Spatial Auditory Blurring and Applications to Multichannel Audio Coding

Adrien Daniel

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : Psychoacoustique et Codage Audio

(École Doctorale Cerveau–Cognition–Comportement)

présentée par

Adrien Daniel

Spatial Auditory Blurring and Applications to Multichannel Audio Coding

Soutenue publiquement le 23 juin 2011 devant le jury composé de :

M$^{\text{me}}$	Dominique Massaloux *Docteur, Télécom Bretagne*	Examinatrice
MM.	Jean-Dominique Polack *Professeur, Université Pierre et Marie Curie*	Examinateur
	Etienne Parizet *Professeur, INSA Lyon*	Rapporteur
	Jens Blauert *Professeur émérite, Ruhr-University Bochum*	Rapporteur
	Stephen McAdams *Professeur, McGill University*	Directeur
	Olivier Warusfel *Docteur, Université Pierre et Marie Curie*	Co-directeur
M$^{\text{lle}}$	Rozenn Nicol *Docteur, Orange Labs*	Encadrante

Orange Labs – Lannion
Technopôle Anticipa
2 Avenue Pierre Marzin
22300 Lannion

CIRMMT
Schulich School of Music
McGill University
555 Sherbrooke St. West
Montreal H3A 1E3
QC, Canada

IRCAM
1 place Igor-Stravinsky
75004 Paris

École Doctorale Cerveau Cognition Comportement
9 Quai Saint Bernard
Bât B, 7ème étage, porte 700, case 25
75005 Paris

Université Pierre et Marie Curie – Paris 6
Bureau d'accueil, inscription des doctorants
Esc G, 2ème étage
15 rue de l'école de médecine
75270 Paris Cedex 06

Contents

Abbreviations **7**

Introduction **9**

Motivations of this Thesis . . . 9

Contributions . . . 10

1 Background on Hearing **11**

1.1 Ear Physiology . . . 11

1.1.1 Body and Outer Ear . . . 11

1.1.2 Middle Ear . . . 11

1.1.3 Inner Ear . . . 12

1.2 Integration of Sound Pressure Level . . . 14

1.2.1 Hearing Area and Loudness . . . 14

1.2.2 Temporal and Frequency Masking Phenomena . . . 18

1.2.3 Critical Bands . . . 19

1.3 Auditory Scene Analysis . . . 20

1.3.1 General Acoustic Regularities used for Primitive Segregation . . . 21

1.3.2 Apparent Motion and Auditory Streaming . . . 22

1.3.3 The Principle of Exclusive Allocation . . . 23

1.3.4 The Phenomenon of Closure . . . 24

1.3.5 Forces of Attraction . . . 26

1.4 Spatial Hearing . . . 27

1.4.1 Localization in Azimuth . . . 27

1.4.2 Localization in Elevation . . . 33

1.4.3 Localization in Distance . . . 33

1.4.4 Apparent Source Width . . . 34

1.4.5 Localization Performance . . . 34

1.4.6 Binaural Unmasking . . . 38

1.5 Localization Cues in Auditory Scene Analysis . . . 40

1.5.1 Sequential Integration . . . 41

1.5.2 Simultaneous Integration . . . 43

1.5.3 Interactions Between Sequential and Simultaneous Integrations . . . 45

1.5.4 Speech-Sound Schemata . . . 46

1.6 Conclusions . . . 47

2 State-of-the-Art of Spatial Audio Coding **49**

2.1 Representation of Spatial Audio . . . 49

2.1.1 Waveform Digitization . . . 49

2.1.2 Higher-Order Ambisonics . . . 50

2.2 Coding of Monophonic Audio Signals 53
2.2.1 Lossless Coding 53
2.2.2 Lossy Coding 54
2.3 Lossless Matrixing 55
2.3.1 Mid/Side Stereo Coding 55
2.3.2 Meridian Lossless Packing 55
2.4 Lossy Matrixing 55
2.4.1 Perceptual Mid/Side Stereo Coding 55
2.4.2 Matrix Encoding 56
2.4.3 Matrixing Based on Channel Covariance 58
2.5 Parametric Spatial Audio Coding 60
2.5.1 Extraction of the Spatial Parameters 62
2.5.2 Computation of the Downmix Signal 66
2.5.3 Spatial Synthesis 67
2.5.4 Quantization of the Spatial Parameters 69
2.6 Conclusions 69

3 Spatial Blurring 71
3.1 Motivations 71
3.2 Terms and Definitions 72
3.3 Paradigm for MAA Assessment 73
3.4 Stimuli 74
3.5 Subjects, Rooms and General Setup 75
3.6 Experiment 1: Spatial Blurring From One Distracter 76
3.6.1 Setup 77
3.6.2 Procedure 77
3.6.3 Data Analysis and Results 81
3.7 Experiment 2: Effect of the Signal-to-Noise Ratio 83
3.7.1 Setup 84
3.7.2 Procedure 85
3.7.3 Tasks 86
3.7.4 Adaptive Method Setup 87
3.7.5 Data Analysis and Results 88
3.8 Experiment 3: Effect of the Distracter Position 91
3.8.1 Setup 91
3.8.2 Procedure 93
3.8.3 Data Analysis and Results 93
3.8.4 Validity of Our Experimental Protocol 94
3.9 Experiment 4: Interaction Between Multiple Distracters 96
3.9.1 Setup 97
3.9.2 Procedure 97
3.9.3 Data Analysis and Results 99
3.10 Summary and Conclusions 103

4 Towards a Model of Spatial Blurring and Localization Blur 105
4.1 Assumptions 105
4.2 Formalism and Overview 106
4.3 Computation of Masking Thresholds 106
4.4 Reference Value of Spatial Blurring 108
4.5 Accounting for the Effect of SNR 110

4.6 Additivity of Distracters 112
4.7 Resulting Localization Blur 118
4.8 Simplification of the Model 118
4.9 Conclusions 118

5 Multichannel Audio Coding Based on Spatial Blurring 121
5.1 Dynamic Bit Allocation in Parametric Schemes 122
5.1.1 Principle Overview 122
5.1.2 Use of our Psychoacoustic Model of Spatial Blurring 123
5.1.3 Bit Allocation of the Spatial Parameters 124
5.1.4 Transmission and Bitstream Unpacking 127
5.1.5 Informal Listening 128
5.2 Dynamic Truncation of the HOA Order 129
5.2.1 Spatial Distortions Resulting from Truncation 130
5.2.2 Principle Overview 132
5.2.3 Modes Of Operation 134
5.2.4 Time-Frequency Transform 134
5.2.5 Spatial Projection 135
5.2.6 Spatio-Frequency Analysis 136
5.2.7 Psychoacoustic Model 137
5.2.8 Space Partitioning 137
5.2.9 Space Decomposition 138
5.2.10 HOA Order Truncation 138
5.2.11 Bit-Quantization by Simultaneous Masking 138
5.2.12 Bitstream Generation 139
5.2.13 Decoding 139
5.3 Conclusions 139

Conclusions 141

A Instructions Given to the Subjects 143
A.1 Left-Right/Right-Left Task 143
A.2 Audibility Task 143

B Study of Inter- and Intra-subject variability 145

C Audio Coding Based on Energetic Masking 149
C.1 Principle Overview 149
C.2 Quantization Errors 150
C.3 Modeling Simultaneous Masking Curves 151
C.4 Computation of Masking Curves 153
C.5 Bit Allocation Strategies 154
C.5.1 A Simple Allocation Method by Thresholding 154
C.5.2 Perceptual Entropy 154
C.5.3 Optimal Bit Allocation for a Fixed Bitrate 155
C.6 Bitstream Format 156

Bibliography 157

Abbreviations

ASA: Auditory Scene Analysis (section 1.3).

BCC: Binaural Cue Coding (section 2.5).

CN: Cochlear Nucleus (section 1.4.1).

DirAC: Directional Audio Coding (section 2.5).

EE: Excitatory-Excitatory cell (section 1.4.1).

EI: Excitation-Inhibition cell (section 1.4.1).

ERB: Equivalent Rectangular Bandwidth (section 1.2.3).

HOA: Higher-Order Ambisonics (section 2.1.2).

HRTF: Head-Related Transfer Function (section 1.4.1).

IC: Interaural Coherence (section 1.4.1).

ICC: Inter-Channel Coherence (section 2.5.1).

ICLD: Inter-Channel Level Difference (section 2.5.1).

ICTD: Inter-Channel Time Difference (section 2.5.1).

ILD: Interaural Level Difference (section 1.4.1).

IPD: Interaural Phase Difference (section 1.4.1).

ITD: Interaural Time Difference (section 1.4.1).

JND: Just-Noticeable Difference (section 1.4.5).

LSO: Lateral Superior Olive (section 1.4.1).

MAA: Minimum Audible Angle (section 3.2).

MDCT: Modified Discrete Cosine Transform (see [PJB87]).

MNTB: Medial Nucleus of the Trapezoid Body (section 1.4.1).

MSO: Medial Superior Olive (section 1.4.1).

PS: Parametric Stereo (section 2.5).

SASC: Spatial Audio Scene Coding (section 2.5).

SNR: Signal-to-Noise Ratio (section 3.7).

SOC: Superior Olivary Complex (section 1.4.1).

STFT: Short-Time Fourier Transform.

VBAP: Vector Base Amplitude Panning (see [Pul97]).

VBIP: Vector Base Intensity Panning (see [Pul97]).

Introduction

Motivations of this Thesis

This thesis deals with several aspects of audio coding. Audio coding concerns the way an audio signal is represented in order to store or transmit it. This work is motivated by the data compression problem in audio coding, that is, the study of representations that produce a significant reduction in the information content of a coded audio signal. This thesis deals specifically with the coding of multichannel audio signals. Multichannel audio aims to reproduce a recorded or synthesized sound scene using a loudspeaker array, usually located on a circle in a horizontal plane around the listener—such as for the 5.1 standard—but this array can also be extended to a sphere surrounding the listener.

As reviewed in chapter 2, there are two main approaches to audio coding: lossless and lossy coding. Lossless coders use mathematical criteria such as prediction or entropy coding to achieve data compression, and ensure perfect reconstruction of the original signal. Lossy coding, however, plays on the precision of coding of the signal to achieve data compression and will thus only approximate the original signal. This type of coder is generally used in combination with perceptual criteria, which drive the alteration of the signal in order to minimize its perceptual impact. For instance, several perceptual coders are based on energetic masking phenomena: components of an audio signal can potentially mask other components that are close to them temporally and/or in frequency. In these coders, a psychoacoustic model of energetic masking drives the bit allocation procedure of each audio sample in the frequency domain such that the noise resulting from the quantization of these samples is kept below masking threshold and thus remains inaudible. In other words, interferences between components are exploited to achieve data compression. Audio coding based on energetic masking is described in appendix C.

Specifically regarding multichannel audio, its coding is very costly given the increase in the number of channels to represent. Lossless and lossy coding methods, including perceptual coding based on energetic masking, have been extended to multichannel audio and achieved significant data compression. However, spatial hearing is not extensively taken into account in the design of these coders. Indeed, as detailed in chapter 1, the auditory spatial resolution—that is, the ability to discriminate sounds in space—is limited. In some parametric coders, a few characteristics of this limitation are used to define the spatial accuracy with which the multichannel audio signal is to be represented. However, the auditory spatial resolution under consideration corresponds to the very specific experimental condition in which the auditory system presents its best performance: the sound scene is composed of a single sound source. Indeed, experimental results have shown that errors of localization increase when the sound scene gets more complex, which suggests that auditory spatial resolution degrades in the case of interferences between simultaneously present sound sources. The first aim of this thesis was to bring to light and study this phenomenon. The second aim was to exploit it in a way similar to the way in which energetic masking is used: by adjusting *dynamically* the spatial accuracy with which the multichannel audio signal is to be represented, depending on auditory spatial resolution. To our knowledge, this approach has not been used in audio coding to date.

Contributions

The contributions of this thesis are divided into three parts. The first part, described in chapter 3, consists of four psychoacoustic experiments aiming to study auditory spatial resolution—also known as “localization blur”—in the presence of distracting sounds. Methodologies which were specifically developed to carry out these experiments are also presented in this chapter. As a result, localization blur increases when these distracters are present, bringing to light what we will refer to as the phenomenon of “spatial blurring”. In these experiments, totaling more than 200 hours of testing, we studied the dependence of spatial blurring on the following variables: the frequencies of both the sound source under consideration and the distracting sources, their level, their spatial position, and the number of distracting sources. Except for the spatial position of distracting sources, all of these variables have been shown to have an effect on spatial blurring.

Secondly, we propose a model of our experimental results. This model, detailed in chapter 4, provides an estimation of localization blur as a function of the sound scene characteristics (number of sources present, their frequency, and their level) by taking into account spatial blurring. It is based on a combination of three simpler models. First, the frequency of the sound under consideration as well as the frequency distance to a single distracter yield a first estimate of the spatial blurring created by this distracter. Second, the sound level difference between these two sounds is accounted for to correct this estimation. And third, the estimates of spatial blurring created by each distracter are combined according to an additivity rule derived from our experimental results.

Finally, in a last part described in chapter 5, we propose two multichannel audio coding schemes taking advantage of spatial blurring to achieve data compression. The general idea is the same for both schemes. The precision with which the spatial aspect of the sound scene can be represented is necessarily limited by the bit pool available to represent the signal. The schemes we propose dynamically adjust the accuracy of the spatial representation of the audio signal in a way that shapes the resulting spatial distortions within localization blur, such that they remain (as much as possible) unnoticeable. Our psychoacoustic model of spatial blurring and localization blur is thus used to drive the spatial accuracy of representation in both schemes. The first coding scheme is integrated into parametric spatial audio coding schemes. In these schemes, the multichannel audio input signal is represented as a downmix signal plus a set of spatial parameters reflecting the spatial organization of the sound scene. The accuracy of the spatial representation depends on the number of bits allocated to code these parameters, which is kept fixed. We propose to dynamically allocate these bits according to our psychoacoustic model. Some informal listening results based on this approach are reported in this chapter. The second coding scheme we investigate is based on the Higher-Order Ambisonics (HOA) representation. It is based on a dynamic truncation of the HOA order of representation according to our psychoacoustic model.

Chapter 1

Background on Hearing

This first chapter deals with the background knowledge on hearing related to this thesis. Sound acquisition by the ears is presented first (section 1.1), after which the integration of sound pressure level (section 1.2) and space (section 1.4) by the auditory system is described. Particular attention is given to spatial hearing in azimuth, because the work presented in this thesis is focused on the azimuthal plane (although it could be extended to elevation). High-level processes which are involved in the production of a description of the auditory scene are also reviewed (sections 1.3 and 1.5).

1.1 Ear Physiology

An overview of the ear is depicted in **figure 1.1**. It is composed of three regions: the outer, the middle, and the inner ear. The outer ear collects sound, in the form of a pressure wave, and transmits it to the middle ear, which converts it into fluid motions in the cochlea. The cochlea then translates the fluid motions into electrochemical signals sent through the auditory nerve. Details about the physiology of hearing can be found for example in [vB60, ZF07].

1.1.1 Body and Outer Ear

The head and shoulders have an important effect before the pressure wave reaches the outer ear: by shadowing and reflections, they distort the sound field. Likewise, the pinna and the ear canal both considerably modify the sound pressure level transmitted to the eardrum in the middle ear. The pinna acts like a filter, and alters mainly the high frequencies of the spectrum of the sound field by reflections. Note that the spectrum resulting from the effects of the pinna, the head and the shoulders is fundamental for sound localization (see section 1.4). The auditory canal, which can be modeled as an open pipe of about 2 cm length, has a primary resonant mode at 4 kHz and thus increases our sensitivity in this area, which is useful for high frequencies in speech (as shown in **figure 1.5**).

1.1.2 Middle Ear

The sensory cells in the inner ear are surrounded by a fluid. The middle ear thus converts pressure waves from the outer ear into fluid motion in the inner ear (see **figure 1.1**). The ossicles, made of very hard bones, are composed of the malleus (or hammer, which is attached to the eardrum), the incus (or anvil) and the stapes (or stirrup), and act as lever and fulcrum to convert large displacements with low force of air particles into

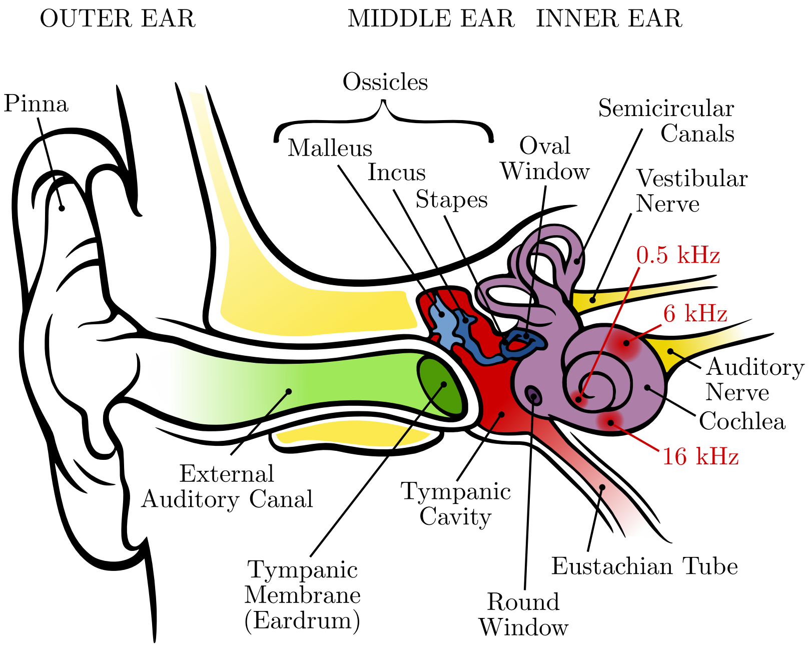


Figure 1.1: General overview of the human ear. (Reproduced from [Fou11].)

small displacements with high force of fluid. Indeed, because of the impedance mismatch between air and fluid, to avoid energy loss at their interface, the middle ear mechanically matches impedances through the relative areas of the eardrum and stapes footplate, and with the leverage ratio between the legs of the hammer and anvil. The best match of impedances is obtained at frequencies around 1 kHz, also useful for speech. Finally, sound waves are transmitted into the inner ear by the stapes footplate through a ring-shaped membrane at its base called the oval window.

Note that the middle-ear space is closed off by the eardrum and the Eustachian tube. Although, the Eustachian tube can be opened briefly when swallowing or yawning. This can be used to resume normal hearing in situations of extreme pressure change in that space.

1.1.3 Inner Ear

The inner ear, also called the labyrinth, is mainly composed of the cochlea (shown in **figure 1.2**), which is a spiral-shaped, hollow, conical chamber of bone. The cochlea is made of three fluid-filled channels, the scalae (depicted in the cross-sectional diagram in **figure 1.2**). Reissner's membrane separates the scala vestibuli from the scala media, but is so thin that the two channels can be considered as a unique hydromechanical unit. Thus, the footplate of the stapes, in direct contact with the fluid in the scala vestibuli, transmits oscillations to the basilar membrane through the fluids. The organ of Corti (see **figure 1.3**) contains the sensory cells (hair cells) that convert these motions into electrochemical impulses and is supported by the basilar membrane. Experimental results from von Békésy [vB60] confirmed a proposition of von Helmholtz: a sound of a particular frequency produces greater oscillations of the basilar membrane at a particular point (tonotopic coding), low frequencies being towards the apex (or helicotrema), and high ones near the oval window, at the base. Consequently, the cochlea acts as a filter bank, as illustrated in **figure 1.4**. Because of the fluid incompressibility, the round window is

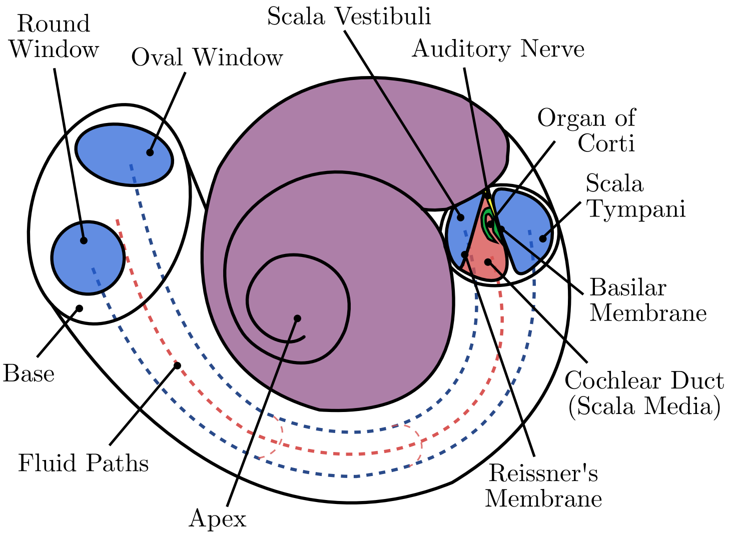


Figure 1.2: Schematic draw of the cochlea. (Reproduced from [Fou11].)

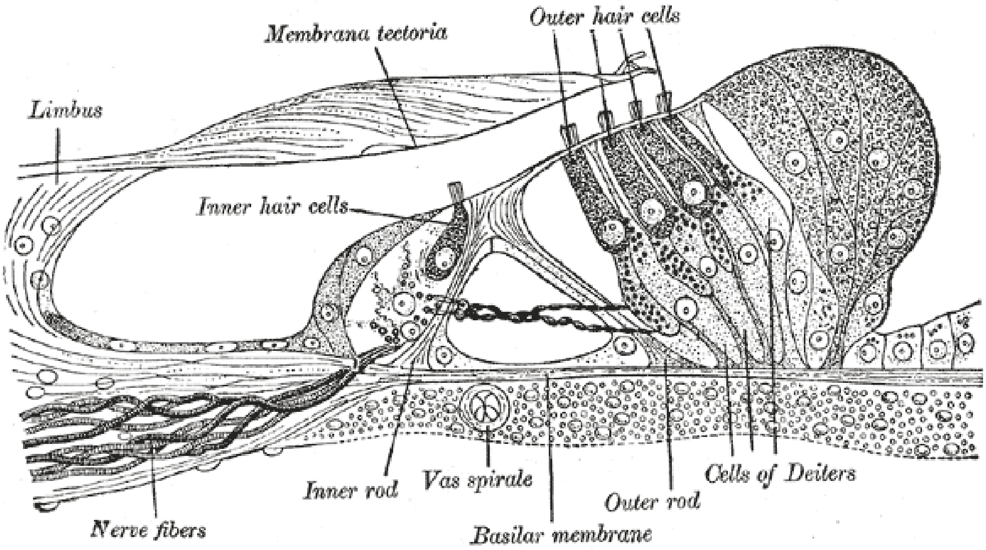


Figure 1.3: The organ of Corti. (After Gray's Anatomy [Fou11].)

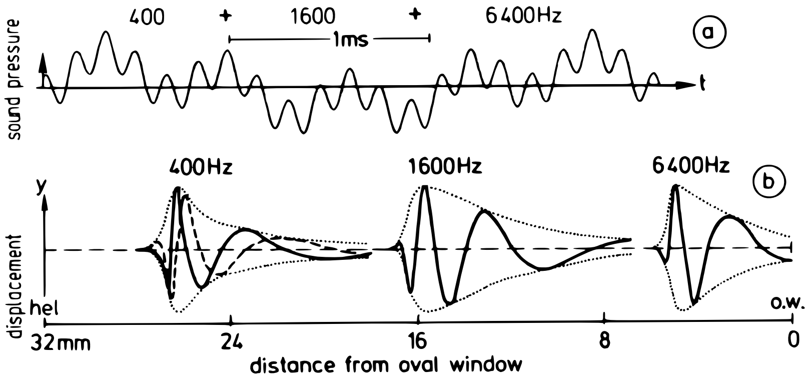


Figure 1.4: Transformation of frequency into place along the basilar membrane. The sound presented in (**a**) is made of three simultaneous tones. The resulting oscillations of the basilar membrane are shown in (**b**). The solid plots represent the instant where the maximum is reached, and the dotted plot at 400 Hz is the instant a quarter period earlier. (After [ZF07].)

necessary to equalize the fluid movement induced from the oval window. The electrical impulses produced by the hair cells are finally transmitted to the brain via the auditory (or cochlear) nerve.

1.2 Integration of Sound Pressure Level

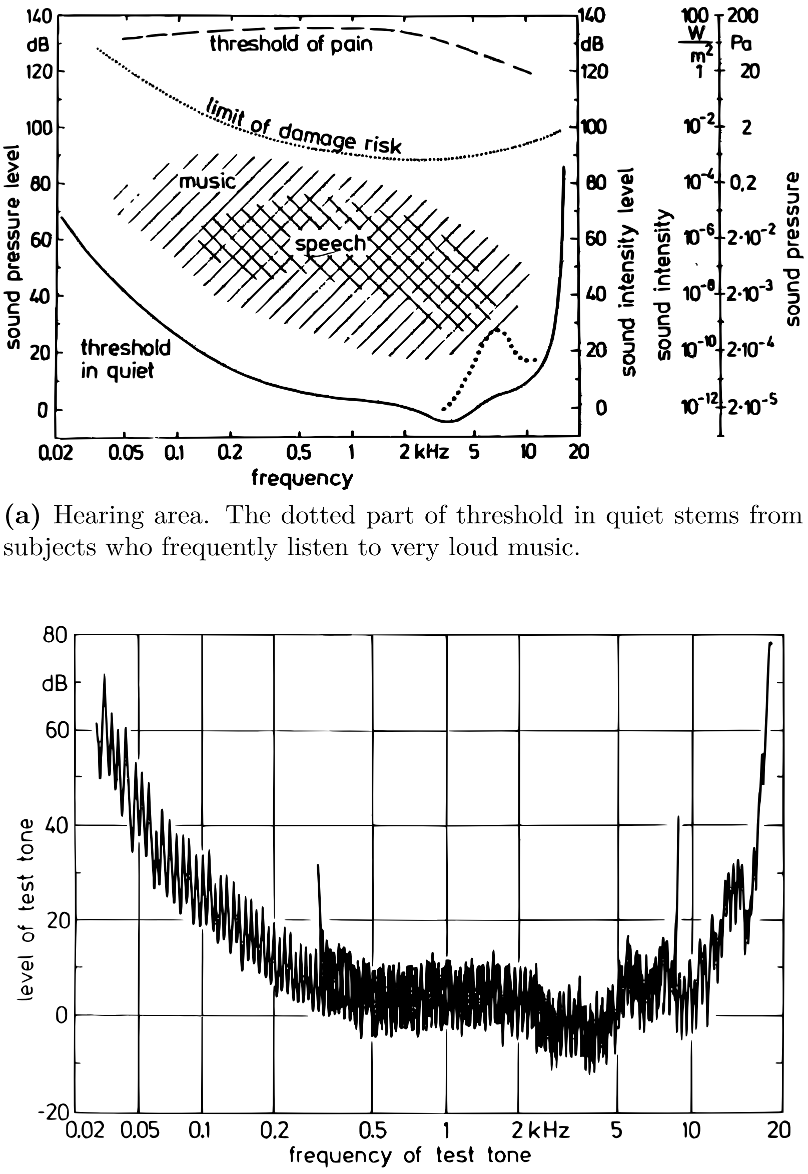
1.2.1 Hearing Area and Loudness

The hearing area (represented in **figure 1.5**) is the region in the SPL/frequency plane in which a sound is audible. In frequency, its range is usually 20 Hz - 20 kHz. In level, it is delimited by two thresholds: the threshold in quiet (or hearing threshold), which represents the minimum necessary level to hear a sound, and the threshold of pain. Their values both depend on frequency. Terhardt [Ter79] proposed an approximation of the threshold in quiet:

$$A(f) = 3.64 f^{-0.8} - 6.5 e^{-0.6(f-3.3)^2} + 10^{-3} f^4, \tag{1.1}$$

where A is expressed in dB and f in kHz.

Loudness, as proposed by Barkhausen [Bar26] in the 1920s, is the subjective perception of sound pressure. The loudness level of a particular test sound, for frontally incident plane waves, is defined as the necessary level for a 1-kHz tone to be perceive as loud as this test sound. Two units exist to express loudness: the sone and the phon, the latter being generally used. 1 phon is equal to 1 dB SPL [1] at a frequency of 1 kHz. The equal-loudness contours (shown in **figure 1.6**) are a way of mapping the dB SPL of a pure tone to the perceived loudness level in phons, and are now defined in the international standard ISO 226:2003 [SMR$^+$03]. According to the ISO report, the curves previously standardized for ISO 226 in 1961 and 1987 as well as those established in 1933 by Fletcher and Munson [FM33] were in error.



(a) Hearing area. The dotted part of threshold in quiet stems from subjects who frequently listen to very loud music.

(b) Békésy tracking method used for an experimental assessment of threshold in quiet. A test tone whose frequency is slowly sweeping is presented to the subject, who can switch between continuously incrementing or decrementing the sound pressure level.

Figure 1.5: Hearing area and its measurement (After [ZF07].)

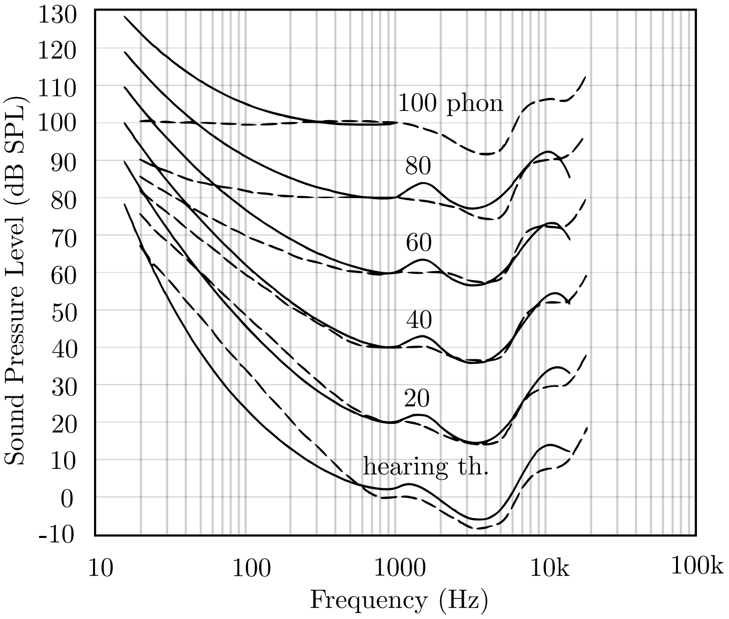


Figure 1.6: Solid plots: equal-loudness contours (from ISO 226:2003 revision). Dashed plots: Fletcher-Munson curves shown for comparison. (Reprinted from [SMR$^+$03].)

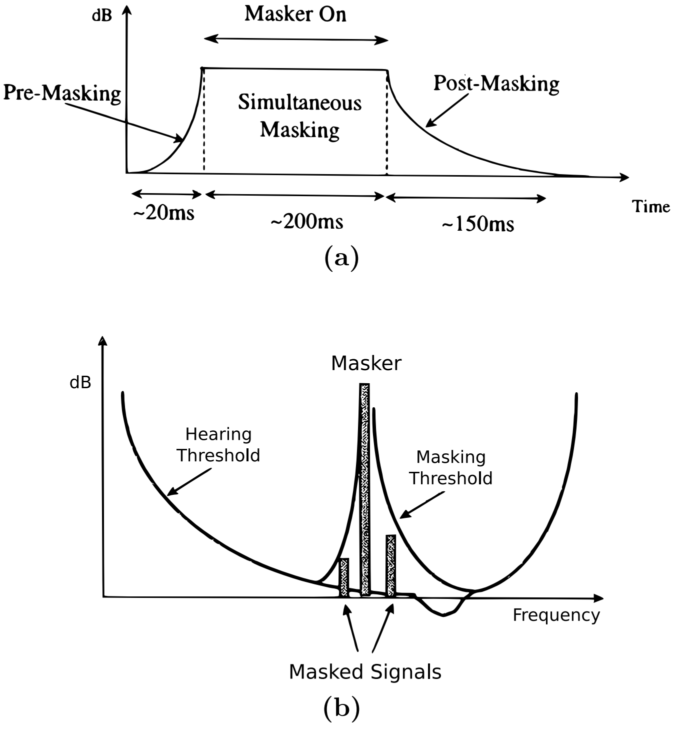


Figure 1.7: Temporal masking **(a)** and frequency masking **(b)** phenomena. (After [BG02].)

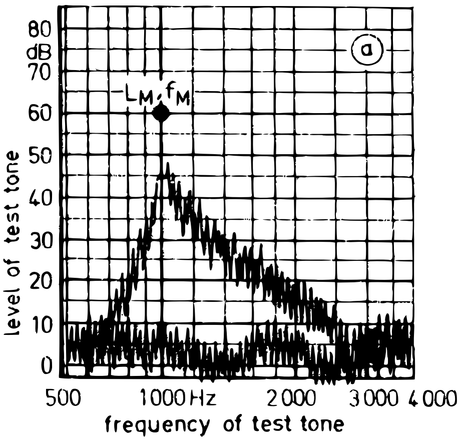
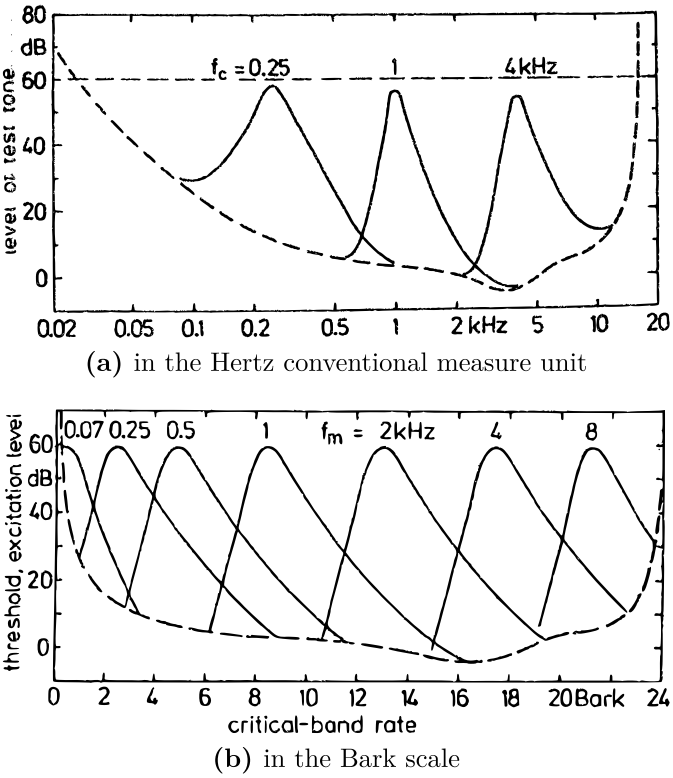


Figure 1.8: Békésy tracking method used for an experimental assessment of a masking curve (and also of the hearing threshold, the bottom curve). In addition to the masking tone, a test tone whose frequency is slowly sweeping is presented to the subject, who can switch between continuously incrementing or decrementing the sound pressure level. (After [ZF07].)



(a) in the Hertz conventional measure unit

(b) in the Bark scale

Figure 1.9: Excitation patterns for narrow-band noise signals centered at different frequencies and at a level of 60 dB. (After [ZF07].)

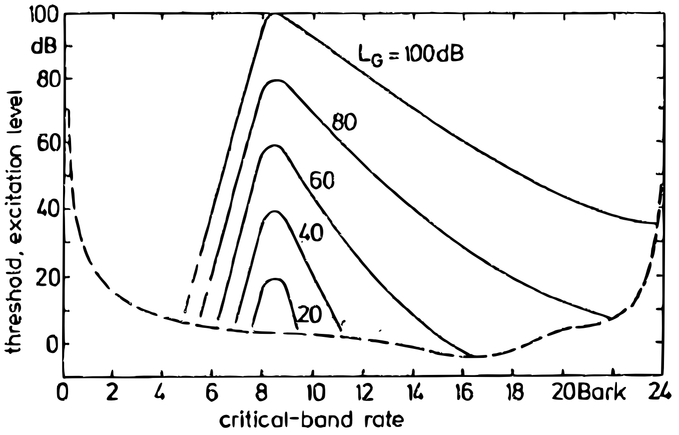


Figure 1.10: Excitation patterns for narrow-band noise signals centered at 1 kHz and at different levels. (After [ZF07].)

1.2.2 Temporal and Frequency Masking Phenomena

Masking occurs when a louder sound (the masker) masks a softer one (the masquee). Two kinds of masking phenomena can be experienced.

The first one is referred to as *temporal masking* and occurs when a masker masks sounds immediately preceding and following it. This phenomenon is thus split into two categories: *pre-masking* and *post-masking* (see **figure 1.7a**). Pre-masking is most effective a few milliseconds before the onset of the masker, and its duration has not been conclusively linked to the duration of the masker. Pre-masking is less effective with trained subjects. Post-masking is a stronger and longer phenomenon occuring after the masker offset, and depending on the masker level, duration, and relative frequency of masker and probe.

The second masking phenomena is known as *frequency masking* or *simultaneous masking*. It occurs when a masker masks simultaneous signals at nearby frequencies. The curve represented in **figure 1.7b** as "masking threshold" represents the altered audibility threshold for signals in the presence of the masking signal. This curve is determined by psychoacoustic experiments similar to the one determining the hearing threshold presented on **figure 1.5**, with the additional presence of the masker, and is highly related to the nature and the characteristics of the masker and the probe (see **figure 1.8**). The level difference at a certain frequency between a signal component and the masking threshold is called *signal-to-mask ratio* (SMR). The shape of the masking curve resulting from a given masker depends on the kind of maskee (tone or narrow-band noise), the kind of masker (idem), but also the sound level and the frequency of the masker. Concerning this last point though, it is admitted that the frequency dependence of experimental masking curves is an artifact of our measurement units. Indeed, when described in frequency units that are linearly related to basilar membrane distances, like the Bark scale (see section 1.2.3), experimental masking curves are independent of the frequency of the masker (see **figure 1.9**). However, the level dependence still remains, as illustrated in **figure 1.10**. The minimum SMR, that is to say the difference between the masker SPL and the maximum of its masking curve, depends on the same characteristics as the shape of the masking curve. All these aspects are studied in-depth in [ZF07, ZF67], and well summarized in [BG02].

[1] 0 dB SPL = 2×10^{-5} Pa, and roughly corresponds to the auditory threshold in quiet at 2 kHz. Note that the definition of the phon (1 phon = 1 dB SPL at 1 kHz) does not imply that 0 phon (i.e., 0 dB SPL at 1 kHz) matches the threshold in quiet; it depends on the hearing threshold curve under consideration.

Models of masking curves are presented in appendix C.3.

1.2.3 Critical Bands

The concept of critical bands was introduced by Fletcher in 1940 [Fle40]: the human ear has the faculty to integrate frequency ranges using auditory filters called critical bands.

Critical bands condition the integration of loudness. When measuring the hearing threshold of a narrow-band noise as a function of its bandwidth while holding its overall pressure level constant, this threshold remains constant as long as the bandwidth does not exceed a critical value, the critical bandwidth. When exceeding this critical bandwidth, the hearing threshold of the noise increases.

Critical bands are related to masking phenomena. If a test signal is presented simultaneously with maskers, then only the masking components falling within the same critical band as the test signal have a maximum masking contribution. In other words, there is a frequency region around a given masking component called critical bandwidth where its masking level is constant. Outside this region, its masking level drops off rapidly. Several methods used for measuring critical bandwidths are described by Zwicker and Fastl in [ZF07]. They propose an analytic expression of critical bandwidths Δf as a function of the masker center frequency f_c:

$$\Delta f = 25 + 75 \left[1 + 1.4(0.001 f_c)^2\right]^{0.69}. \tag{1.2}$$

For center frequencies below 500 Hz, their experimental results report that critical bandwidths are frequency independent and equal to 100 Hz. Then, for frequencies above 500 Hz, critical bandwidths roughly equal 20% of the center frequency. Critical bandwidths are independent of levels, although bandwidths increase somewhat for levels above about 70 dB.

It should be noted that some authors (e.g. [GDC$^+$08], see section 1.4.6) postulate that listeners are able to widen their effective auditory filters greater than a critical band in response to the variability of a stimulus.

The Bark scale

A critical band is an auditory filter that can be centered on any frequency point. However, adding one critical band to the next in such a way that the upper limit of the lower critical band corresponds to the lower limit of the next higher critical band leads to 24 abutting critical bands. The scale produced in this way is called critical band rate, z, and has the unit “Bark.” Hence, $z = 1$ Bark corresponds to the upper limit of the first and the lower limit of the second critical band, $z = 2$ Bark to the upper limit of the second and the lower limit of the third, and so forth. It has been experimentally demonstrated in [ZF67] that the ear, in order to establish the hearing threshold of a wide-band noise, actually divides the hearing area in 24 abutting critical bands. The Bark scale is then a mapping of frequencies onto a linear distance measure along the basilar membrane, using the critical bandwidth as unit. It is possible to approximate the critical band rate $z(f)$ using the following expression from [ZF07]:

$$z(f) = 13 \arctan\left(\frac{0.76 f}{0.001}\right) + 3.5 \arctan\left[\left(\frac{f}{0.0075}\right)^2\right]. \tag{1.3}$$

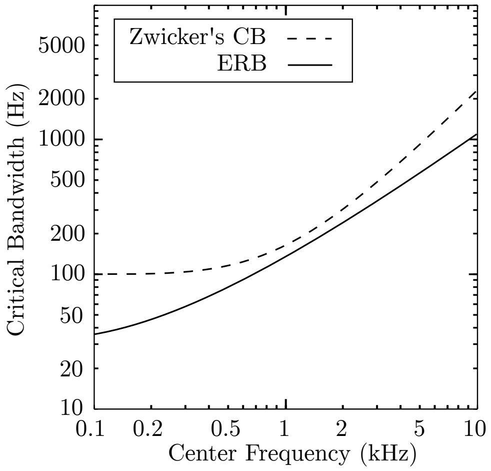


Figure 1.11: Comparison of Zwicker's model of critical bandwidth as defined on equation (1.2), and Moore and Glasberg's equivalent rectangular bandwidth as defined on equation (1.5).

Equivalent Rectangular Bandwidth (ERB)

Several authors [Gre61, Sch70], using different measurement methods, do not agree with the equation (1.2) proposed by Zwicker, particularly for center frequencies below 500 Hz. Moore and Glasberg [MG96] used a notched-noise method [GM90], for which the measure of masking is not affected by the detection of beats or intermodulation products between the signal and masker. Besides, the effects of off-frequency listening are taken into account. The Bark scale is then replaced by what they call the *Equivalent Rectangular Bandwidth* (ERB) scale:

$$\mathrm{ERBS}(f) = 21.4 \log_{10}(0.00437f + 1), \tag{1.4}$$

and the critical bandwidth is given by:

$$\mathrm{ERB}(f) = 0.108f + 24.7. \tag{1.5}$$

As can be seen in **figure 1.11**, the equivalent rectangular bandwidths are particularly narrower than the critical bandwidths proposed by Zwicker for frequencies below 500 Hz. Auditory filters following the ERB scale are often implemented using gammatone filters [PRH$^+$92].

1.3 Auditory Scene Analysis

Auditory scene analysis (ASA) describes the way the auditory system organizes the auditory inputs to build a description of the components of the auditory scene. To reveal the mechanisms underlying this complex process, Bregman [Bre94] proposed the concept of "auditory streams," which tries to explain how the auditory system groups acoustic events into perceptual units, or *streams*, both instantaneously and over time. An auditory stream is different from a sound (an acoustic event) in the sense that it represents a single happening, so that a high-level mental representation can be involved in the stream segregation process. As an example, a series of footsteps can form a single experienced event, even if each footstep constitutes a separate sound. The spatial aspects of ASA are discussed in section 1.5.

As we will see, many of the phenomena occurring in audition have an analogy in vision, and some parts of the concept of auditory stream are inspired by studies of the Gestalt psychologists [Ash95].

The stream segregation ability of our auditory system seems both innate and learned. An experiment by Demany [Dem82], using the habituation and dishabituation method, [2] demonstrated that infants are already able to segregate high tones and low ones within a sequence into two streams. Later, McAdams and Bertoncini [MB97] obtained results suggesting that newborn infants organize auditory streams on the basis of source timbre and/or spatial position. Although these results also showed that newborns have limits in temporal and/or pitch resolution when discriminating tone sequences, which suggest that stream segregation is also a learned ability, in the same way musicians improve their segregation capability (of instruments for instance) by practicing music. Thereby, innate abilities seem to act as a “bootstrap” for the acquisition of an accurate segregation system.

The effects of the unlearned constraints of the auditory scene analysis process are called by Bregman “primitive segregation,” and those of the learned ones “schema-based segregation.” The primitive segregation is based on relations between the acoustic properties of a sound, which constitute general acoustic regularities.

1.3.1 General Acoustic Regularities used for Primitive Segregation

As explained above, these regularities of the world are used for scene analysis, even if the listener is not familiar with the signal. Bregman reports four of them that have been identified as utilized by the auditory system:

1. It is extremely rare that sounds without any relations between them start and stop precisely at the same time.

2. Progression of the transformation:
   i. The properties of an isolated sound tend to change continuously and slowly.
   ii. The properties of a sequence of sounds arising from the same source tend to change slowly.

3. When a sounding object is vibrating at a repeated period, its vibrations give rise to an acoustic pattern with frequency components that are multiples of a common fundamental frequency.

4. “Common fate”: Most of modifications arising from an acoustic signal will affect all components of the resulting sound, identically and simultaneously.

The first general regularity is used by the auditory system through what Bregman calls the “old-plus-new” heuristic: when a spectrum suddenly becomes more complex, while holding its initial frequency components, it is interpreted by the nervous system as a continuation of a former signal to which is added a new signal.

The second regularity is based on two rules, of which the auditory system takes advantage. They are related to the sequential modification of sounds of the environment. The first rule concerns the “sudden transformation” of the acoustic properties of a signal, which are interpreted as the beginning of a new signal. It is guided by the old-plus-new heuristic. The suddenness of the spectral transformation acts as any other cue in scene analysis: the greater it is, the more it affects the grouping process. The second rule leads to

[2] This is a kind of method based on a rewarding process which is used with subjects, typically infants, who cannot directly tell if they consider two stimuli as the same or as different.

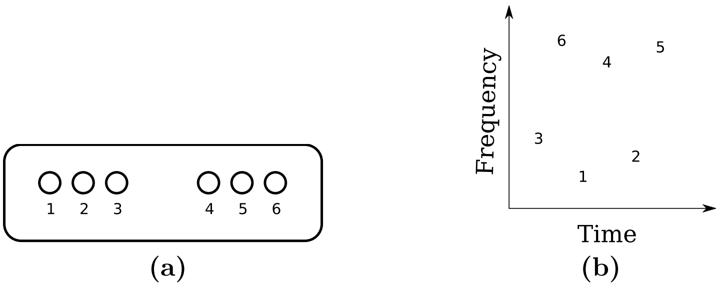


Figure 1.12: **(a)** A device that can be used to show the apparent motion effect, and **(b)** a looping sequence that can be used to show the auditory streaming effect (Reprinted from [Bre94].)

"grouping by similarity." Similarity of sounds is not well understood, but is related to the fundamental frequency, the timbre (spectrum shape) and spatial localization. Similarity (as well as proximity, which is related) is discussed in section 1.3.5.

The third regularity is firstly used by the auditory system to assess the pitch of each sound from a mixture of sound. To do so, it uses the harmonic relations between the partials of a sound. But it is also used to segregate groups of partials from each other, as demonstrated by an experiment of Darwin and Gardner [DG86]. Considering the spectrum of a vowel, they mistuned one of its low-frequency harmonics. For a mustuning of 8%, the identity of the vowel is altered, and the mistuned harmonic is heard as a distinct sound.

The fourth regularity, also called amplitude modulation or AM, has been observed as efficient for grouping spectral components with the same pattern of intensity variation, but the way it is exploited by the auditory system is not well known yet. It can be illustrated by a series of words pronounced by a person, which form a distinct pattern of modification. It is brought to light in laboratory by a phenomenon called the *comodulation masking release* (CMR) [HHF84].

It should be noted that synchronous frequency modulation or FM of partials, which is also a case of "common fate" regularity, is not reported by researches as a relevant regularity for the auditory system, despite its apparent usefulness. Indeed, works on this subject did not show that FM reinforces the tendency for harmonics to group, or causes non-harmonic partials to be perceptually fused. Therefore, harmonics in a mixture that are affected by the same FM are seemingly not grouped together because of their motion, but because of the third regularity.

1.3.2 Apparent Motion and Auditory Streaming

Körte formulated [KÏ5] several laws about the impression of movement that we can get with a panel of electric light bulbs in sequence alternatively flashed. His third law states that when the distance between the lamps increases, it is necessary to slow down the alternation of flashes to keep a strong impression of motion. An experiment implying the switch at a sufficient speed of the lights of a device like the one depicted in **figure 1.12a**, according for example to the pattern 142536, should show an irregular motion between members of the two separate sets of lamps. But as the speed increases, the motion will appear to split into two separate streams, one occuring in the right triplet (123), and the other in the left one (456). This phenomenon occurs because the distance between lamps of each triplet is too great for a move between triplets to be plausible, as predicted by Körte's law. We get exactly the same phenomenon of streaming in audition, when listening

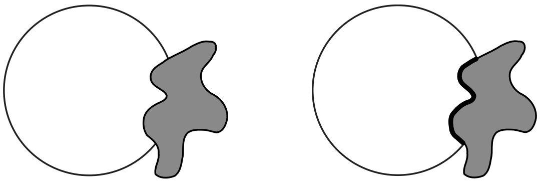


Figure 1.13: An example of belongingness: the dark portion of the line shown on the right figure seems, on the left figure, to belong to the irregular form. (After [Bre94].)

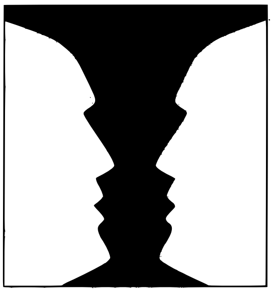


Figure 1.14: An example of exclusive allocation of evidence: a vase at the center or two faces at the sides can be seen, depending on the choice of allocation of the separating edges. (After [Bre94].)

at high speed to the looping sequence presented in **figure 1.12b**: the heard pattern is not 361425 as it is at a lower speed, but is divided into two streams, 312 (corresponding to the low tones) and 645 (corresponding to the high tones). According to Körte's law, with melodic motion taking the place of spatial motion, the distance in frequency between the two groups of tone is too great regarding the speed of movement between them.

1.3.3 The Principle of Exclusive Allocation

On the left side of **figure 1.13**, the part of the drawing at which the irregular form overlaps the circle (shown with a bold stroke on the right side of the figure) is generally seen as part of the irregular shape: it *belongs* to the irregular form. It can be seen as part of the circle with an effort. Be that as it may, the principle of "belongingness," introduced by the Gestalt psychologists, designates the fact that a property is always a property *of* something.

This principle is linked to that of "exclusive allocation of evidence," illustrated in **figure 1.14**, which states that a sensory element should not be used in more than one description at a time. Thus, on the figure, we can see the separating edges as exclusively allocated to the vase at the center, or to the faces at the sides, but never to both of them. So, this second principle corresponds to the belongingness one, with an unique allocation at a time.

These principles of vision can be applied to audition as well, as shown by the experiment by Bregman and Rudnicky [BR75] illustrated in **figure 1.15**. The task of the subject was to determine the order of the two target tones A and B: high-low or low-high. When the pattern AB is presented alone, the subject easily finds the correct order. But, when the two tones F (for "flankers") are added, such that we get the pattern FABF, subjects have difficulty hearing the order of A and B, because they are now part of an auditory stream. However, it is possible to assign the F tones to a different perceptual stream than to that of the A and B tones, by adding a third group of tones, labeled C for "captors."

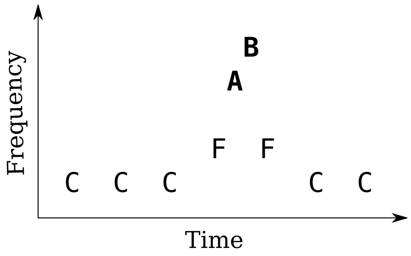


Figure 1.15: The tone sequence used by Bregman and Rudnicky to underline the exclusive allocation of evidence. (Reprinted from [Bre94].)

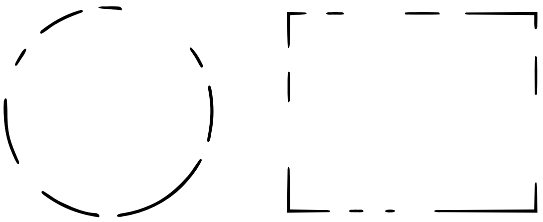


Figure 1.16: An example of the closure phenomenon. Shapes are strong enough to complete evidences with gaps in them. (After [Fou11].)

When the C tones were close to the F tones in frequency, the latter were captured[3] into a stream CCCFFCC by the former, and the order of AB was clearer than when C tones were much lower than F tones. Thus, when the belongingness of the F tones is switched, the perceived auditory streams are changed.

1.3.4 The Phenomenon of Closure

Also proposed by the Gestalt psychologists, the phenomenon of closure represents the tendency to close certain "strong" perceptual forms such as circles or squares, by completing evidences with gaps in them. Examples can be seen on the left of **figure 1.13** and in **figure 1.16**.

However, when the forces of closures are not strong enough, as shown in **figure 1.17**, the presence of the mask could be necessary to provide us informations about which spaces have been occluded, giving us the ability to discriminate the contours that have been produced by the shape of the fragments themselves from those that have been produced by the shape of the mask that is covering them. This phenomenon is called the phenomenon of "perceived continuity," and has an equivalent in audition. **Figure 1.18** presents an experiment where an alternatively rising and falling pure-tone glide is periodically interrupted. In this case, several short rising and falling glides are heared. But in the presence of a loud burst of broad-band noise exactly matching the silences, a single continuous sound is heard. Note that to be successful, the interrupting noise must be loud enough and have the right frequency content, corresponding to the interrupted portion of the glide. This is also an illustration of the old-plus-new heuristic (see section 1.3.1).

[3] This capture is reinforced by the regular rhythmic pattern of the F and C tones.

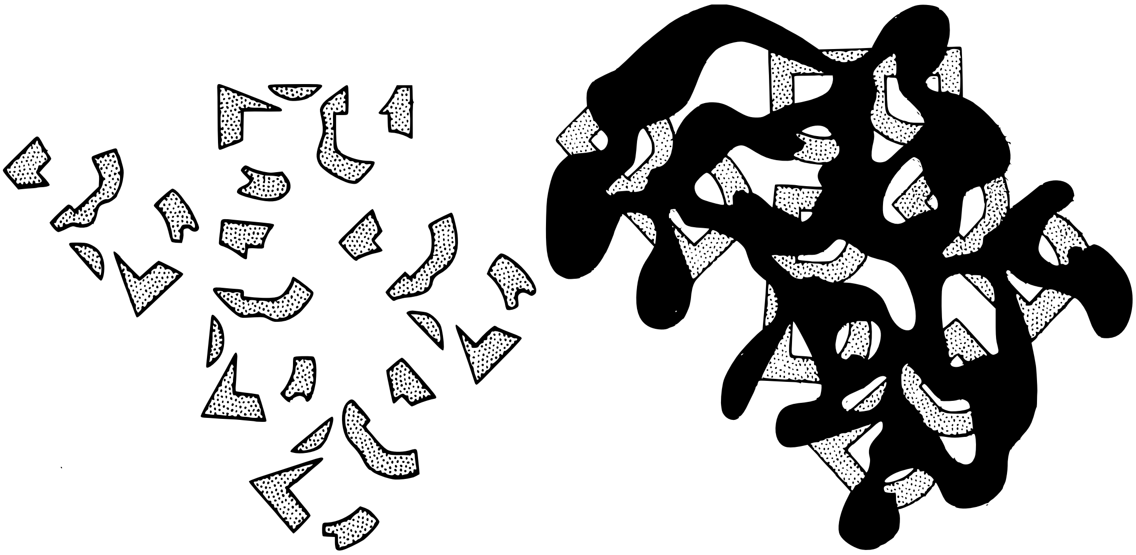


Figure 1.17: Objects occluded by a masker. On the left, fragments are not in good continuation with one another, but with the presence of the masker, on the right, we get informations about occlusion, and then fragments are grouped into objects. (After [Bre94].)

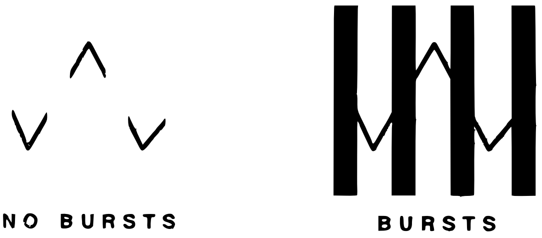


Figure 1.18: The illusion of continuity. (After [Bre94].)

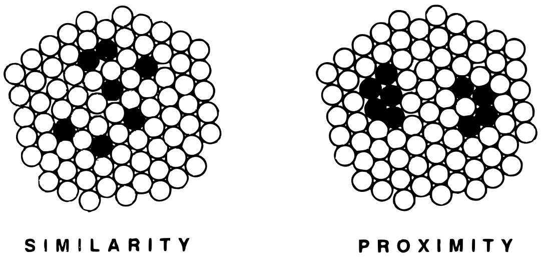


Figure 1.19: Two process of perceptual organization. (After [Bre94].)

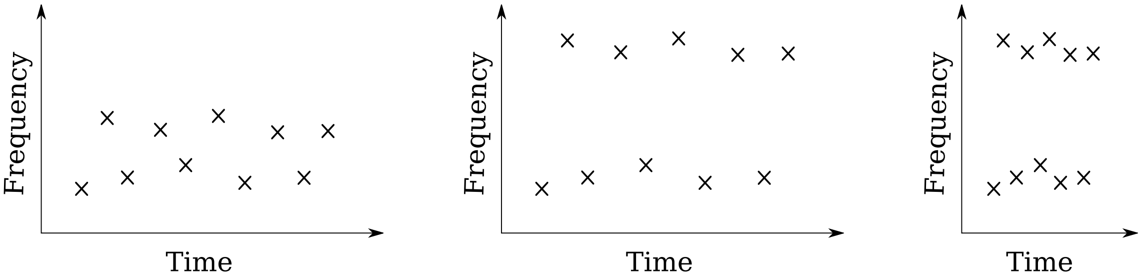


Figure 1.20: Stream segregation by proximity in frequency and time. The segregation is poor for the left sequence, greater for the middle one, and greatest for the right one. (Reprinted from [Bre94].)

1.3.5 Forces of Attraction

Let's make another analogy with vision. In **figure 1.19**, two processes of perceptual organization are highlighted. The first one (shown on the left side of the figure) concerns the similarity grouping: because of the similarity of color, and thus of the contrast between the black and white blobs, two clusters appear, as in audition when sounds of similar timbre group together.

The second process of perceptual organization is about grouping by proximity and is shown on the right side of the figure. Here, the black blobs fall into two separate clusters, because each member of one cluster is closer to its other members than to those of the other one. This Gestalt law has a direct analogy in audition. In **figure 1.20**, an experiment is illustrated, in which two sets of tones, one high and the other low in frequency, are shuffled together. As visually when looking at the figures, the listening of the third one (on the right) will show greater perceptual segregation than the second, and the second than the first.

Thereby, forces of attraction are applying to perceptually group objects together, the most important being the time and frequency proximity in audition (corresponding to distance in vision). Note that indeed, only one factor—time—is really implied in sound proximity, since frequency is highly related to time. So, two kinds of perceptual grouping coexist: one "horizontal," the sequential grouping, related to time and melody, and one "vertical," the simultaneous grouping, related to frequency and harmony ; and these grouping factors interact between them since time is implied in both. But what if these forces are contrary? An experiment [BP78] by Bregman and Pinker discuss this and is displayed in **figure 1.21**. It consist of a repeating cycle formed by three pure tones A, B and C arranged in such a way that A and B tones frequencies are grossly in the same area, as well as B and C are roughly synchronous. The experiment showed that it was possible

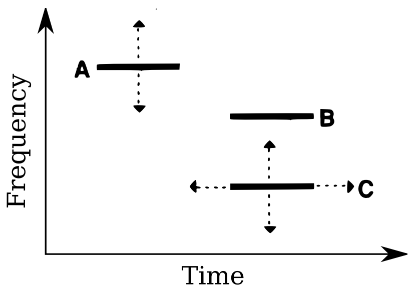


Figure 1.21: Forces of attraction competiting in an experiment by Bregman and Pinker. (After [Bre94].)

to hear the sequence in two different ways. In the first one, A and B tones are streamed together, depending on their proximity in frequency. And as a second way, B and C tone are fused in a complex sound if their synchrony is sufficient. It was as if A and C were competing to see which one would get to group with B.

Finally, our perception system tries to integrate these grouping laws in order to build a description of the scene. Though, the built of this description is not always totally right, as shown by an illusion set up by Diana Deutsch [Deu74a]. The listener is presented with a continuously repeating alternation of two events. The first event is a low tone presented to the left ear synchronously with a high tone (one octave above) to the right ear. The second event is the reverse: left/high and right/low. However, many listeners described another experience. They heard a single sound bouncing back and forth between the ears, and alternating between high and low pitch. The explanation comes from the fact that, assuming the existence of a single tone, the listeners derived of it two different descriptions from two different types of perceptual analyzes, and put them together in a wrong way.

1.4 Spatial Hearing

The auditory system is able, even when sight is not available, to derive more or less precisely a position of sound sources in three dimensions, thanks to its two auditory sensors, the ears. A spherical coordinate system is useful to represent each sound source of the auditory scene with three coordinates relative to the center of the listener's head: the azimuth, the elevation, and the distance, thereby defining the three dimensions of sound localization. Localization in azimuth (see section 1.4.1) is mainly attributed to a binaural processing of cues, based on the integration of time and intensity differences between ear inputs, whereas localization in elevation (see section 1.4.2) is explained by the use of monaural cues. Although, monaural cues play a role as well in localization in azimuth. Localization in distance (see section 1.4.3) is more related to characteristics of the sources, like spectral content and coherence.

1.4.1 Localization in Azimuth

The duplex theory

As illustrated in **figure 1.22**, from a physical point of view, if one considers a single monochromatic sound source, the incident sound wave will directly reach the closest ear (the ipsilateral ear). Before reaching the other ear (the contralateral ear), the head of the listener constitutes an obstacle to the wave propagation, and depending on its wavelength,

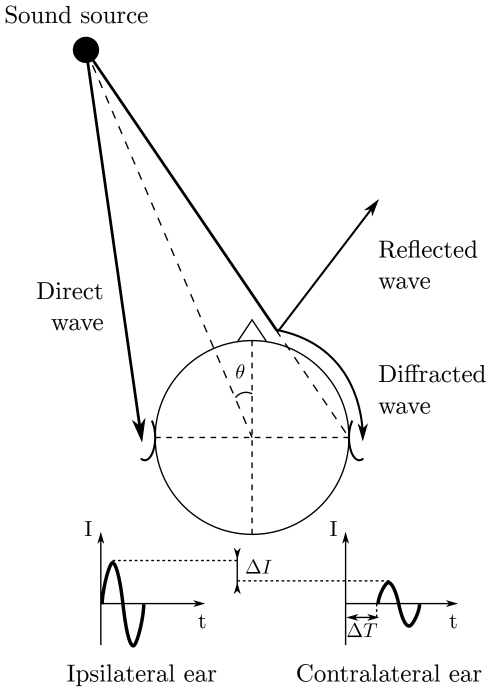


Figure 1.22: Interaural time and intensity differences for a monochromatic sound source.

the wave is subject to be partly diffracted and partly reflected by the head of the listener. The greater distance of the contralateral ear from the sound source, in conjunction with the diffraction of the wave by the head, induces a delay between the time of arrival of the wave to each ear, namely an *interaural time difference* (ITD). The reflection by the head attenuates the wave before reaching the contralateral ear, resulting in an *interaural intensity difference* (IID), also known as *interaural level difference* (ILD). The duplex theory, proposed by Lord Rayleigh [Ray07] in 1907, states that our lateralization ability (localization along only one dimension, the interaural axis) is actually based on the integration of these interaural differences. It has been confirmed by more recent studies [Bla97] that indeed ITDs and ILDs are used as cues to derive the position of sound sources in azimuth.

Neural processing of interaural differences

This section aims to bring a physiological justification of the binaural cues ITD and ILD. In the continuity of section 1.1 describing the ear physiology, the inner hair cells of the organ of Corti convert the motions occuring along the basilar membrane into electrical impulses which are transmitted to the brain through the auditory (or cochlear) nerve. Hence, each fiber in the nerve is related to a particular band of frequencies from the cochlea and has a particular temporal structure depending on impulses through the fiber. When phase-locking is effective (that is for low frequencies [PR86]), discharges through the fiber occur within a well-defined time window relative to a single period of the sinusoid. The signal coming from the cochlea passes through several relays in the auditory brainstem (see **figure 1.23**) before reaching the auditory cortex. At each relay, the initial tonotopic coding from the cochlea is projected, as certain neurons respond principally to components close to their best frequency. Note that the two parts (left ear and right ear) of this brainstem are interconnected, allowing for binaural processing of information. The center that interests us is the *superior olivary complex* (SOC). In most mammals, two major

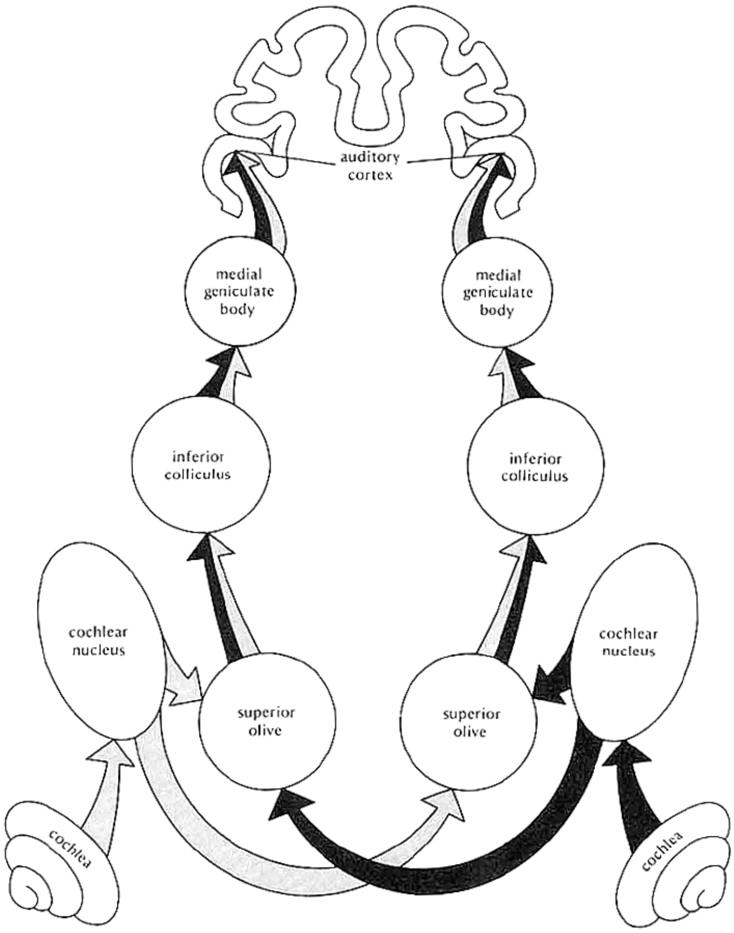


Figure 1.23: A highly schematic view of the auditory brainstem. Only basic contralateral connections are represented. (After [Yos94].)

types of binaural neurons are found within this complex.

In 1948, Jeffress [Jef48] proposed a model of ITD processing which is consistent with more recent studies [JSY98, MJP01]. A nucleus of the SOC, the *medial superior olive* (MSO), hosts cells designated as *excitatory-excitatory* (EE) because they receive excitatory input from the *cochlear nucleus* (CN) of both sides. An axon from one CN and an axon from the contralateral CN are then connected to an EE cell. An EE cell is a "coincidence detector" neuron: its response is maximum for simultaneous inputs. Each axon from a CN having its own conduction time, this CN-EE-CN triplet is sensitive to a particular ITD, and the whole set of such triplets finally acts as a cross-correlator. Consequently, phase-locking is an essential prerequisite for this process to be effective.

The second type of binaural neuron [Tol03] is a subgroup of cells of a nucleus of the SOC, the *lateral superior olive* (LSO), which are excited by the signals from one ear and inhibited by the signals from the other ear, and thus are designated as *excitation-inhibition* (EI) type. To do so, the signal coming from the contralateral CN is presented to the *medial nucleus of the trapezoid body* (MNTB), another nucleus of the SOC, which makes it inhibitory and presents it to the LSO. Also, the LSO receives an excitatory signal from the ipsilateral CN, and thus acts as a subtractor of patterns from the two ears. The opposite influence of the two ears makes these cells sensitive to ILD. It is also believed that cells from the LSO are involved in the extraction of ITDs by envelope coding [JY95].

The processing from MSO and LSO is then transmitted to the *inferior colliculus* (IC), where further processing takes place before transmission to the *thalamus* and the *auditory cortex* (AC).

Validity of interaural differences

Depending on frequency, ITD and ILD cues will be more or less exploitable by the auditory system. At low frequencies, where the wavelength is important compared to the head radius, the sound wave is reflected to a negligible degree, and thus the ILD is almost nil. [4] As the wave frequency increases, the wavelength gets smaller with respect to the head radius, and the reflected part of the sound wave increases, until being completely reflected at high frequencies. By definition, the ILD equals:

$$ILD = 10 \log_{10} \frac{|p_L|^2}{|p_R|^2}, \tag{1.6}$$

where p_L and p_R are respectively the acoustic pressure on the left and right eardrum.

At low frequencies, the ITD can be equally described as an *interaural phase difference* (IPD), approximated by [Kuh77]:

$$IPD_{\mathrm{lo}} = 3ka \sin\theta, \tag{1.7}$$

with $k = 2\pi/\lambda$ being the wave number, a the head radius (modeled as a rigid sphere), λ the wavelength, c the sound speed in air, and θ the source azimuth. Under the assumption that $(ka)^2 \ll 1$, the IPD can be expressed as an ITD independent of frequency:

$$ITD_{\mathrm{lo}} = \frac{IPD_{\mathrm{lo}}}{\omega} = 3\frac{a}{c}\sin\theta, \tag{1.8}$$

[4] However, in the proximal region (i.e., within one meter of the listener's head), because the sound wave can no longer be treated as planar but as spherical, and thus because of the inverse relationship of sound pressure and distance, the difference distance between each ear and the source implies a significant difference in pressure between the ears, even if no head shadowing occurs [SSK00].

where ω is the angular velocity. Above 1500 Hz, the wavelengths fall below the interaural distance, which is about 23 cm, and thus delays between the ears can exceed a period of the wave and become ambiguous. Moreover, the phase-locking ability of the auditory system (that is its ability to encode the phase information in the auditory nerve and neurons) decreases with increasing stimulus frequency [PR86], and is limited to frequencies below 1.3 kHz [ZF56]. However, it has been shown [MP76] that when listening to a signal at one ear, and to an envelope-shifted version of it at the other ear, the ITD is still effective, even if the carrier is above 1500 Hz (but this ability decreases for frequencies above 4 kHz [MG91]). In that case, the ITD is well described by Woodworth's model [WS54], which is independent of frequency:

$$ITD_{\mathrm{hi}} = \frac{a}{c}[sin(\theta) + \theta]. \tag{1.9}$$

So, two mechanisms are involved in the integration of the ITD: phase delay at low frequencies, and envelope delay at higher frequencies. Note that formulas (1.8) and (1.9) are assuming a spherical head (which is actually more oval) and ears at ±90° (which are actually a few degrees backward). Moreover, these models do not depend on elevation, which implies that cones of constant azimuth share the same ITD value. A more realistic model has been designed by Busson [Bus06].

The formulas above for ILD and ITD are physical descriptions of interaural cues, but the way these cues are integrated by the auditory system is still unclear. It is generally accepted given the neural processing of interaural differences, that ITD and ILD are processed in narrow bands by the brain before being combined with information from other modalities (vision, vestibules, etc.) to derive the position of the identified auditory objects (see section 1.5). Besides, additional results support a frequency-specific encoding of sound locations (see section 1.5.2). Such processing in narrow bands ensures the ability to localize concurrent sound sources with different spectra. Therefore, the integration of interaural differences is often simulated by processing the two input channels through filter banks and by deriving interaural differences within each pair of sub-bands [BvdPKS05].

Finally, the ITD and the ILD are complementary: in most cases, at low frequencies (below 1000 Hz), the ITD gives most informations about lateralization, and roughly above this threshold, the ITD becoming ambiguous, the ILD reliability is increasing, to take the lead above about 1500 kHz. Using synthetic and conflicting ITD and ILD, Wightman and Kistler [WK92] showed that the ITD is prominent for the low-frequency lateralization of a wide-band sound. However, for a sound without any low frequencies (below 2500 Hz), the ILD is prevailing. Gaik showed [Gai93] that conflicting cues induce artifacts of localization and modifications of the perception of tone color.

Limitations of ITD and IID

Assuming the simple models of ILD and ITD described above, these two cues do not depend on frequency, and especially they do not depend on elevation either. Hence, particular loci, called "cones of confusion", were introduced by Woodworth; they are centered on the interaural axis and correspond to an infinite number of positions for which the ITD and ILD are constant (see **figure 1.24**). Actually, these cones do not stricly make sense, and would rather in reality correspond to a set of points of equal ITD/ILD pair. Indeed, ITD and ILD are more complex than simple models, and measured iso-ITD or iso-ILD curves are not strictly cone-shaped (see them on **figure 1.24b**). Anyhow, the necessary threshold to detect small variations of position (see section 1.4.5) increases the number of points of equal ITD/ILD pair. The term "cones of confusion" can also refer to a single binaural cue, ITD *or* ILD, considering the iso-ITD *or* the iso-ILD curves only.

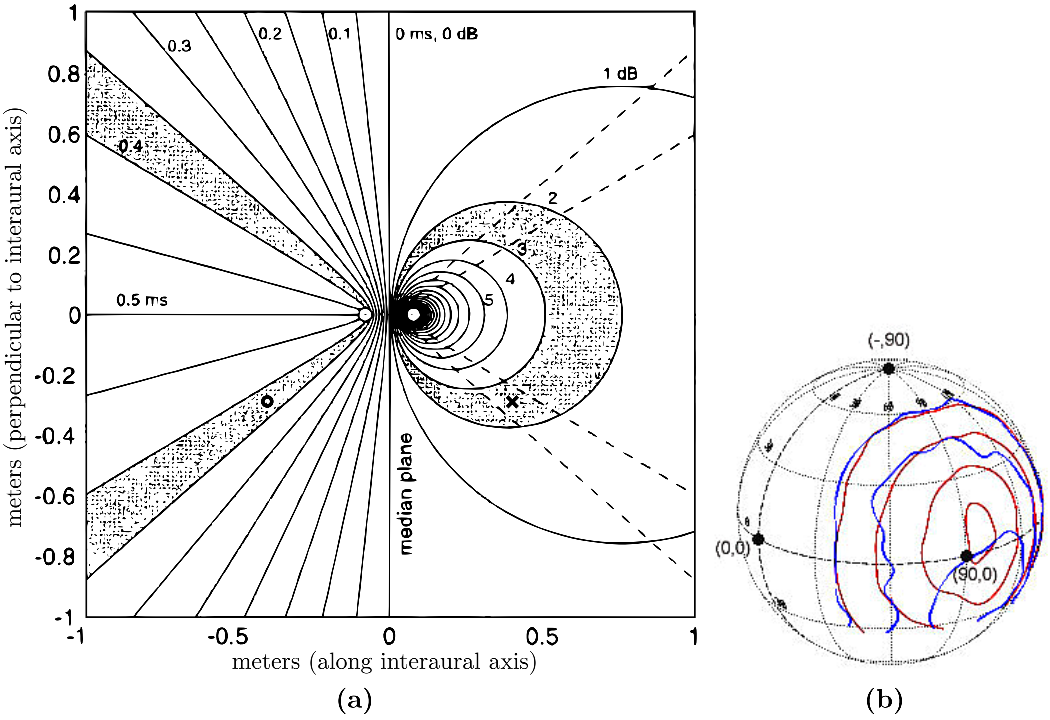


Figure 1.24: **(a)** Iso-ITD (left side) and iso-IID (right side) contours in the horizontal plane, in the proximal region of space. In the distal region, however, iso-ITD and iso-IID surfaces are similar. (After [SSK00].) **(b)** Superimposition of measured iso-ITD (in red, quite regular and roughly concentric, steps of 150 μs) and iso-ILD (in blue, irregular, steps of 10 dB) curves in the distal region. (Reprinted from [WK99].)

Thus, the duplex theory does not explain our discrimination ability along these "cones of confusion," which implies a localization in elevation and in distance (also with an extracranial perception). The asymetrical character of the ILD could be a front/back discrimination cue. However that may be, these limitations suggested the existence of other localization cues, the monaural spectral cues, which are discussed in the next section. It has also been shown that monaural cues intervene in localization in azimuth for some congenitally monaural listeners [SIM94], and thus might be used as well by normal listeners. It is especially believed that monaural cues are used to reduce front/back confusions.

1.4.2 Localization in Elevation

So far, the presented localization cues, based on interaural differences, were not sufficient to explain the discrimination along cones of confusion. Monaural cues (or spectral cues) put forward an explanation based on the filtering of the sound wave of a source, due to reflections and diffractions by the torso, the shoulders, the head and the pinnae before reaching the tympanic membrane. The resulting colorations for each ear of the source spectra, depending on both direction and frequency, could be a localization cue.

Assuming that $x_L(t)$ and $x_R(t)$ are the signals of the left and right auditory canal inputs of a $x(t)$ source signal, this filtering can be modeled as:

$$x_L(t) = h_L * x(t), \text{ and } x_R(t) = h_R * x(t), \tag{1.10}$$

where h_L and h_R designate the impulse responses of the wave propagation from the source to the left and right auditory canals, and thus the previously mentioned filtering phenomena. Because of the direction-independent transfert functions from the auditory canals to the eardrums, these are not included in h_L and h_R. The equivalent frequency domain filtering model is given by:

$$X_L(f) = H_L \cdot X(f), \text{ and } X_R(f) = H_R \cdot X(f). \tag{1.11}$$

The H_L and H_R filters are called *head-related transfer functions* (HRTF), whereas h_L and h_R are called *head-related impulse responses* (HRIR). The postulate behind localization in elevation is that this filtering induces peaks and valleys in X_L and X_R (the resulting spectra of x_L and x_R) varying with the direction of the source as a "shape signature", especially in high frequencies [LB02]. The auditory system would first learn these shape signatures, and then use this knowledge to associate a recognized shape with its corresponding direction (especially in elevation). Consequently, this localization cue requires a certain familiarity with the original source to be efficient, especially in the cases of static sources with no head movements. In the case of remaining confusions of source position in cones of constant azimuth, due to similar spectral contents, the ambiguity can be solved by left/right and up/down head movements [WK99]. Also, these movements improve localization performance [Wal40, TMR67]. Actually, a slow motion of the source in space is sufficient to increase localization performance, implying that the knowledge of the relative movements is not necessary.

1.4.3 Localization in Distance

The last point in localization deals with the remaining coordinate of the spherical system: the distance. The available localization cues are not very reliable, which is why our perception of distance is quite imprecise [Zah02]. Four cues are involved in distance perception [Rum01]. First, the perceived proximity increases with the source sound level.

Second, the direct field to reverberated field energy ratio gets high values for closer sources, and this ratio is assessed by the auditory system through the degree of coherence between the signals at the two ears. Third, the high frequencies are attenuated with air absorption, thus distant sources have less high frequency content. And finally, further away sources have less difference between arrival of direct sound and floor first reflections.

1.4.4 Apparent Source Width

The apparent source width (ASW) has been studied for the acoustics of concert halls and deals with how large a space a source appears to occupy from a sonic point of view. It is related to *interaural coherence* (IC) for binaural listening or to *inter-channel coherence* (ICC) for multichannel reproduction, which are defined as the maximum absolute value of the normalized cross-correlation between the left (x_l) and right (x_r) signals:

$$IC = \max_{\Delta t} \frac{\left|\sum_{n=-\infty}^{n=\infty} x_l[n] \cdot x_r[n+\Delta t]\right|}{\sqrt{\sum_{n=-\infty}^{n=\infty} x_l^2[n] \cdot \sum_{n=-\infty}^{n=\infty} x_r^2[n+\Delta t]}}, \tag{1.12}$$

When $IC = 1$, the signals are coherent, but may have a phase difference (ITD) or an intensity difference (ILD), and when $IC = 0$ the signals are independent. Blauert [Bla97] studied the ASW phenomenon with white noises and concluded that when $IC = 1$, the ASW is reduced and confined to the median axis; when IC is decreasing, the apparent width increases until the source splits up into two distinct sources for $IC = 0$.

1.4.5 Localization Performance

The estimation by the auditory system of sound source attributes (as loudness, pitch, spatial position, etc.) may differ to a greater or lesser extent from the real characteristics of the source. This is why one usually differentiates *sound events* (physical sound sources) from *auditory events* (sound sources as perceived by the listener) [Bla97]. Note that a one-to-one mapping does not necessarily exist between sound events and auditory events. The association between sound and auditory events is of particular interest in what is called *auditory scene analysis* (ASA, see section 1.3) [Bre94].

Localization performance covers two aspects. *Localization error* is the difference in position between a sound event and its (supposedly) associated auditory event, that is to say the accuracy with which the spatial position of a sound source is estimated by the auditory system. *Localization blur* is the smallest change in position of a sound event that leads to a change in position of the auditory event, and thereby is a measure of sensitivity. It reflects the extent to which the auditory system is able to spatially discriminate two positions of the same sound event, that is the auditory spatial resolution. When it characterizes the sensitivity to an angular displacement (either in azimuth or in elevation), the localization blur is sometimes expressed as a *minimum audible angle* (MAA). MAAs in azimuth constitute the main topic of this thesis and are especially studied in chapter 3. We will see in the following that both localization error and localization blur mainly depend on two parameters characterizing the sound event: its position and its spectral content.

Localization errors have been historically studied by Lord Rayleigh [Ray07] using vibrating tuning forks, after which several studies followed. Concerning localization in azimuth, studies [Pre66, HS70] depicted in **figure 1.25** using white noise pulses have shown that localization error is the smallest in the front and back (about 1°), and much greater for lateral sound sources (about 10°). Carlile *et al.* [CLH97] reported similar trends using broadband noise bursts. According to Oldfield and Parker [OP84], localization error in

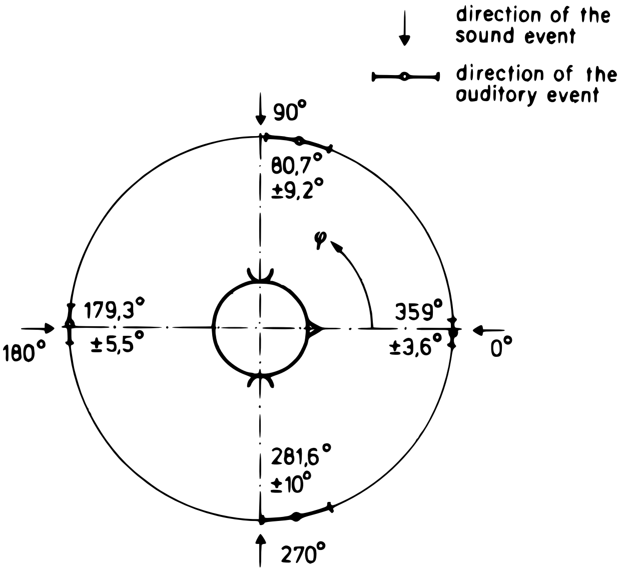


Figure 1.25: Localization error and localization blur in the horizontal plane with white noise pulses [Pre66, HS70]. (After [Bla97].)

azimuth is almost independent of elevation. Localization blur in azimuth follows the same trend as localization error (and is also depicted in **figure 1.25**), but is slightly worse in the back (5.5°) compared to frontal sources (3.6°), and reaches 10° for sources on the sides. Perrott [PS90], using click trains, got a smaller mean localization blur of 0.97°. The frequency dependence of localization error in azimuth can be found for example in [SN36], where maximum localization errors were found around 3 kHz using pure tones, these values declining for lower and higher frequencies. This type of shape for localization performance in azimuth as a function of frequency, showing the largest values at mid frequencies, is characteristic for both localization error and localization blur and is usually interpreted as an argument supporting the duplex theory. Indeed, in that frequency range (1.5 kHz to 3 kHz), the frequency is too high for phase locking to be effective (which is necessary for ITD), and the wavelength is too long for head shadowing to be efficient (which is necessary for ILD), thereby reducing the available localization cues [MG91]. The frequency dependence of localization blur has been studied by Mills [Mil58] (see **figure 1.26**) and varies between 1° and 3° for frontal sources. Boerger [Boe65] got similar results using Gaussian tone bursts of critical bandwidth. Important front/back confusions for pure tones, compared to broadband stimuli, are reported in [SN36]. Moreover, the study from Carlile *et al.* [CLH97] confirms that only a few front/back confusions are found with broadband noise. This is a characteristic result concerning narrow band signals, given that monaural cues are poor in such cases and cannot help discriminate the front from the back by signal filtering.

Concerning localization error in elevation, several studies report worse performance than in azimuth. Carlile *et al.* [CLH97] reported 4° on average with broadband noise bursts. Damaske and Wagener [DW69], using continuous familiar speech, reported localization error and localization blur in the median plane that increases with elevation (see **figure 1.27**). Oldfield and Parker [OP84], however, announce an error independent of the elevation. Blauert [Bla70] reported a localization blur of about 17° for forward sources, which is much more than Damaske and Wagener's estimation (9°), but using unfamiliar speech. This supports the idea that a certain familiarity with the source is necessary to

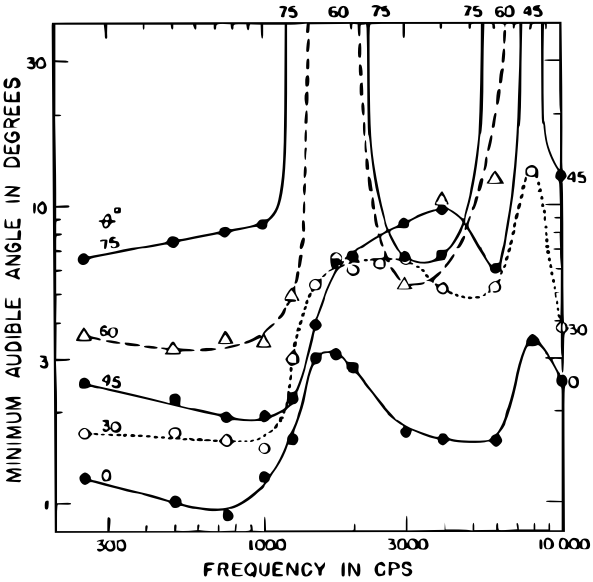


Figure 1.26: Frequency dependence of localization blur in azimuth (expressed here as a "minimum audible angle") using pure tones, as a function of the sound source azimuth position θ. (After [Mil58].)

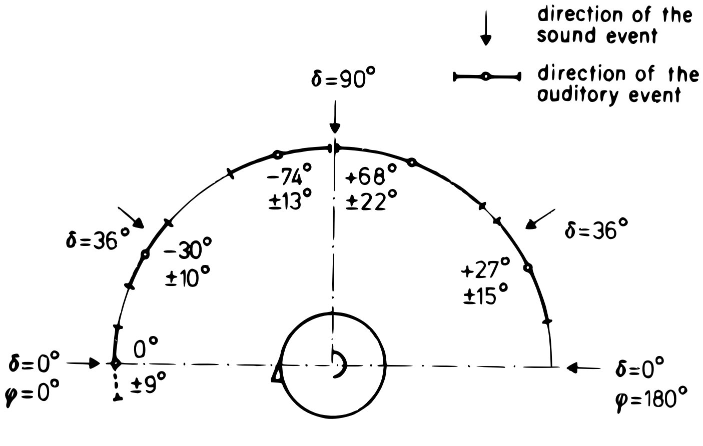


Figure 1.27: Localization error and localization blur in the median plane using familiar speech [DW69]. (After [Bla97].)

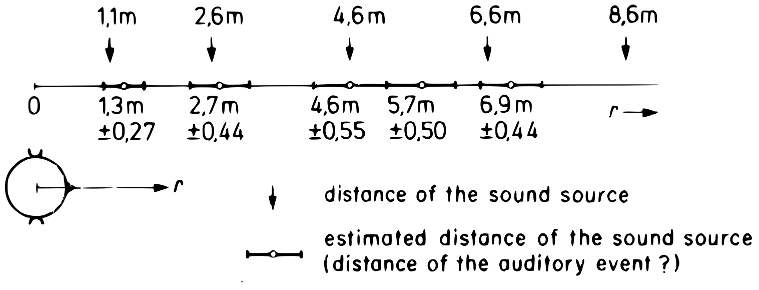

1,1m
2,6m
4,6m
6,6m
8,6m
0
1,3m ±0,27
2,7m ±0,44
4,6m ±0,55
5,7m ±0,50
6,9m ±0,44
r
r
distance of the sound source
estimated distance of the sound source (distance of the auditory event ?)

Figure 1.28: Localization error and localization blur for distances using impulsive sounds [Hau69]. (After [Bla97].)

obtain optimal localization performance in elevation. Perrott [PS90], using click trains, got a mean localization blur of 3.65°. He also performed measures of localization blur for oblique planes, and reported values below 1.24° as long as the plane is rotated more than 10° away from the vertical plane. Grantham [GHE03], on the contrary, reported a higher localization blur for a 60° oblique plane than for the horizontal plane, but still a lower blur than for the vertical plane. Blauert [Bla68] brought to light an interesting phenomenon concerning localization in the median plane of narrow-band signals (bandwidth of less than 2/3 octave): the direction of the auditory event does not depend on the direction of the sound event, but only on the frequency of the signal. Once again, this is justified by the fact that monaural cues are non-existent for such narrow-band signals.

Finally, studies dealing with localization in distance suggest that performance depends on the familiarity of the subject with the signal. Gardner [Gar69] studied localization in the range of distance from 0.9 to 9 m with a human speaker whispering, speaking normally, and calling out loudly. For normal speaking, performance is excellent, whereas distances for whispering and calling out voices are under- and an over-estimated, respectively. Good performance for the same range of distances was reported by Haustein [Hau69] using impulsive sounds (see **figure 1.28**), but using test signals that were demonstrated beforehand from different distances. On the contrary, Zahorik [Zah02] reports a compression phenomenon: sources closer than one meter are over-estimated, whereas far-away sources are under-estimated.

Just-Noticeable Differences

Beside the sensitivity to a physical displacement of a sound source, which is characterized by the notion of localization blur, some research has studied the sensitivity to the cues underlying the localization process, namely ITD, ILD, IC, and spectral cues. In this section, we will focus on results concerning the sensitivity to changes in the binaural cues only (ITD, ILD and IC). This can be measured by manipulating localization cues to generate a pair of signals (specific to each ear) and playing them over headphones to a listener. For instance, an artificial ITD can be introduced between the two signals to test the smallest variation of ITD a listener is able to detect, i.e., the *just-noticeable difference* (JND) of ITD. The sensitivity to a given cue can potentially depend on the following main parameters: the reference value of this cue (the initial value from which the sensitivity to a slight variation is tested), the frequency (content) of the stimulus, the level of the stimulus (apart from the potential presence of an ILD), and the actual values of the other localization cues.

For stimuli with frequencies below 1.3 kHz, the ITD can be described as a phase difference (IPD, see section 1.4), and for a null reference IPD, the JND of IPD does not depend on frequency and equals about 0.05 rad [KE56]. This sensitivity tends to increase as the reference IPD increases [Yos74, HD69]. There does not seem to be an effect of the stimulus level on the JND of ITD [ZF56], although there is a decrease in sensitivity for very low levels [HD69]. Finally, the sensitivity to a change of ITD decreases with an increasing ILD [HD69].

For a null reference ILD, the JND of ILD is relatively independent of stimulus level [Yos72] (except again for very low levels [HD69]) and of stimulus frequency [Gra84]. As the reference ILD increases, the sensitivity to a change of ILD decreases [Mil60, RT67, YH87]: from between 0.5 and 1 dB for a reference ILD of 0 dB, to between 1.5 and 2 dB for a reference ILD of 15 dB. The dependence of ILD sensitivity on the reference ITD is not clear, since Yost [Yos72] reported that this sensitivity increases when the ITD increases, whereas no dependence is reported in [HD69].

Concerning the JND of IC, a strong dependence on the reference IC has been shown [RJ63, LJ64, GC81, CCS01]: from 0.002 for a reference IC of +1, to about a 100 times larger necessary variation for a reference IC of 0. The sensitivity to IC change does not depend on stimulus level [HH84], except for low levels.

1.4.6 Binaural Unmasking

Central masking

Two kinds of masking can happen when a target signal is mixed with competing sources. The first one, known as *energetic masking* (EM), occurs when a masker (for instance a noise) renders a target totally or partially inaudible, as presented in section 1.2.2. It results from the competition between the target and the masker in the auditory periphery, i.e., overlapping excitation patterns in the cochlea or auditory nerve, which corresponds to *peripheral masking* (PM). This situation must be differentiated from *central masking* (CM; also called *nonenergetic masking*), which takes place more centrally in the auditory pathway (see **figure 1.23**). Most central masking consists of *informational masking* (IM),[5] which is caused by two main factors: target-masker similarity and stimulus (masker) uncertainty. In such a situation, although the target is audible, it is confused with the masker, leading, as in the case of energetic masking, to a decrease in the threshold of detectability of the target. Target-masker similarity is encountered when target and masker[6] share a number of similar characteristics, such as spectro-temporal similarities (a typical example being speech on speech), even without frequency overlap, such that EM is reduced. Those similar characteristics recall incidentally the grouping cues used to form auditory streams (see section 1.3). Stimulus uncertainty occurs when one or several characteristics (such as frequency, intensity,[7] or spatial location) of the masker vary rapidly in time to appear random[8] to the listener. As explained by Durlach in [DMKJ$^+$03], these main factors interact: stimulus uncertainty is neither necessary nor sufficient to produce non-energetic masking;

[5] The extent to which CM and IM are related is currently left open. In particular, questions arise about which central limitations that lead to threshold elevation should be included under CM, and which components of CM should be included in IM. This is discussed in [DMKJ$^+$03].

[6] In fact, some authors consider that what is really important is not the similarity between the masker M and the target T, but rather the similarity between M and M+T, because of the design of the task in a detection experiment (see [DMKJ$^+$03]).

[7] In this particular case, it is the overall level—of the masker *and* the target—which vary rapidly, in order to keep a constant TMR (target-to-mask ratio).

[8] In order to be effective, the range of the randomization must *not* be small compared to the listener's resolution.

the effects of uncertainty are reduced by decreasing target-masker similarity. Furthermore, Fan showed [FSD08] that compared both to the effects of spectral uncertainty and to the effects of overall-level uncertainty, the effects of spatial uncertainty are relatively small. Neff and Dethlefs [ND95] and Oxenham *et al.* [OFMKJ03] report that some listeners are resistant to the effects of uncertainty. In the case of frequency uncertainty, Gallun *et al.* [GDC^{+}08] report that some listeners, less affected by IM than others, *seem* to react as they do for EM. The proposed explanation of these results relies on a power-spectrum model of masking release (more details are given in [DMG^{+}05]). Briefly, it postulates that listeners who are adversely affected by the variability in the masking stimulus (i.e., its uncertainty) widen their effective auditory filters (larger than a critical band), which leads to an increase of its masking ability. This idea was first suggested by Lutfi [Lut93].

Binaural release from masking

In both situations (energetic and informational masking), listening performances can be improved when the sources arise from different spatial locations. This leads to "binaural release from masking," which is also called *spatial unmasking.* The change in the detection threshold of the target compared to when sources are presented with identical spatial positions is called *binaural masking level difference* (BMLD). As an example, a masking tone presented in phase to both ears with a pure-tone target presented out-of-phase to each ear simultaneously—this situation is called a NoSπ condition—results in a lower threshold level for detecting the target than in the case when both the masker and the target are presented in phase (a NoSo condition). This reflects spatial unmasking, because adding a phase difference to the target corresponds to the generation of a constant ITD, that is to say a perceptual spatial displacement of the target only. Generally, and as described in [Shi05], three distinct mechanisms are associated with spatial unmasking: acoustic better-ear effects, binaural processing, and spatial attention.

Better-ear effects are a direct consequence of ILD. When a source is to the side of the listener and contains significant energy above 2 kHz, sound reflection by the head increases the energy level at one ear. Thus, if the masker is spatially displaced from the target, the signal-to-mask ratio (SMR, see section 1.2.2) increases at one ear and decreases at the other, which leads to great improvements in the intelligibility of the target.

Binaural processing can improve the target audibility when this one remains masked despite the help provided by better-ear effects. As explained in [Shi05], when the target and the masker arise from the same direction, the interaural coherence does not change significantly when the target is present. However, in the case of different directions, the interaural coherence is decreased. Hence, the listener identifies this temporary decrease in correlation as being caused by an (otherwise inaudible) source at that time and frequency. Therefore, binaural processing allows listeners to detect the presence of target energy in a particular time and frequency band if the target and masker induce different interaural time and/or intensity differences.

The third mechanism associated with spatial unmasking is known as spatial attention, and has been largely studied in vision. It is based on the fact that listeners can orient their attention spatially to enhance the detection and identification of simple targets. Several authors report that when target and masker are spectro-temporally similar, perceived spatial separation contributes to spatial unmasking, but not when target and masker are dissimilar and easily segregated. For instance, Kidd and colleagues have conducted experiments with competing signals that have very little spectro-temporal overlap [AMKJ05, KJMBH05]. Because there is almost no energetic masking, one could predict little or no spatial unmasking, but it was actually prominent when target and masker were statistically similar and negligible when the masker was steady-state noise. Further, Best got similar results with spectro-temporally complex birdsongs [BOG+05]. Therefore, spatial attention seems to be efficient in cases of informational masking.

An interesting hypothesis combining the last two mechanisms would be that binaural processing actually creates binaural spatial channels, corresponding to different interaural configurations (ITD/ILD pairs). Then spatial attention would consist of nulling out the masker rather than focusing on the target. Although the results from Fan [FSD08] reported above conclude that the effect of spatial uncertainty in informational masking is weak, it is not negligible. That could lead to the conclusion that masked detection performance is still vulnerable to randomization of the spatial properties of the masker, which would be consistent with this hypothesis.

As explained by Durlach in [DMKJ+03], there could hypothetically be two extreme types of listeners, the *Listener-Max* corresponding to an archetypal analytic listener attempting to maximize the T/M ratio by maximizing T (with an acceptance filter focused on T), and *Listener-Min* corresponding to an archetypal holistic listener attempting to maximize the T/M ratio by minimizing M (with a multiple notch-rejection filter matched to M). With this distinction, when M is uncertain, *Listener-Max* should do best, whereas when T is uncertain, *Listener-Min* should do best.

But is the binaural release from energetic or informational masking due to a perceived difference in location or only to the presence of interaural difference cues (which could be then processed separately)? Gallun *et al.* in [GDC+08] report that for both types of masking, the obtained detection thresholds are similar in either the case of "reinforcing" ITDs and ILDs (one ear advanced in phase and higher in level) or in the case of "opposing" ITDs and ILDs (one ear advanced in phase and the other one higher in level), leading to an irrelevance of differences in the perceived location for both energetic and informational masking.

In the continuity of the model of auditory filter widening described above, Gallun *et al.* [GDC+08] also postulate that, in the case of informational masking, interaural differences reduce uncertainty and allow listeners to focus their effective filter more appropriately, the lower limit being the width of one critical band.

Kopčo and Shinn-Cunningham studied spatial unmasking in the case of nearby target and masker [KS03]. They evaluated the relative contribution of better-ear effects and binaural processing to spatial unmasking, showing a large contribution of better-ear effects. Interestingly, they noticed that there are cases where detection performance is actually worse when the sources are spatially separated compared to when they are at the same location, constituting a case of "spatial masking." This situation arises in particular when both the target and the masker are located at 90°, the masker being farther away than the target.

1.5 Localization Cues in Auditory Scene Analysis

This section concerns the integration of the spatial cues in the auditory scene analysis (ASA) process. As explained by Bregman in [Bre94] (see section 1.3), the auditory system, in its grouping process, seems to act as a voting system based on heuristic criteria. Spatial cues form part of theses criteria, but do not necessarily overpower other heuristics when in conflict with them—for instance, we are able to segregate different voices from a monophonic record.

The sequential and simultaneous integration of spatial cues are discussed in separate parts, but because these two kinds of integration interact between them, a third part deals

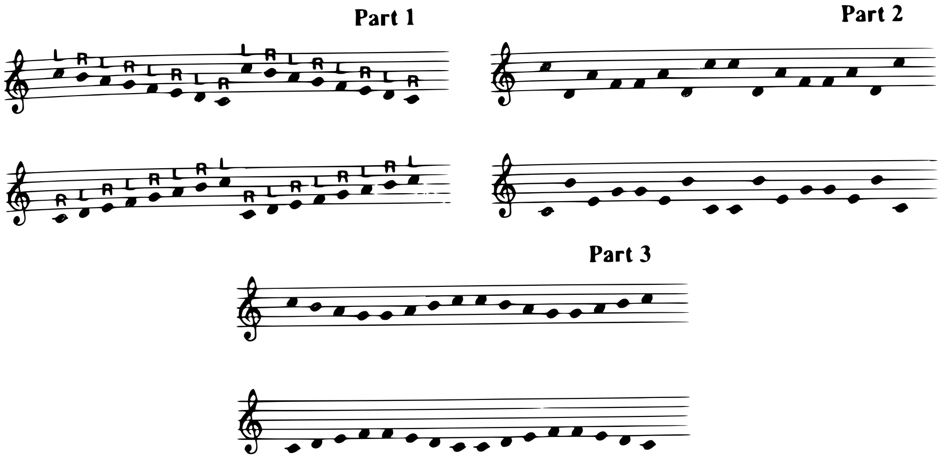


Figure 1.29: Illustration of the scale illusion from Deutsch [Deu74b]. Part 1: the stimulus, the letters show the ear of presentation. The tones of upper and lower staffs are presented simultaneously. Part 2: The stimulus resulting in each ear (upper staff for left ear, lower staff for right ear). Part 3: the sequences, as they are perceived by most of the subjects. Tones seem to have been grouped by frequency range. (After [Bre94].)

with this interaction. The last part treats the particular case of speech.

1.5.1 Sequential Integration

The sequential integration of spatial cues concerns how (much) the spatial location is involved in the streaming phenomenon.

Roger Shepard introduced the "psychophysical complementarity" principle in [She81]: because the auditory system can segregate sounds by their location, as space is physically a continuum (an object moving from one point to another must pass through all the intermediate positions), the perceptual representation of space must also have this property. Thus, if two identical sounds arise from sufficiently far-away locations in space without a continuous movement joining them, they would probably be treated as coming from different sources.

In two different experiments, Norman [Nor66] and van Noorden [vN75] reported a strong segregation by spatial origin. Norman observed that when a trill is presented with each of its alternate tones to a different ear, the continuous up-and-down movement of the pitch does not occur. Van Noorden did an introspective experiment and concluded that the temporal links between the tones in the two ears were weak. Using sequences with tones of identical frequency, he reported that judging if the tones sent to one ear were exactly halfway in time between the tones sent to the other ear was a hard task (depending on the tone rate), which is not the case with a diotic presentation. These experiments would lead to the conclusion that spatial separation is strong with respect to other grouping criteria.

Nevertheless, the proposals made by Norman and van Noorden have not been confirmed by later research. Indeed, criteria in conflict with spatial cues can lead to stream segregation, so that hearing of dichotically alternating tones as an integrated sequence is possible, as shown by the Deutsch's "scale illusion" [Deu74b]. The C major scale is presented to each ear of the subject over headphones in its ascending and descending form according to the score of part 1 of **figure 1.29**. The alternation between left and right ears results in two complex melodies illustrated in the second part of the figure. However,

Pattern A	Channel 1	... 1 - 4 - 1 - 4 - ...
	Channel 2	... - 2 - 3 - 2 - 3 ...
Pattern B	Channel 1	... 1 - 4 - 1 - 4 - ...
	Channel 2	... - 3 - 2 - 3 - 2 ...

Figure 1.30: The patterns presented by Judd to achieve a segregation by spatial location. Hyphens represent silences, and the numbers 1 to 4 represent four tones within an eight-semitone range.

most listeners actually perceive the pattern shown in the third part of the figure: the two complex patterns are turned into two much simpler patterns, meaning that the tones were grouped by their frequency range rather than by their ear of presentation. Bregman noticed that this illusion is effective even for rates of up to 20 tones per second, which means that its efficiency is *not* due to the slowness of the sequence (which otherwise would allow the listeners to switch their attention back and forth, and track the tonal sequence on the basis of frequency proximities).

Judd [Jud77] gave an explanation of the failure of spatial grouping to effectively compete with frequency proximity in Deutsch's scale illusion, which is based on the simultaneity of tones at the two ears. In an experiment, subjects must choose which of the two patterns in **figure 1.30** they heard over headphones. Because the sequences of channel 1 are identical in both cases, and the two sequences of channel 2 are the same but with a different starting point, the distinction of the two patterns requires channels 1 and 2 *not* to be segregated (by spatial location) in order to integrate the whole pattern and use channel 1 as a comparison point for channel 2. In such conditions, listeners had difficulty to distinguishing the two patterns. But Judd showed that by replacing the silences with a white noise with an amplitude equal to that of the tones, it was easier for listeners to distinguish the two patterns, thereby suggesting a reduction of the perceived ILD and thus a weakening of the spatial localization cues, thanks to the contralateral white noise. Judd also showed that by not playing the tones of the scale illusion simultaneously, but with a contralateral white noise, one can observe a segregation of streams according to the ear of arrival.

Later, Deutsch confirmed [Deu79] Judd's results with an identification task of eight-tone patterns in which the component tones switch between the ears. The task is very difficult as it is, but with a contralateral "drone" (like Judd's white noise), the melodies could be recognized more easily. Also, an experiment by Schubert and Parker [SP55] reports similar results with speech switching between the ears.

These results prove that different spatial locations of sounds can cause the auditory system to assign them to separate streams. Nevertheless, we saw that stream segregation by spatial location is not overpowering when in conflict with other bases for grouping. Bregman explains these unexpected observations by the fact that spatial cues are not really reliable for grouping because several sources can come from the same direction (but should not be grouped together), and also because sounds echoe, reverberate and wrap around obstructions, leading to unclear spatial origins of a sound. Evidence shows that location differences have a powerful multiplying effect when they are consistent with other bases for grouping (see [Bre94], ch. 7). Moreover, it should be noted that the ITD localization cue is not present in the illusions and experiments presented in this section, therefore leading to a weakening of the spatial basis for segregation.

Distortion of time

A surprising phenomenon has been related by Axelrod and Guzy [AG68]. Subjects are presented with series of rapid clicks uniformly spaced in time over headphones in two different forms: dichotic (clicks alternate between left- and right-ear), and diotic (each click is sent to both ears). With the dichotic presentation, the sequence is perceived as slower. Huggins [Hug74] did further experiments to assess the resulting distortion of time, by asking subjects to adjust the diotic sequence to make it sound as fast as the dichotic one. At slow rates (below eight tones per second), his subjects adjusted the diotic sequence to the same rate as the total event rate in the dichotic sequence. Then, for click rates from 8 to 20 per second, the diotic sequence was continuously slowed-down relative to the dichotic sequence. Finally, for higher rates, the diotic sequence was adjusted to one-half the dichotic sequence rate. Although for low and high rates, it is possible to argue that the dichotic sequence was either integrated as one single stream or segregated into two separate streams, respectively, the stream concept somewhat fails to explain the results with intermediate rates, all the more so as this distortion of time does *not* occur when using frequency grouping cues instead of spatial ones. Nevertheless, Bregman proposed the hypothesis that because perceptual organization tends to be unstable when the cues favoring segregation are not decisive, at intermediate rates, on some trials the clicks formed an integrated stream and on others they segregated into two streams.

1.5.2 Simultaneous Integration

Kubovy in [Kub81] proposed that in audition, as in vision, segregation of simultaneous signals relies on *indispensable attributes*. An indispensable attribute is a feature allowing the perception of the twoness of two simultaneous signals that are identical except for that feature. Kubovy states that indispensable auditory attributes are time and frequency, but not space or loudness. Whereas this is true for the loudness attribute, Bregman notes that concerning space, in cases of complex scenes, two sounds that differ only in spatial location can be heard separately and not fused together. He gives as an example the following experiment. Using an algorithm that simulates the spatial position of a source on headphones (binaural synthesis), he synthesized one pure tone at 600 Hz at $+45°$, and another one still at 600 Hz at $-45°$, 0° being in front of the listener. Each of the two tones was played at irregular intervals, leading to unsynchronised onsets and offsets and substantial overlaps in time. One could predict that during time overlaps, because of the balance of energy, only one tone (positionned at 0°) should be heard resulting from the fusion of the two tones at $\pm 45°$. However, no tone is heard in the middle, and the two tones are still perceived at their initial position. This result is due to the timing difference at the two ears when changes (onset or offset) occur, which constitutes a strong cue for spatial localization, and therefore for spatial segregation.

A frequency-specific ear-comparison process

Bregman provides evidence that ear comparisons must be frequency specific. This evidence is important because it justifies the auditory system's ability to derive separate locations for simultaneous sources. Besides, it justifies the choice made in the design of some multichannel audio coders to analyze spatial cues in the input signals into frequency subbands (see section chapter 2).

The first point states that lesions of the auditory cortex of a cat lead to the inability for it to tell where a short burst of a certain frequency is coming from in space, but its spatial localization ability may remain intact and normal for other frequencies [JM84]. As similar results have been reported for humans [ZP01, SS11], this would confirm that the auditory system is capable of frequency-specific localization.

Another point concerns the tolerance in mistuning of two tones at different locations for them to be fused together. As reported by Perrott and Barry [PB69], in the range from 250 Hz to 4 kHz, it is never greater than about 7%, which implies that spatial fusion requires quite a fine tuning to operate.

Conversely, an experiment by Kubovy and Howard [KH76] shows the existence of a computation of location that is separate for each frequency, but also has a separate memory for each. To do that, they presented listeners over headphones with a chord composed of six pure tones of separate spatial position in the horizontal plane by introducing an ITD cue for each. The resulting overall sound was perceptually confused. Then, after a brief pause, they presented the same chord except that the horizontal position of one of the components was changed. This time, the moved tone was salient, and its pitch was heard separated from the blurred mass. Note that the first chord was obviously indispensable for this experiment to succeed, and for most of the listeners, the silence must not exceed 1.5 seconds.

Cramer and Huggins made an experiment [CH58] tending to demonstrate the auditory system's ability to separate different spectral components on the basis of their points of origin in space. Using a diotic broad-band noise from which the low-frequency part (below 1000 Hz) has been phase delayed in one ear relative to the other (the remainder being identical at the two ears), the perceived result is that the pitch of the low-filtered part emerges slightly from the rest of the broad-band noise.

Concerning the fusion observed between a sound event and its echo (known as the "precedence effect" [LCYG99, WNR49]), the stream notion seems to be a valid interpretation of this scene-analysis process, since only one sonic event has actually occured. An interesting remark, however, is that this process seems to rely on a "delay-resisting" common fate process between the original sound event and its echo, which follow identical variations in frequency or intensity, but are delayed in time.

There are experiments showing that segregation by spatial location can be quite compelling. For example, Green [GKJP83] noticed that, given two spectra with different overall loudness, listeners can either recognize two spectra with identical shapes, or differentiate two spectra with different shapes, even if their shapes differ only by a modest increase in the intensity of one spectral component (at 949 Hz) relative to the others. However, if the 949 Hz component is separated and sent to the opposite ear, listeners could no longer integrate the left- and right-ear sounds into a single qualitative analysis.

Interaction with other grouping cues

The fact that two sounds coming from the same direction do not necessarily originate from the same acoustic event (a sound source might be acoustically transparent), and that spatial cues (binaural cues and spectral cues) are not infallible (for example in a reverberating environment), leads the auditory system to consider other grouping cues, as well (see section 1.3.5). Because, for instance, the fundamental frequency of a sound event, or its internal harmonic relations, are not modified by the environment, these cues can override spatial cues in the decision made by the auditory system.

An illustration of this is a pattern made at IRCAM[9] by Reynolds and Lancino for use in a musical piece. It consists of the synthesis of even and odd harmonics of an oboe on

[9] Institut de Recherche et Coordination Acoustique/Musique, Paris, France.

separate speaker channels, with FM micromodulations specific to each channel imposed on the harmonics. When left and right micromodulations are synchronized and identical, the listener heard a single centered oboe, but as the micromodulations are gradually made independent from left to right, but synchronously within each speaker, the sound seemed to split into two separate sounds, even harmonics on one channel, and odd harmonics on the other one. Here, in the first case, ear cues are overcome by spectral cues (harmonic relations) and sequential cues (common fate with micromodulations).

An experiment by Steiger and Bregman [SB82] which tries to assess the extent to which the fusion of partials between two simultaneous sound events can occur in dichotic and diotic situations reports that the requirements for the fusion of partials are less stringent within an ear than across ears. Bregman justifies this idea by the fact that nature is able to send the same sound to different ears with loudness lowered in one ear, but not pitch. To the contrary, most natural sounds contain partials that are not exactly harmonically related, so within-ear fusion can accept less exact relations between partials.

An important point, illustrated by the illusion from Deutsch described in section 1.3.5, is that spatial cues do not particularly override other grouping cues in the presence of discrepancies. Moreover, in those situations, the auditory system tries to build a consistent, but erroneous, description of the auditory scene.

One last interesting interaction is with visual cues. The ventriloquism effect [WW80] occurs when the visual and the auditory spatial locations of a multimodal event do not coincide. In this case, the typical perceptual result is the choice of an intermediate location for the sound. However, when several visual event occur at the same time, the decision seems to rely upon at least two criteria: the angular distance between the auditory and the potential visual event (which must not be too important) and their temporal synchrony (which must at least be weak).

1.5.3 Interactions Between Sequential and Simultaneous Integrations

Despite the small influence of spatial cues on simultaneous integration, the streaming of sound elements over time are more strongly influenced by spatial location, particularly for speech intelligibility. Darwin and Hukin [DH00] asked subjects to report a target word contained in a target carrier phrase, while a second carrier phrase was presented simultaneously. Thereby, two candidate target words were presented simultaneously during a time-aligned temporal gap present in both the target and competing carrier phrases. Despite the presence of grouping cues opposed to spatial cues, the subjects reported the target word that spatially matched the target phrase.

Shinn-Cunningham [Shi05] proposed a simplistic view of how spatial cues affect auditory scene analysis:

1. Spatial cues do not influence grouping of simultaneous sources. Instead other sound features determine how simultaneous or near-simultaneous sounds are grouped locally in time and frequency, forming “snippets” of sound.

2. Once a sound snippet is formed, its spatial location is computed, based primarily on the spatial cues in the sound elements grouped into that snippet.

3. Sound snippets are then pieced together across time in a process that relies heavily on perceived location of the snippets.

However, it has been shown by Darwin and Hukin [DH97, DH98] that spatial cues can also influence simultaneous grouping when other grouping cues are ambiguous. With

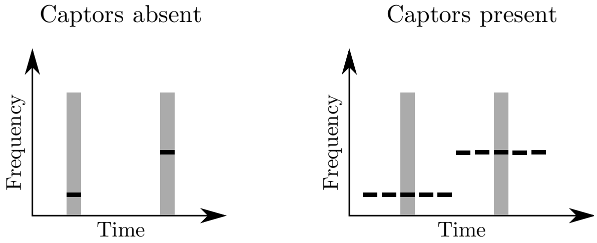


Figure 1.31: Stimuli used by Bregman and Steiger. The vertical bars represent noise bursts (played on a loudspeaker) and the dark horizontal bars are tones (presented over headphones). (Reprinted from [BS80].)

stimuli in which a target tone could logically fall into one of two streams, one a sequence of repeated tones and the other a simultaneous harmonic complex, they measured the degree to which the ambiguous target was heard as part of the harmonic complex. The results showed that when the spatial cues in the target and the harmonic complex matched, the tone was heard more prominently in the harmonic complex than when the spatial cues were uninformative. But this study also shows that spatial cues can influence grouping when top-down listener expectations also influence grouping: the results of a given stimulus depend on what subjects heard in past trials. In the same spirit, research conducted by Lee and Shinn-Cunningham [LSO05] provided evidence that the perceptual organization of a mixture depends on what a listener is attending to. In particular, with the previous tone paradigm and by changing which object a given listener was asked to attend to (holding the same stimuli), they found that there was no predictive relationship between the degree to which the target was *in* one auditory object and the degree to which it was *out* of the other.

An experiment [BS80] conducted by Bregman and Steiger is interesting because it involves spatial location in elevation. As illustrated in **figure 1.31**, it uses white noise bursts played on a loudspeaker simultaneously with a high or low pure tone (presented over headphones) in order to color it, leading to a localization of the noise burst higher or lower in space (see sections 1.4.2 and 1.4.5). However the experiment shows that when captors of the frequency of the pure tone are present before and after the mixture in order to capture the pure tone into a separate stream, the perceived position of the noise burst relies only upon the real position of the loudspeaker. This shows the effects of perceptual organization on the actual perceived location of sounds.

As in the continuity illusion (see section 1.3.4), sequential integration overcomes the simultaneous integration (and more precisely spectral integration). There is a simple experiment, designed by Bregman, that illustrates that sequential integration can overcome simultaneous integration (actually binaural integration in this case) by playing the same pure tone on the left and right channels, but with varying intensity in one ear (left, for example) from silence to the level of the right ear (which is fixed). If the intensity variation is low, the perception is one tone which seems to move from the right to the center, as expected. However, when increasing variation speed, two tones are heard, one at the right, and the other one pulsing at the left.

1.5.4 Speech-Sound Schemata

In the particular case of speech, differences in spatial origin rarely prevent the auditory system from grouping the information from different spatial locations, because it uses powerful schemata for the recognition of speech sounds, which could not care about the spatial origins of acoustic components.

This could be illustrated by an experiment from Broadbent [Bro55] in which listeners are presented with high-frequency components of speech to one ear and low-frequency components to the other. As a result, the listeners fused the two parts and heard a single sound. Broadbent explains this result by the commonalities of the two signals: fundamental frequency, onset/offset synchrony, harmonicity, etc. Bregman notes that this fusion is possible because spatial location can be assessed independently in different frequency bands (as suggested in sections 1.4.1 and 1.5.2) and results from an heuristic vote comprising spatial estimates among many other factors.

Similar results were reported by Cutting [Cut76] with formants presented to different ears. One interesting point is that when the fundamentals of the formants of a two-formant synthetic syllable were only one-third of a semitone apart (100 and 102 Hz), the formants were almost always heard as two sounds. Such a little difference in frequency would probably not have led to a segregation if the sounds had not been in separate ears. This tends to show that spatial estimation and fundamental frequency cues increase their own segregation power when acting together. Moreover, in the case of different fundamentals, the listeners reported hearing two sounds, but identified the speech sounds anyway. It is possible to break the phonetic fusion by repeating the syllable over and over, but only if the two sounds are differentiated both by location *and* fundamental frequency. Hence, phonetic integration is difficult to disrupt probably because of speech-sound schemata. Bregman explains the relative insensitivity of the phonetic schemata to factors such as perceived location and fundamental frequency by the fact that probably neither of these factors is used by the recognition schemata since neither is involved in the definition of a speech sound, in opposition to the time-varying spectral pattern. However, they may be important in situations where several speech-sound schemata are in competition, assuming that speech sounds come from different spatial directions.

A last experiment, by Scheffers [Sch83], which is not related to spatial cues lets us suppose a particular way in which these speech-sound schemata work. Listeners are presented with pairs of vowels, each vowel with a given fixed fundamental frequency. Results show that a difference in the fundamental frequency facilitates vowel segregation. It could be explained by the fact that the auditory system probably grouped the set of harmonics for each vowel to reconstitute its spectral shape. But this experiment also reports that using the same fundamental frequency, with same spatial location and onset/offset synchronization, listeners recognized the vowels 68% of the time. So, as noted by Bregman, it seems that speech-sound schemata integrate spectral information without the cues used for scene analysis, simply by taking the desired material from a dense mixture of sound.

1.6 Conclusions

Localization cues, as described in section 1.4, constitute a low-level spatial assessment of each frequency component of the auditory input, but the grouping into auditory streams (and the derivation of their associated spatial locations) is performed at a higher level and in combination with other modalities (such as vision, for example) to yield a final description of the auditory scene. Bregman believes that the auditory system, as with other perceptual systems, acts as a voting system based on *heuristic criteria*, including localization cues as well as criteria such as those described in section 1.3. Because of the variability of the environment, each of these criteria alone is not guaranteed to succeed, but when put together, a good description of the scene can result. At the extremes, when all criteria vote in a different way, the solution is ambiguous, whereas when they are unanimous, the description is stable and clear.

Studies that investigated localization performance have been presented. Auditory spatial resolution (or localization blur) has been especially studied for target sources in quiet, that is without any interfering source. However, as we will show in chapter 3, this resolution is subject to deteriorate when multiple sources are present.

Chapter 2

State-of-the-Art of Spatial Audio Coding

In this chapter we provide a panorama of the technologies used in audio coding and more specifically in *spatial* audio coding. Here, the implicit meaning of coding is “achieving a bit-rate reduction”. The approach we have chosen is driven by the concepts behind audio coding rather than the actual codecs available today. The monophonic case is treated briefly first, to focus afterwards on stereophonic and multichannel signals.

When dealing with multi-channel signals, the input signals are assumed to be two or more channels either acquired using a microphone array or synthesized using spatialization techniques such as vector-base amplitude panning (VBAP) [Pul97], wave field synthesis (WFS) [dV09], higher-order Ambisonics (HOA) [DNM03] or even binaural rendering. Information about recording and synthesis can be found in [Rum01].

Particular attention has been given to HOA as well as to parametric spatial audio coding methods, as they both constitute the bases of the developments proposed in chapter 5.

2.1 Representation of Spatial Audio

2.1.1 Waveform Digitization

The most usual digital representation of a monophonic waveform is Linear Pulse-Code Modulation (LPCM) encoding, often referred to simply as PCM. The waveform is discretized at a given sampling rate using a uniform quantization. The fidelity of the representation is given both by the sampling rate and the quantization step size. The Nyquist-Shannon sampling theorem [Sha49] asserts that the uniformly spaced discrete samples are an exact representation of the signal if this bandwidth is less than half the sampling rate. So, in LPCM, the highest represented frequency component is half the sampling rate of the signal. The quantization step size depends on the number of bits with which each sample is coded [GG91]. The original waveform is thus approximated depending on these two parameters.

LPCM is especially used in the Compact Disc Digital Audio (CD-DA) standard, created by Sony and Philips in 1980 and regarded as a reference format in terms of quality. On a CD, a waveform is encoded using a sample rate of 44100 Hz and each sample is coded on 16 bits, meaning that a bit rate of $44100 \times 16 = 700$ kbit/s is needed. Therefore, it appears that a large bit rate is necessary to represent a monophonic signal using LPCM at CD quality.

Usually, to digitize multichannel signals (including stereophonic signals), LPCM is used on each channel separately.

2.1.2 Higher-Order Ambisonics

Representation of the acoustic field

Higher-Order Ambisonics (HOA) is an evolution by Bamford [Bam95] and Daniel [Dan01] of the *Ambisonics* sound spatialization technique initiated notably by Gerzon [Ger85]. HOA aims to represent the acoustic field in the vicinity of one point, assumed to be the listening position, and provides methods to reproduce this field in 2- and 3-dimensions using an array of loudspeakers.

As explained in [DNM03], HOA is a representation model of an acoustic wave based on a development of this wave on the eigenfunctions of the acoustic wave equation in spherical coordinates, with r being the radius, φ the azimuth angle, and θ the elevation angle. These eigenfunctions combine several functions to describe both the radial and the angular dependencies of the acoustic wave. The radial dependencies are described by spherical Bessel functions of the first kind $j_m(kr)$ and of the second kind (or Neumann functions) $n_m(kr)$, and/or spherical Hankel functions of the first kind $h_m^+(kr)$ (converging progressive wave). The angular dependencies are described by spherical harmonics $Y_{mn}^\sigma(\varphi, \theta)$.

Generally, the sound scene that is to be represented is only composed of converging waves, and the development of the acoustic pressure $p(\vec{r}, \omega)$, where ω is the angular velocity, and k the wave number, can be expressed as:

$$\underbrace{p(\vec{r},\omega)}_{\substack{\textbf{acoustic}\\ \textbf{pressure}}} = \sum_{m=0}^{+\infty} i^m \underbrace{j_m(kr)}_{\substack{\text{radial}\\ \text{dependencies}}} \sum_{n=0}^{m} \sum_{\sigma=\pm 1} \underbrace{B_{mn}^\sigma(\omega)}_{\substack{\textbf{HOA}\\ \textbf{coefficients}}} \underbrace{Y_{mn}^\sigma(\varphi,\theta)}_{\substack{\text{angular}\\ \text{dependencies}}} . \tag{2.1}$$

The acoustic pressure, and therefore the acoustic wave, is thus fully described by the $B_{mn}^\sigma(\omega)$ coefficients, which are called the *B-format* signals and constitute the HOA representation of the acoustic field. In practice, this HOA representation has to be truncated to a given order M, which leads to a representation of the 3D audio scene with $K_{\text{3D}} = (M+1)^2$ B_{mn}^σ components.

One can also restrict the representation to the horizontal plane, thus considering only the development of the acoustic wave in a 2D representation. This can be done by expressing the sound field in terms of its cylindrical harmonics, and in this case the representation is only composed of $K_{\text{2D}} = 2M+1$ B_{mm}^σ components:

$$\begin{aligned} p(\vec{r},\omega) = B_{00}^{+1} j_0(kr) &+ \sum_{m=1}^{+\infty} j_m(kr) B_{mm}^{+1} \sqrt{2}\cos(m\varphi) \\ &+ \sum_{m=1}^{+\infty} j_m(kr) B_{mm}^{-1} \sqrt{2}\sin(m\varphi). \end{aligned} \tag{2.2}$$

Figure 2.1 gives a 3D representation of the spherical harmonic functions. High m order functions have a higher angular variability. Ambisonics, as developed by Gerzon, corresponds to the first order Ambisonics case, that is, the W, X, Y and Z components. **Figure 2.2** shows the contribution of the spherical Bessel functions $j_m(kr)$ as a function of the distance from center O. Components of a higher m order contribute to the description of the farther sound field, regarding the wavelength.

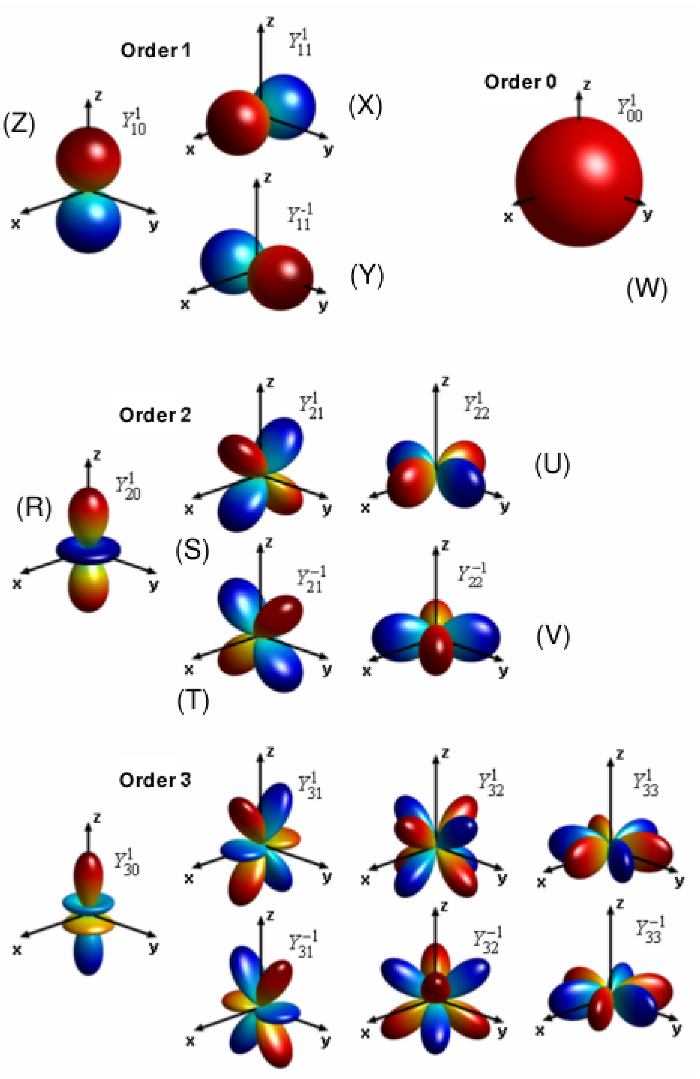


Figure 2.1: 3D view of spherical harmonic functions for orders $m = 0$, 1, 2 and 3, and their usual associated ambisonic components designation. Red and blue colors correspond to positive and negative values, respectively. (Reprinted from [Mor06].)

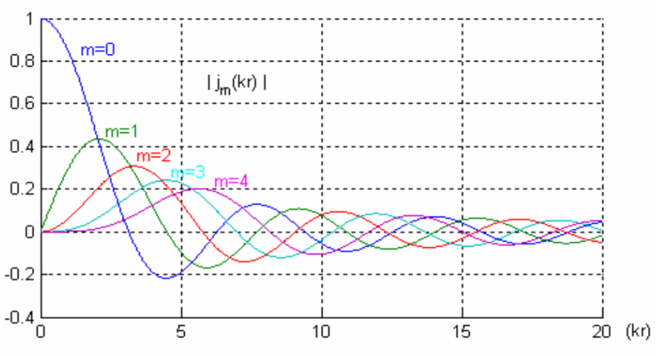


Figure 2.2: Spherical Bessel functions. (After [Mor06].)

Reproduction of the acoustic field

To acoustically reproduce the wavefield over a specific loudspeaker layout, it is necessary to derive the appropriate loudspeakers signals, which constitute the so-called *D-format*, from the ambisonic signals (i.e., B-format). This is done using the "re-encoding principle." We assume an array of N loudspeakers that are far enough from the center of the listening area such that their signals S_i are encoded as plane waves with coefficient vectors $\mathbf{c_i}$, recomposing the encoded Ambisonics components as $\tilde{B}^{\sigma}_{mn}$:

$$\mathbf{c_i} = \begin{bmatrix} Y^{+1}_{00}(\theta_i, \delta_i) \\ Y^{+1}_{11}(\theta_i, \delta_i) \\ Y^{-1}_{11}(\theta_i, \delta_i) \\ \vdots \\ Y^{\sigma}_{mn}(\theta_i, \delta_i) \\ \vdots \end{bmatrix}, \tilde{\mathbf{B}} = \begin{bmatrix} \tilde{B}^{+1}_{00} \\ \tilde{B}^{+1}_{11} \\ \tilde{B}^{-1}_{11} \\ \vdots \\ \tilde{B}^{\sigma}_{mn} \\ \vdots \end{bmatrix}, \mathbf{S} = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_N \end{bmatrix}. \tag{2.3}$$

Then, the re-encoding principle stands as below:

$$\tilde{\mathbf{B}} = \mathbf{C} \times \mathbf{S}, \tag{2.4}$$

with $\mathbf{C} = [\mathbf{c_1}, \cdots, \mathbf{c_N}]$ being the re-encoding matrix. The original ambisonic signal $\mathbf{B}$ is matrixed with a decoding matrix $\mathbf{D}$ to derive the decoded signals $\mathbf{S}$ that will feed the loudspeakers:

$$\mathbf{S} = \mathbf{D} \times \mathbf{B}. \tag{2.5}$$

To get $\tilde{\mathbf{B}} = \mathbf{B}$, equation (2.4) is inverted:

$$\mathbf{D} = \mathbf{C}^{+} = \mathbf{C}^{\mathrm{T}} \times \left(\mathbf{C} \times \mathbf{C}^{\mathrm{T}}\right)^{-1}, \tag{2.6}$$

assuming that $N \geq K_{\mathrm{2D}}$ or $N \geq K_{\mathrm{3D}}$ (that is to say, there are enough loudspeakers). Hence, $\mathbf{D}$ relies on the loudspeaker layout. For regular layouts, the expression of the decoding matrix is simplified and is given by [DNM03]:

$$\mathbf{D} = \frac{1}{N}\mathbf{C}^{\mathrm{T}}. \tag{2.7}$$

For instance, in the case of a horizontal-only regular array, we get:

$$\mathbf{D} = \frac{\sqrt{2}}{M} \begin{bmatrix} \frac{1}{\sqrt{2}} & \cdots & \frac{1}{\sqrt{2}} \\ \cos(\phi_1) & \cdots & \cos(\phi_N) \\ \sin(\phi_1) & \cdots & \sin(\phi_N) \\ \cos(2\phi_1) & \cdots & \cos(2\phi_N) \\ \sin(2\phi_1) & \cdots & \sin(2\phi_N) \\ \vdots & \cdots & \vdots \\ \cos(K_{\mathrm{2D}}\phi_1) & \cdots & \cos(K_{\mathrm{2D}}\phi_N) \\ \sin(K_{\mathrm{2D}}\phi_1) & \cdots & \sin(K_{\mathrm{2D}}\phi_N) \end{bmatrix}^{\mathrm{T}}, \tag{2.8}$$

where ϕ_i is the angle of the i^{th} loudspeaker. For an encoding/decoding of order M and reproduced over $2N+1$ loudspeakers in a circular and regular array, the resulting wave field error stays below -15 dB as long as the following relationship is respected [WA01]:

$$M \geq kr, \tag{2.9}$$

where r is the radius of the reproduction area. The disk of radius $r \leq M/k$ is generally referred to as the "sweet spot". From this formula one can conclude:

1. that the higher the frequency of the acoustic wave to be represented over a given area, the higher the truncation order needed to properly represent that wave over this area;

2. that the wider the desired area of proper reconstruction of this acoustic wave, the higher the truncation order needed.

It should be noted that research exists concerning the optimization of the decoding phase in situations where the satisfying acoustical reconstruction area is too limited regarding the listening area, for instance when the listener moves away from the ideal listening position or when the order M is too small. In those cases, it is possible to improve the subjective quality of source localization by optimizing objective criteria. As an example, two important criteria have been introduced by Gerzon for first-order Ambisonics signals, but which can be extended to higher orders: the *velocity* and *energy* vectors. For further reading on this optimization process, see [Ger92, Mor06].

Useful properties of HOA signals

The HOA format is universal in the sense that it is able to describe any acoustic wave. It has two interesting properties. First, it is independent of the recording and reproduction systems, which makes it very flexible and suitable in several situations. In particular, a sound field represented in the HOA domain can be reproduced in theory on any type of loudspeaker layout.

Second, the HOA description of the scene is hierarchical; that is, the first-order components are sufficient to represent the acoustic wave, and the higher-order components only extend the bandwidth and the accurate area of reproduction of this wave. This property makes the HOA format *scalable*, which is very useful in a network transmission context to adapt the transmission rate to the available network bandwidth; it is possible, given a HOA representation of order M of the sound field, to give priority to the transmission of the B_{mn}^{σ} components associated with the first L orders ($L \leq M$), affecting the size of the accurate listening area and the reproduced bandwidth. In the same way, it is possible, given an order M with $2M + 1$ channels (2D HOA), to adapt the reproduction to the listener's setup by considering only a subset of $2L + 1$ channels with an order L representation.

2.2 Coding of Monophonic Audio Signals

As explained in section 2.1.1, to ensure CD quality using the LPCM representation of a waveform, a bit rate of 700 kbit/s is needed. Audio coding solutions exist to reduce this reference bit rate with or without loss of information.

2.2.1 Lossless Coding

Starting from the fact that the transmitted audio signals are rarely full-scale white noise (i.e., a random signal), redundancies in the waveform can be accounted for in order to reduce the bit rate without a loss of information. The basic idea is to predict future samples based on previous ones using linear prediction [Mak75] and to transmit only the prediction error (i.e., the residuals) and the prediction rules (if applicable). If the input signal is self-correlated, because the entropy of the residuals is less important than that of the original signal, a potential coding gain can be achieved. For example, in *Differential*

Pulse-Code Modulation (DPCM), the waveform is differentiated (the previous sample is subtracted from the next one) and then entropy coded. In a second form of DPCM, a local model of the input is made, the prediction is subtracted from the original signal, and the residuals are entropy coded. This principle is used for example in *MPEG-4 Audio Lossless Coding*, or in DTS.

Note that since the entropy of the residuals is unknown, any lossless coding will always have a variable bit rate on normal audio.

2.2.2 Lossy Coding

Perceptual Coding

Another approach, know as perceptual coding (see appendix C for a detailed description), is to accept to lose information during the coding process, all the while ensuring that this loss is as imperceptible as possible by the auditory system. Therefore, perceptual coding models and exploits properties of hearing, such that less perceptible information can be coded with less precision. In the frequency domain, or frequency subbands domain, the quantization of each sample induces a quantization noise. The energy of this noise depends on the coding precision of the considered sample: the higher the number of bits allocated to it, the lower the noise. As we saw in section 1.2.2, each frequency component generates a mask, and any other component whose energy falls below this mask will be inaudible. A cumulative mask, resulting from the presence of all frequency components, is modeled and used to compute how far the quantization noise would be from the masking threshold. The bit allocation to each sample is then done in order to keep the quantization noise as much below the mask as possible. This principle is especially used in the MPEG-1/2 Layer 3 (MP3) codec [BG02, Bra99].

Spectral Band Replication

When operating at very low bit rates, the quantization noise might not be kept below the masking threshold, resulting in audible and unpleasant distortions in the signal. A method named *Spectral Band Replication* (SBR) proposed recently [DLKK02], offers a way of reducing the bandwidth of the signal without altering the perceptual rendering too much. It is based on the idea that low and mid frequencies can be used to reconstruct the untransmitted higher frequencies (usually over 5 kHz) by replication/transposition, plus a set of parameters describing the spectral envelopes of the noise and the harmonic content. SBR can potentially be associated with any perceptual coder and is especially used in MPEG-2/4 as *High-Efficiency Advanced Audio Coding* (HE-AAC) [WKHP03].

Parametric Representation of the Signal

Again, at very low bit rates and for the same reasons as mentioned above, a parametric representation of the signal components might be more efficient. For example, in *Harmonic and Individual Lines plus Noise* (HILN) [PEF98] using a time-frequency transform, the important harmonic content is matched and tracked across frames to group them into harmonic lines and individual sinusoids. Individual sinusoids are described as amplitude and frequency, harmonic lines as fundamental frequency, amplitude, and spectral envelope of the partials, and the noise as amplitude and spectral envelope. Only the differences between the components of a track are coded, resulting in a coding gain, especially for long tracks. A perceptual model is used as well to select only the most audible components given the bit rate constraint. HILN is used down to 6 kbit/s in MPEG-4 Parametric Audio Coding.

This type of coding is known, however, to present issues in describing transients of the original signal. This issue has been addressed in [BA04].

2.3 Lossless Matrixing

The same linear prediction strategies, described for monophonic signals in section 2.2.1, can be applied to each channel separately. However, a lossless matrixing technique can be used first to reduce the inter-channel correlations.

2.3.1 Mid/Side Stereo Coding

Mid/Side Stereo Coding [JF92] is a form of matrixing used to reduce correlations between the two channels of a stereo signal. The principle is to transform the left and right channels into a sum channel, called mid channel, $m[n] = \frac{1}{\sqrt{2}}(l[n] + r[n])$, and a difference channel, called side channel, $s[n] = \frac{1}{\sqrt{2}}(l[n] - r[n])$, which carries the residuals. This way, the m and s channels are less correlated than the original l and r channels, and in particular, the entropy of s is reduced. At the decoding phase, by using the same operations on the transformed channels, the original stereo channels can be fully recovered, provided that no additional lossy coding has been used on either of the transformed channels (see section 2.4.1). M/S coding is one of the two stereo joining techniques, together with ISC (presented in section 2.5), and is usually applied for coding low frequencies, whereas ISC is used for high-frequency coding.

2.3.2 Meridian Lossless Packing

For the multi-channel audio case, in [GCLW99], Meridian Audio proposes an invertible matrixing technique, which reduces the inter-channel correlations prior to applying linear prediction on each channel separately. The result of this combination is called *Meridian Lossless Packing* (MLP) and is widely used for audio and video DVD, but also by Dolby's AC-3 codec on Blu-ray discs. MLP typically provides a 2:1 compression for music content.

2.4 Lossy Matrixing

2.4.1 Perceptual Mid/Side Stereo Coding

As stated at section 2.3, M/S stereo coding cannot be used as is in conjunction with a lossy coding. Indeed, considering an input signal x, the quantization noise $Q(x)$ resulting from perceptual coding (see section 2.2.2) can be modeled as an adaptive noise source $N(x)$:

$$Q(x) = x + N(x). \tag{2.10}$$

When perceptual coding is used on a monophonic signal x, the power spectrum of $N(x)$ can be kept below the mask induced by x. But when perceptual coding is applied on M/S matrixed signals m and s, the reconstructed stereo signals l' and r' are:

$$\left\{ \begin{array}{lcl} l' & = & \frac{1}{\sqrt{2}}[Q(m) + Q(s)] = l + \frac{1}{\sqrt{2}}[N(m) + N(s)] \\ r' & = & \frac{1}{\sqrt{2}}[Q(m) - Q(s)] = r + \frac{1}{\sqrt{2}}[N(m) - N(s)] \end{array} \right. , \tag{2.11}$$

resulting in a sum of uncorrelated quantization noises $N(m)$ and $N(s)$ of which the overall power is equally present in l' and r'. However, as explained in section 1.4.6, binaural unmasking occurs when a target source, masked in the presence of a collocated concurrent masking source, becomes audible when these sources are presented at different positions. Therefore, if the masking thresholds induced by l and r are spectrally unequal, the quantization noise can become audible. To avoid this, a solution is proposed, for example, in [tK96], but the usual strategy, described in [JF92], is to dynamically switch between two modes: using a time-frequency representation of the input signals, if the masking thresholds of l and r in a given frequency subband differ by less than 2 dB, m and s will be quantized and transmitted (M/S mode), otherwise l and r will be (L/R mode).[1] Then, in M/S mode, the stereo masking contribution of the m and s signals is computed, and the signals are quantized such that their respective noises stay below their respective stereo masks.

In Dolby Stereo, M/S coding is used in the time domain, but in digital codecs it is performed independently in separate frequency subbands, as in Dolby AC-3 [Dav99] (under the term *Rematrixing*). In MPEG-2 AAC [BBQ+97], M/S coding is used in a multi-channel context by applying it to pairs of opposite channels with respect to the front/back axis.

2.4.2 Matrix Encoding

The matrix encoding principle is based on the downmixing of a finite number of audio channels (e.g., 5) into a lesser number of channels (e.g., 2) to reduce the necessary bit rate. This encoding is done such that after transmission the signal can either be played as is, or, with an appropriate decoder, upmixed to the original number of channels, which are then approximated. An encoding-transmission-decoding chain is usually noted as I:D:O, where I is the number of input channels (i.e., prior to the encoding), D the number of channels in the downmix, and O the number of output channels (i.e., after decoding). In the common case where the transmitted downmix is a pair of stereo channels, those channels are usually noted L_t and R_t for "Left Total" and "Right Total", respectively. Few matrixing/unmatrixing techniques examples are presented hereafter, and more details can be found in [Rum01].

Ambisonics UHJ

Gerzon [Ger85] proposed a 4:2:4 scheme to transmit Ambisonics signals (see section 2.1.2) in a stereo-compatible way. The encoding equations are:

$$\left\{\begin{array}{lcl} S & = & 0.940\ W + 0.186\ X \\ D & = & j(-0.342\ W + 0.510\ X) + 0.656\ Y \\ L_t & = & \frac{1}{2}(S + D) \\ R_t & = & \frac{1}{2}(S - D) \end{array}\right. , \tag{2.12}$$

[1] In any case, if these masking thresholds differ too much, it means that l and r are uncorrelated, and hence, M/S coding would not be efficient.

where j is a +90° phase shift, and W, X, Y are the original Ambisonics B-format signals. L_t and R_t compose the stereo downmix. The received downmix can be decoded with:

$$\left\{\begin{array}{lcl} S & = & \frac{1}{2}(L_t + R_t) \\ D & = & \frac{1}{2}(L_t - R_t) \\ W' & = & 0.982\ S + j\ 0.164\ D \\ X' & = & 0.419\ S - j\ 0.828\ D \\ Y' & = & 0.763\ D + j\ 0.385\ S \end{array}\right. . \tag{2.13}$$

Dolby Surround

Dolby Motion Picture (MP) matrix, also known as *Dolby Stereo* [Dre00], is a 4:2 encoding scheme based on the assumption that the sound scene is organized as a set of primary sources in front (carried by the front left channel L, the front right channel R, and the center channel C), and secondary sources and room effects/reverberation at the rear (carried by the rear surround channel S). As its name states, this configuration is especially adapted to motion picture material. The encoding equations are:

$$\left\{\begin{array}{lcl} L_t & = & L + \frac{1}{\sqrt{2}}(C + jS) \\ R_t & = & R + \frac{1}{\sqrt{2}}(C - jS) \end{array}\right. , \tag{2.14}$$

The decoder part, referred to as *Dolby Surround*, uses the following equations:

$$\left\{\begin{array}{lcl} L' & = & L_t \\ R' & = & R_t \\ C' & = & \frac{1}{\sqrt{2}}(L_t + R_t) \\ S' & = & \frac{1}{\sqrt{2}}(L_t - R_t) \end{array}\right. , \tag{2.15}$$

If no center speaker is present in the reproduction system, and because C is equally present in L_t and R_t (and thus in L' and R'), a phantom center speaker will be perceived by the listener, thereby reinforcing the frontal image, assuming that the listener is positioned equidistant from the left and right speakers.

The effective separation resulting from the matrixing-unmatrixing process is unequal between pairs of opposite channels, and pairs of adjacent channels. Indeed, the separation is perfect between L' and R', since L and R are not mixed together. Also, because S is present in phase opposition between L' and R', the presence of S cancels out when computing C', thus yielding a perfect separation between C' and S'. However the separation is poor (3 dB) between any adjacent pair of speakers, causing spatial degradations like a narrowed frontal scenery or inaccurate room effects. Also, any difference between the original L and R channels will leak into S', but this issue is generally improved by taking advantage of the precedence effect (see for example [LCYG99, WNR49]): by adding a time delay between the front channels and the surround channel, front sources will be perceived as coming from the front (first wavefront), the late signal coming from the rear being perceived as a room effect. This delay can be optimized if the distance between the listening position and the front and surround speakers is known. But this trick also reinforces the presence of S in L' and R'. However, Dolby justifies this choice by the nature of the target material (motion picture). Indeed, sounds are expected to come predominantly from the screen direction. Besides, signals associated with the surround track usually are not associated with specific source locations.

Active Matrix Decoding

In order to further improve channel separation, passive decoders like Dolby Surround, can be replaced by active ones. These are based on the idea that the signals obtained by the passive decoder can be used to detect a predominant source and that a better separation of the channels can be achieved by manipulating those signals. Therefore, active decoders continuously monitor the encoded channels for soundfield dominance and use an adaptive matrix to enhance the directionality of the predominant source at a given time. This principle is known as *steering logic* and is used, for example, in *Dobly Pro Logic*, *Lexicon Logic7* or *SRS Circle Surround*. In Dolby Pro Logic [Dre00], the predominant source direction is expressed continuously on an XY-coordinate plane by computing the energy ratios between L' and R' (X-axis), and between C' and S' (Y-axis). Then converting from rectangular to polar coordinates gives the dominance expressed as a vector quantity, of which the magnitude represents its relative dominance and the angle its direction. To improve the separation of the dominant source with other sources, techniques such as cancellation are used: for example, if a frontal source is detected, its leakage into the L' and R' channels is removed by inverting the polarity of one of them and adding them together. Clearly this method blends L' and R' together and looses their separation, but the dominant source is now played only in the front speaker.

A system with broadband steering like Pro Logic is known to create unpleasant and unnatural effects like "pumping" or gating. These effects can be attenuated by steering separately in at least two frequency subbands, thereby allowing each subband to be steered toward a different side. Besides, this design allows the use of an optimal time constant for each subband without introducing distortions. As an example, Circle Surround uses three subband steering generators: low, mid, and high frequencies.

2.4.3 Matrixing Based on Channel Covariance

Yang *et al.* [YAKK03, YKK04] proposed a high bit-rate coding model based on inter-channel redundancy removal called *modified AAC with Karhunen-Loève transform* (MAAC-KLT). In this method, the input channels are statistically decorrelated by applying a *Karhunen-Loève Transform* (KLT). The interest of this transform is that most of the energy is compacted into the first several resulting channels, allowing for significant data compression by entropy coding or by selecting the channels associated with the highest variances. Energy masking thresholds are computed based on the transformed signals, and their frequency components are then bit-quantified. As its name states, MAACKLT is designed to be incorporated into the AAC coding scheme [BBQ+97].

The Karhunen-Loève transform, also known as *principal components analysis* (PCA), is a linear transformation projecting data onto the eigenvector basis of their covariance. In our case, data are the n (correlated) channels of the signal, represented by the $n \times k$ matrix V (k is the number of samples of a temporal frame):

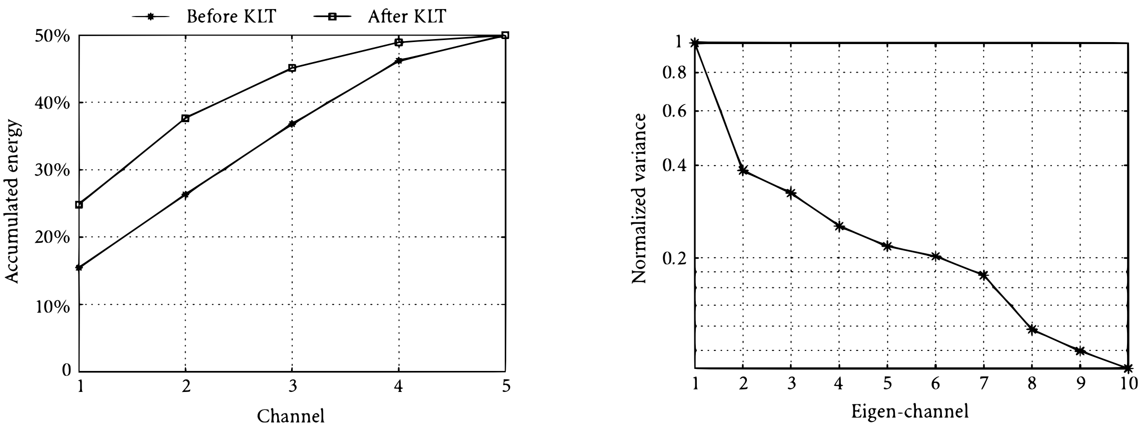
$$V = [V(1), \dots, V(i), \dots, V(k)], \text{ with } V(i) = [x_1, x_2, \dots, x_n]^T. \tag{2.16}$$

The covariance matrix C_V of V is defined as:

$$C_V = \mathrm{E}\left[(V - \mu_V)(V - \mu_V)^T\right] = \frac{\sum_{i=1}^{k}[V(i) - \mu_V][V(i) - \mu_V]^T}{k}, \tag{2.17}$$

where the mean vector μ_V is defined as:

$$\mu_V = \mathrm{E}[V] = \frac{\sum_{i=1}^{k} V(i)}{k}. \tag{2.18}$$



(a) Comparison of accumulated energy distribution for a five-channel audio source.

(b) Normalized variances (eigenvalues) for a ten-channel audio source

Figure 2.3: MAACKLT performances (After [YKK04].)

The KLT matrix, M, is defined as the eigenvectors $m_1, m_2, \ldots, m_n$ of the covariance matrix C_V:

$$M = [m_1, m_2, \ldots, m_n]^T. \tag{2.19}$$

Finally, the matrix of transformed channels U is given via an orthogonal transform:

$$\begin{array}{ccccc} \begin{matrix}\mathit{KLT}\\ \mathit{matrix}\\ M\end{matrix} & & \begin{matrix}\mathit{Correlated}\\ \mathit{component}\\ V\end{matrix} & & \begin{matrix}\mathit{Decorrelated}\\ \mathit{component}\\ U\end{matrix} \\ \begin{pmatrix} - & - & - & - & - \\ - & - & - & - & - \\ - & - & - & - & - \\ - & - & - & - & - \\ - & - & - & - & - \end{pmatrix} & \times & \begin{pmatrix} - \\ - \\ - \\ - \\ - \end{pmatrix} & = & \begin{pmatrix} - \\ - \\ - \\ - \\ - \end{pmatrix} \end{array} \tag{2.20}$$

The transformed channels in U are called eigenchannels. The transform produces statistically decorrelated channels in the sense of having a diagonal covariance matrix filled with eigenvalues of C_V for transformed signals U. Since C_V is real and symmetric, the matrix is formed by normalized eigenvectors which are orthonormal. Thus the inverse transform matrix is equal to its transpose, so that $V = M^T U$ in reconstruction, which is an “immediate” computation.

Each eigenvalue represents the variance of the projection onto its corresponding eigenvector, that is to say the variance of the eigenchannel. As a result of this transform, most of the energy is compacted into the first several eigenchannels (see **figure 2.3**), and experimental results showed that the compaction efficiency increases with the number of input channels. This property allows a great deal of data compression by selecting in M the eigenvectors with largest eigenvalues, and transmitting only their associated eigenchannel, which minimizes the error in the least-square-error sense between the original and the reconstructed channels. Besides, this energy compaction property implies that the entropy of the last eigenchannels is reduced, ensuring an additional coding gain. In that sense, the KLT applied to a set of channels can be thought of as a generalization of the M/S stereo coding presented in section 2.3.1. In consequence to all of this, MAACKLT can be used as a scalable format: the higher the number of transmitted channels, the better the reconstruction quality, knowing that perfect reconstruction is reached (apart from the quantization noise) when all channels are transmitted.

This transform can be done in either the time or the frequency domain. However, Yang explains that globally the frequency domain offers better performance. This is due to two factors. First, in general, the signal energy is compacted in the low part of the spectra. Second, time domain signals may contain effects like delay or reverberation among different channels, which affect the time domain KLT decorrelation capability.

In order to achieve a maximum decorrelation of channels, a temporal adaptive KLT can be done. This involves updating the KLT matrix M each adaptation period or block. Because M has to be transmitted for each block, the temporal adaptive KLT increases the bitrate. Anyway, it is possible to reduce it by transmitting the covariance matrix C_V of V instead of M, because it is real and symmetric, which implies that its lower (or higher) triangular part is sufficient to compute M on the decoder side. Considering a covariance matrix quantized to 16 bits per element, the overhead bit rate is given by:

$$r_{\text{overhead}} = \frac{8(n+1)}{K}, \tag{2.21}$$

where K is the adaptation period in seconds, and n the number of channels. Yang estimated that an adaptation period of about 10 seconds is optimal, whereas for K less than 10 seconds the bitrate is too important regarding the decorrelation efficiency.

2.5 Parametric Spatial Audio Coding

So far, in matrixing techniques, the effort was put on trying to transmit all the channels by combining them or by eliminating redundancies between them. A rather different approach, presented in this section, is to transmit a parametric representation of the spatial attribute of the sound scene. As depicted in **figure 2.4**, the general idea is to use a time-frequency representation to extract from the input channels a set of "spatial" parameters describing the spatial organization of the scene/channels, on the one hand (see section 2.5.1), and to group all these channels using a downmixing technique (see section 2.5.2) to form a single mono or stereo signal thereby reducing inter-channel redundancies, on the other hand. The spatial parameters are then used in the decoding phase (see section 2.5.3) to reconstruct from the downmix an approximation of the original channels, or even to generate a new set of channels adapted to another loudspeaker setup. These schemes especially rely on the supposition that the transmission of the parameters takes only a few kbits/s, which is very small compared to the bit rate dedicated to the audio channel(s). This means that the quantization process of these parameters has to be performed carefully to ensure a reliable spatial representation with only a few bits (see section 2.5.4). Besides, as already evoked in section 2.4.1, if the downmix is perceptually encoded prior to transmission, binaural unmasking has to be taken into account to avoid noise unmasking [tK96].

This approach has been followed by several methods, using two main strategies, depending on the nature of the spatial parameters to be extracted. First, these parameters can represent the inter-channel differences that are relevant to the perception of space, with respect to the auditory localization cues (see section 1.4). *Frequency joining* methods constitute the first step in this direction, but to a limited extent. Initially designed for a pair of stereo channels under the term *Intensity Stereo Coding* (ISC) [HBL94], and later extended to multi-channel as *Channel Coupling* [BBQ$^+$97], the frequency join consists of

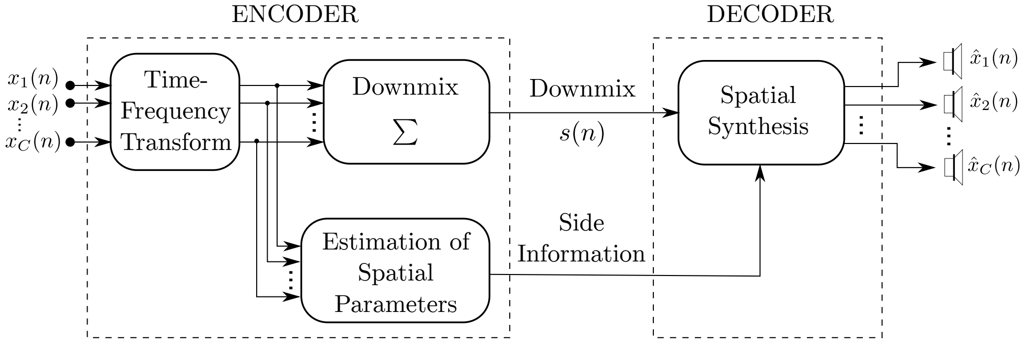


Figure 2.4: Generic scheme of parametric spatial audio coding.

generating a "coupling channel" which is an in-phase sum of the input channels, plus a set of parameters, named scale factors, representing the power ratio between each input channel and the coupling channel, for each frequency subband of the current frame. Channel coupling is used exclusively to represent high frequencies, assuming that localization of such components is based on ILD and on interaural delay of the signal energy envelope (high-frequency ITD). The fine temporal structure information, which is lost with this representation, is not necessary anyway for localization at these frequencies. In other words, each channel will share the same set of spectral values contained in the coupling channel, but their relative power ratio will be reconstructed from frame to frame with the scale factors, ensuring a conservation of the high-frequency localization cues. Concerning the low frequencies, however, this information is necessary because localization in this case is based on IPD/ITD, and therefore the corresponding spectral coefficients cannot be coupled. For stereo signals, ISC is usually applied in conjunction with M/S stereo coding (see section 2.3.1) for low frequencies, forming together the joint stereo technique. ISC is used for example in MPEG-1/2 Layer 3 (MP3) [BG02, Bra99] for stereo signals, and in MPEG-2 AAC [BBQ$^+$97] for stereo signals and multi-channel signals, by using it on pairs of opposite channels with respect to the front/back axis. MPEG-2 AAC also offers the use of channel coupling instead of ISC, among different pairs of M/S coded channels. ISC and channel coupling are also used in Dolby AC-3 [Dav99] and to a lesser extent in DTS (scale factors are not transmitted). As an example, compared to an LPCM encoded stereo signal (see section 2.1.1), an MP3 encoded stereo signal will be about 7 times smaller when used with a bit rate of 192 kbit/s, which is considered as the transparent quality bit rate.

Because the phase information is lost in the frequency joining process, issues arise when uncorrelated components are concerned. Therefore this approach has been improved in *Parametric Stereo* (PS) [BvdPKS05] and *Binaural Cue Coding* (BCC) [Fal04] with a full parametric description using parameters carrying all the (azimuth) sound localization cues used by the auditory system, that is ITD, ILD, and IC (see section 1.4), expressed as inter-channel differences. This time, the parametric description is valid on the full bandwidth, and therefore spectral coefficients for the whole spectrum are shared via a downmix signal. PS is intended for stereo signals, while BCC applies to multi-channel signals as well. However, because these types of representation carry the inter-channel differences, they are somehow restrained to stick to the original channel setup at the decoding phase. The MPEG Surround standard [BF07], which is based on these representations, proposes mechanisms to correct this issue. PS and BCC are very similar, and since BCC is not restricted to stereo signals, only BCC will be considered in the following.

In a second category, the spatial parameters would rather represent a spatial description of the scene, by means of angular positions notably. *Spatial Audio Scene Coding* (SASC) [GJ08] and *Directional Audio Coding* (DirAC) [Pul07] both belong to this category. SASC and DirAC are general frameworks that can be used for other purposes than transmission with a reduced bit rate. Particularly, they offer possibilities of format conversion or enhancement. Indeed, unlike methods from the first category, they present the advantage of being format-agnostic, in the sense that their representation of the acoustic field is generic and not tied to the input format. This property allows a more flexible spatial synthesis potentially compatible with any loudspeaker arrangement, or even with binaural rendering. However, this is beyond the scope of this thesis and hence in the following we will focus only on the transmission context, that is when a downmix channel and a set of parameters are sent. A notable peculiarity of DirAC, compared to BCC and SASC, is that the input signals are assumed to be first-order ambisonics components in B-format, that is the W, X, Y and Z components (see section 2.1.2). In DirAC and SASC, the input and output loudspeaker setups are assumed to be known during the encoding and decoding phases, respectively.

2.5.1 Extraction of the Spatial Parameters

Once a time-frequency representation of the input signals is obtained, a spatial analysis is performed to extract the set of parameters associated with each frequency subband. Here is a summary of the parameters that are extracted for each method (their description is given in the next sections).

BCC:

- Inter-Channel Coherence c
  (1 coefficient)
- Inter-Channel Level Difference ΔL
  ($C - 1$ coefficients)
- Inter-Channel Time Difference τ
  ($C - 1$ coefficients)

where C is the number of input channels.

DirAC:

- Diffuseness coefficient Ψ
- Direction vector $\mathbf{D}$

SASC:

- Ambient energy fraction λ
- Primary localization vector $\mathbf{d_P}$
- Ambient localization vector $\mathbf{d_A}$

Temporal and frequency resolution of the spatial analysis

Parametric spatial audio coding methods all rely on two assumptions which justify the choice of the frequency and temporal resolution of the extraction of the parameters. These assumptions are motivated by psychoacoustic results.

The first assumption is that the finest temporal resolution at which the auditory system can track binaural localization cues is comprised between 30 and 100 milliseconds, a phenomenon referred to as *binaural sluggishness* [HKK98, KG90]. Therefore, typically, spatial parameters are extracted around every 20 milliseconds, to ensure a finer temporal resolution than the auditory system.

Second, it is assumed that the listener cannot discriminate two simultaneous sources in space which spectrally belong to the same critical band (see section 1.2.3), while spatially arising from different locations [Pul07, BvdPKS05]. Consequently, the spatial parameters describing the scene that are transmitted along with the downmix are usually computed for each band of a set of adjacent critical bands, assuming the presence of a single auditory event in each band. This means that two spatially separated frequency components within the same critical band will share the same spatial descriptors. This second assumption is a fair approximation, since spatial discrimination ability certainly is reduced for small frequency differences (as in the case of components within the same critical band) [Per84b]. However, note that this might result in an inaccurate representation of space since counterexamples to this assumption can be found in the literature. For instance, Perrott [Per84a] reports a performance exceeding 90% in the spatial discrimination of two simultaneous pure tones around 500 Hz within the same critical band when separated by an angle of at least 26° symmetrically about the subject's median plane.

Two types of time-frequency transform can be used to divide signals into frequency bands. If a filter-bank technique is used, the input signals are filtered with a number of narrowband filters, of which the center frequency and the bandpass width mimic human auditory resolution in terms of critical bands. A second technique is to use a short-time Fourier transform (STFT). In that case, the frequency subbands of a given channel are grouped in such a way that each resulting band has a bandwidth approximately equal to the critical band corresponding to its center frequency. STFT implies a fixed temporal accuracy at all frequencies, which do not match the frequency-dependent temporal resolution of humans, whereas filter banks allow for the choice of the adequate temporal resolution for each filter. On the other hand, STFT is better suited for low-complexity applications, such as teleconferencing. So the choice between the two methods is a trade-off between quality and complexity.

Depending on the chosen method, the computation of the parameters might slightly differ. All parametric spatial audio coding methods offer both implementations, but in the following, only the filter-bank implementation will be described. The computation is the same in each subband, so for the sake of clarity, in the following the subband index is omitted. l and c respectively represent the time and channel indices.

Diffusion analysis

With each method, the diffuse part of the input signal is estimated for later reproduction, as this is necessary to preserve a faithful spatial imagery. Non-diffuse or primary components are highly correlated between channels (such as discrete pairwise-panned sources) and represent the point-like sources of the scene. Diffuse or ambient components are uncorrelated between channels, and represent less localized content, such as reverberation and wide sources. A parameter representing this diffusion is transmitted.

In BCC, as the approach is to carry inter-channel differences, and based on the fact that IC (see section 1.4) is the main cue related to the perception of diffuse sounds, the *Inter-Channel Coherence* (ICC) parameter, c, is computed between each pair of channels $\tilde{x}_1(l)$ and $\tilde{x}_2(l)$:

$$c_{12}(l) = \max_d |\Phi_{12}(d,l)|, \tag{2.22}$$

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

$$\Phi_{12}(d,l) = \frac{p_{\tilde{x}_1\tilde{x}_2}(d,l)}{\sqrt{p_{\tilde{x}_1}(l-d_1)p_{\tilde{x}_2}(l-d_2)}}, \tag{2.23}$$

where

$$\begin{aligned} d_1 &= \max\{-d, 0\} \\ d_2 &= \max\{d, 0\}, \end{aligned} \tag{2.24}$$

$p_{\tilde{x}_c}(l)$ is a short-time estimate of the power of $x_c(l)$, and $p_{\tilde{x}_1\tilde{x}_2}(d,l)$ is a short-time estimate of the mean of $\tilde{x}_1(l-d_1)\tilde{x}_2(l-d_2)$. $c_{12}(l)$ has a range of $[0,1]$, zero meaning that the channels have no correlation within the time extent d, and a value of one meaning that they are totally correlated (although with a potential delay). In the case of more than 2 channels, and in order to reduce the bit rate, in each band at each time index only ICC cues between the two channels with the most energy are estimated and transmitted.

In DirAC, a parameter Ψ corresponding to a diffuseness coefficient is computed from the B-format input (noted as w, x, y and z signals) as:

$$\Psi(l) = 1 - \frac{\sqrt{2}\left\|\sum_{m=a_1}^{b_1} w(l+m)\mathbf{v}(l+m)W_1(m)\right\|}{\sum_{m=a_1}^{b_1}[|w(l+m)|^2 + |\mathbf{v}(l+m)|^2/2]W_1(m)}, \tag{2.25}$$

where W_1 is a window function defined between constant time values $a_1 \leq 0$ and $b_1 > 0$ for short-time averaging, and $\mathbf{v}$ is the particle velocity vector, approximated by

$$\mathbf{v} = x\mathbf{e}_x + y\mathbf{e}_y + z\mathbf{e}_z, \tag{2.26}$$

where $\mathbf{e}_x$, $\mathbf{e}_y$, and $\mathbf{e}_z$ represent Cartesian unit vectors. Diffuseness gets a value of zero with plane waves from a single direction, and reaches one in a field where there is no net transport of acoustic energy.

In SASC, a primary-ambient decomposition [Goo08] is performed to separate the primary components from the ambient components, which allows them to be analyzed separately. From this decomposition, a parameter λ corresponding to the ambient energy fraction is also computed:

$$\lambda(l) = \frac{|\mathbf{A}(l)|^2}{|\mathbf{X}(l)|^2}, \tag{2.27}$$

where $\mathbf{A}$ is the multichannel ambient signal, and $\mathbf{X}$ is the downmix signal.

Localization analysis

This part of the analysis deals with pointlike sources. Each method extracts one or more parameters in each subband determining the direction of the auditory event associated with that subband.

For BCC, based on the fact that the direction of a source is provided by the ITD and ILD cues (see section 1.4), two parameters carrying those cues are computed (see also equations (2.23) and (2.24) for notations):

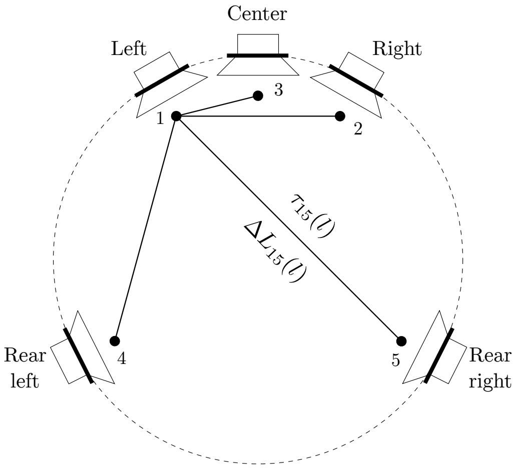


Figure 2.5: ICTD and ICLD are defined between the reference channel 1 and each of the other $C-1$ channels. (Reprinted from [Fal04].)

- Inter-Channel Time Difference (ICTD) [samples]:

$$\tau_{12}(l) = \arg\max_{d} \{\Phi_{12}(d,l)\}\,. \tag{2.28}$$

- Inter-Channel Level Difference (ICLD) [dB]:

$$\Delta L_{12}(l) = 10\log_{10}\left(\frac{p_{\tilde{x}_2}(l)}{p_{\tilde{x}_1}(l)}\right). \tag{2.29}$$

Note that the ICLD is the equivalent of the scale factors that were extracted in ISC (see section 2.5). In the case of more than 2 channels, and in order to reduce the bit rate, ICTD and ICLD are defined between a reference channel (e.g. channel 1) and the other channels, as depicted in **figure 2.5**. The relation between these parameters and the auditory system localization cues, make this representation particularly suited for binaural inputs. Indeed, binaural signals are characterized by the phase and level relationships between the left and right signals, which describe the filtering of a given source by the head-related transfer function (HRTF).

In DirAC, a single parameter is computed, the direction vector $\mathbf{D}$, defined as the opposite direction of the instantaneous intensity vector:

$$\mathbf{D}(l) = -\sum_{m=a_2}^{b_2} w(l+m)\mathbf{v}(l+m)W_2(m), \tag{2.30}$$

where W_2 is a window function for short-time averaging $\mathbf{D}$, and a_2 and b_2 are defined similarly to a_1 and b_1, respectively. Note that $\mathbf{D}$ is a *direction* vector and not a localization vector (as in SASC). Therefore, its meaning is in its angle (which is transmitted), and its magnitude is ignored.

In SASC, since each input channel has been decomposed into a primary and an ambient component, a localization vector is extracted from each of these components. The Gerzon velocity vector [Ger92] is used for the primary component:

$$\mathbf{g}_P(l) = \frac{\sum_{c=1}^{C} |\alpha_c(l)|\mathbf{q}_c}{\sum_{i=1}^{C} |\alpha_i(l)|}, \tag{2.31}$$

where α_c indicates the relative primary component in channel c, and $\mathbf{q}_c$ is a unit vector in the spatial direction of the c^{th} input channel (given by the loudspeaker input format). However, as explained in [JMG+07], the Gerzon vector magnitude is bounded by the polygon whose vertices correspond to the loudspeakers positions, which leads to inaccurate radius values and thus inaccurate spatial reproduction. Therefore, the magnitude boundary of the Gerzon vector $\mathbf{g}_P$ is expanded to the whole circle (or sphere) to give the primary localization vector parameter $\mathbf{d}_P$:

$$\begin{aligned} r_P &= \|((\mathbf{E}_{ij})^{-1}\mathbf{g}_P(l)\|_1 && (2.32)\\ \mathbf{d}_P(l) &= r_P\ \mathbf{g}_P(l)/\|\mathbf{g}_P(l)\|, && (2.33)\end{aligned}$$

with $\mathbf{E}_{ij} = [\mathbf{e}_i\ \mathbf{e}_j]$ representing the positions of the two input loudspeakers bracketing the Gerzon vector. The length r_P of $\mathbf{d}_P$ indicates the extent to which the component is pairwise panned, and therefore expresses its radial position. A length of 0 means that the component is non-directional, whereas a value of 1 corresponds to a discrete pairwise panned component. For the ambient component, the Gerzon energy vector is directly computed to give the ambient localization vector parameter $\mathbf{d}_A$:

$$\mathbf{d}_A(l) = \mathbf{g}_A(l) = \frac{\sum_{c=1}^{C} |A_c(l)|^2\mathbf{q}_c}{\sum_{i=1}^{C} |A_i(l)|^2}, \qquad (2.34)$$

where A_c indicates the relative ambient component in channel c. As stated in [GJ08], the Gerzon vector does not need to be expanded for ambient components, and therefore:

$$r_A = \|\mathbf{g}_A\|. \qquad (2.35)$$

One can conclude from this analysis that DirAC and SASC extract parameters that describe physical properties of the acoustic field, whereas the parameters of BCC are based on a perceptual description.

2.5.2 Computation of the Downmix Signal

During this process, all the C input channels are mixed together in a single monophonic signal, in order to transmit a sum signal containing all signal components of the input signals. This process is common to all methods. Note however that it can be skipped in the case of DirAC, because with the input signals being in B-format, the omnidirectional signal W is already a monophonic representation of the spatial audio scene containing all components of the directional signals X, Y and Z. W thus constitutes a downmix.

Once a time-frequency representation of the input signals is obtained, in each frequency subband, the signals of each input channel are added. As proposed by Faller in [Fal04], in order to prevent attenuation or amplification of signal components due to certain phase interactions, each subband signal $\tilde{x}_c(l)$ is multiplied by a factor $e(l)$, resulting in an equalization of the downmix signal. That is:

$$\tilde{s}(l) = e(l)\sum_{c=1}^{C}\tilde{x}_c(l), \qquad (2.36)$$

with

$$e(l) = \sqrt{\frac{\sum_{c=1}^{C} p_{\tilde{x}_c}(l)}{p_{\tilde{x}}(l)}}, \qquad (2.37)$$

where $\tilde{s}(l)$ is the equalized subband under consideration, $p_{\tilde{x}_c}(l)$ is a short-time estimate of the power of $\tilde{x}_c(l)$, and $p_{\tilde{x}}(l)$ is a short-time estimate of the power of $\sum_{c=1}^{C}\tilde{x}_c(l)$. This

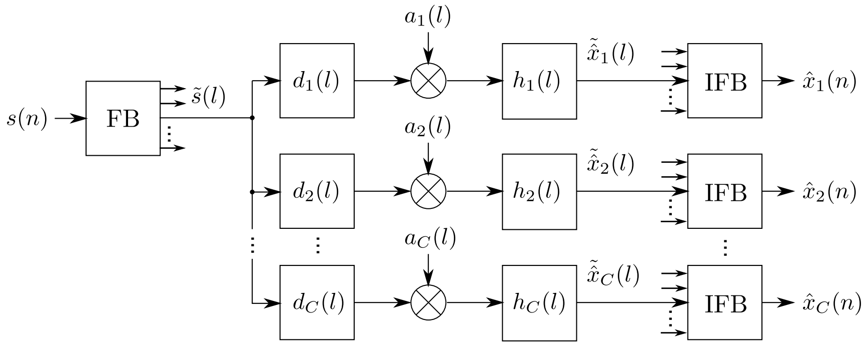


Figure 2.6: Synthesizing of spatial cues: ICTD (with delays d_c), ICLD (with scalings a_c) and ICC (with decorrelation filters h_c). (Reprinted from [Fal04].)

ensures that the power of signal components in the summed signal is approximately the same as the corresponding power in all input channels. Finally the equalized subbands are transformed back to the time domain, resulting in a summed signal $s(n)$ that is ready to be transmitted.

Specific methods exist for downmixing a stereophonic signal into a monophonic signal. They can be found in [BvdPKS05, SPSG06].

2.5.3 Spatial Synthesis

This decoding process aims to generate a set of output signals that reproduce the input sound scene. This is done using the downmix signal plus the spatial parameters. First, the downmix signal $s(n)$ is decomposed into frequency subbands (employing the same time-frequency transform used at the encoding phase), and then the idea is to distribute each subband of the downmix to its appropriate location using the spatial parameters. In the case of BCC, the output channels are an approximation of the input channels since the parameters correspond to the inter-channel differences in terms of localization cues. For DirAC and SASC, the number of output channels and their setup can differ from the input, the spatial parameters carrying a spatial representation of the scene.

For BCC, in order to recreate channels with spatial cues, the spatial synthesis is performed by applying delays d_c (for ICTD), scale factors a_c (for ICLD), and decorrelation filters h_c (for ICC) to each subband. This is illustrated in **figure 2.6**.

- ICTD:

$$d_c = \begin{cases} -\frac{1}{2}\left(\max_{2\le i\le C}\tau_{1i}(l) + \min_{2\le i\le C}\tau_{1i}(l)\right), & c=1 \\ \tau_{1c}(l) + d_1, & 2\le c\le C \end{cases} \tag{2.38}$$

  The delay for the reference channel d_1 is chosen such that the maximum delay is minimized, because minimizing the modifications in a subband leads to minimizing the risk of artifacts.

- ICLD:
  Because of the equation (2.29), a_c must satisfy $a_c/a_1 = 10^{\Delta_{1c}(l)/20}$. Also, the output subbands of reference channel 1 are chosen such that the sum of the power of all output channels is equal to the power of the summed input signals. Hence:

$$a_c = \begin{cases} 1/\sqrt{1+\sum_{i=2}^{C} 10^{\Delta L_{1i}/10}}, & c=1 \\ 10^{\Delta L_{1c}/20} a_1, & 2\le c\le C \end{cases} \tag{2.39}$$

- ICC:
  Two methods described in [Fal04] are proposed by Faller to design the filters h_c. The aim of these filters is to reduce correlation between the subbands of a given channel pair, while taking care of not affecting ICTD and ICLD. The first method is to vary ICTD and ICLD in each critical band as a function of frequency, all the while keeping the same average value. This method relies heavily on the assumption, stated earlier, that localization cues are derived over critical bands by the auditory system, and particularly assumes that ITD and ILD are derived as an average within a critical band. The second method is to add an artificial (late) reverberation to each output channel as a function of time and frequency. Moreover, when using this second method, as only the ICC value between the two most energetic channels is transmitted, ICC between the other channel pairs is produced by choosing the ratio between the power of late reverberation to sum signal $\tilde{s}(l)$ for each of the $C-2$ remaining channels to be the same as for the second most energetic channel. This way, the single transmitted ICC parameter per subband describes the overall coherence between all audio channels.

In DirAC, the downmix is first divided into diffuse and non-diffuse streams by multiplying it by $\sqrt{\Psi}$ and $\sqrt{1-\Psi}$, respectively, in each subband in order to get the energy of the original diffuse and non-diffuse components back. The diffuse signals to send to each loudspeaker are obtained by decorrelating the diffuse stream: the sound for each loudspeaker is convolved with a random sequence (this sequence is different for each loudspeaker) as explained in [Pul07]. Several methods can be used for the reproduction of point-like sources (from the non-diffuse stream), and vector base amplitude panning (VBAP) [Pul97] is one of them. Thus using the direction vector parameter, VBAP is used to compute the gain factors associated with each loudspeaker to apply to the considered frequency subband of the non-diffuse stream. This allows for a pairwise pan of each frequency component to its original location. In order to maintain a balance between the methods used for reproduction of the diffuse sound and of point-like sources, a scaling factor of $1/\sqrt{C}$ (where C is the number of output loudspeakers) is used for the random sequences used in decorrelation.

In SASC (as in DirAC), several methods can be used for the reproduction of sources. The method we will present here is to use amplitude-panning techniques. The design of the spatial synthesis of SASC is driven by a consistency requirement, which is that if the output signals were to be analyzed by SASC, they should yield the same set of parameters as those used to synthesize the scene. As explained in [JMG$^+$07], a set of panning weights $\boldsymbol{\gamma}$ bound to the output channels which satisfy this requirement can be expressed as:

$$\boldsymbol{\gamma} = r\boldsymbol{\rho} + (1-r)\boldsymbol{\epsilon}, \tag{2.40}$$

where r is the radius of the localization vector $\mathbf{d}$ (either primary or ambient, see equations (2.32) to (2.35)). $\boldsymbol{\rho}$ is a pairwise panning weight vector, which contains non-zero coefficients only for the two output channels bracketing the direction of the localization vector $\mathbf{d}$. A solution for the $\boldsymbol{\rho}$ weights can be found using VBAP [Pul97] or vector base intensity panning (VBIP) [JLP99] for primary and ambient components, respectively. $\boldsymbol{\epsilon}$ is a non-directional panning weight vector that yields a Gerzon vector of zero magnitude. This vector can be found by using an optimization algorithm [GJ06] yielding weights (called null weights) for an arbitrary loudspeaker setup. Therefore, the interpolation via r expressed in equation (2.40) reflects the distance from the listening point of the component described by $\mathbf{d}$. The ambient energy fraction parameter λ is used to separate the downmix into a primary and an ambient stream, and once the weights $\boldsymbol{\gamma}$ for both primary and ambient components are obtained, they are then used to feed the loudspeakers. In order to reproduce an accurate perception of the ambient components, however, decorrelation filters such as those proposed in [Ken95] are applied to at least some of the ambient signal channels, prior to their combination with the primary signal channels. SASC presents the peculiarity that, thanks to the primary-ambient decomposition, primary and ambient components are analyzed and synthesized separately. This allows for a directional distribution of the ambient components, which is not proposed in DirAC when using a monophonic downmix signal, as explained in [Pul07].

2.5.4 Quantization of the Spatial Parameters

The quantization (or bit allocation) process takes place prior to transmission and aims to determine the precision with which each parameter will be coded, that is, the number of bits used to represent each parameter value.

The simplest approach is to use a uniform and independent quantization of each parameter. This is the approach used in all three methods BCC, DirAC and SASC. However, as explained in [BvdPKS05], a non-uniform quantization can be used based on the sensitivity to changes of a given parameter, i.e., just-noticeable differences (JND). There exist experimental results concerning ILD, IPD and IC, which are described in section 1.4.5. The reader is invited to seek the justifications of the following in this section. Those results particularly apply to coders like BCC, which rely on inter-channel differences based on these localization cues. Hence, ICLD, ICPD and ICC can be quantized according to perceptual criteria, introducing just-inaudible quantization errors. More specifically, ICLD quantization steps are non-uniform, increasing with ICLD values. ICC quantization steps are non-uniform as well, large when coherence is around 0 (uncorrelated channels) and small when it is around 1. IPD quantization steps, however, have to remain uniform.

Concerning DirAC and SASC, however, no perceptual quantization rule of the parameters has been proposed yet. As presented in chapters 3, 4 and 5, we ran perceptual experiments based on minimum audible angles (MAA), which can be interpreted to derive quantization rules for angle parameters like those used in both DirAC and SASC, and which can be exploited in BCC as well.

2.6 Conclusions

In this chapter several spatial audio coding techniques have been presented. Matrix encoding techniques such as Dolby Surround, perhaps outdated, were appropriate for blending multi-channel information into a stereo compatible stream, constrained in that sense. Digital formats are more flexible, and especially benefit from the advantage of using time-frequency representations of the signal. Parametric spatial audio coding techniques are promising in terms of bit-rate reduction since a single audio stream is transmitted, the spatial information being parameterized.

Interestingly, the attempt to reduce the bit rate without affecting the rendering quality has not been yet fulfilled, according to subjective listening tests recently conducted by the European Broadcasting Union (EBU) [KMMS07]. Indeed, when operated at their appropriate bit rate, the parametric spatial audio coding method MPEG Surround (96 kbits/s with HE-AAC) obtains a “good” score on average, whereas Dolby AC-3 (448 kbits/s) is scored as “excellent”. Furthermore, MPEG Surround presents issues with specific material, such as applause sequences, whereas Dolby AC-3 is more consistent over content. Still according to these tests, matrix encoding techniques like Dolby Surround Pro Logic (in its most advanced form) are outperformed by digital techniques like Dolby AC-3 or MPEG Surround.

Perceptual coding, as it is used to code monophonic signals, relies on interference between frequency components, exploiting energy masking phenomena. This concept has been extended to space by taking into account binaural unmasking phenomena and deriving masking rules depending on the relative spatial location of the components. Unfortunately, it resulted in a decrease in the masking ability of the signal components and therefore in the reduction of the coding gain. However, interference between components in terms of their spatial perception (rather than their audibility) has not been considered for perceptual coding yet, although it can potentially lead to a considerable coding gain. Indeed, as presented in chapter 3, interference between frequency components reduces auditory spatial resolution, a phenomenon which is modeled in chapter 4, and exploited in perceptual coding techniques proposed in chapter 5.

Chapter 3

Spatial Blurring: Auditory Spatial Resolution in the Presence of Distracting Sound Sources

This chapter presents our psychoacoustic studies dealing with localization blur and spatial blurring. The motivations for this work are presented first, after which the common points of all experimental protocols are described before going into the details of each experiment separately. Indeed, the experimental protocols used for these experiments are very similar, and therefore a specific choice is assumed to remain unchanged from one experiment to the next, unless otherwise stated. Four psychoacoustic experiments were carried out to address the effect of different variables. First in section 3.6 the actual increase of localization blur associated with a target sound source in the presence of a single distracting sound source is studied. In this experiment, the effect of the target and distracter center frequencies is tested. In a second experiment presented in section 3.7, the effect of the relative sound level of the target and the distracter on spatial blurring is assessed. The effect of the distracter position on spatial blurring is addressed both in the third and fourth experiments (sections 3.8 and 3.9, respectively), although the main focus of the fourth experiment is the interaction between multiple distracters.

The effect of each factor at stake is studied in this chapter, but only in terms of significance. Effect sizes are not considered at this point, because this particular aspect is dealt with in chapter 4.

3.1 Motivations

Even if the studies presented in this chapter are interesting from a purely psychoacoustical point of view, this work is first motivated by the spatial audio coding aspects presented in chapter 2, aiming at bitrate reduction. However space is represented by a given coding scheme, in practice and in order to ensure a reasonable transmission bit rate, the precision of this representation is necessarily limited by some quantization process(es). This induces quantization errors in the position coding of the "components" of the sound scene which, at listening phase, are perceived as changes in position. These spatial distortions are time-varying and result in a general spatial instability. The concept of localization blur (introduced in section 1.4.5) precisely characterizes the sensitivity to position changes of sound events, and in order to be as undetectable as possible spatial distortions should thus be held within the actual localization blur. Consequently, localization blur is an adequate descriptor to drive the quantization process involving the spatial representation of the sound scene, a strategy which is discussed in chapter 5.

Localization blur has been extensively studied (see section 1.4.5) for a target source in quiet, that is without any interfering source. However, as the scene gets more complex, localization errors seem to increase [BH02], and more interestingly, sensitivity to localization cues decreases [HR10]. This suggests that auditory spatial resolution is likely to deteriorate as well when multiple sources are present, thus motivating the assessment of localization blur in the presence of distracting sources. Besides, very recent work confirms an increase of the localization blur in such situations [CG10, KT03], especially when attention is divided [MGM11].

This chapter is thus dedicated to the assessment of auditory spatial resolution via measures of localization blur. We chose to focus our work on *angular* localization blur, and more specifically in *azimuth*. This choice is motivated by the practical fact that most of the mass-market sound reproduction systems are only two-dimensional. Nevertheless, this work could be extended to elevation as well. Radial localization blur is not considered either, because the encoding of distance cues (see section 1.4.5) is generally not responsible for any bitrate increase. [1] The underlying aim is to gather a sufficient amount of data to build a model of localization blur that describes auditory spatial resolution in a maximum number of conditions, and in particular in the presence of distracting sounds. The design of such a model is addressed in chapter 4.

From an audio coding point of view, at the sound reproduction phase the listener is assumed to be free to move his or her head to face any direction within the reproduction system. As auditory spatial resolution is the finest for sources in front of the listener (see section 1.4.5), only localization blur associated with frontal sound sources is investigated in the following experiments, which constitutes the worst-case scenario in terms of audio coding.

In all of our experiments, localization blur is assessed using real sources only, meaning that no phantom sources are produced, because it has been shown that when using spatialization techniques, localization performance differs compared to when real sources are used, and in particular it depends on the spatialization technique and the speaker layout employed [MPM07, MPM].

3.2 Terms and Definitions

"Localization blur" is a generic term designating the smallest separation between two positions of the same sound event that the auditory system is able to discriminate. It can either be expressed as an angle (e.g., azimuth, elevation) or a distance (e.g., a radius). Localization blur is dependent on the characteristics of the considered sound event (frequency content, level, position, etc.), and thus is expressed *for* this specific sound event (which is usually referred to as "the target" in an experimental protocol). As we will see, it also depends on the condition in which this sound event is presented to the listener, such as the other components' frequency content or energy, but also the characteristics of the listening room, etc. Therefore localization blur is expressed *under* a specific condition. When expressed as an angle (either in azimuth, elevation, or even a combination of both), localization blur is usually assessed through a *measure* called "minimum audible angle"

[1] As opposed to the increase in number of channels necessary to improve the spatial resolution either in azimuth or in elevation. However, component position radii are specifically coded in SASC for instance (see section 2.5).

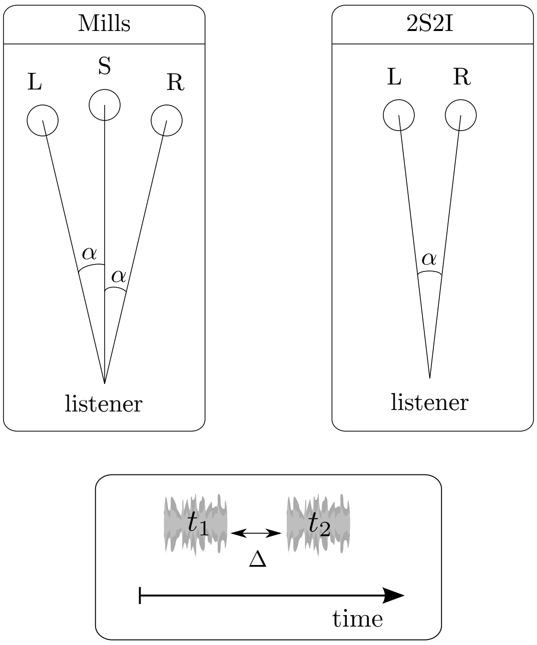


Figure 3.1: Top left: the experimental paradigm used by Mills [Mil58]. The sequences for the targets t_1 and t_2 presented to the subject are S-L (standard-left) and S-R (standard-right). Top right: the experimental paradigm proposed by Hartmann and Rakerd [HR89], and which we used in our experiments. The sequences of targets presented to the subject are L-R (left-right) and R-L (right-left). Bottom: the temporal organization of a trial.

(MAA), which is described in section 3.3. We define the "auditory spatial resolution"—under a given condition—as the localization blur associated under that condition with any given sound event. When considering only angular localization blur, the auditory spatial resolution can thus be seen theoretically as the infinite set of MAAs describing all possible sound event characteristics. In particular, we will show in the following that the auditory spatial resolution deteriorates in the presence of distracting sound sources. As a result, a distracting source indeed exerts what will be referred to in the following as a "spatial blurring" of the target source. Spatial blurring is thus the additional localization blur observed in the presence of distracters compared to localization blur in quiet.

3.3 Paradigm for MAA Assessment

As first proposed by Mills [Mil58], the minimum audible angle is the smallest angle between two positions of the same sound event that the auditory system is able to discriminate. To assess MAAs, Mills proposed the paradigm illustrated in the top left of **figure 3.1**. A listener is presented with a sequence made of two target sounds separated by a silence. The two sounds share the same characteristics except for their angular position: for each trial, the first target is played at the standard position (noted S, which is right in front of the listener), and the second target is played separated from the first one with an angle α, either leftward (position noted L) or rightward (position noted R). The task of the listener is to indicate if the sequence heard corresponded to a standard-right or a standard-left displacement, following a two-alternative forced-choice (2AFC) procedure. The experimenter assumes that at each trial listeners are trying to discriminate L from S, or R from S, and thus that he is testing the sensitivity of listeners to the angle α.

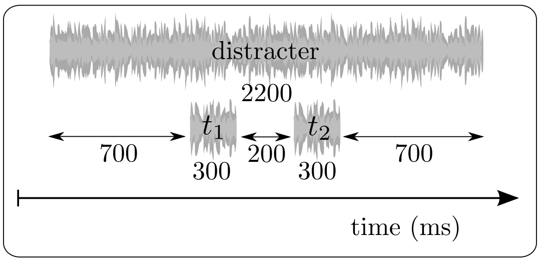


Figure 3.2: The temporal organization used in our experiments.

However, as mentioned by Hartmann and Rakerd [HR89], the listener may learn to identify the left and right source positions absolutely, therefore ignoring the standard sound (identification task), instead of trying to discriminate the standard position from the second target position at each trial (discrimination task). In such a case, the tested angle would then be 2α instead of α. The choice of using one strategy or the other is up to the listener, and Hartmann and Rakerd especially showed that Mills' paradigm seems to be performed by listeners as an identification task. To avoid this ambiguity, we decided to follow their proposed "two sources, two intervals" (2S2I) paradigm instead, which is illustrated on the top right of **figure 3.1**. This procedure is similar to that of Mills, but without the standard source between the left (L) and right (R) sources. This way, the two types of sequences the subject is presented with are left-right and right-left, instead of standard-left and standard-right.

As the point of our experiments is to assess localization in the presence of distracting sound sources, in most of the experimental conditions a distracting sound is played during the target sequence, as shown in **figure 3.2**. On a given trial, the distracter was played first and held throughout the whole trial. Its onset was followed after 700 ms by the onset of the first target of 300 ms, then followed after 200 ms by the onset of the second 300 ms target. When several distracters were present, they were played synchronously, following this scheme. The subject's task was to indicate, for each trial, whether the sequence made of the two target sounds was a left-right or a right-left sequence, using two keys on a regular keyboard. No feedback was provided to the listener.

3.4 Stimuli

The choice of stimuli is justified by experiment 1, where the effect of the relative frequency difference between the target and the distracter is investigated, implying the use of narrowband stimuli. The same stimuli were kept for the other experiments as well, in order to facilitate the comparison between experiments.

The stimuli were white noises filtered through a single ERB band, according to Glasberg and Moore's recommendations [GM90]. These filters, which simulate the cochlear filtering of the input sound into critical bands (see section 1.2.3), were implemented using the gammatone filters proposed by Patterson *et al.* [PRH+92]. The choice of critical bandwidth stimuli results from that parametric spatial audio coders compute a set of spatial parameters in critical bands (see section 2.5), and also from the fact that the integration of sound pressure level performed by the auditory system also seems to correspond to processing by auditory filters one critical band wide (see section 1.2.3). Target and distracting stimuli only differ by their duration: 300 ms for targets, and 2200 ms for distracters, including raised cosine rise and decay ramps of 5 ms. The stimuli are then characterized by a single parameter: the center frequency of the gammatone filter. For the targets, three center frequencies were considered in total for all of the experiments: 700 Hz, 1400 Hz, and 5000 Hz. These three values were chosen by following previous results in sound localization and localization blur (see section 1.4.5), and according to the three main ways binaural cues are supposedly combined by the auditory system in the duplex theory: ITD only at low frequencies, ILD only at high frequencies, and both at mid frequencies.

Besides each target stimulus, a group of distracting stimuli were generated. Each distracter of a group is then linked to its "reference target" in the following way: the center frequency of each distracter of the group was chosen in order to get a distracter located a certain number of adjacent critical bands (-4, -3, -1, 0, 1, 3, or 4) away in frequency from the reference target. The characteristics of the stimuli used for all experiments are summarized in **table 3.1**, where the rows describing the reference targets are in boldface. When a distracter was generated using the same center frequency as its reference target (that is for a distracter 0 critical bands away from the target), they shared the same characteristics.

Each stimulus (target or distracter) was generated ten times using independent white noise samples. At playback, one of these ten samples was picked randomly, which prevented the listener from learning the stimuli; this would have helped to recognize potential spectral characteristics for each loudspeaker, which is to be avoided. This solution is practically equivalent to the generation of stimuli at playback time.

Prior to each experiment, the stimuli were scaled in amplitude in order to produce an equal perceptual loudness using the following process. Using the same setup, in the same listening conditions as for the considered experiment, and using the center loudspeaker (located at 0°, right in front of the listener's head), listeners distinct from those who participated in the actual experiment were asked to equalize in loudness each distracter stimulus with its reference target stimulus played at 52 dBA, using an adjustment method. This was necessary to remove the effect of loudness from the results. Six listeners participated in the equalization process of experiment 1, seven in experiment 2, and ten in experiments 3 and 4. The resulting average equalization gains are reported in **table 3.1**.

3.5 Subjects, Rooms and General Setup

For all experiments, the recruited subjects had to pass a complete audiogram before taking part in the actual experiment, to ensure that they had normal hearing. We also made sure that their professional occupations were not related to the sound or hearing areas. All subjects were paid. A different group of subjects took part in each experiment. Eleven subjects (six females, five males) participated in the first experiment, ten subjects (six females, four males) in the second, twelve subjects (six females, six males) in the third, and finally fifteen subjects (six females, nine males) in the fourth experiment. No subject participated in more than one of the experiments. All subjects in a given experiment performed all the experimental conditions, according to a *repeated-measures* scheme.

Two different testing rooms were used in these experiments. Experiment 1 and 2 were run in the Perceptual Testing Lab at the *Centre for Interdisciplinary Research in Music Media and Technology* (CIRMMT)[2], McGill University, in Montreal, Quebec, Canada. The room is a 4.6 m × 5.6 m × 3.6 m (h) semi-anechoic room, acoustically treated on the walls and the ceiling with foam wedges, and fully covered by carpet on the floor. The subject sat in a chair. Experiments 3 and 4 were run in the anechoic chamber of Orange

[2] Thanks are due to Yves Méthot, Julien Boissinot and Harold Kilianski at CIRMMT, and to Matthieu Berjon at Orange Labs for their great help with the technical setup of these experiments.

Center	ΔERB	ERB (Hz)	Loudness equalization gain (dB)			
frequency (Hz)			Exp. 1	Exp. 2	Exp. 3	Exp. 4
374	-4	47	-1.2	-	-	-
605	-1	90	-4.7	-	-	-
700	**0**	**100**	**0**	**-**	**-**	**-**
806	1	112	-0.4	-	-	-
1202	4	155	-0.6	-	-	-
828	-4	114	2.7	-	-	1.6
949	-3	127	-	-	-	1.9
1233	-1	158	2.5	-	-	-
1400	**0**	**176**	**0**	**0**	**0**	**0**
1586	1	196	-0.7	1.2	-	-
2024	3	243	-	-	-	-2.0
2281	4	271	-3.5	0.7	-2.55	-4.0
3164	-4	366	-1.0	-	-	-
4464	-1	507	0.2	-	-	-
5000	**0**	**565**	**0**	**-**	**-**	**-**
5597	1	629	-0.3	-	-	-
7829	4	870	0.7	-	-	-

Table 3.1: Center frequency and bandwidth of the stimuli used in all experiments, as well as the loudness equalization gain applied to each of them. Within each of the three groups of stimuli, the boldface row highlights the characteristics of the reference target, the other rows describing the distracting stimulus. ΔERB is the frequency separation between a distracter and its reference target, expressed in critical bands (ERB-rate).

Labs[2] (France Telecom R&D) in Lannion, France. This room is 7.6 m × 8.6 m × 8 m (h), and is acoustically treated on the walls, the ceiling and the floor with mineral wool. In this room, the subject sat in a chair vertically located halfway between the floor and the ceiling.

It has been shown that small head movements might improve localization performances [WK99]. Therefore, to ensure realistic listening conditions, listeners were instructed to face the 0° direction but their head was not restrained (see appendix A for the exact instructions that were given to the subjects).

The whole software used for running the following experiments was developed from scratch for this purpose. It consisted of a `puredata` patch for audio I/O and user interactions, used in conjunction with Java code handling the experimental procedure. The communication between `puredata` and Java was ensured by `pdj` [Gau10]. A screenshot (of experiment 4) is given in **figure 3.14**.

3.6 Experiment 1: Spatial Blurring From One Distracter[3]

The primary goal of this experiment was to show that the localization blur associated with a target sound source increases in the presence of a single distracting sound source. As a secondary goal, the effect of the target center frequency was assessed, as well as the frequency separation between the distracter and the target.

[3] This experiment was carried out in collaboration with Catherine Guastavino and Georgios Marentakis from McGill University.

Speaker number	Position (°)	Speaker number	Position (°)
1	−12.50	12	0.35
2	−10.51	13	1.05
3	−8.71	14	1.75
4	−7.09	15	2.45
5	−5.65	16	3.33
6	−4.40	17	4.40
7	−3.33	18	5.65
8	−2.45	19	7.09
9	−1.75	20	8.71
10	−1.05	21	10.51
11	−0.35	22	12.50

Table 3.2: Azimuth position of target speakers in experiment 1.

3.6.1 Setup

A schematic of the experimental setup is represented in **figure 3.3c**. Due to the room size constraints, we had to use very small handmade loudspeakers in order to reach a spatial resolution of 0.7° in front of the listener. Therefore, we used twenty-three Phoenix SE transducers from Harman International (which supported this work), made of a 45 mm diaphragm and enclosed in 120 mm × 104 mm × 54.6 mm extruded aluminum boxes, with inner surfaces covered with foam. The target speakers were positioned according to **table 3.2**, spaced in a linear increasing fashion. The distracter loudspeaker was positioned at 0° in azimuth and 0.9° in elevation (because of the presence of the target speakers, see **figure 3.3b**). The target and distracter speakers output levels were adjusted in order to produce 52 dBA at the center of the listener's head when fed with one of the target stimuli. A soft overall level (52 dBA) was chosen for the experiment to be painless over time, especially in the presence of an additional distracter.

Because of their small diameter, at high input level, the target speakers tend to add harmonic distortions to the original signal. We verified that these distortions were below hearing threshold. This is necessary because undesired frequency components can first add unintended localization cues to the stimuli, but could also deepen each speaker's characteristics and thus help listeners in their task by allowing them to identify each speaker separately.

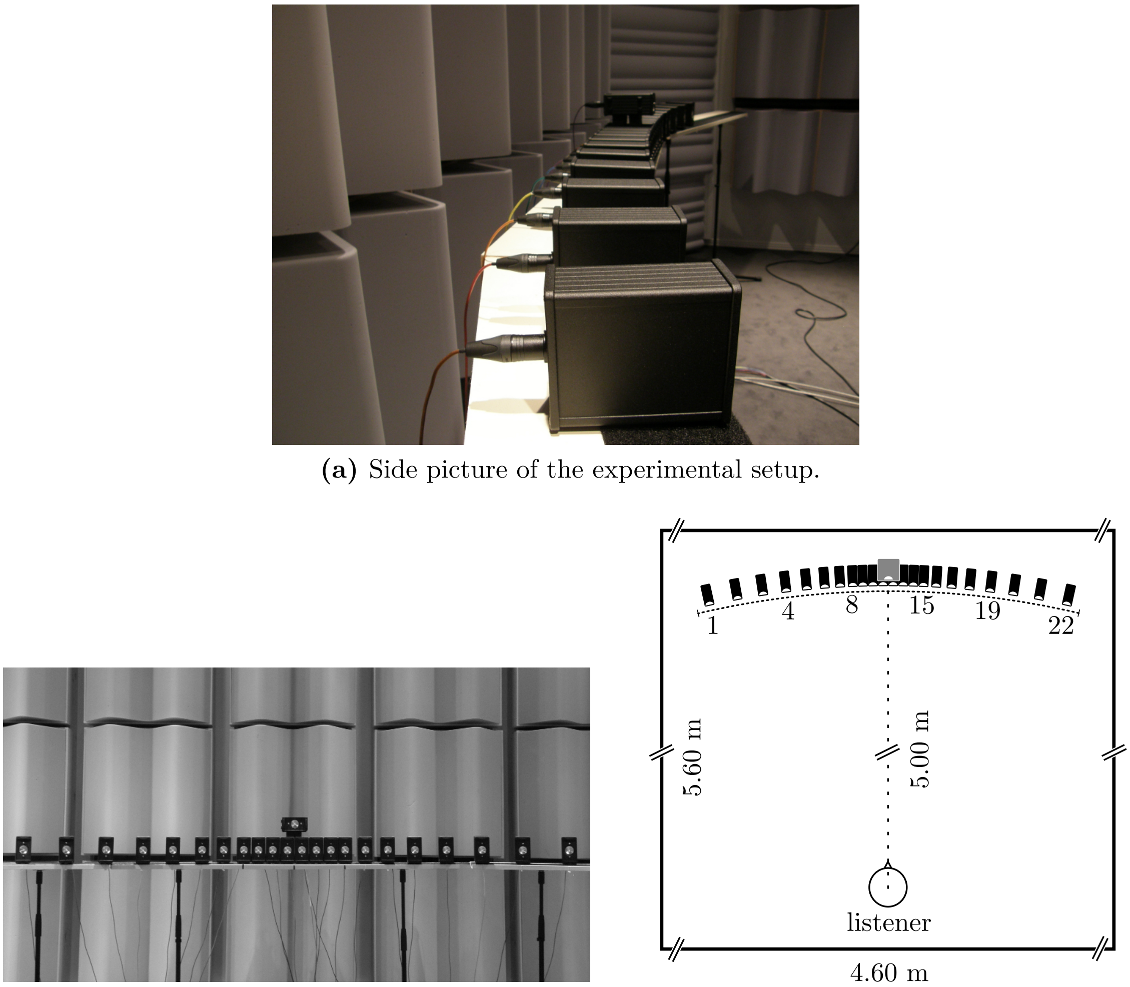
3.6.2 Procedure

A given experimental condition is defined by:

- the target center frequency: either 700 Hz, 1400 Hz, or 5000 Hz;
- the distracter center frequency relative to the target, expressed in critical bands: -4, -1, 0, 1, 4 or "no distracter" (see **table 3.1**).

Hence, each subject was tested in 3 × 6 = 18 conditions. When present, the distracter was always located at 0°, in front of the listener. The variable is the angular separation between the two targets (following the MAA task paradigm presented in section 3.3).

Depending on the tested angular separation, the two targets were played through different speakers, according to **table 3.3**. When the separation was asymmetrical (that is, when the distracter could not be exactly in the middle of a pair of speakers), two pairs



(a) Side picture of the experimental setup.

(b) Front picture of the experimental setup, from the subject's point of view. The loudspeakers were placed on a curved wood plate. The raised loudspeaker was used for the distracter.

(c) Top view of the experimental setup (see **table 3.2** for the exact target speakers positions). The gray speaker at the center is the distracter speaker. Its misalignment with the two most centered target speakers is for drawing purposes only.

Figure 3.3: Experimental setup for experiment 1.

Separation number	Angular separation (°)	Pair(s) of speakers
1	**0.7**	**11,12**
2	1.4	10, 12 & 11, 13
3	**2.1**	**10,13**
4	2.8	9, 13 & 10, 14
5	**3.5**	**9,14**
6	4.2	8, 14 & 9, 15
7	**4.9**	**8,15**
8	5.8	7, 15 & 8, 16
9	**6.7**	**7,16**
10	7.7	6, 16 & 7, 17
11	**8.8**	**6,17**
12	10.1	5, 17 & 6, 18
13	**11.3**	**5,18**
14	12.7	4, 18 & 5, 19
15	**14.2**	**4,19**
16	15.8	3, 19 & 4, 20
17	**17.4**	**3,20**
18	19.2	2, 20 & 3, 21
19	**21.0**	**2,21**
20	23.0	1, 21 & 2, 22
21	**25.0**	**1,22**

Table 3.3: Angular separations used in experiment 1, and the involved pair(s) of speakers. The symmetrical separations are represented in boldface characters.

of speakers were used in a balanced random order, in order to counterbalance the effect of this asymmetry.

We chose to use the *method of constant stimuli* for this experiment. In this method, the psychometric curve of performance of a given subject is assessed by presenting repeatedly a fixed set of angular separations (an example can be seen in **figure 3.4**), assuming that the non-asymptotic part of the curve (the part between 60% to 90%) is comprised in that fixed set. However, in a pilot experiment, we noticed that subjects varied greatly in their performance, meaning that it was not possible to present the same set of angular separations to each subject[4] (see appendix B for an illustration of this point). Therefore, a specific protocol was designed to adjust the set of tested angular separations until finding the proper area of testing for a given experimental condition. It was organized in blocks of trials. Within a block, the condition parameters (that is, the target frequency and, if present, the relative distracter band) remained fixed, and a given set of angular separations was used. The order of presentation of the separations was randomized within a block. The number of left-right and right-left sequences was balanced for each separation. Hence, the method of constant stimuli was used within each block.

As schematized in **figure 3.5**, for a given experimental condition, the subject was first presented with a training block, made of seven angular separations, covering from the smallest separation (0.7°) to a very large one (21°), each separation being presented 12 times. Then another block made of the same parameters (target center frequency, relative distracter band, set of angular separations, and number of repetitions) was presented to the subject, and used as an "overview" block. The goal of this block was to get a rough

[4] Otherwise, the number of angular separation per set necessary to ensure an accurate estimation of the curve would have been too great, increasing the duration of the experiment.

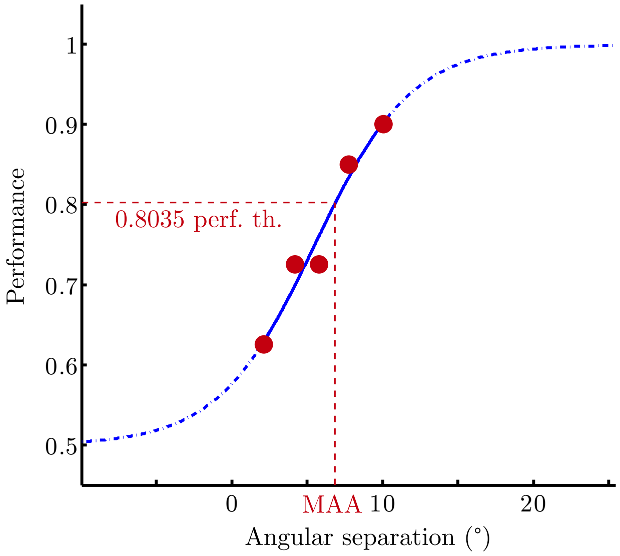


Figure 3.4: Example of curve fitting in experiment 1, for subject 1 (target center frequency: 5000 Hz, relative distracter band: 0).

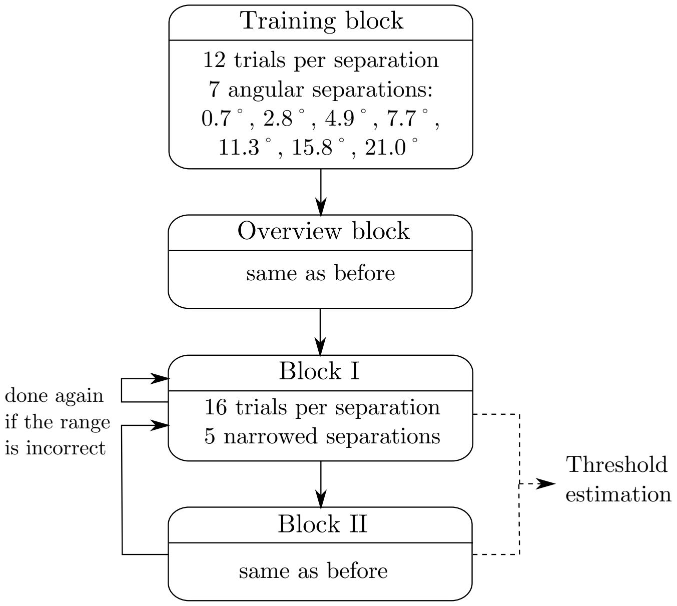


Figure 3.5: The experimental protocol of experiment 1 for a given condition.

idea of the range of separations necessary to obtain a good estimation of the subject's psychometric curve of performance. Thus five separations were chosen for the next block by narrowing the tested range of angular separations. The number of repetitions was also increased to 16 per separation. From this point on, each time the tested range was adjusted, the corresponding five angular separations were yielded by an automatic procedure. This procedure ensured that the proposed angular separations were distributed as uniformly as possible over the tested range. The most informative part of a psychometric curve is between performances of 60% and 90%, thus ideally, we want the smallest separation of the set to give about 60% correct answers, and the largest one about 90%. So the new range of angular separations was chosen by the experimenter from the overview block in order to approach this requirement. Once executed, if the range indeed followed these expectations, another block made of the same parameters was presented once again. If not, the tested range was adjusted and a new block was executed. Thus, the tested range is adjusted until two successive blocks gave consistent results. Then the performances from these two blocks were averaged, meaning that each data point is an average value over 32 repetitions. The coefficients of the underlying psychometric curve of performance were estimated by fitting a logistic function to the data points using `psignifit` [Hil07], a Matlab toolbox which implements the maximum-likelihood method described in [WH01]. The MAA was then extracted as the angular separation corresponding to a performance of 80.35% [5] on the resulting psychometric function. An example of this process is illustrated in **figure 3.4**.

The whole experiment in itself was organized in the following way. The different conditions were presented grouped by target frequencies, and the resulting sets were presented in a different order for each subject. As three target frequencies were tested and twelve subjects participated, each possible order of target frequency set has then been presented to two different subjects. For each subject, within each of these sets, the six distracter bands relative to the target frequency (including the case without distracter) were ordered randomly. Concretely, for each subject, the experiment was divided into five sessions of one and a half hours each.

3.6.3 Data Analysis and Results

In addition to the eleven subjects who successfully completed this experiment, one was discarded because of an inability to do the task in too many conditions (7 out of 18), and an overall low level of performance compared to the other subjects. Using Grubbs' test, 5 outliers out of 198 samples (that is, 2.5% of the dataset) were replaced by the means of the corresponding group. All post-hoc tests were corrected for multiple comparisons using Holm-Bonferroni correction. Unless otherwise stated, non-parametric tests (Wilcoxon's signed-ranks test and Cuzick's test) yielded results similar to those given in this section.

Figure 3.6 gives the mean thresholds across all participants. From this plot we can observe three points, which we will test statistically:

1. the presence of the distracter seems to globally increase the MAA (compare for instance, on any of the three plots, the "no distracter" case with a distracter in the same critical band as the target);

2. the performance level seems to depend on the target center frequency;

[5] The choice of this threshold value, which usually is 75%, is motivated by constraints on this value in the next experiments, as explained in section 3.7.4. Therefore the same threshold value has been chosen for all experiments to allow straightforward comparisons between them.

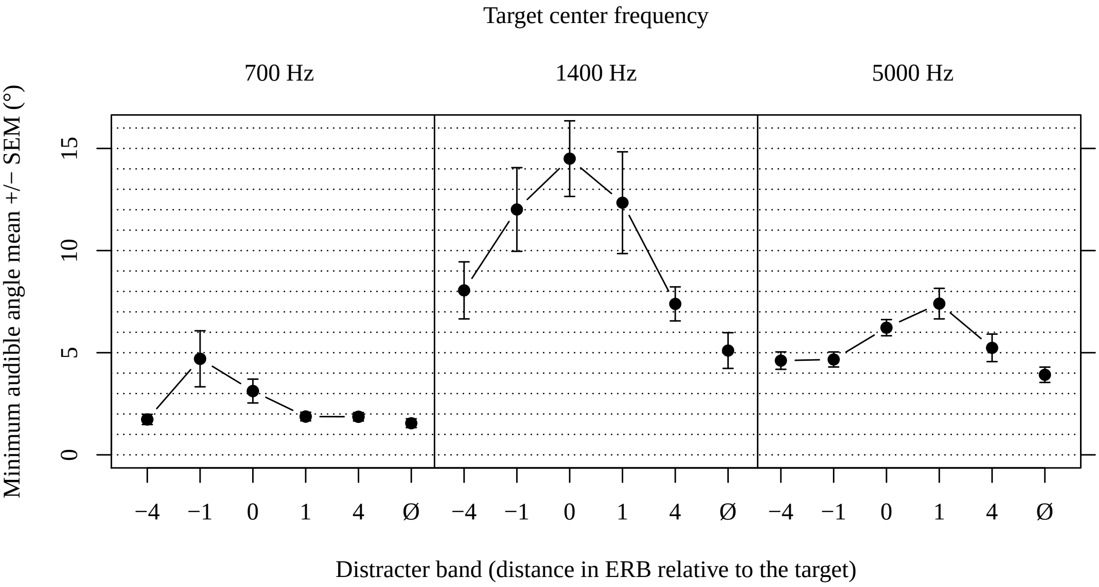


Figure 3.6: The mean thresholds across all participants for experiment 1. The vertical bars represent ± the standard error of the mean. ∅ indicates the condition with no distracter.

3. the effect of the frequency distance between the target and the distracter seems to differ from one target frequency to another (the respective shape of each of the three plots is different);

Let us begin with the last observation. If indeed this point is true, it would mean that there exists an interaction between the target frequency and the distracter band, which would oblige us to test the effect of the distracter separately for each target frequency. Partly to test this hypothesis, we carried out a two-way factorial repeated-measures ANOVA, the target center frequency (TF) and the relative distracter band (DB) treated as independent variables (fixed factors), and the MAA being the dependent variable. It returned a significant effect for TF ($F(2, 20) = 26.2$, $p < .001$), DB ($F(5, 50) = 13.6$, $p < .001$), as well as for their interaction ($F(10, 100) = 3.9$, $p < .01$).

The significant effect of TF suggests that our second observation is true: two TF groups at least have different group means. Paired t-tests allow us to confirm that all the three group means are different from each other ($p < .001$ in all cases). This result is in accordance with those of Mills [Mil58] and Boerger [Boe65] (see section 1.4.5): localization blur is the smallest for a low-frequency target, larger for high frequencies, and the largest for mid frequencies. Note, however, that when comparing all the "no distracter" (∅) DB groups between them, no significant effect is found between ∅ at 1400 Hz and ∅ at 5000 Hz (whereas $p < .01$ and $p < .001$ when comparing ∅ at 700 Hz with ∅ at 1400 Hz and at 5000 Hz, respectively).

The significant effect of the interaction between TF and DB confirms our third observation stated above: the DB effect differs between at least two of the three TF groups. Therefore, we had to test the effect of DB separately for each TF group. This also implies that not much can be drawn from the significant effect of DB reported by the ANOVA.

We can now focus on our first observation, which is the main question of interest here: does the presence of the distracter increase localization blur? To answer this question, within each TF group and using one-tailed paired t-tests, we looked for a significant mean difference by testing the ∅ DB group against the DB group which produces the greatest mean variation of MAA, that is, -1, 0 and 1 for the 700 Hz, 1400 Hz and 5000 Hz TF groups, respectively. In all cases, this difference is significant ($p < .05$, $p < .01$ and $p < .001$ for the 700 Hz, 1400 Hz and 5000 Hz TF groups, respectively). Therefore, we can answer the question posed above by saying that the presence of the distracter does increase localization blur, and does so for the three target center frequencies we considered in this experiment. This does not mean, however, that the presence of the distracter had a significant effect for all DB.

The effect of DB—that is, the way the relative band of the distracter affects the increase of localization blur—has been assessed using quadratic trend analyses within each TF group. As a result, a significant quadratic trend was found for all TF groups ($p < .01$ for the 700 Hz TF group, and $p < .001$ for the 1400 Hz and 5000 Hz TF groups). These results confirm the effect of DB within each TF group. Now, the specific asymmetric shapes within the 700 Hz and 5000 Hz TF groups are also confirmed by significant linear trends when considering only the -1, 0 and 1 DB groups ($p < .05$ for the 700 Hz group [6] and $p < .001$ for the 5000 Hz group). Note that for the 1400 Hz group, there is no significant quadratic trend anymore (even without correction) when considering only these DB groups (-1, 0 and 1), which suggests that these three data points are not significantly different from each other. These last results confirm that the dependence of spatial blurring upon the distracter relative center frequency varies with the target center frequency. In other words, there is an interaction between the distracter relative center frequency and the target frequency.

As a conclusion, we can restate our previous observations in the following way:

1. spatial blurring phenomenon: for all the tested target center frequencies, at least one of the tested distracters induced an increase of localization blur compared to when no distracter was presented;

2. there is a global effect of the target center frequency on localization blur;

3. the effect of the frequency separation between a target and a distracter on spatial blurring is dependent on the target center frequency.

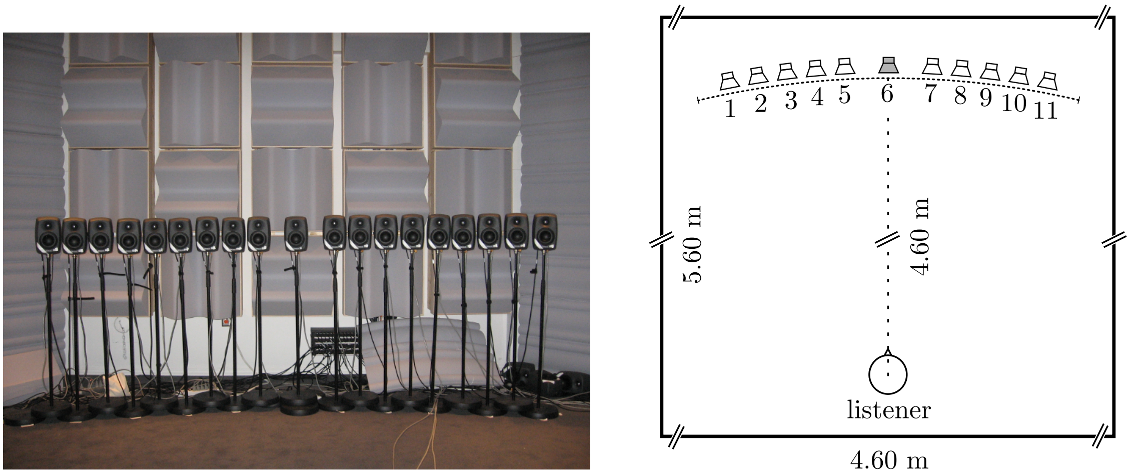
3.7 Experiment 2: Effect of the Signal-to-Noise Ratio [7]

Part of the following was published in the proceeding of the 10th french congress of acoustics (CFA) [DGM10].

The previous experiment showed the effect of the target and the distracter frequencies on localization blur. The levels of the stimuli were kept fixed such that the target was always played at 52 dBA and that the level of the distracter was perceptually matched in loudness with the target. However, from an audio coding point of view, the components of a sound scene are likely to have different levels, and thus the previous results might not apply in most cases. Therefore, this second study mainly aimed to assess the effect of the *signal-to-noise ratio* (SNR, which here is the ratio between the target and the distracter levels) on localization blur. This study also provides an assessment of the audibility threshold of the target in the presence of a distracter (or masking threshold) for several angular separations. From this experiment on, we will only test the target center frequency of 1400 Hz for two reasons: first, testing other target frequencies would increase

[6] The 700 Hz TF group did not show a significant linear trend when performing a non-parametric test.

[7] This experiment was carried out in collaboration with Catherine Guastavino from McGill University.



(a) Front picture of the experimental setup, from the subject's point of view.

(b) Top view of the experimental setup (see **table 3.4a** for the exact speaker positions). The gray speaker at the center is the distracter speaker.

Figure 3.7: Experimental setup for experiment 2.

Speaker number	Position	Speaker number	Position
1	−11°	7	3°
2	−9°	8	5°
3	−7°	9	7°
4	−5°	10	9°
5	−3°	11	11°
6	**0°**		

(a) Azimuth position of the speakers. The distracter speaker is in boldface. All speakers have a null elevation (0°).

Separation number	Angular separation	Pair of speakers
1	6°	5, 7
2	10°	4, 8
3	14°	3, 9
4	18°	2, 10
5	22°	1, 11

(b) Angular separations and their corresponding pairs of speakers. The target-distracter separation is obtained by dividing these angles by two.

Table 3.4: Loudspeaker positions and angular separations used in experiment 2.

the duration of experiments too much, and second, from experiment 1, it seems to be the most promising center frequency in terms of spatial blurring.

3.7.1 Setup

A schematic of the experimental setup is depicted in **figure 3.7b**. All speakers (Genelec 8020A) were equalized beforehand at the position of the distracter speaker (#6, represented in gray) to 65 dBA using a pink noise, with a microphone positioned three meters away. They were then positioned according to **table 3.4a** (with an elevation of 0° in relation to the listener's head). The overall level was adjusted in order to produce 52 dBA at the center of the listener's head when a speaker is fed with the target stimulus at its "base" level. That is to say, when the SNR equals 0 dB, both the target and the distracter are played at this base level (thus with a respective gain of 0 dB). This overall level was chosen in accordance with our previous experiment.

3.7.2 Procedure

The main aim of this experiment was to obtain an estimation of the MAA as a function of the SNR. In the previous experiment, we dealt with the inter-subject variability using a specific experimental protocol designed to adjust the set of tested angular separations until finding the proper range for testing (see section 3.6.2). This procedure unfortunately has two issues. First, from a practical point of view, the process of deciding the next testing range is very time-consuming. But more importantly, because the experimenter makes this decision, the process is not totally reliable, nor reproducible. Consequently, for this second experiment, we wanted to change the assessment method by using an *adaptive method*, which is described in detail in section 3.7.4. Adaptive methods present the advantage of being more flexible than the method of constant stimuli, because they can be used with subjects showing very different performance levels.

As this experiment aims to assess the effect of the SNR on the MAA, the most straightforward way to do so would be to estimate directly the MAA for several fixed SNR values. However, directly assessing the MAA in a given condition requires the discretization of space with an array of loudspeakers at fixed positions[8] (as the one used in the previous experiment). Unfortunately, the adaptive method we wanted to use requires a continuous dependent variable. This is why we chose instead to fix a set of angular separations, and to assess for each separation the SNR for which this separation is effectively the MAA. This way, the SNR becomes the dependent variable and, presenting the advantage of being continuous, an adaptive method can thus be applied to vary it and extract a threshold value.

Nonetheless, we noted that in particular situations and for large angular separations, the spatial discrimination seems to be ensured so long as the target remains audible. In such situations, the criterion of feasibility of the task is no longer the spatial discriminability, but the target audibility, and we are thus measuring the actual audibility threshold of the target. Therefore, it was necessary to distinguish these particular conditions from those for which the subject clearly hears the target but has difficulty discriminating the presented angular separation. This first remark motivates the separate measure of the audibility threshold of the target in the presence of the distracter as a point of comparison.

Second, since this measure of the audibility threshold of the target in the presence of the distracter is conducted for several angular separations between the target and the distracter, it is possible to derive the resulting *spatial unmasking* (or *binaural release from masking*, see section 1.4.6), which is a useful additional information in terms of audio coding (as explained for example in section 2.4.1). Indeed, one can expect this audibility threshold—we could equally call it masking threshold—to decrease with an increasing angular distance to the distracter.

This experiment was organized in blocks. One block consisted of an adaptive procedure yielding a threshold value for a given experimental condition. An experimental condition was made of three parameters: task type (discrimination of the direction of the change in spatial location or audibility threshold), relative distracter band (0, 1 or 4 ERB bands above the target), and angular separation (6, 10, 14, 18 or 22 degrees). The target center frequency was always 1400 Hz, and the distracter was always located at 0°, in front of the listener. As explained above, the variable was the SNR, that is to say the ratio between the target and distracter levels.[9] Using a repeated measures scheme, the experiment was thus made of 30 blocks for each subject.

[8] Unless one has access to a mechanical system allowing one to move a pair of loudspeakers in quiet.

[9] Therefore, a positive SNR means that the target level is greater than the distracter level.

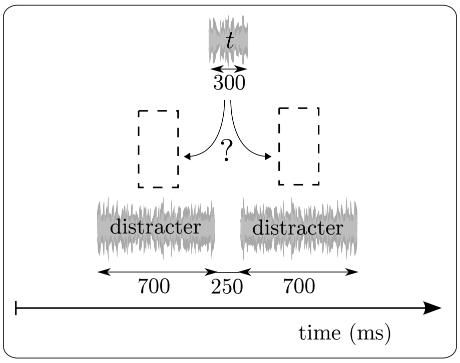


Figure 3.8: Temporal organization of a trial for the audibility task. The target sound is presented on either the first or second presentation of the distracter sound.

The order of presentation of the blocks was organized on three levels, using three nested loops: a first loop on the three distracter bands, then a second one on the two task types, and finally a third one on the five angular separations. For each subject, the distracter band loop, the three task type loops, and the six angular separation loops, were all separately randomized. In other words, a given subject did all the angular separations for a given task type and a given distracter band before moving on to the other task type of this distracter band. Moreover, he or she had to do the two task types of a distracter band before moving on to the next distracter band.

The amount of training given to each subject was very small. When switching from a given distracter band to another, each subject performed around 20 training trials, no matter what the first task type was. It means that when switching the task type within a given distracter band, no additional training was given to the subject. In practice, for each subject, the experiment was divided into three sessions of one and a half hours each.

3.7.3 Tasks

As explained above, this experiment involved two tasks for the subjects. The first task has already been described in section 3.3, and remains unchanged. We will refer to this task as the “LR/RL task”. In brief, the subject had to indicate, for each trial, whether the sequence made of the two target sounds described a displacement from the left to the right or the opposite. The SNR threshold leading a given angular separation to be a MAA was obtained using the adaptive method detailed in section 3.7.4.

As illustrated in **figure 3.8**, and following the example of the LR/RL task, this second task used a 2AFC procedure. On each trial, the subject was presented with two distracter/masker intervals of 700 ms separated by a silence of 250 ms. In one of these two intervals a target sound of 300 ms, temporally centered within the interval, was added. The target was played randomly through the left or the right speaker of the pair of speakers associated with the desired angular separation (see **table 3.4b**). The task of the subject was to indicate which of the two intervals contained the target using a keyboard. No feedback was provided about the accuracy of the answer. We will refer to this task as the “audibility task”. The SNR threshold of audibility of the target was obtained using the adaptive method detailed in section 3.7.4.

3.7.4 Adaptive Method Setup

The idea behind adaptive methods is to vary from trial to trial the stimulus intensity (the studied variable) as a function of the subject's previous answers. Depending on the rule of adaptation of the stimulus intensity which is chosen, this intensity is supposed to oscillate around a value corresponding to a certain theoretical performance of the subject. Therefore, in this method, the psychometric curve is no longer estimated, but rather a small portion of it located around a certain targeted performance point.

For each of the two tasks described above, the adjusted variable was the target level, using an adaptive adjustment method made of staircases with fixed step sizes. The distracter level was kept constant (loudness matched with the target stimulus at 52 dBA). Therefore, the SNR we will display in our results is simply the gain applied to the target stimulus by the adaptive method. Hence, when SNR equaled 0 dB, the target was played at a level which produced 52 dBA at the listening position, and the distracter was played at its loudness matched level. The staircases used in this experiment were based on a 2-down/1-up underlying rule, with unequal step up and down sizes (namely $\Delta^+ = 3$ dB and $\Delta^- = 1.6464$ dB), targeting a 80.35% performance threshold. Each block was composed of two interwoven staircases of 14 reversals each. The reasons for these choices are given below. Before running the two staircases, a preliminary phase was made of a 1-down/1-up staircase beginning way above the presumed threshold, and stopped after 3 reversals, in order to approach the threshold region quickly. Then the pair of staircases took over with the next target stimulus level. No boundaries were imposed on the target level, but in practice, the SNR values lay between 26 dB and -50 dB. The peaks and valleys of all reversals (apart from the preliminary phase) and from the two interwoven staircases were averaged to estimate the 80.35% performance threshold values of SNR. In a resulting couple {angular separation, SNR}, the angular separation can then be interpreted as the MAA we would have obtained if we were assessing the MAA associated with that fixed SNR value, instead of the reverse, as we did in practice.

We used two interwoven staircases instead of a single one. This means that for a given run, two staircases with identical parameters were run in parallel, randomly switching from one to the other. This presents the advantage of hiding from subjects the underlying rule of the staircases, which might otherwise bias their responses.

As reported by García-Pérez [Gar98, Gar00], the staircase parameters must be chosen carefully in order to ensure a reliable convergence to the targeted theoretical percent-correct. This theoretical percent-correct depends on the underlying rule of the staircase, but García-Pérez [Gar98] used simulations of staircases of several thousand trials each to show that the percent-correct targeted in reality depends also on the ratio between Δ^+ and the spread σ of the underlying psychometric function.[10] More interestingly, the level of dependence varies with the choice of Δ^+/Δ^-. In particular, $\Delta^+/\Delta^- = 1$ is the worst choice because, depending on Δ^+/σ, the percent-correct targeted in reality oscillates between 60% and 90% correct. The difficulty comes from the fact that σ is unknown and might depend on the task type, the experimental condition and the subject. Thus this variability impedes both averaging between subjects and comparing between experimental conditions. At the other extreme, depending on the underlying rule of the staircase, there is an optimal Δ^+/Δ^- ratio with which the targeted percent-correct remains constant whatever the Δ^+/σ ratio is. In our case, we chose a 2-down/1-up rule, and the corresponding optimal Δ^+/Δ^- ratio is 0.5488. By observing this ratio, our staircases

[10] The spread σ of a psychometric function is the extent over which it displays non-asymptotic behavior, measured in whichever units are relevant (in decibels in our case). More specifically, in his simulations García-Pérez defines it as the part of the curve between performances of 51% and 99%.

should effectively target a 80.35 percent-correct point.

Another point studied by García-Pérez [Gar00] is the bias induced by the fact that in practice we use small-sample staircases (as opposed to several thousand reversals). He proposes a measure of the accuracy of a staircase depending on the spread σ of the underlying psychometric curve: a given staircase setup has a “small” probabilistic error when simulations show that it targets 80% of the time a percent-correct point which deviates less than 0.32σ from its theoretical value. An interesting result of his is that as long as the optimal Δ^+/Δ^- ratio is observed, the accuracy of a staircase only depends on two parameters: the number of reversals and the Δ^+/σ ratio. According to his results, with less that 18 reversals, staircases yield biased estimations. From about 30 reversals, this bias is totally absent. Therefore, to avoid this bias, we chose to use two interwoven staircases each made of 14 reversals, reaching 28 reversals in total. The second dependence—the Δ^+/σ ratio—is intuitive and represents the dependence on the step sizes.[11] This dependence on Δ^+/σ must be differentiated from the one previously stated. By choosing an optimal Δ^+/Δ^- ratio, we ensured that our staircase targets the same percent-correct whatever the Δ^+/σ ratio equals. Here, due to the limited duration of our staircases, there might be some bias (around the targeted percent-correct) in our estimations, and this potential bias depends on Δ^+/σ as well. If Δ^+ is large compared to σ, that would lead to a large standard error around the targeted threshold. If it is too small, the duration of the experiment is uselessly long. According to García-Pérez' simulations, Δ^+ is optimal when Δ^+/σ is between 1/3 and 1, but once again σ is unknown, and might depend on the task type, the experimental condition and the subject; and so would Δ^+. However, we tried to find a single Δ^+ value suitable in all cases. To do so, we first used the method of constant stimuli with a group of separate subjects to get a very rough idea of the shapes of the psychometric curves, and derived a first value for Δ^+ by targeting $\Delta^+/\sigma = 1$. Then we started piloting with this value, and in order to refine it, we compared the block durations (in terms of trials) to the number of trials that this staircase would take if Δ^+/σ was indeed 1,[12] according to García-Pérez' simulations. The final Δ^+ value of 3 dB (leading to a Δ^- value of 1.6464 dB) gave satisfying results. According to García-Pérez results, a 2-down/1-up staircase of 30 reversals should have an average number of trials around 95 when $\Delta^+/\sigma = 1$. A larger number of trials in practice would mean that the actual Δ^+/σ ratio was smaller than 1 (and in that case the estimation is more accurate). The blocks resulting from this experiment had an average trial number of 99.82 for the MAA task and 97.97 for the the LR/RL task, with a respective standard deviation of 14.63 and 12.41 trials, which means that our threshold estimations should be accurate. Besides, the similar statistics between the two tasks show that using the same step sizes for both of them was acceptable.

3.7.5 Data Analysis and Results

In addition to the ten subjects who successfully completed this experiment, three were discarded because of their inability to do the task for certain conditions and an overall low level of performance compared to the other subjects. Using Grubbs' test, 3 outliers out of 150 threshold values (that is, 2% of the dataset) were replaced by the mean of the corresponding group for the LR/RL task, and one outlier out of 150 threshold values (0.6% of the dataset) for the audibility task. All post-hoc tests were corrected for multiple

[11] Note that because Δ^+ and Δ^- are tied by the optimal ratio described above, this second dependence can either be described by the Δ^+/σ ratio or by the Δ^-/σ ratio.

[12] This would correspond to a deviation of the estimation of about 0.30σ. We did not target a smaller value of Δ^+/σ, because it would increase the necessary number of trials.

	DB →	0	1	4
SEP	6°	·	*	***
	10°	n.s.	n.s.	**
	14°	n.s.	n.s.	*
	18°	n.s.	n.s.	n.s.
	22°	n.s.	n.s.	n.s.

Table 3.5: Pairwise comparisons of the mean differences between the two task types using paired t-tests (Holm-Bonferroni correction was applied within each DB group). Significance codes: *** $< .001$, ** $.001 - .01$, * $.01 - .05$. The dot $(\cdot)$ denotes a nonsignificant trend ($.1 > p > .05$).[13]

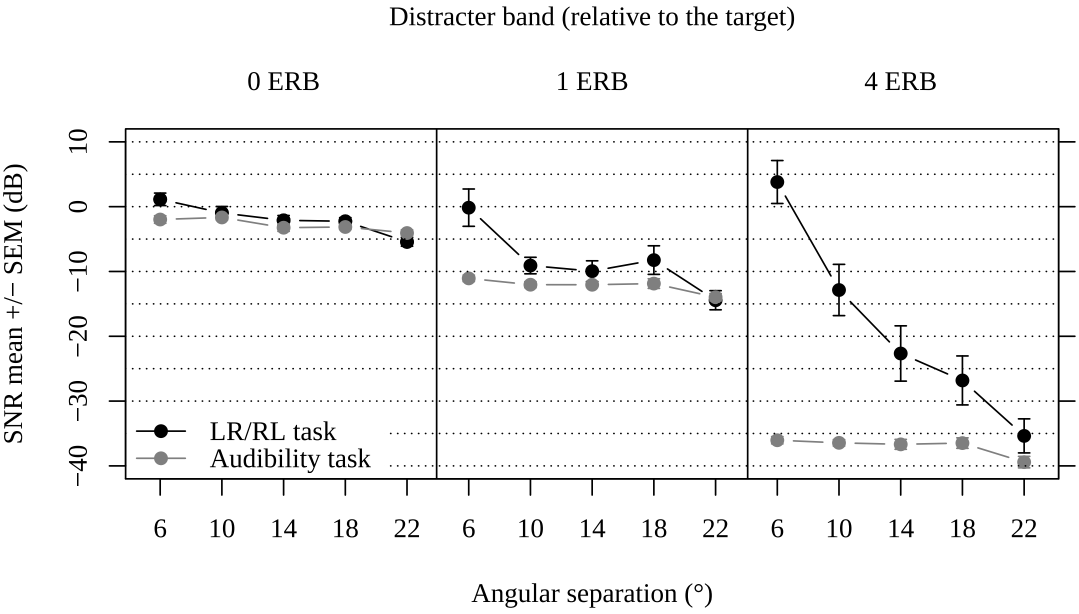
comparisons using Holm-Bonferroni correction. Unless otherwise stated, non-parametric tests (Wilcoxon's signed-ranks test and Cuzick's test) yielded results similar to those given in this section.

Figure 3.9 gives the mean thresholds across all participants, either grouped **(a)** by distracter band or **(b)** by angular separation. As expected, for each distracter band, from some given angular separation the performance in the LR/RL task seems to be limited by the audibility threshold. For the distracter band 4 ERB above the target, where the audibility threshold is relatively low, the SNR decreases with the increasing angular separation, suggesting an effect of the SNR on the MAA.

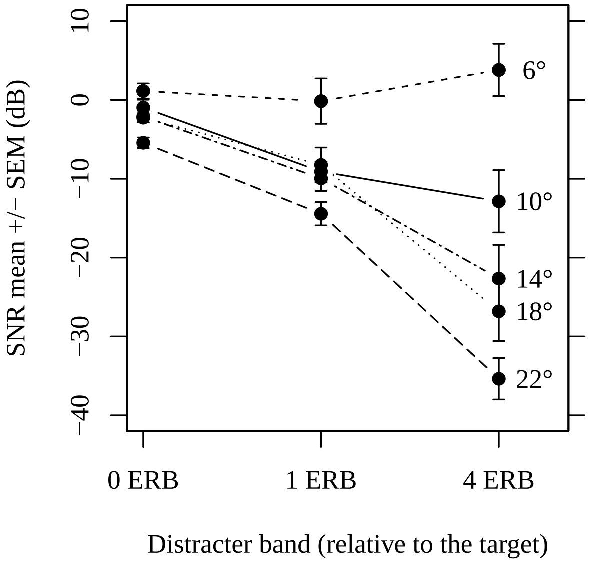
Two separate repeated-measures factorial analyses of variance (ANOVA) were carried out (one for each task type), the distracter band (DB) and the angular separation (SEP) treated as independent variables, and the SNR being the dependent variable. As a result, DB and SEP both have a significant effect for the two tasks, $F(2, 18) = 36.1$, $p < .001$ and $F(4, 36) = 58.7$, $p < .001$, respectively, for the LR/RL task, and $F(2, 18) = 1310.7$, $p < .001$ and $F(4, 36) = 24.7$, $p < .001$, respectively, for the audibility task. Their interaction is also significant for the LR/RL task ($F(8, 72) = 11.9$, $p < .001$); however, it is not for the audibility task ($F(8, 72) = 1.68$, n.s.).

Let us focus first on the LR/RL task. Given that the ANOVA reported a significant interaction between DB and SEP, post-hoc tests have to be done either between groups of same DB or between groups of same SEP, but by no means we can pool together groups of same DB or groups of same SEP. Moreover, a particular SNR threshold value is only meaningful when the measured threshold is significantly greater than the audibility threshold, otherwise that would mean that an audibility task is underlying the LR/RL task (that is, so long as the subject hears the target, he or she is able to discriminate the two sound locations, as explained in section 3.7.2). Therefore, pairwise comparisons were conducted using paired t-tests to look for significant differences between means of the two task types for all {DB, SEP} condition pairs. The resulting significance levels are given in **table 3.5**. As expected, for each distracter band, from some given angular separation the difference between the obtained thresholds of the two task types becomes insignificant. Continuing with this rationale, we assessed the effect of SEP by considering only those SNR threshold values of the LR/RL task that were significantly greater than the audibility threshold. Concretely, this mean that we kept the following {DB, SEP} condition pairs: {1, 6°}, {4, 6°}, {4, 10°} and {4, 14°}. The resulting effect of SEP in the LR/RL task is the following. Since for the 0 ERB DB group no point is concerned, and that a single point is concerned for the 1 ERB DB group ({1, 6°}), we could not perform any test for these two DB groups. Concerning the 4 ERB DB group, a trend analysis on

[13] Note that, when performing a non-parametric test, the condition {0, 6°} showed significant mean differences between the two task types ($p < .05$).



(a) Mean SNR thresholds for both tasks, grouped by distracter band.



(b) Mean SNR thresholds for the LR/RL task only, grouped by angular separation.

Figure 3.9: Mean signal-to-noise ratio (SNR) thresholds across all participants for a target center frequency of 1400 Hz, in experiment 2. This SNR value represents the ratio between the target the distracter levels. A SNR of 0 dB corresponds to the target played at its base level (52 dBA) and the distracter played at its loudness-equalized level (see section 3.4). The vertical bars represent ± the standard error of the mean.

the three concerned points showed that the effect of SEP follows a significant linear trend ($p < .001$). These results show the effect of the angular separation on the SNR threshold, and taken the other way around, they show the effect of the SNR on the MAA. In other words, the SNR has an effect on spatial blurring: as the SNR decreases, spatial blurring increases.

We will now focus on the audibility task. Since the ANOVA did not report a significant effect of the interaction between DB and SEP, we can pool together all SEP groups when looking at the effect of DB on the SNR, and in the same way, we can pool together all DB groups when looking at the effect of SEP on the SNR (see **figure 3.9a**). Trend analyses revealed a significant linear trend in both cases ($p < .001$). These results confirm the expected spatial unmasking, and show the decrease in the audibility (masking) threshold as the difference between the distracter and the target frequencies increases.

As a conclusion, the main result of this second experiment is the significant effect of the SNR (the ratio between the target and the distracter levels) on spatial blurring. As secondary results, we also confirmed the binaural release from masking of the target, as well as the effect of the difference in frequency between the distracter and the target on the masking threshold (it decreases as the difference increases).

3.8 Experiment 3: Effect of the Distracter Position

The main aim of this third study is the effect of the position of a single distracter on localization blur. Remember that we are not interested in the effect of the target position on localization blur, since it is already known that it is significant (see section 1.4.5), and more importantly because from an audio coding point of view, we assume that the listener is facing the target sound source (see section 3.1).

In addition to the effect of the distracter position, the design of this experiment was also motivated by a question that the previous experiment raised. In experiment 2, we assessed the effect of the SNR on the MAA by varying the target level and keeping the distracter level fixed. An interesting point however is that in quiet (that is without any distracting source), localization ability depends on the target level [SMM05, SR01]. Therefore it is reasonable to wonder whether a part of the effect we observed in the previous experiment might actually be attributable to the variation of the absolute target level rather than to the variation of the SNR between the target and the distracter levels. This question is addressed in section 3.8.4.

3.8.1 Setup

A schematic of the experimental setup is represented in **figure 3.10**. All speakers (custom Studer) were equalized beforehand at the listening position, positioned according to **table 3.6** (with an elevation of 0°), by recording their frequency response, inverting it, and then filtering each speaker with the corresponding 4096-sample long FIR filter. This ensured as flat a response as possible for each speaker, and consequently a similar response for all speakers. The overall level was adjusted in order to produce 52 dBA at the center of the listener's head when a speaker is fed with the target stimulus at its "base" level. That is to say, when the SNR equals 0 dB, both the target and the distracter are played at this base level. This overall level was chosen in accordance with our previous experiments.

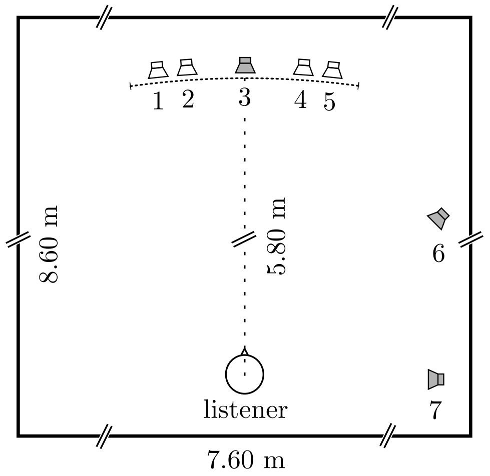


Figure 3.10: Top view of the experimental setup for experiment 3 (see table 3.6 for the exact target speakers positions). The gray speakers are the distracter speakers.

Speaker number	Position	Speaker number	Position
1	−6°	4	4°
2	−4°	5	6°
3	**0°**	**6**	**45°**
		7	**90°**

Table 3.6: Azimuth position of the speakers. The distracter speakers are in boldface. All speakers have a 0° elevation.

3.8.2 Procedure

The main aim of this study was to estimate the effect of the distracter position on the MAA. Since we have shown the dependence of the MAA on the SNR in our previous experiment, we assumed that it is possible to show the effect of a given variable on the MAA by testing its effect on the SNR corresponding to a 80.35% performance with a fixed angular separation. We therefore chose to retain the procedure we used for experiment 2, and thus a significant effect of a given variable on the SNR will be interpreted as being significant on the MAA as well.

The same tasks as for experiment 2 (LR/RL and audibility, see section 3.7.3) were used in this experiment. However, the audibility task was slightly different in this experiment, because the target was always played in the center speaker (speaker 3). This task still allowed an estimation of spatial unmasking, but contrarily to experiment 2, the target is now in front of the listener and the position of the distracter is varied. The audibility task, however, was mainly used to ensure that the LR/RL task was not in reality driven by an audibility task, as explained in section 3.7.2. Since the adaptive method gave accurate results in experiment 2 (see section 3.7.4), and the task types remained unchanged for this experiment, the same setup was used for the adaptive method in this experiment.

This experiment was organized in blocks. One block consisted of an adaptive procedure yielding a threshold value for a given experimental condition. An experimental condition was made of one or two parameters, depending on the task type (LR/RL or audibility). For the LR/RL task, two parameters were explored: the distracter position (0°, 45° or 90°), and the angular separation between targets (8° or 12°), whereas for the audibility task, only the distracter position was explored (0°, 45° or 90° as well). Both tasks also included a condition in which the distracter was not present (in that case, the two intervals in the audibility task were signaled by two flashes on a screen in the subject's sight, see section A.2). Using a repeated-measures design, the experiment was thus made of 12 blocks for each subject.

The order of presentation of the blocks was organized in two parts, based on the task type, all the LR/RL conditions being presented first, followed by all the audibility conditions. Within each of those two parts, all conditions were randomized.

The amount of training given to the subject was very small for the LR/RL task, and null for the audibility task. The experiment was divided into two sessions of one hour and fifteen minutes each, and the subjects were presented with a few training trials at the beginning of each session.

3.8.3 Data Analysis and Results

In addition to the twelve subjects who successfully completed this experiment, one was discarded, because of an inability to do the task for certain conditions, and an overall low level of performance compared to the other subjects. No outliers were detected in the dataset using Grubbs' test. All post-hoc tests were corrected for multiple comparisons using Holm-Bonferroni correction. Non-parametric tests (Wilcoxon's signed-ranks test and Cuzick's test) yielded results similar to those given in this section.

Figure 3.11 gives the mean thresholds for both tasks across all participants, grouped by target angular separation. At first sight, it seems that the angular separation has some effect on the SNR. There also seems to be an effect of the presence of the distracter on the SNR (for each curve compare the right-most symbol with the others). However this effect looks stronger for the 8° angular separation than for the 12° angular separation.

A two-way factorial repeated-measures ANOVA was carried out for the LR/RL task

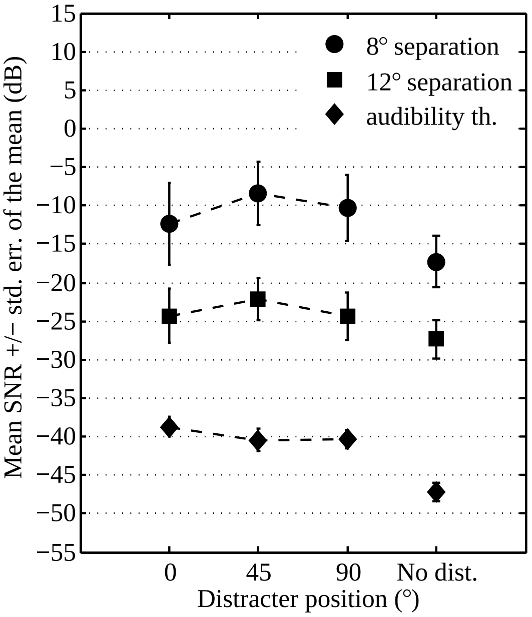


Figure 3.11: Mean signal-to-noise ratio (SNR) thresholds across all participants for a target center frequency of 1400 Hz, in experiment 3. This SNR value represents the ratio between the target the distracter levels. A SNR of 0 dB corresponds to the target played at its base level (52 dBA) and the distracter played (when present) at its loudness adjusted level (see section 3.4). The vertical bars represent ± the standard error of the mean.

type, with distracter position (DP) and angular separation (SEP) treated as independent variables, and SNR being the dependent variable. Note that the "no distracter" condition was not included in this analysis. As a result, it indicated that SEP is significant ($F(1, 11) = 42.4$, $p < .001$), but neither DP nor its interaction with SEP were significant ($F(2, 22) = .4$, n.s. and $F(2, 22) = .2$, n.s., respectively). Even without correction for multiple comparisons, paired t-tests did not show any significant differences between DP groups, both for the 8° and 12° SEP groups. In the same way, no significant quadratic trend of the effect of DP was found for either SEP group. Consequently, no significant effect of the distracter position can be found in our data. Two reasons can be proposed to explain this result. First, the effect of the distracter is reduced with such an important frequency distance between the target and the distracter (4 ERB critical bands), which might make the effect of the distracter position difficult to observe in a statistically significant way. Second, the limited bandwidth of the distracter might also limit the effect of its position. As a consequence, the effect of the distracter position on the SNR was investigated in the next experiment: 1) for several relative distracter bands, and 2) with multiple simultaneous distracters, therefore spanning several critical bands (see section 3.9).

3.8.4 Validity of Our Experimental Protocol

As formulated above, an additional concern was the validity of the experimental protocol used in experiment 2 and in the present experiment, and more specifically the choice of varying the target level. The question was to determine whether the observed effect of angular separation on the SNR was explained by the variation of the actual ratio between the target and the distracter levels, or only by the variation of the absolute target level.

To examine this question, we began by testing the effect of the presence of the distracter on the SNR using one-tailed paired t-tests with linear contrasts. In other words, we tested the 0°, 45° and 90° DP groups together against the "no distracter" DP group, and this for each SEP group. A significant difference was found for the 8° SEP group ($p < .05$), but not for the 12° SEP group. Therefore the presence of the distracter had a significant effect, but only for the 8° angular separation. In addition, we tested the effect of SEP for each DP group using paired t-tests, which showed a significant effect in each case: $p < .01$ for the 0° DP group, $p < .01$ for 45°, $p < .001$ for 90°, and $p < .01$ for the "no distracter" group. The fact that angular separation has a significant effect in the condition without distracter shows that the MAA also depends on target level. This extends to localization blur the conclusions of several studies [SMM05, SR01] about sound localization at near-threshold levels. Indeed, it has been found that target level has an effect on localization performance: it decreases starting from levels below 30 dB SL [14] and particularly below 20 dB SL. In the present experiment, the fact that the presence of the distracter has no significant effect on the SNR for an angular separation of 12° suggests that most of the effect of the angular separation is due to the difference in target level, and not to the difference in SNR as we assumed. According to the results cited above, this would make sense, since for the 12° separation, the resulting SNR is around −25 dB, which corresponds to a target level around $47 - 25 = 22$ dB SL (in **figure 3.11**, the threshold in quiet is around an SNR of −47 dB), and for the 8° separation, the resulting SNR is around −12 dB, which corresponds to a target level around $47 - 12 = 35$ dB SL. Therefore, for the 12° angular separation, the target level is below 30 dB SL, which means that localization is probably affected, whereas for the 8° angular separation, the target level is above 30 dB SL, and thus localization should not be much affected. Another point is that the decrease in localization performances observed in [SMM05, SR01] for near-threshold levels concerns broadband stimuli, and is usually explained by the fact that parts of the target spectrum (especially low frequencies) fall below the hearing threshold while other parts remain audible, which would result in the loss of localization cues associated with those inaudible parts. On the opposite, our stimuli are narrow-band, and therefore the dependence on the target absolute level should be reduced in our case.

Consequently, this experimental protocol might be biased when a specific experimental condition leads the SNR to reach values below 30 dB SL. In other words, to avoid a potential bias when using this protocol, it is safer to choose the experimental conditions such that the resulting SNR stays above 30 dB SL. In experiment 2, concerning the 4 ERB relative distracter band, this is the case for angular separations 6° and 10°, but not for the remaining ones: assuming a threshold in quiet around an SNR of −47 dB, [15] the target level corresponding to the condition with a 10° angular separation is around $47 - 14 = 33$ dB SL, whereas it is around 25 dB SL, 21 dB SL and 12 dB SL for the 14°, 18°, and 22° separations, respectively. Note that a paired t-test showed significantly different means between the 6° and 10° conditions ($p < .01$), so that our conclusions from experiment 2 remain valid: the angular separation has a significant effect on the SNR

[14] The level in dB SL ("sensation level") is the perceived intensity of a stimulus by an individual. It can be roughly approximated by the intensity in dB HL ("hearing level"), which is an average perceived intensity. This latter value is obtained by subtracting the average threshold in quiet to the stimulus presentation level.

[15] This is the threshold in quiet we obtained in experiment 3, but it could also be derived by looking at **figure 1.6**: the threshold in quiet is around 5 dB SPL, which is equivalent to 5 dBA at 1400 Hz, according to the A-weighting standard. In experiment 2, a target level at a SNR of 0 dB corresponds to 52 dBA. Consequently the corresponding threshold in quiet of the target stimulus indeed is at an SNR of $5 - 52 = -47$ dB.

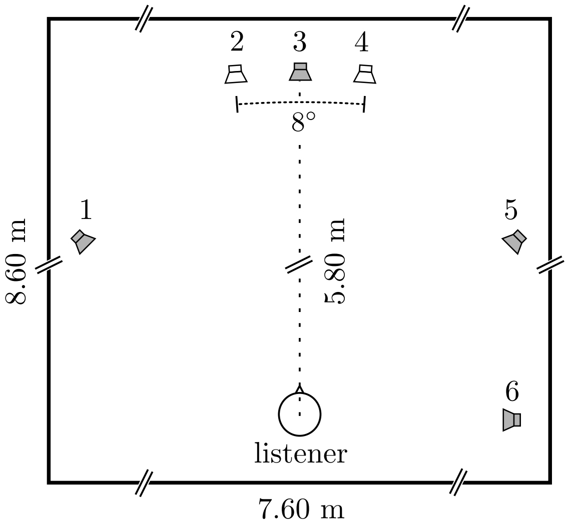


Figure 3.12: Top view of the experimental setup (see table 3.7 for the exact target speakers positions). The gray speakers are the distracter speakers.

or, the other way around, the SNR has a significant effect on the MAA. However, when computing the effect size of the SNR on the MAA, only the 6° and 10° data points should be considered, because other points are biased by the effect of the target level, which is then mixed with the effect of the SNR.

3.9 Experiment 4: Interaction Between Multiple Distracters

Our previous experiments have shown that the presence of one distracter may create spatial blurring on a target sound. From an audio coding point of view, it would be interesting to know if beyond a single distracter, the presence of additional distracters further increases spatial blurring or not. This question mainly guided the design of this last experiment, and hence localization blur was estimated in the presence of one, two, and four distracters, but also in quiet without any distracter. As a consequence, using the estimation in quiet as a reference, this allows a comparison of spatial blurring as a function of the number of distracters present.

In experiment 3, we did not find any significant effect of the distracter position on localization blur. This might be due to the frequency distance between the target and the distracter (4 critical bands) or also to its limited bandwidth. Thus, the present experiment also extends experiment 3 by testing the effect of the position of multiple distracters in separate critical bands. Indeed, these multiple distracters can be seen as a single, wideband distracter, as long as they are played synchronously and from the same position. A "spread" condition is also included, were several distracters are played synchronously, but from different positions.

A last concern was to study the relation between spatial blurring and energetic masking. If they have a simple relation, that would mean that spatial blurring could be predicted from energetic masking, which would be very convenient in terms of audio coding.

Speaker number	Position	Speaker number	Position
1	**-45°**	4	4°
2	−4°	**5**	**45°**
3	**0°**	**6**	**90°**

Table 3.7: Azimuth position of the speakers. The distracter speakers are in boldface. All speakers have a 0° elevation.

3.9.1 Setup

A schematic of the experimental setup is represented in **figure 3.12**. All speakers (custom Studer) were equalized beforehand at the listening position, positioned according to **table 3.7** (with an elevation of 0°), by recording their frequency response, inverting it, and then filtering each speaker with the corresponding 1024-sample long FIR filter. This ensured as flat a response as possible for each speaker, and consequently a similar response for all speakers. The overall level was adjusted in order to produce 52 dBA at the center of the listener's head when a speaker is fed with the target stimulus at its "base" level. That is to say, when the SNR equals 0 dB, both the target and the distracter are played at this base level. This overall level was chosen in accordance with our previous experiments.

3.9.2 Procedure

The main aim of this study was to estimate the effect of multiple distracters on the MAA. As explained for experiment 3, since we have shown in experiment 2 the dependence of the MAA on the SNR, we assumed that it is possible to show the effect of a given variable on the MAA by testing its effect on the SNR corresponding to a 80.35% performance with a fixed angular separation. This is why we chose to retain the procedure we used for experiments 2 and 3. In order to observe the conditions stated in section 3.8.4 and ensure that this experimental protocol is not biased, we presented an 8° angular separation, avoiding having the resulting SNR fall below 30 dBSL.[16] In the conditions where several distracters were presented simultaneously, all distracters were played at their loudness-matched level, and the SNR value displayed in our results is still the gain applied to the target stimulus by the adaptive method. This means that this SNR value does *not* represent the ratio between the target level and the energy sum of all distracters, but the ratio between the target level and the level of *any* of the distracters present. This allows us to directly interpret a variation of SNR in the results as an underlying variation of spatial blurring. The same tasks as for experiment 2 (LR/RL and audibility, see section 3.7.3) were used in this experiment, but as for experiment 3, in the audibility task the target was always played in the center speaker (speaker 3). Once again, the audibility task was mainly used to ensure that the LR/RL task is not in reality driven by an audibility task, as explained in section 3.7.2.

This experiment was organized in blocks. One block consisted of an adaptive procedure yielding a threshold value for a given experimental condition. An experimental condition was composed of several parameters: the task type (LR/RL or audibility), the number of distracters present (0, 1, 2 or 4), the center frequency of each of them (-4, -3, 3 or 4 ERB above the target center frequency), and their position (-45, 0, 45, 90 degrees, or "spread"). For the LR/RL task, the presented angular separation was always 8°. For each

[16] Experiment 3 showed that the SNR is likely to fall below 30 dBSL when no distracter or a single distracter is present for a 1400 Hz target center frequency.

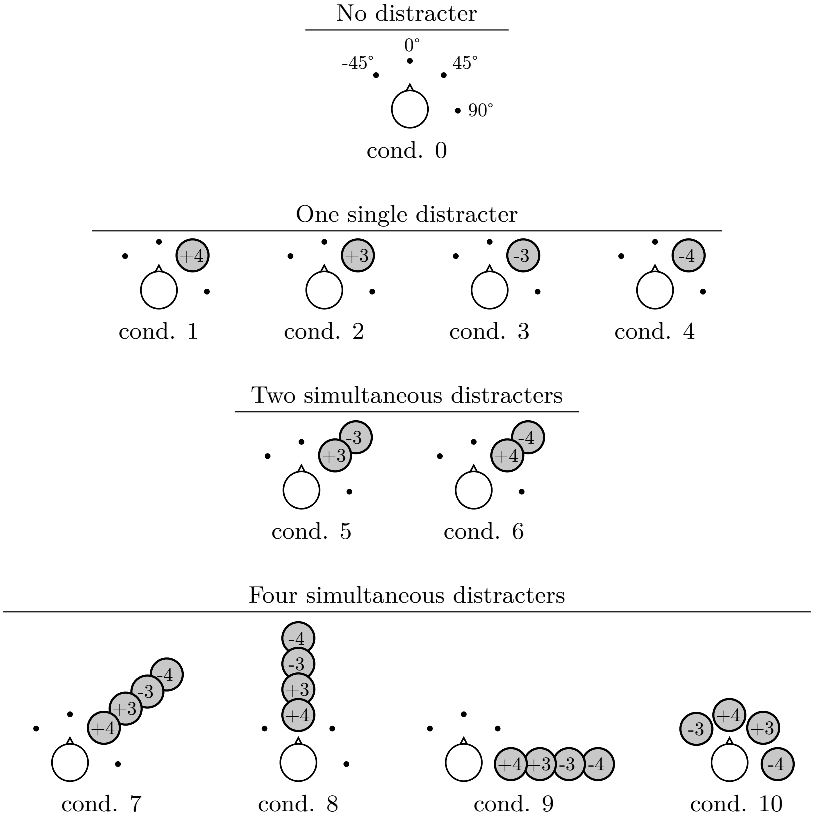


Figure 3.13: Experimental conditions of experiment 4: angular position and relative critical band (in ERB) of the distracter(s). The radius has no meaning: distracters at the same angular position were played at the same level. The targets were always presented in front of the listener with an angular separation of 8°.

subject, using a repeated-measures design, the experiment was made of 11 conditions for each task type, that is 22 conditions in all. These 11 experimental conditions are depicted in **figure 3.13**.

The order of presentation of the blocks was organized in two main parts, based on the task type: all the audibility conditions were presented first, followed by all the LR/RL conditions. Within each of these two parts, the presentation orders of all conditions were randomized. Further, to improve this randomization, and to approach as much as possible an optimal presentation order, the randomized sets were manually modified in a fashion that minimizes the number of appearances of a given condition at a given serial position.

Compared to experiments 2 and 3, the setup of the adaptive method was changed by decreasing the number of reversals of the staircases from 14 to 8. Concerning the audibility task, our previous experiments showed that the inter-subject variability is very reduced. Moreover, we did not necessarily need a precise estimation of audibility thresholds for our purposes, so two interwoven staircases of 8 reversals each (that is 16 reversals in total) should be sufficient to get accurate enough estimations. Concerning the LR/RL task, however, we needed very accurate estimations to be able to show significant effects. Therefore, each experimental condition was run twice, in two separate sessions, using each time two interwoven staircases of 8 reversals each. This means that compared to experiments 2 and 3, we broke a long staircase ($2 \times 14 = 28$ reversals) into two shorter ones ($2 \times 8 = 16$ reversals each), which has the advantage of avoiding fatigue. Thus, combining these two sessions, the threshold corresponding to a given LR/RL condition was obtained by averaging the peaks and valleys over 32 reversals, which is more than the 28 reversals used in experiments 2 and 3.[17] The first session followed the original presentation order that was assigned to the subject, and the second one—run on a different day—followed the same order, but flipped. This allowed to average out training effects due to the order of presentation of the conditions and to reduce intra-subject variability (this second point is specifically discussed in appendix B).

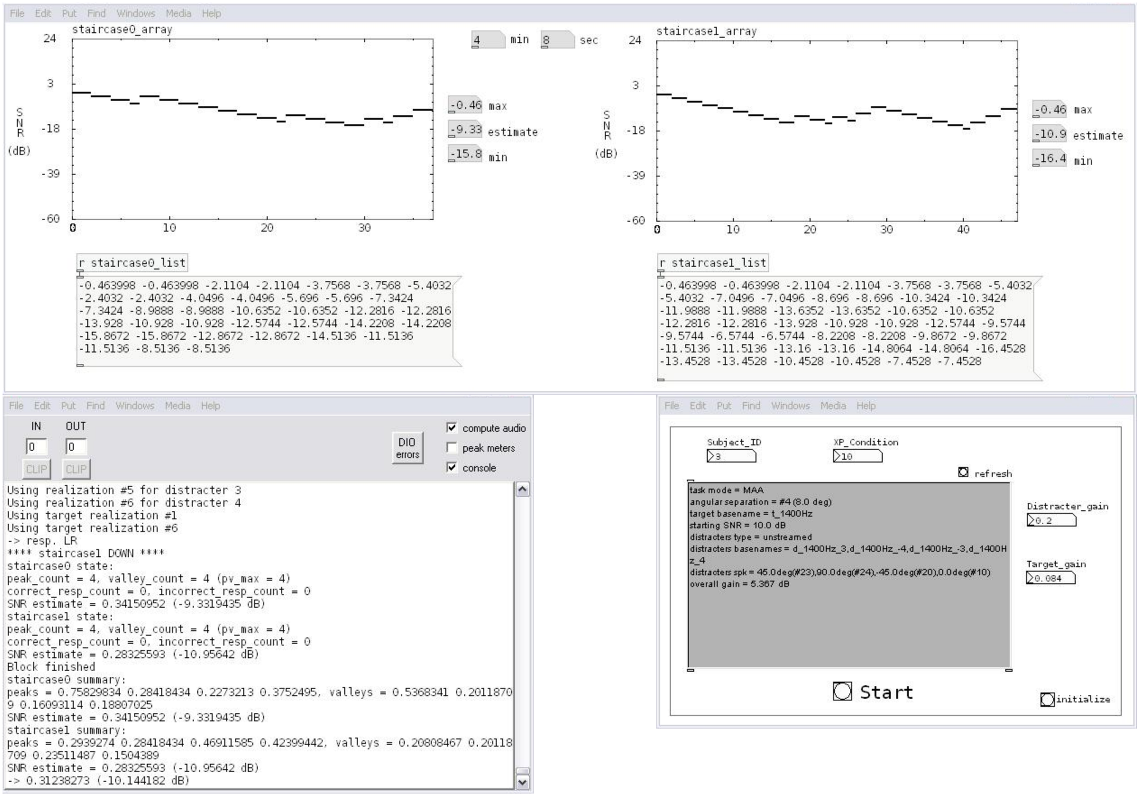
The amount of training given to the subject was very small for the LR/RL task, and null for the audibility task. The experiment was composed of three sessions. The first session, lasting one and a half hours, was dedicated to a hearing test, a quick ability test concerning the LR/RL task, and all of the 11 audibility conditions. The second and third sessions, lasting one hour and fifteen minutes each, were both dedicated to the LR/RL task.

3.9.3 Data Analysis and Results

In addition to the fifteen subjects who successfully completed this experiment, three were discarded, because of their inability to do the task for certain conditions, and an overall low level of performance compared to the other subjects. No outliers were detected in the dataset using Grubbs' test. All post-hoc tests were corrected for multiple comparisons using Holm-Bonferroni correction. Unless otherwise stated, non-parametric tests (Wilcoxon's signed-ranks test and Cuzick's test) yielded results similar to those given in this section.

Figure 3.15 gives the mean thresholds for both tasks across all participants in each experimental condition. We can draw a few observations from it, that we will test statistically. Note that in the following, we will consider that a change in SNR is equivalent

[17] We slightly increased the total number of reversals because García-Pérez reported that running several short staircases is less efficient statistically than running a single long staircase with the same total number of reversals [Gar00].



```
r staircase0_list
-0.463998 -0.463998 -2.1104 -2.1104 -3.7568 -3.7568 -5.4032
-2.4032 -2.4032 -4.0496 -4.0496 -5.696 -5.696 -7.3424
-7.3424 -8.9888 -8.9888 -10.6352 -10.6352 -12.2816 -12.2816
-13.928 -10.928 -10.928 -12.5744 -12.5744 -14.2208 -14.2208
-15.8672 -15.8672 -12.8672 -12.8672 -14.5136 -11.5136
-11.5136 -8.5136 -8.5136
```

```
r staircase1_list
-0.463998 -0.463998 -2.1104 -2.1104 -3.7568 -3.7568 -5.4032
-5.4032 -7.0496 -7.0496 -8.696 -8.696 -10.3424 -10.3424
-11.9888 -11.9888 -13.6352 -13.6352 -10.6352 -10.6352
-12.2816 -12.2816 -13.928 -10.928 -10.928 -12.5744 -9.5744
-9.5744 -6.5744 -6.5744 -8.2208 -8.2208 -9.8672 -9.8672
-11.5136 -11.5136 -13.16 -13.16 -14.8064 -14.8064 -16.4528
-13.4528 -13.4528 -10.4528 -10.4528 -7.4528 -7.4528
```

```
Using realization #5 for distracter 3
Using realization #6 for distracter 4
Using target realization #1
Using target realization #6
-> resp. LR
**** staircase1 DOWN ****
staircase0 state:
peak_count = 4, valley_count = 4 (pv_max = 4)
correct_resp_count = 0, incorrect_resp_count = 0
SNR estimate = 0.34150952 (-9.3319435 dB)
staircase1 state:
peak_count = 4, valley_count = 4 (pv_max = 4)
correct_resp_count = 0, incorrect_resp_count = 0
SNR estimate = 0.28325593 (-10.95642 dB)
Block finished
staircase0 summary:
peaks = 0.75829834 0.28418434 0.2273213 0.3752495, valleys = 0.5368341 0.2011870
9 0.16093114 0.18807025
SNR estimate = 0.34150952 (-9.3319435 dB)
staircase1 summary:
peaks = 0.2939274 0.28418434 0.46911585 0.42399442, valleys = 0.20808467 0.20118
709 0.23511487 0.1504389
SNR estimate = 0.28325593 (-10.95642 dB)
-> 0.31238273 (-10.144182 dB)
```

Subject_ID 3

XP_Condition 10

```
task mode = MAA
angular separation = #4 (8.0 deg)
target basename = t_1400Hz
starting SNR = 10.0 dB
distracters type = unstreamed
distracters basenames = d_1400Hz_3,d_1400Hz_-4,d_1400Hz_-3,d_1400H
z_4
distracters spk = 45.0deg(#23),90.0deg(#24),-45.0deg(#20),0.0deg(#10)
overall gain = 5.367 dB
```

Distracter_gain 0.2

Target_gain 0.084

Figure 3.14: Screenshot of experiment 4 running.

to a change in localization blur, as justified in section 3.9.2. First, spatial blurring seems to increase with the number of distracters present (compare for example conditions 4, 6 and 7). Second, there seems to be an asymmetry between distracters located below and above the target center frequency: distracters located below tend to create more spatial blurring on the target (compare conditions 2 and 3 for instance). Third, distracters located closer to the target center frequency seem to create more spatial blurring on the target (compare for example conditions 3 and 4, or conditions 5 and 6). Fourth, even if globally a distracter which has a strong energetic masking ability tends to have a strong spatial blurring ability, there seem to be several counter-examples in our data (compare for example conditions 2 and 3: they show a similar energetic masking ability, but different spatial blurring abilities). Finally, the position of the distracters (conditions 7, 8, 9 and 10) does not seems to have much effect, although the 0° position seems to differ from the other cases.

A one-way repeated-measures ANOVA was carried out for the LR/RL task type. Because the design of this experiment is not factorial, the experimental condition was the only considered factor, and the SNR was the dependent variable. It revealed that the experimental condition has a significant effect on the SNR ($F(10, 140) = 25.1$, $p < .001$).

We started by looking for significant linear trends within groups of conditions with an increasing number of distracters, that is among the condition groups {2, 5, 7}, {3, 5, 7}, {1, 6, 7}, and {4, 6, 7}. All of them showed a significant linear trend ($p < .001$ in all cases, except the {3, 5, 7} group, for which $p < .05$ [18]). This confirms our first observation: spatial blurring increases with the number of distracters, at least up to four independent

[18] The {3, 5, 7} group did not show a significant linear trend when performing a non-parametric test.

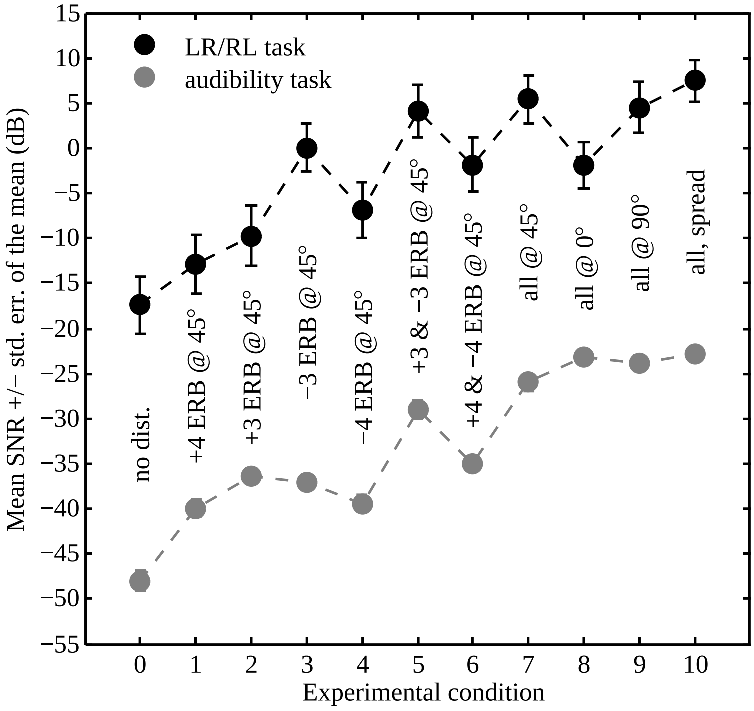


Figure 3.15: Mean signal-to-noise ratio (SNR) thresholds across all participants for a target center frequency of 1400 Hz, in experiment 4. This SNR value represents the ratio between the target level and *any* of the distracters present. A SNR of 0 dB corresponds to the target played at its base level (52 dBA) and each of the distracters played (when present) at its loudness-matched level (see section 3.4). The vertical bars represent ± the standard error of the mean.

distracters of a critical bandwidth each. A modeling of the additivity of distracters is proposed in chapter 4.

Paired t-tests were used to test the second and third remarks stated above. Conditions 2 and 3 have significantly different means ($p < .01$), suggesting that there is an asymmetry between distracters located below and above the target center frequency. Moreover, conditions 3 and 4, as well as conditions 1 and 2,[19] have significantly different means ($p < .01$ and $p < .05$, respectively), which would confirm that distracters located closer to the target center frequency create more spatial blurring on the target, both for distracters located below and above the target center frequency.

When the number of distracters present increases, energetic masking increases because the total energy of the distracters increases, and of course spatial blurring increases as well for the same reason. However, apart from that particular case, an increase in energetic masking does not necessarily imply an increase in spatial blurring. For instance, as conditions 2 and 3 are significantly different, we can state that spatial blurring is more important in condition 3 than in condition 2. However, energetic masking follows the opposite trend in these two conditions. This means that the blurring ability of a distracter is not correlated with its masking ability, which suggests that spatial blurring and energetic masking rely on different auditory processes.

Finally, concerning the effect of the position of the distracters, paired t-tests among conditions 7, 8, 9 and 10 only showed a significant mean difference between the 0° and the "spread" conditions ($p < .01$). Note that without the correction for multiple comparisons, the 45° and 90° conditions were also significantly different from the 0° condition, though. The insignificant difference between the 45° and 90° conditions found in experiment 3 was also observed in this experiment, thereby extending this result to wideband distracters.[20] Since the "spread" condition does not significantly differ from the 45° and 90° conditions, having a sound scene with spread components does not significantly increase spatial blurring, although there is a slight tendency to do so. Only the 0° position for all distracters seems to be a peculiar experimental condition in which subjects perform better than with any other distracter positioning. Actually, this is probably more likely due to the fact that the distracters are in between the two targets than that the distracters are in front of the listener. This result may then be explained by a repulsion effect [LDS09] between the distracters and each of the two targets. Indeed, when a target is played, the presence of a simultaneous distracter "pushes" the localization of the target away from the distracter, and since in our case the distracters are in between the two targets, a repulsion effect would increase the perceived angle between them, and thus facilitate their spatial discrimination. This singularity concerning the distracter position was also observable as a nonsignificant trend in results from experiment 3 (see **figure 3.11**). In this fourth experiment, note however that the "spread" condition has the highest mean value, suggesting that this effect is only effective when there are no other sources outside the angular area between the two targets. This would make sense since the presence of other sources would reduce this repulsion effect.

[19] Conditions 1 and 2 did not show a significant mean difference when performing a non-parametric test.

[20] As already stated, since all distracters are collocated and synchronous, they can also be considered as a single, wideband distracter.

3.10 Summary and Conclusions

From all these four experiments, we gathered several results concerning localization blur and spatial blurring that can be summarized as follows:

1. the localization blur associated with a specific sound source depends on the frequency content of that source [exp. 1];
2. the localization blur associated with a specific sound source increases when its level decreases, but this effect appears to be restricted to levels near the hearing threshold (that is below 30 dB SL) [exp. 3];
3. spatial blurring phenomenon: a distracting sound source can increase the localization blur associated with a target sound source [exp. 1];
4. the spatial blurring a distracter might create on a target depends on their frequency separation [exp. 1 & 4] as well as on their energy ratio [exp. 2];
5. a distracter might not exert the same spatial blurring depending on whether it is located below or above the target in frequency [exp. 4];
6. the effect of the frequency separation between a target and a distracter on spatial blurring is dependent on the target center frequency [exp. 1];
7. spatial blurring increases with the number of distracters, at least up to four independent distracters of a critical bandwidth each [exp. 4];
8. the distracter position does not seem to affect spatial blurring [exp. 3 & 4] except when all the distracters present are specifically collocated with the target, [21] where their effect is significantly reduced [exp. 4];
9. the blurring ability of a distracter is not correlated with its masking ability, which suggests that spatial blurring and energetic masking rely on different auditory processes [exp. 4].

By the nature of the stimuli and the experimental conditions used in these experiments, these results are low-level results, they do not give any indications about what would happen in an auditory scene analysis context (see section 1.3), that is a real sound scene context where auditory streams are formed by the auditory system. Moreover, we only considered continuous distracters, although temporal variations seem to affect their blurring ability, as reported for example by Croghan and Grantham using pulsating sources [CG10].

However, we used the few results we got from our experiments to build a model of localization blur and spatial blurring which is presented in chapter 4. Then we propose in chapter 5 two spatial audio coding schemes which, based on this model, take advantage of spatial blurring to provide a bitrate reduction.

[21] Or, from an experimental point of view, when all distracters are positioned between the two targets.

Chapter 4

Towards a Model of Spatial Blurring and Localization Blur

In this chapter, we propose a psychoacoustic model of spatial blurring and localization blur based on the experimental data obtained in chapter 3. Indeed, spatial blurring estimations for a target and a single distracter assuming an SNR of zero are provided by experiment 1, while the effect of SNR on spatial blurring will be extracted from experiment 2. Finally, we will derive from experiment 4 a rule of additivity for multiple distracters. In chapter 5, we will propose two spatial audio coding schemes based on this psychoacoustic model.

Given the small amount of experimental data at our disposal, some of the assumptions we will make in the following are difficult to verify or justify. Further studies will be necessary to refine and correct this model.

4.1 Assumptions

In our psychoacoustic experiments, localization blur was always assessed in front of the listener. As explained in section 3.1, localization blur is the smallest for frontal sources, which therefore constitutes the worst-case scenario in terms of audio coding. As a consequence, our psychoacoustic model will always return an MAA value assuming that the target source is located in front of the listener.

Among results of our fourth experiment (see section 3.9.3), the distracter position did not have a significant effect on spatial blurring, except in the experimental condition in which the distracter was centered between the two targets, in which case spatial blurring was reduced. In practice, this corresponds to the situation in which the target and distracter under consideration are collocated. On the other hand, results from the same experiment showed that this effect seems to disappear when sources are present outside the angular area between the two targets, which happens in practice when there are sources that are *not* collocated with the target. As a consequence, this effect seems to occur only when all sources are collocated, otherwise it is likely to be greatly reduced. We decided in this model to ignore the position of the distracter, by always assuming that distracters are collocated with the target, which thus corresponds to the worst-case scenario in terms of audio coding. This solution also suits coding schemes within which MAAs are considered relatively (as in the scheme presented in section 5.1). Indeed, either all sources are really collocated, in which case spatial blurring in fact is reduced for all these sources, or some of the sources are not collocated, and in that case spatial blurring is no longer reduced for the isolated sources, while potentially still being reduced for the others, but by a negligible amount due to the presence of the isolated sources. Thus, by assuming that all distracters are collocated with the target, we globally underestimate spatial blurring, but in the same proportion for all sources.

As explained in section 1.2.3, loudness is integrated by the auditory system within critical bands, and localization cues are frequency specific (see section 1.4.1). As a result, a source will be described in our model as a certain level of energy in a given critical band.

Our psychoacoustic experiment studied localization blur associated with primary [1] sources. Nevertheless, it is likely that using ambient sources would have led to different results. For this reason, we will anticipate such results by including the type of source as an input argument of our model, although this notion is not considered further in the following.

4.2 Formalism and Overview

A source s_k is described by three parameters $\{f_c, I, t\}$, where f_c is the center frequency of the sub-band to which the source belongs, I is the associated energy level of the source, and t is its type (primary or ambient). The psychoacoustic model is fed with *all* sources $\{s_k, k \in 1..K\}$ present in the sound scene via a call to function $\psi\left(\{s_k, k \in 1..K\}\right)$. It consists of the following set of functions:

1. Θ returns the masking threshold associated with each source in the presence of all other sources (see section 4.3);

2. b_0 returns the reference value of spatial blurring created on a given target source by a given distracter, assuming a SNR of 0 dB between them (see section 4.4);

3. b_τ corrects the estimation given by b_0 by taking into account the actual SNR τ between the target and the distracter (see section 4.5);

4. b_Σ computes the additive spatial blurring resulting from the simultaneous presence of the sources (see section 4.6);

5. Ω returns the localization blur associated with a given source by combining the localization blur in quiet with the total spatial blurring created on this source (see section 4.7).

For each source s_k, the model calls $\Omega\left(s_k, \{s_1, \ldots, s_{k-1}, s_{k+1}, \ldots, s_K\}\right)$, which considers s_k as a target in the presence of all other $K-1$ sources considered as distracters, and returns the MAA α_k associated with s_k. As a result, the psychoacoustic model yields the localization blur associated with each input source as a set of MAAs $\{\alpha_k, k \in 1..K\}$. This process is illustrated in **figure 4.1**, for a sound scene composed of three sources.

4.3 Computation of Masking Thresholds

Function $\Theta\left(\{s_k, k \in 1..K\}\right)$ is called a single time by the model, and returns the set of masking thresholds $\{\theta_k, k \in 1..K\}$ associated with each source s_k in the presence of the other $K-1$ sources. This function is mainly necessary to know which sources are inaudible—that is, below their masking threshold. Since these sources are inaudible, they will not be considered in the following as potential distracters. For the same reason, they

[1] By primary sources, we mean sources created by correlated signals between speakers. To the contrary, ambient sources correspond to uncorrelated signals between loudspeakers.

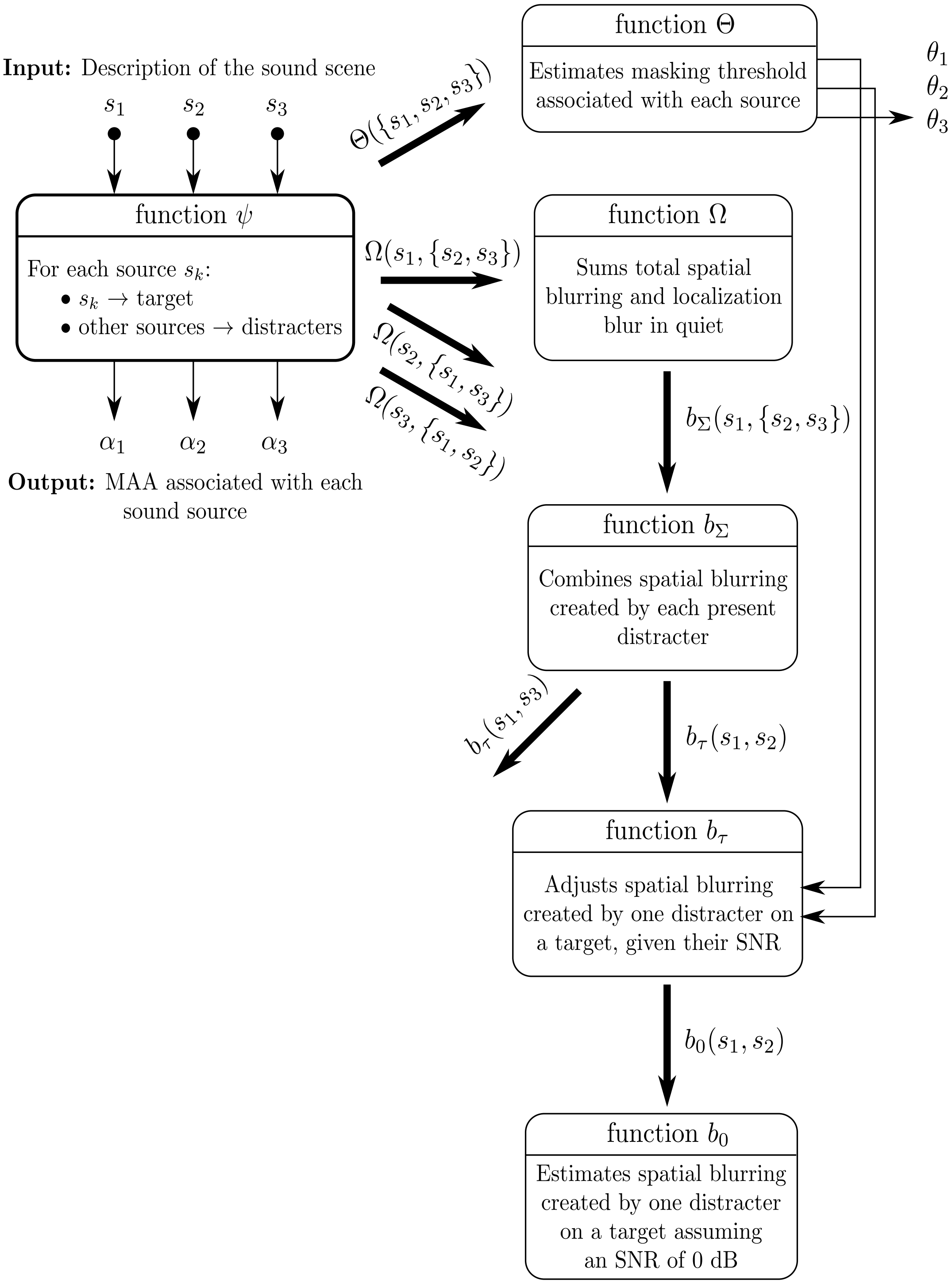

Input: Description of the sound scene
s_1
s_2
s_3
function ψ
For each source s_k:
• $s_k \rightarrow$ target
• other sources $\rightarrow$ distracters
α_1
α_2
α_3
Output: MAA associated with each sound source
$\Theta(\{s_1, s_2, s_3\})$
function Θ
Estimates masking threshold associated with each source
θ_1
θ_2
θ_3
$\Omega(s_1, \{s_2, s_3\})$
$\Omega(s_2, \{s_1, s_3\})$
$\Omega(s_3, \{s_1, s_2\})$
function Ω
Sums total spatial blurring and localization blur in quiet
$b_\Sigma(s_1, \{s_2, s_3\})$
function b_Σ
Combines spatial blurring created by each present distracter
$b_\tau(s_1, s_3)$
$b_\tau(s_1, s_2)$
function b_τ
Adjusts spatial blurring created by one distracter on a target, given their SNR
$b_0(s_1, s_2)$
function b_0
Estimates spatial blurring created by one distracter on a target assuming an SNR of 0 dB

Figure 4.1: Illustration of the function calls (thick arrows) made within the psychoacoustic model when it is invoked with a sound scene composed of three sources s_1, s_2 and s_3.

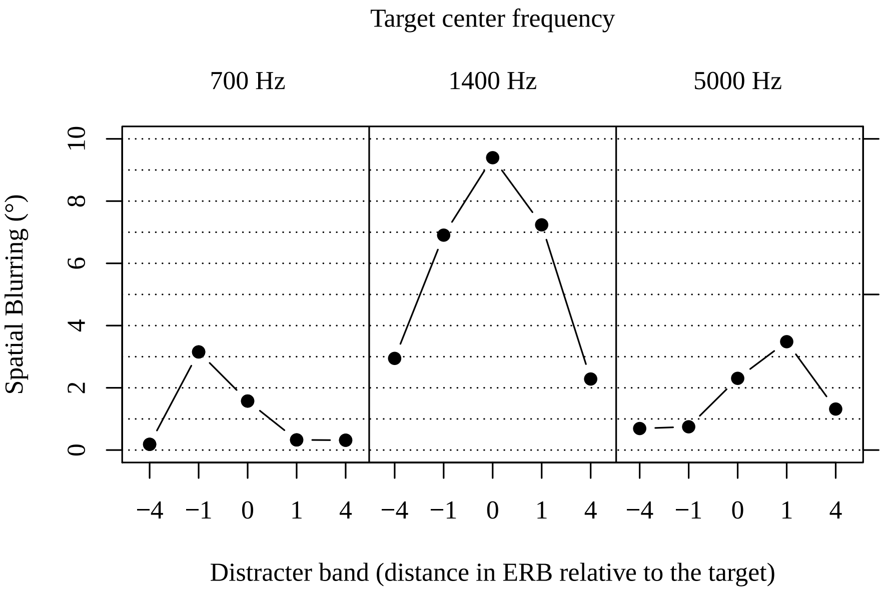


Figure 4.2: Experimental spatial blurring obtained from results of experiment 1 (see **figure 3.6**) after subtraction of localization blur in quiet.

do not have any associated MAA, which by convention will be set to $+\infty$.[2] The interest is double. First, having fewer sources to consider reduces computation. Second, having sources with an infinite associated MAA will necessarily lead to better performances in the coding schemes we will present in chapter 5. These considerations will be useful for the model of function b_τ, detailed in section 4.5.

We gathered data concerning energetic masking and binaural unmasking, especially in our second experiment (see section 3.7). In particular, for a target with a center frequency of 1400 Hz, linear regressions yielded a binaural unmasking of -0.3 dB per degree, and a variation of the masking threshold with the target-masker frequency distance of -8.48 dB per ERB. However our experiments did not aim primarily to study energetic masking, and these results are of course insufficient to build a model of it, especially in terms of additivity of the maskers. Hence a more complete model should be considered for this module. This is beyond the scope of this thesis, and more information about such models can be found for instance in [BG02].

4.4 Reference Value of Spatial Blurring

For the following, we need to define d_z as the distance on an ERB scale (see equation (1.4)) between a target s_t and a distracter s_d:

$$d_z(f_t, f_d) = \mathrm{ERBS}(f_d) - \mathrm{ERBS}(f_t), \tag{4.1}$$

where f_t and f_d are the center frequencies of the critical bands to which the target and the distracter belong, respectively. So when the distracter belongs to a critical band above that of the target, d_z gets a positive value.

[2] However, the worst case scenario in term of binaural unmasking has to be considered here, to avoid the source being unmasked if its position is affected by the coding scheme.

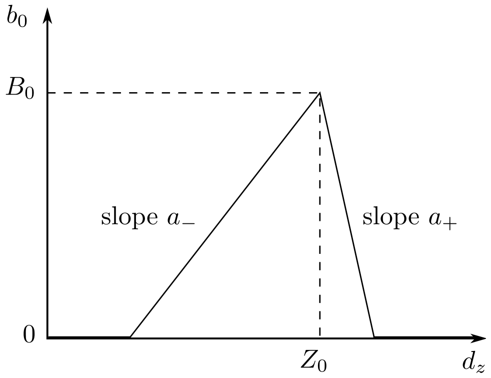


Figure 4.3: Characteristics of function b_0 for a given value of f_t.

Function $b_0(s_t, s_d)$ returns the spatial blurring created on a target s_t of center frequency f_t by a distracter s_d located d_z ERBs away assuming an SNR between them of 0 dB. To do so, we will model data from our first experiment (see section 3.6) depicted in **figure 3.6** in order to be able to express the spatial blurring created by a distracter of any frequency on a target of any frequency. Recall that, as in the experimental conditions of chapter 3, an SNR of 0 dB means that the target and the distracter are loudness equalized (see section 3.4). The estimation yielded by b_0 will be considered as an estimation of reference which will be corrected afterwards by function b_τ (see section 4.5).

From our experimental data, we know the localization blur in quiet for targets with center frequencies of 700 Hz, 1400 Hz and 5000 Hz, respectively. We also know the localization blur resulting from the presence of a distracter located -4, -1, 0, 1, and 4 ERBs away from these targets. In section 3.2, we defined spatial blurring as the difference between localization blur in the presence of distracters and localization blur in quiet. We can compute the spatial blurring resulting from the presence of the distracter by subtracting the localization blur in quiet from each of the three curves, that is 1.5° at 700 Hz, 5.1° at 1400 Hz, and 3.9° at 5000 Hz. The resulting curves are given in **figure 4.2**. The shape of each curve follows the same characteristics, depicted in **figure 4.3**. Spatial blurring is close to zero for large values of $|d_z|$, and we will assume that it eventually becomes nil at some frequency distance between the target and the distracter. We will also assume that it never becomes negative, which would otherwise mean that the presence of the distracter enhances localization performance. A maximum spatial blurring B_0 is reached for a certain frequency distance Z_0 between the target and the distracter. For $d_z < Z_0$, spatial blurring follows a slope a_-, whereas for $d_z > Z_0$, it follows a slope a_+. Coordinates $\{Z_0, B_0\}$ as well as slopes a_- and a_+ are dependent on f_t. As a result, the spatial blurring created on a target s_t of center frequency f_t by a distracter s_d located d_z critical bands away can be modeled as a triangular function b_0:

$$b_0(s_t, s_d) = \max\left(0, \left(a_- \times \theta(d_{Z_0}) + a_+ \times \theta(-d_{Z_0})\right) \times d_{Z_0} + B_0(f_t)\right), \tag{4.2}$$

where $d_{Z_0} = d_z - Z_0(f_t)$ is the distance of the distracter band from the band inducing maximum spatial blurring, and $\theta(d_{Z_0})$ is the step function equaling one for positive values of d_{Z_0} and zero otherwise.

Modeling b_0 consists of modeling each of these characteristics. Let us first consider the modeling of coordinates $\{Z_0, B_0\}$ for any target center frequency. From our experimental data, these coordinates are the following:

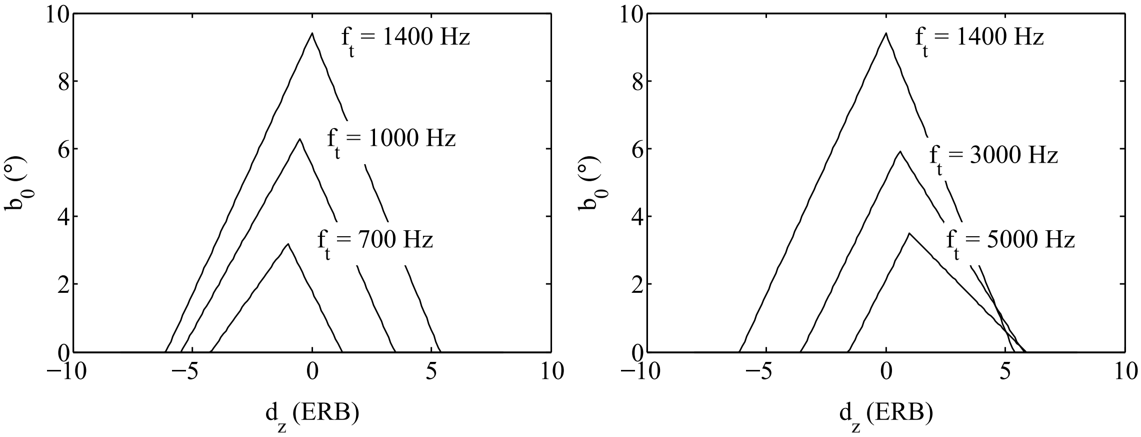


Figure 4.4: Function b_0 for different target center frequencies: 700 Hz, 1000 Hz and 1400 Hz (left panel); 1400 Hz, 3000 Hz and 5000 Hz (right panel).

f_t	B_0	Z_0
700	3.2	-1
1400	9.4	0
5000	3.5	1

To obtain a value of B_0 for any value of f_t, we interpolated linearly on an ERB scale between each of the three experimental values of B_0. For f_t smaller than 700 Hz we held B_0 constant at a value of 3.2, and for f_t greater than 5000 Hz we held B_0 constant at a value of 3.5. We performed the same operations to obtain a value of Z_0 for any value of f_t, with boundary values of -1 for f_t smaller than 700 Hz and 1 for f_t greater than 5000 Hz.

Concerning the modeling of a_- and a_+, we performed linear regressions on our data to obtain these slopes for each of our experimental f_t values. Depending on the target center frequency f_t, this regression was not performed on points corresponding to the same values of d_z. The resulting values for a_+ and a_- are the following (the d_z values considered for each regression are provided as well):

f_t = 700 Hz		f_t = 1400 Hz		f_t = 5000 Hz	
$d_z = -4, -1$	$d_z = -1, 0, 1$	$d_z = -4, -1, 0$	$d_z = 0, 1, 4$	$d_z = -1, 0, 1$	$d_z = 1, 4$
$a_- = 0.99$	$a_+ = -1.41$	$a_- = 1.54$	$a_+ = -1.75$	$a_- = 1.37$	$a_+ = -0.72$

As previously, we interpolated linearly on an ERB scale between all these experimental results to obtain values of a_- and a_+ for any value of f_t. For f_t less than 700 Hz, we held a_- and a_+ constant at 0.99 and -1.41, respectively, and for f_t greater than 5000 Hz we held them constant at 1.37 and -0.72, respectively.

Function b_0 is illustrated in **figure 4.4** for various target center frequencies, and an overview plot of b_0 is given in **figure 4.5**. The spatial blurring created on a target s_t of center frequency f_t by a distracter s_d located d_z critical bands away is thus obtained by a call to $b_0(s_t, s_d)$.

4.5 Accounting for the Effect of SNR

Function b_0 returns an estimation of the spatial blurring between a target and a distracter assuming an SNR of zero. We will present in this section function $b_\tau(s_t, s_d)$ that corrects this estimation by taking into account the actual SNR τ between the energy level I_t of

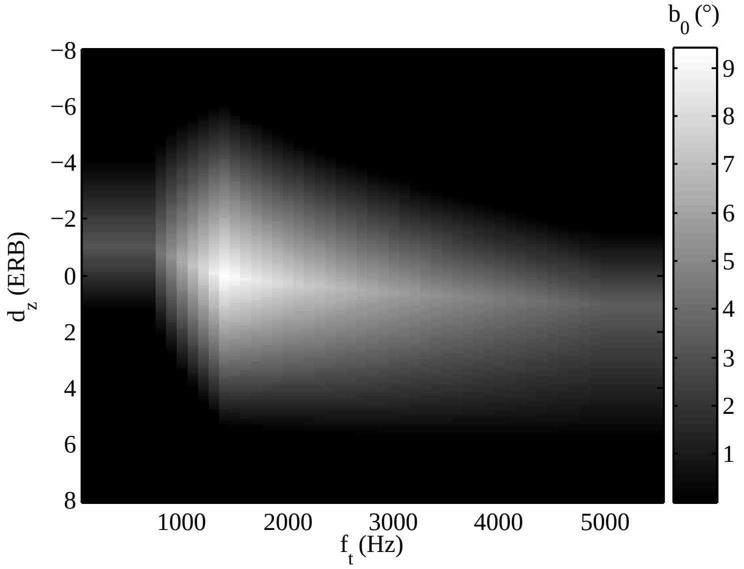


Figure 4.5: Overview plot of function b_0. The value of b_0 in degrees is represented according to the grayscale at the right. The value of b_0 is plotted for combinations of target frequency, f_t, and the distance d_z of the distracter from the target in ERBs.

a target s_t and that of a distracter s_d, I_d. We studied the relation between SNR and localization blur in experiment 2. Since in the experimental stimuli we used, distracters were loudness equalized with the target, it is first necessary to apply a correction on I_t and I_d to match our experimental conditions. This can be done, for example, by applying an A-weighting coefficient, $A(f)$, which is an offset added to the energy levels in decibels, depending on the frequency of the signal under consideration:

$$A(f) = 2.0 + 20\log_{10}(R_A(f)), \tag{4.3}$$

with

$$R_A(f) = \frac{12200^2 f^4}{(f^2+20.6^2)\sqrt{(f^2+107.7^2)(f^2+737.9^2)}(f^2+12200^2)}. \tag{4.4}$$

Thus we can define:

$$I_{t_A} = I_t + A(f_t), \text{ and } I_{d_A} = I_d + A(f_d). \tag{4.5}$$

We will use our data from experiment 2 to model the relation between SNR and spatial blurring. From this experiment, we are only looking for the slope of the relation, such that the likelihood of an offset between results from experiment 1 and 2 due to different experimental protocols, setups and subjects is not an issue. We propose to model b_τ linearly as:

$$b_\tau(s_t, s_d) = D(b_0(s_t, s_d)) \times \tau + b_0(s_t, s_d), \tag{4.6}$$

where $\tau = I_{t_A} - I_{d_A}$. The choice of b_0 as an intersept is justified by the fact that b_τ must equal b_0 for $\tau = 0$. The dependence of the slope D on b_0 is assumed but cannot be assessed or even verified from our limited experimental data. Further studies would be necessary to address this point. Indeed, we can only use two points out of the entire dataset of experiment 2, because either the other points are not significantly above masking threshold (see section 3.7.5) or they are too close to the audibility threshold in quiet (see section 3.8.4), meaning that these points are potentially biased. Moreover, we did not

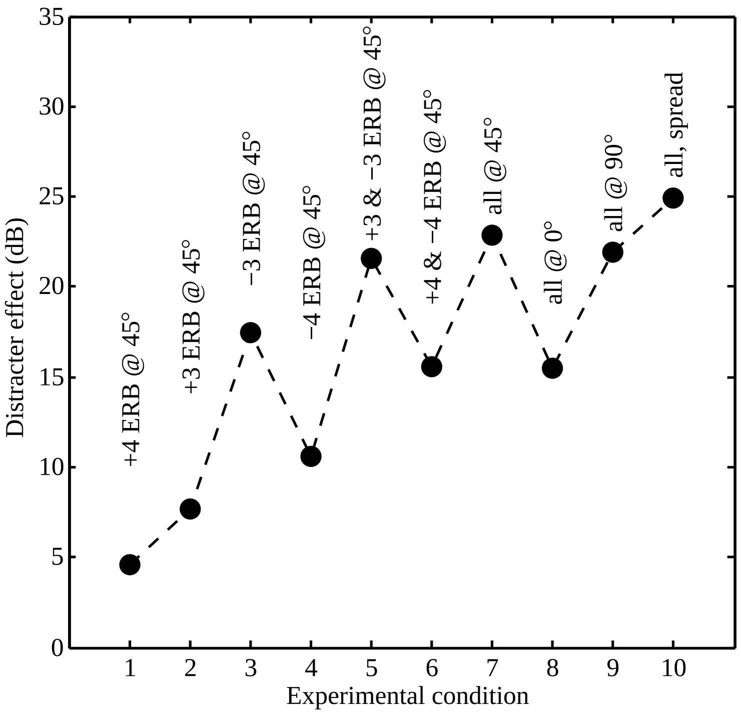


Figure 4.6: Effect of distracter(s), obtained from results of experiment 4 by subtracting the SNR obtained in quiet (condition 0) from the SNR obtained in the presence of the distracters under consideration.

test any target frequencies other than 1400 Hz in that experiment. The two remaining unbiased points at our disposal correspond to angular separations of 6° and 10°, for a target center frequency of 1400 Hz and a distracter four critical bands above the target: $\{3.8\text{ dB}, 6°\}$ and $\{-12.9\text{ dB}, 10°\}$. This leads to a slope D of $-.24°$ per decibel in that specific condition. In the absence of other estimations of D in different conditions, we will use this value of D in all cases.

The expression of $b_\mathcal{T}$ given in equation (4.6) is only valid under specific conditions. First, the distracter level has to be above its masking threshold, as returned by function Θ. Otherwise, since the distracter is inaudible we can assume that its effect is null, and thus that $b_\mathcal{T} = 0$. In the same way, the level of the target has to be above its masking threshold, otherwise, since the target is inaudible we can assume by convention that the spatial blurring created on it is infinite, that is, $b_\mathcal{T} = +\infty$. Finally, we will assume that the level of the target is not close to its threshold of audibility in quiet. Indeed, as already mentioned in section 3.8.4, the level of the target would otherwise have an effect on its associated localization blur [SMM05, SR01]. In such cases, the expression of $b_\mathcal{T}$ given in equation (4.6) would have to be revised.

4.6 Additivity of Distracters

Function b_Σ computes the additive spatial blurring resulting from the simultaneous presence of the sources. We studied the additivity of simultaneous distracters in experiment 4 (see section 3.9). We found a significant increase of spatial blurring as the number of distracters presented simultaneously increases, at least up to four distracters. From this experiment, we can observe the spatial blurring created by distracter(s) through the increase in the resulting SNR when the distracters are present, that is, by subtracting the SNR obtained in quiet from the SNR obtained in the presence of the distracters under consideration. The result of this operation is depicted in **figure 4.6**. Let us call Δ_c the difference observed in condition c. It is recalled that the SNR resulting from a given condition is the SNR between the target and any of the distracters, all distracters being presented at the same, fixed, loudness-matched level. To model b_Σ we must find a prediction function f such that:

$$\begin{aligned} \Delta_5 &= f(\{\Delta_2, \Delta_3\}), & (4.7)\\ \Delta_6 &= f(\{\Delta_1, \Delta_4\}), & (4.8)\\ \Delta_7 &= f(\{\Delta_1, \Delta_2, \Delta_3, \Delta_4\}), & (4.9)\\ \Delta_7 &= f(\{\Delta_5, \Delta_6\}). & (4.10) \end{aligned}$$

Note that Δ_7 can be predicted in two ways: either from Δ_1, Δ_2, Δ_3 and Δ_4, or from Δ_5 and Δ_6. We investigated different types of additivity between distracters. First, the resulting effect might be that of the distracter producing the greatest effect, that is:

$$f(\{\Delta_c, c \in C\}) = \max(\Delta_c, c \in C), \tag{4.11}$$

where C is the set of conditions under consideration. A plot of this prediction model is given in **figure 4.7**. In this figure, the predicted effect of the distracters is plotted against the observed effect, such that the closer data points are to the diagonal line—denoting "predicted equals observed"—the more accurate is the prediction. We can see that the data points are scattered away from the diagonal, mainly below it, which suggests an underestimation of additivity. Moreover, Pearson's linear correlation coefficient yields $r(58) = .69$ ($p < .001$) for a root mean square error of 8.0 dB. All of this confirms our observation concerning experiment 4 that spatial blurring does increase with the addition of distracters, even weaker ones.

A second option is that distracters might add as statistically independent sources (which they actually are). In this case:

$$f(\{\Delta_c, c \in C\}) = 10 \log_{10} \sum_{c \in C} \left(10^{\Delta_c/20}\right)^2. \tag{4.12}$$

This second formula seems to better match our experimental results, as can be seen in **figure 4.8**. Indeed, Pearson's linear correlation coefficient yields $r(58) = .74$ ($p < .001$) for a root mean square error of 6.9 dB. However, this prediction model still has a tendency to underestimate the additivity of distracters. Actually, additivity is predicted more accurately by assuming that the distracters are correlated sources (which they are not):

$$f(\{\Delta_c, c \in C\}) = 10 \log_{10} \left(\sum_{c \in C} 10^{\Delta_c/20}\right)^2. \tag{4.13}$$

Figure 4.9 gives a plot of the predicted effect of distracters against the observed effect for all subjects assuming correlated distracters. Predicted and observed effects are much better correlated using this model of additivity: Pearson's linear correlation coefficient reaches $r(58) = .75$ ($p < .001$) for a root mean square error down to 6.6 dB. This was unexpected. As a consequence, spatial blurring created by distracters seems to over-add,

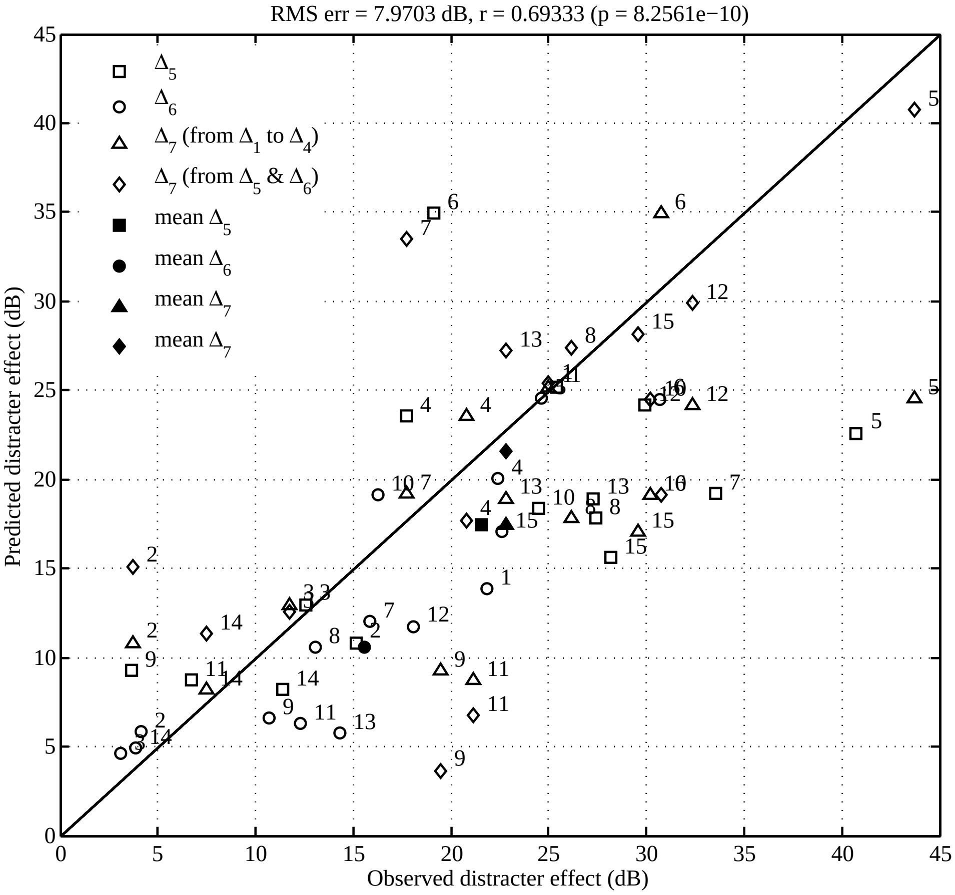


Figure 4.7: Plot of the predicted effect of distracters against the observed effect from experiment 4, assuming that the resulting effect is that of the distracter producing the greatest effect (see equation (4.11)). Δ_c is the difference between the SNR obtained in quiet (condition 0) and the SNR obtained in the presence of the distracters under consideration in condition c. The empty symbols represent the values from individual subjects, whereas the filled symbols represent the mean over all fifteen subjects.

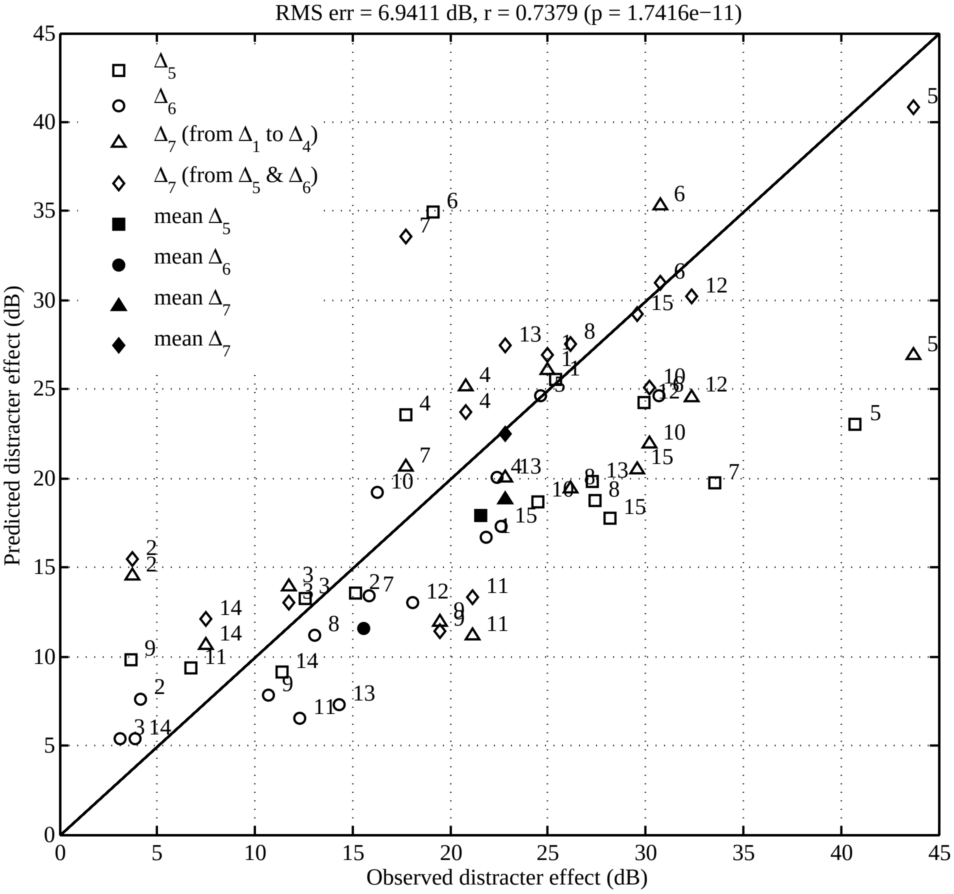


Figure 4.8: Plot of the predicted effect of distracters against the observed effect from experiment 4, assuming statistically independent distracters (see equation (4.12)). Δ_c is the difference between the SNR obtained in quiet (condition 0) and the SNR obtained in the presence of the distracters under consideration in condition c. The empty symbols represent the values from individual subjects, whereas the filled symbols represent the mean over all fifteen subjects.

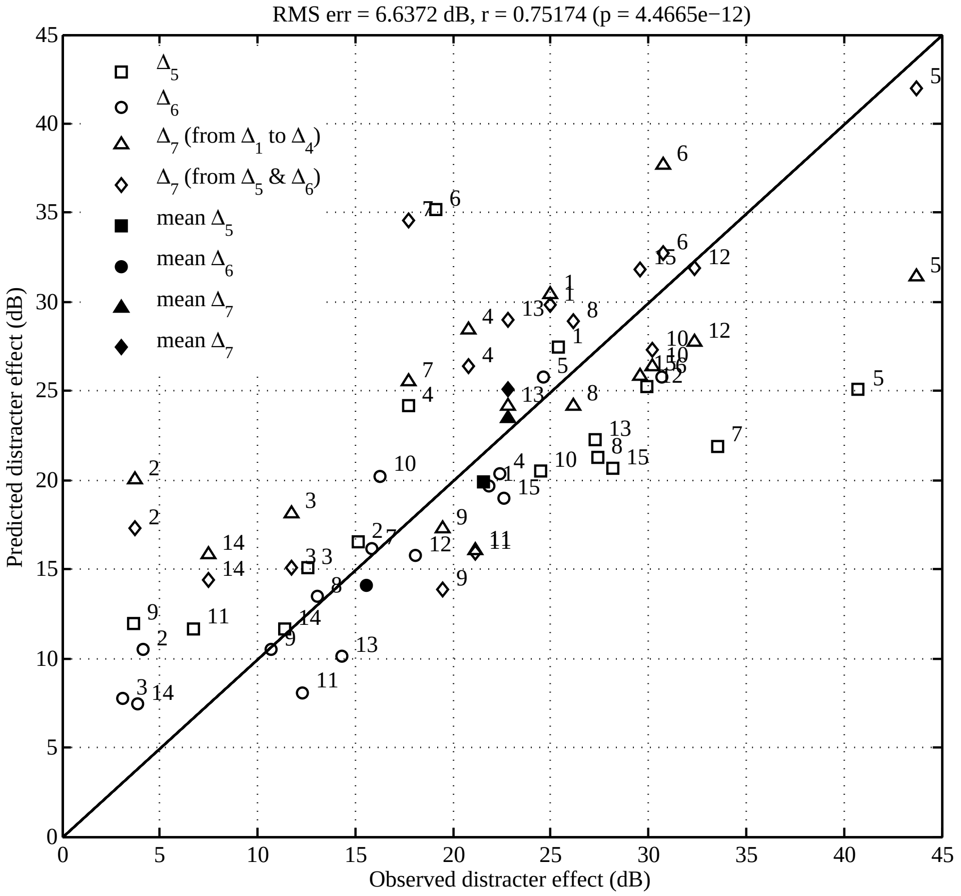


Figure 4.9: Plot of the predicted effect of distracters against the observed effect from experiment 4, assuming correlated distracters (see equation (4.13)). Δ_c is the difference between the SNR obtained in quiet (condition 0) and the SNR obtained in the presence of the distracters under consideration in condition c. The empty symbols represent the values from individual subjects, whereas the filled symbols represent the mean over all fifteen subjects.

since distracters are in reality statistically independent, which suggests an interaction between them. We have not yet investigated the possible reasons behind such an additivity rule.

Note also that additivity obtained in condition 7 seems to be less accurately predicted from conditions 5 and 6 than from conditions 1 to 4 (the diamond symbols in **figure 4.9** are farther away from the diagonal line than the triangular symbols), which suggests an overestimation of the additivity in the first of these two cases. This can probably be attributed to the fact that in this case, the distracters to be added have a bandwidth exceeding a single critical band. Interestingly, as can be seen on **figure 4.8**, this specific case—and therefore the additivity of wideband distracters—seems to be best predicted using the rule from equation (4.12).

In experiment 4, all distracters in a given condition were presented with the same loudness-matched level. We will assume that the additivity-prediction rule stated in equation (4.13) applies to distracters with non-loudness-matched levels as well. Besides, we will assume that for the range of SNRs concerning experiment 4, the relation between spatial blurring and SNR is linear, since we do not have enough data from experiment 2 to assess more precisely this relation. We will thus assume that an increase in SNR for a given angular separation is equivalent to a proportional increase in spatial blurring for a fixed SNR. Consequently, we can state, given a target s_t and a set of distracters $\{s_{d_p}, p \in 1..P\}$, that:

$$b_\Sigma(s_t, \{s_{d_p}, p \in 1..P\}) = \sum_{p=1}^{P} b_\tau(s_t, s_{d_p}) + \delta, \text{ with } \delta > 0, \tag{4.14}$$

where δ is the additional spatial blurring resulting from the over-additivity of distracters. If the relation between SNR and spatial blurring was further shown to be non-linear, equation (4.14) would have to be adjusted accordingly. If distracter effects had been shown to add as independent sources—that is, according to equation (4.12)—δ would have been null. There is no direct approach that allows us to derive an expression of δ from our experimental results. However, another way to take into account this over-additivity is to correct the amplitude level associated with each of the distracters by a factor γ prior to assessing the effect of SNR on spatial blurring with b_τ. Let us define $i_p = 10^{I_p/20}$ as the amplitude level of distracter s_p, where I_p is its associated energy level. Because distracters are independent but seem to have effects that add as correlated sources, according to equations (4.12) and (4.13) we are looking for the expression of γ verifying:

$$\sum_{p=1}^{P} (\gamma\ i_p)^2 = \left(\sum_{p=1}^{P} i_p \right)^2, \text{ with } \gamma > 0. \tag{4.15}$$

That is:

$$\gamma = \frac{\sum_{p=1}^{P} i_p}{\sqrt{P \sum_{p=1}^{P} (i_p)^2}}. \tag{4.16}$$

We can thus define I'_p as the corrected energy level from distracter s_p:

$$\begin{aligned} I'_p &= 20 \log_{10}(\gamma\ i_p) && (4.17) \\ &= I_p + \mathit{\Gamma}, && (4.18) \end{aligned}$$

with

$$\mathit{\Gamma} = 20 \log_{10}(\gamma) = 20 \log_{10} \frac{\sum_{p=1}^{P} i_p}{\sqrt{P \sum_{p=1}^{P} (i_p)^2}}. \tag{4.19}$$

The spatial blurring created on a target s_t by a set of distracters $\{s_{d_p}, p \in 1..P\}$ is finally given by:

$$b_\Sigma(s_t, \{s_{d_p}, p \in 1..P\}) = \sum_{p=1}^{P} b_\mathcal{T}(s_t, s'_{d_p}), \tag{4.20}$$

where s'_p denotes distracter s_p with corrected energy level I'_p. Given that experiment 4 showed this additivity effect up to four distracters, a potentially necessary precaution might be to limit this expression to the four greatest effective distracters.

4.7 Resulting Localization Blur

The last function $\Omega(s_t, \{s_{d_p}, p \in 1..P\})$ returns the localization blur associated with a target s_t in the presence of a set of distracters $\{s_{d_p}, p \in 1..P\}$ based on the spatial blurring they create on s_t, obtained by a call to $b_\Sigma(s_t, \{s_{d_p}, p \in 1..P\})$, and on the localization blur in quiet $\alpha_{\min}(s_t)$ associated with s_t.

In section 3.2, we defined spatial blurring as the difference between localization blur in the presence of distracters and localization blur in quiet, such that:

$$\Omega(s_t, \{s_{d_p}, p \in 1..P\}) = b_\Sigma(s_t, \{s_{d_p}, p \in 1..P\}) + \alpha_{\min}(s_t). \tag{4.21}$$

A value of $\alpha_{\min}$ can be computed for any target frequency by interpolating linearly on an ERB scale between each of the three experimental values of $\alpha_{\min}$ we obtained in experiment 1, that is, 1.5° at 700 Hz, 5.1° at 1400 Hz, and 3.9° at 5000 Hz.

4.8 Simplification of the Model

When MAAs are not to be used absolutely, but rather relatively (as in the coding scheme proposed in section 5.1), this model can be roughly approximated. Indeed, the main result from the estimations yielded by this model is that sources with greater associated energy levels are more likely to have a smaller associated localization blur. Therefore, in practical situations where complexity is critical, this model can be greatly simplified by associating with a given source a localization blur that is inversely proportional to its audible energy—that is, above masking threshold. This avoids the assessment of the spatial blurring created by each of the distracters on each target, which represents a great deal of decrease in complexity.

Moreover, by looking at **figure 4.2**, one can see that spatial blurring seems to be correlated to localization blur in quiet, that is, the greater the localization blur in quiet, the more important spatial blurring seems to be, whatever the distracter under consideration. Consequently, this simplified model based on energy could be refined by weighting the estimated localization blur with the value of $\alpha_{\min}$ corresponding to the target.

4.9 Conclusions

This chapter provided a modeling of spatial blurring within the limits of our experimental data. The interest of the approach we proposed is that it divides the modeling of a complex phenomenon—spatial blurring—involving several variables (target center frequency, distracter center frequency, SNR, number of distracters) into the modeling of three simpler effects involving less variables. First, spatial blurring with loudness-matched target and distracter (SNR of 0 dB), involving solely the modeling of the target center frequency and the distracter relative center frequency. Second, the modeling of the effect of SNR on spatial blurring, involving, as we assume, only SNR and the spatial blurring created with an SNR of 0 dB. And finally, the modeling of the additivity of distracters, involving the number of distracters and the spatial blurring that each of them create when presented alone. As a result, this approach does not require the acquisition of psychoacoustic data that test all these variables simultaneously, which would be difficult in practice.

Further studies are needed to refine and correct this model. For instance, the temporal aspect of distracting sources was not considered, but recent investigations by Croghan and Grantham [CG10] using pulsating and non-pulsating interferers suggest that spatial blurring is greatest with pulsating sources compared to non-pulsating sources. Studies by Kopčo *et al.* [KBS07] showed that non-simultaneous distracters can also affect localization, which suggests that localization blur might be affected as well. Moreover, the perception of sources in terms of auditory scene analysis (see section 1.3) is not accounted for either in our model. For example, it is not clear if two sources—in the sense of our model—fused by the auditory system can be considered as potential distracters for each other or not. Also, these fusion phenomena might increase localization blur, because the auditory system seems to derive a location only after the fusion process (see section 1.5). Therefore, our present model is a low-level model that might need to be modified to include the results of higher-level auditory processes.

Chapter 5

Multichannel Audio Coding Based on Spatial Blurring

As briefly presented in section 2.2.2 and reviewed in more detail in appendix C, perceptual coding takes advantage of interference between frequency components, exploiting energy masking phenomena. To extend this concept to multichannel audio signals, the phenomenon of binaural unmasking has been taken into account, which resulted in a decrease of the masking ability of the signal components, and consequently limited the bitrate reduction. The coding of the spatial attribute of a sound scene has a cost, as explained in chapter 2, and strategies (based on auditory perception or not) have already been proposed to reduce this cost. However, interference between components in terms of their perception in space by the auditory system has not been considered so far in spatial audio coding. Indeed, we have seen from our experiments presented in chapter 3 that auditory spatial resolution is likely to decrease as the sound scene gets complex, that is, when several sources are present simultaneously. In this chapter, we put forward two spatial audio coding schemes that take advantage of this decrease in resolution to ease the precision of representation of the spatial sound scene, leading to a bitrate reduction. The first scheme applies in the context of parametric spatial audio coding (see section 2.5), while the second one concerns the HOA representation (see section 2.1.2). As in chapter 3, we will limit our reasoning to the azimuthal plane, but this work could be extended to elevation as well.

Localization Blur in Spatial Audio Coding

In the two coding schemes we will consider in this chapter, the accuracy of representation of the spatial attribute of the scene has a cost. For instance, in parametric spatial audio coding schemes, space is explicitly encoded as a set of discrete spatial parameters. The cost is in the precision of coding of these parameters, upon which depends the precision of representation of space. In order to reduce the necessary transmission bitrate, these parameters are quantized and coded with only few bits, and thus the representation of space is degraded. In HOA, the cost is in the order of representation of the scene. The higher this order, the larger the area of accurate reconstruction of the sound scene, but also the larger the number of coefficients to encode. Therefore, the representation is truncated to a certain order to reduce the necessary transmission bitrate.

In both cases—quantization or truncation—the consequence is a degraded spatial representation of the components of the sound scene which, at the listening phase, are perceived as changes in position. As localization blur characterizes the sensitivity to position

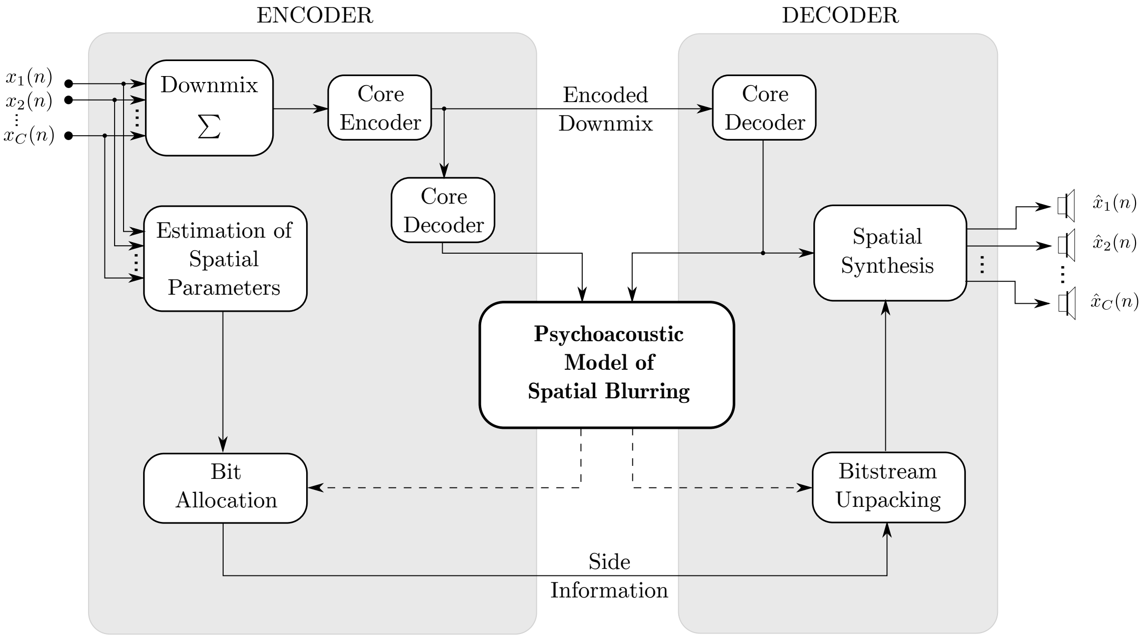


Figure 5.1: Overview of the proposed parametric spatial audio coding scheme based on spatial blurring.

changes of sound events, if the erroneous change in position of a given component is greater than its associated localization blur, it will be perceived by the auditory system. On the contrary, if this change in position occurs within localization blur, it will not be perceived. For this reason, localization blur could play an important role in spatial audio coding. Moreover, the experiments we carried out (see chapter 3) have shown that a given component of the sound scene is susceptible to increase the localization blur associated with other components in its frequency vicinity. This is why we propose to use localization blur resulting from spatial blurring in exactly the same way masking thresholds resulting from energetic masking are used (see appendix C): by shaping spatial distortions to contain them (as much as possible) within localization blur, such that they are not perceived by the listener. Therefore we will use our psychoacoustic model of spatial blurring (described in chapter 4) to estimate localization blur at each time instant, and this will determine the spatial distortion we can afford for each component of the sound scene. This strategy thus implies a dynamic adjustment of the accuracy of representation of the sound scene. The interest can be seen in two ways: either the perceived quality remains the same (transparent coding) and the bitrate is reduced whenever possible, or the same bitrate is kept (constraint bitrate coding), but the perceived quality is improved.

5.1 Dynamic Bit Allocation in Parametric Schemes

A patent is pending for this work [DN11].

5.1.1 Principle Overview

Three parametric spatial audio coding schemes have been presented in section 2.5: BCC (along with PS), SASC and DirAC. Potentially, the dynamic bit allocation process we propose in this section can be inserted into any of these schemes, as depicted in **figure 5.1**.

The input channels are downmixed into a mono (or stereo) channel(s) in a way that ensures the conservation of total energy from all channels. Also a set of spatial parameters is computed from the input channels within each of a group of frequency subbands. [1] These spatial parameters interest us as they provide a description of the spatial attribute of the scene. They are coded with the same precision in all subbands and at each time instant, which is not optimal from a perceptual point of view since localization blur is likely to differ from one subband to another due to spatial blurring and is also likely to vary in time.

The optimization principle we propose is to use our psychoacoustic model of spatial blurring to guide the bit-allocation procedure of the spatial parameters such that the spatial distortions resulting from this quantization are shaped according to localization blur. This new quantization process is thus dynamic, since quantization will vary in time and depending on the context (the content of the input signals).

An original point in this scheme is that the psychoacoustic scheme is used both on encoding and decoding sides. As we will explain in section 5.1.4, this trick will save us the transmission of bit-allocation information of the spatial parameters.

Both the quantized spatial parameters and the downmix signal are then transmitted. The downmix signal can optionally be encoded using any external core coder (such as MP3, AAC, etc.) prior to the transmission. At decoding, the spatial parameters are dequantized, the downmix signal is decoded, if necessary, and these are used to synthesize a set of output channels.

5.1.2 Use of our Psychoacoustic Model of Spatial Blurring

The spatial parameters concerned by this dynamic bit allocation procedure are any parameter describing the angular azimuthal position of a sound source. For DirAC, it is straightforward since the direction vector $\mathbf{D}$ represents the direction of the source and is coded in practice as an angle. Therefore $\mathbf{D}$ can benefit from our dynamic bit allocation. In SASC, the primary and ambient localization vectors $\mathbf{d_P}$ and $\mathbf{d_A}$ are usually coded in polar coordinates, that is as angle and radius. The bit allocation for the angular value of each vector can thus be optimized. Finally, in BCC the angular position of the source is not directly represented by any of the spatial parameters; it is rather the localization cues used by the auditory system to derive this position that are coded. But still, a quantization error in an ICLD or an ICTD value will be perceived as a change in the angular position of the source, and therefore our dynamic allocation process can be applied here as well. In the following, we will not distinguish between the three considered coding schemes, and potentially other schemes might be concerned as well.

As explained in section 2.5.1, the frequency subbands within which spatial parameters are computed following a set of adjacent critical bands. The spatial parameters associated with a subband describe a single position, and this because it is assumed that the listener cannot discriminate two simultaneous sources in space, which spectrally belong to a same critical band while spatially arising from different locations. Since our psychoacoustic model defines a source as energy in a critical band (see section 4.1), we can directly use it to estimate the MAA associated with each subband, considered as a source. More exactly, if the primary/ambient energy ratio of each subband is known, each subband consists of two sources: a primary source and an ambient source. Depending on the concerned coding scheme, this ratio is known if the following parameter is computed and transmitted: the

[1] The time-frequency transform used for this purpose is not represented in **figure 5.1**. More information about this transform are given in section 2.5.1.

inter-channel coherence c for BCC, the diffuseness coefficient Ψ for DirAC, and the ambient energy fraction λ in SASC. Otherwise, each subband consists of a single source, considered as primary. The submission of this additional information to our psychoacoustic model has two consequences. First, it allows the model to treat primary and ambient components separately instead of pooling together the energy from both components, thus avoiding an overestimation of the energy associated with a given source. But this also refines the estimation of the MAA, because spatial blurring might vary depending on whether the target and the distracters are primary or ambient components. [2]

As justified in section 4.1, neither the position of the source under estimation, nor the positions of the distracting sources are needed by the model to yield an estimation of MAA; only the energy associated with each source is required. Indeed, our psychoacoustic results showed that the position of a given source does not have an effect on the spatial blurring that this source creates, and in our model the target source is assumed to be in front of the listener. Consequently, the MAA associated with a given source can be estimated solely from the downmix signal, since the total energy from the original input channels is conserved. It is important for this estimation to be done after encoding-decoding by the external core coder (if any), as explained in section 5.1.4.

Each source s is characterized by three parameters $\{f_c, I, t\}$, where f_c is the center frequency of the subband to which the source belongs, I is the associated energy level of the source, and t is its type (primary or ambient). If K is the total number of sources, the psychoacoustic model is called via $\psi\left(\{s_k, k \in 1..K\}\right)$. It returns an estimation of the MAA associated with each source s_k considered as a target, based on the spatial blurring created on it by all other $K-1$ sources, considered as simultaneous distracters.

As a result, the psychoacoustic model yields two estimations per subband, one corresponding to the MAA associated with the primary component of the subband, and the other one corresponding to the MAA associated with the ambient component. However the latter is exploitable only in SASC because it is the only scheme transmitting parameters that describe the spatial position of ambient components. For the two other schemes, the benefit is a more accurate estimation of MAAs, but only primary components will be considered for the dynamic bit-allocation procedure. In any case, note that the number of sources remains the same from one temporal frame to another; it equals either once or twice the number of subbands, depending on whether the primary/ambient energy ratio of each subband is known.

5.1.3 Bit Allocation of the Spatial Parameters

Bit allocation is the module that optimizes the representation of space as a function of the available bitrate. The bit-allocation process is split into two parts. The first part consists of a certain number of bits equally distributed among all sources—we will refer to them as “fixed” bits—this basic allocation ensuring a minimal quality of spatial reproduction for all sources. The second part however consists of bits distributed as a function of the associated localization blur of each source—we will refer to them as “floating” bits.

For the sake of simplicity, we will illustrate the bit-allocation strategy assuming a single parameter to code for each source, but the process remains the same for other parameters. To lighten the language a bit, we will sometimes use expressions similar to “to code source s”, which is shorthand for “to code the parameter associated with source s”. As this process is applied identically and independently for each temporal frame, we

[2] However, as explained in chapter 4, further psychoacoustic experiments would be necessary to see if the primary or ambient nature of components affects spatial blurring.

will only describe it for a single frame. Once again, if parameters specifically describing the position of the ambient component of each subband exist (which is the case in SASC), then primary and ambient sources are both concerned by this procedure—meaning that the bit pool is shared between primary and ambient sources—otherwise only primary sources are concerned, since there is no parameter to code for ambient sources. Let us introduce the following variables:

K: total number of sources (primary and ambient);
P: number of sources to code;
N: total number of bits to allocate;
n_{fixed}: minimum number of bits allocated to each source's parameter (i.e., the number of fixed bits);
s_p: p^{th} sound source to code, $p \in \{1, \ldots, P\}$;
α_p: MAA associated with the p^{th} source in the presence of the $K-1$ other sources, as given by the psychoacoustic model;
n_p: number of floating bits allocated to the parameter of s_p;
$n'_p = n_{\text{fixed}} + n_p$: total number of bits allocated to the parameter of s_p.

First of all, we need to express the consequence of allocating an additional bit for the precision of coding of the parameter. The precision of coding of a parameter is given by the number of values over which it is quantized. It is possible to represent 2^n quantization values with n bits, so whatever the distribution of these quantization values is (uniform or not), we will consider that adding a coding bit doubles the number of quantization values. [3] Thus adding a coding bit doubles the precision of representation of the parameter.

We propose to avoid trying to link directly quantization errors with angular changes in position. The first reason for this choice is because this link might change depending on the nature of the concerned spatial parameter, but besides, it would be unnecessarily complicated. The basic idea of our dynamic bit allocation is the following. Instead of considering the MAA associated with each source absolutely and trying to allocate the necessary number of coding bits to each source to ensure that spatial distortions stay within localization blur, we will rather simply ensure that the smaller the MAA associated with a given source, the larger the number of bits allocated to it. In other words, the source with the smallest associated MAA will be coded the most precisely, and the one with the largest associated MAA will be coded the most roughly. In between, sources will be coded more or less precisely depending on their associated MAA. If the total number of bits available is sufficient, then spatial distortions will be constrained within localization blur. If not, spatial distortions will be perceptually minimized in an homogeneous fashion among all sources. This approach is especially necessary in a situation of constrained bitrate, which we will describe first. However, the unconstrained bitrate case, studied afterwards, also benefits from the simplicity of this approach.

Constrained bitrate

Having a constrained bitrate means that besides the number of fixed bits, the number of floating bits is constrained too. So let us call N_{float} the number of floating bits we have to distribute between all sources according to the psychoacoustic model. As stated above, the source with the smallest associated MAA—let us call its index m—should be coded with the maximum precision, which then equals 2^{n_m}. Moreover, the relative coding

[3] If this assumption is not verified, equations (5.1) and (5.7) stated thereafter must be adjusted consequently.

precision between s_m and any other source s_p should be inversely proportional to their relative MAA. This can be expressed as:

$$\frac{2^{n_p}}{2^{n_m}} = \frac{\alpha_m}{\alpha_p}, \text{with } n_p, n_m \in \mathbb{N}^+, \text{and } \alpha_p, \alpha_m \in \mathbb{R}^{+*}. \tag{5.1}$$

Thus:

$$n_p = n_m + \log_2\left(\frac{\alpha_m}{\alpha_p}\right). \tag{5.2}$$

Moreover, the sum of floating bits from each source must equal the total number of available floating bits N_{float}:

$$\sum_{p=1}^{P} n_p = N_{\text{float}}. \tag{5.3}$$

Thus, by replacing in this equation the expression of n_p given in equation (5.2), we obtain:

$$n_m = \frac{N_{\text{float}} - \log_2\left(\prod_{p=1}^{P} \frac{\alpha_m}{\alpha_p}\right)}{P}. \tag{5.4}$$

Equations (5.2) and (5.4) above give a first approximation of the number of bits to allocate to the parameter of sources s_p and s_m, respectively. Since all n_p have to be integers, after rounding, some bits to allocate might remain ($\sum_{p=1}^{P} n_p < N_{\text{float}}$), or too many bits might have been allocated ($\sum_{p=1}^{P} n_p > N_{\text{float}}$). The following heuristic—a greedy type algorithm—allows the allocation process of floating bits to be finalized [4] with an iterative procedure. Let us define Δ_p from equation (5.1) as a distance between the optimal bit allocation [5] for source s_p and its current bit allocation:

$$\Delta_p = \frac{\alpha_m}{\alpha_p} - \frac{2^{n_p}}{2^{n_m}}. \tag{5.5}$$

Thus Δ_p reflects how much precision is lacking in the coding of s_p. At each iteration, the index of the source to/from which the next bit should be allocated (or taken back) is determined by $\text{argmax}_p\Delta_p$ (or $\text{argmin}_p\Delta_p$). Δ_p is computed again after each operation on one bit. The allocation process is finalized when the total number of allocated floating bits equals exactly N_{float}.

It might happen at some point that Δ_p is null for all p whereas the total number of allocated bits does not equal N_{float}. This means that the allocation is optimal given the number of allocated bits although the allocation process is not finished yet, because bits remain to be allocated or withdrawn. In that case, the source that should receive (or from which should be taken back) the next bit is the source with the smallest (or greatest) associated MAA. Then the iterative process can go on until N_{float} bits have been allocated.

Finally, the total number of bits n'_p allocated to the coding of the parameter of source s_p is given by:

$$n'_p = n_{\text{fixed}} + n_p. \tag{5.6}$$

[4] Note that the whole allocation process can be done using this algorithm exclusively.

[5] The optimal bit allocation is reached when equation (5.1) is satisfied for all sources.

Unconstrained bitrate

In an unconstrained bitrate context, a specific coding quality is targeted. Let us define a fictive source $s_{\bar{\alpha}}$, which will serve as a reference. The MAA associated with this fictive source ideally is the mean MAA $\bar{\alpha}$ observed over all sources and temporal frames of the signal. In practice, $\bar{\alpha}$ can be either assumed or estimated (over the next few temporal frames, for instance). The coding quality is then defined by the number of bits $\bar{n}$ allocated to $s_{\bar{\alpha}}$, and this choice is up to the user. The aim of this bit allocation procedure is to obtain, depending on the accuracy of the estimation of $\bar{\alpha}$, a practical mean number of bits allocated to each source close to $\bar{n}$.

A source with an MAA equal to that of the reference source $s_{\bar{\alpha}}$ will thus be coded with $\bar{n}$ bits, that is, with a precision of $2^{\bar{n}}$. Then a source with an associated MAA greater than $\bar{\alpha}$ should be coded less precisely than $s_{\bar{\alpha}}$, and *vice versa*. Therefore, the relative coding precision between $s_{\bar{\alpha}}$ and any other source s_p should be inversely proportional to their relative MAA. This can be expressed as:

$$\frac{2^{n_p}}{2^{\bar{n}}} = \frac{\bar{\alpha}}{\alpha_p}, \text{with } n_p, \bar{n} \in \mathbb{N}^{+*}, \text{and } \alpha_p, \bar{\alpha} \in \mathbb{R}^{+*}. \tag{5.7}$$

The number of floating bits to allocate to the parameter of source s_p is thus given by:

$$n_p = \bar{n} + \log_2\left(\frac{\bar{\alpha}}{\alpha_p}\right). \tag{5.8}$$

Equation (5.6) gives the total number of bits to allocate to the parameter of source s_p.

5.1.4 Transmission and Bitstream Unpacking

Once the bit allocation for each parameter is known, parameters can be quantized and transmitted. As quantization is now dynamic due to the account for spatial blurring, it would be necessary *a priori* to transmit more information about how each parameter has been quantized—the number of bits on which each parameter has been coded—in order to properly dequantize them at decoding. The cost resulting from the transmission of this additional information would not be affordable given the very small number of bits these parameters are aimed to be coded on. Nevertheless, since localization blur can be estimated solely from the downmix signal, it is possible to run at decoding the same bit allocation process as the one run at encoding to know exactly how many bits were allocated to each parameter. The counterpart of this solution is that the psychoacoustic model has to be known by the decoder.[6] Moreover, if the downmix has been encoded with an external core coder prior to transmission, the psychoacoustic model has to be fed at encoding with the downmix signal encoded then decoded by this core coder to ensure that the same downmix signal is submitted to it at decoding, and thus leads to the same bit allocation.

If it is transmitted, the primary/ambient energy ratio—which is coded on a fixed number of bits—has to be transmitted before any dynamically quantized parameter. Indeed, this parameter is necessary to separate the energy part associated with the primary and ambient components. Then information about each source can be submitted to the psychoacoustic model at decoding prior to dequantizing the dynamically quantized parameters.

[6] The real implication is that any modified version of the psychoacoustic model in the encoder needs its specific revised decoder.

A last interesting possibility is that if n_{fixed} is nonzero, it is possible to get a first approximation of each of the spatial parameters without knowing the number of allocated bits to each of them. Indeed, if the bitstream is organized such that the most significant n_{fixed} bits of each source are sent first, followed by the remaining n_p floating bits of each source, then these n_{fixed} most significant bits approximate each transmitted value. This might be useful in case any further experimental study would show that some basic information about source position is in fact necessary to estimate MAAs more accurately. In such a case, the downmix signal would no longer be enough for MAA estimation, and these approximate parameter values would help maintain the possibility of MAA estimation at decoding (meaning that at encoding MAA estimation has to be done using the same approximate parameter values). Therefore, the larger n_{fixed}, the better the approximation of the parameters.

5.1.5 Informal Listening

We carried out informal listening tests to get a first idea of advantages and drawbacks of our dynamic bit-allocation approach in comparison with a static bit allocation—that is, when a parameter is coded with the same number of bits for all sources. We implemented our procedure within a Parametric Stereo coding scheme (see section 2.5), using a short-time Fourier transform (STFT) with temporal frames of 10-ms duration. Inter-Channel Coherence (ICC) was not transmitted, and the effect of the quantization of either Inter-Channel Level Difference (ICLD) or Inter-Channel Phase Difference (ICPD) was assessed separately. This means that when ICLD was quantized, ICPD was not, and *vice versa*. ICLD and ICPD were quantized according to results given in section 1.4.5, as explained in section 2.5.4. Thus ICPD was uniformly quantized, and for ICLD, quantization steps increased as the reference ICLD increased. ICLD parameters were computed for subbands over the full bandwidth (up to 22050 Hz), whereas ICPD parameters were only computed for subbands up to 1500 Hz (the 11 first subbands), as is usually the case given that phase becomes ambiguous for higher frequencies. No subband pre-selection by energetic masking was applied, thus parameters of all subbands were coded. We chose to test the constrained bitrate allocation procedure only, applied with different sizes of bit pool. n_{fixed} was always equal to zero, so only N_{float} was varied in our tests. Two types of signals were tested: intensity stereo panning music and binaural recordings of sound scenes.

A general result is that our dynamic bit-allocation procedure is especially interesting for very low bitrates. Indeed, assuming a symmetrical scalar quantizer,[7] one or two bits[8] at least per parameter are technically necessary when using a static bit-allocation procedure. However, when using our dynamic procedure, a basic stereo image can be obtained even with less than a single bit per parameter in average. In fact, bits will be allocated only to the perceptually most important subbands, while the remaining subbands will not get any bits.

Concerning the bit allocation for the ICLD parameter, mainly tested on intensity stereo panning music, in general our dynamic procedure produced an equivalent stereo image with fewer bits compared to the static bit allocation. In other words, the stereo image is richer with our procedure when the same number of bits as in the static bit allocation is used.

[7] Note that we do not consider vector quantization, differential coding or other advanced methods, the purpose here being to compare dynamic and static bit-allocation procedures.

[8] Two bits if the quantizer includes the zero output value (which was the case in our tests), one bit otherwise.

However, our dynamic procedure produces artifacts in certain cases when applied to ICLD, especially when all of the energy of an auditory source—in the ASA sense (see section 1.3)—is panned on only one of the two channels (situation of hard-panning). In that case, the fact that some subbands of the source, which should only be present in one channel, are also present in the other one is perceived as some sort of incomplete phantom source. This effect is almost imperceptible when listening over loudspeakers, though. This is probably due to the fact that when listening over headphones, such a phantom source is perceived at one ear, whereas the complete source is perceived at the other ear. Hence, the complete source is fused by the auditory system at one ear, whereas at the other one the fusion operates only on the energy of this source that is erroneously present. As a consequence, two different sources are perceived, one at each ear. When listening over loudspeakers, crosstalk greatly attenuates this effect. In fact, it seems that fusion happens between the incomplete source and crosstalk, such that in the end the incomplete source is fused with the complete one, and a single source is perceived. This flaw can probably be attributed to the fact that the ICLD parameter of each of the originally hard-panned subbands is coded with zero bits, meaning that ICLD is considered null in these subbands, and thus energy is equally present in the two channels. This can be avoided by setting a nonzero value to n_{fixed}, but this reduces the dynamic aspect of the bit allocation. Therefore, the setting of our dynamic procedure for ICLD is a trade-off between n_{fixed} and N_{float}.

The bit allocation for the ICPD parameter was mainly tested on the binaural recordings, since the information of phase difference is essential for this type of signal. From a general point of view, quantization of ICPD induces audible artifacts with both methods, although it seems that with fewer bits, our dynamic allocation procedure produces less artifacts compared to the static bit allocation. This tendency gets stronger when using the same number of bits as in the static bit allocation.

As a conclusion, our approach seems to provide interesting results, even if there is room for improvement. Of course, as the number of bits allocated to each parameter on average is increased, differences between the two allocation methods become less perceptible, since the number of bits is enough to keep spatial distortions within localization blur even with static bit allocation.

5.2 Dynamic Truncation of the HOA Order

Part of the following was published in the proceeding of the 40th AES conference [DNM10], and is patent-pending [DN10].

This second coding scheme based on spatial blurring takes advantage of the HOA representation to shape spatial distortions according to localization blur. One advantage of HOA is that it represents accurately several spatially separated sound sources lying in a same critical band, which is not the case of parametric coders (see section 2.5.1). We will start by studying spatial distortions resulting from truncation of the order of representation using an approach based on simulations, and we will propose an adjusted notion of sweet spot based on perceptual criteria. Then the actual coding scheme by dynamic truncation of the HOA order will be described.

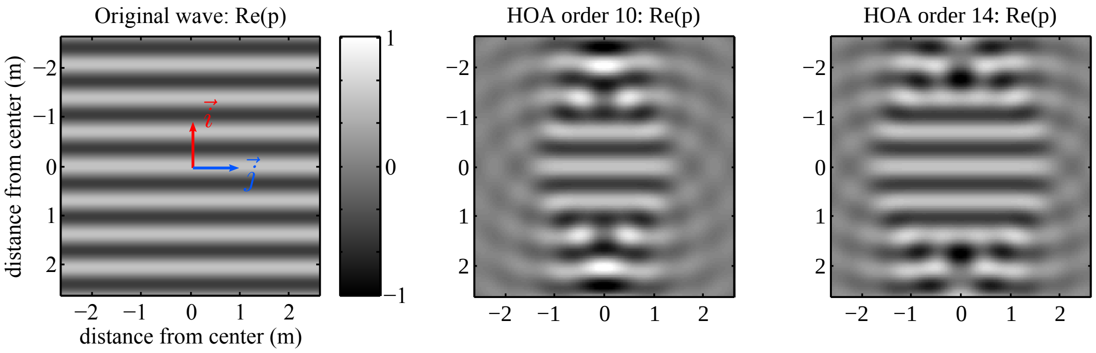


Figure 5.2: Simulation of the acoustic field in the HOA domain for the orders of representation 10 and 14 for a single 500-Hz plane wave arriving at 0° azimuth (that is, from the middle top of each graph). The real part of the pressure p is plotted. The non-erroneous area of representation of the acoustic wave widens as the order of representation increases.

5.2.1 Spatial Distortions Resulting from Truncation

The principle of HOA is detailed in section 2.1.2. HOA provides a representation of the acoustic wave expressed in the azimuthal plane as:

$$\begin{aligned} p(\vec{r},\omega) = B_{00}^{+1} j_0(kr) &+ \sum_{m=1}^{+\infty} j_m(kr) B_{mm}^{+1} \sqrt{2}\cos(m\varphi) \\ &+ \sum_{m=1}^{+\infty} j_m(kr) B_{mm}^{-1} \sqrt{2}\sin(m\varphi), \end{aligned} \tag{5.9}$$

where k is the wave number, r is the radius, ω is the angular velocity, ϕ is the azimuthal angle, $p(\vec{r},\omega)$ is the acoustic pressure, and $j_m(kr)$ are the spherical Bessel functions of the first kind. The B_{mm}^{σ} coefficients constitute the HOA representation of the acoustic wave. In practice, the HOA representation has to be truncated to a given order M, with $2M+1$ components B_{mm}^{σ} ($m = 0, 1, \ldots, M$; $\sigma = \pm 1$).

The accuracy of the field represented by the HOA signal depends both on the frequency bandwidth of the signal and on the order of truncation of the HOA representation. The error resulting from the truncation of the representation can be estimated from an acoustical point of view as:

$$\frac{|p(kr,\varphi) - p_M(kr,\varphi)|^2}{|p(kr,\varphi)|^2}, \tag{5.10}$$

A negligible error of representation of the field (less than 15 dB) is ensured as long as:

$$M \geq kr \tag{5.11}$$

The listening area within the centered circle of radius r is usually called the “sweet spot”. The sweet spot is thus a suitable listening area from an acoustical point of view. The evolution of the sweet spot as a function of the HOA order is depicted in **figure 5.2**.

Angular distortion and perceptual sweet spot

The truncation error may have various effects on the perception of the scene. Among others, a sound source might not be perceived at its original position. As proposed by Makita [Mak62], we can assume that the apparent direction of a sound source is that of

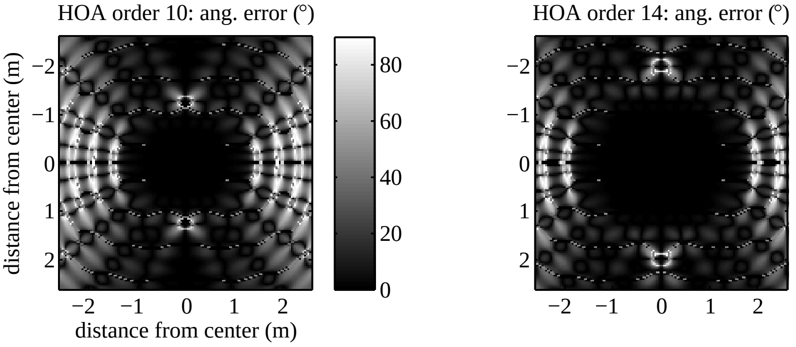


Figure 5.3: Angular error resulting from the truncation to orders 10 and 14 for a single 500-Hz plane wave arriving at 0° azimuth.

the normal to the wave front. This direction can be obtained by computing the gradient of the phase of the pressure p at several positions in the reconstructed acoustic field:

$$\overrightarrow{\nabla} \arg p = \frac{\partial \arg p}{\partial x}\vec{\imath} + \frac{\partial \arg p}{\partial y}\vec{\jmath} \tag{5.12}$$

Finally, if we consider a plane wave with a 0° azimuth of incidence, the error of incidence of the wave, e, is given by the angle between $\overrightarrow{\nabla} \arg p$ and the unit vector $\vec{\imath}$, assuming that $\vec{\imath}$ points toward 0° azimuth, and $\vec{\jmath}$ toward 90°:

$$e = \angle\left(\overrightarrow{\nabla} \arg p, \vec{\imath}\right). \tag{5.13}$$

Because the radius of the sweet spot depends on the signal bandwidth, the angular error of representation of a given wave field will be maximized by considering only the highest frequency component present in this field. Therefore, in our simulations, the field is composed of a single plane wave, which will represent this critical component. The angular error resulting from the truncation to orders 10 and 14 for a single 500-Hz plane wave is depicted in **figure 5.3**.

If we place ourselves in an audio coding context, we can assume that the original signal to encode is already truncated to a given order M_{ori}, and we are interested in the additional angular error of encoding e_{enc} resulting from a further truncation to order $M < M_{\text{ori}}$:

$$e_{\text{enc}} = \max(e_M - e_{M_{\text{ori}}}, 0), \tag{5.14}$$

where $e_{M_{\text{ori}}}$ and e_M are the angular errors resulting from truncation to order M_{ori} and M, respectively. If the angular error with order M is smaller than the angular error with order M_{ori}, it is considered that there is no additional error. The angular error of encoding resulting from the truncation to encoding orders ranging from 10 to 13, with an original order of representation of 14, for a single 500-Hz plane wave arriving at 0° azimuth is depicted in **figure 5.4**.

It might be interesting to compute this angular error of encoding along circles of increasing radius. The mean error along a circle of a given radius r can be computed as:

$$\epsilon_r = \left(\frac{1}{N}\sum_{i=1}^{N} e_i^{\gamma}\right)^{\frac{1}{\gamma}}, \tag{5.15}$$

where e_i are the encoding angular error values of the N points of the circle. Depending on the value of γ, more or less weight in the sum will be given to the high error values

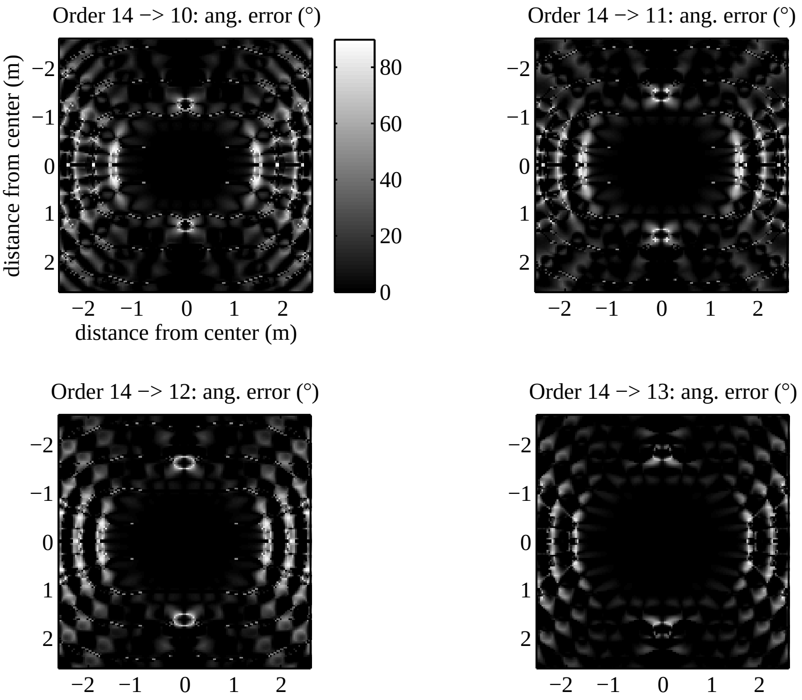


Figure 5.4: Angular error of encoding resulting from the truncation to encoding orders ranging from 10 to 13, with an original order of representation of 14, for a single 500-Hz plane wave arriving at 0° azimuth.

of the circle. For $\gamma = 1$, ϵ_r will represent the arithmetic mean error along the circle. For $\gamma = 2$, it will represent the quadratic mean, and if $\gamma = \infty$, it will represent the maximum error value on the circle. For our simulations, we chose a value of $\gamma = 1$.

Figure 5.5 illustrates these simulations for a 500-Hz plane wave and for encoding orders ranging from 10 to 13, with an original representation order of 14. For the sake of clarity, the represented angular error of a given radius r is the maximum error value of the averaged circles of radius less than or equal to r. It can be seen that globally the angular error gets larger as the encoding order decreases relative to the original order. Depending on the scene content and on the desired area of accurate reproduction, it is possible to select the minimum encoding order that ensures a large enough reproduction area within which the mean angular error of encoding is less than the MAA. As an example, in **figure 5.5**, if the desired radius of accurate reproduction is 2 m and the MAA is 15°, then the minimum order of encoding is 13.

Therefore it is possible with such simulations to compute a table giving, for a pair (r, α) representing the desired radius r of accurate reproduction and the MAA α, the corresponding minimum order of encoding for the scene. The disk of radius r can then be seen as a “perceptual sweet spot” representing a suitable listening area from a psychoacoustic point of view. In other words, the angular distortion resulting from the encoding will not be perceived by the listener within the perceptual sweet spot.

5.2.2 Principle Overview

The coding scheme is given in **figure 5.6**. To give an overview of the principle, a multichannel audio signal encoded in the *Higher Order Ambisonics* (HOA) domain is given as

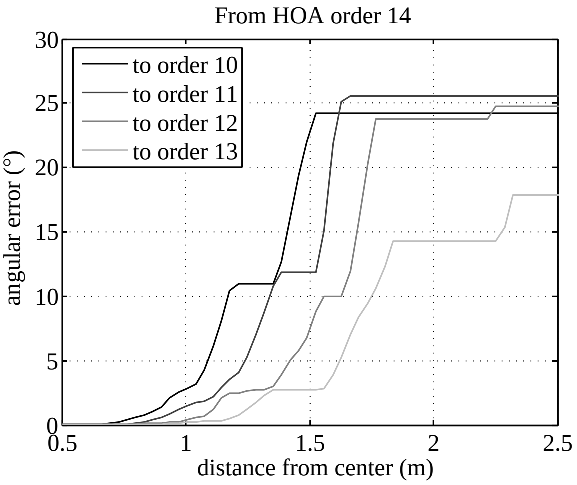


Figure 5.5: Simulation of the encoding angular error of the gradient of the pressure phase for a single 500-Hz plane wave arriving at 0° azimuth. within a given radius, for encoding orders ranging from 10 to 13, with an original representation order of 14. The angular error globally increases as the encoding order decreases relative to the original order.

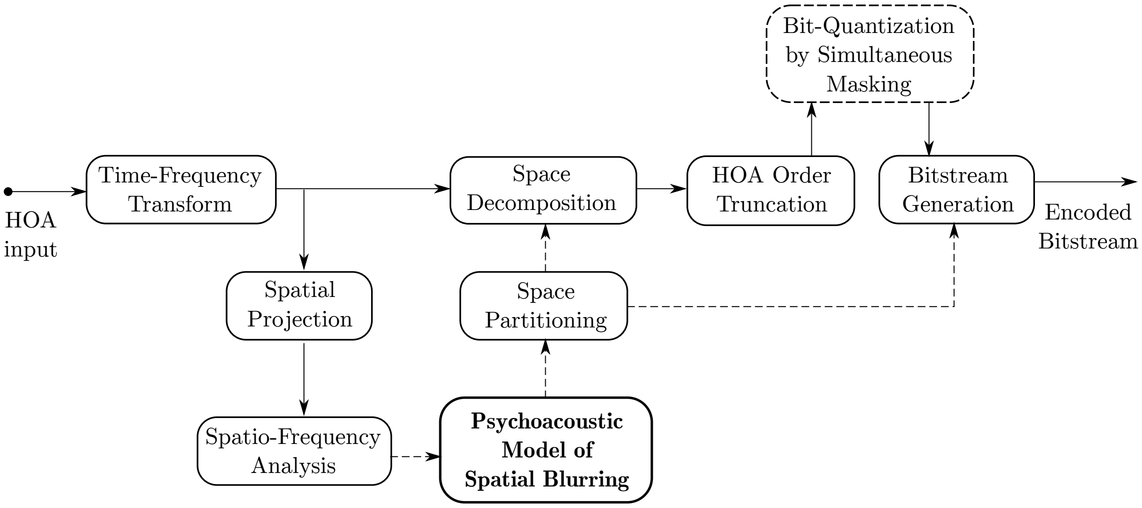


Figure 5.6: The proposed multichannel audio coding scheme, based on a psychoacoustic model used to properly truncate the HOA order of representation of the input signal.

an input. This signal is fed into a time-frequency transform module (see section 5.2.4), which divides each channel into temporal frames and transforms them into the frequency domain. A spatial projection (see section 5.2.5) is then made of the resulting signals in order to obtain the spatial evolution of the energy within a frequency bin on a circle around the listening point. From the projected signals, a spatio-frequency analysis is performed (see section 5.2.6). This module aims at identifying, based on a temporal correlation analysis, the sound sources present in each critical band, and more specifically their center of energy, their width, and their energy level. This information is then fed into the psychoacoustic model (see chapter 4 and section 5.2.7) which computes, for the current temporal frame of the sound scene, the localization blur associated with each source, in the presence of the other sources considered as distracters. This localization blur corresponds to the minimum spatial resolution needed to ensure an accurate spatial representation of the scene. In each critical band, the HOA representation is accordingly divided into smaller "chunks" of space requiring the same spatial resolution (see sections 5.2.8 and 5.2.9). The final module truncates the order of the HOA representation of each chunk of space, according to localization blur (see section 5.2.10). The spatially degraded chunks of space are then encoded into a bitstream (see section 5.2.12), and a compression gain is obtained as a result of the truncation process.

5.2.3 Modes Of Operation

As explained in section 5.2.1, prior to transmission, it is necessary to provide information to the coder concerning the desired radius r_{per} of accurate reproduction (i.e., the radius of the perceptual sweet spot) for each frequency. This can be done in two ways: transparent mode and constrained mode.

Transparent mode

In this mode, for a given frequency, the perceptual sweet spot radius r_{per} will be equal to the sweet spot radius of the initial order for this frequency, $r_{\text{ini}} = \frac{M}{k}$. If the original order M is 14, the sweet spot radius roughly equals 1.5 m. As depicted in **figure 5.5**, if the current MAA is 12°, then it is possible to truncate the HOA representation to order 11 to keep a perceptual sweet spot of 1.5 m. In that case, one could discard three orders out of 14, which represents a raw [9] coding gain of about 20%.

Constrained mode

In this mode, the perceptual sweet spot radius r_{per} will be constrained for all frequencies, as much possible, to equal a fixed value r_0, and the accurate order of representation for each frequency will be derived from it. This approach can be used for example to reduce the accurate listening area (and thus the order of representation) associated with the low frequencies, which is typically too wide compared to that of the high frequencies.

5.2.4 Time-Frequency Transform

This transform, and thus the analysis into critical bands, is justified by the second assumption of our psychoacoustic model (see chapter 4).

[9] The additional cost due to the partitioning of space is not taken into account (see section 5.2.8).

Because the input signals can have different forms equivalent to equation (5.9), let us assume that the HOA components are in the temporal domain, and thus let us name them b_{mn}^{σ}. M is the encoding order of the HOA representation.

In this module, these signals are transformed from the temporal domain to the frequency domain, using a Fourier-related transform, such as the *modified discrete cosine transform* (MDCT) [PJB87]. In this case, we can divide each channel $b_{mn}^{\sigma}[p]$ into frames of P samples, and using analysis windows of $2P$ samples with 50% overlap, we obtain the channels in the frequency domain:

$$B_{mn}^{\sigma}[k] = \sum_{p=0}^{2P-1} b_{mn}^{\sigma}[p] w[p] \cos\left[\frac{\pi}{P}\left(p + \frac{1}{2} + \frac{P}{2}\right)\left(k + \frac{1}{2}\right)\right], \tag{5.16}$$

with $w[p]$ being the temporal form of the analysis window, such as the Kaiser-Bessel derived (KBD) window:

$$w[p] = \sin\left(\frac{\pi}{2}\sin^2\left[\frac{\pi}{2P}\left(p + \frac{1}{2}\right)\right]\right). \tag{5.17}$$

The main advantage of using this type of transform is the *time-domain aliasing cancellation* (TDAC), ensuring a perfect reconstruction of the temporal signals.

The value of P must be large enough to ensure an accurate representation of low frequencies, but it must not exceed the time integration window of the auditory system spatial analysis, which is about 30-60 ms according to [BF07] (see section 2.5.1).

5.2.5 Spatial Projection

This module is necessary to obtain a representation of the energy within a frequency bin on a circle around the listening point. This is equivalent to decoding the HOA signals into the D-format (or loudspeaker format, see section 2.1.2). However, in our case, the D-format signals will not be used to feed loudspeakers, but rather to perform a spatial analysis of the scene at a certain number of nodes on the circle.

We can choose a regular distribution of the nodes on the circle. Therefore the decoding matrix simply has the following form:

$$\mathbf{D} = \frac{\sqrt{2}}{M}\begin{bmatrix} \frac{1}{\sqrt{2}} & \cdots & \frac{1}{\sqrt{2}} \\ \cos(\phi_1) & \cdots & \cos(\phi_K) \\ \sin(\phi_1) & \cdots & \sin(\phi_K) \\ \cos(2\phi_1) & \cdots & \cos(2\phi_K) \\ \sin(2\phi_1) & \cdots & \sin(2\phi_K) \\ \vdots & \cdots & \vdots \\ \cos(K\phi_1) & \cdots & \cos(K\phi_K) \\ \sin(K\phi_1) & \cdots & \sin(K\phi_K) \end{bmatrix}^{\mathrm{T}}, \tag{5.18}$$

where $K = 2M + 1$ is the number of nodes, and ϕ_i is the angle of the i^{th} node. The decoding process is performed with a matrix product:

$$\mathbf{S} = \mathbf{D} \times \mathbf{B}, \tag{5.19}$$

where $\mathbf{S}$ are the decoded signals, and $\mathbf{B}$ the HOA signals of the current frame resulting from the time-frequency transform.

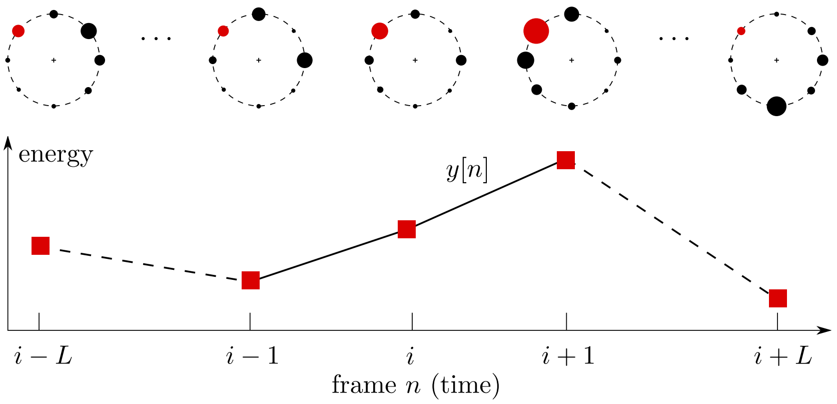


Figure 5.7: Spatio-frequency analysis for a given node on the circle around the listener.

5.2.6 Spatio-Frequency Analysis

The aim of this module is to analyze the spatial distribution of the energy within each critical band of the current frame in order to roughly identify sources. A first optional step can be to separate primary from ambient components, as this might subsequently refine results from our psychoacoustic model. This can be done, for example, by applying a primary-ambient decomposition such as the one proposed by Goodwin in [Goo08]. The following applies equally to primary and ambient components.

We need to group frequency bins into subbands of critical bandwidth in order to follow our psychoacoustic model. To do so, we can define critical bands using a set of adjacent *Equivalent Rectangular Bandwidths* (ERB) by following the recommendations from Glasberg and Moore [GM90]. Then the energy within one subband is given by summing the energy of each of its constituent frequency bins.

Let us call $X_i[b]$ the total energy of a given node, for the current temporal frame i in subband b. As illustrated in **figure 5.7**, it is possible to define a downsampled temporal signal

$$y[n] = \{X_{i-L}[b], X_{i-L+1}[b], ..., X_i[b], X_{i+1}[b], ..., X_{i+L}[b]\}, \tag{5.20}$$

where L is the desired temporal extent of the signal in samples.

Because our psychoacoustic experiments have shown that the MAA depends on the level of the sources present, we need to know the total level of each identified source. If two nodes are correlated in time, it means that they are parts of the same source, and that their energies sum in time. This analysis can be done by computing a correlation coefficient, such as the Pearson sample correlation coefficient, between the two nodes. Between two sets of samples X and Y, written as x_i and y_i, where $i = 1, 2, ..., n$, this coefficient is expressed as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \tag{5.21}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, s_x and s_y are their sample standard deviations.

We can start by finding the node that has the most associated energy, because this node is likely to be close to a sound source in the field. To know the spatial extent of the identified source, it is possible to compute a correlation coefficient between the downsampled temporal signal of the current node $y_s[n]$ and that of the two nodes in its vicinity $y_1[n]$ and $y_2[n]$. When the correlation coefficients are greater than or equal to a given threshold of correlation τ_c, then the initial source node can be spatially extended to the corresponding nodes. Then the same process can be applied between the initial node and the nodes in the vicinity of the extended source, and so on until all the correlation coefficients are below τ_c. By repeating this process until all the nodes are associated with a source, or until no node with energy remains, it is possible to identify several uncorrelated sources within each subband and for the current temporal frame. For each subband, the total energy associated with a given identified source is the sum of energies from all nodes associated with this source.

Note that in our model, reflections of a source are separated from it, which is why a correlation coefficient is used instead of a cross-correlation coefficient.

5.2.7 Psychoacoustic Model

Our psychoacoustic model is called to estimate the MAA associated with each source, which can be interpreted as the maximum spatial distortion that can be allowed in the direction of each source to accurately represent the encoded sound scene. As justified in section 4.1, it is not necessary to submit the position of each source to the model. Indeed, our psychoacoustic results showed that the position of a given source does not have an effect on the spatial blurring that this source creates, and in our model the target source is assumed to be in front of the listener.

More formally, each identified source s is characterized by three parameters $\{f_c, I, t\}$, where f_c is the center frequency of the subband to which the source belongs, I is the associated energy level of the source, and t is its type (primary or ambient). If no primary/ambient decomposition has been performed beforehand, all identified sources are considered as primary sources. If K is the number of identified sources, the psychoacoustic model is called via $\psi\left(\{s_k, k \in 1..K\}\right)$. It returns an estimation of the MAA associated with each source s_k considered as a target, based on the spatial blurring created on it by all other $K-1$ sources, considered as simultaneous distracters.

5.2.8 Space Partitioning

This step aims to partition the representation space of the scene, depending on the resolution needed by each node. The general idea is to cluster adjacent nodes needing approximately the same spatial resolution.

The partitioning of the space must be optimized as a function of the number of components needed to represent the scene, i.e., the order of representation of each chunk of space. The order of representation depends on the targeted spatial resolution within that chunk. The targeted spatial resolution is that of the node needing the finest resolution in the chunk under consideration, depending on the MAA of the source this node is associated with. The order of representation M necessary to ensure this resolution can be obtained from the simulation table proposed in section 5.2.1, considering only the highest frequency boundary of the current subband. The cost of each chunk is its order of representation M_i, and the cost of the whole divided space can be expressed as:

$$\sum_{i=0}^{n} M_i \tag{5.22}$$

where n is the number of chunks. An important point is that the smaller the chunk of space, the higher the order of the spherical harmonics necessary to represent that chunk. This constrained selection of the minimum order of representation of the scene can be expressed, for a given chunk i, as:

$$M_i = \max\left(\max\{M_{ij}, j = 1..k\}, M_{i0}\right), \tag{5.23}$$

where M_{ij} is the minimum order of representation for node j in chunk i, k is the number of nodes in chunk i, and M_{i0} is the minimum order of representation for chunk i given its size. Also, one can assume that the total number of HOA components needed to represent all chunks is potentially greater than the initial number of components (see section 5.2.9). Thus, dividing the space into smaller chunks has a cost that can be expressed as:

$$\lambda(n-1), \tag{5.24}$$

where λ is a scalar value representing the cost of one division of the space. From expressions (5.22) and (5.24), we can derive the overall cost of the representation of the scene for a given partition:

$$c = \lambda(n-1) + \sum_{i=0}^{n} M_i. \tag{5.25}$$

The optimal partitioning of the space is the one leading to the minimum cost value. In order to decrease the complexity of the optimization, it is possible to constrain the number of chunks in the partition. It is also possible to completely avoid this dynamic partitioning step by using a static partitioning of the space, but this solution would provide a weaker coding gain than would dynamic partitioning.

5.2.9 Space Decomposition

One method that can be used is that proposed by Pomberger and Zotter [PZ09]. This method represents an auditory scene on a smaller portion of the space by finding a new set of eigenfunctions dedicated to this space. These functions are derived by an eigendecomposition of the initial eigenfunctions. This method suits our needs because the inverse transform (the extension back to the whole space) does not introduce phantom sources or other artifacts. However, of course, the information initially outside of the targeted portion of space is lost during the transform.

5.2.10 HOA Order Truncation

This module truncates the HOA order of each chunk within each subband, given the minimum order of representation M_i of the considered chunk i obtained from the previous step. All the frequency bins within a subband are truncated at the same order. The truncation simply consists of assigning zeros to the components associated with the orders greater than M_i.

For practical use, especially if considering the Pomberger and Zotter space decomposition method, it might be easier to rather truncate the HOA order of the scene prior to the space decomposition step. As a result, the discarded eigenfunctions will not be used in the eigendecomposition, which will then need less associated components.

5.2.11 Bit-Quantization by Simultaneous Masking

This optional module provides the possibility of applying a bit-quantization process to the signal associated with each of the HOA components using a simultaneous (or frequency) masking model (see appendix C). However, it is very important to take into consideration the phenomenon of binaural release from masking (see sections 1.4.6 and 2.4.1). Indeed, the masking effect of a signal component on another component depends on their relative spatial position. Particularly, the masking effect decreases when the two components are no longer collocated.

However, this process has to be performed carefully, because modifying the bit-quantization of a component belonging to a chunk of space affects all the sources within that space. Moreover, the sources present in a given chunk have to be considered as located potentially anywhere in this chunk. Therefore, when such lack of information arises, the most unfavorable case (in terms of masking) should be considered in order to avoid audible quantization noise.

Another issue is that each sound source of the scene is represented across several HOA components. Therefore, the bit-quantization has to be distributed across the HOA components. Some work has already been done along these lines [HSS08].

5.2.12 Bitstream Generation

The bitstream contains the coefficients associated with the eigenfunctions of each chunk of space for each temporal frame, as well as the partitioning of the space. The main characteristic of the generated bitstream is that the components that were given values of zero at the truncation step are highly compressed by entropy coding (see section C.6). This is where the coding gain is achieved.

5.2.13 Decoding

The decoding process consists of reproducing each chunk in the full spatial domain and summing the resulting signals together. To do so, it is necessary beforehand to perform the eigendecompositions of the eigenfunctions to obtain the new set of eigenfunctions again, exactly as in the module described in section 5.2.9. This way, it is not necessary to transmit the new eigenfunctions of each partition in the bitstream, but only the partitioning of the space itself.

5.3 Conclusions

The two proposed coding schemes provide a way to encode an acoustic field that takes advantage of the limits of the auditory system concerning its spatial resolution. It is promising, because in complex scenes most of the spatial information is not perceived accurately. The dynamic bit allocation procedure for parametric coding schemes is probably more suited for low bitrates. From our listening tests, it appeared that a basic stereo image can be obtained even by allocating less than a single bit per parameter on average.

On the other hand, dynamic truncation of the HOA order is more appropriate for high bitrates. First, contrary to parametric coding schemes, several sources at different positions can be present in the same critical band. Second, it benefits from all the advantages of HOA: the representation of the acoustic field is independent of the recording and reproduction systems, and the representation remains hierarchical, allowing an adaptation of the encoded signal determined upon the available bandwidth.

Conclusions

In this thesis, we first studied localization blur, or auditory spatial resolution, in situations of complex listening in which multiple sources are present simultaneously. We have demonstrated through psychoacoustic experiments that the localization blur associated with a target sound source is greater when it is assessed in the presence of distracting sources compared to when it is assessed in quiet. The concept of "spatial blurring" that a distracter creates on a target was proposed to denote this phenomenon. This main point was accompanied by several other results concerning localization blur and spatial blurring. Previous findings from other investigators concerning the dependence of localization blur on the frequency content of the target source were confirmed. We showed that spatial blurring depends on the frequency separation between the target and the distracter, and that this effect of the frequency separation on spatial blurring is dependent on the target center frequency. The spatial blurring that a distracter creates on a target source also depends on its energy level relative to that of the target. We also demonstrated that the spatial blurring created on a target increases with the complexity of the sound scene, that is with the number of distracters present, at least up to four independent narrow-band distracters. This increase follows an over-additivity rule, meaning that the spatial blurring created by a set of distracters is greater that the sum of the spatial blurring that each of these distracters would create when presented alone. The distracter positions do not seem to affect spatial blurring, except specifically when they are all collocated with the target, where it is significantly reduced; otherwise, distracters scattered around the listener have a tendency to increase spatial blurring compared to when all distracters are collocated. Another interesting result is that the spatial blurring that a distracter creates is not correlated with the energetic masking ability of this distracter, which suggests that spatial blurring and energetic masking rely on different auditory processes.

The main aim of this thesis was to exploit this degradation of the auditory spatial resolution due to spatial blurring in a multichannel audio coding scheme. This required modeling the psychoacoustic results. In the model we proposed, we made a point of modeling as separately as possible the variables on which spatial blurring depends. The interest of this approach is that it avoids the necessity of psychoacoustic data testing all these variables simultaneously, which would otherwise present practical issues. As a result, this model comprises three parts. It first models the spatial blurring created by a single distracter on a target depending on their frequency separation and on the target center frequency, assuming that their sound levels are matched in loudness. The second part models the effect of their relative energy level on spatial blurring, which is used to correct the previous estimation. Finally, the spatial blurring created by all distracters on a given target source is combined according to the over-additivity rule we derived previously.

This thesis proposed also two multichannel audio coding schemes based on spatial blurring. They both exploit the original idea that it is possible to achieve data compression by reducing the spatial accuracy of representation of the sound scene while shaping the resulting spatial distortions according to localization blur, thereby minimizing their perceptual impact in an homogeneous fashion among all sources, or even making these distortions imperceptible. Therefore, these two audio coding schemes use our psychoacoustic model of spatial blurring to dynamically drive the spatial accuracy of representation of the sound scene. One of them is based on parametric spatial audio coding schemes. We implemented it within a parametric stereo scheme, and this approach seems to give promising results since informal listening tests revealed that it is possible to obtain a basic stereo image by coding each spatial parameter on less that one bit on average. The second proposed coding scheme, which is based on the HOA representation, relies on a dynamic truncation of the HOA order of representation. It has not yet been tested, but future work will be devoted to further implementation and formal testing of the coding schemes we proposed, as well as to finding other applications of our psychoacoustic model. For instance, an interesting application could be for situations in which spatial sound scenes have to be synthesized in real-time (as in video games for example, see [MBT$^+$07]). Indeed, the sources composing the scene could be grouped (to the extent of their associated localization blur) into clusters sharing a single rendering position, therefore reducing the complexity of the synthesis.

Our work constitutes a first step in the study of spatial blurring. Our psychoacoustic model of spatial blurring will require additional experimentation before it can be fully validated. The type of sources—targets and distracters—we studied is limited since they were all continuous noises. We did not consider the temporal characteristics of the stimuli, or test the impact of the temporal organization of sources (for instance, Perrott *et al.* [PMMS89] showed that localization blur increases in situations where the precedence effect operates). Given the simplicity of these stimuli, driven by the need to solve primary problems first, the experimental conditions we tested can clearly be extended to more complex situations using the methodologies we have developed in this thesis. In particular, the stimuli will need to be complexified because the streaming and fusion processes performed by the auditory system are limited in the cases tested, and since they play an important role in spatial perception, they will need to be taken into account in future work. For this reason our model in its current only considers low-level processes, and further studies will assess how spatial blurring behaves when higher-level processes are involved.

Appendix A

Instructions Given to the Subjects

A.1 Left-Right/Right-Left Task

The experiment is organized in experimental blocks of about 5 minutes. On each trial, you will be presented with a sequence consisting of the exact same **short sound played twice**, but **from different locations in space**. Your task is to **listen to the sequence and to indicate the direction of the change in position** between the two sounds.

Note that the sounds will only be presented in the horizontal plane in front of you. For some blocks, an additional sound will be played in the background. You have to **ignore that sound** and **focus on the two sounds of the sequence**.

Each sequence will only be played once, and you will have to indicate your answer on the keyboard using the **'←' and '→' keys** which mean "from right to left" and "from left to right," respectively. You do not have to wait until the end of the background sound (when present) to answer. Once you have answered, the next trial will be played automatically, and so on until the end of the block.

During the experiment, your head must be right above the white cross marked on the floor. Sit up straight, resting on the back of the chair, and look right in front of you. **Try to keep the same position throughout the whole experiment.**

At the end of the block, feel free to take a break and **wait for the experimenter to set up the next block.**

A.2 Audibility Task

[With distracters]

The experiment is organized in experimental blocks of about 5 minutes. On each trial, you will be presented with a sequence consisting of the exact same **background sound played twice**. **One** of these two sounds **contains an additional target sound**. Your task is to **listen to the sequence and to indicate which one of the two background sounds contained the target sound**.

Each sequence will be played only once, and you will have to indicate your answer on the keyboard using the **'1' and '2' keys** which mean "the first interval contained the target sound" and "the second interval contained the target sound," respectively. You **must** wait until the end of the whole sequence to answer. Once you have answered, the next trial will be played automatically, and so on until the end of the block.

During the experiment, your head must be right above the white cross marked on the floor. Sit up straight, resting on the back of the chair, and look right in front of you. **Try**

to keep the same position throughout the whole experiment.

At the end of the block, feel free to take a break and **wait for the experimenter to set up the next block.**

[Without distracters]

The experiment is organized in experimental blocks of about 5 minutes. On each trial, **a square will flash two times on the screen**. **A target sound will be played during one** of these two flashes. Your task is to **indicate during which of the two flashes the target sound was played**.

Each sequence will be played only once, and you will have to indicate your answer on the keyboard using the **'1' and '2' keys** which mean “the first interval contained the target sound” and “the second interval contained the target sound,” respectively. You **must** wait until the end of the whole sequence to answer. Once you have answered, the next trial will be played automatically, and so on until the end of the block.

During the experiment, your head must be right above the white cross marked on the floor. Sit up straight, resting on the back of the chair, and look right in front of you. **Try to keep the same position throughout the whole experiment.**

At the end of the block, feel free to take a break and **wait for the experimenter to set up the next block.**

Appendix B

Study of Inter- and Intra-subject variability

This appendix aims to briefly illustrate differences in performance observed in our psychoacoustic experiments presented in chapter 3.

Differences in performance between subjects for the same experimental condition (i.e., *inter*-subject variability) can be observed in the results from experiment 1. Each row in **figure B.1** depicts two psychometric curves obtained from two subjects for the same experimental condition. The bottom row of the figure illustrates that, for the same experimental condition, the spread of these psychometric functions (their non-asymptotic part) can be limited (subject 9) as well as extended (subject 11). Hence, even if these two subjects share almost the same MAA (around 5°), if the angular separations used for subject 11 had been presented to subject 9, the estimated curve would have been imprecise (a single point would have been within the non-asymptotic part). To the contrary, as can be seen in the middle row of the figure, finding a common set of angular separations when the psychometric functions of two subjects have approximately the same spread might also cause troubles if their respective MAAs differ too much (more than 10° in this case). Indeed, switching the two sets of angular separations would have necessitated in both cases an extrapolation of the curve to reach the 80.35% performance point, which is not desirable. Further, increasing the number of angular separations to present would increase the block duration. The top row of the figure illustrates that the spread of the psychometric functions and the MAA can also vary at once from one subject to another for the same experimental condition, again making difficult to find a common set of angular separations to present which suits both cases. Note finally that the two left plots in the top and middle rows of the figure are from the same subject, but for different experimental conditions, and show curves varying both in spread and MAA. All these remarks justify the protocol we proposed for experiment 1 (see section 3.6.2) to adjust the set of angular separations to present to each subject and for each experimental condition, until finding the proper area of testing. Examples of inter-subject variability can also be found in the LR/RL task of experiments 2, 3 and 4, and justify the use of an adaptive method, which, when properly parameterized, is less sensitive to variations in psychometric functions (see section 3.7.4). For instance, **figure B.2** shows these inter-subject variations for experiment 4, an extreme case being the experimental condition 8 (in which the four distracters are presented in front of the subject, in between the two targets). Note that inter-subject variability is very reduced for the audibility task.

Another type of variability that can be seen in **figure B.2** is the *intra*-subject variability, that is, how greatly two estimates for the same subject and experimental condition can differ from each other. Focusing on the LR/RL task in experiment 4, for a given experimental condition, the difference between the estimates of each of the two sessions (see section 3.9.2) is represented on the figure with error bars. The estimate of each session is represented with different symbols: - for the first session and $*$ for the second one, the average being represented by an open circle. An interesting point is that the extent of this variability differs from one subject to another: subjects 5 and 10, for instance, are very unstable, whereas subjects 12 and 15 are quite consistent. Intra-subject variability could be due to training effects, as it might be the case for subject 8 (most of the second session estimates correspond to lower SNR values). However, this pattern is not found for other subjects. Another cause could be the order of presentation of the experimental conditions (which is flipped between the first and the second sessions), or the fact that the two sessions are done on different days. Anyway, the design in two sessions we used for this fourth experiment permitted to average out this variability. Pilot results did not show such a variability for the audibility task, thus a single estimation was performed.

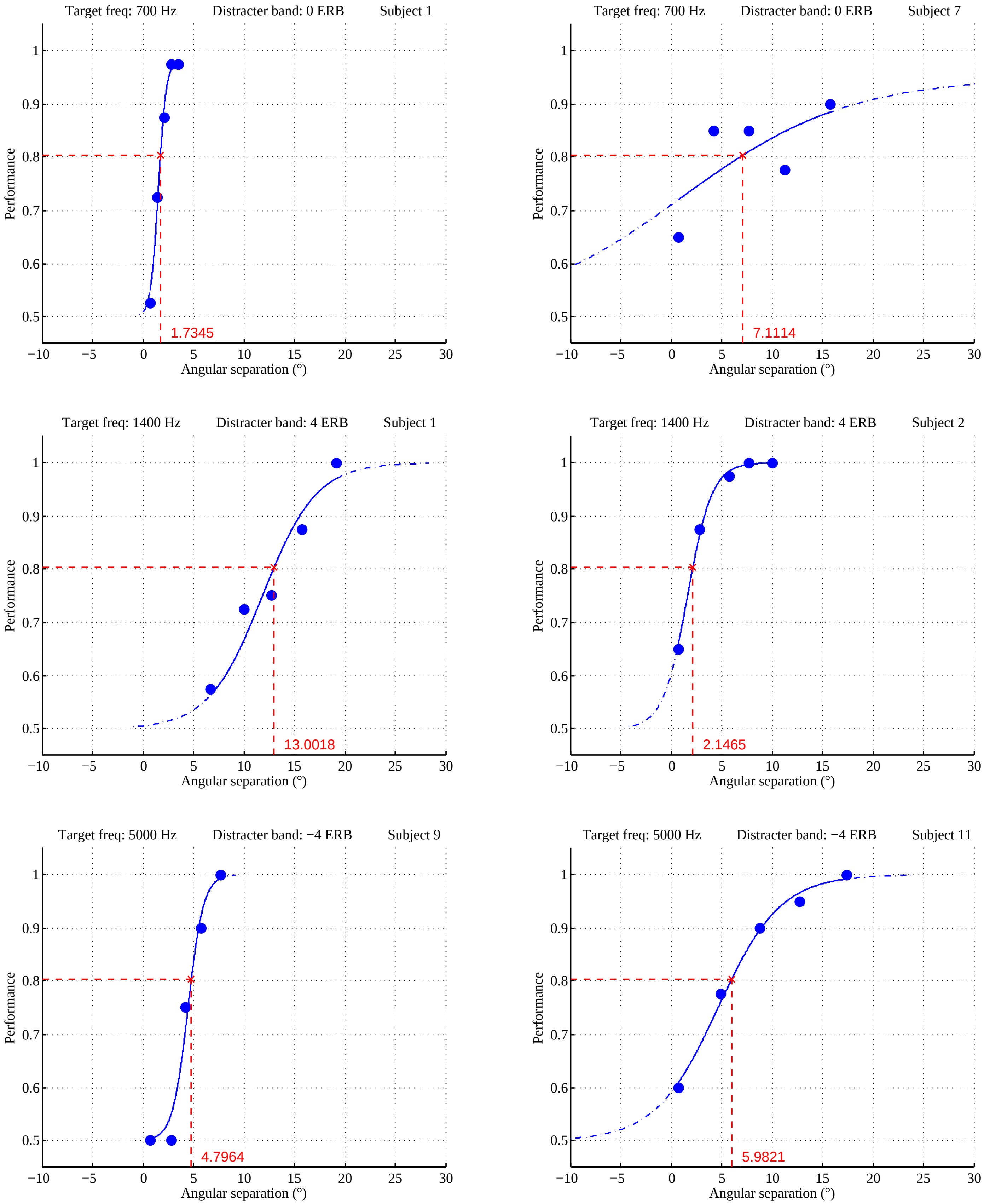


Figure B.1: Representative psychometric curves from experiment 1 for several experimental conditions and subjects. Two plots on the same row are from two different subjects in the same experimental condition.

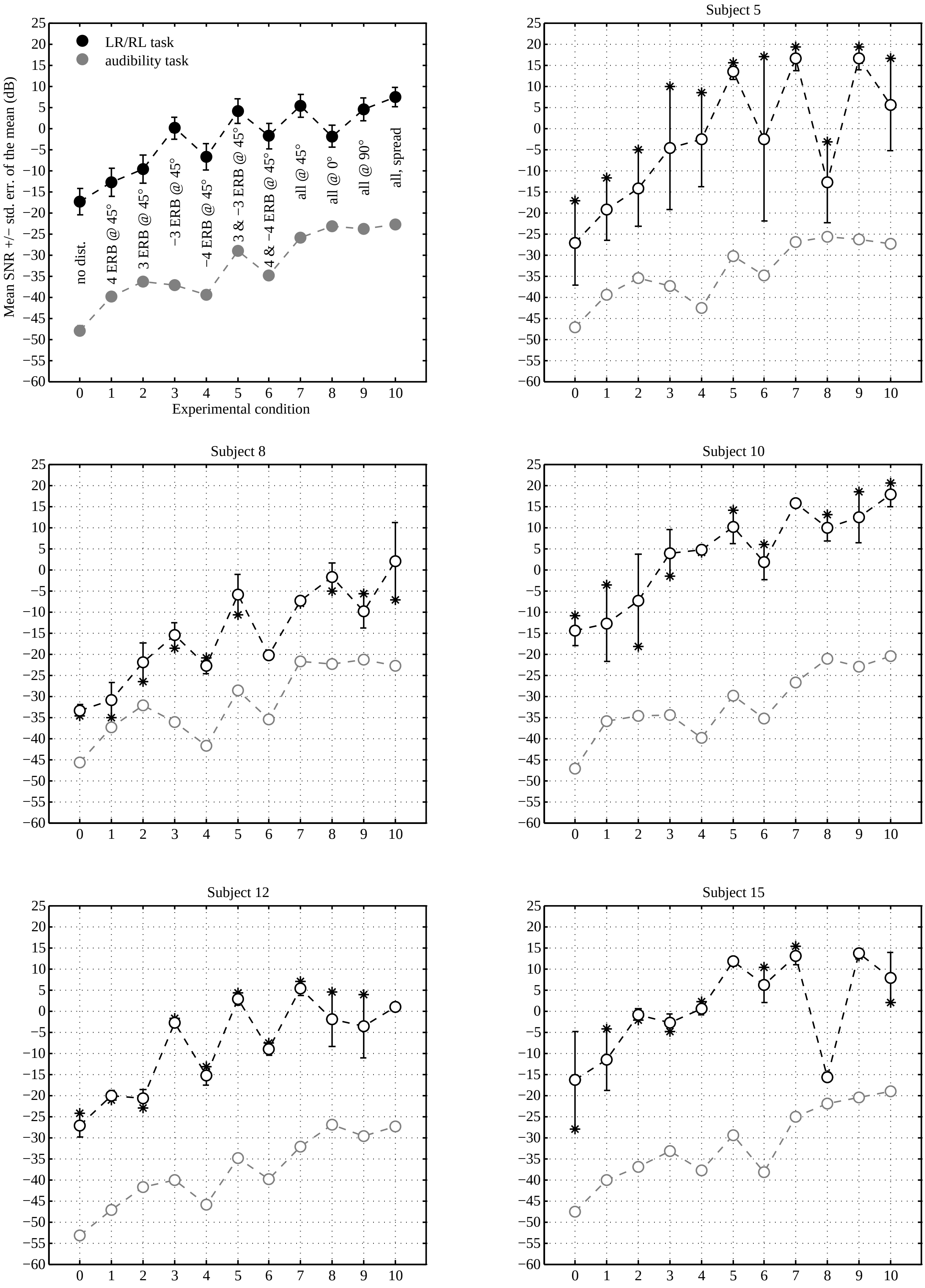


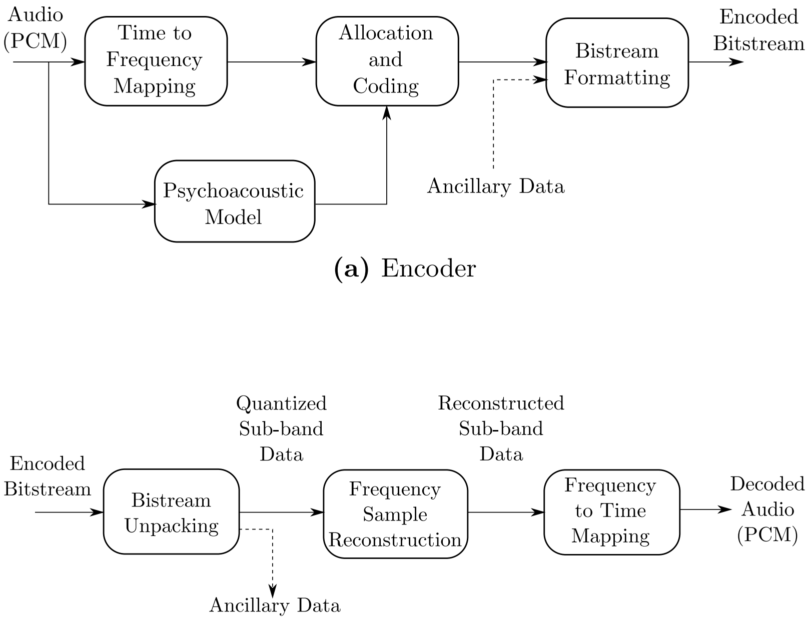
Figure B.2: Mean results along with some individual results from experiment 4 (target center frequency of 1400 Hz). Error bars represent the difference between estimates of the first (-) and the second (∗) session.

Appendix C

Audio Coding Based on Energetic Masking

This appendix deals with conventional audio coders known as “perceptual coders,” that is to say coders that make use of the masking phenomena presented in section 1.2.2 to increase the coding gain. We will focus on quantization error, modeling of psychoacoustic masking curves, and bit allocation strategies. Further information can be found in Bosi’s book [BG02], on which this description is mainly based.

C.1 Principle Overview



(a) Encoder

(b) Decoder

Figure C.1: Schematic of a perceptual audio coder based on energetic masking. (Reprinted from [BG02].)

An overview of a perceptual audio coder is given in **figure C.1**. The encoder is fed with an audio PCM input signal. Then, the signal is represented in the frequency domain rather than time domain via a filter bank or a transform like PQMF or MDCT [PJB87] to take advantage of harmonic redundancy removal. A psychoacoustic model is also fed with the input signal and computes simultaneous and temporal masking curves according to experimental data (see sections 1.2.2 and C.3). Given the data rate constraint and the masking curves, frequency samples coming from time-to-frequency mapping are quantized in such a way that quantization noise is kept below (or as close as possible to) their masking thresholds (see section C.5). Finally, the encoded bitstream is made of the coded audio data with quantified samples and bit allocation information, control parameters such as block length and type of windowing, and also ancillary data like time-synchronization stamps and error correction codes.

At the decoder side, the encoded bitstream is unpacked into audio data, control parameters and ancillary data. The original frequency samples are reconstructed using the bit allocation information. The noise resulting from this dequantization should be below (or as close as possible to) masking thresholds computed by the psychoacoustic model at the encoder side, and thus should be (almost) inaudible. After that, the frequency domain signal is mapped back to the time domain using the appropriate synthesis bank or inverse transform such as PQMF or IMDCT, and converted into an audio PCM output data stream. For some audio coding schemes, such as AC-3, the bit allocation routine is computed both at the encoder and decoder sides, which decreases the transmitted allocation side information, but increases the complexity of the decoder.

C.2 Quantization Errors

Quantization covers two steps in an audio coding scheme. The first one, which actually occurs prior to the encoding process (and consequently is not represented in **figure C.1**), concerns the mapping of an analog signal onto codes that can be represented with a finite number of bits. The resulting quantized signal constitutes the input signal of the encoder. The second step, which interests us, is represented by the box labeled “Allocation and Coding” in the figure, and also corresponds to a representation of the signal with codes, but rather from an already quantized version of it (in the frequency domain) than from its analog form. As we will explain in section C.5, the idea is to use fewer bits to encode each temporal frame by assigning a different number of bits from one frequency sample to another, but (as much as possible) without introducing quantization noise.

We will focus on errors resulting from a quantization process. One generally measures quantization error in terms of *signal-to-noise ratio* (SNR) measured in decibels (dB):

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{\langle x_{\mathrm{in}}^2 \rangle}{\langle q^2 \rangle} \right), \tag{C.1}$$

where q is defined as the power in the difference between the input signal x_{in} and the output signal x_{out}:

$$q(t) = x_{\mathrm{out}} - x_{\mathrm{in}}. \tag{C.2}$$

Quantization can lead to two kinds of error: round-off error and overload or clipping error. The latter occurs when signal amplitudes are too high in level for the quantizer. As a result, these amplitudes are clipped to the highest and lowest quantizer steps.

Round-off error is more interesting, in the sense that it is the kind of error resulting from the second quantization process mentioned above. It happens when mapping ranges of input signal amplitudes onto a single code (which thus corresponds to a single output level). In the case of a uniform quantizer, if we call δ the size of the input range per code, assuming that the amplitude falls randomly into each quantization bin, the probability density of the error signal $q(t)$ at any time is approximately equal to $\frac{1}{\delta}$ in the range between $-\frac{\delta}{2}$ and $\frac{\delta}{2}$ and zero elsewhere (i.e., the round-off error is equally likely to be any value between $-\frac{\delta}{2}$ and $\frac{\delta}{2}$). Given the error probability distribution, the expected error power for the quantizer can be expressed as:

$$\left\langle q^2 \right\rangle = \int_{-\infty}^{\infty} q^2 p(q) dq = \int_{-\frac{\delta}{2}}^{\frac{\delta}{2}} q^2 \frac{1}{\delta} dq = \frac{\delta^2}{12}. \tag{C.3}$$

$\delta \approx 2 \times \frac{x_{\max}}{2^R}$ in the case of a uniform quantizer with R bits, where $x_{\max}$ represents the highest quantizer step. So:

$$\left\langle q^2 \right\rangle = \frac{x_{\max}^2}{3 \times 2^{2R}}. \tag{C.4}$$

By substituting this result in equation (C.1), we get the approximately expected SNR in dB from the quantizer when fed with an input power equal to $\langle x_{\text{in}}^2 \rangle$:

$$\text{SNR} \approx 10 \log_{10} \left(\frac{\langle x_{\text{in}}^2 \rangle}{x_{\max}^2} \right) + 6.021 \times R + 4.771. \tag{C.5}$$

C.3 Modeling Simultaneous Masking Curves

The following sections explain how experimental data briefly introduced in section 1.2.2 are used to compute an overall simultaneous (or frequency) masking curve for each time frame.

Models of the Spreading Function

As demonstrated by Bosi, the SPL hearing threshold of a test sound B depending on the SPL of a masker A can be written as:

$$\text{SPL}_B = \text{SPL}_A + \underbrace{10 \log_{10} \left(\frac{\ln(10)}{10} \Delta\text{L}_{\min} \right)}_{\substack{\text{down-shifting by a} \\ \text{constant } \Delta}} + \underbrace{10 \log_{10} \left(F(z(f)) \right)}_{\text{spreading function}}, \tag{C.6}$$

where $\Delta\text{L}_{\min}$ is the smallest perceptible change in dB in the basilar membrane excitation pattern, $F(z)$ is a function describing the shape of the original signal excitation pattern, and $z(f)$ represents the location along the basilar membrane of the peak excitation resulting from a signal at frequency f. Hence, the masking curve relative to a masker can be derived at each frequency location from the SPL of the masker A by down-shifting it by a constant Δ that depends on $\Delta\text{L}_{\min}$ evaluated for masker A, and adding a frequency-dependent function that describes the spread of the masker excitation energy along the basilar membrane.

Because the codec quantification noise is spectrally complex rather than tonal, the masking-curve models should reproduce experimental narrow-band noise masked by narrow-band noise or tone masking curves. However, Bosi reports that there is very little data in the literature that address this issue.

Many spreading function models are given in the literature, the most simple being the triangular function, written in terms of the Bark scale difference between the maskee and masker frequencies $dz = z(f_{\text{maskee}}) - z(f_{\text{masker}})$:

$$10 \log_{10} \left(F(dz, L_M) \right) = \left(-27 + 0.37 \max\{L_M - 40, 0\} \theta(dz) \right) |dz|, \tag{C.7}$$

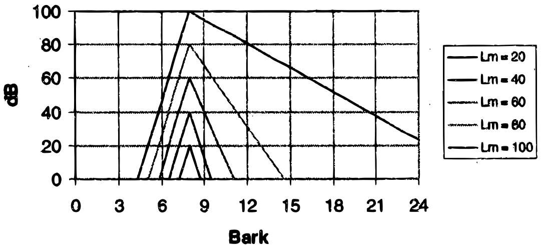


Figure C.2: Spreading function described by the two slopes derived from narrow-band noise masking data for different levels of the masker. (After [BG02].)

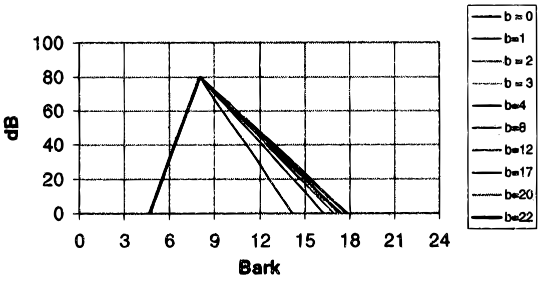


Figure C.3: Terhardt spreading function for different masker center frequencies superimposed at 8 Bark. (After [BG02].)

where L_M is the masker SPL and $\theta(dz)$ is the step function equal to zero for negative values of dz and equal to one for positive values of dz. This spreading function is plotted in **figure C.2**.

Despite the remarks stated in section 1.2.2 concerning the independence of the induced masking curve on the frequency of the masker when described on the Bark scale, there may be some frequency dependence according to some experimental data. Terhardt [Ter79] proposed a modified version of the triangular approximation that takes this frequency dependence into account (see the result in **figure C.3**):

$$10\log_{10}\left(F(dz, L_M)\right) = \left(-24 + \left(0.2L_M + \frac{230\text{ Hz}}{f}\right)\theta(dz)\right)dz. \tag{C.8}$$

As described in section 1.2.2, an important parameter is the minimum SMR, corresponding in equation (C.6) to a down-shift Δ from the masker SPL by an amount that hinges upon $\Delta\mathrm{L}_{\mathrm{min}}$, which varies with the type of masker. Zwicker suggested a value of 1dB for $\Delta\mathrm{L}_{\mathrm{min}}$, implying that the peak of the masking curve should be about 6 dB below the SPL level of the masker, which corresponds to experimental data for narrow-band noise masking tones. Concerning tone maskers, Moore proposed $\Delta\mathrm{L}_{\mathrm{min}} = 0.1$ dB, which leads to $\Delta = 16$ dB. In general, when the masker is tonal, the minimum SMR levels are higher than when the masker is noise-like. Futhermore, experimental data show that our ability to detect changes in excitation level is reduced at low frequencies, implying that Δ increases with increasing frequency. This variation also depends on whether the masker is noise-like or tone-like. For example, in [JJS93], Jayant proposed:

$$\begin{aligned}\Delta_{\text{tone masking noise}} &= 14.5 + z \text{ dB},\\ \Delta_{\text{noise masking tone}} &= C \text{ dB},\end{aligned} \tag{C.9}$$

where C varies between 3 and 6 dB depending upon experimental data. Note however that, as shown on the left of **figure 1.9**, there seems to be a frequency dependence for Δ even in the case of noise masking tones.

Addition of Masking Curves

Typical audio sounds contain more than one masker. The computation of individual masking curves at a certain time interval, consists of identifying tone-like and noise-like components and deriving their related masking patterns. This section deals with the computation of the overall masking curve, that is to say, how the individual masking curves interact or "add". The masking-curve summation effect can be described according to the following formula:

$$I_N = \left(\sum_{n=0}^{N-1} I_n^{\alpha} \right)^{\frac{1}{\alpha}} , \ 1 \leq \alpha \leq \infty, \tag{C.10}$$

where I_N represents the intensity of the masking curve that results from the combination of N individual masking curves with intensities I_n for $n = 0, \ldots, N-1$, and α is a parameter that defines the way the curves add. By choosing $\alpha = 1$, the resulting addition rule is the intensity addition of individual masking curves, whereas for $\alpha \to \infty$, the highest masking curve is used at each frequency location. By setting α lower than 1, we would assume that masking curves over-add in a non-linear way.

Even if over-adding of masking curves seems to be a real empirical effect, there is not complete agreement on this subject. Lufti in [Lut83] proposes a value of 0.33 for α in cases of up to four maskers. Nevertheless, adding numerous masking curves this way would lead to nonsensical situations. For example, 10 maskers combine to be equivalent to a masking curve 30 dB higher than each individual curve, meaning that in this case these maskers are masking themselves. In fact, we saw in section 1.2.3 that the loudness is integrated within critical bands, which means that tones will not be treated by the ear as independent maskers when their separation in frequency is less than a critical bandwidth.

Maskers close in frequency pose a problem, because beating can cause some unmasking effects, leading to an erroneous description of the perceived signal in current models.

Concerning the combination of the overall masking curve with the threshold in quiet, it is generally assumed that the maximum value between them is kept as the masked threshold.

C.4 Computation of Masking Curves

A high frequency resolution is necessary to locate frequency masking components and their resulting masking curve. Because the time-to-frequency mapping used by the coder may not have that accuracy, the psychoacoustic model uses for each block a high-resolution *discrete Fourier transform* (DFT). Hence, to make sure that this high-resolution block is time-synchronized with the data block being quantized,[1] the high-resolution block is centered on it.

A runtime problem arises when computing the masking curve of each component by convolving it with an appropriate spreading function. Bosi exhibits two typical solutions which are detailed in her book: 1) limiting the number of maskers, and 2) creating the masking curves using convolutions rather than a loop over maskers.

[1] Indeed, this high-resolution block is necessarily longer (in terms of time samples) than the block being quantized.

The last remaining issue concerns the mapping of high-resolution masking curves into SMRs to associate with each frequency bands of the "time-to-frequency mapping" module in the coder's main path. A difficulty appears because the coder frequency bandwidths are linear in frequency whereas critical bandwidths follow a nearly logarithm-like frequency progression, so that the coder frequency bands are wide compared to critical bands at low frequencies and *vice versa* at high frequencies. Since masking effects tend to be constant within a critical band, one solution is to choose:

- the average masking level in the critical band containing the coder frequency band when the coder band is narrow compared with critical bands ;
- the lowest masking level in the coder frequency band when the coder band is wide compared with critical bands, so that the masking level represents the most sensitive critical band in that coder band.

Then, for each frequency band of the coder, the SMR is computed, based on the amplitude of the largest spectral line in the band and on the masking level previously determined.

C.5 Bit Allocation Strategies

Given the constraint of a fixed bitrate and hence a fixed bit pool, this section is about the strategies of allocation of this pool to code each frequency subband.

C.5.1 A Simple Allocation Method by Thresholding

A first possibility of tonal redundancy removal could be to set up a threshold and to transmit only data for subbands whose signal amplitude is above this threshold. This implies passing a bit for each subband to tell the decoder whether the subband is coded or not. Although such a method is convenient for a variable bitrate coder, because each block will contain a different number of exceeding subbands, a slightly different method must be used for a fixed bitrate: the subband amplitudes can be sorted in descending order and then the bits distributed to the highest amplitudes first, until the bit pool is consumed.

C.5.2 Perceptual Entropy

Assuming that within each critical band the noise resulting from an R-bit quantization will not be audible as long as the SNR is higher than the SMR, it is possible to use the masked level values to allocate the quantization noise. Johnston [Joh88] defined the notion of perceptual entropy as the average minimum number of bits per frequency sample needed to code a signal while keeping SNR higher than SMR, and thus without introducing any perceptual distortion regarding the original signal:

$$PE = \frac{1}{N}\sum_{i=0}^{N-1} \max\left\{0, \log_2\left(\sqrt{\frac{I(f_i)}{I_T(f_i)}}\right)\right\} \approx \frac{1}{N}\sum_{i=0}^{N-1} \log_2\left(1+\sqrt{\frac{I(f_i)}{I_T(f_i)}}\right), \tag{C.11}$$

where, at each frequency f_i, I is the signal intensity, I_T is the relative intensity of the masked threshold, for a total of N frequencies in the signal representation. This could be a solution but, as previously, it is only convenient for a variable bitrate coder, since the quantization precision of each frequency sample will vary with time (depending on the masking threshold).

C.5.3 Optimal Bit Allocation for a Fixed Bitrate

In this section the case of a fixed bitrate is discussed, that is to say, given a fixed bit pool per block, how does one distribute bits throughout the spectrum of the signal in order to minimize the quantization noise. Because audio signals are generally colored, such a distribution leads to a coding gain.

Redundancy Removal Consideration Only

For the moment, we do not make use of psychoacoustic masking effects resulting from a perceptual model. Bosi demonstrates that in the case of uniform quantization, we minimize distortions by allocating the same number of bits to each spectral sample that we pass through the coder. Concerning the floating-point quantization, the expected error power for a spectral sample k quantized with a mantissa of R_k bits with a spectral sample amplitude x_k is roughly given by (see equation (C.4)):

$$\left\langle \epsilon^2 \right\rangle = \frac{x_k^2}{3 \times 2^{2R_k}}. \tag{C.12}$$

The average block squared error can be expressed as:

$$\left\langle q^2 \right\rangle_{\text{block}} = \frac{1}{K} \sum_{k=0}^{K-1} \left\langle \epsilon_k^2 \right\rangle, \tag{C.13}$$

where ϵ_k is the quantization error for spectral sample k and $\langle \epsilon_k^2 \rangle$ is the expected power of this quantization error. Replacing $\langle \epsilon_k^2 \rangle$ by its value from equation (C.12) we get:

$$\left\langle q^2 \right\rangle_{\text{block}} = \frac{1}{K} \sum_{k=0}^{K-1} \frac{x_k^2}{3 \times 2^{2R_k}} = \frac{1}{K} \sum_{k=0}^{K-1} \frac{1}{3 \times 2^{2R_k - \log_2(x_k^2)}}. \tag{C.14}$$

The fact that 2^{-Q} is convex prevents reducing the average block squared error by varying R_k from one frequency sample to another, because $2^{-(Q+\delta)} + 2^{-(Q-\delta)} \geq 2 \times 2^{-Q}$. So, the error is minimized when the exponent in the denominator is equal for all terms, that is to say:

$$2R_k - \log_2(x_k^2) = C, \text{ or} \tag{C.15}$$

$$R_k = \frac{1}{2}\left(C + \log_2(x_k^2)\right), \tag{C.16}$$

where C is a constant depending on the number of bits available to allocate to the mantissas in the block. This result suggests that it is necessary to allocate more bits to spectral samples with higher amplitudes.

It is possible to relate C to the size of the bit pool and the signal spectrum by averaging equation (C.15) over the K_p passed samples of the K spectral samples:

$$C = 2\left(\frac{P}{K_p}\right) - \frac{1}{K_p} \sum_{\text{passed k}} \log_2(x_k^2) = 2\left(\frac{P}{K_p}\right) - \log_2\left(\left|\prod_{\text{passed k}} x_k^2\right|^{\frac{1}{K_p}}\right), \tag{C.17}$$

where P designates the bit pool for the mantissas. The spectral samples that are not passed are allocated zero mantissa bits. By substituting this value of C in equation (C.16), Bosi reports the following optimal bit allocation result:

$$R_k^{\text{opt}} = \left(\frac{P}{K_p}\right) + \frac{1}{2}\log_2\left(x_k^2\right) - \left\langle \frac{1}{2}\log_2\left(x_k^2\right)\right\rangle_{\text{passed k}} \tag{C.18}$$

for all k bands with non-zero bit allocations.

For transform coders, spectral samples are grouped into subbands containing multiple spectral samples, and block floating point quantize the subband. In the case of B subbands indexed by b with N_b spectral samples in subband b and with maximum value of x_k^2 for that subband denoted as $x^2_{\mathrm{max}_b}$, the bit allocation equation for the spectral lines in subband b becomes:

$$R_b^{\mathrm{opt}} = \left(\frac{P}{K_p}\right) + \frac{1}{2}\log_2\left(x^2_{\mathrm{max}_b}\right) - \frac{1}{K_p}\sum_{\text{passed b}} N_b \frac{1}{2}\log_2\left(x^2_{\mathrm{max}_b}\right). \tag{C.19}$$

Integration of Perceptual Models

So far, we did not use perceptual thresholding informations in the optimal bit allocation. Now, besides redundancy removal, we would like to distribute the resulting quantization noise below masking curves. If the data rate constraint does not permit that, the *noise-to-mask ratio* (NMR), which is the difference between the SMR and the SNR, must be kept as low as possible in order to minimize the perceived noise. The only difference with the previous situation rests with the average block squared error equation (C.13), which now stands for a measure of perceptible distortion:

$$\left\langle q^2\right\rangle^{\text{percept}}_{\text{block}} = \frac{1}{K}\sum_{k=0}^{K-1}\frac{\langle\epsilon_k^2\rangle}{M_k^2}, \tag{C.20}$$

where M_k is the amplitude equivalent to the masking level evaluated at frequency index k. Then the optimal bit allocation can be derived from previous results (equation (C.19)):

$$R_b^{\mathrm{opt}} = \left(\frac{P}{K_p}\right) + \frac{1}{2}\log_2\left(\frac{x^2_{\mathrm{max}_b}}{M_b^2}\right) - \frac{1}{K_p}\sum_{\text{passed b}} N_b \frac{1}{2}\log_2\left(\frac{x^2_{\mathrm{max}_b}}{M_b^2}\right), \tag{C.21}$$

for all b with non-zero bit allocations (i.e., passed samples) where M_b is the amplitude corresponding to the masking level assumed to apply in subband b. This equation can be rewritten in terms of each subband SMR:

$$R_b^{\mathrm{opt}} = \left(\frac{P}{K_p}\right) + \frac{\ln(10)}{20\ln(2)}\left(\mathrm{SMR}_b - \frac{1}{K_p}\sum_{\text{passed b}} N_b\mathrm{SMR}_b\right), \tag{C.22}$$

where SMR_b represents the SMR that applies to subband b.

C.6 Bitstream Format

The decoder needs to know how to decode the data sent by the encoder, and in particular how many bits code each sample. So the bitstream contains, beside the sample quantized values themselves, the bit allocation information in order to dequantize them. The bitstream also contains other information such as scale factors (see section 2.5), or global information concerning the encoded data, such as sampling rate or copyrights.

The number of bits needed to describe this information is reduced using entropy coding methods [Huf52, RL79]. As an example, when using Huffman coding, since samples coded on zero bits are likely to appear frequently in the bitstream, the length of the code used to indicate a zero bit length sample will be short. All these correspondence codes are stored in a dictionary called the codebook. To avoid its transmission in the bitstream as well, it is usually predefined by presenting the coder with a variety of input signals, and defining the most common frequencies of occurrence.

Bibliography

[AG68] S. Axelrod and L.T. Guzy. Underestimation of dichotic click rates: results using methods of absolute estimation and constant stimuli. *Psychonomic Science*, 12(4):133–134, 1968. *Cited page 43*

[AMKJ05] T.L. Arbogast, C.R. Mason, and G. Kidd Jr. The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 117:2169–2180, 2005. *Cited page 39*

[Ash95] M.G. Ash. *Gestalt Psychology in German Culture, 1890-1967: Holism and the Quest for Objectivity.* Cambridge University Press, 1995. *Cited page 21*

[BA04] R. Boyer and K. Abed-Meraim. Audio modeling based on delayed sinusoids. *IEEE Transactions on Speech and Audio Processing*, 12(2):110–120, 2004. *Cited page 55*

[Bam95] J.S. Bamford. *An analysis of ambisonic sound systems of first and second order.* M.Sc. thesis, University of Waterloo, 1995. *Cited page 50*

[Bar26] H. Barkhausen. Ein neuer Schallmesser für die Praxis (A new sonometer for the practice). *Zeitschrift fur Technische Physik*, 7:599–601, 1926. *Cited page 14*

[BBQ+97] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa. ISO/IEC MPEG-2 Advanced Audio Coding. *Journal of the Audio Engineering Society*, 45(10):789–814, 1997. *Cited pages 56, 58, 60, and 61*

[BF07] J. Breebaart and C. Faller. *Spatial Audio Processing - MPEG Surround and Other Applications.* Engineering and Computer Science. Wiley, 2007. *Cited pages 61 and 135*

[BG02] M. Bosi and R.E. Goldberg. *Introduction to Digital Audio Coding and Standards.* Engineering and Computer Science. Kluwer Academic Publishers, 2002. *Cited pages 16, 18, 54, 61, 108, 149, and 152*

[BH02] J. Braasch and K. Hartung. Localization in the presence of a distracter and reverberation in the frontal horizontal plane. I. Psychoacoustical data. *Acta Acustica united with Acustica*, 88:942–955, November 2002. *Cited page 72*

[Bla68] J. Blauert. Ein Beitrag zur Theorie des Vorwärts-Rückwärts-Eindruckes beim Hören (A contribution to the theory of the front-back impression in hearing). In *Proceedings of the 6th International Congress on Acoustics*, pages A–3–10, Tokyo, 1968. *Cited page 37*

[Bla70] J. Blauert. Ein Versuch zum Richtungshören bei gleichzeitiger optischer Stimulation (An experiment in directional hearing with simultaneous optical stimulation). *Acustica*, 23:118–119, 1970. *Cited page 35*

[Bla97] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* Mit Press, 1997. *Cited pages 28, 34, 35, 36, and 37*

[Boe65] G. Boerger. *Die Lokalisation von Gausstönen (The localization of Gaussian tones).* Dissertation, Technische Universität Berlin, 1965. *Cited pages 35 and 82*

[BOG$^+$05] V. Best, E. Ozmeral, F.J. Gallun, K. Sen, and B.G. Shinn-Cunningham. Spatial unmasking of birdsong in human listeners: Energetic and informational factors. *The Journal of the Acoustical Society of America*, 118:3766–3773, 2005. *Cited page 40*

[BP78] A.S. Bregman and S. Pinker. Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, 32(1):19–31, 1978. *Cited page 26*

[BR75] A.S. Bregman and A.I. Rudnicky. Auditory segregation: Stream or streams? *Journal of Experimental Psychology*, 1(3):263–267, 1975. *Cited page 23*

[Bra99] K. Brandenburg. MP3 and AAC explained. *AES 17th International Conference on High Quality Audio Coding*, 1999. *Cited pages 54 and 61*

[Bre94] A.S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound.* The MIT Press, September 1994. Published: Paperback. *Cited pages 20, 22, 23, 24, 25, 26, 27, 34, 40, 41, and 42*

[Bro55] D.E. Broadbent. A note on binaural fusion. *The Quarterly Journal of Experimental Psychology*, 7(1):46—47, 1955. *Cited page 47*

[BS80] A.S. Bregman and H. Steiger. Auditory streaming and vertical localization: interdependence of “what” and “where” decisions in audition. *Perception & Psychophysics*, 28(6):539—546, 1980. *Cited page 46*

[Bus06] S. Busson. *Individualisation d’indices acoustiques pour la synthèse binaurale (Individualization of acoustic cues for binaural synthesis).* Ph.D. thesis, Université de la Méditerranée, Aix-Marseille II, January 2006. *Cited page 31*

[BvdPKS05] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers. Parametric coding of stereo audio. *EURASIP Journal on Applied Signal Processing*, 2005:1305–1322, January 2005. ACM ID: 1287201. *Cited pages 31, 61, 63, 67, and 69*

[CCS01] J.F. Culling, H.S. Colburn, and M. Spurchise. Interaural correlation sensitivity. *The Journal of the Acoustical Society of America*, 110(2):1020–1029, 2001. *Cited page 38*

[CG10] N.B.H. Croghan and D.W. Grantham. Binaural interference in the free field. *The Journal of the Acoustical Society of America*, 127(5):3085–3091, May 2010. *Cited pages 72, 103, and 119*

[CH58] E.M. Cramer and W.H. Huggins. Creation of pitch through binaural interaction. *The Journal of the Acoustical Society of America*, 30:413–417, 1958. *Cited page 44*

[CLH97] S. Carlile, P. Leong, and S. Hyams. The nature and distribution of errors in sound localization by human listeners. *Hearing Research*, 114(1-2):179–196, December 1997. *Cited pages 34 and 35*

[Cut76] J.E. Cutting. Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, 83(2):114–140, 1976. *Cited page 47*

[Dan01] J. Daniel. *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia (Representation of acoustic fields, application to the transmission and to the reproduction of complex sound scenes in a multimedia context).* Ph.D. thesis, Université Paris 6, July 2001. *Cited page 50*

[Dav99] G.A. Davidson. Digital audio coding: Dolby AC-3. In *Digital Signal Processing Handbook.* CRC Press LLC, V.K. Madisetti and D.B. Williams edition, 1999. *Cited pages 56 and 61*

[Dem82] L. Demany. Auditory stream segregation in infancy. *Infant Behavior and Development*, 5(26):1—76, 1982. *Cited page 21*

[Deu74a] D. Deutsch. An auditory illusion. *Nature*, 251(5473):307–309, 1974. *Cited page 27*

[Deu74b] D. Deutsch. An illusion with musical scales. *The Journal of the Acoustical Society of America*, 56:S25, 1974. *Cited page 41*

[Deu79] D. Deutsch. Binaural integration of melodic patterns. *Perception and Psychophysics*, 25(5):339–405, 1979. *Cited page 42*

[DG86] C.J. Darwin and R.B. Gardner. Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality. *The Journal of the Acoustical Society of America*, 79:838–845, 1986. *Cited page 22*

[DGM10] A. Daniel, C. Guastavino, and S. McAdams. Effet du rapport signal-sur-bruit sur l'angle minimum audible en présence d'un son distracteur (Effect of signal-to-noise ratio on minimum audible angle in the presence of a distracting sound). *Actes du dixième Congrès Français d'Acoustique*, April 2010. *Cited page 83*

[DH97] C.J. Darwin and R.W. Hukin. Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *The Journal of the Acoustical Society of America*, 102:2316–2324, 1997. *Cited page 45*

[DH98] C.J. Darwin and R.W. Hukin. Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset asynchrony. *The Journal of the Acoustical Society of America*, 103:1080–1084, 1998. *Cited page 45*

[DH00] C.J. Darwin and R.W. Hukin. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *The Journal of the Acoustical Society of America*, 107:970–977, 2000. *Cited page 45*

[DLKK02] M. Dietz, L. Liljeryd, K. Kjorling, and O. Kunz. Spectral band replication, a novel approach in audio coding. *AES 112th Convention*, Preprint 5553, April 2002. *Cited page 54*

[DMG+05] N.I. Durlach, C.R. Mason, F.J. Gallun, B.G. Shinn-Cunningham, H.S. Colburn, and G. Kidd Jr. Informational masking for simultaneous nonspeech stimuli: Psychometric functions for fixed and randomly mixed maskers. *The Journal of the Acoustical Society of America*, 118:2482–2497, 2005. *Cited page 39*

[DMKJ+03] N.I. Durlach, C.R. Mason, G. Kidd Jr, T.L. Arbogast, H.S. Colburn, and B.G. Shinn-Cunningham. Note on informational masking (L). *The Journal of the Acoustical Society of America*, 113:2984–2987, 2003. *Cited pages 38 and 40*

[DN10] A. Daniel and R. Nicol. Compression de flux audio multicanal (Compression of multichannel audio stream). *French patent application FR 1051420 (pending)*, February 2010. *Cited page 129*

[DN11] A. Daniel and R. Nicol. Allocation par sous-bandes de bits de quantification de paramètres d'information spatiale pour un codage paramétrique (Subband bit allocation for the quantization of spatial information parameters in parametric coding). *French patent application FR 1152602 (pending)*, February 2011. *Cited page 122*

[DNM03] J. Daniel, R. Nicol, and S. Moreau. Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging. *114th AES Convention, Amsterdam*, pages 22–25, 2003. *Cited pages 49, 50, and 52*

[DNM10] A. Daniel, R. Nicol, and S. McAdams. Multichannel audio coding based on minimum audible angles. In *Proceedings of the AES 40th Conference on Spatial Audio*, October 2010. *Cited page 129*

[Dre00] R. Dressler. Dolby surround pro logic decoder principles of operation. *Dolby Laboratories*, 2000. *Cited pages 57 and 58*

[dV09] D. de Vries. Wave field synthesis. *AES Monograph. AES, New York*, 2009. *Cited page 49*

[DW69] P. Damaske and B. Wagener. Richtungshörversuche über einen nachgebildeten Kopf (Investigations of directional hearing using a dummy head). *Acustica*, 21:30–35, 1969. *Cited pages 35 and 36*

[Fal04] C. Faller. *Parametric coding of spatial audio.* PhD thesis, École Polytechnique Fédérale de Lausanne, 2004. *Cited pages 61, 65, 66, 67, and 68*

[Fle40] H. Fletcher. Auditory patterns. *Reviews of Modern Physics*, 12(1):47–65, 1940. *Cited page 19*

[FM33] H. Fletcher and W.A. Munson. Loudness of a complex tone, its definition, measurement and calculation. *The Journal of the Acoustical Society of America*, 5:65, 1933. *Cited page 14*

[Fou11] The Wikimedia Foundation. Commons. http://commons.wikimedia.org, 2011. *Cited pages 12, 13, and 24*

[FSD08] W.L. Fan, T.M. Streeter, and N.I. Durlach. Effect of spatial uncertainty of masker on masked detection for nonspeech stimuli (L). *The Journal of the Acoustical Society of America*, 124:36–39, 2008. *Cited pages 39 and 40*

[Gai93] W. Gaik. Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. *The Journal of the Acoustical Society of America*, 94:98–110, 1993. *Cited page 31*

[Gar69] M.B. Gardner. Distance estimation of 0° or apparent 0°-Oriented speech signals in anechoic space. *The Journal of the Acoustical Society of America*, 45:47–53, 1969. *Cited page 37*

[Gar98] M.A. García-Pérez. Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Research*, 38(12):1861–1881, June 1998. *Cited page 87*

[Gar00] M.A. García-Pérez. Optimal setups for forced-choice staircases with fixed step sizes. *Spatial Vision*, 13(4):431–448, 2000. PMID: 11310536. *Cited pages 87, 88, and 99*

[Gau10] P. Gauthier. pdj - java interface for pure-data. http://www.le-son666.com/software/pdj/, 2010. *Cited page 76*

[GC81] K.J. Gabriel and H.S. Colburn. Interaural correlation discrimination: I. bandwidth and level dependence. *The Journal of the Acoustical Society of America*, 69(5):1394–1401, May 1981. *Cited page 38*

[GCLW99] M.A. Gerzon, P.G. Craven, M.J. Law, and R.J. Wilson. The MLP lossless compression system. *In Proceedings of the AES 17th Conference on High-Quality Audio Coding*, 1999. *Cited page 55*

[GDC+08] F.J. Gallun, N.I. Durlach, H.S. Colburn, B.G. Shinn-Cunningham, V. Best, C.R. Mason, and G. Kidd Jr. The extent to which a position-based explanation accounts for binaural release from informational masking. *The Journal of the Acoustical Society of America*, 124:439–449, 2008. *Cited pages 19, 39, and 40*

[Ger85] M.A. Gerzon. Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society*, 33(11):859–871, 1985. *Cited pages 50 and 56*

[Ger92] M.A. Gerzon. General metatheory of auditory localisation. *AES 92nd Convention*, Preprint 3306, March 1992. *Cited pages 53 and 65*

[GG91] A. Gersho and R.M. Gray. *Vector Quantization and Signal Compression.* Springer, 1 edition, November 1991. *Cited page 49*

[GHE03] D.W. Grantham, B.W.Y. Hornsby, and E.A. Erpenbeck. Auditory spatial resolution in horizontal, vertical, and diagonal planes. *The Journal of the Acoustical Society of America*, 114(2):1009–1022, 2003. *Cited page 37*

[GJ06] M. Goodwin and J.M. Jot. Analysis and synthesis for universal spatial audio coding. In *121th AES Convention, San Francisco, USA. Preprint*, volume 6874, 2006. *Cited page 68*

[GJ08] M. Goodwin and J.M. Jot. Spatial audio scene coding. *125th AES Convention*, October 2008. *Cited pages 62 and 66*

[GKJP83] D.M. Green, G. Kidd Jr, and M.C. Picardi. Successive versus simultaneous comparison in auditory intensity discrimination. *The Journal of the Acoustical Society of America*, 73:639–643, 1983. *Cited page 44*

[GM90] B.R. Glasberg and B.C.J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2):103–38, August 1990. PMID: 2228789. *Cited pages 20, 74, and 136*

[Goo08] M.M. Goodwin. Primary-ambient decomposition and dereverberation of two-channel and multichannel audio. *42nd Asilomar Conference on Signals, Systems and Computers*, pages 797–800, October 2008. *Cited pages 64 and 136*

[Gra84] D.W. Grantham. Interaural intensity discrimination: Insensitivity at 1000 hz. *The Journal of the Acoustical Society of America*, 75(4):1191–1194, April 1984. *Cited page 38*

[Gre61] D.D. Greenwood. Critical bandwidth and the frequency coordinates of the basilar membrane. *The Journal of the Acoustical Society of America*, 33(10):1344–1356, October 1961. *Cited page 20*

[Hau69] B.G. Haustein. Hypothesen über die einohrige Entfernungswahrnehmung des menschlichen Gehörs (Hypotheses about the perception of distance in human hearing with one ear). *Hochfrequenztech. u. Elektroakustik*, (78):46–57, 1969. *Cited page 37*

[HBL94] J. Herre, K. Brandenburg, and D. Lederer. Intensity stereo coding. *AES 96th Convention*, Preprint 3799, February 1994. *Cited page 60*

[HD69] R.M. Hershkowitz and N.I. Durlach. Interaural time and amplitude jnds for a 500-Hz tone. *The Journal of the Acoustical Society of America*, 46(6B):1464–1467, December 1969. *Cited page 38*

[HH84] J.W. Hall and A.D.G. Harvey. NoSo and NoSpi thresholds as a function of masker level for narrow-band and wideband masking noise. *The Journal of the Acoustical Society of America*, 76(6):1699–1703, December 1984. *Cited page 38*

[HHF84] J.W. Hall, M.P. Haggard, and M.A. Fernandes. Detection in noise by spectro-temporal pattern analysis. *The Journal of the Acoustical Society of America*, 76:50–56, 1984. *Cited page 22*

[Hil07] N.J. Hill. psignifit - free software package for fitting psychometric functions to psychophysical data. http://bootstrap-software.org/psignifit/, 2007. *Cited page 81*

[HKK98] I. Holube, M. Kinkel, and B. Kollmeier. Binaural and monaural auditory filter bandwidths and time constants in probe tone detection experiments. *The Journal of the Acoustical Society of America*, 104(4):2412–2425, October 1998. *Cited page 63*

[HR89] W.M. Hartmann and B. Rakerd. On the minimum audible angle—A decision theory approach. *The Journal of the Acoustical Society of America*, 85(5):2031–2041, May 1989. *Cited pages 73 and 74*

[HR10] L.M. Heller and V.M. Richards. Binaural interference in lateralization thresholds for interaural time and level differences. *The Journal of the Acoustical Society of America*, 128(1):310–319, July 2010. *Cited page 72*

[HS70] B.G. Haustein and W. Schirmer. Messeinrichtung zur Untersuchung des Richtungslokalisationsvermögens (A measuring apparatus for the investigation of the faculty of directional localization). *Hochfrequenztech. u. Elektroakustik*, 79:96–101, 1970. *Cited pages 34 and 35*

[HSS08] E. Hellerud, A. Solvang, and U.P. Svensson. Quantization of 2D higher order ambisonics. *124th AES Convention*, May 2008. *Cited page 139*

[Huf52] D.A. Huffman. A method for the construction of Minimum-Redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101, 1952. *Cited page 156*

[Hug74] A.W.F. Huggins. On perceptual integration of dichotically alternated pulse trains. *The Journal of the Acoustical Society of America*, 56:939–943, 1974. *Cited page 43*

[Jef48] L.A. Jeffress. A place theory of sound localization. *J. Comp. Physiol. Psychol*, 41(1):35–39, 1948. *Cited page 30*

[JF92] J.D. Johnston and A.J. Ferreira. Sum-difference stereo transform coding. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 569–572, Los Alamitos, CA, USA, 1992. IEEE Computer Society. *Cited pages 55 and 56*

[JJS93] N. Jayant, J. Johnston, and R. Safranek. Signal compression based on models of human perception. *Proceedings of the IEEE*, 81(10):1385–1422, 1993. *Cited page 152*

[JLP99] J.M. Jot, V. Larcher, and J.M. Pernaux. A comparative study of 3-D audio encoding and rendering techniques. In *16th Audio Engineering Society International Conference: Spatial Sound Reproduction*, 1999. *Cited page 68*

[JM84] W.M. Jenkins and M.M. Merzenich. Role of cat primary auditory cortex for sound-localization behavior. *Journal of Neurophysiology*, 52(5):819–847, November 1984. *Cited page 44*

[JMG$^+$07] J.M. Jot, J. Merimaa, M. Goodwin, A. Krishnaswamy, and J. Laroche. Spatial audio scene coding in a universal Two-Channel 3-D stereo format. In *123th Audio Engineering Society Convention*, 2007. *Cited pages 66 and 68*

[Joh88] J.D. Johnston. Estimation of perceptual entropy using noise masking criteria. *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 2524–2527, 1988. *Cited page 154*

[JSY98] P.X. Joris, P.H. Smith, and T.C. Yin. Coincidence detection in the auditory system: 50 years after jeffress. *Neuron*, 21(6):1235–1238, December 1998. PMID: 9883717. *Cited page 30*

[Jud77] T. Judd. An explanation of deutsch's scale illusion. *Unpublished manuscript, Department of psychology, Cornell University*, 1977. *Cited page 42*

[JY95] P.X. Joris and T.C. Yin. Envelope coding in the lateral superior olive. i. sensitivity to interaural time differences. *Journal of Neurophysiology*, 73(3):1043 –1062, March 1995. *Cited page 30*

[KÏ5] A. Körte. Kinomatoscopishe Untersuchungen (Kinematoscopic investigations). *Zeitschrift für Psychologie der Sinnesorgane*, 72:193–296, 1915. *Cited page 22*

[KBS07] N. Kopco, V. Best, and B.G. Shinn-Cunningham. Sound localization with a preceding distractor. *The Journal of the Acoustical Society of America*, 121(1):420–432, January 2007. *Cited page 119*

[KE56] R.G. Klumpp and H.R. Eady. Some measurements of interaural time difference thresholds. *The Journal of the Acoustical Society of America*, 28(5):859–860, 1956. *Cited page 38*

[Ken95] G.S. Kendall. The decorrelation of audio signals and its impact on spatial imagery. *Computer Music Journal*, 19(4):71–87, December 1995. *Cited page 69*

[KG90] B. Kollmeier and R.H. Gilkey. Binaural forward and backward masking: Evidence for sluggishness in binaural detection. *The Journal of the Acoustical Society of America*, 87(4):1709–1719, April 1990. *Cited page 63*

[KH76] M. Kubovy and F.P. Howard. Persistence of a pitch-segregating echoic memory. *Journal of Experimental Psychology: Human Perception and Performance*, 2(4):531–537, 1976. *Cited page 44*

[KJMBH05] G. Kidd Jr, C.R. Mason, A. Brughera, and W.M. Hartmann. The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acustica united with Acta Acustica*, 91(3):526–536, 2005. *Cited page 39*

[KMMS07] F. Kozamernik, D. Marston, A. Mason, and G. Stoll. Ebu tests of Multi-Channel audio codecs. In *122nd Audio Engineering Society Convention*, May 2007. *Cited page 69*

[KS03] N. Kopčo and B.G. Shinn-Cunningham. Spatial unmasking of nearby puretone targets in a simulated anechoic environment. *The Journal of the Acoustical Society of America*, 114:2856–2870, 2003. *Cited page 40*

[KT03] M.C. Kelly and A.I. Tew. A novel method for the efficient comparison of spatialization conditions. *Audio Engineering Society; 1999*, 2003. *Cited page 72*

[Kub81] M. Kubovy. Concurrent-pitch segregation and the theory of indispensable attributes. *Perceptual organization*, pages 55–98, 1981. *Cited page 43*

[Kuh77] G.F. Kuhn. Model for the interaural time differences in the azimuthal plane. *The Journal of the Acoustical Society of America*, 62:157–167, 1977. *Cited page 30*

[LB02] E.H.A. Langendijk and A.W. Bronkhorst. Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, 112(4):1583–1596, October 2002. *Cited page 33*

[LCYG99] R.Y. Litovsky, H.S. Colburn, W.A. Yost, and S.J. Guzman. The precedence effect. *The Journal of the Acoustical Society of America*, 106:1633–1654, 1999. *Cited pages 44 and 57*

[LDS09] A.K.C. Lee, A. Deane-Pratt, and B.G. Shinn-Cunningham. Localization interference between components in an auditory scene. *The Journal of the Acoustical Society of America*, 126(5):2543–2555, November 2009. *Cited page 102*

[LJ64] T.L. Langford and L.A. Jeffress. Effect of noise crosscorrelation on binaural signal detection. *The Journal of the Acoustical Society of America*, 36(8):1455–1458, 1964. *Cited page 38*

[LSO05] A.K.C. Lee, B.G. Shinn-Cunningham, and A.J. Oxenham. The missing target: Evidence of a tone's inability to contribute to the auditory foreground'. *Proc. Mid-winter meeting of the ARO, New Orleans*, 2005. *Cited page 46*

[Lut83] R.A. Lutfi. Additivity of simultaneous masking. *Journal of the Acoustical Society of America*, 73(1):262–267, 1983. *Cited page 153*

[Lut93] R.A. Lutfi. A model of auditory pattern analysis based on component-relative-entropy. *The Journal of the Acoustical Society of America*, 94:748–758, 1993. *Cited page 39*

[Mak62] Y. Makita. On the directional localization of sound in the stereophonic sound field. *EBU Review*, pages 102–108, 1962. *Cited page 130*

[Mak75] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975. *Cited page 53*

[MB97] S. McAdams and J. Bertoncini. Organization and discrimination of repeating sound sequences by newborn infants. *The Journal of the Acoustical Society of America*, 102(5):2945–2953, 1997. *Cited page 21*

[MBT$^+$07] T. Moeck, N. Bonneel, N. Tsingos, G. Drettakis, I. Viaud-Delmon, and D. Alloza. Progressive perceptual audio rendering of complex scenes. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, I3D '07, page 189–196, Seattle, Washington, 2007. ACM. ACM ID: 1230133. *Cited page 142*

[MG91] J.C. Middlebrooks and D.M. Green. Sound localization by human listeners. *Annual Review of Psychology*, 42(1):135–159, 1991. *Cited pages 31 and 35*

[MG96] B.C.J. Moore and B.R. Glasberg. A revision of zwicker's loudness model. *Acta Acustica United with Acustica*, 82(2):335–345, 1996. *Cited page 20*

[MGM11] G. Marentakis, C. Griffiths, and S. McAdams. Detecting spatial displacement in musical scenes. In *Forum Acusticum*, Aalborg, Denmark, 2011. *Cited page 72*

[Mil58] A.W. Mills. On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4):237–246, 1958. *Cited pages 35, 36, 73, and 82*

[Mil60] A.W. Mills. Lateralization of High-Frequency tones. *The Journal of the Acoustical Society of America*, 32(1):132–134, 1960. *Cited page 38*

[MJP01] D. McAlpine, D. Jiang, and A.R. Palmer. A neural code for low-frequency sound localization in mammals. *Nat Neurosci*, 4(4):396–401, April 2001. *Cited page 30*

[Mor06] S. Moreau. *Étude et réalisation d'outils avancés d'encodage spatial pour la technique de spatialisation sonore Higher Order Ambisonics : microphone 3D et contrôle de distance (Study and realization of advanced spatial encoding tools for the Higher Order Ambisonics sound spatialization method: 3D microphone and distance control).* Ph.D. thesis, Université du Maine, July 2006. *Cited pages 51 and 53*

[MP76] D. McFadden and E.G. Pasanen. Lateralization at high frequencies based on interaural time differences. *The Journal of the Acoustical Society of America*, 59(3):634–639, March 1976. *Cited page 31*

[MPM] G. Marentakis, N. Peters, and S. McAdams. Sensitivity to sound displacement for virtual sources. *(in preparation).* *Cited page 72*

[MPM07] G. Marentakis, N. Peters, and S. McAdams. Auditory resolution in auditory virtual environments. *The Journal of the Acoustical Society of America*, 122(5):3054, November 2007. *Cited page 72*

[ND95] D.L. Neff and T.M. Dethlefs. Individual differences in simultaneous masking with random-frequency, multicomponent maskers. *The Journal of the Acoustical Society of America*, 98:125–134, 1995. *Cited page 39*

[Nor66] D.A. Norman. Rhythmic fission : Observations on attention, temporal judgments and the critical band. *Unpublished manuscript, Center for Cognitive Studies, Harvard University*, 1966. *Cited page 41*

[OFMKJ03] A.J. Oxenham, B.J. Fligor, C.R. Mason, and G. Kidd Jr. Informational masking and musical training. *The Journal of the Acoustical Society of America*, 114(3):1543–1549, 2003. *Cited page 39*

[OP84] S.R. Oldfield and S.P.A. Parker. Acuity of sound localization: a topography of auditory space. i. normal hearing conditions. *Perception*, 13(5):581–600, 1984. *Cited pages 34 and 35*

[PB69] D.R. Perrott and S.H. Barry. Binaural fusion. *The Journal of Auditory Research*, 3:263–269, 1969. *Cited page 44*

[PEF98] H. Purnhagen, B. Edler, and C. Ferekidis. Object-Based Analysis/Synthesis audio coder for very low bit rates. *AES 104th Convention*, Preprint 4747, 1998. *Cited page 54*

[Per84a] D.R. Perrott. Concurrent minimum audible angle: A re-examination of the concept of auditory spatial acuity. *The Journal of the Acoustical Society of America*, 75(4):1201–1206, April 1984. *Cited page 63*

[Per84b] D.R. Perrott. Discrimination of the spatial distribution of concurrently active sound sources: Some experiments with stereophonic arrays. *The Journal of the Acoustical Society of America*, 76(6):1704–1712, December 1984. *Cited page 63*

[PJB87] J. Princen, A. Johnson, and A. Bradley. Subband/transform coding using filter bank designs based on time domain aliasing cancellation. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87.*, 12, 1987. *Cited pages 7, 135, and 149*

[PMMS89] D.R. Perrott, K. Marlborough, P. Merrill, and T.Z. Strybel. Minimum audible angle thresholds obtained under conditions in which the precedence effect is assumed to operate. *The Journal of the Acoustical Society of America*, 85(1):282–288, 1989. *Cited page 142*

[PR86] A.R. Palmer and I.J. Russell. Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing Research*, 24(1):1–15, 1986. *Cited pages 28 and 31*

[Pre66] R. Preibisch-Effenberger. *Die Schallokalisationsfähigkeit des Menschen und ihre audiometrische Verwendbarkeit zur klinischen Diagnostik (The human faculty of sound localization and its audiometric application to clinical diagnostics).* Habilitationsschrift, Medizinische Akademie, Dresden, 1966. *Cited pages 34 and 35*

[PRH+92] R.D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. Complex sounds and auditory images. *Auditory Physiology and Perception*, pages 429–446, 1992. *Cited pages 20 and 74*

[PS90] D.R. Perrott and K. Saberi. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *The Journal of the Acoustical Society of America*, 87(4):1728–1731, April 1990. *Cited pages 35 and 37*

[Pul97] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc*, 45(6):456–466, 1997. *Cited pages 8, 49, and 68*

[Pul07] V. Pulkki. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516, June 2007. *Cited pages 62, 63, 68, and 69*

[PZ09] H. Pomberger and F. Zotter. An ambisonics format for flexible playback layouts. *Ambisonics Symposium 2009, Graz*, June 2009. *Cited page 138*

[Ray07] L. Rayleigh. On our perception of sound direction. *Philos. Mag*, 13:214–232, 1907. *Cited pages 28 and 34*

[RJ63] D.E. Robinson and L.A. Jeffress. Effect of varying the interaural noise correlation on the detectability of tonal signals. *The Journal of the Acoustical Society of America*, 35(12):1947–1952, December 1963. *Cited page 38*

[RL79] J. Rissanen and G.G. Langdon. Arithmetic coding. *IBM Journal of Research and Development*, 23:149–162, March 1979. ACM ID: 1664217. *Cited page 156*

[RT67] R.C. Rowland and J.V. Tobias. Interaural intensity difference limen. *J Speech Hear Res*, 10(4):745–756, December 1967. *Cited page 38*

[Rum01] F. Rumsey. *Spatial Audio*. Engineering and Computer Science. Focal Press, Oxford, 2001. *Cited pages 33, 49, and 56*

[SB82] H. Steiger and A.S. Bregman. Competition among auditory streaming, dichotic fusion, and diotic fusion. *Perception & Psychophysics*, 32(2):153–162, 1982. *Cited page 45*

[Sch70] B. Scharf. Critical bands. In *Foundations of modern auditory theory*, volume 1, page 159–202. Academic Press, New York, J. V. Tobias edition, 1970. *Cited page 20*

[Sch83] M.T.M. Scheffers. Sifting vowels. *Unpublished doctoral dissertation, Groningen University*, 1983. *Cited page 47*

[Sha49] C.E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949. *Cited page 49*

[She81] R.N. Shepard. Psychophysical complementarity. *Perceptual organization*, pages 279–341, 1981. *Cited page 41*

[Shi05] B.G. Shinn-Cunningham. Influences of spatial cues on grouping and understanding sound. *Proceedings of Forum Acusticum*, 29:1539–1544, 2005. *Cited pages 39 and 45*

[SIM94] W.H. Slattery III and J.C. Middlebrooks. Monaural sound localization: Acute versus chronic unilateral impairment. *Hearing Research*, 75(1-2):38–46, May 1994. *Cited page 33*

[SMM05] A.T. Sabin, E.A. Macpherson, and J.C. Middlebrooks. Human sound localization at near-threshold levels. *Hearing Research*, 199(1-2):124–134, January 2005. PMID: 15574307. *Cited pages 91, 95, and 112*

[SMR$^+$03] Y. Suzuki, V. Mellert, U. Richter, H. Moller, L. Nielsen, R. Hellman, K. Ashihara, K. Ozawa, and H. Takeshima. Precise and full-range determination of two-dimensional equal loudness contours. Technical report, Tohoku University, Japan, 2003. *Cited pages 14 and 16*

[SN36] S.S. Stevens and E.B. Newman. The localization of actual sources of sound. *The American Journal of Psychology*, 48(2):297–306, April 1936. *Cited page 35*

[SP55] E.D. Schubert and C.D. Parker. Addition to cherry's findings on switching speech between the two ears. *The Journal of the Acoustical Society of America*, 27:792–794, 1955. *Cited page 42*

[SPSG06] E.K. Samsudin, N.B. Poh, F. Sattar, and F. George. A stereo to mono dowmixing scheme for MPEG-4 parametric stereo encoder. *Proc. ICASSP'06*, 2006. *Cited page 67*

[SR01] T.I. Su and G.H. Recanzone. Differential effect of near-threshold stimulus intensities on sound localization performance in azimuth and elevation in normal human subjects. *Journal of the Association for Research in Otolaryngology: JARO*, 2(3):246–256, September 2001. PMID: 11669397. *Cited pages 91, 95, and 112*

[SS11] M.L. Sutter and S.A. Shamma. The relationship of auditory cortical activity to perception and behavior. In *The Auditory Cortex*, page 617–641. Springer US, 2011. *Cited page 44*

[SSK00] B.G. Shinn-Cunningham, S. Santarelli, and N. Kopco. Tori of confusion: Binaural localization cues for sources within reach of a listener. *The Journal of the Acoustical Society of America*, 107:1627–1636, 2000. *Cited pages 30 and 32*

[Ter79] E. Terhardt. Calculating virtual pitch. *Hearing Res*, 1(2):155–182, 1979. *Cited pages 14 and 152*

[tK96] W.R.T. ten Kate. Compatibility matrixing of multichannel Bit-Rate-Reduced audio signals. *Journal of the Audio Engineering Society*, 44(12):1104–1119, December 1996. *Cited pages 56 and 60*

[TMR67] W.R. Thurlow, J.W. Mangels, and P.S. Runge. Head movements during sound localization. *The Journal of the Acoustical Society of America*, 42:489, 1967. *Cited page 33*

[Tol03] D.J. Tollin. The lateral superior olive: A functional role in sound source localization. *The Neuroscientist*, 9(2):127–143, April 2003. *Cited page 30*

[vB60] G. von Békésy. *Experiments in hearing (Wever, EG, Trans.)*. New York, NY: McGraw-Hill, 1960. *Cited pages 11 and 12*

[vN75] L. van Noorden. Temporal coherence in the perception of tone sequences. *Unpublished doctoral dissertation, Eindhoven University of Technology*, 1975. *Cited page 41*

[WA01] D.B. Ward and T.D. Abhayapala. Reproduction of a plane-wave sound field using an array of loudspeakers. *IEEE Transactions on Speech and Audio Processing*, 9(6):697–707, 2001. *Cited page 52*

[Wal40] H. Wallach. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27(4):339–368, 1940. *Cited page 33*

[WH01] F.A. Wichmann and N.J. Hill. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, 63(8):1293–1313, 2001. *Cited page 81*

[WK92] F.L. Wightman and D.J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, 91:1648–1661, 1992. *Cited page 31*

[WK99] F.L. Wightman and D.J. Kistler. Resolution of front—back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America*, 105:2841–2853, 1999. *Cited pages 32, 33, and 76*

[WKHP03] M. Wolters, K. Kjorling, D. Homm, and H. Purnhagen. A closer look into MPEG-4 high efficiency AAC. In *115th Convention of the Audio Engineering Society*, October 2003. *Cited page 54*

[WNR49] H. Wallach, E.B. Newman, and M.R. Rosenzweig. The precedence effect in sound localization. *American Journal of Psychology*, 62:315–336, 1949. *Cited pages 44 and 57*

[WS54] R.S. Woodworth and H. Schlosberg. Experimental psychology (Rev. ed.). *New York: Holt*, 1954. *Cited page 31*

[WW80] R.B. Welch and D.H. Warren. Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, 88(3):638–667, 1980. *Cited page 45*

[YAKK03] D. Yang, H. Ai, C. Kyriakakis, and C. Kuo. High-fidelity multichannel audio coding with Karhunen-Loève transform. *IEEE Transactions on Speech and Audio Processing*, 11(4):365–380, 2003. *Cited page 58*

[YH87] W.A. Yost and E.R. Hafter. Lateralization. In *Directional hearing*, page 49–84. Springer, New York, 1987. *Cited page 38*

[YKK04] D.T. Yang, C. Kyriakakis, and C.-C.J. Kuo. *High-Fidelity Multichannel Audio Coding*. Hindawi, 2004. *Cited pages 58 and 59*

[Yos72] W.A. Yost. Weber's fraction for the intensity of pure tones presented binaurally. *Perception & Psychophysics*, 11(1):61–64, January 1972. *Cited page 38*

[Yos74] W.A. Yost. Discriminations of interaural phase differences. *The Journal of the Acoustical Society of America*, 55(6):1299–1303, June 1974. *Cited page 38*

[Yos94] W.A. Yost. *Fundamentals of Hearing : An introduction*. San Diego : Academic Press, 1994. *Cited page 29*

[Zah02] P. Zahorik. Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America*, 111(4):1832–1846, April 2002. *Cited pages 33 and 37*

[ZF56] J. Zwislocki and R.S. Feldman. Just noticeable differences in dichotic phase. *The Journal of the Acoustical Society of America*, 28(5):860–864, 1956. *Cited pages 31 and 38*

[ZF67] E. Zwicker and R. Feldtkeller. *Das Ohr als Nachrichtenempfänger (The ear as a receiver of information)*. Hirzel Verlag, Stuttgart, 1967. *Cited pages 18 and 19*

[ZF07] E. Zwicker and H. Fastl. *Psychoacoustics - Facts and Models*. Springer-Verlag, Berlin, 2007. *Cited pages 11, 14, 15, 17, 18, and 19*

[ZP01] R.J. Zatorre and V.B. Penhune. Spatial localization after excision of human auditory cortex. *The Journal of Neuroscience*, 21(16):6321–6328, 2001. *Cited page 44*

Abstract

Spatial Auditory Blurring and Applications to Multichannel Audio Coding

This work concerns the telecommunications area, and more specifically the transmission of multichannel audio signals. Four psychoacoustic experiments were carried out in order to study the spatial resolution of the auditory system—also known as *localization blur*—in the presence of distracting sounds. As a result, localization blur increases when these distracters are present, bringing to light what we will refer to as the phenomenon of “spatial blurring”. These experiments assess the effect of several variables on spatial blurring: the frequencies of both the sound source under consideration and the distracting sources, their level, their spatial position, and the number of distracting sources. Except for the spatial position of distracting sources, all of these variables have been shown to have an effect on spatial blurring.

This thesis also deals with the modeling of this phenomenon, in order to predict auditory spatial resolution as a function of the sound scene characteristics (number of present sources, their frequency and their level).

Finally, two multichannel audio coding schemes are proposed that take advantage of this model in order to reduce the information to be transmitted: one is based on a parametric representation (downmix + spatial parameters) of the multichannel signal and the other is based on the Higher-Order Ambisonics (HOA) representation. These schemes are both based on the original idea of dynamically adjusting the spatial accuracy of representation of the multichannel signal in a way that shapes the resulting spatial distortions within localization blur, such that they remain unnoticeable.

Keywords Psychoacoustics, Spatial Hearing, Spatial Audio Coding, Multichannel Audio, Minimum Audible Angle, Localization Blur, Parametric Coding, Higher-Order Ambisonics, Auditory Scene Analysis

Résumé

Floutage Spatial Auditif et Applications au Codage Audio Multicanal

Ce travail se place en contexte de télécommunications, et concerne plus particulièrement la transmission de signaux audio multicanaux. Quatre expériences psychoacoustiques ont été menées de façon à étudier la résolution spatiale du système auditif — également appelée *flou de localisation* — en présence de sons distracteurs. Il en résulte que le flou de localisation augmente quand ces distracteurs sont présents, mettant en évidence ce que nous appellerons le phénomène de « floutage spatial » auditif. Ces expériences estiment l’effet de plusieurs variables sur le floutage spatial : la fréquence de la source sonore considérée ainsi que celles des sources distractrices, leur niveau sonore, leur position spatiale, et le nombre de sources distractrices. Exceptée la position des sources distractrices, toutes ces variables ont montré un effet significatif sur le floutage spatial.

Cette thèse aborde également la modélisation de ce phénomène, de sorte que la résolution spatiale auditive puisse être prédite en fonction des caractéristiques de la scène sonore (nombre de sources présentes, leur fréquence, et leur niveau).

Enfin, deux schémas de codage audio multicanaux exploitant ce modèle à des fins de réduction de l’information à transmettre sont proposés : l’un basé sur une représentation paramétrique (downmix + paramètres spatiaux) du signal multicanal, et l’autre sur la représentation Higher-Order Ambisonics (HOA). Ces schémas sont tous deux basés sur l’idée originale d’ajuster dynamiquement la précision de la représentation spatiale du signal multicanal de façon à maintenir les distorsions spatiales résultantes dans le flou de localisation, afin que celles-ci restent indétectables.

Mots-clefs Psychoacoustique, Écoute Spatialisée, Codage Audio Spatialisé, Son Multicanal, Angle Minimum Audible, Flou de Localisation, Codage Paramétrique, Ambisonie d’Ordres Supérieurs, Analyse de Scène Auditive