the cue selection has to yield ITD and ILD cues similar to the free-field cues of each of the speech sources in order to correctly predict the directions of the perceived auditory events. Three simulations were carried out with 2, 3, and 5 concurrent speech sources. The signal of each source consisted of a different phonetically balanced sentence from the Harvard IEEE list (IEEE 1969) recorded by the same male speaker. As the first case, 2 speech sources were simulated at azimuthal angles of $\pm 40°$. Figure 7.2 shows the IC, ILD, and ITD as a function of time for the critical bands with center frequencies of 500 Hz and 2 kHz. The free-field cues which would occur with a separate simulation of the sources at the same angles are indicated with the dashed lines. The selected ITD and ILD cues (7.7) are marked with bold solid lines. Thresholds of $c_0 = 0.95$ and $c_0 = 0.99$ were used for the 500 Hz and 2 kHz critical bands, respectively, resulting in 65 % and 54 % selected signal power (7.10). The selected cues are always close to the free-field cues, implying perception of two auditory events located at the directions of the sources, as reported in the literature. As expected, due to the

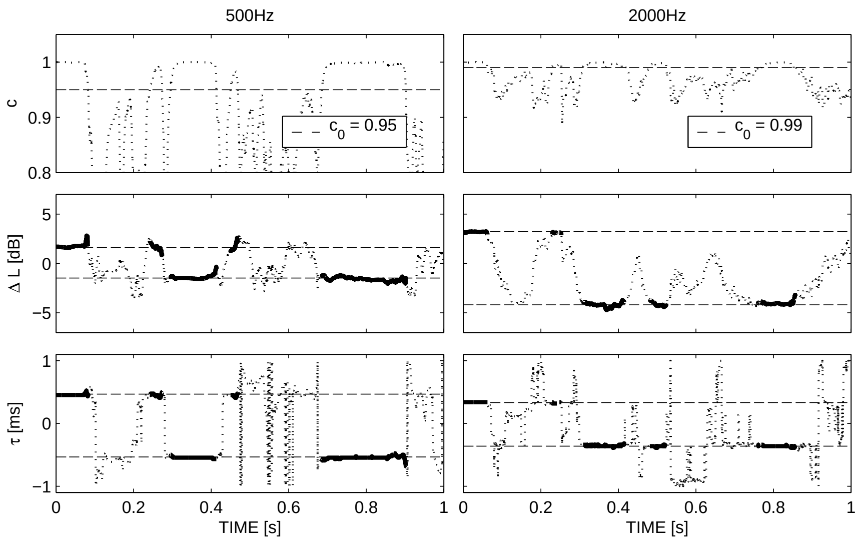


Figure 7.2: IC, ILD, and ITD as a function of time for two independent speech sources at $\pm 40^\circ$ azimuth. Left column: 500 Hz, and right column: 2 kHz critical band. The cue selection thresholds (top row) and the free-field cues of the sources (middle and bottom rows) are indicated with dashed lines. Selected cues are marked with bold solid lines.

neural transduction IC has a smaller range at the 2 kHz critical band than at the 500 Hz critical band. Consequently, a larger c_0 is required.

The performance of the cue selection was assessed as a function of c_0 for the same two speech sources and the critical bands with center frequencies of 250, 500, 1000, 2000, and 3000 Hz. Figure 7.3 shows the ITD and ILD biases (7.8) and standard deviations (7.9), as well as the fraction of signal power corresponding to the selected cues (7.10) as a function of c_0. The biases and standard deviations were computed for both sources separately, as described earlier, and then averaged over 1.6 s of the signals. The graphs indicate that both the biases and the standard deviations decrease with increasing c_0. Thus, the larger the c_0, the closer the obtained cues are to the reference free-field values. Furthermore, the selected signal power decreases gradually until fairly high values of c_0. The general trend of having higher absolute ILD errors at high frequencies is related to the overall larger range of ILDs occurring at high frequencies due to more efficient head shadowing.

The simulation with 3 independent talkers was performed with speech sources at 0° and $\pm 30^\circ$ azimuth, and the simulation of 5 talkers with two additional sources at $\pm 80^\circ$ azimuth. In both cases the results were fairly similar at different critical bands, so the data are only shown for the 500 Hz band. Panels (A) and (B) of Figure 7.4 show PDFs of ITD and ILD without the cue selection for the 3 and 5 speech sources, respectively. Panels (C) and (D) of Figure 7.4 show similar PDFs of the selected cues. The selection threshold was set at $c_0 = 0.99$ corresponding to 54 % selected signal power for the 3 sources and 22 % for the 5 sources. In both cases, even the PDFs considering all

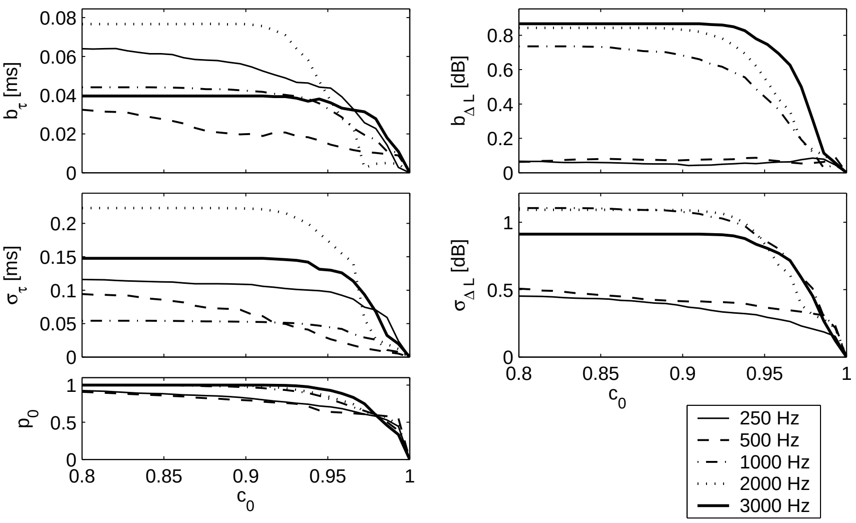


Figure 7.3: ITD and ILD bias, standard deviation, and relative power of the selected signal portions as a function of the cue selection threshold c_0 for two independent speech sources. Data is shown for the 250, 500, 1000, 2000, and 3000 Hz critical bands.

cues show ITD peaks at approximately correct locations, and the cue selection can be seen to enhance the peaks. With the cue selection, the widths of the peaks (i.e. the standard deviations of ITD and ILD) in the 3 source case are as narrow as in separate one-source free-field simulations, which implies robust localization of three auditory events corresponding to the psychophysical results of Hawley *et al.* (1999) and Drullman and Bronkhorst (2000). In the case of 5 sources, the peaks get slightly broader. The ITD peaks are still narrow and correctly located but at the 500 Hz critical band, the range of ILD cues is insufficient for distinct peaks to appear along the ILD axis. This result is also in line with the classic duplex theory (Rayleigh 1907) of sound localization, stating that at low frequencies ITD cues are more salient than ILD cues.

Click-train and noise

Good and Gilkey (1996) and Good *et al.* (1997) studied the localization of a click-train target in the presence of a simultaneous noise distracter. Using loudspeaker reproduction in an anechoic chamber, localization performance was shown to degrade monotonously with a decreasing target-to-distracter ratio (T/D). The investigated T/D ratios were defined relative to the individual detection threshold of each listener for the case when the target sound was presented from the same direction as the distracter. With a target level just a few dB above the detection threshold, localization performance in the left-right direction (e.g. frontal horizontal plane) was still found to be nearly as good as without the distracter. The degradation started earlier and was more severe for the up-down and front-back directions. The results for the left-right direction were later confirmed by Lorenzi *et al.* (1999), who conducted a sim-

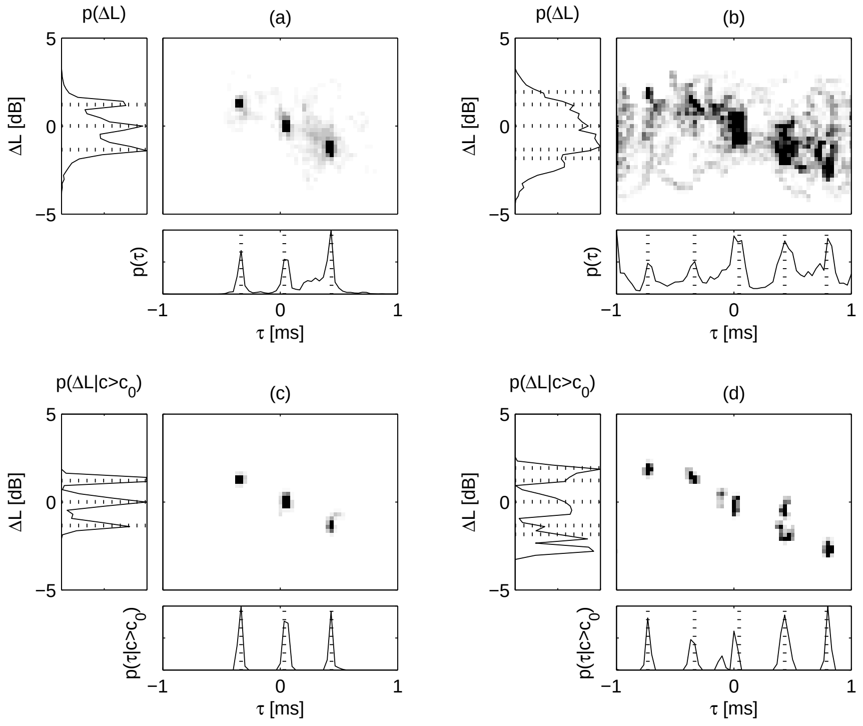


Figure 7.4: PDFs of ITD and ILD for 3 (A) and 5 (B) independent speech sources and corresponding PDFs when cue selection is applied (C and D). The values of the free-field cues for each source are indicated with dotted lines. Data is shown for the 500 Hz critical band.

ilar experiment with sound sources in the frontal horizontal plane. However, the detection levels of Lorenzi *et al.* (1999) were slightly higher, maybe due to utilization of a sound-treated chamber instead of a strictly anechoic environment. Furthermore, Lorenzi *et al.* (1999) found a degradation in performance when the stimuli were lowpass filtered at 1.6 kHz, unlike when the stimuli were highpass filtered at the same frequency.

A simulation was carried out with a white noise distracter directly in front of the listener and a click-train target with a rate of 100 Hz located at 30° azimuth. Assuming a detection level of -11 dB (the highest value in Good *et al.* 1997), the chosen absolute T/Ds of -3, -9, and -21 dB correspond to the relative T/Ds of 8, 2, and -10 dB, respectively, as investigated by Good and Gilkey (1996). The PDFs for the critical band centered at 500 Hz did not yield a clear peak corresponding to the direction of the click train. Motivated by the fact that in this case higher frequencies are more important for directional discrimination (Lorenzi *et al.* 1999), we investigated further the 2 kHz critical band. Panels (A)-(C) of Figure 7.5 show PDFs of ITD and ILD without the cue selection for the selected T/D ratios. Corresponding PDFs obtained by the cue selection (7.7) are shown in panels (D)-(F). The thresholds for the panels (D)-(F) were $c_0 = 0.990$, $c_0 = 0.992$, and $c_0 = 0.992$, respectively, resulting in

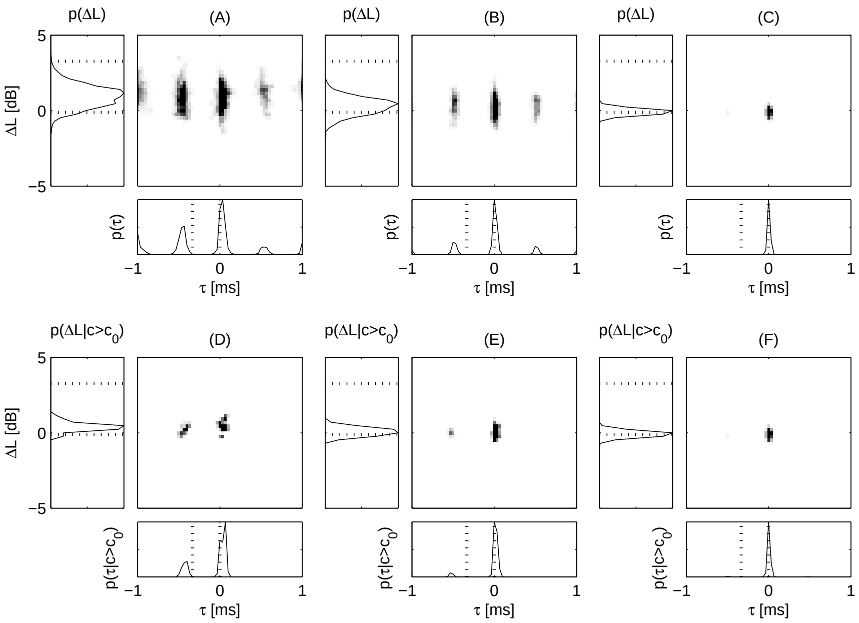


Figure 7.5: PDFs of ITD and ILD for a click-train and white Gaussian noise at different T/D ratios: -3, -9, -21 dB (A-C), and the corresponding PDFs when cue selection is applied (D-F). The values of the free-field cues are indicated with dotted lines. Data is shown for the 2 kHz critical band.

3 %, 9 %, and 99 % of the signal power being represented by the selected cues.

The PDFs in Figure 7.5 imply that the target is localized as a separate auditory event for the T/D ratios of -3 dB and -9 dB. However, for the lowest T/D ratio the target click-train is no longer individually localizable, as also suggested by the results of Good and Gilkey (1996). In panels (A) and (B), ITD peaks are seen to rise at regular intervals due to the periodicity of the cross-correlation function, while the cue selection suppresses the periodical peaks as shown in panels (D) and (E). Note that when the click-train is individually localizable, only the recovered ITD cues are close to the free-field cues of both sources, whereas a single broad ILD peak appears. This is in line with the findings of Braasch (2003) that in the presence of a distracter, ILDs are less reliable cues for localization, and that ITDs also gain more importance in the subjective localization judgment. The ITD peaks corresponding to the click-train are also shifted away from the distracter. Such a pushing effect caused by a distracter in front of the listener was observed for one listener in a similar experiment (Lorenzi *et al.* 1999) and for most listeners when the target was an independent noise signal (Braasch and Hartung 2002). On the contrary, Good and Gilkey (1996) reported a pulling effect, which was also the case for two listeners in the experiment of Lorenzi *et al.* (1999).

7.3.2 Precedence effect

This section illustrates the cue selection within the context of the precedence effect. Pairs of clicks are used to demonstrate the results for wideband signals (in this case a signal with at least the width of a critical band). Sinusoidal tones are simulated with different onset rates and the cues obtained during the onset are shown to agree with results reported in the literature.

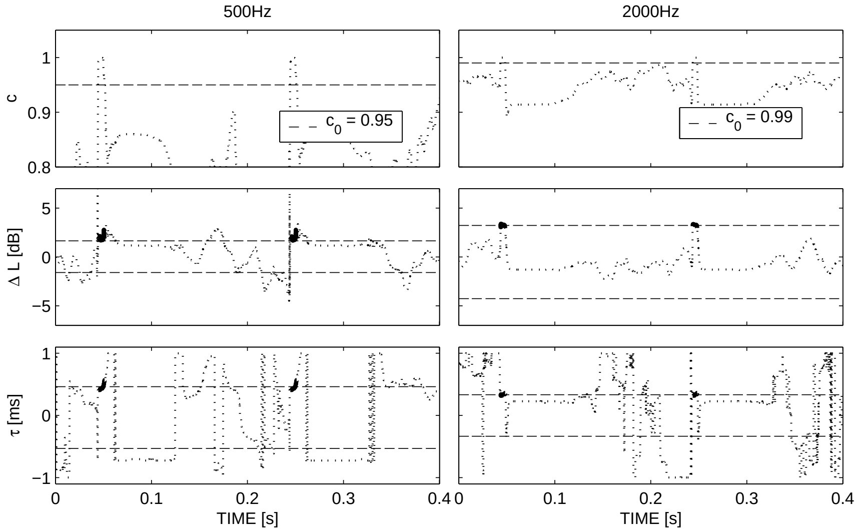


Figure 7.6: IC, ILD, and ITD as a function of time for a lead/lag click-train with a rate of 5 Hz and an ICI of 5 ms. Left column: 500 Hz, and right column: 2 kHz critical band. The cue selection thresholds (top row) and the free-field cues of the sources (middle and bottom rows) are indicated with dashed lines. Selected cues are marked with bold solid lines.

Pairs of clicks

In a classical precedence effect experiment, a lead/lag pair of clicks is presented to the listener (Blauert 1997, Litovsky *et al.* 1999). The leading click is first emitted from one direction, followed by another identical click from another direction after an *interclick interval* (ICI) of a few milliseconds. As discussed in Section 7.1, the directional perception changes depending on ICI.

Figure 7.6 shows IC, ILD, and ITD as a function of time for a click train with a rate of 5 Hz analyzed at the critical bands centered at 500 Hz and 2 kHz. The lead source is simulated at 40° and the lag at −40° azimuth with an ICI of 5 ms. As expected based on earlier discussion, IC is close to one whenever only the lead sound is within the analysis time window. As soon as the lag reaches the ears of the listener, the superposition of the two clicks reduces the IC. The cues obtained by the selection with $c_0 = 0.95$ for the 500 Hz and $c_0 = 0.985$ for the 2 kHz critical band are shown in the figure, and the free-field cues of both sources are indicated with dashed lines. The selected cues are close to the

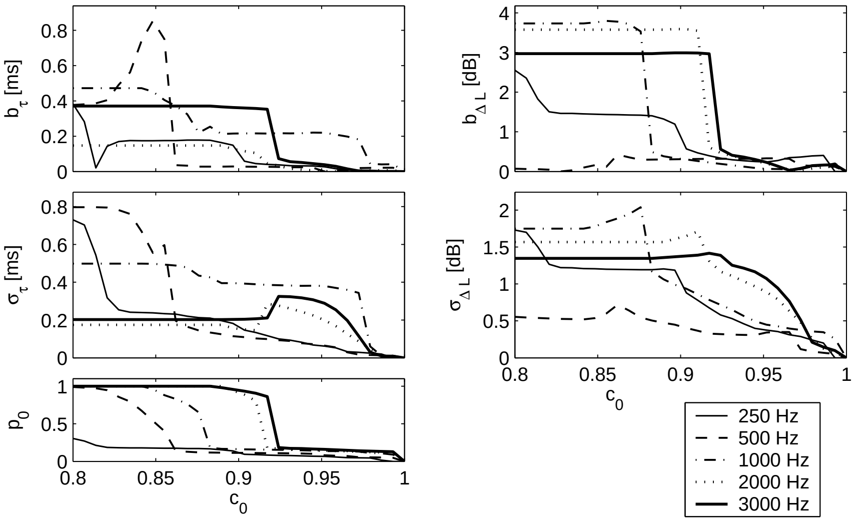


Figure 7.7: ITD and ILD bias, standard deviation, and relative power of the selected signal portions as a function of the cue selection threshold c_0 for lead/lag click-train. Data is shown for the 250, 500, 1000, 2000, and 3000 Hz critical bands.

free-field cues of the leading source and the cues related to the lag are ignored, as is known to happen based on psychophysical studies (Litovsky *et al.* 1999). The fluctuation in the cues before each new click pair is due to the internal noise of the model.

The performance of the cue selection was again assessed as a function of c_0 for the critical bands with center frequencies of 250, 500, 1000, 2000, and 3000 Hz. The statistical measures were calculated from a 1.6 s signal segment. Figure 7.7 shows ITD and ILD biases (7.8) and standard deviations (7.9), as well as the power of the selected cues (7.10) as a function of c_0. The biases and standard deviations were computed related to the free-field cues of the leading source, since localization of the lag should be suppressed if the selection works correctly. Both the biases and standard deviations decrease as c_0 increases. Thus the larger the cue selection threshold c_0, the more similar the selected cues are to the free-field cues of the leading source.

At a single critical band, the energy of the clicks is spread over time due to the gammatone filtering and the model of neural transduction. Therefore, with an ICI of 5 ms, a large proportion of the critical band signals related to the clicks of a pair is overlapping, and only a small part of the energy of the lead click appears in the critical band signals before the lag. Consequently, the relative signal power corresponding to the selected cues is fairly low when requiring small bias and standard deviation, as can be seen in the left bottom panel of Figure 7.7.

Localization as a function of ICI The previous experiment was repeated for ICIs in the range of 0 – 20 ms using the 500 Hz critical band. The chosen range of delays includes summing localization, localization suppression, and

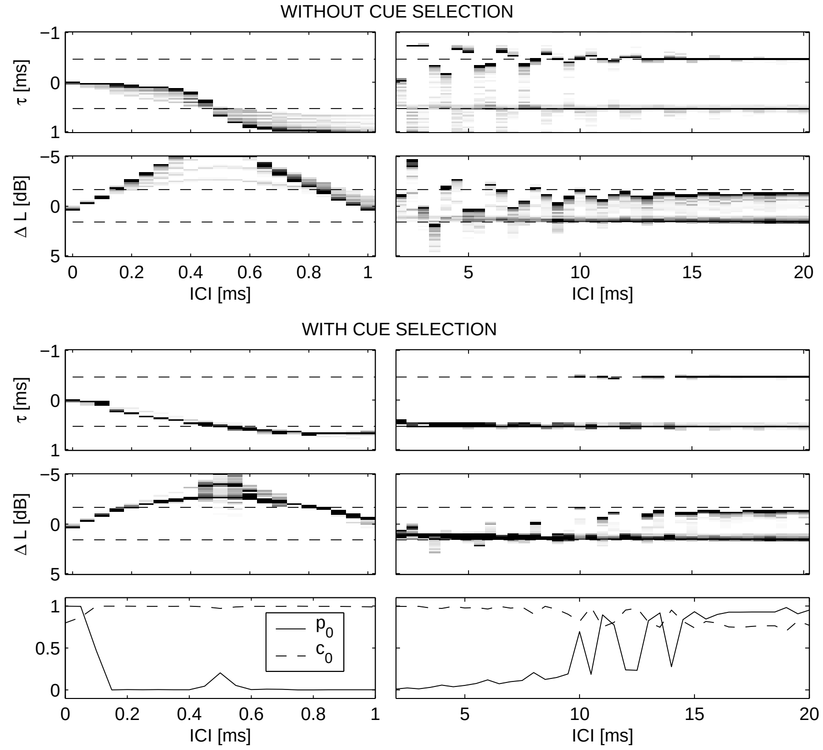


Figure 7.8: PDFs of ITD and ILD as a function of ICI for a click-train: Without cue selection (rows 1 and 2) and with cue selection (rows 3 and 4). Cue selection threshold c_0 and relative power p_0 of the selected signal portion (bottom row).

independent localization of both clicks without the precedence effect (Litovsky *et al.* 1999). For all previous simulations, a suitable c_0 was chosen as a compromise between similarity of the cues to free-field cues and how frequently cues are selected. Here, each ICI corresponds to a different listening situation, since the different delays of the lag imply different acoustical environments. It is thus expected that the most effective c_0 may also differ depending on ICI.

Several different criteria for determining c_0 were assessed. Indeed, using the same c_0 for all ICIs did not yield the desired results. The criterion of adapting c_0 such that the relative power of the selected cues (7.10) had the same value for each simulation did not yield good results either. Thus, a third criterion was adopted. The cue selection threshold c_0 was determined numerically for each simulation such that σ_τ (the narrowness of the peaks in the PDFs of ITD) was equal to 15 μs. This could be explained with a hypothetical auditory mechanism adapting c_0 in time with the aim of making ITD and/or ILD standard deviation sufficiently small. Small standard deviations indicate small fluctuations of the selected cues in time and thus non-time-varying localization of auditory events. The resulting PDFs of ITD and ILD as a function of ICI with and without the cue selection are shown in Figure 7.8.

The PDFs without the cue selection (rows 1 and 2 in Figure 7.8) indicate two independently localized auditory events for most ICIs above 1 ms. Furthermore, the predicted directions depend strongly on the delay. On the contrary, the PDFs with the cue selection show that the selected cues correctly predict all the three phases of the precedence effect (summing localization, localization suppression, and independent localization). At delays less than approximately 1 ms the ITD peak moves to the side as the delay increases, as desired, but the ILD cues do not indicate the same direction as the ITD cues. However, this is also in line with existing psychophysical literature. Anomalies of the precedence effect have been observed in listening tests with bandpass filtered clicks (Blauert and Cobben 1978), suggesting a contribution of the extracted misleading ILDs to the localization judgment. For delays within the range of approximately $1 - 10$ ms there is only one significant peak in the PDFs, indicating localization in the direction of the lead. For larger delays two peaks appear, suggesting two independently localized auditory events. Note that the fusion of two clicks has been found to sometimes break down earlier, but 10 ms is within the range of reported critical thresholds for localization dominance (Litovsky *et al.* 1999, Litovsky and Shinn-Cunningham 2001).

The bottom row of Figure 7.8 shows the selection threshold c_0 and the relative power p_0 of the signal corresponding to the selected cues as a function of the ICI. For most ICIs up to approximately 8 ms, the relative power of the selected signal portion almost vanishes. However, there is one characteristic peak of p_0 at approximately 0.5 ms. The experiment was repeated for a number of critical bands in the range of 400 to 600 Hz with the observation that the location of this peak moves along the ICI axis as a function of the center frequency of the considered critical band. Furthermore, the general trends of the selected cues were very similar to those at the 500 Hz band in that they all strongly implied the three phases of the precedence effect. Thus, by considering a number of critical bands the three phases of the precedence effect can indeed be explained by the cue selection such that at each ICI a signal portion with non-vanishing power is selected.

Cue selection threshold and precedence buildup For the previous experiment, it was hypothesized that the criterion for determining c_0 is the standard deviation of ITD and/or ILD. The computation of these quantities involves determining the number of peaks (i.e. the number of individually localized auditory events) adaptively in time, which might be related to the buildup of precedence. A buildup occurs when a lead/lag stimulus with ICI close to the echo threshold is repeated several times. During the first few stimulus pairs, the precedence effect is not active and two auditory events are independently perceived. After the buildup, the clicks merge to a single auditory event in the direction of the lead (Freyman *et al.* 1991). An adaptive process determining c_0 would require a certain amount of stimulus activity and time until an effective c_0 is determined and it could thus explain the time-varying operation of the precedence effect.

The precedence effect literature also discusses a breakdown of precedence when, for instance, the directions of the lead and lag are suddenly swapped (Clifton 1987, Blauert 1997, Litovsky *et al.* 1999). However, more recent results of Djelani and Blauert (2001, 2002) indicate that the buildup is direction-specific, suggesting further that what has been earlier reported as breakdown of precedence is rather a consequence of precedence not being built up for a new lag direction. Djelani and Blauert (2002) also showed that without stimulus activity the effect of the buildup decays slowly by itself, which supports the idea of an adaptive c_0. In order to model the direction-specific buildup, c_0 would also need to be defined as a function of direction. However, testing and developing the corresponding adaptation method is beyond the scope of this thesis and will be part of the future work.

Onset rate of a sinusoidal tone

Rakerd and Hartmann (1986) investigated the effect of the onset time of a 500 Hz sinusoidal tone on localization in the presence of a single reflection. In the case of a sinusoidal tone, the steady state ITD and ILD cues result from the coherent sum of the direct and reflected sound at the ears of a listener. Often these cues do not imply the direction of either the direct sound or the reflection. Rakerd and Hartmann (1986) found that the onset rate of the tone was a critical factor in determining how much the misleading steady state cues contributed to the localization judgment of human listeners. For fast onsets, localization was based on the correct onset cues, unlike when the level of the tone raised slowly. The cue selection cannot, as such, explain the discounting of the steady state cues, which always have IC close to one. However, considering just the onsets the following results reflect the psychophysical findings of Rakerd and Hartmann (1986).

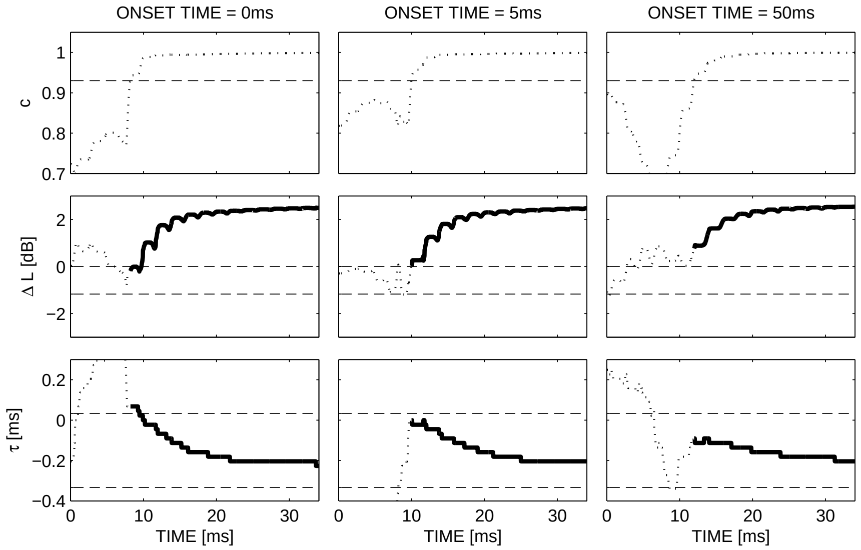


Figure 7.9: IC, ILD, and ITD as a function of time for a 500 Hz sinusoidal tone and one reflection. The columns from left to right show results for onset times of 0 ms, 5 ms, and 50 ms. The cue selection threshold of $c_0 = 0.95$ (top row) and the free-field cues of the source and the reflection (middle and bottom rows) are indicated with dashed lines. Selected cues are marked with bold solid lines. Data is shown for the 500 Hz critical band.

Figure 7.9 shows IC, ILD, and ITD as a function of time for a 500 Hz tone with onset times of 0, 5, and 50 ms. The simulated case corresponds approximately to the "WDB room" and "reflection source 6" condition reported by Rakerd and Hartmann (1986). The direct sound is simulated in front of the listener, and the reflection arrives with a delay of 1.4 ms from an azimuthal angle of 30°. A linear onset ramp is used and the steady state level of the tone is set to 65 dB SPL. The ITD and ILD cues selected with a threshold of $c_0 = 0.93$ are marked with bold solid lines and the free-field cues of the direct sound and the reflection are indicated with dashed lines. Note that the direct sound reaches the ears of the listener at approximately 7 ms. For onset times of 0 and 5 ms, ITD and ILD cues are similar to the free-field cues at the time when IC reaches the threshold. However, with an onset time of 50 ms the ITD and ILD cues no longer correspond to the free-field cues, which is suggested by the degraded localization performance in the experiment of Rakerd and Hartmann (1986).

In order to predict the final localization judgment, another selection mechanism would be needed to only include the localization cues at the time instants when the cue selection becomes effective. The dependence on the onset rate can be explained by considering the input signals of the binaural processor. During the onset, the level of the reflected sound follows that of the direct sound with a delay of 1.4 ms. Thus, the slower the onset, the smaller the difference. The critical moment is when the level of the direct sound rises high enough above the level of the internal noise to yield IC above the selection threshold. If the reflection has non-negligible power at that time, localization cues will be biased to the steady state direction already when the selection begins.

7.3.3 Independent sources in reverberant environment

As a final test for the model, the localization of 1 and 2 speech sources was simulated in a reverberant environment. The utilized BRIRs were measured with a Neumann KU 80 dummy head in an empty lecture hall with reverberation times of 2.0 and 1.4 s at the octave bands centered at 500 and 2000 Hz, respectively. The same phonetically balanced speech samples as used in Section 7.3.1 were convolved with BRIRs simulating sources at 30° azimuth for the case of one source and ±30° for the two sources. The case of two talkers included again two different sentences uttered by the same male speaker. For computing the free-field cues, the BRIRs were truncated to 2.3 ms, such that the effect of the reflections was ignored.

The chosen hall is a very difficult case for localization due to lots of diffuse reflections from the tables and benches all around the simulated listening position. At the 500 Hz critical band, the ITD and ILD cues prior to the selection did not yield any meaningful data for localization. The cue selection resulted in high peaks close to the free-field cues, but it was not able to suppress all other peaks implying different directions. A subsequent investigation showed that these erroneous peaks appear at different locations at different critical bands. Thus, processing of localization information across critical bands should be able to further suppress them. At 2 kHz, the results for a single critical band were clearer and they will be illustrated here.

Panels (A) and (B) of Figure 7.10 show PDFs of ITD and ILD without the cue selection, and panels (C) and (D) show the corresponding PDFs of

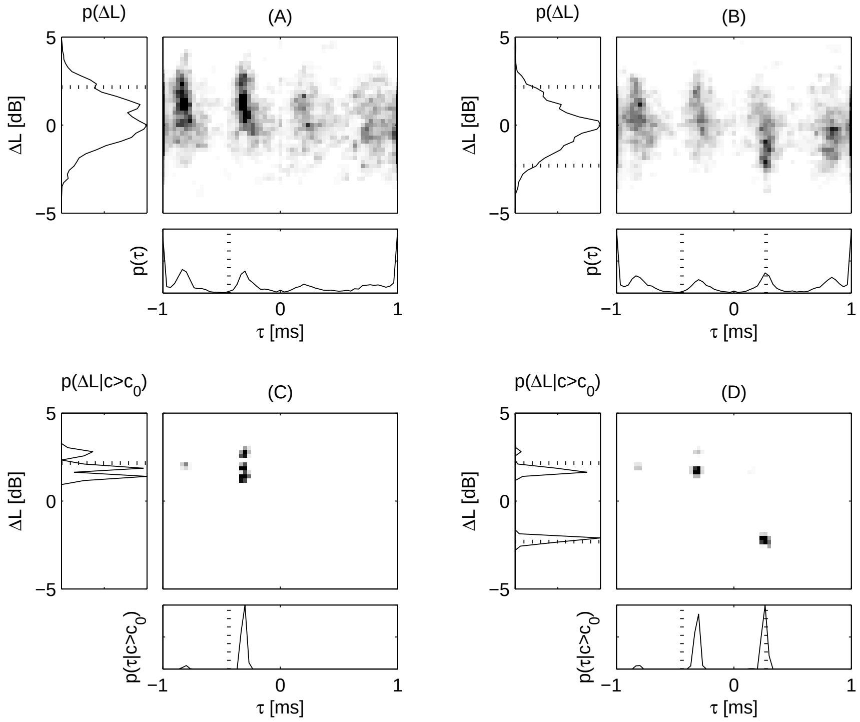


Figure 7.10: PDFs of ITD and ILD for 1 (A) and 2 (B) speech sources in a reverberant hall and the corresponding PDFs when cue selection is applied (C and D). The values of the free-field cues for each source are indicated with dotted lines. Data is shown for the 2 kHz critical band.

the selected cues. Since the cue selection in this case samples the ITD and ILD relatively infrequently, the PDFs were computed considering 3 s of signal. Similar results are obtained when the PDFs are computed from different time intervals. The cue selection criterion for both the 1 and 2 source scenarios was $c_0 = 0.99$, resulting in 1 % of the signal power corresponding to the selected cues. Without the cue selection, the PDFs do not yield much information for localization in either of the cases. Periodicity of the cross-correlation function is clearly visible and it is difficult to distinguish between the one and two source cases. However, with the cue selection sharp peaks arise relatively close to the free-field cues. In the two source case, the right source is practically correctly localized, whereas the ITD cues of the left source are slightly biased towards the center. Note that contrary to the results in Section 7.3.1, the localization is in this case shifted towards the competing sound source. As discussed, also this kind of a pulling effect has been reported in psychoacoustical studies (Good and Gilkey 1996, Lorenzi *et al.* 1999, Braasch and Hartung 2002).

7.4 General discussion

In the previous sections, the selection of ITD and ILD cues based on IC was introduced into a localization model and applied to simulations of a number of complex listening scenarios. In comparison to several existing localization models, a significant difference in the proposed method is the way that the signal power at each time instant affects the localization judgment. In models not designed for complex listening situations, the localization cues and subsequently the final localization judgment are often derived from a time window including the whole stimulus, or of a time integration of a binaural activity pattern computed with running non-normalized cross-correlation. In such cases, the contribution of each time instant to the final localization depends on the instantaneous power. In our approach, only the cues during the selected time instants contribute to localization. Thus the model can in many cases neglect localization information corresponding to time instants with high power, if the power is high due to concurrent activity of several sound sources (or concurrent activity of sources and reflections). The relative power of individual sources also affects how often ITD and ILD cues corresponding to each source are selected.

The proposed model also bears resemblance to earlier models of the precedence effect. The temporal inhibition of the model of Lindemann (1986*a*) tends to hold the highest peaks of the running cross-correlation function (calculated with the stationary inhibition that incorporates ILDs into the model). The higher a peak (i.e. the higher the IC at the corresponding time instant), the stronger the temporal inhibition. The cue selection achieves a somewhat similar effect without a need for an explicit temporal inhibition mechanism, since the localization suppression is directly related to the IC estimated with a similar time window. However, the effect can also be quite different in some scenarios. Whereas the model of Lindemann (1986*a*) only "remembers" the peaks corresponding to a high IC for a short time (time constant of 10 ms), the cue selection with a slowly varying c_0 has a much longer memory. The frequency of the time instants when the direct sound of only one source dominates within a critical band depends on the complexity of the listening situation. In complex cases (e.g. Section 7.3.3), only a small fraction of the ear input signals contribute to localization, and new localization information may be acquired relatively infrequently. We, nevertheless, assume that it is the cues at these instants of time that determine the source localization. During the time when no cues are selected, the localization of the corresponding auditory events is assumed to be determined by the previously selected cues, which is in principle possible. Localization of sinusoidal tones based only on their onsets (Rakerd and Hartmann 1985, 1986) and a related demonstration called the "Franssen effect" (Franssen 1960, Hartmann and Rakerd 1989) show that a derived localization judgment can persist for several seconds after the related localization cues have occurred. In precedence effect conditions (Section 7.3.2) the cue selection naturally derives most localization information from signal onsets, as is explicitly done in the model of Zurek (1987) (see also Martin 1997). However, the cue selection is not limited to getting information from onsets only, and it does not necessarily include all onsets.

Throughout the chapter, the resulting ITD and ILD cues were considered separately instead of deriving a combined localization judgment. The mutual role of ITDs and ILDs is often characterized with time-intensity trading ratios (Blauert 1997) or in the form of the classic duplex theory (Rayleigh 1907): ITD cues dominate localization at low frequencies and ILDs at high frequencies. However, in complex listening situations the relative weights of these cues may change. Wightman and Kistler (1992) have shown that in the presence of conflicting ITD and ILD cues, ITD cues will dominate the localization judgment of broadband noise as long as low frequency energy is present. Furthermore, Braasch (2003) has found that the presence of a distracting sound source even strengthens the weight of ITD cues. Nevertheless, the results of Rakerd and Hartmann (1986) suggest that steady-state ITDs can sometimes be completely neglected, unlike ILD cues. Considering the relative weights of ITD and ILD cues in more detail is beyond the scope of this thesis. However, in future work it will be interesting to assess whether the proposed cue selection reflects the relative importance of ITD and ILD cues, i.e. whether the cue selection, for example, recovers more reliably ITD cues in cases where they are weighted more than ILD cues, and vice versa.

The cue selection mechanism could be seen to perform a function that Litovsky and Shinn-Cunningham (2001) have characterized as "a general process that enables robust localization not only in the presence of echoes, but whenever any competing information from a second source arrives before the direction of a previous source has been computed." For the purposes of this chapter, ITD and IC cues were analyzed using a cross-correlation model, whereas ILDs were computed independently. Similar cue selection could also be implemented in other localization models, such as the excitation-inhibition (EI) model of Breebaart *et al.* (2001) involving joint analysis of ITD and ILD cues within a physiologically motivated structure. In the EI model, full coherence is not represented by maximum activity but by zero activity. Thus, as opposed to specifying a lower bound of IC for cue selection, an upper bound of activity would need to be determined.

As shown in Section 7.3, the cue selection model was able to simulate most psychophysical results reviewed in the introduction by using a selection threshold adapted to each specific listening scenario. Although this chapter is limited to localization based on binaural cues, it should be mentioned that the precedence effect has also been observed in the median sagittal plane where the localization is based on spectral cues instead of interaural differences (Blauert 1971, Litovsky *et al.* 1997). Thus, the cue selection model does not fully describe the operation of the precedence effect. Furthermore, the model cannot as such explain the discounting of ITD and ILD cues occurring simultaneously with a high IC during the steady state sound in two scenarios: A sinusoidal tone presented in a room (Rakerd and Hartmann 1985, 1986, Hartmann and Rakerd 1989) and two independent noise sources with the same envelopes presented from different directions (Braasch 2002). The psychophysical results of Litovsky *et al.* (1997) show that the localization suppression is somewhat weaker in the median plane than in the horizontal plane, which could be interpreted as evidence for another suppression mechanism, possibly operating simultaneously with a binaural mechanism such as the proposed cue selection. Indeed, simulating all the results cited in this paragraph would appear to require some additional form of temporal inhibition.

7.5 Discussion related to binaural cue coding (BCC)

Previously, the cue selection model was only applied to "natural" listening scenarios, i.e. it was assessed in terms of localization of physically existing sources. For using the cue selection model for predicting the performance of BCC, "real-world" stereo or multi-channels signals are considered. Specific recording or mixing techniques are applied for generating such signals with a desired auditory spatial image. The number of auditory events and their locations do not directly depend on the number of playback transducers (headphones or loudspeakers), but is determined by the recording engineer through choice of mixing and microphone techniques. In Section 7.5.1, the cue selection model is applied for predicting localization of auditory events evoked by playing back stereo audio signals through headphones or loudspeakers (independently of BCC). Section 7.5.2 discusses the cue selection model for BCC synthesized stereo signals. A number of simulations compare predicted localization of auditory events of stereo signals and corresponding BCC synthesized stereo signals. In earlier subjective tests (Section 3.5.1), we observed that BCC without ICC synthesis leads to reduced width of the auditory spatial image and non-stable auditory event locations. In Section 7.5.3, the cue selection model is applied to such signals in an attempt to predict the reduced width and instability. Some concluding remarks are given in Section 7.5.4.

7.5.1 Cue selection and stereo audio playback

In Section 2.2.5, it was described that when pairs of ear entrance signals or loudspeaker signals, each evoking a single auditory event, are superimposed the resulting composite signal pair will often result in perception of auditory events at the same locations where the single auditory events are perceived when playing back the corresponding separate signals. This principle was illustrated in Figures 2.5 and 2.6. Recording and mixing techniques for generating stereo or multi-channel audio signals assume the validity of this principle. In the following, the cue selection model is applied for predicting this phenomenon. Signal pairs evoking one auditory event when played back with headphones and loudspeakers are denoted *one-auditory-event* signals. Without loss of generality, only stereo signals are considered.

Figures 7.11 and 7.12 illustrate the different cues which are referred to in the following discussion and simulations for stereo headphone and loudspeaker playback, respectively. The inter-channel time difference (ICTD) and inter-channel level difference (ICLD) which are inherent between the two channels of a one-auditory-event stereo signal are denoted the *mixing cues*. In free-field, these mixing cues and the specific playback setup determine the ITD and ILD cues which result in a specifically localized auditory event. These ITD and ILD cues are considered as the reference for the following simulations and are denoted *reference cues*. The motivation for using the reference cues and not the free-field cues is that in the best case it is expected that localization of auditory events in stereo signals is as good as localization of the single auditory events when corresponding one-auditory-event stereo signals are played back over headphones or loudspeakers in free-field. Thus, it is meaningful to compare the auditory event localization to this case by means of comparing the values of the selected ITD and ILD cues to the reference cues.

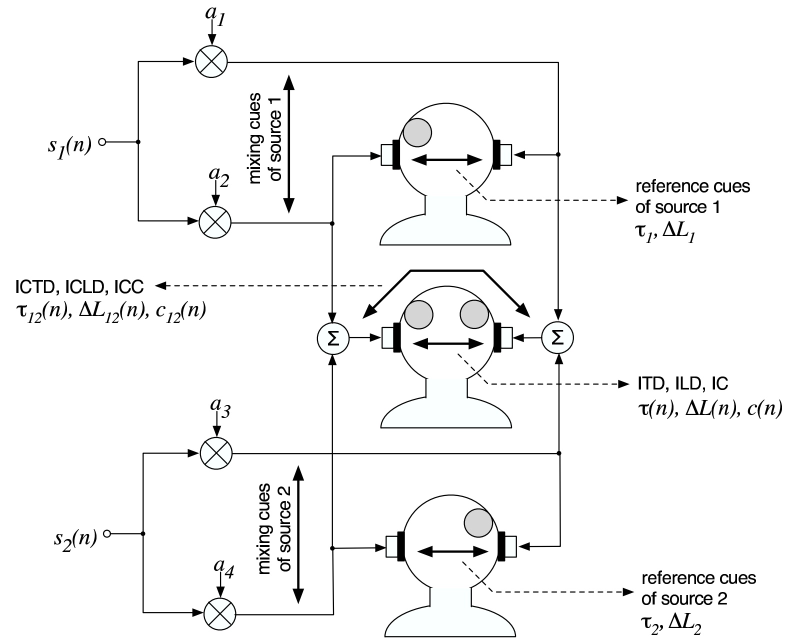


Figure 7.11: Artificially mixed stereo signal and headphone playback. The different cues used for the simulations are indicated: Mixing cues, reference cues, inter-channel cues (ICTD, ICLD, ICC), and binaural cues (ITD, ILD, and IC).

As mentioned in Section 7.2.3, naturally occurring (one-source) free-field cues are assumed to be the reference for the auditory system. Using the reference cues in the following does not imply that it is assumed that here these cues are the reference for the auditory system. The use of the reference cues is only for the purpose of comparing localization to localization of one-auditory-event signals.

Headphone playback

Similarly to the case of the binaural cues considered in the previous simulations, a large inter-channel coherence (ICC) between a stereo channel pair implies that the corresponding signal component was caused by a single source (e.g. amplitude panned source or a source picked up by coincident-pair microphones). Thus, the corresponding ICTD and ICLD cues are equal to the mixing cues. Furthermore, since for headphone playback the ITD, ILD, and IC cues occurring in the auditory system before neural transduction (ideally) are equal to the ICTD, ICLD, and ICC cues, it is expected that the cue selection results in cues similar to the mixing cues (modified by considering the effect of the neural transduction). In this case, the references cues are also similar to the modified mixing cues and the cue selection is expected to be effective.

A simulation with 2 sources was performed with a piano and organ playing together. Amplitude panning with ±8 dB was used for generating a stereo signal where the piano and organ are located towards the left and right sides.

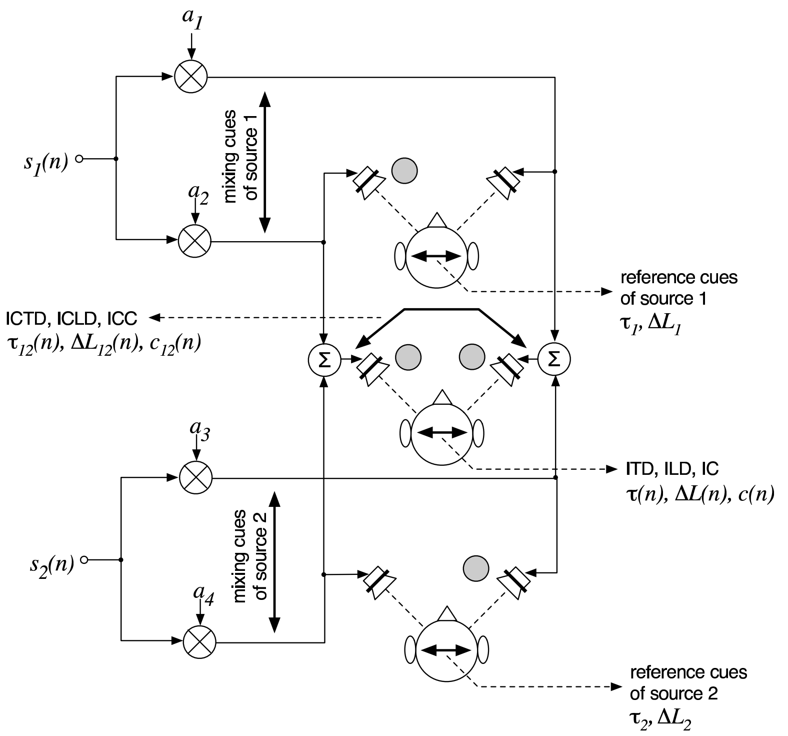


Figure 7.12: Artificially mixed stereo signal and loudspeaker playback. The different cues used for the simulations are indicated: Mixing cues, reference cues, inter-channel cues (ICTD, ICLD, ICC), and binaural cues (ITD, ILD, and IC).

Figure 7.11 illustrates the experiment and the relation between the different cues as described previously. The left column of panels in Figure 7.13 show IC, ITD, and ILD as a function of time for the stereo audio signal played back with headphones for the critical band at 500 Hz. A cue selection threshold of $c_0 = 0.995$ was used and the selected cues are indicated with bold solid lines. Note that since the stereo signal was generated with amplitude panning (with no time delays), most ITD cues are close to zero. Since the selected ITD and ILD cues are always close to the reference cues the cue selection model correctly predicts auditory events localized similarly as if the corresponding one-auditory-event stereo signals were played back separately.

The results of this simulation were also visualized with PDFs. The left top panel in Figure 7.14 shows PDFs of the ITD and ILD cues which occur during the previously described simulation. The left bottom panel in Figure 7.14 shows the PDFs of the selected ITD and ILD cues. The peaks in the PDFs are located close to the indicated reference cues as desired.

The middle and right columns of panels in Figures 7.14 and 7.13 show similar simulations carried out with BCC synthesized signals, as discussed later in Sections 7.5.2 and 7.5.3.

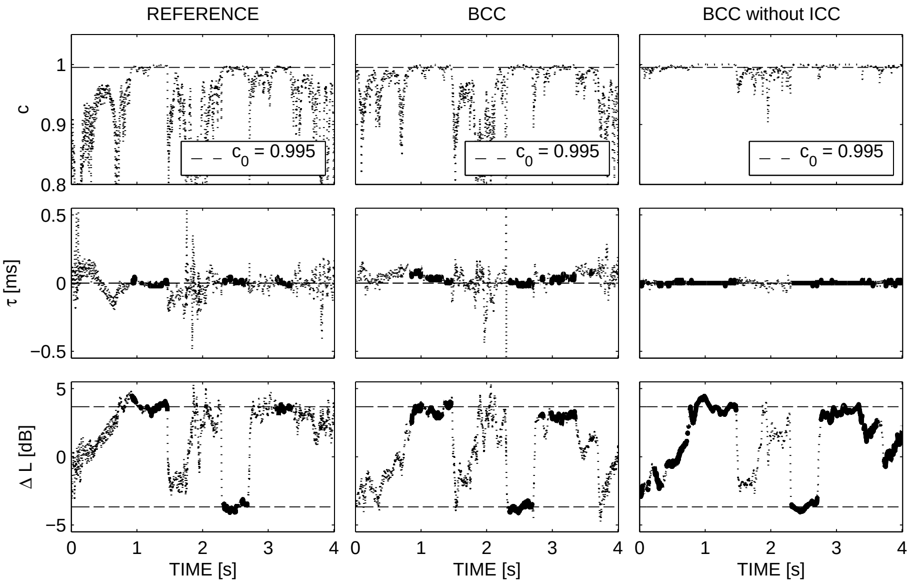


Figure 7.13: Headphone playback of a stereo signal: IC, ILD, and ITD as a function of time for two amplitude panned instruments playing together for the 500 Hz critical band. Left column: Reference stereo signal, middle column: BCC signal, and right column: BCC without ICC synthesis signal. The cue selection thresholds (top row) and the reference cues (middle and bottom rows) are indicated with dashed lines. Selected cues are marked with bold solid lines.

Loudspeaker playback

Stereo loudspeaker playback relies on the phenomenon of summing localization (Blauert 1997), i.e. by emitting specific coherent signals from a loudspeaker pair, an auditory event can be made appear at a desired position between the loudspeaker pair. Summing localization results from the fact that ICTD and ICLD cues translate into ITD and ILD cues which crudely approximate the cues caused by a real source at the location of the summing localization auditory event. In this case, the reference cues are computed from the ear entrance signals occurring when the one-auditory-event stereo signals are played back separately from the loudspeakers in free-field.

Figure 7.12 illustrates the relation between the different cues when playing back the same stereo signal that was used for the previous simulations through a pair of loudspeakers. The following considerations imply that the cue selection model is also applicable to stereo loudspeaker playback:

- ICC coincides with ICTD and ICLD similar to the mixing cues.
- These mixing cues evoke ITD and ILD cues similar to the reference cues.
- A time instant with a large ICC between the left and right stereo signal pair is likely to translate into a time instant with large IC between the

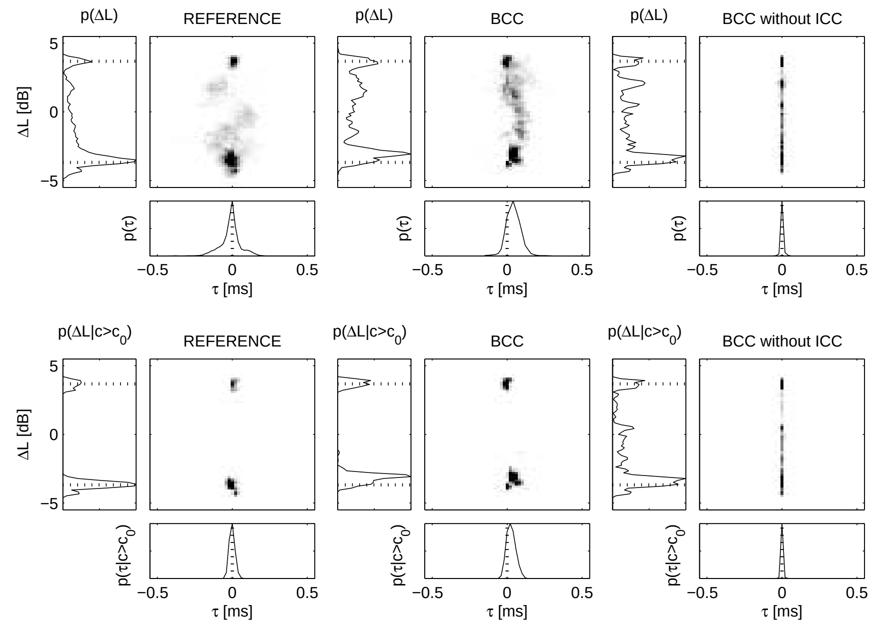


Figure 7.14: Headphone playback of a stereo signal: PDFs of ITD and ILD for reference stereo signal, BCC signal, and BCC without ICC signal (top panels) and the corresponding PDFs when cue selection is applied (bottom panels). The values of the reference cues are indicated with dotted lines. Data is shown for the 500 Hz critical band.

ear entrance signals.

From these properties it follows that large IC cues coincide with cues similar to the reference cues, which would suggest that the cue selection model is able to predict localization of auditory events for stereo loudspeaker playback.

A simulation was carried out, using the same stereo signal as previously, but played back over loudspeakers in free-field. The loudspeaker playback in free-field was simulated by filtering the left and right channel of the stereo signal with HRTFs corresponding to sources at $\pm 30^\circ$ azimuth and adding the resulting two pairs of left and right ear entrance signals. The critical band at 500 Hz was considered.

The left column of panels in Figure 7.15 shows the resulting IC, ITD, and ILD cues. The reference cues are indicated with dashed lines. A cue selection threshold of $c_0 = 0.995$ was used and the selected cues are indicated with bold solid lines. Since the selected ITD and ILD cues are always close to the reference cues the cue selection model correctly predicts auditory events localized similarly as if the corresponding one-auditory-event stereo signals were played back separately.

The data was also visualized with PDFs. The left top panel in Figure 7.16 shows PDFs of ITD and ILD for the cues occurring in the described simulation. The PDFs of the selected cues are shown in the left bottom panel. The peaks close to the reference cues indicate correctly predicted localization of the two

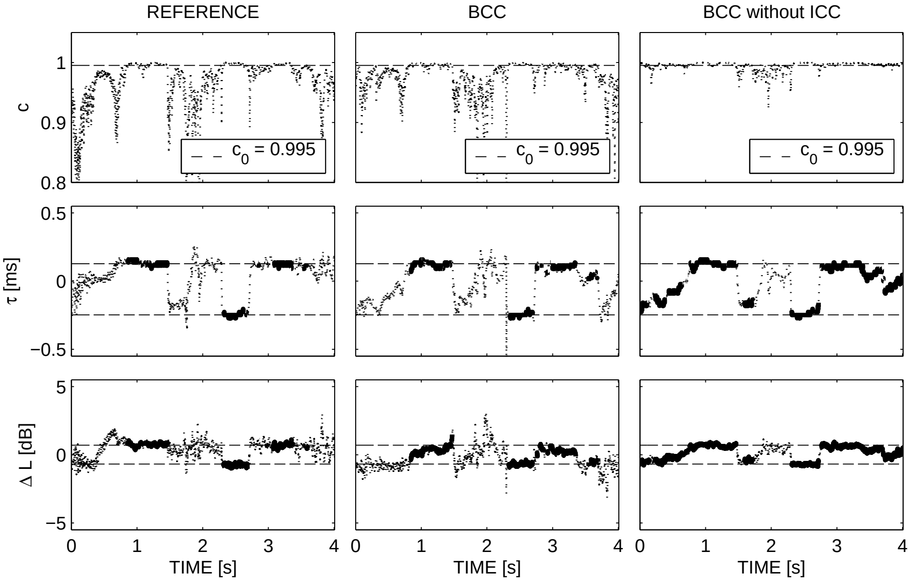


Figure 7.15: Loudspeaker playback of a stereo signal: IC, ILD, and ITD as a function of time for two amplitude panned instruments playing together for the 500 Hz critical band. Left column: Reference stereo signal, middle column: BCC signal, and right column: BCC without ICC synthesis signal. The cue selection thresholds (top row) and the reference cues (middle and bottom rows) are indicated with dashed lines. Selected cues are marked with bold solid lines.

auditory events.

Note that the middle and right columns of panels in Figures 7.16 and 7.15 show similar simulations carried out with BCC synthesized signals, as discussed in Sections 7.5.2 and 7.5.3.

7.5.2 BCC and the cue selection

Given the sum signal of all channels of an audio signal, it would be a daunting task to attempt to re-create the specific complex mixes of each of the audio channels. Binaural cue coding (BCC) does not at all attempt to recover the original waveforms of the audio channels. However, it is often able to generate audio signals which are perceived as being very similar to the original audio signals. The stereo and multi-channel audio signals are generated by blindly synthesizing ICLD, ICTD, and ICC cues between channel pairs as a function of time and frequency such that they approximate the corresponding cues of the original audio signal.

Earlier in this section, ICTD, ICLD, and ICC cues in a stereo signal were related to ITD, ILD, and IC cues and the cue selection. The discussion and simulations implied that the cue selection model may be able predict localization of auditory events for stereo headphone and loudspeaker playback. Thus, if BCC

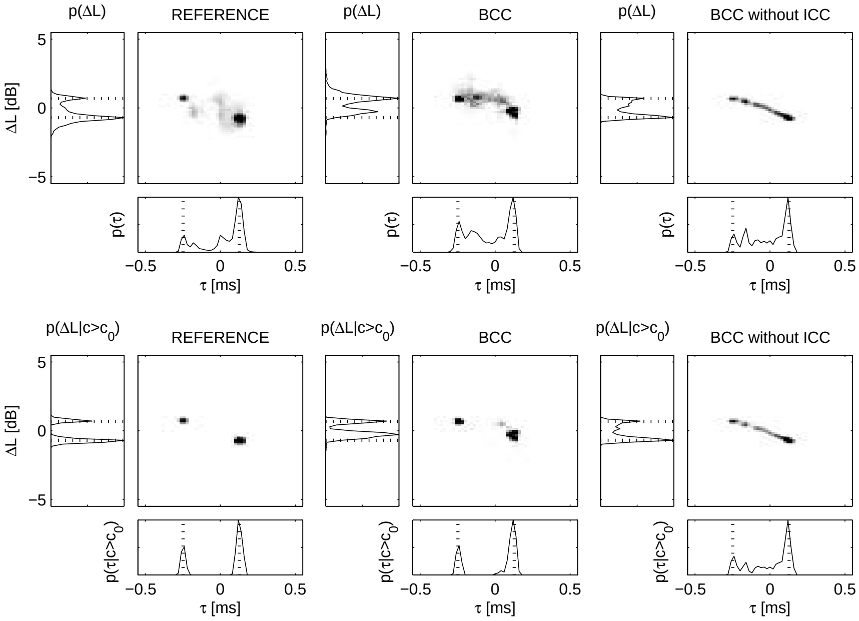


Figure 7.16: Loudspeaker playback of a stereo signal: PDFs of ITD and ILD for reference stereo signal, BCC signal, and BCC without ICC synthesis signal (top panels) and the corresponding PDFs when cue selection is applied (bottom panels). The values of the reference cues are indicated with dotted lines. Data is shown for the 500 Hz critical band.

would synthesize ICTD, ICLD, and ICC perfectly it would be expected that the cue selection model would predict similar localization of auditory events for BCC synthesized signals as for corresponding reference signals.

However, for practical reasons BCC synthesizes the cues only every 4−16 ms (Section 3.5.1). Therefore, whenever ICTD, ICLD, and ICC in the reference signal fluctuate within the time interval in which the cues are estimated, the BCC synthesized signal will have different (time-smoothed) ICTD, ICLD, and ICC cues. ICC is expected to be smaller than the largest ICC of the reference signal in that time interval. Thus, whenever the cues fluctuate within the estimation interval it is likely that the cues are not selected in that interval (due to the lower IC resulting from the lower ICC). The estimated “averaged” ICTD and ICLD cues are arbitrary and do not correspond to the mixing cues (and thus ITD and ILD do not correspond to the reference cues). The cue selection is expected to ignore the evoked “wrong” ITD and ILD cues and thus will imply correct localization of auditory events for BCC synthesized signals in this case (assuming that cues without fluctuation in cue estimation intervals occur for every auditory event).

For precedence effect conditions, the cue selection has only a similar effect for the original and BCC synthesized signal when lead and lag do not fall within the period in which only one set of cues is estimated. For real-world precedence effect conditions, caused by the direct sound and the first reflection, updating the cues every 4 ms should in most cases be enough, since this implies the likely case that the direct path of sound is at least about 1.2 m shorter than the path of any reflection. Furthermore, if it is assumed that the microphones were placed not very close to the sources and a wall or obstacle, a short lead/lag delay implies that direct sound and reflection arrive from approximately the same direction. In such a case, lead and lag falling within the same cue estimation time interval does not hurt since the resulting ITD and ILD cues of both are approximately the same. Thus, for many cases, updating the cues only every $4-16$ ms may suffice and the generally good performance of BCC for real-world audio signals may be explained by this.

Time instants when the precedence effect becomes active may also occur when there are independent concurrently active sound sources from different directions. By chance, two transients emanating from two different sources may follow each other within a short time, leading to a condition where the precedence effect is active. Assuming independent sources with transients being emitted randomly, such conditions occur at random time instants with random lead/lag delay. The random lead/lag delays result in that at some time instants lead and lag lie within one cue estimation interval and at other time instants lead and lag are in different cue estimation intervals. The ITD and ILD cues evoked by the BCC synthesized signal only coincide with high IC when the lead and lag were in different estimation intervals. The cue selection is expected to ignore the “wrong” cues evoked when the time resolution of BCC was not high enough, i.e. when lead and lag were in the same estimation interval. Thus, it is expected that the BCC synthesized signal leads to a signal where the cue selection yields cues similar to the reference cues (assuming that for every auditory event some transients are in cue estimation intervals without transients from other sources).

The previous simulations for headphone and loudspeaker stereo playback were repeated with a BCC synthesized version of the same stereo audio signal. BCC was used with synthesizing cues every 13 ms. The subbands of BCC were chosen to be 2 ERB wide. The results are shown in the middle columns of panels in Figures 7.13, 7.14, 7.15, and 7.16. These results suggest that the BCC synthesized stereo signal results in auditory events localized very similarly compared to the original stereo signal (left columns of panels in the figures), as predicted by the cue selection model.

7.5.3 ICC and auditory spatial image width and stability

Initially, the cue selection model was inspired by the intuition that ICC synthesis in BCC is not only necessary for mimicking attributes related to spatial impression. Informal experiments implied that ICC synthesis was also important for synthesizing signals with a stable and as-wide-as-original auditory spatial image. The listening test results of BCC without ICC synthesis, described in Section 3.5.1, imply that BCC without ICC synthesis suffers from auditory spatial image width reduction and instability.

The previous simulations with two amplitude panned sources result in an auditory spatial image with two well-focused auditory events “without” an associated spatial impression (no reverberation, no room impression). Thus, in this case the ICC cues have no importance with respect to spatial impression. By comparing such a stereo signal with the same stereo signal but without

ICC cues (setting ICC = 1, i.e. BCC without ICC synthesis), one may get an indication about the role ICC plays independently of being related to spatial impression.

The previous simulations were repeated for the same stereo audio signal synthesized with BCC without ICC cues. The results are shown in the right columns of panels in Figures 7.13, 7.14, 7.15, and 7.16. For headphone playback and loudspeaker playback, the cue selection model is less effective in this case and often selects cues which are more to the center than the reference cues. This may not only explain the observed reduced width of the auditory spatial image, but also the instability (there is significant variation of the selected ITD and ILD values in time).

7.5.4 Discussion

All simulations presented in this section were also carried out for critical bands at lower and higher frequencies than the previously used 500 Hz. The results were similar for all tested center frequencies. For the sake of brevity, only the results for the 500 Hz critical band were shown.

Our implementation of BCC has been using cue estimation and synthesis at regular time intervals. To prevent cases when two transients fall within the estimation of a single set of cues, one could adapt the update rate of the cue estimation and synthesis when lead and lag would fall within one estimation interval. Practically, this can be implemented using a filterbank with "block-switching", i.e. switching to a higher time-resolution when it is advantageous to do so. The parametric stereo coder by Schuijers *et al.* (2003) incorporates such block switching. Or one could use a filterbank with a higher time resolution and only adapt the rate of cue analysis/synthesis, as has been done by Schuijers *et al.* (2004).

7.6 Conclusions

A cue selection mechanism was presented for modeling source localization in complex listening scenarios. The cue selection can simulate both localization of several concurrently active independent sources and suppression of the localization of reflected sound by considering ITD and ILD cues only when IC at the corresponding critical band is larger than a certain threshold. It was shown that at time instants when this occurs, ITD and ILD are likely to represent the direction of one of the sources. Thus, by looking at the different ITD and ILD values during the selected time instants one can obtain information about the direction of each source.

The proposed cue selection mechanism was implemented in the framework of a binaural model considering the known periphery of the auditory system. The remaining parts of the model were analytically motivated for the sake of focus on the cue selection method without having to consider the specific properties and limitations of existing physiologically motivated models. Nevertheless, it was pointed out in the discussion that in principle the proposed cue selection method is physiologically feasible.

The binaural model with the proposed cue selection was verified with the results of a number of psychophysical studies from the literature. The simulation results suggest relatively reliable localization of concurrent speech sources both in anechoic and reverberant environments. The effect of target-to-distracter ratio corresponds qualitatively to published results of localization of a click-train in the presence of a noise distracter. Localization dominance is correctly reproduced for click pairs and for the onsets of sinusoidal tones. It was also hypothesized that the buildup of precedence may be related to the time the auditory system needs to find a cue selection threshold which is effective for the specific listening situation. As a final test, the model was applied for source localization in a reverberant hall with one and two speech sources. The results suggest that also in this most complex case the model is able to obtain cues corresponding to the directions of the sources.

The cue selection was also discussed related to BCC. It may explain why localization of auditory events in BCC synthesized signals is often not significantly impaired. Also, the selected cues imply auditory spatial image width reduction and instability when BCC is used without considering coherence cues. This is in line with the listening test results presented in Section 3.5.1.

Chapter 8

Conclusion

8.1 Thesis summary

A coding scheme for stereo and multi-channel audio signals was proposed, denoted binaural cue coding (BCC). The properties of the auditory spatial image which is perceived when playing back stereo or multi-channel audio signals is largely determined by certain difference measures between audio channel pairs. These include level difference, time difference, and a statistical similarity measure (coherence) considered as a function of time and frequency. Considering this, BCC represents stereo and multi-channel audio signals as a single audio channel containing the signal components of all input audio channels. Additionally, BCC estimates level difference, time difference, and coherence between the original audio channels and transmits these as side information to the decoder along with the single audio channel. The decoder generates the decoded audio signal by processing the transmitted single channel such that the mentioned difference measures between the output audio channels approximate those of the original audio signal.

The representation of stereo and multi-channel audio signals as a single channel plus side information enables coding of audio signals at a bitrate nearly as low as the bitrate necessary for coding of a mono audio signal. The reason for this is that the side information contains about two orders of magnitude less information than the audio channel waveforms themselves. Compared to conventional state-of-the-art audio coders this represents a significant reduction in the bitrate necessary for coding of such signals.

The results of a number of subjective tests indicate that BCC achieves good audio quality. It has been shown that BCC combined with a conventional mono audio coder for coding of the single transmitted audio channel outperforms existing audio coders.

Since BCC is based on transmission of a single valid audio channel, existing mono systems can be extended for stereo and multi-channel audio playback by additionally transmitting the side information. A variation of BCC which transmits more than one audio channel enables not only backwards compatible extension of mono systems but also extension of stereo systems to multi-channel surround. Also, compatible coding between different surround formats can be implemented with this scheme.

Earlier versions of BCC did not consider coherence cues between the audio channels. Subjective tests and informal listening indicated that the auditory spatial image evoked when playing back BCC synthesized signals without considering coherence cues is impaired in three ways: Firstly, the spatial impression is modified, i.e. the impression of "space" is largely lost. Secondly, auditory events at the sides appear more towards the center such that the overall width of the auditory spatial image in the left/right dimension is reduced. Thirdly, the locations of the auditory events are often non-stable. The first impairment can be directly related to spatial impression and its relation to interaural coherence. The second and third impairments motivated further thinking about the role interaural coherence plays related to source localization in the presence of concurrent sound, i.e. in the presence of other sources and reflections. This lead to the proposition of a model for source localization in complex listening scenarios. The model was verified with the results of a number of previously published psychophysical studies. Additionally, the model was applied for predicting auditory event localization of auditory events during stereo audio playback, comparing stereo signals and corresponding BCC synthesized stereo signals. The results indicate that, according to this model, BCC synthesized signals result in similarly localized auditory events as for the corresponding non-coded stereo signals.

8.2 Impact

Arguably, the most relevant contribution of this thesis may be that it was shown that high quality audio coding by means of representing a stereo or multichannel audio as a single channel plus perceptually motivated side information is possible. Previous related techniques, e.g. intensity stereo coding, resulted in very limited audio quality.

Our early BCC papers have inspired other activities by other researchers in this field. There is a renewed interest in spatial audio coding, which has also led to a separate session for this topic at Audio Engineering Society Conventions. A BCC-like stereo extension to a parametric coder has recently been standardized in ISO/IEC MPEG.

The author collaborated with researchers from Fraunhofer IIS, Erlangen, Germany on 5-to-2 BCC (as part of a joint development agreement between Agere Systems and Fraunhofer IIS), which resulted in a backwards compatible 5.1 surround extension to the popular MP3 audio coder denoted "MP3 Surround". 5-to-2 BCC is also being applied to various other audio coders in an attempt to provide backwards compatible extension to 5.1 surround for digital radio broadcasting. Fraunhofer IIS also initiated an ISO/IEC MPEG call for proposals for BCC-like techniques for "spatial audio coding".

8.3 Future research

The author carried out a number of informal experiments comparing binaural recordings to BCC synthesized binaural recordings. It seems that not only the original signals result in externalization, but to a large degree also the BCC synthesized signals. This needs further investigation and may lead to a parametrization of head related transfer functions (HRTFs) or binaural room impulse responses (BRIRs) with cues similar as are used in BCC.

When BCC is used for stereo signals and headphone playback, the interaural cues inherent is such signals can be manipulated by means of manipulation of the cues used by BCC. Given a specific “real-world” stereo signal which is coded transparently by BCC, it may be argued that in this case BCC restores all the interaural cues which are relevant for the perception of the auditory spatial image. The cues then could be systematically manipulated and listening experiments could assess the impact of different manipulations. This could be a way of conducting listening experiments with complex multi-source stimuli.

The proposed model for source localization features one parameter c_0 (the cue selection threshold) which is assumed to adapt to a specific listening situation. For simplicity, this threshold was determined for most simulations manually. Automatic adaptation of c_0 is expected to be very complex. A “top-down” process, i.e. a process operating at higher stages of the auditory system, is expected to adapt this parameter such that plausible cues are obtained by the model. More work needs to be done to add automatic adaptation of c_0 to the model.

The application of the proposed model for predicting auditory event localization when stereo and multi-channel audio signals are played back needs further and more thorough study. The discussion in this thesis was limited to stereo audio playback and a single signal with two amplitude panned sources. The cue selection model possibly may also be applied for predicting attributes of the auditory spatial image other than only auditory event localization.

Bibliography

Akeroyd, M. A. (2001), ‘A binaural cross-correlogram toolbox for MATLAB’. http://www.biols.susx.ac.uk/home/Michael_Akeroyd/download2.html.

Algazi, V. R., Duda, R. O., Thompson, D. M. and Avendano, C. (2001), The CIPIC HRTF Database, *in* ‘IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.’, New Paltz, NY, USA, pp. 99–102.

Ando, Y. and Kurihara, Y. (1986), ‘Nonlinear response in evaluating the subjective diffuseness of sound fields’, *J. Acoust. Soc. Am.* **80**(3), 833–836.

Atal, B. S. and Hanauer, S. L. (1971), ‘Speech analysis and synthesis by linear prediction of the speech wave’, *J. Acoust. Soc. Am.* **50**, 637–655.

Atal, B. S. and Schroeder, M. R. (1966), ‘Nachahmung der Raumakustik durch Elektronenrechner [Simulation of room acoustics using electronic computers’, *Gravesaner Bl”atter* **27/28**, 125–137. (Cited in Blauert 1997).

Avendano, C. and Jot, J.-M. (2002), Ambience extraction and synthesis from stereo signals for multi-channel audio up-mix, *in* ‘Proc. ICASSP, Orlando, Florida’, Vol. 2, pp. 1957–1960.

Barron, M. and Marshall, A. H. (1981), ‘Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure’, *J. of Sound and Vibration* **77**(2), 211–232.

Bauer, B. B. (1961*a*), ‘Phasor analysis of some stereophonic phenomena’, *J. Acoust. Soc. Am.* **33**, 1536–1539.

Bauer, B. B. (1961*b*), ‘Stereophonic earphones and binaural loudspeakers’, *J. Audio Eng. Soc.* **9**, 148–151.

Baumgarte, A., Faller, C. and Kroon, P. (2004), Audio coder enhancement using scalable binaural cue coding with equalized mixing, *in* ‘Preprint 116th Conv. Aud. Eng. Soc.’.

Baumgarte, F. and Faller, C. (2002*a*), Design and evaluation of Binaural Cue Coding schemes, *in* ‘Preprint 113th Conv. Aud. Eng. Soc.’.

Baumgarte, F. and Faller, C. (2002*b*), Estimation of auditory spatial cues for Binaural Cue Coding (BCC), *in* ‘Proc. ICASSP 2002’, Vol. 2, pp. 1801–1804.

Baumgarte, F. and Faller, C. (2002*c*), Why Binaural Cue Coding is better than Intensity Stereo Coding, *in* 'Preprint 112th Conv. Aud. Eng. Soc.'.

Baumgarte, F. and Faller, C. (2003), 'Binaural Cue Coding - Part I: Psychoacoustic fundamentals and design principles', *IEEE Trans. on Speech and Audio Proc.*

Benesty, J., Gänsler, T., Morgan, D. R., Sondhi, M. M. and Gay, S. L. (2001), *Advances in Network and Acoustic Echo Cancellation*, Springer.

Bennett, J. C., Barker, K. and Edeko, F. O. (1985), 'A new approach to the assessment of stereophonic sound system performance', *J. Audio Eng. Soc.* **33**(5), 314–321.

Beranek, L. L. (1962), *Music, Acoustics and Architecture*, NewYork: Wiley.

Bernfeld, B. (1973), Attempts for better understanding of the directional stereophonic listening mechanism, *in* 'Preprint 44th Conv. Aud. Eng. Soc.'.

Bernstein, L. R. and Trahiotis, C. (1996), 'The normalized correlation: Accounting for binaural detection across center frequency', *J. Acoust. Soc. Am.* **100**(6), 3774–3784.

Bernstein, L. R., van de Par, S. and Trahiotis, C. (1999), 'The normalized interaural correlation: Accounting for NoSπ thresholds obtained with Gaussian and "low-noise" masking noise', *J. Acoust. Soc. Am.* **106**(2), 870–876.

Blauert, J. (1971), 'Localization and the law of the first wavefront in the median plane', *Acustica* **50**(2 Pt. 2), 466–470.

Blauert, J. (1997), *Spatial Hearing: The Psychophysics of Human Sound Localization*, revised edn, The MIT Press, Cambridge, Massachusetts, USA.

Blauert, J. and Cobben, W. (1978), 'Some consideration of binaural cross correlation analysis', *Acustica* **39**(2), 96–104.

Blumlein, A. (1931), 'Improvements in and relating to sound transmission, sound recording and sound reproduction systems', *British Patent Specification 394325.* Reprinted in *Stereophonic Techniques*, Aud. Eng. Soc., New York, 1986.

Boehnke, S. E., Hall, S. E. and Marquadt, T. (2002), 'Detection of static and dynamic changes in interaural correlation', *J. Acoust. Soc. Am.* **112**(4), 1617–1626.

Boerger, G., Laws, P. and Blauert, J. (1977), 'Stereophone Kopfhörerwiedergabe mit Steuerung bestimmter Übertragungsfaktoren durch Kopfdrehbewegung [Sereophonic headphone reproduction with variation of various transfer factors by means of rotational head movements]', *Acustica* **39**, 22–26.

Boland, S. and Deriche, M. (1995), High quality audio coding using multipulse LPC and wavelet decomposition, *in* 'Proc. Int. Conf. Acoust., Speech, and Signal Proc.', IEEE, Detroit, USA, pp. 3067–3069.

Bolia, R. S., Ericson, M. A., McKinley, W. T. and Simpson, B. D. (1999), 'A cocktail party effect in the median plane?', *J. Acoust. Soc. Am.* **105**, 1390–1391.

Bolia, R. S., Nelson, W. T., Ericson, M. A. and Simpson, B. D. (2000), 'A speech corpus for multitalker communications research', *J. Acoust. Soc. Am.* **107**(2), 1065–1066.

Bosi, M., Brandenburg, K., Quackenbush, S., Fielder, L., Akagiri, K., Fuchs, H., Dietz, M., Herre, J., Davidson, G. and Oikawa, Y. (1997), 'ISO/IEC MPEG-2 advanced audio coding', *J. Audio Eng. Soc.* **45**(10), 789–814.

Braasch, J. (2002), 'Localization in the presence of a distracter and reverberation in the frontal horizontal plane. II. Model algorithms', *Acta Acustica United with Acustica* **88**(6), 956–969.

Braasch, J. (2003), 'Localization in the presence of a distracter and reverberation in the frontal horizontal plane. III. The role of interaural level differences', *Acta Acustica United with Acustica* **89**(4), 674–692.

Braasch, J. and Blauert, J. (2003), 'The precedence effect for noise bursts of different bandwidths. II. Comparison of model algorithms', *Acoust. Sci. & Tech.* **24**(5), 293–303.

Braasch, J. and Hartung, K. (2002), 'Localization in the presence of a distracter and reverberation in the frontal horizontal plane. I. Psychoacoustical data', *Acta Acustica United with Acustica* **88**(6), 942–955.

Braasch, J., Blauert, J. and Djelani, T. (2003), 'The precedence effect for noise bursts of different bandwidths. I. Psychoacoustical data', *Acoust. Sci. & Tech.* **24**(5), 233–241.

Bradley, J. S. (1994), 'Comparison of concert hall measurements of spatial impression', *J. Acoust. Soc. Am.* **96**(6), 3525–3535.

Bradley, J. S. and Soulodre, B. A. (1995), 'Objective measures of listener envelopment', *J. Acoust. Soc. Am.* **98**, 2590–2597.

Brandenburg, K. and Stoll, G. (1994), 'ISO-MPEG-1 audio: a generic standard for coding of high-quality digital audio', *J. Audio Eng. Soc.* pp. 780–792.

Brandenburg, K., Langenbucher, G. G., Schramm, H. and Seitzer, D. (1982), A digital signal processor for real time adaptive transform coding of audio signal up to 20 khz bandwidth, *in* 'Proc. ICCC', pp. 474–477.

Breebaart, J., van de Par, S. and Kohlrausch, A. (2001), 'Binaural processing model based on contralateral inhibition. I. Model structure', *J. Acoust. Soc. Am.* **110**(2), 1074–1088.

Bronkhorst, A. and Houtgast, T. (1999), 'Auditory distance perception in rooms', *Nature* **397**, 517–520.

Bruekers, A. A. M. L., Oomen, W. J. and van der Vleuten, R. J. (1996), Lossless coding for DVD audio, *in* 'Preprint 100th Conv. Aud. Eng. Soc.'.

Brungart, D. S. and Rabinowitz, W. M. (1999), 'Auditory localization of nearby sources I: Head-related transfer functions', *J. Acoust. Soc. Am.* **106**, 1465–1479.

Cellier, C., Chênes, P. and Rossi, M. (1993), Lossless audio data compression for real time applications, *in* 'Preprint 95th Conv. Aud. Eng. Soc.'.

Chernyak, R. I. and Dubrovsky, N. A. (1968), Pattern of the noise images and the binaural summation of loudness for the different interaural correlation of noise, *in* 'Proc. 6th Int. Congr. on Acoustics Tokyo', Vol. 1, pp. A–3–12. (See Blauert 1997, Fig. 3.24).

Clifton, R. K. (1987), 'Breakdown of echo suppression in the precedence effect', *J. Acoust. Soc. Am.* **82**(5), 1834–1835.

Coleman, P. D. (1962), 'Failure to localize the source distance of an unfamiliar sound', *J. Acoust. Soc. Am.* **34**, 345–346.

Culling, J. F., Colburn, H. S. and Spurchise, M. (2001), 'Interaural correlation sensitivity', *J. Acoust. Soc. Am.* **110**(2), 1020–1029.

Damaske, P. (1967/68), 'Subjektive Untersuchungen von Schallfeldern [Subjective investigations of sound fields]', *Acustica* **19**, 198–213. (See Blauert 1997, Fig. 3.43).

den Brinker, B., Schuijers, E. and Oomen, W. (2002), Parametric coding for high-quality audio, *in* 'Preprint 112th Conv. Aud. Eng. Soc.'.

Dietz, M., Liljeryd, L., Kjörling, K. and Kunz, O. (2002), Spectral band replication - a novel approach in audio coding, *in* 'Preprint 112th Conv. Aud. Eng. Soc.'.

Djelani, T. and Blauert, J. (2001), 'Investigations into the build-up and breakdown of the precedence effect', *Acta Acustica - ACUSTICA* **87**(2), 253–261.

Djelani, T. and Blauert, J. (2002), Modelling the direction-specific build-up of the precedence effect, *in* 'Forum Acusticum', Sevilla, Spain.

Dressler, R. (2000), Dolby Surround Prologic II Decoder - Principles of operation, Technical report, Dolby Laboratories. www.dolby.com/tech/.

Drullman, R. and Bronkhorst, A. W. (2000), 'Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation', *J. Acoust. Soc. Am.* **107**(4), 2224–2235.

Edler, B., Purnhagen, H. and Ferekidis, C. (1996), An analysis/synthesis audio codec (asac), *in* 'Preprint 100th Conv. Aud. Eng. Soc.'.

Enkl, F. (1958), 'Die übertragung räumlicher Schallfeldstrukturen über einen kanal mit hilfe unterschwelliger pilotfrequenzen [The transmission of spatial sound field structures over one channel aided by pilot tones below the threshold]', *Elektron. Rdsch.* **12**, 347–349.

Faller, C. (2003), ‘Parametric multi-channel audio coding: Synthesis of cross-correlation cues’, *IEEE Trans. on Speech and Audio Proc.* (submitted Dec. 2003).

Faller, C. and Baumgarte, F. (2001), Efficient representation of spatial audio using perceptual parametrization, *in* ‘Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.’, pp. 199–202.

Faller, C. and Baumgarte, F. (2002*a*), Binaural Cue Coding: A novel and efficient representation of spatial audio, *in* ‘Proc. ICASSP’, Vol. 2, pp. 1841–1844.

Faller, C. and Baumgarte, F. (2002*b*), Binaural Cue Coding applied to audio compression with flexible rendering, *in* ‘Preprint 113th Conv. Aud. Eng. Soc.’.

Faller, C. and Baumgarte, F. (2002*c*), Binaural Cue Coding applied to stereo and multi-channel audio compression, *in* ‘Preprint 112th Conv. Aud. Eng. Soc.’.

Faller, C. and Baumgarte, F. (2003), ‘Binaural Cue Coding - Part II: Schemes and applications’, *IEEE Trans. on Speech and Audio Proc.*

Faller, C. and Merimaa, J. (2004), ‘Source localization in complex listening situations: Selection of binaural cues based on interaural coherence’, *J. Acoust. Soc. Am.* (submitted Jan. 2004).

Fielder, L. D., Bosi, M., Davidson, G., Davis, M., Todd, C. and Vernon, S. (1996), AC-2 and AC-3: Low-complexity transform-based audio coding, *in* N. Gilchrist and C. Grewin, eds, ‘Collected Papers on Digital Audio Bit-Rate Reduction’, Audio Engineering Society Inc., pp. 54–72.

Franssen, N. V. (1960), Some considerations on the mechanism of directional hearing, PhD thesis, Technische Hogeschool, Delft, The Netherlands.

Freyman, R. L., Clifton, R. K. and Litovsky, R. Y. (1991), ‘Dynamic processes in the precedence effect’, *J. Acoust. Soc. Am.* **90**(2), 874–884.

Freyman, R. L., Helfer, K. S., McCall, D. D. and Clifton, R. K. (1999), ‘The role of perceived spatial separation in the unmasking of speech’, *J. Acoust. Soc. Am.* **106**(6), 3578–3588.

Gabriel, K. J. and Colburn, H. S. (1981), ‘Interaural correlation discrimination: I. Bandwidth and level dependence’, *J. Acoust. Soc. Am.* **69**(5), 1394–1401.

Gaik, W. (1993), ‘Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling’, *J. Acoust. Soc. Am.* **94**(1), 98–110.

Gardner, W. G. (1995), ‘Efficient convolution without input-output delay’, *J. Audio Eng. Soc.* **43**(3), 127–136.

Gardner, W. G. (1998), Reverberation algorithms, *in* M. Kahrs and K. Brandenburg, eds, 'Applications of Digital Signal Processing to Audio and Acoustics', Kluwer Academic Publishing, Norwell, MA, USA, chapter 2.

Gerzon, M. (1990), 'Three channels: the future of stereo?', *Studio Sound* pp. 112–125.

Gerzon, M. A., Graven, P. G., Stuart, J. R., Law, M. J. and Wilson, R. J. (1999), The MLP lossless compression system, *in* 'Proc. AES 17th Int. Conf.: High-Quality Audio Coding', AES, Florence, Italy, pp. 61–75.

Giguère, C. and Abel, S. M. (1993), 'Sound localization: Effects of reverberation time, speaker array, stimulus frequence, and stimulus rise/decay', *J. Acoust. Soc. Am.* **94**(2), 769–776.

Glasberg, B. R. and Moore, B. C. J. (1990), 'Derivation of auditory filter shapes from notched-noise data', *Hear. Res.* **47**, 103–138.

Good, M. D. and Gilkey, R. H. (1996), 'Sound localization in noise: The effect of signal to noise ratio', *J. Acoust. Soc. Am.* **99**(2), 1108–1117.

Good, M. D., Gilkey, R. H. and Ball, J. M. (1997), The relation between detection in noise and localization in noise in the free field, *in* R. H. Gilkey and T. R. Anderson, eds, 'Binaural and Spatial Hearing in Real and Virtual Environments', Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 349–376.

Grantham, D. W. (1982), 'Detectability of time-varying interaural correlation in narrow-band noise stimuli', *J. Acoust. Soc. Am.* **72**(4), 1178–1184.

Grill, B. (1999), The MPEG-4 general audio coder, *in* 'Proc. AES 17th Int. Conf.: High-Quality Audio Coding', AES, Florence, Italy, pp. 147–156.

Härmä, A. and Laine, U. K. (1999), Warped low-delay celp for wide-band audio coding, *in* 'Proc. of the AES 17th Int. Conference: High-Quality Audio Coding', pp. 207–215.

Härmä, A., Laine, U. K. and Karjalainen, M. (1997), An experimental audio codec based on warped linear prediction of complex valued signals, *in* 'Proc. ICASSP 1997', Vol. 1, pp. 323–327.

Hartmann, W. M. (1983), 'Localization of sound in rooms', *J. Acoust. Soc. Am.* **74**(5), 1380–1391.

Hartmann, W. M. (1997), Listening in a room and the precedence effect, *in* R. H. Gilkey and T. R. Anderson, eds, 'Binaural and Spatial Hearing in Real and Virtual Environments', Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp. 349–376.

Hartmann, W. M. and Rakerd, B. (1989), 'Localization of sound in rooms, IV: The Franssen effect', *J. Acoust. Soc. Am.* **86**(4), 1366–1373.

Hartung, K. and Trahiotis, C. (2001), 'Peripheral auditory processing and investigations of the "precedence effect" which utilize successive transient stimuli', *J. Acoust. Soc. Am.* **110**(3), 1505–1513.

Hawley, M. L., Litovsky, R. Y. and Colburn, H. S. (1999), 'Speech intelligibility and localization in a multi-source environment', *J. Acoust. Soc. Am.* **105**(5), 3436–3448.

Hedelin, P. (1981), A tone-oriented voice-excited vocoder, *in* 'Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.', Vol. I, IEEE, Atlanta, USA, pp. 205–208.

Herre, J., Brandenburg, K. and Lederer, D. (1994), 'Intensity stereo coding', *96th AES Conv., Feb. 1994, Amsterdam (preprint 3799).*

Herre, J., Faller, C., Ertel, C., Hilpert, J., Hoelzer, A. and Spenger, C. (2004), MP3 Surround: Efficient and compatible coding of multi-channel audio, *in* 'Preprint 116th Conv. Aud. Eng. Soc.'.

Hilson, J. (2004), Mixing with Dolby Pro Logic II Technology, Technical report, Dolby Laboratories. www.dolby.com/tech/PLII.Mixing.JimHilson.html.

Holube, I., Kinkel, M. and Kollmeier, B. (1998), 'Binaural and monaural auditory filter bandwidths and time constants in probe tone detection experiments', *J. Acoust. Soc. Am.* **104**(4), 2412–2425.

Hull, J. (1999), Surround sound past, present, and future, Technical report, Dolby Laboratories. www.dolby.com/tech/.

Huopaniemi, J. (1999), Virtual Acoustics and 3D Sound in Multimedia Signal Processing, PhD thesis, Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland. Rep. 53.

IEEE (1969), 'IEEE recommended practice for speech quality measurements', *IEEE Trans. Audio Electroacoust.* **17**(3), 137–148.

ISO 389 (1975), 'Acoustics — standard reference zero for the calibration of pure-tone audiometers'.

ISO/IEC (1993), *Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio*, ISO/IEC 11172-3 International Standard. JTC1/SC29/WG11.

ISO/IEC (1997), *Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding*, ISO/IEC 13818-7 International Standard. JTC1/SC29/WG11.

ISO/IEC (1999), *MPEG-4 Audio Version 2*, ISO/IEC 14496-3 International Standard. JTC1/SC29/WG11.

Jain, M., Gallagher, D. T., Koehnke, J. and Colburn, H. S. (1991), 'Fringed correlation discrimination and binaural detection', *J. Acoust. Soc. Am.* **90**(4), 1918–1926.

Janovsky, W. H. (1948), 'An apparatus for three-dimensional reproduction in electroacoustical presentations', *German Federal Republic Patent No. 973570.* (cited in Blauert 1997).

Jayant, N. S. and Noll, P. (1984), *Digital Coding of Waveforms*, Prentice-Hall Signal Processing Series.

Jeffress, L. A. (1948), 'A place theory of sound localization', *J. Comp. Physiol. Psych.* **61**, 468–486.

Johnston, J. D. (1988), Estimation of perceptual entropy using noise masking criteria, *in* 'Proc. ICASSP-88'.

Johnston, J. D. and Ferreira, A. J. (1992), Sum-difference stereo transform coding, *in* 'Proc. ICASSP-92', pp. 569–572.

Karjalainen, M. and Järveläinen, H. (2001), More about this reverberation science: Perceptually good late reverberance, *in* 'Preprint 111th Conv. Aud. Eng. Soc.'.

Kleijn, W. B. and Paliwal, K. K. (1995), *An Introduction to Speech Coding*, Elsevier, Amsterdam.

Koehnke, J., Colburn, H. S. and Durlach, N. I. (1986), 'Performance in several binaural-interaction experiments', *J. Acoust. Soc. Am.* **79**(5), 1558–1562.

Kollmeier, B. and Gilkey, R. H. (1990), 'Binaural forward and backward masking: Evidence for sluggishness in binaural detection', *J. Acoust. Soc. Am.* **87**(4), 1709–1719.

Kurozumi, K. and Ohgushi, K. (1983), 'The relationship between the cross-correlation coefficient of two-channel acoustic signals and sound image quality), and apparent source width (asw) in concernt halls', *J. Acoust. Soc. Am.* **74**(6), 1726–1733.

Kuttruff, H. (1991), *Room Acoustics*, Elsevier Science Publishers, Barking, UK.

Langendijk, E. H. A., Kistler, D. J. and Wightman, F. L. (2001), 'Sound localization in the presence of one or two distracters', *J. Acoust. Soc. Am.* **109**(5), 2123–2134.

Lauridsen, H. (1954), 'Nogle forsog med forskellige former rumakustic gengivelske', *Ingenioren* **47**, 906. (cited in Schroeder 1961 and Blauert 1997).

Lauridsen, H. and Schlegel, F. (1956), 'Stereofonie und richtungsdiffuse Klangwiedergabe [Stereophony and directionally diffuse reproduction of sound]', *Gravesaner Blätter* **5**, 28–50. (cited in Blauert 1997).

Lindemann, W. (1986*a*), 'Extension of a binaural cross-correlation model by means of contralateral inhibition, I Simulation of lateralization of stationary signals', *J. Acoust. Soc. Am.* **80**(6), 1608–1622.

Lindemann, W. (1986*b*), 'Extension of a binaural cross-correlation model by means of contralateral inhibition, II The law of the first wave front', *J. Acoust. Soc. Am.* **80**(6), 1623–1630.

Litovsky, R. Y. and Shinn-Cunningham, B. G. (2001), 'Investigation of the relationship among three common measures of precedence: Fusion, localization dominance, and discrimination suppression', *J. Acoust. Soc. Am.* **109**(1), 346–358.

Litovsky, R. Y., Colburn, H. S., Yost, W. A. and Guzman, S. J. (1999), 'The precedence effect', *J. Acoust. Soc. Am.* **106**(4), 1633–1654.

Litovsky, R. Y., Rakerd, B., Yin, T. C. T. and Hartmann, W. M. (1997), 'Psychophysical and physiological evidence for a precedence effect in the median sagittal plane', *J. Neurophysiol.* **77**(4), 2223–2226.

Lochner, J. P. A. and de V. Keet, W. (1960), 'Stereophonic and quasi-stereophonic reproduction', *J. Acoust. Soc. Am.* **32**, 392–401.

Lorenzi, C., Gatehouse, S. and Lever, C. (1999), 'Sound localization in noise in normal-hearing listeners', *J. Acoust. Soc. Am.* **105**(5), 1810–1820.

Makhoul, J. and Berouti, M. (1979), High-frequency regeneration in speech coding systems, *in* 'Proc. ICASSP', Vol. 428-431.

Malvar, H. S. (1992), *Signal processing with lapped transforms*, Artech House.

Martin, K. D. (1997), Echo suppression in a computational model of the precedence effect, *in* 'IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.', New Paltz, NY, USA.

Mason, R. (2002), Elicitation and measurement of auditory spatial attributes in reproduced sound, PhD thesis, University of Surrey. A review of existing measurements that relate to spatial impression.

McAulay, R. J. and Quatieri, T. F. (1986), 'Speech analysis/synthesis based on a sinusoidal representation', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **34**(4), 744–754.

Mershon, D. H., and King, L. E. (1975), 'Intensity and reverberation as factors in the auditory perception of egocentric distance', *Perception & Psychophysics* **18**(6), 409–415.

Mershon, D. H. and Bowers, J. N. (1979), 'Absolute and relative cues for the auditory perception of egocentric distance', *Perception* **8**, 311–322.

Morimoto, M. and Maekawa, Z. (1989), Auditory spaciousness and envelopment, *in* 'Proc. 13th Int. Congr. on Acoustics', Vol. 2, Belgrade, pp. 215–218.

Okano, T., Beranek, L. L. and Hidaka, T. (1998), 'Relations among interaural cross-correlation coefficient ($\mathrm{IACC_E}$), lateral fraction ($\mathrm{LF_E}$), and apparent source width (asw) in concernt halls', *J. Acoust. Soc. Am.* **104**(1), 255–265.

Oppenheim, A. V. and Schaefer, R. W. (1989), *Discrete-Time Signal Processing*, Signal Processing Series, Prentice Hall.

Patterson, R. D., Allerhand, M. H. and Giguère, C. (1995), 'Time-domain modeling of peripheral auditory processing: A modular architecture and software platform', *J. Acoust. Soc. Am.* **98**(4), 1890–1894.

Pollack, I. and Trittipoe, W. (1959*a*), 'Binaural listening and interaural noise cross correlation', *J. Acoust. Soc. Am.* **31**(9), 1250–1252.

Pollack, I. and Trittipoe, W. (1959*b*), 'Interaural noise correlation: Examination of variables', *J. Acoust. Soc. Am.* **31**(12), 1616–1618.

Pulkki, V. (1997), 'Virtual sound source positioning using Vector Base Amplitude Panning', *J. Audio Eng. Soc.* **45**, 456–466.

Pulkki, V. (2001*a*), 'Localization of amplitude-panned sources I: Stereophonic panning', *J. Audio Eng. Soc.* **49**(9), 739–752.

Pulkki, V. (2001*b*), 'Localization of amplitude-panned sources II: Two- and three-dimensional panning', *J. Audio Eng. Soc.* **49**(9), 753–757.

Purat, M., Liebchen, T. and Noll, P. (1997), Lossless transform coding of audio signals, *in* 'Preprint 102th Conv. Aud. Eng. Soc.'.

Purnhagen, H. and Meine, N. (2000), HILN - The MPEG-4 parametric audio coding tools, *in* 'Proc. ISCAS'.

Purnhagen, H., Meine, N. and Edler, B. (2002), Sinusoidal coding using loudness-based component selection, *in* 'Proc. ICASSP'.

Rakerd, B. and Hartmann, W. M. (1985), 'Localization of sound in rooms, II: The effects of a single reflecting surface', *J. Acoust. Soc. Am.* **78**(2), 524–533.

Rakerd, B. and Hartmann, W. M. (1986), 'Localization of sound in rooms, III: Onset and duration effects', *J. Acoust. Soc. Am.* **80**(6), 1695–1706.

Rayleigh (1907), 'On our perception of sound direction', *Philos. Mag.* pp. 13:214–232. (J. W. Strutt).

Rec. ITU-R BS.1116.1 (1997), *Methods for Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Surround Systems*, ITU. http://www.itu.org.

Rec. ITU-R BS.1534 (2003), *Method for the subjective assessment of intermediate quality levels of coding systems*, ITU. http://www.itu.org.

Rec. ITU-R BS.562.3 (1990), *Subjective Assessment of Sound Quality*, ITU. http://www.itu.org.

Rec. ITU-R BS.775 (1993), *Multi-Channel Stereophonic Sound System with or without Accompanying Picture*, ITU. http://www.itu.org.

Rec. ITU-R G.722.1 (1997), *Coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss*, ITU. http://www.itu.org.

Rec. ITU-T G.711 (1993), *Pulse Code Modulation (PCM) of Voice Frequencies*, ITU. http://www.itu.org.

Rec. ITU-T G.729 (1996), *Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP)*, ITU. http://www.itu.org.

Robinson, T. (1994), Simple lossless and near-lossless waveform compression, *in* 'Technical report CUED/F-INFENG/TR.156, Cambridge University, Engineering Department'.

Rumsey, F. (2001), *Spatial Audio*, Focal Press, Music Technology Series.

Rumsey, F. (2002), 'Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm', *J. Audio Eng. Soc.* **50**(9), 651–666.

Sabine, W. C. (1922), *Collected Papers on Acoustics*, University Press Harvard. Reprinted by Dover, New York, 1964.

Schroeder, M. R. (1958), 'An artificial stereophonic effect obtained from a single audio signal', *J. Acoust. Soc. Am.* **6**, 74–79.

Schroeder, M. R. (1961), 'Improved quasi-stereophony and "colorless" artificial reverberation', *J. Acoust. Soc. Am.* **33**, 1061–1064.

Schroeder, M. R. (1962), 'Natural sounding artificial reverberation', *J. Aud. Eng. Soc.* **10**(3), 219–223.

Schroeder, M. R. (1965), 'New method of measuring reverberation time', *J. Acoust. Soc. Am.* **37**, 409–412.

Schuijers, E., Breebaart, J., Purnhagen, H. and Engdegard, J. (2004), Low complexity parametric stereo coding, *in* 'Preprint 117th Conv. Aud. Eng. Soc.'.

Schuijers, E., Oomen, W., den Brinker, A. C. and Gerrits, A. J. (2002), Advances parametric coding for high-quality audio, *in* 'Proc. MPCA'.

Schuijers, E., Oomen, W., den Brinker, B. and Breebaart, J. (2003), Advances in parametric coding for high-quality audio, *in* 'Preprint 114th Conv. Aud. Eng. Soc.'.

Shinn-Cunningham, B. G. (2000), Distance cues for virtual auditory space, *in* 'Proc. 1st IEEE Pacific-Rim Conf. on Multimedia, Sydney, Australia', pp. 227–230.

Singh, P. K., Ando, Y. and Kurihara, Y. (1994), 'Individual subjective diffuseness responses of filtered noise sound fields', *Acustica* **80**, 471–477.

Singhal, S. (1990), High quality audio coding using multipulse LPC, *in* 'Proc. Int. Conf. Acoust., Speech, and Signal Proc.', Vol. I, IEEE, Albuquerque, USA, pp. 1101–1104.

Sinha, D., Johnston, J. D., Dorward, S. and Quackenbush, S. (1997), The perceptual audio coder (PAC), *in* V. Madisetti and D. B. Williams, eds, 'The Digital Signal Processing Handbook', CRC Press, IEEE Press, Boca Raton, Florida, chapter 42.

Slaney, M. (1998), 'Auditory toolbox: Version 2'. Technical Report No. 1998-010. http://rvl4.ecn.purdue.edu/∼malcolm/interval/1998-010/.

Smith, J. O. and Serra, X. (1987), PARSHL: An analysis/synthesis program for nonharmonic sounds based on sinusoidal representation, *in* 'Proc. Int. Computer Music Conf.', pp. 290–297. Cited from (**?**).

Sondhi, M. M., Morgan, D. R. and Hall, J. L. (1995), 'Stereophonic acoustic echo cancellation - an overview of the fundamental problem', *IEEE Signal Processing Lett.* **2**, 148–151.

Spieth, W., Curtis, J. F. and Webster, J. C. (1954), 'Responding to one of two simultaneous messages', *J. Acoust. Soc. Am.* **26**(3), 391–396.

Steinberg, J. and Snow, W. (1934), 'Auditory perspectives - physical factors', *Stereophonic Techniques* pp. 3–7. Audio Engineering Society.

Stoll, G. (1996), ISO-MPEG-2 audio: A generic standard for the coding of two-channel and multichannel sound, *in* N. Gilchrist and C. Grewin, eds, 'Collected Papers on Digital Audio Bit-Rate Reduction', Audio Engineering Society Inc., pp. 43–53.

Streicher, R. and Everest, F. A. (1998), *The New Stereo Soundbook - Second Edition*, Audio Engineering Associates, Pasadena, CA.

Theile, G. (1981*a*), 'Zur Kompatibilität von Kunstkopfsignalen mit intensitätsstereophonen Signalen bei Lautsprecherwiedergabe: Die Klangfarbe [On the compatibility of dummy-head signals with intensity stereophony signals in loudspeaker reproduction: Timbre', *Rundfunktech. Mitt.* **25**, 146–154. (See Blauert 1997, p. 360-364).

Theile, G. (1981*b*), 'Zur Kompatibilität von Kunstkopfsignalen mit intensitätsstereophonen Signalen bei Lautsprecherwiedergabe: Die Richtungsabbildung [On the compatibility of dummy-head signals with intensity stereophony signals in loudspeaker reproduction: Directional imaging', *Rundfunktech. Mitt.* **25**, 67–73. (See Blauert 1997, p. 360-364).

Theile, G. (1981*c*), 'Zur Theorie der optimalen Wiedergabe von stereophonen Signalen über Lautsprecher und Kopfhörer [on the theory of the optimal reproduction of stereophonic signals over loudspeakers and headphones', *Rundfunktech. Mitt.* **25**, 155–169. (See Blauert 1997, pp. 360-364).

Theile, G. and Plenge, G. (1977), 'Localization of lateral phantom sources', *J. Audio Eng. Soc.* **25**(4), 196–200.

Tsutsui, K., Suzuki, H., Shimoyoshi, O., Sonohara, M., Akagiri, K. and Heddle, R. M. (1996), ATRAC: Adaptive transform acoustic coding for MiniDisc, *in* N. Gilchrist and C. Grewin, eds, 'Collected Papers on Digital Audio Bit-Rate Reduction', Audio Engineering Society Inc., pp. 95–101.

van de Par, S., Trahiotis, C. and Bernstein, L. R. (2001), 'A consideration of the normalization that is typically included in correlation-based models of binaural detection', *J. Acoust. Soc. Am.* **109**(2), 830–833.

Viemeister, N. F. and Wakefield, G. H. (1991), 'Temporal integragion and multiple looks', *J. Acoust. Soc. Am.* **90**(2), 858–865.

von Hornbostel, E. M. and Wertheimer, M. (1920), ‘über die Wahrnehmung der Schallrichtung [On the perception of the direction of sound]’, *Sitzungsber. Akad. Wiss. Berlin* pp. 388–396.

Waal, R. and Veldhuis, R. (1991), ‘Subband coding of stereophonic digital audio signals’, *Proc. IEEE ICASSP 1991* pp. 3601–3604.

Wightman, F. L. and Kistler, D. J. (1992), ‘The dominant role of low-frequency interaural time differences in sound localization’, *J. Acoust. Soc. Am.* **91**(3), 1648–1661.

Yin, T. C. T. and Chan, J. C. K. (1990), ‘Interaural time sensitivity in medial superior olive of cat’, *J. Neurophysiol.* **64**(2), 465–488.

Zelinski, R. and Noll, P. (1977), ‘Adaptive transform coding of speech signals’, *IEEE Trans. Acoust. Speech, and Signal Processing* **ASSP-25**, 299–309.

Zurek, P. M. (1987), The precedence effect, *in* W. A. Yost and G. Gourevitch, eds, ‘Directional Hearing’, Springere-Verlag, New York, pp. 85–105.

Zwicker, E. and Fastl, H. (1999), *Psychoacoustics: Facts and Models*, Springer, New York.

Curriculum Vitae

Christof Faller

Mobile Terminals Division
Agere Systems (Lucent Technologies spin-off)
Allentown, PA 18109, USA

Audiovisual Communications Laboratory
Swiss Federal Institute of Technology Lausanne (EPFL)
1015 Lausanne, Switzerland

Personal

Date of birth: January 2, 1974.
Nationality: Swiss.
Civil status: Single.

Education

1994 - 1997	Faculty of Electrical Engineering, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland.
1998	Faculty of Electrical Engineering, ČVUT Praha (Czech Technical University), Prague, Czech Republic.
1999	Faculty of Electrical Engineering, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland.
2000	Music school Jiřina Marková, Prague, Czech Republic.
2001-2004	Ph.D. candidate in Department of Communication Systems, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland.
2001	Faculty of Electrical Engineering, Columbia University, New York, USA (attended one course in networking).

Professional Experience

1994 - 1996	**IT consultant and software developer**, Thomas Kurer Architects, Zürich, Switzerland.

1997	**Internship** (neural model applied for modeling of a plasma process), Swiss Federal Laboratories for Materials Testing and Research (EMPA), Dübendorf, Switzerland.
1994 - 1998	**Independent entrepreneur**, developed audio processing software *HDR-Studio* and *SoundWorks* (concept, development, design, user's manual, localization to English, German, French, Portuguese, and Japanese). Sold about 2000 times worldwide.
1999 - 2000	**Consultant**, Acoustics and Speech Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ, USA (2000: remotely from Czech Republic while pursuing other interests in music).
2000 - 2001	**Principle Investigator**, Acoustics and Speech Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ, USA (while pursuing in parallel Ph.D. for EPFL).
2001 - now	**Principle Investigator**, Mobile Terminals Division, Agere Systems (Lucent Technologies spin-off), Allentown, PA, USA (while pursuing in parallel Ph.D. for EPFL).

Academic activities

1996	Teaching Assistant, Institute for Mechanics, ETH Zürich, Switzerland.
1997 - 1998	Teaching Assistant, Institute for Signal and Information Processing, ETH Zürich, Switzerland.
2002	Workshop chair, Audio Engineering Society (AES) 113th Convention, Los Angeles, USA, Oct. 2002. Title: *"Coding of Spatial Audio: Yesterday, Today, and Tomorrow"*.
2003	Workshop chair, Audio Engineering Society (AES) 114th Convention, Amsterdam, NL, Mar. 2003. Title: *"Low Bitrate Coding of Spatial Audio"*.
2003	Research visits at Acoustics Laboratory, Helsinki University of Technology (HUT), Espoo, Finland (Jan.-March and Nov.-Dec. 2003) (while employed by Agere Systems and pursuing Ph.D.).
2000-2004	Regularly supervised EPFL students for audio related diploma projects.
2000-2004	Regularly reviewed papers for several journals and conferences.

Awards and honors

1. Won the first price at a Swiss youth science contest which was held in honor of the 100 year anniversary of Brown Boveri Corporation (BBC) (now: Asea Brown Boveri, ABB), 1991.

Publications and patents

Journal papers

1. E. M. Moser, C. Faller, S. Pietrzko, and F. Eggimann, "Modeling the functional performance of plasma polymerized thin films," *Thin Solid Films, Elsevier*, pp. 49–54, 1999.
2. J. N. Laneman, C.-E. W. Sundberg, and C. Faller, "Huffman code based error screening and channel code optimization for error concealment in perceptual audio coding (PAC) algorithms," *IEEE Trans. on Broadcasting*, vol. 48, no. 3, pp. 193–206, Sept. 2002.
3. C. Faller, B.-H. Juang, P. Kroon, H.-L. Lou, S. Ramprashad, and C.-E. W. Sundberg, "Technical advances in digital audio radio broadcasting," *Proceedings of the IEEE*, vol. 90, no. 8, pp. 1303–1333, Aug. 2002.
4. F. Baumgarte and C. Faller, "Binaural Cue Coding - Part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, Nov. 2003.
5. C. Faller and F. Baumgarte, "Binaural Cue Coding - Part II: Schemes and applications," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, Nov. 2003.
6. C. Faller and J. Chen, "Suppressing acoustic echo in a sampled envelope space," *IEEE Trans. on Speech and Audio Proc.*, submitted in Aug. 2003 (accepted, in revision).
7. C. Faller, "Parametric multi-channel audio coding: Synthesis of cross-correlation cues," *IEEE Trans. on Speech and Audio Proc.*, submitted in Dec. 2003.
8. C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, submitted in Jan. 2004 (accepted, in revision).

Conference papers

1. M. Erne, G. Moschytz, and C. Faller, "Best wavelet-packet bases for audio coding using perceptual and rate-distortion criteria," in *Proc. ICASSP*, Mar. 1999, vol. 2, pp. 909–912.
2. C. Faller, R. Dvořaková, and P. Horaček, "Short-time prediction of the power consumption," in *Proc. AUTOS 99, Plzen, Czech Republic*, June 1999, pp. 5–10.
3. C. Faller, R. Dvořaková, and P. Horaček, "Short-term load forecasting based on a multi-model," in *Proc. IFAC Symposium on Power Plants and Power Systems Control, Brussels, Belgium*, April 2000.
4. B. Edler, C. Faller, and G. Schuller, "Perceptual audio coding using a time-varying linear pre- and post-filter," in *Preprint 109th Conv. Aud. Eng. Soc.*, Sept. 2000.

5. C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parametrization," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, Oct. 2001, pp. 199–202.
6. C. Faller, "Audio coding using perceptually controlled bitstream buffering," in *Preprint 111th Conv. Aud. Eng. Soc.*, Sept 2001.
7. C. Faller and F. Baumgarte, "Binaural Cue Coding: A novel and efficient representation of spatial audio," in *Proc. ICASSP*, May 2002, vol. 2, pp. 1841–1844.
8. F. Baumgarte and C. Faller, "Estimation of auditory spatial cues for Binaural Cue Coding (BCC)," in *Proc. ICASSP*, May 2002, vol. 2, pp. 1801–1804.
9. C. Faller and F. Baumgarte, "Binaural Cue Coding applied to stereo and multi-channel audio compression," in *Preprint 112th Conv. Aud. Eng. Soc.*, May 2002.
10. F. Baumgarte and C. Faller, "Why Binaural Cue Coding is better than Intensity Stereo Coding," in *Preprint 112th Conv. Aud. Eng. Soc.*, May 2002.
11. C. Faller, R. Haimi-Cohen, P. Kroon, and R. Rothweiler, "Perceptually based joint program coding," in *Preprint 113th Conv. Aud. Eng. Soc.*, October 2002.
12. F. Baumgarte and C. Faller, "Design and evaluation of Binaural Cue Coding schemes," in *Preprint 113th Conv. Aud. Eng. Soc.*, Oct. 2002.
13. C. Faller and F. Baumgarte, "Binaural Cue Coding applied to audio compression with flexible rendering," in *Preprint 113th Conv. Aud. Eng. Soc.*, Oct. 2002.
14. C. Faller, "Perceptually motivated low complexity acoustic echo control," in *Preprint 114th Conv. Aud. Eng. Soc.*, Mar. 2003.
15. C. Faller, "Binaural Cue Coding: Rendering of sources mixed into a mono signal," in *Proc. DAGA 2003, Aachen, Germany*, Mar. 2003 (invited).
16. A. Härmä and C. Faller, "Spatial decomposition of time-frequency regions: Subbands or sinusoids," in *Preprint 116th Conv. Aud. Eng. Soc.*, May 2004.
17. A. Baumgarte, C. Faller and P. Kroon, "Audio Coder Enhancement using Scalable Binaural Cue Coding with Equalized Mixing," in *Preprint 116th Conv. Aud. Eng. Soc.*, May 2004.
18. J. Herre, C. Faller, C. Ertel, J. Hilpert, A. Hoelzer, and C. Spenger, "MP3 Surround: Efficient and compatible coding of multi-channel audio," in *Preprint 116th Conv. Aud. Eng. Soc.*, May 2004.
19. V. Pulkki and C. Faller, "The directional effect of crosstalk in multichannel sound reproduction," in *Proc. 18th Int. Congr. on Acoust. ICA*, April 2004.

20. F. Wallin and C. Faller, "Perceptual quality of hybrid echo canceler/suppressor," in *Proc. ICASSP*, May 2004.
21. J. Herre, P. Kroon, C. Faller, and S. Geyersberger, "Spatial audio coding - an enabling technology for bitrate-efficient and compatible multichannel audio broadcasting," in *Proc. IBC, Amsterdam*, Sept. 2004.

Book chapters

1. E. M. Moser, C. Faller, S. Pietrzko, M. Amberg, K. Kurz, and F. Eggimann, "Neural network modeling of plasma processes," in *Nachhaltige Material und Systemtechnik*, W. Muster and K. Schläpfer, Eds. EMPA Akademie, Dübendorf, 2001.

Patents

1. C. Faller, "Perceptual audio coder bit allocation scheme providing improved perceptual quality consistency," *United States Patent No. 6,499,010*, Jan. 2002, Filed: Jan. 2000.
2. C. Faller, "Method and apparatus for detecting noise-like signal components," *United States Patent No. 6,647,365*, Nov. 2003, Filed: June 2000.
3. B. Edler and C. Faller, "Perceptual coding of audio signals using cascaded filterbanks for performing irrelevancy reduction and redundancy reduction with different spectral/temporal resolution," *United States Patent No. 6,678,647*, Jan. 2004, Filed: June 2000.
4. C. Faller, "Perceptual synthesis of auditory scenes," *United States Patent Application No. 20030026441*, Feb. 2003, Filed: May 2001 (pending).
5. C. Faller and R. Haimi-Cohen, "Method and apparatus for frame-based buffer control in a communication system," *United States Patent Application No. 20030002609*, Jan. 2003, Filed: June 2001 (pending).
6. C. Faller, "Method and apparatus for controlling buffer overflow in a communication system," *United States Patent Application No. 20030002588*, Jan. 2003, Filed: June 2001 (pending).
7. C. Faller, "Distortion-based method and apparatus for buffer control in a communication system," *United States Patent Application No. 20030061038*, Mar. 2003, Filed: Sept. 2001 (pending).
8. F. Baumgarte, J. Chen, and C. Faller, "Backwards-compatible perceptual coding of spatial cues," *United States Patent Application No. 20030035553*, Feb. 2003, Filed: Nov. 2001 (pending).
9. F. Baumgarte and C. Faller, "Coherence-based audio coding and synthesis," *United States Patent Application No. 20030219130*, Nov. 2003, Filed: May 2002 (pending).
10. C. Faller, "Suppression of echo signals and the like," *United States Patent Application No. 20040057574*, Mar. 2004, Filed: Sept. 2002 (pending).

11. Three US patent applications are not published yet and thus not included in this list.