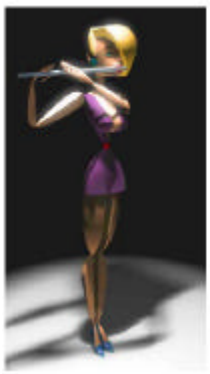
Helsinki University of Technology Laboratory of Acoustics and Audio Signal Processing
Espoo 1999 Report 53

VIRTUAL ACOUSTICS AND 3-D SOUND IN MULTIMEDIA SIGNAL PROCESSING

Jyri Huopaniemi



Helsinki University of Technology Laboratory of Acoustics and Audio Signal Processing
Espoo 1999 Report 53

VIRTUAL ACOUSTICS AND 3-D SOUND IN MULTIMEDIA SIGNAL PROCESSING

Jyri Huopaniemi

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission for public examination and debate in Auditorium S4, Department of Electrical and Communications Engineering, Helsinki University of Technology (Espoo, Finland) on the 5th of November, 1999, at 12 o'clock noon.

Helsinki University of Technology
Deparment of Electrical and Communications Engineering
Laboratory of Acoustics and Audio Signal Processing

Teknillinen korkeakoulu
Sähkö- ja tietoliikennetekniikan osasto
Akustiikan ja äänenkäsittelytekniikan laboratorio

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
P.O.Box 3000
FIN-02015 HUT
Tel. +358 9 4511
Fax +358 9 460 224
E-mail lea.soderman@hut.fi

ISBN 951-22-4706-2
ISSN 1456-6303

Libella Oy
Espoo, Finland 1999

Abstract

In this work, aspects in real-time modeling and synthesis of three-dimensional sound in the context of digital audio, multimedia, and virtual environments are studied. The concept of virtual acoustics is discussed, which includes models for sound sources, room acoustics, and spatial hearing.

Real-time virtual acoustics modeling is carried out using a real-time parametric room impulse response rendering technique proposed by the author. The algorithm uses a well-known time-domain hybrid model, in which the direct sound and early energy are modeled using geometrical acoustics, and diffuse late energy is modeled using a recursive reverberator. Sound source directivity, and dynamic room acoustical phenomena such as time-varying early reflections from boundary materials and air absorption are incorporated in the image-source model using low-order digital filter approximations. A novel technique for estimation of source directivity based on the reciprocity method is introduced.

Optimization of binaural systems is discussed. Novel methods for head-related transfer function approximation based on frequency warping and balanced model truncation are proposed by the author. Objective and subjective methods for estimating the quality of virtual source generation in headphone listening are introduced. For objective analysis, a binaural auditory model is used, which predicts well the expected performance of different filter designs. Subjective listening tests have been carried out to compare design techniques. The results show that binaural filter design based on auditory criteria results in more efficient and more perceptually relevant processing. This is the main result of the thesis.

Crosstalk canceled loudspeaker reproduction of 3-D sound is reviewed. Stereo widening and virtual loudspeaker algorithms are discussed. An efficient, high-quality design method of virtual surround filters for playback of multichannel audio on two loudspeakers based on a warped shuffler structure is presented. The performance of the method has been verified in subjective listening tests.

The methods and results presented in this thesis are applicable to digital audio signal processing in general, and especially to implementation of virtual acoustic environments in multimedia and virtual reality systems.

Keywords: 3-D sound, auralization, head-related transfer function, binaural technology, crosstalk canceling, spatial hearing, digital signal processing, auditory displays, virtual acoustics, virtual environments

Preface

This work is a result of research carried out at the Laboratory of Acoustics and Audio Signal Processing of the Helsinki University of Technology (HUT) in Espoo, Finland, during the years 1995-1998, at the Center for Computer Research in Music and Acoustics (CCRMA) of Stanford University during the year 1998, and finalized in 1998-1999 at Nokia Research Center in Helsinki, Finland. I am most grateful to my supervisor, Professor Matti Karjalainen, for his continuous interest and support in my work over the years, and for our productive collaboration in all research areas of this thesis. I'm also very grateful to Prof. Julius O. Smith III for inviting and supervising my research work at CCRMA and for many intriguing discussions and collaboration on HRTF filter design. Furthermore, I'd like to thank the pre-examiners of my thesis, Prof. Richard Duda (San Jose State University) and Dr. Tapio Lahti (Insinööritoimisto Akukon) for positive feedback and very valuable comments.

During my thesis work, I had close cooperation with many researchers. A number of early publications on physical modeling I co-authored with Prof. Karjalainen and Dr. Vesa Välimäki formed the basis for my research into virtual acoustics. I'd like to thank Matti and Vesa for very fruitful collaboration, and for guidance, reading and commenting most of my publications, including this thesis. The Digital Interactive Virtual Acoustics (DIVA) group, headed by Prof. Tapio Takala (HUT Laboratory of Telecommunications Software and Multimedia), has been a project which has influenced my research greatly. I'd like to thank the DIVA group, especially Mr. Lauri Savioja and Mr. Tapio Lokki for collaboration on implementation issues and on many excellent research articles. I worked with the 3-D sound group at HUT Acoustics Lab on many interesting topics in 1996-1998. I'd like to thank Ms. Riitta Väänänen for our productive collaboration in MPEG-4 scene description, in the DIVA project, and in many publications. Mr. Klaus Riederer was the key person in obtaining the HRTF data which is the basis for many parts of this work. Mr. Martti Rahkila and I have collaborated on many internet and web related projects. I admire Mara's style and net knowledge. Mr. Tero Tolonen's expertise in DSP, physical modeling and music has been inspiring. I've enjoyed Dr. Unto Laine's wisdom in warping, audio DSP, and philosophy. Mr. Ville Pulkki's ideas and our cooperation in binaural auditory modeling have been very fruitful. Mr. Aki Härmä and I have shared common in-

terests on research of frequency warping and its applications. Mr. Jussi Hynninen created the ultimate listening test software, GP2, which was used in the virtual surround test. Mr. Antti Järvinen has introduced me to many interesting people, and we've collaborated on many research issues. Ms. Lea Söderman has assisted me in many practical issues over the years. I'm very grateful to these fine persons and the entire HUT Acoustics Lab personnel for cooperation, friendships and an inspiring working atmosphere during the five years I spent there as a researcher.

My colleagues at Nokia Research Center (NRC), Speech and Audio Systems Laboratory, are thanked for professional and academic support in this work. In particular I'd like to thank Dr. Petri Haavisto, Mr. Mauri Väänänen, Mr. Matti Hämäläinen and Mr. Nick Zacharov. I've co-authored research papers with Nick and Matti, and I've enjoyed Nick's expertise in subjective experiments and Matti's knowledge in audio DSP. I'd also like to thank the entire NRC Speech and Audio Systems Lab's 3-D Audio Group for hard-working research efforts and for making a great working atmosphere. It has been a pleasure to collaborate with Mr. Juha Backman (Nokia Mobile Phones) during many years. His knowledge of physics, acoustics and audio is remarkable. I'd like to thank Dr. Durand Begault and Dr. Elizabeth Wenzel (NASA Ames Research Center) for their continuing interest in our virtual acoustics research, and for very inspiring discussions. Dr. Jonathan Mackenzie is acknowledged for collaboration in HRTF modeling using the BMT method. I've had the pleasure to work together with Mr. Eric Scheirer (MIT Media Lab) on MPEG-4 audio scene description related work, and we've collaborated on two publications. Dr. Jean-Marc Jot (Creative Labs) and I have had many very fruitful discussions on virtual acoustics, and we've worked together on MPEG-4 audio rendering issues. I'd like to sincerely thank these great researchers for their interest in my work.

There are also many friends and my family I'd like to acknowledge. Very special thanks go to Mr. Saku Heiskanen, Ms. Kaisu Iisakkila, Mr. Martti Rahkila and Ms. Outi Kosonen for friendship and great company. I've played music with Saku in a number of bands for 15 years, and it continues to bring a lot of joy into my life. The friends I lived with in Toukola, the HC Finland, the Munich people, the Mielentila/Mimedia crew, the Lautta musicians are acknowledged for making me do, see and hear other things than work every now and then. Most importantly, I'd like to thank my girlfriend, Ms. Patty Huang, for everything, including valuable comments to the thesis manuscript. Finally, I'd like to thank my parents and their families for support all through my years of research. Special thanks go to my mother, Ms. Sinikka Huopaniemi, Mr. Vesa Hatakka, and my father, Mr. Hannu Huopaniemi, Ms. Annariitta Huopaniemi, and my brothers, Mikko and Juha. The financial support of Nokia Research Center, the Academy of Finland, the Emil Aaltonen Foundation, the Helsinki University of Technology Foundation, and the Nokia Foundation is greatly acknowledged.

Lauttasaari, Helsinki, October 13, 1999 Jyri Huopaniemi

Table of Contents

List of Symbols

a	radius of human head; attenuation in dB per meter
$\alpha(\omega)$	absorption coefficient
c	speed of sound in air
$c(n)$	cepstrum
$C(z)$	crosstalk canceling filter
δf_{CE}	ERB scale frequency bandwidth
δf_{CB}	Bark scale frequency bandwidth
f	temporal frequency
f_s	sampling frequency
f_{ro}	oxygen relaxation frequency
f_{rn}	nitrogen relaxation frequency
$F_k(z)$	listener model filter block
g	amplification factor
h	molar concentration of water vapor
$h(n)$	impulse response
$h_A, \ldots, h_E$	HRTF filter coefficients
$h_m(kr)$	spherical Hankel function of order m in kr
$h'_m(ka)$	derivative of spherical Hankel function of order m with respect to ka
$H(e^{j\omega})$	frequency response
λ	warping factor
n	integer variable
p_r	reference ambient atmospheric pressure
P_m	Legendre polynomial of order m
ϕ	elevation angle
r	distance from source to listener
RT_{60}	reverberation time
$R(j\omega)$	reflectance
ρ_0	density of air
S	sound source
t	time variable
T	sampling interval
$T_k(z)$	auralization filter block

T	temperature in Kelvin
T_0	reference air temperature
τ_{high}	high-frequency interaural time delay
τ_{group}	group-delay interaural time delay
τ_{low}	low-frequency interaural time delay
θ	azimuth angle
θ_e	azimuth angular error
ω	angular frequency
$x(n)$	input signal
$y(n)$	output signal
z	z-transform variable
$Z(j\omega)$	normalized impedance

List of Abbreviations

ANCOVA	analysis of covariance
ANOVA	analysis of variance
AR	auto-regressive
ARMA	auto-regressive moving average
BEM	boundary element method
BIFS	binary format for scenes
BRIR	binaural room impulse response
BMT	balanced model truncation
CCRMA	Center for Computer Research in Music and Acoustics
CD	compact disc; committee draft
CF	Caratheodory-Fejer
CT	cross-talk
DF	direct form
DFT	discrete Fourier transform
DIVA	digital interactive virtual acoustics
DRIR	direct room impulse response
DSP	digital signal processing
DTF	directional transfer function
DVD	digital versatile disc
ERB	equivalent rectangular bandwidth
FDTD	finite difference time domain
FEM	finite element method
FFT	fast Fourier transform
FIR	finite impulse response
FM	frequency modulation
GTFB	gammatone filterbank
GUI	graphical user interface
HOA	Hankel norm optimal approximation
HRIR	head-related impulse response
HRTF	head-related transfer function
HSV	Hankel singular values
HUT	Helsinki University of Technology
IACC	interaural cross-correlation
IFFT	inverse fast Fourier transform
IIR	infinite impulse response

ILD	interaural level difference
IPD	interaural phase difference
IRCAM	Institut de Recherche et Coordination Acoustique/Musique
ISO	international standardization organization
ITD	interaural time difference
JND	just noticeable difference
KLE	Karhunen-Loeve expansion
LFE	low frequency energy
LMS	least-mean square
LS	least squares
LTI	linear time-invariant
MA	moving average
MPEG	moving picture experts group
PCA	principal components analysis
PRIR	parametric room impulse response
RIR	room impulse response
SEA	statistical energy analysis
SER	signal-to-error ratio
SPL	sound pressure level
SPSS	Statistical Package for Social Sciences
STFT	short-time Fourier transform
SVD	singular value decomposition
TAFC	two alternatives forced choice
TTS	text-to-speech synthesis
VAD	virtual auditory display
VHT	virtual home theatre
VL	virtual loudspeaker
VRML	virtual reality modeling language
VS	virtual speaker
WFIR	warped finite impulse response
WIIR	warped infinite impulse response
WLS	weighted least squares
WSF	warped shuffler filter

Chapter 1

Introduction

The topic of this thesis is physically-based modeling of room acoustical systems and human spatial hearing by means of digital signal processing (DSP). A concept of *virtual acoustics* is explored, which deals with three major subsystems in acoustical communication as shown in Fig. 1.1: source modeling, transmission medium modeling, and listener modeling (Begault, 1994, p. 6). Underlying DSP concepts for efficient and auditorily relevant multimedia sound processing are studied, and a general framework for virtual acoustics modeling is proposed. The goal in this thesis is to develop technology for delivering an acoustical message in a virtual reality system from the source to the receiver as it would happen in a real-world situation.

1.1 Background

The characteristics of human hearing form the source of information for the research of 3-D sound and virtual acoustics (Blauert, 1997). Within this framework,

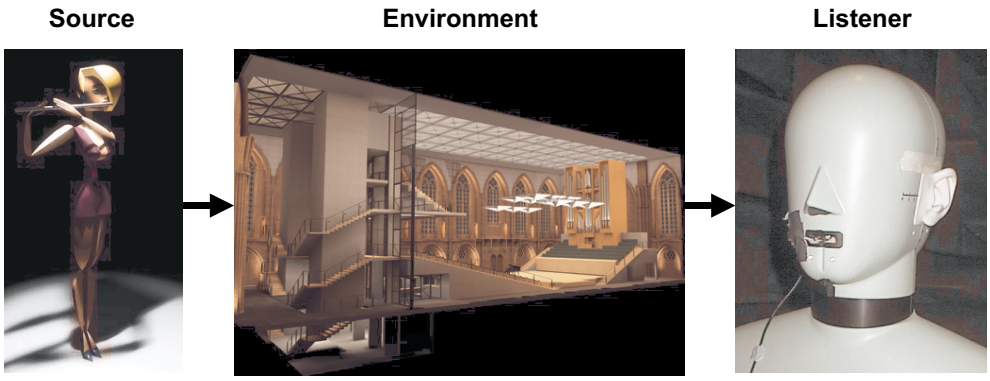


Figure 1.1: The concept of virtual acoustics comprises source modeling, environment modeling, and listener modeling.

Term	Explanation
Monotic or Monaural	stimulus presented to only one ear
Dichotic	stimuli that are different at the two ears
Diotic	stimuli that are identical at the ears
Binaural	any stimulus that is diotic or dichotic

Table 1.1: Terminology of binaural technology (Yost and Gourevitch, 1987, p. 50).

distinctive terms describing methods for spatial localization and processing of sounds have been defined. Fundamental definitions of binaural terminology are presented in Table 1.1 (Yost and Gourevitch, 1987). In this work, the term binaural is constrained to mean only such processing that a three-dimensional auditory illusion is created for headphone listening (Møller, 1992). The term crosstalk canceled binaural processing is used to refer to manipulation of binaural information for listening with two loudspeakers [1]. The term auralization is understood as rendering audible a modeled acoustical space in such a way that the three-dimensional sound illusion is preserved (Kleiner et al., 1993). By virtual acoustics the concept of auralization is expanded to include virtual reality aspects such as dynamic movement, dynamic source and acoustic space characterization, and immersion using advanced control systems (Takala et al., 1996; Savioja et al., 1999). Auditory display is a method or device used to display information to human observers via the auditory system (Shinn-Cunningham et al., 1997, p. 611).

Sound has many roles in a multimedia or virtual reality system. In most cases, sound is used primarily for communicating (informational, navigational speech and non-speech audio) or entertaining (music, voice, ambience) purposes. This thesis deals with the processing and manipulation of any input sound, modifying it so that the outcome resembles a physical or artificial three-dimensional auditory environment. There is a long history of research on spatial hearing and three-dimensional sound (Blauert, 1997). Lord Rayleigh (J.W. Strutt, 3rd Baron of Rayleigh) presented the first influential theory of spatial hearing, called the "duplex theory," in the turn of the 20th century (Rayleigh, 1907). The first virtual acoustic concepts were explored and examined in the field of teleoperation. Spandöck (1934) designed scale models and auralized concert hall designs in the 1930's [2]. Even prior to that, during World War I, enemy aircraft detection was enhanced by imitating spatial hearing cues for accurate three-dimensional source localization using a pseudophone (magnified ear canal and pinnae) [3]. Both of

[1] The term *transaural* that refers to same type of processing is trademarked by Cooper and Bauck (1989). See (Gardner, 1998a) for review of crosstalk canceling technology.

[2] See (Kleiner et al., 1993; Kuttruff, 1993; Begault, 1994; Blauert, 1997) for review of auralization and spatial hearing.

[3] See (Wenzel, 1992; Shinn-Cunningham et al., 1997) for review of auditory displays.

these early applications indeed had the same major goals as virtual acoustics has today: *virtual design* of audiovisual spaces, i.e., exploration of acoustical behavior virtually, and accurate three-dimensional simulation and reproduction of auditory events. In other words, the aim is to augment the hearing sense when using unideal reproduction devices such as headphones or a pair of loudspeakers. With the rapid advance of multimedia technology, the virtual acoustics concept has in recent years expanded to cover even more areas, ranging from physical modeling of sound sources, via room acoustics rendering, to modeling of spatial hearing cues.

The application area of virtual acoustic technology is wide, ranging from entertainment and telecommunications to virtual reality and acoustics simulation. Three-dimensional audio systems are often associated with virtual acoustic displays (VAD). The VADs play an important role in simulations and the virtual reality technology, where source, environment, and listener models are combined to produce an immersive experience (Begault, 1994). Game and multimedia technology are today the leading edge in consumer applications of 3-D sound. Teleconferencing and videoconferencing are another popular application area utilizing three-dimensional sound processing. Virtual environment technology, with standards like Virtual Reality Modeling Language (VRML) (ISO/IEC, 1997) and MPEG-4 (ISO/IEC, 1999), is further expanding the use of virtual acoustics in the context of audiovisual simulation. The growth of positional 3-D sound technology and applications that use binaural headphone or loudspeaker processing has been significant in the last few years, especially with the emergence of multichannel audio formats. Multichannel audio, for example in DVD movies, is often played back in domestic environments using only two channels. Therefore, the concepts of *virtual surround* (or virtual home theater, VHT) for headphone and loudspeaker reproduction have been developed to enhance the three-dimensional listening experience.

From the viewpoint of sound reproduction, the research conducted in this thesis also involves improving the capabilities of traditional two-channel stereophony. Stereophonic system implementation can be divided into three stages: sound pickup (microphone), sound mixing (processing), and sound reproduction (Theile, 1991). This research is focused on the last two stages of stereophonic reproduction. Virtual acoustic technology and binaural processing for headphone and loudspeaker listening has produced a new era of virtual audio reality to two-channel stereophony. With the aid of modern signal processing methods rewiewed and presented in this thesis, it is possible to design and implement three-dimensional virtual auditory environments and audio effects for consumer audiovisual products used in computer, television and hi-fi systems.

1.2 Overview of the Thesis

The topics covered in this thesis can be considered interdisciplinary. They range from psychoacoustics and spatial hearing to digital signal processing aspects and implementation, and from room acoustics and the behavior of musical instruments to real-time system design and specification. The aim of this thesis is not to seamlessly cover all the interesting and unexplored research problems in the cross-section of all the areas in virtual acoustics. Instead, this thesis is written more in the spirit of presenting "selected topics in virtual acoustics," concentrating on the new innovations and contributions by the author related to the field. Some of the underlying theory is discussed in great detail, especially that of spatial hearing and manipulation of head-related transfer functions, to produce a complete view of the problem. Some background topics (for example, geometrical room acoustics modeling and implementation) are, however, only briefly introduced to the reader. Therefore a moderate to good background knowledge of the topic area is expected.

The scope of this work has been twofold. In the first part, topics in source and room acoustics modeling related to real-time processing of virtual acoustic environments are covered (Savioja et al., 1999). New methods are presented for source and environment modeling and filter design. In the second part, head-related transfer function (HRTF) modeling, approximation, and evaluation techniques for headphone and loudspeaker reproduction are presented. The work reviews existing theories and methods for realization of a real-time auralization system, and presents new methodology, filter design and implementation issues of HRTFs and virtual acoustics. Furthermore, novel methods and results of subjective and objective evaluation of 3-D audio systems for both headphone and loudspeaker listening are presented. The author motivates the use of auditory criteria in HRTF-based filter design for binaural and crosstalk canceled binaural processing. The results are verified by carefully conducted subjective listening experiments, and a binaural auditory model is used for predicting further filter design and implementation performance. One result of this thesis has been the author's contribution to the auralization specification and implementation of the DIVA (Digital Interactive Virtual Acoustics) Virtual Audio Reality System (Takala et al., 1996; Savioja et al., 1999) developed at the Helsinki University of Technology. The DIVA system is a real-time virtual concert hall performance environment including MIDI-to-movement mapping, animation, physical modeling, room acoustics modeling, and 3-D sound simulation. Furthermore, the author has actively contributed to the MPEG-4 version 2 standardization in the field of virtual acoustic parametrization (Swaminathan et al., 1999; Väänänen and Huopaniemi, 1999; Scheirer et al., 1999). The guidelines for HRTF filter design provided by the author in this thesis give a scientifically solid framework for optimized reproduction of 3-D sound in a wide range of applications such as virtual reality, teleconferencing, and professional and domestic audio systems.

This thesis is divided as follows. An overview of virtual acoustic technology is given in Chapter 2. In Chapter 3, topics in sound source parametrization and modeling are studied, and new methods for source modeling are presented. Chapter 4 deals with real-time room acoustical issues, concentrating on novel filter design aspects for phenomena such as air absorption and reflections from room boundaries. Chapter 5 is devoted to HRTF approximation and filter design for measurements at both close distance and far distance, and on dummy heads as well as human subjects. Optimization of binaural processing for headphone listening is discussed, and new methods and results for subjective and objective evaluation of binaural filter design are presented. In Chapter 6, binaural techniques are extended for loudspeaker reproduction by crosstalk canceling mechanisms and virtual loudspeaker technology. A new algorithm and filter design for stereo widening and virtual surround processing is presented. Subjective evaluation has been carried out to evaluate the quality of the designed virtual surround filters. Finally, conclusions are drawn and directions for further work are given in Chapter 7.

1.3 Contributions of the Author

The author's contributions to the field of 3-D audio and virtual acoustic technology, presented in this thesis, can be summarized as follows. In virtual acoustic environment related issues (Chapter 2), the author has participated in the DIVA group research and contributed to system architecture definition and both non-real-time and real-time auralization filter design and implementation (Huopaniemi et al., 1994; Takala et al., 1996; Huopaniemi et al., 1996; Savioja et al., 1997, 1999). The work in the DIVA group has been a result of long cooperation with Mr. Lauri Savioja, Ms. Riitta Väänänen, and Mr. Tapio Lokki. The author has also actively contributed to MPEG-4 standardization work and introduced parametrization strategies for physically-based virtual acoustic rendering (Swaminathan et al., 1999; Scheirer et al., 1999; Väänänen and Huopaniemi, 1999).

In source modeling (Chapter 3), the author investigated and was the first to introduce the use of physical musical instrument models as sound sources in a virtual acoustic simulation environment (Huopaniemi et al., 1994). Concepts and modeling strategies for sound source directivity have been presented by the author in (Huopaniemi et al., 1994, 1995; Karjalainen et al., 1995; Välimäki et al., 1996; Huopaniemi et al., 1999a). This work has been a result of fruitful cooperation with Prof. Matti Karjalainen and Dr. Vesa Välimäki. A novel method for measuring and modeling sound radiation from the mouth has been suggested by the author in (Huopaniemi et al., 1999a).

In room acoustics modeling (Chapter 4), the author has contributed to filter design and real-time simulation of boundary reflections and air absorption (Huopaniemi et al., 1997) as well as reverberation algorithm design (Huopaniemi

et al., 1994; Väänänen et al., 1997).

The author's contribution to research in HRTF filter design and binaural systems (Chapters 5 and 6) has resulted in new binaural filter design methods based on frequency warping (Huopaniemi and Karjalainen, 1996a,b, 1997) and balanced model truncation (Mackenzie et al., 1997) (in cooperation with Prof. Karjalainen, Dr. Välimäki, and Dr. Jonathan Mackenzie). The author has also investigated distance-dependency in HRTFs (Huopaniemi and Riederer, 1998). Furthermore, objective and subjective evaluation of auditory binaural audio signal processing has been performed (Huopaniemi and Smith, 1999; Pulkki et al., 1999; Huopaniemi et al., 1999b) and practical implementations (Karjalainen et al., 1999; Zacharov et al., 1999; Zacharov and Huopaniemi, 1999) have been created by the author.

Practical and useful results of the thesis work include methods for real-time virtual acoustics rendering, virtual loudspeaker design, enhanced stereo imaging, and widening of the listening area in two-channel loudspeaker reproduction.

1.4 Related Publications

During the research work, the author has contributed to the following publications, parts of which have been used in this thesis.

Reviewed Journal Articles

Savioja, L., Huopaniemi, J., Lokki, T. and Väänänen, R. 1999. Creating interactive virtual acoustic environments, *Journal of the Audio Engineering Society* **47**(9): 675–705, September 1999.

Scheirer, E., Väänänen, R., Huopaniemi, J. 1999. AudioBIFS: The MPEG-4 standard for audio effects processing, *IEEE Transactions on Multimedia* **1**(3): 237–250, September 1999.

Huopaniemi, J., Zacharov, N., and Karjalainen, M. 1999b. Objective and subjective evaluation of head-related transfer function filter design, *Journal of the Audio Engineering Society* **47**(4): 218–239, April 1999.

Pulkki, V., Karjalainen, M. and Huopaniemi, J. 1999. Analyzing virtual sound sources using a binaural auditory model, *Journal of the Audio Engineering Society* **47**(4): 203–217, April 1999.

Mackenzie, J., Huopaniemi, J., Välimäki, V. and Kale, I. 1997. Low-order modelling of head-related transfer functions using balanced model truncation, *IEEE Signal Processing Letters* **4**(2): 39–41.

Välimäki, V., Huopaniemi, J., Karjalainen, M. and Jánosy, Z. 1996. Physical modeling of plucked string instruments with application to real-time sound synthesis, *Journal of the Audio Engineering Society* **44**(5): 331–353.

International Conference Articles

Zacharov, N., and Huopaniemi, J. 1999. Results of a round robin subjective evaluation of virtual home theatre sound systems, *Presented at the 107th Convention of the Audio Engineering Society*, preprint 5760, New York.

Huopaniemi, J., Kettunen, K. and Rahkonen, J. 1999. Measurement and modeling techniques for directional sound radiation from the mouth, *Proceedings of the 1999 IEEE Workshop of Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, New York, pp. 183–186.

Huopaniemi, J. and Smith, J. O. 1999. Spectral and time-domain preprocessing and the choice of modeling error criteria in binaural digital filter design, *Proceedings of the 16th Audio Engineering Society International Conference*, Rovaniemi, Finland, April 10-12, 1999, pp. 301–312.

Zacharov, N., Hämäläinen, M. and Huopaniemi, J. 1999. Round robin subjective evaluation of virtual home theatre systems at the AES 16th international conference, *Proceedings of the 16th Audio Engineering Society International Conference*, Rovaniemi, Finland, April 10-12, 1999, pp. 544–556.

Väänänen, R. and Huopaniemi, J. 1999. Spatial presentation of sound in scene description languages, *Presented at the 106th Convention of the Audio Engineering Society*, preprint 4921, Munich, Germany.

Huopaniemi, J., Zacharov, N. and Karjalainen, M. 1998. Objective and subjective evaluation of head-related transfer function filter design, *Presented at the 105th Convention of the Audio Engineering Society*, preprint 4805, San Francisco, CA, USA (invited paper).

Huopaniemi, J. and Riederer, K. 1998. Measuring and modeling the effect of distance in head-related transfer functions, *Proceedings of the joint meeting of the Acoustical Society of America and the International Congress on Acoustics*, Washington, USA, pp. 2083–2084.

Huopaniemi, J., Savioja, L. and Karjalainen, M. 1997. Modeling of reflections and air absorption in acoustical spaces: a digital filter design approach, *Proceedings of the 1997 IEEE Workshop of Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, New York.

Karjalainen, M., Härmä, A., Laine, U. and Huopaniemi, J. 1997. Warped filters and their audio applications, *Proceedings of the 1997 IEEE Workshop of Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, New York.

Väänänen, R., Välimäki, V., Huopaniemi, J. and Karjalainen, M. 1997. Efficient and parametric reverberator for room acoustics modeling, *Proceedings of the International Computer Music Conference*, Thessaloniki, Greece, pp. 200–203.

Hiipakka, D. G. J., Hänninen, R., Ilmonen, T., Napari, H., Lokki, T., Savioja, L., Huopaniemi, J., Karjalainen, M., Tolonen, T., Välimäki, V., Välimäki, S. and Takala, T. 1997. Virtual Orchestra Performance, *Visual Proceedings of SIGGRAPH'97*, ACM SIGGRAPH, Los Angeles, p. 81.

Huopaniemi, J. and Karjalainen, M. 1997. Review of digital filter design and implementation methods for 3-D sound, *Presented at the 102nd Convention of the Audio Engineering Society*, preprint 4461, Munich, Germany.

Huopaniemi, J., Savioja, L. and Takala, T. 1996. DIVA virtual audio reality system, *Proceedings of the International Conference on Auditory Display*, Palo Alto, California, USA.

Takala, T., Hänninen, R., Välimäki, V., Savioja, L., Huopaniemi, J., Huotilainen, T. and Karjalainen, M. 1996. An integrated system for virtual audio reality, *Presented at the 100th Convention of the Audio Engineering Society*, preprint 4229, Copenhagen, Denmark.

Huopaniemi, J. and Karjalainen, M. 1996a. Comparison of digital filter design methods for 3-D sound, *Proceedings of the IEEE Nordic Signal Processing Symposium*, Espoo, Finland, pp. 131–134.

Huopaniemi, J. and Karjalainen, M. 1996b. HRTF filter design based on auditory criteria, *Proceedings of the Nordic Acoustical Meeting (NAM'96)*, Helsinki, Finland, pp. 323–330.

Karjalainen, M., Huopaniemi, J. and Välimäki, V. 1995. Direction-dependent physical modeling of musical instruments, *Proceedings of the International Congress on Acoustics*, Vol. 3, Trondheim, Norway, pp. 451–454.

Huopaniemi, J., Karjalainen, M. and Välimäki, V. 1995. Physical models of musical instruments in real-time binaural room simulation, *Proceedings of the International Congress on Acoustics*, Trondheim, Norway, pp. 447–450.

Huopaniemi, J., Karjalainen, M., Välimäki, V. and Huotilainen, T. 1994. Virtual instruments in virtual rooms – a real-time binaural room simulation environment for physical modeling of musical instruments, *Proceedings of the International Computer Music Conference*, Aarhus, Denmark, pp. 455–462.

Parts of this present work have been published in the same or modified form in:

Huopaniemi, J. 1997. *Modeling of Human Spatial Hearing in the Context of Digital Audio and Virtual Environments*, Helsinki University of Technology, Department of Electrical and Communications Engineering, Laboratory of Acoustics and Audio Signal Processing, p. 101. Lic.Tech. Thesis.

Chapter 2

Virtual Acoustics

Virtual acoustics is a general term for the modeling of acoustical phenomena and systems with the aid of a computer. It may comprise many different fields of acoustics, but in the context of this work the term is restricted to describe a system ranging from sound source and acoustics modeling in rooms to spatial auditory perception simulation in humans.

The virtual acoustic concept is closely related to other media for many reasons. The audiovisual technology (audio, still picture, video, animation etc.) is rapidly integrating into a single interactive media. Research on audiovisual media modeling has increased dramatically in the last decade. Standardization of advanced multimedia and virtual reality rendering and definition has been carried out for several years. Such examples are the Moving Pictures Expert Group MPEG-4 (ISO/IEC, 1999) and the Virtual Reality Modeling Language (VRML) (ISO/IEC, 1997) standards. The progress of multimedia has also introduced new fields of research and application into audio and acoustics, one of which is virtual acoustic environments[1] and their relation to graphical presentations.

In this chapter, an overview of virtual acoustics modeling concepts is given. The system that has been designed in the course of this study is outlined.

2.1 Virtual Acoustic Modeling Concepts

The design and implementation of virtual acoustic displays (VADs) can be divided into three different stages as depicted in Fig. 2.1 (Huopaniemi, 1997; Savioja et al., 1999): a) definition, b) modeling, and c) reproduction. The definition of a virtual acoustic environment includes the prior knowledge of the system to be implemented, that is, information about the sound sources, the room geometry, and the listeners. The definition part is normally carried out off-line prior to the

[1] Virtual acoustic environments are also called virtual auditory displays, auditory virtual environments, virtual auditory space (Begault, 1994; Shinn-Cunningham et al., 1997).

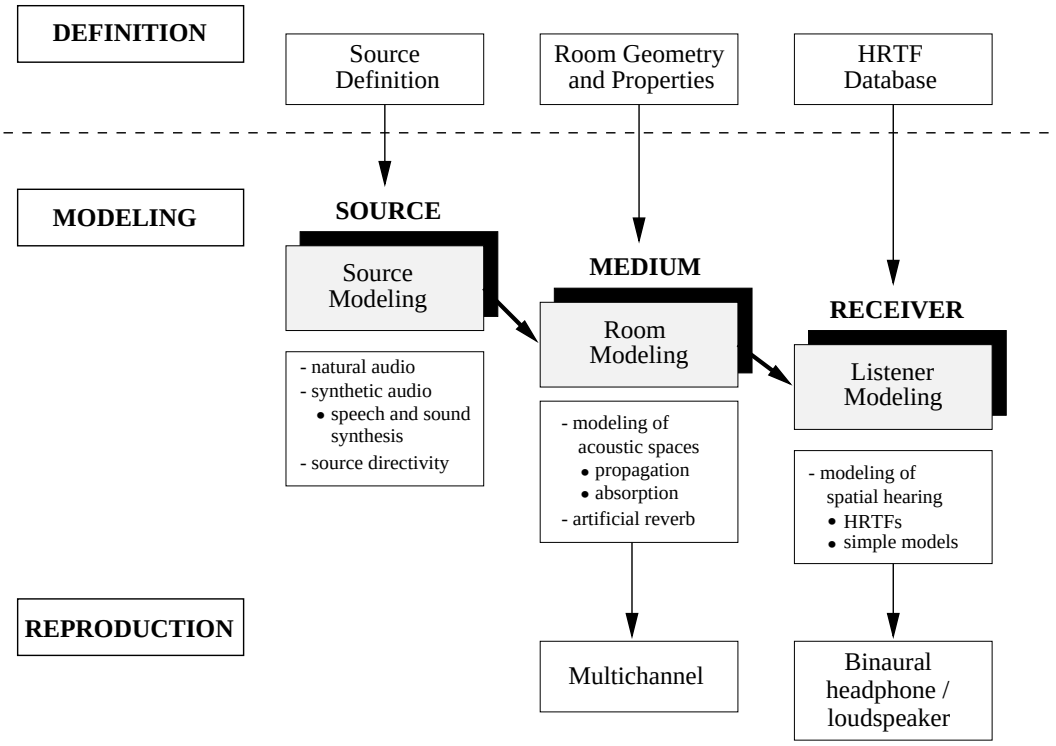


Figure 2.1: Modeling concept of virtual acoustics (Huopaniemi, 1997; Savioja et al., 1999).

real-time simulation process. In the modeling part the rendering is divided into three tasks (Wenzel, 1992; Begault, 1994):

- Source Modeling
- Transmission Medium (Room Acoustics) Modeling
- Receiver (Listener) Modeling

This division is also often called the *source-medium-receiver concept* (Wenzel, 1992; Begault, 1994), which is widely known in many communication systems[2]. Interactive virtual acoustic environments and binaural room simulation systems have been studied by, e.g., (Foster et al., 1991; Wenzel, 1992; Begault, 1994; Shinn-Cunningham et al., 1997; Jot, 1999; Lehnert and Blauert, 1992; Vian and Martin, 1992; Martin et al., 1993), but until recently the physical relevance and the relation of acoustical and visual content has not been of major interest. The term *auralization* (Kleiner et al., 1993) is understood as a subset of virtual acoustics referring to modeling and reproduction of sound fields (shown in Fig. 2.1).

[2] In communication technology, the concept is often called source-channel-receiver.

Source modeling (Fig. 2.1) consists of methods which produce (and possibly add physical character to) sound in an audiovisual scene. The most straightforward method is to use pre-recorded digital audio. In most auralization systems sound sources are treated as omnidirectional point sources. This approximation is valid for many cases, but more accurate methods are, however, often needed. For example, most musical instruments have radiation patterns which are frequency-dependent. Typically sound sources radiate more energy to the frontal hemisphere whereas sound radiation is attenuated and lowpass filtered when the angular distance from the on-axis direction increases. In Chapter 3, methods for efficient modeling of sound source directivity are presented.

The task of modeling sound propagation behavior in acoustical spaces (Fig. 2.1) is an integral part of virtual acoustic simulation. The goal in most room acoustic simulations has been to compute an energy time curve (ETC) of a room (squared room impulse response) and, based on that, to derive room acoustical attributes such as reverberation time (RT_{60}). The ray-based methods, the ray tracing (Krokstad et al., 1968; Kulowski, 1985) and the image-source methods (Allen and Berkley, 1979; Borish, 1984), are the most often used modeling techniques. Recently computationally more demanding wave-based techniques like finite element method (FEM), boundary element method (BEM), and finite-difference time-domain (FDTD) methods have also gained interest (Kuttruff, 1995; Botteldooren, 1995; Savioja et al., 1995; Savioja, 1999). These techniques are, however, suitable only for low-frequency simulation (Kleiner et al., 1993). In real-time auralization, the limited computational capacity calls for simplifications, modeling only the direct sound and early reflections using geometrical acoustics and rendering the late reverberation by recursive digital filter structures (Schroeder, 1962; Moorer, 1979; Jot, 1999; Gardner, 1998b; Savioja et al., 1999).

In listener modeling (Fig. 2.1), the properties of human spatial hearing are considered. Simple means for giving a directional sensation of sound are the interaural level and time differences (ILD and ITD), but they cannot resolve the front-back confusion[3]. The head-related transfer function (HRTF) that models the reflections and filtering by the head, shoulders and pinnae of the listener, has been studied extensively during the past decade. With the development of HRTF measurement techniques (Wightman and Kistler, 1989; Møller et al., 1995) and efficient filter design methods (Kendall and Rodgers, 1982; Huopaniemi et al., 1999b) real-time HRTF-based 3-D sound implementations have become applicable in virtual environments.

[3] See (Begault, 1994; Blauert, 1997; Møller, 1992; Kleiner et al., 1993; Kendall, 1995) for fundamentals on spatial hearing and auralization.

2.2 Methods for 3-D Sound Reproduction

The illusion of three-dimensional sound fields can be created using various methods. The goal in 3-D sound simulation is to recreate any static or dynamic natural or imaginary sound field using a desired amount of transducers and proper signal processing techniques. In this work, three methods have been considered:

- binaural reproduction ,
- crosstalk canceled binaural reproduction ,
- multichannel reproduction .

In order to more thoroughly understand the different aspects involved in 3-D sound processing, the two binaural choices for implementation are first considered, namely binaural processing for headphone and loudspeaker listening. Figure 2.2 illustrates the differences between these two processing methods.

2.2.1 Headphone Reproduction

In the case of binaural filter design, the HRTF measurement may directly be approximated by various filter design methods (see Chapter 5) provided that proper equalization is carried out. In the example of Fig. 2.2a, a monophonic time-domain signal $x_m(n)$ is filtered with two HRTF filter approximations $H_l(z)$ and $H_r(z)$ to create an image of a single virtual source. Known advantages of binaural reproduction are the (trivial) facts that the acoustics of the listening room and the positioning of the listener do not affect the perception. On the other hand, head-tracking is required to compensate for head movements. Furthermore, it has been found that individual HRTFs should be used to create the best performance in localization and externalization (Wenzel et al., 1993; Møller et al., 1996) and that care must be taken in the equalization and placing of headphones in order to obtain an immersive 3-D soundscape.

2.2.2 Loudspeaker Reproduction

Loudspeaker reproduction of binaural information is critically different from headphone reproduction, as can be seen in Fig. 2.2. The directional characteristics that are present in binaural signals are exposed to crosstalk when loudspeaker reproduction is used. Crosstalk occurs because the sound from the left loudspeaker is heard at both left and right ears, and vice versa. The theory of crosstalk canceling for binaural reproduction was presented in the 1960's by Bauer (1961) and practically implemented by Schroeder and Atal (1963), and has been later investigated intensively by the research community and industry. A taxonomy of crosstalk canceling technologies is given in (Gardner, 1998a).

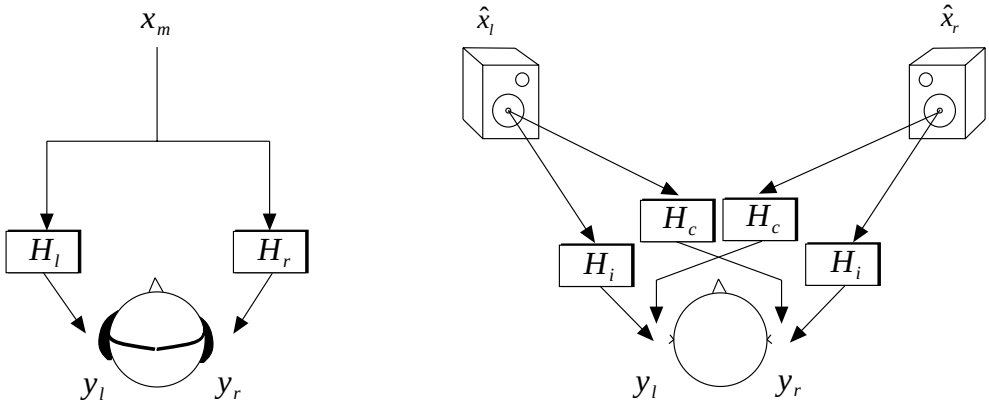


Figure 2.2: Signal flowgraph of a) binaural and b) crosstalk canceled binaural audio processing.

In binaural synthesis for loudspeaker listening (see 2.2b), the crosstalk canceled binaural signals $\hat{x}_l(n)$ and $\hat{x}_r(n)$ are applied to the speakers to yield the desired outputs $y_l(n)$ and $y_r(n)$. The direction-dependent loudspeaker-to-ear transfer functions $H_i(z)$ and $H_c(z)$ (we consider here only the symmetrical listening position[4]) have to be taken into account in order to obtain a similar effect as in headphone listening. The filtering can here be understood as a cascaded process, in which HRTF filters are designed and implemented separately from the crosstalk canceling filters. Another alternative is to combine these processes and design virtual speaker filters by using, e.g., shuffler structures (Cooper and Bauck, 1989). Shuffler structures and other binaural crosstalk canceling systems will be studied in more detail in Chapter 6. Crosstalk canceled binaural systems have two main limitations: 1) critical listening position, and 2) critical listening room conditions. The full spatial information can be retained only in anechoic chambers or listening rooms and the "sweet spot" for listening is very limited. On the other hand, at its best crosstalk canceled binaural reproduction is capable of auralizing a sound field in a very convincing and authentic way.

2.2.3 Multichannel Techniques

A natural choice to create a two- or three-dimensional auditory space is to use multiple loudspeakers in the reproduction. With this concept the problems in retaining spatial auditory information are reduced to placement of the N loudspeakers and panning of the audio signals according to the direction (Gerzon, 1992). The problem of multiple loudspeaker reproduction and implementation of panning rules can be formulated in a form of vector base amplitude panning (VBAP) (Pulkki, 1997) or by using decoded 3-D periphonic systems such as

[4] The subscript i means ipsilateral (same side), and c means contralateral (opposite side).

Ambisonics (Gerzon, 1973; Malham and Myatt, 1995). The VBAP concept introduced by Pulkki (1997) gives the possibility of using an arbitrary loudspeaker placement for three-dimensional amplitude panning. The rapid progress of multichannel surround sound systems for home and theater entertainment during the past decade has opened wide possibilities also for multiloudspeaker auralization. Digital multichannel audio systems for movie theaters and home entertainment offer three-dimensional spatial sound that has been either decoded from two-channel material (such as Dolby ProLogic) or uses discrete multichannel decoding (such as Dolby Digital). The ISO/MPEG-2 AAC audio coding standard offers 5.1 discrete transparent quality channels at a rate of 320 kb/s (compression rate 1:12) (Brandenburg and Bosi, 1997). Another multichannel compression technique that is already widespread in the market is Dolby Digital which provides similar compression rates to MPEG-2 AAC (Brandenburg and Bosi, 1997).

A general problem with multichannel ($N > 2$) sound reproduction especially in domestic setups is the amount of needed hardware (6 speakers for a discrete 5.1 surround setup) and their placing. On the other hand, intuitively, the listening area should be larger and the localization effect more stable than with two-channel binaural loudspeaker systems. Another advantage of multichannel reproduction over binaural systems is that no listener modeling (use of HRTFs) is necessary, and thus it is in general computationally less demanding.

2.3 Implementation of Virtual Acoustic Systems

The previous section dealt with rendering free-field three-dimensional auditory cues using different reproduction systems. However, when an interactive time-varying virtual acoustic processing scheme (see Fig. 2.1) is considered, the reproduction models interact and are preceded by different modeling techniques for source and room acoustics. Implementation issues of virtual acoustic systems have been studied by the DIVA group (Takala et al., 1996; Savioja et al., 1999), including major contributions from the author of this work. In this section, an outline for a geometrical acoustics-based modeling approach is given.

Real-time rendering of virtual acoustics poses several implementational and computational challenges. The most crucial ones in terms of perception and immersion are dynamic (time-varying) room impulse processing and interactivity of the user with the virtual acoustic model. To gain understanding of room acoustics modeling techniques, Fig. 2.3 shows different methods used in many current systems (Savioja et al., 1999).

Kleiner et al. (1993) have divided the modeling techniques for room acoustics between computational simulation and scale modeling. Scale models are widely used by architects and acousticians in the design of concert halls. However, computers have been used for thirty years to model room acoustics (Krokstad et al., 1968; Schroeder, 1970, 1973) and computational modeling has become a

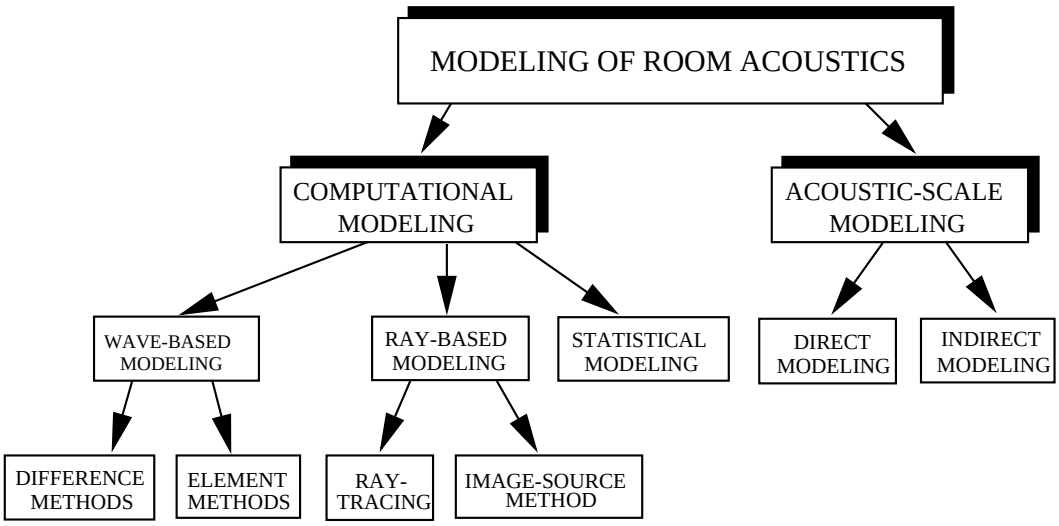


Figure 2.3: Different methods for room acoustics modeling and auralization (Savioja et al., 1999).

widely accepted tool in room acoustic design (Kuttruff, 1995).

There are three different approaches in computational modeling of room acoustics as illustrated in Fig. 2.3. In ray-based methods the effect of wavelength is neglected and sound is supposed to act like rays (Kuttruff, 1991). All phenomena due to the wave nature, such as diffraction and interference, are excluded from the model. This assumption is valid when the wavelength of the sound is small compared to the dimensions of surfaces and obstacles in the room and large compared to their roughness. Ray tracing and image-source methods are the most often used algorithms in ray-based modeling of room acoustics. Enhancements to geometrical acoustics modeling principles that take into account diffusion (Dalenbäck, 1995; Naylor and Rindel, 1994) and edge diffraction (Svensson et al., 1997) have been proposed in recent years.

More accurate results can in principle be achieved with wave-based methods which are based on solving the wave equation. An analytical solution for the wave equation can be found only in rare cases such as rectangular rooms with rigid walls. Therefore, numerical wave-based methods must be applied. These methods are, however, computationally very expensive. Element methods, such as the finite element method (FEM) and boundary element method (BEM), are widely used in vibration mechanics, and, despite the heavy computational requirements, also in room acoustics. As an alternative to element methods, finite-difference time-domain (FDTD) simulation has been recently found to be attractive for room acoustics simulation (Botteldooren, 1995; Savioja et al., 1995).

In addition there are also statistical modeling methods, such as statistical energy analysis (SEA) (Lyon and DeJong, 1995). Those methods are mainly used in prediction of noise levels in coupled systems where sound transmission through

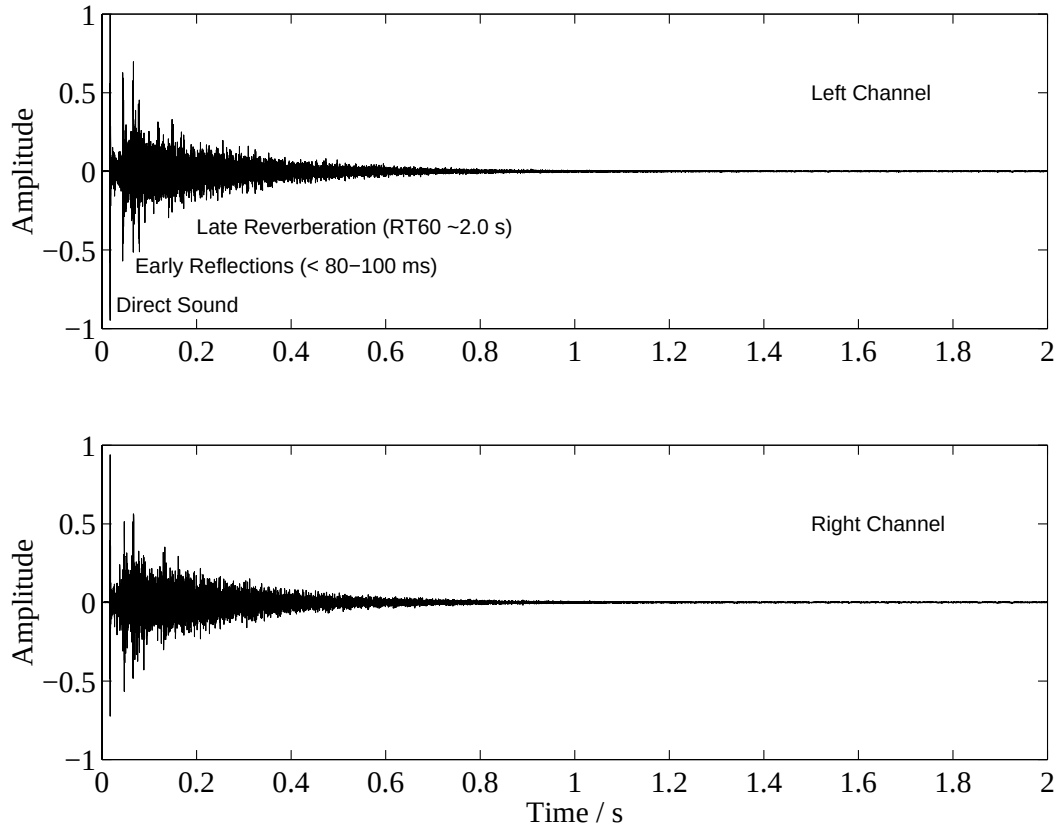


Figure 2.4: A binaural room impulse response measured in a concert hall. The room impulse response can be divided into the direct sound, early reflections, and late reverberation.

structures is an important factor, but they are not suitable for auralization purposes.

2.3.1 Real-Time Geometrical Room Acoustics Approach

Although the geometrical acoustics modeling approach has known limitations, it is practically the only relevant technique suitable for real-time or near-real-time rendering. The output from all virtual acoustic simulators is essentially a *binaural room impulse response* (BRIR), an example of which is shown in Fig. 2.4. This is a two-channel response dependent on the positions and orientations of the source and the listener and on the properties of the room. An interactive auralization model should therefore produce output which depends on the dynamic properties of the source, receiver, and environment. In principle, there are two different approaches to achieve this goal. The methods presented in the following are called *direct room impulse response rendering* and *parametric room impulse response rendering*.

Direct Room Impulse Response Rendering

The direct room impulse response (DRIR) rendering technique is based on a priori stored BRIRs, obtained either from simulations or from measurements[5]. The BRIRs are defined at a grid of listening points. The auralized sound is produced by convolving the excitation signal with BRIRs of both ears. In interactive movements this convolution kernel is formed by interpolating BRIRs of neighboring listening points. The spatial (positional) accuracy of the simulation is dependent on the grid density of BRIR approximation points. The direct convolution can be carried out in the time or frequency domain or by using hybrid methods (Gardner and Martin, 1994; Jot et al., 1995; McGrath, 1996). Nevertheless, this technique is computationally expensive and has high memory requirements for BRIR storage. A further setback of this method is the fact that the source, room, and receiver parameters can not be extracted from the measured BRIR. This means that changes in any of these features will require a new set of BRIR measurements or simulations.

Parametric Room Impulse Response Rendering

A more robust way for interactive auralization is to use a parametric room impulse response (PRIR) rendering method (Savioja et al., 1999). In this technique the BRIRs in different positions in the room are not predetermined. The responses are formed in real time during interactive simulation. The actual rendering process is carried out in several parts decomposed in the time domain. The first part consists of direct sound and early reflections modeling, both of which are time- and place-variant. The latter part of rendering represents the diffuse reverberant field, which can be treated as a time-invariant digital filter. In practice this means that the late reverberator can be predetermined, but the direct sound and early reflections are auralized according to parameters obtained using a real-time room acoustic model. The parameters for the artificial late reverberation algorithm are derived from measurements or from room acoustic simulations. This modeling framework is illustrated in Fig. 2.5 (Savioja et al., 1999). An obvious drawback of this real-time image-source based model is, however, that it lacks diffusion and diffraction models. This fact is recognized and can sometimes degrade the physical accuracy of the simulation significantly.

The PRIR rendering technique has been further divided into two categories[6] (Väänänen and Huopaniemi, 1999; Jot, 1999):

- Physical modeling approach
- Perceptual modeling approach

[5] Dummy head recordings which were the early form of binaural reproduction would also fall into this category.

[6] Jean-Marc Jot. Personal communication. CCRMA, Stanford University, Aug. 6, 1998.

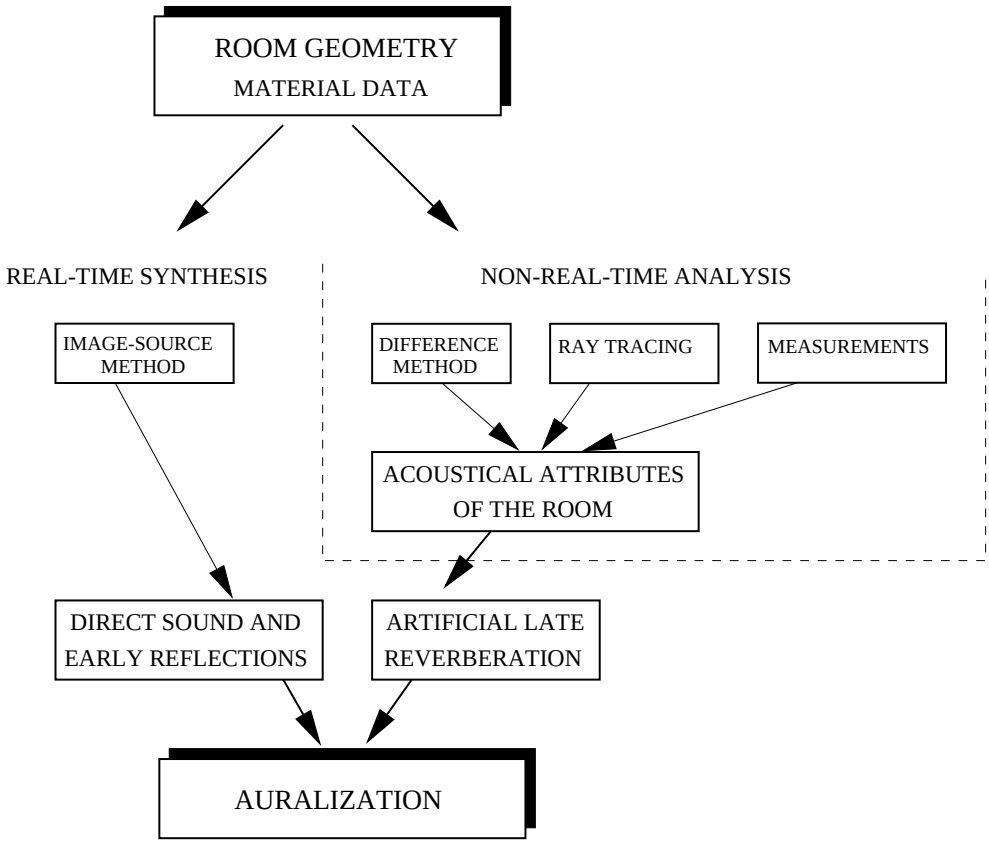


Figure 2.5: Computational room acoustic rendering methods used in the DIVA system. The model is a combination of a real-time image-source method and artificial reverberation which is parametrized according to room acoustical parameters obtained by simulation or measurements (Savioja et al., 1999).

The physical modeling approach, which is the main topic area of this work, aims at capturing both the macroscopic (reverberation time, room volume, absorption area, etc.) and microscopic (reflections, material absorption, air absorption, directivity, diffraction, etc.) room acoustical features by a geometrical acoustics approach, and diffuse late reverberation using statistical means. The goal in perceptual parametrization, on the other hand, is to find an orthogonal set of features using which a normative virtual acoustic rendering algorithm can be controlled to produce a desired auditory sensation. Therefore the output is user-controllable, not environment-controllable as in the physical approach. The perceptual approach uses physical properties for the direct sound, macroscopic room acoustical features and a static image-source method for early and late early reflections, and a statistical late reverberation module for late reverberation (Jot et al., 1995; Jot and Chaigne, 1996; Jot et al., 1998a; Jot, 1999).

This division clearly separates the two main application goals for parametric

room impulse rendering techniques. The first approach is suitable for accurate physical simulation of spaces such as concert halls, auditoria, and for auditorily accurate audiovisual rendering. The latter approach is computationally more efficient, has more intuitive perceptually based control parameters, and is thus more suitable to virtual reality applications and game technology.

The physical PRIR rendering method for binaural reproduction (headphones or a pair of loudspeakers), which is the main area of interest in this work, will now be investigated more thoroughly. This technique has been used in the DIVA system (Takala et al., 1996; Savioja et al., 1999)[7]. In Fig. 2.6, this modeling framework is shown, which is based on the well-known and established image-source method. If we denote the monophonic input signal by $x_m(n)$ and the binaural output from the direct sound and early reflections module as $\mathrm{y}(n) = [y_l(n)\ y_r(n)]^T$, the system transfer function $\mathrm{H}(z) = \mathrm{Y}(z)/X(z)$, where $X(z)$ and $\mathrm{Y}(z$ are z-transforms of $x_m(n)$ and $\mathrm{y}(n)$, can be written as:

$$\mathrm{H}(z) = \sum_{k=0}^{N} T_k(z)\mathrm{F}_k(z)z^{-d_k}, \tag{2.1}$$

where $T_k(z)$ are cascaded filter structures comprising source directivity (studied in Chapter 3), material reflection, and air absorption filtering (studied in Chapter 4), $\mathrm{F}(_k z) = [F_{lk}(z)F_{rk}(z)]^T$ are the binaural filters for the direct sound and early reflections (studied in Chapter 5), and the delay term z^{-d_k} is the travel time delay of the direct sound or the corresponding early reflection. The direct sound is fed (and possibly pre-filtered (Jot et al., 1995)) to the late reverberation module $R(z)$, which produces diffuse late reverberant energy according to the given parameter specifications (see, e.g., (Jot et al., 1995; Väänänen et al., 1997; Savioja et al., 1999) for more details). Furthermore, if listening is carried out using a pair of loudspeakers, an additional crosstalk canceling filter network $C(z)$ is required. Methods and structures for crosstalk canceling are presented in Chapter 6.

2.4 DIVA System

In this section, a short overview of the DIVA system is given. The goal in the research of the DIVA group has been to make a real-time environment for immersive audiovisual processing. The system should integrate the audio signal processing chain from sound synthesis through room acoustics simulation to spatialized reproduction. This is combined with synchronized animated motion. A practical application of this project has been a virtual concert performance (Hiipakka et al.,

[7] Similar DSP methods have been proposed and used for real-time auralization by many researchers. See, e.g., (Foster et al., 1991; Begault, 1994; Jot et al., 1995; Shinn-Cunningham et al., 1997).

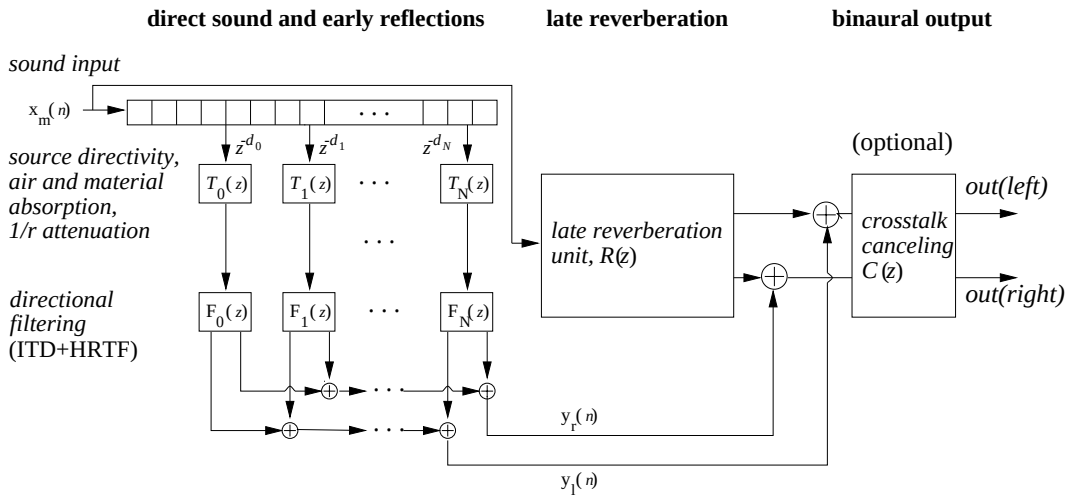


Figure 2.6: Implementation of virtual acoustics using geometrical acoustics modeling. The direct sound, early reflections and late reverberation are processed separately, enabling a fully parametric implementation.

1997; Lokki et al., 1999) and virtual acoustic (and visual) simulations of concert halls and other acoustical spaces (Savioja et al., 1999).

In Fig. 2.7 the architecture of the DIVA virtual concert performance system is presented. There may be two simultaneous users in the system, a conductor and a listener, who can both interact with the system. The conductor wears a tail coat with magnetic sensors for tracking. With hand movements the conductor controls the orchestra, which may contain both real and virtual musicians. The aim of the conductor gesture analysis is to synchronize electronic music with human movements. In the graphical user interface (GUI) of the DIVA system, animated human models are placed on stage to play music from MIDI files. The virtual musicians play their instruments at tempo and loudness shown by the conductor. Instrument fingerings are found from predefined grip tables. All the joint motions of human hand are found by inverse kinematics calculations, and they are synchronized to perform exactly each note on an animated instrument model. A complete discussion of computer animation in the DIVA system is presented in (Hänninen, 1999).

At the same time, a listener may freely fly around in the concert hall. The GUI sends the listener position data to the auralization unit which renders the sound material provided by physical models and a MIDI synthesizer. The auralized output is reproduced either through headphones or loudspeakers.

In order to perform real-time auralization according to Fig. 2.6, the following information is needed before the startup of the processing (Savioja et al., 1999):

- Geometry of the room
- Materials of the room surfaces

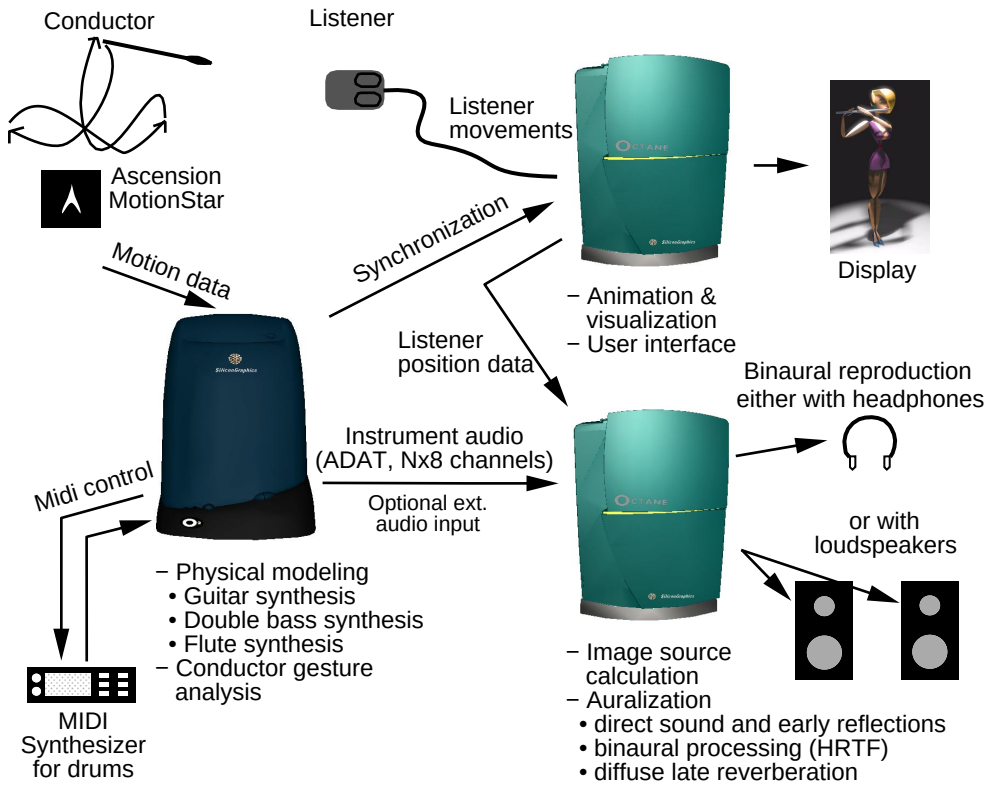


Figure 2.7: The DIVA virtual concert performance system architecture (Savioja et al., 1999).

- Location and orientation of each sound source
- Location and orientation of the listener

Using this information, the early reflections can be dynamically computed using the image-source method as shown in Fig. 2.6. The parameters concerning each visible image source that are sent to the auralization module consist of (Takala et al., 1996; Savioja et al., 1999):

- Distance from the listener
- Azimuth and elevation angles with respect to the listener
- Source orientation with respect to the listener
- Set of filter coefficients which describe the material properties in reflections

This parametrization used in the DIVA system has also formed the basis for the author's contributions to MPEG-4 standardization of three-dimensional audio in virtual environments. The concepts of scene description in multimedia standards will be briefly discussed in the next section.

2.5 Virtual Acoustics in Object-Based Multimedia

One prominent future application area of virtual acoustics is virtual three-dimensional environments. Currently, there are several specifications that can be used to create and render audiovisual scenes in multimedia applications that require efficient and parametric coding of objects. Such applications are, for example, those used across a computer network or other bandwidth restricted delivery channels, and that need to have interaction and real-time rendering capability. New multimedia standards such as MPEG-4 (ISO/IEC, 1999) and VRML97 (ISO/IEC, 1997) (also commonly known as VRML2.0) specify hierarchical scene description languages that can be used to build up 3-D audiovisual scenes where the user can interact with objects. Both VRML and MPEG-4 scene description rely on a scene graph paradigm to describe the organization of audiovisual material. A scene graph represents content as a set of hierarchically related nodes. Each node in the scene graph represents an object (visual objects such as a cube or image, or a sound object), a property of an object (such as the textural appearance of a face of a cube, or the acoustical properties of the material), or a transformation of a part of the scene (such as a rotation or scaling operation). By connecting multiple nodes together, object-based hierarchies are formed.

The VRML97 standard allows the inclusion of short audio clips to the scenes in a specified spatial location to improve the immersiveness and interaction between the user and the virtual environment audio. The MPEG-4 audio standard provides a powerful toolset for both natural and synthetic audio coding and for audio post-processing and three-dimensional rendering. A binary format scene description language (binary format for scenes, or BIFS) defined in the systems level of the MPEG-4 standard enables simultaneous real-time decoding of video and audio objects and presentation of these elementary stream objects based on the scene description to the end user. This requires more efficiency from the scene decoding, but allows inclusion of more natural audiovisual objects where the sound is streamed from the server and can, for example, be real-time encoded audio, or synthetic speech or audio presented in a parametric form.

The MPEG-4 version 2 standard will provide parameters for virtual acoustic rendering that enable room-dependent effects (reverberation time, material reflections, air absorption, Doppler effect) and source-dependent features (directivity, attenuation) (Scheirer et al., 1999; Väänänen and Huopaniemi, 1999). In other words, the desire is to visualize and auralize an MPEG-4 scene with the given parameters. Listener modeling (HRTF reproduction, crosstalk canceling, multichannel panning), however, will not be defined (and therefore normative) in the standard. Spatial presentation of sound in scene description languages has been studied in greater detail in (Väänänen and Huopaniemi, 1999; Scheirer et al., 1999). The author of the present work has been a major contributor to

virtual acoustic parametrization of MPEG-4 version 2. However, these issues will not be dealt with further in this thesis.

2.6 Conclusions

In this chapter, a framework for virtual acoustic modeling was outlined. The virtual acoustic concept covers source, medium, and listener modeling. The geometrical room acoustics modeling paradigm, although not physically correct especially at lower frequencies, allows for using a parametric room impulse rendering technique. This time-domain hybrid impulse response reconstruction divided the modeling into three parts: direct sound, early reflections, and late reverberation rendering. A structure for dynamic interactive simulation was presented, which is capable of high-quality virtual acoustic rendering.

In the next chapter, topics in real-time modeling of sound sources in virtual acoustic systems are covered.

Chapter 3

Sound Source Modeling

In this chapter, topics in modeling of sound sources in relation to real-time rendering are studied. The role of sound sources in virtual acoustic environments is discussed, and efficient sound source directivity rendering techniques are presented. A novel method for modeling sound radiation from the mouth using the reciprocity method is introduced.

Sound source modeling in the virtual acoustic concept refers to attaching sound to an environment and giving it properties such as directivity. The simplest approach has traditionally been to use an anechoic audio recording or a synthetically produced sound as an input to the auralization system (Kleiner et al., 1993). In an immersive virtual reality system, however, a more general and thorough modeling approach is desired that takes into account directional characteristics of sources in an effective way. General qualitative requirements for audio source signals in a virtual acoustic system (used for physical modeling) are:

- Each sound source signal should be "dry," not containing any reverberant or directional properties, unless it is explicitly desired in the simulation (e.g., simulating stereophonic listening in a listening room).
- The audio source inputs are normally treated as point sources in the room acoustical calculation models (image-source method, ray-tracing method, discussed in Chapter 2). This requires the elementary source signals to be monophonic. A stereophonic signal emanating from loudspeakers can thus be modeled as two point sources.
- The quality (signal-to-noise ratio, quantization, sampling rate) of the sound source signal should be adequately high so as not to cause undesired effects in auralization.

From the coding and reproduction point of view, audio source signals may be divided into two categories:

- Natural Audio

- Synthetic Audio

The concept of natural audio refers to sound signals that have been coded from an existing waveform. This category includes all forms of digital audio, raw or bitrate-reduced, and thus the basic requirements for usage in virtual reality sound are those stated in the previous section. By using various bitrate reduction techniques, relatively efficient data transfer rates for high-quality digital audio may be achieved (Brandenburg and Bosi, 1997).

Purely synthetic audio may be explained as sound signal definition and creation without the aid of an a priori coded sound waveform. Exceptions are sound synthesis techniques that involve wavetable data and sampling, which both use a coded waveform of some kind. These methods can be regarded as hybrid natural/synthetic sound generation methods. On the other hand, many sound synthesis techniques such as FM (frequency modulation), granular synthesis, additive synthesis and physical modeling are purely parametric and synthetic[1].

Text-to-speech synthesis (TTS) techniques use a set of parameters that describe the input (often pure text) as phonemes and as prosodic information. This also allows for very low bitrate transmission, because the speech output is rendered in the end-user terminal. Furthermore, it is possible to attach face animation parameters to TTS systems to allow for creation of "talking heads" in advanced virtual reality modeling systems such as MPEG-4 (ISO/IEC, 1999).

The advantage of using synthetic audio as sound sources in virtual acoustic environments is the achieved reduction in the amount of data transferred when compared to natural digitally coded audio. This results, however, in added computation and complexity in the audio rendering. Parametric control provided by synthetic audio is also a remarkable advantage, allowing for flexible processing and manipulation of the sound (Vercoe et al., 1998).

An attractive sound analysis and synthesis method, physical modeling, has become a popular technique among researchers and manufacturers of synthesizers in the past few years (Smith, 1997, 1998; Tolonen et al., 1998). Physical model based sound synthesis methods are particularly suitable to virtual acoustics simulation due to many reasons. First, one of the very aims in virtual acoustics is to create models for the source, the room, and the listener that are based on their physical properties and can be controlled in a way that resembles their physical behavior. Second, source properties such as directivity may be incorporated into the physical modeling synthesizers (Huopaniemi et al., 1994; Karjalainen et al., 1995).

[1] See (Smith, 1991; Roads, 1995; Tolonen et al., 1998) for a taxonomy of modern sound synthesis techniques.

3.1 Properties of Sound Sources

Modeling the radiation properties and directivity of sound sources is an important (yet often forgotten) topic in virtual reality systems. During the course of this study, the directional properties of musical instruments (Huopaniemi et al., 1994; Karjalainen et al., 1995; Savioja et al., 1999) and the human head (Huopaniemi et al., 1999a) were investigated, and real-time algorithms for frequency-dependent directivity filtering for a point-source approximation were implemented. In general, mathematical modeling of the directivity characteristics of sound sources (loudspeakers, musical instruments) can be a complex and time-consuming task. Therefore, in many cases measurements of directivity are used to obtain numerical data for simulation purposes (Meyer, 1978).

3.1.1 On the Directivity of Musical Instruments

String instruments exhibit complex sound radiation patterns due to various reasons. The resonant mode frequencies of the instrument body account for most of the sound radiation (Fletcher and Rossing, 1991). Each mode frequency of the body has a directivity pattern such as monopole, dipole, quadrupole, or their combination. The sound radiated from the vibrating strings, however, is weak and can be neglected in the simulation. In wind instruments, the radiation properties are dominated by outcoming sound from various parts of the instrument (the finger holes or the bell). For example, in the case of the flute, the directivity is caused by radiation from the embouchure hole (the breathier noisy sound) and the toneholes (the harmonic content) as discussed in (Karjalainen et al., 1995).

Another noticeable factor in the modeling of directivity is the directional masking and reflections caused by the player of the instrument. Masking plays an important role in virtual environments where the listener and sound sources are freely moving in a space. A more detailed investigation of musical instrument directivity properties can be found, e.g., in (Meyer, 1978). Comparisons between instrument directivity with and without the player have been published by Cook and Trueman (1998).

3.1.2 Source Directivity Models

For real-time simulation purposes, it is necessary to derive simplified source directivity models that are efficient from the signal processing point of view and as good as possible from the perceptual point of view. Different strategies for implementing directivity in physical models of musical instruments have been proposed by the author in several studies (Huopaniemi et al., 1994; Karjalainen et al., 1995) in relation to sound sources in virtual environments. Of these directivity modeling methods, there are two strategies that are attractive to virtual reality audio source simulation in general: 1) directional filtering and 2) a set of elementary

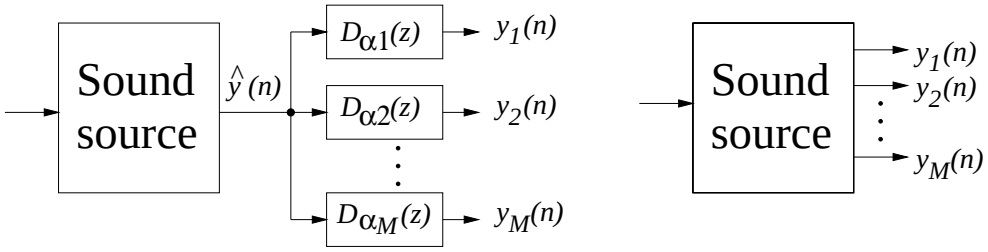


Figure 3.1: Two methods for incorporating directivity into sound source models: (a) directional filtering, (b) set of elementary sources (Karjalainen et al., 1995; Savioja et al., 1999).

sources. In Fig. 3.1, these two methods are illustrated. To obtain empirical data, measurements of source directivity in an anechoic chamber were conducted for two musical instruments: the acoustic guitar and the trumpet (Huopaniemi et al., 1994; Karjalainen et al., 1995; Välimäki et al., 1996)[2]. In both cases, two identical microphones were placed at a 1 meter distance from the source and the player, one being in the reference direction (normally 0° azimuth and elevation) and the other being in the measured direction. An impulsive excitation was used and the responses $H_{\alpha_0}(z)$ and $H_{\alpha_m}(z)$ were registered simultaneously. For modeling purposes, the directivity *relative to main-axis radiation* was examined, that is, the deconvolution of the reference and measured magnitude responses was computed for each direction:

$$|D_{\alpha_m}(z)| = \frac{|H_{\alpha_m}(z)|}{|H_{\alpha_0}(z)|} \tag{3.1}$$

In the following, the two main approaches for real-time source directivity approximation are studied.

3.1.3 Directional Filtering

The directivity properties of sound sources may be efficiently modeled in conjunction with geometrical acoustics methods such as the image-source method. In this technique, the monophonic sound source output $\hat{y(n)}$ is fed to M angle-dependent digital filters $D_{\alpha_m}(z)$ $(m = 1, \ldots, M)$ representing each modeled output direction from the source (see Fig. 3.1a). Generally, the lowpass characteristic of the direction filter increases as a function of angle, and low-order filters have been found to suffice for real-time models.

As an example, a set of first-order IIR filters modeling the directivity characteristics of the trumpet (and the player) is represented in Fig. 3.2. The directivity

[2] More extensive instrument radiation databases have been measured and published by Cook and Trueman (1998) and Semidor and Couthon (1998).

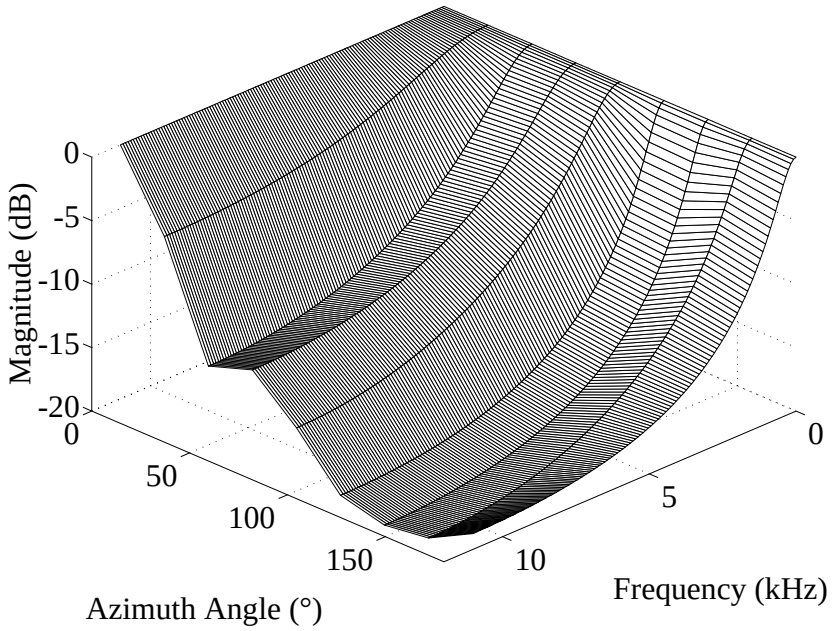


Figure 3.2: Modeling the directivity characteristics of the trumpet with first-order IIR filters. The azimuth interval of modeled responses is 22.5° (Karjalainen et al., 1995).

functions were obtained using Eq. (3.1). A first-order IIR filter was fitted to each result. The filters were designed with a minimum-phase constraint. In our measurements, we analyzed directivity only on the horizontal plane (0° elevation).

3.1.4 Set of Elementary Sources

The radiation and directivity patterns of sound sources can also be distributed, and a point-source approximation may not be adequate for representation. Such examples could be, e.g., a line source. So far only a point source type has been considered and more complex sources have been modeled as multiple point sources. The radiation pattern of a musical instrument may thus be approximated by a small number of elementary sources. These sources are incorporated in, e.g., the physical model and each of them produces an output signal as shown in Fig. 3.1b. (Huopaniemi et al., 1994; Karjalainen et al., 1995). This approach is particularly well suited to flutes, where there are inherently two point sources of sound radiation, the embouchure hole and the first open tone hole.

3.2 Sound Radiation from the Mouth

In this section, a case study on modeling sound radiation from the human mouth is presented. Mathematical and empirical methods are discussed, and a novel technique based on the principle of reciprocity is introduced (Huopaniemi et al., 1999a).

The free-field radiated sound spectrum from the mouth varies as a function of distance and angle. The distance attenuation is known to be caused by the transmitting medium, air, whereas the directional attenuation is caused by the human head and torso. The angular dependence influences the perceived speech spectrum and intelligibility. In recent years, incorporating directional properties of sound sources have become attractive in virtual reality and multimedia systems (Savioja et al., 1999). An example of such specification is the MPEG-4 Systems scene description language (BIFS) (Koenen, 1999; Swaminathan et al., 1999) and its advanced audio rendering capabilities (Scheirer et al., 1999). MPEG-4 features human face and body animation, text-to-speech synthesis technology as well as advanced audiovisual scene description capabilities, thus interactive virtual reality modeling of human speakers is becoming more and more attractive with a wide range of application areas.

Directional filtering caused by the human head and torso has been investigated by Flanagan (1960, 1972). Meyer (1978), Marshall and Meyer (1985), did further work on practical analysis of the directivity of singers. Sugiyama and Irii (1991) extended a spherical head directivity model to an analytical model of a prolate spheroid in an infinite baffle, which was claimed to provide a closer approximation to the radiation directivity properties of human or dummy head mouth. Generally it is possible to take both an analytical and empirical approach to the problem of obtaining directivity data. Simple analytical models are relatively easy to compute and produce moderately good results when compared to empirical data (Flanagan, 1960). Empirical directivity measurements, on the other hand, can be fairly easily accomplished using a dummy head, but measurements using human test subjects are very difficult due to, for example, problems in transducer placement at the mouth opening (if real speech is not used as input).

An interesting observation can be made here concerning spherical head directivity and its relation to spherical head HRTFs (head-related transfer functions). Many authors have introduced simplified HRTFs based on approximating pressure distribution around a rigid sphere the size of a human head (Rayleigh, 1904; Cooper, 1982; Rabinowitz et al., 1993; Blauert, 1997; Duda and Martens, 1998). But, based on the *principle of reciprocity*, we are able to use the same data also for approximating directivity caused by the human head by exchanging the source and receiver positions (Morse and Ingard, 1968, pp. 342-343). This statement is valid if a simple spherical point source approximation is used. The main novelty in this study (Huopaniemi et al., 1999a) has been the application of the reci-

procity method to practical directivity measurements; by exchanging the source and receiver positions a very straightforward measurement can be carried out.

3.2.1 Analytical Formulation of Head Directivity

Analytically, the spherical head directivity can be estimated by examining the diffraction around the rigid head and exchanging the source and receiver positions by introducing the principle of reciprocity. The boundary conditions to be fulfilled are (Blauert, 1997):

- At distance $r >> a$ wave propagation is assumed to be planar (r is the distance from the source to the center of the sphere, and a is the radius of the sphere).
- At the surface of the sphere the normal component of the volume velocity is zero.

A formula for range-dependent diffraction coefficient calculation has been derived in (Rabinowitz et al., 1993) and (Duda and Martens, 1998). The pressure p_s along the surface of a sphere of radius a is given by Duda and Martens (1998, Eqs. 2-3):

$$p_s(r, a, f, \theta, t) = \frac{i\rho_0 c S_\omega}{4\pi a^2} \Psi e^{-i2\pi f t}, \tag{3.2}$$

where ρ_0 is the density of air, c is the speed of sound, f is frequency, θ is the azimuth angle, and S_ω is the magnitude of flow of an ideal point source. The variable Ψ is the series expansion

$$\Psi = \sum_{m=0}^{\infty} (2m+1) P_m(\cos\theta) \frac{h_m(kr)}{h'_m(ka)}, r < a, \tag{3.3}$$

where P_m is the Legendre polynomial of order m, $k = 2\pi f/c$ is wave number, $h_m(kr)$ is the spherical Hankel function of order m in kr, and $h'_m(ka)$ is the derivative of spherical Hankel function of order m with respect to ka. By exchanging the source and receiver positions, the directivity relative to main-axis radiation ($\theta = 0°$) can thus be calculated as:

$$d_\theta = \frac{p_s(r, a, f, \theta, t)}{p_s(r, a, f, \theta = 0°, t)}. \tag{3.4}$$

In Fig. 3.3, a three-dimensional plot of the spherical head approximation directivity for azimuth angles $0° - 180°$ is shown. The lowpass characteristic behavior is clearly seen[3]. In the next section, the empirical measurement method for head directivity is discussed, and results for the reciprocity measurement principle are provided.

[3] The "lobing" or "bright spot" effect at 180° azimuth is clearly visible in Fig. 3.3. This is caused by superposition of arriving signals in-phase, and is likely to be reduced in real measurements.

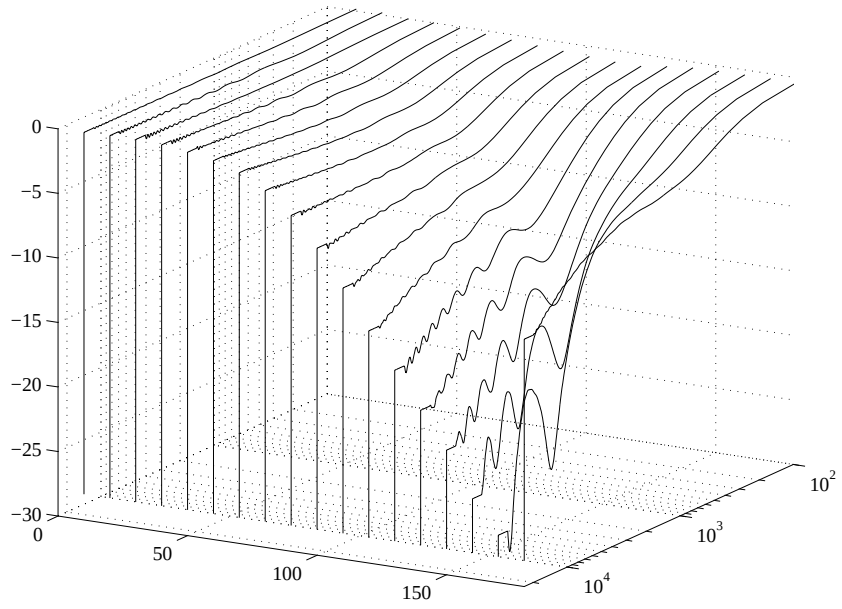


Figure 3.3: Spherical head approximation directivity at 2 meter distance. X-axis: Azimuth angle (degrees), Y-axis: Frequency (Hz), Z-axis: Magnitude (dB).

3.2.2 Measurement of Head Directivity

In order to compare and investigate the measurement of radiation directivity characteristics from a human mouth in relation to analytical methods described in the previous section, a set of measurements was carried out. The goal was to compare a direct measurement of directional radiation to the reciprocity method where the source and receiver positions are exchanged. The measurements were performed in the large anechoic chamber of the Acoustics Laboratory at the Helsinki University of Technology. The measurements consisted of two parts, in both of which a Brüel&Kjaer BK4128 dummy head was placed on an electronic turntable which was rotated at 10° azimuth angular intervals.

The test signal used in the measurements was a 8192 samples long pseudo-random (flat spectrum random phase) signal with a sampling frequency of 48000 Hz. Averaging of 30 measurements was used to improve the signal-to-noise ratio. The measurements were carried out using a signal processing and measurement program QuickSig (Karjalainen, 1990) in the Apple Macintosh platform. The test signal was generated with a dual channel 16-bit AD/DA card installed in the computer. (Possible interferences of the sound card were eliminated by using the other channel as a reference by connecting it directly from output to input.) The signal was then amplified (Luxman M-120A) and transmitted through a loud-speaker (a 150mm diameter spherical shaped loudspeaker cabinet with a 4-inch

full-range Audax AT100M0 element (Riederer, 1998a)). The sound pressure was picked up by a miniature measurement microphone (Sennheiser KE 4-211-2), and then amplified with a microphone preamplifier and transmitted to the input of the sound card.

First the radiation from the mouth was studied in the direct measurement. Figure 3.4 illustrates the measurement principle. The signal was transmitted using the built-in loudspeaker in the dummy head's mouth (BK4128). A microphone was set at the same height at the distance of 2 meters from the center of the head. After registering the data and rotating the turntable, the measurement was repeated at 10° azimuth increments for a full circle.

To test the principle of reciprocity, another set of measurements was required. Figure 3.5 illustrates this measurement principle. The microphone was now placed in the dummy head's mouth at two positions: a) about 1.4 centimeters from the lip plane, and b) at the mouth opening entrance. The spherical loudspeaker (Audax AT100M0) was used and placed at the same height at a 2 meter distance. In this case the sound pressure was registered in the mouth and the measurements were repeated as earlier. Theoretically the measurement conditions for the reciprocity method should be equivalent, i.e., the same transducers and measuring positions should be used. However, as Eq. 3.4 illustrates, the directivity is computed relative to main axis radiation, thus the effects of the transducers are not influencing the final result.

The impulse responses obtained as measurement results were converted from QuickSig (Karjalainen, 1990) to Matlab (Mathworks, 1994) for analysis. The measurement results were limited to the range of 100 Hz to 15000 Hz. For both tests, the angular frequency responses were divided by the reference response measured at 0° as derived in Eq. 3.4. Figures 3.6–3.7 illustrate the measurement results for the direct case (dashed line) and the reciprocal case (solid line). Figure 3.6 corresponds to the measurement point 14mm inside the mouth opening, and Fig. 3.7 refers to the microphone position at the mouth opening. Generally, for head directivity, it is clear that as the angle of incidence increases, increased lowpass filtering results (as is well known from earlier studies), despite of the slight undulation in the responses. The attenuation of high frequencies becomes especially significant as the wavelength approaches, and decreases beyond, the dimensions of the human head. The lobing effect, clearly present in the analytical model results, is also shown in the empirical measurement results, although not as prominently.

Comparing the direct measurement results to the responses from the reciprocity measurement, it can be seen that they match very well (within 1-2 dB accuracy) up to about 5 kHz. After this the placing of the microphone inside the dummy head's mouth becomes interfering, since these wavelengths approach the dimensions of the mouth and the microphone. The difference in this frequency range can be as much as 10 dB (not shown in the figures). However, because of the good results with frequencies from 100 Hz to 5500 Hz, which are significant

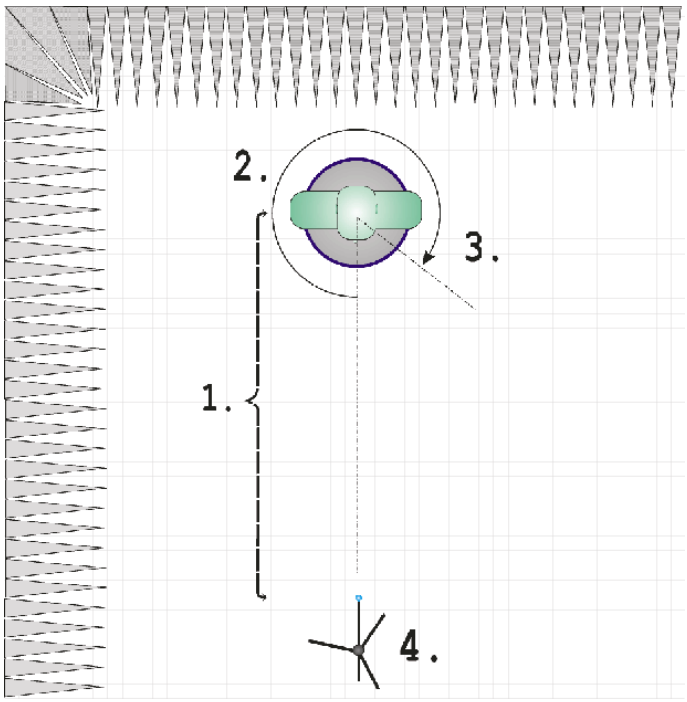


Figure 3.4: Measurement installation in the anechoic chamber for the direct method. **1.** Distance to source (the mouth speaker of the dummy head): 2 meters. **2.** Azimuth angle. **3.** Rotational direction of the BK4128 dummy head placed on a turntable. **4.** Measurement microphone (Sennheiser KE 4-211-2) on stand (Huopaniemi et al., 1999a).

for sound radiation in normal speech bandwidth, it can be said that measurement based on the principle of reciprocity has very satisfactory performance. The measurement position in the reciprocity technique did not seem to remarkably affect the results in the frequency range of interest, below 5 kHz, as was expected.

3.2.3 Modeling and Measurement Result Comparison

In the previous two sections, the analytical and empirical approaches for obtaining head directivity data were discussed. In this section, the aim is to compare and discuss the two approaches and their validity. A similar plot to Figs. 3.6–3.7 for directivity characteristics of the spherical head model is shown in Fig. 3.8. It can clearly be seen that the overall macroscopic characteristics such as lowpass filtering as the angle on incidence increases and the lobing effect at the symmetry point in the back of the head are similar in both methods. However, it is also noticeable that the overall directivity contours by frequency band as a function of azimuth angle differ remarkably. This is clearly due to the fact that a simple spherical model is not an adequate replica of a real or dummy head. It has been

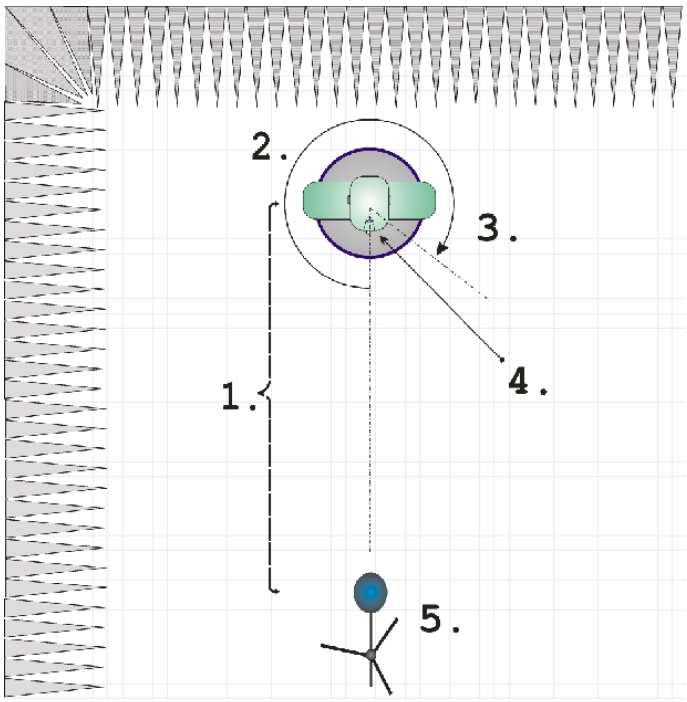


Figure 3.5: Measurement installation in the anechoic chamber during reciprocity method. **1.** Distance to source (spherical loudspeaker): 2 meters. **2.** Azimuth angle. **3.** Rotational direction of the BK4128 dummy head. **3.** Measurement microphone (Sennheiser KE 4-211-2) installed in the mouth of the dummy head. **5.** Spherical loudspeaker (Audax AT100M0) on stand (Huopaniemi et al., 1999a).

argued in the literature (Sugiyama and Irii, 1991) that a prolate spheroid resembles the human head shape more accurately and would predict the directivity cues more reliably.

3.3 Discussion and Conclusions

In this chapter, topics in sound source characterization and directivity modeling were overviewed, and efficient digital filter simulations of musical instrument and human head directivity were illustrated. A novel measurement and modeling technique for sound radiation from the human mouth has been developed by the author. Results of empirical and analytical verification of the method were provided.

In general, sound source models in virtual acoustics have been often neglected. In this work, methods for sound source characterization have been presented that can be used particularly in virtual reality applications. Due to the simplicity of many of the models, these methods may not be suitable for reproducing complex source radiation patterns. For practical applications in the context of this thesis,

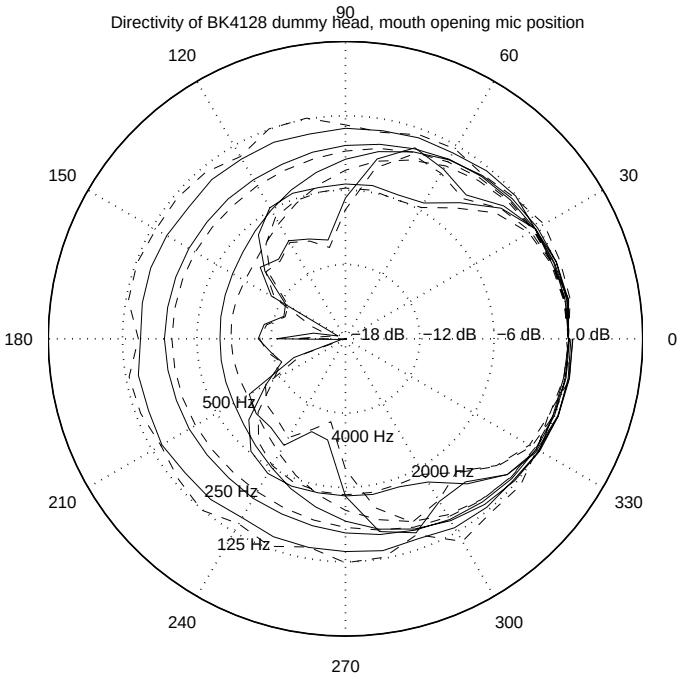


Figure 3.6: Directivity of the BK4128 dummy head using the direct method (dashed line) and the reciprocity method (solid line). The reciprocal microphone position is at the mouth transducer position (14 mm inside the mouth opening).

however, the framework and results have proven to be very relevant.

In the next chapter, topics in geometrical room acoustics modeling are discussed. Emphasis is placed on a parametric room impulse response rendering technique, which is particularly suitable for dynamic interactive auralization.

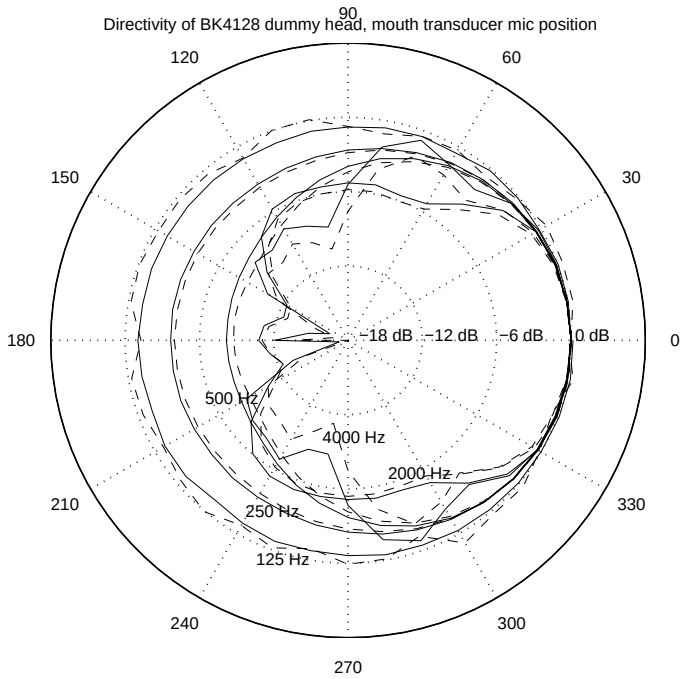


Figure 3.7: Directivity of the BK4128 dummy head using the direct method (dashed line) and the reciprocity method (solid line). The reciprocal microphone position is at the mouth opening.

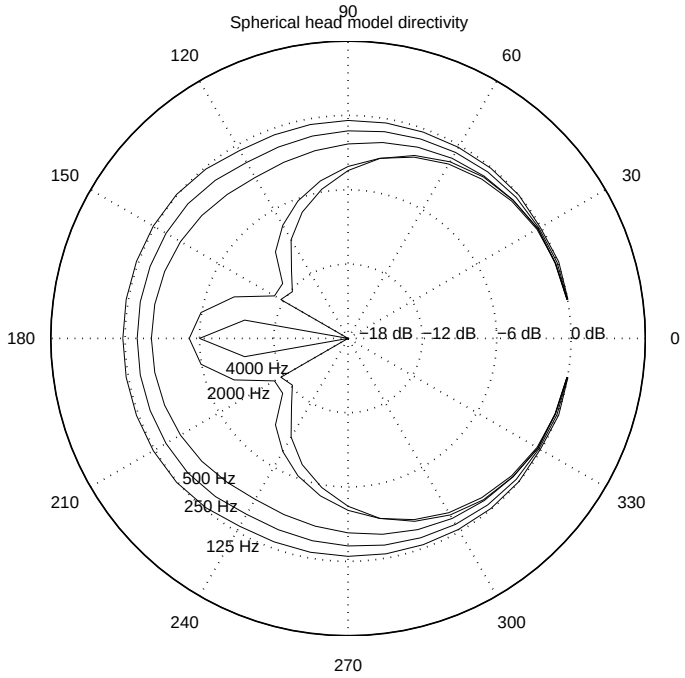


Figure 3.8: Directivity of the analytical spherical head approximation.

Chapter 4

Enhanced Geometrical Room Acoustics Modeling

This chapter deals with added features to a real-time geometrical room acoustics modeling scheme presented in Chapter 2. Among the modeled features are reflection and air absorption behavior. Material absorption values are used, e.g., for room impulse response (RIR) prediction by computational methods such as the ray-tracing and the image-source technique (Kuttruff, 1995; Savioja et al., 1999). If the material data is given as a measured impulse response or transfer function, the problem is to reduce it for efficient simulation. The distance-dependent air absorption is also a significant factor that needs to be taken into account in accurate simulation. The parametric room impulse response rendering technique, discussed in Chapter 2.3.1, separates the processing of direct sound and early reflections from late reverberation. In a dynamic virtual acoustic environment, the role of the first reflections is very important in adding to the perception of the space. In the following, a technique where boundary material and air absorption characteristics are modeled by low-order minimum-phase digital filters is presented. The filter design takes into account the logarithmic distribution of octave frequencies as well as the non-uniform frequency resolution of the human ear by using frequency warping or weighting functions. A spectral error measure is used to estimate the needed filter order. The filter design techniques are particularly useful in efficient room acoustics simulation and real-time implementation of virtual acoustic environments.

It is assumed that the sound traveling in the air and reflecting from surfaces behaves linearly and that the system is time invariant. Such a system may be modeled by any linear and time-invariant (LTI) technique. Digital filtering is an especially efficient LTI technique since DSP algorithms have been developed for fast execution. With this approach the problem appears as how to simulate the boundary reflections and air absorption in an efficient, accurate, and physically plausible way. If a transfer function—reflection and/or propagation—is given based on measured impulse response or transfer function data or on an analytical

model, the filter design problem is simply to search for an optimal match of data to given filter type and order and to given error criteria. If the data is available in a magnitude–only form, such as absorption coefficients for octave bands, there is an additional task of finding proper phase characteristics as well as interpolation of the sparsely given data in an acoustically meaningful way.

There exist many 'standard' modeling techniques using digital filtering approach based on AR (all-pole), MA (FIR), or ARMA (pole-zero) modeling. For example, in Matlab (Mathworks, 1994), such filter design functions as `yulewalk`, `invfreqz`, and `cremez`, are available. The selection of a method depends on the available response data and target criteria of the design. In the experiments it was found that the least mean squares fit to sparsely sampled data of absorption coefficients or air absorption[1] (Huopaniemi et al., 1997). The magnitude response is first converted to minimum-phase data, which is then converted to an IIR filter of desired order.

4.1 Acoustical Material Filters

Measurement and modeling of acoustic materials is an important part of room acoustical simulation. The traditional methods for material measurements are the standing wave tube technique (ISO, 1996) and the reverberation room measurement (ISO, 1985). Methods that need digital signal processing are the intensity measurement technique and the transfer function method (Fahy, 1995; Chung and Blaser, 1980). Results may be given in various forms: reflection (impulse) response, reflection transfer function (reflection "coefficient"), impedance or admittance data, or absorption data. In the literature, absorption coefficients are generally given in octave bands from 125 Hz to 4000 Hz, specifying only the magnitude of reflection.

The problem of modeling the sound wave reflection from acoustic boundary materials is a complex one. The temporal or spectral behavior of reflected sound as a function of incident angle, the scattering and diffraction phenomena, etc., makes it impossible to use numerical models that are accurate in all aspects. Depending on the application, more or less approximation is needed (Huopaniemi et al., 1997).

In this work the focus was on DSP-oriented techniques to simulate sound signal behavior in reflection and propagation. Thus many important issues were ignored, such as the angle dependency of reflection, which could be included and approximated by various methods. Also, attention was not paid to the fact that material data measured in diffuse field or in impedance tube should be treated differently.

The most common characterization of acoustic surface materials is based on absorption coefficients, given for octave bands 125 Hz to 4000 Hz. On the con-

[1] For example, the `invfreqz` function in Matlab was found to work well.

trary, the DSP-based measurement methods yield the reflection impulse response $r(t)$ or the complex-valued reflection transfer function (reflectance) $R(j\,\omega) = F\{r(t)\}$, where F is the Fourier transform. Since the absorption coefficient is the power ratio of the absorbed and incident powers, the relation between the absorption coefficient $\alpha(\omega)$ and the reflectance $R(j\,\omega)$ is given by

$$\alpha(\omega) = 1 - |R(j\,\omega)|^2 \tag{4.1}$$

whereby $|R(j\,\omega)| = \sqrt{1-\alpha(\omega)}$ can be used to obtain the absolute value of the reflectance when absorption coefficients are given[2]. The relation between the normalized impedance $Z(j\,\omega)$ and the reflectance $R(j\,\omega)$ is

$$R(j\,\omega) = \frac{Z(j\,\omega) - 1}{Z(j\,\omega) + 1} \tag{4.2}$$

which can be used to compute $R(j\,\omega)$ when the material impedance (or admittance, its inverse value) is given. This equation is valid for a normally incident plane wave reflecting from an infinite planar surface. Based on the equations above, material data can be converted to $R(j\,\omega)$ or $r(t)$ for filter approximation. The application of absorption material filter design to concert hall auralization is discussed next. The polygon model of the venue under study, Sigyn Hall (Lahti and Möller, 1996), located in Turku, Finland, is shown in Fig. 4.1. The absorption coefficient data depicted in Table 4.1 (Naylor and Rindel, 1994; Lahti and Möller, 1996) was taken from the concert hall design. The material names and corresponding octave-band absorption coefficients are shown in Table 4.1. In Fig. 4.2, fitting of low-order filters to absorption data is shown. In the figure, the magnitude responses of 16 IIR filters designed to match the corresponding target values are plotted (the target response at Nyquist frequency was approximated from the 4 kHz octave band absorption coefficient). The filters were designed using a weighted least squares (LS) approximation using ERB scale weighting. From the plot it can be seen that for most absorption properties a first-order IIR filter is an adequate approximation. For some data, a higher-order (order 3) filter was needed.

[2] A negative value of the square root is possible in theory but practically never happens in practice.

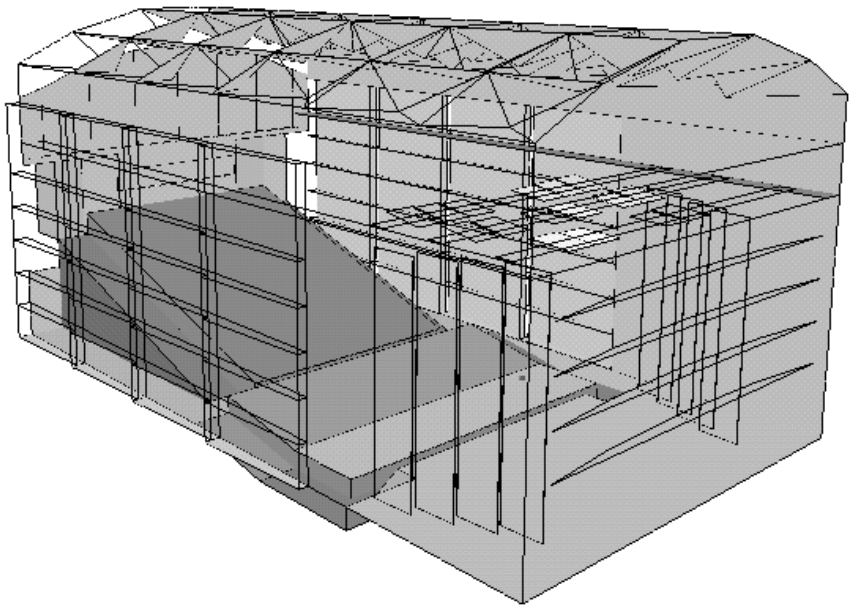


Figure 4.1: Computer model of Sigyn Hall, Turku, Finland (Lahti and Möller, 1996).

125 Hz	**250 Hz**	**500 Hz**	**1000 Hz**	**2000 Hz**	**4000 Hz**	**Description**
0.01	0.01	0.01	0.02	0.02	0.02	smooth beton
0.08	0.11	0.05	0.03	0.02	0.03	gypsum 13mm
0.15	0.1	0.06	0.04	0.04	0.05	gypsum 2x13mm
0.55	0.7	0.65	0.65	0.6	0.6	gypsum 710mm
0.11	0.13	0.05	0.03	0.02	0.03	gypsum 9mm
0.02	0.02	0.04	0.05	0.05	0.1	cork
0.08	0.04	0.03	0.03	0.02	0.02	glass 1k
0.1	0.07	0.05	0.03	0.02	0.02	glass 2k
0.3	0.15	0.1	0.05	0.03	0.02	glass 610
0.25	0.15	0.1	0.09	0.08	0.07	wooden panels22
0.15	0.11	0.1	0.07	0.06	0.07	wooden floor
0.08	0.2	0.55	0.65	0.5	0.4	wooden wall4
0.02	0.02	0.03	0.04	0.04	0.05	linoleum
0.15	0.7	0.6	0.6	0.75	0.75	wool 50mm
0.15	0.7	0.6	0.6	0.85	0.9	wool 50mmkas
0.16	0.24	0.56	0.69	0.81	0.78	audience

Table 4.1: Sigyn Hall absorption coefficient data.

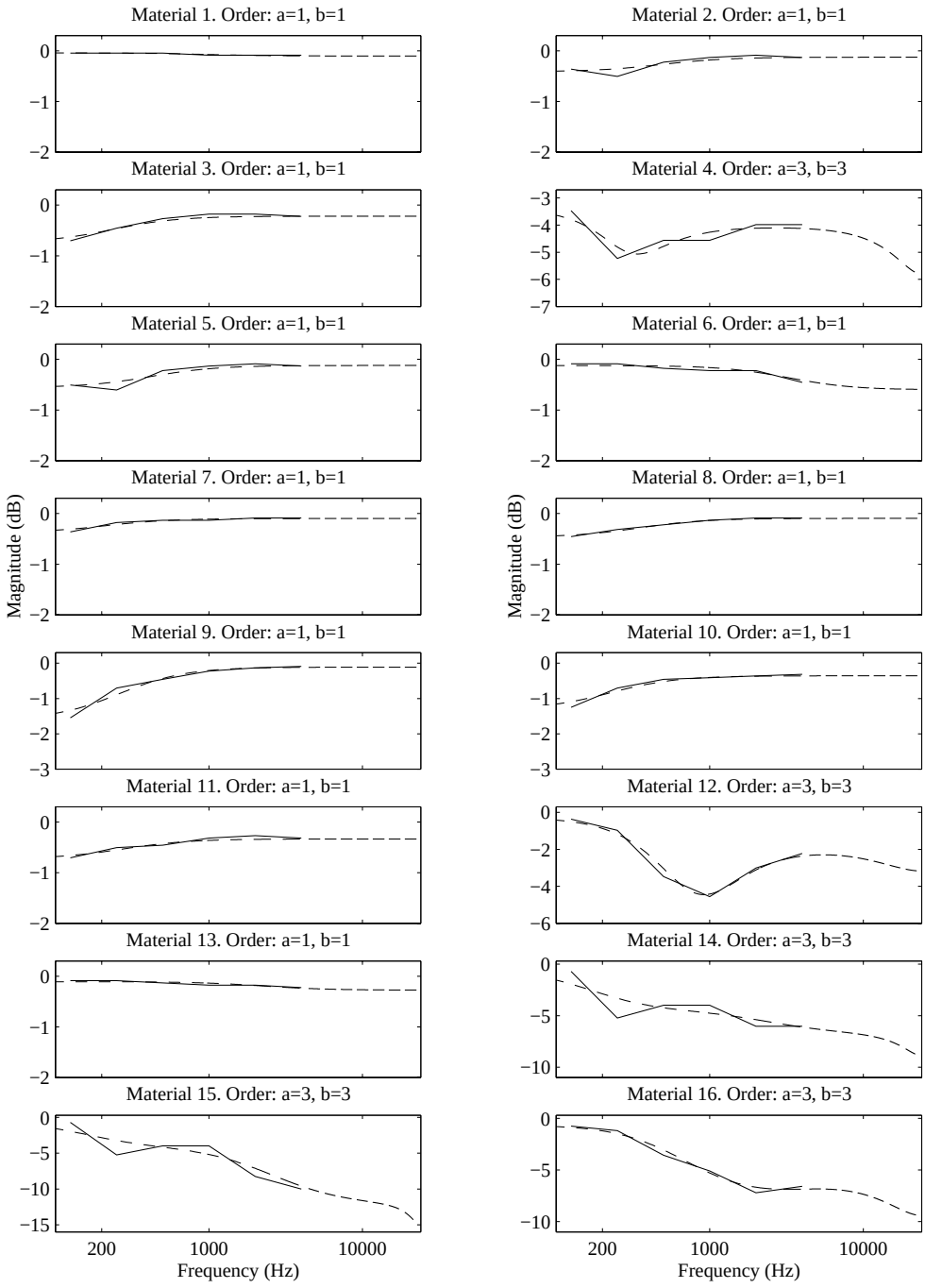


Figure 4.2: First-order and third-order minimum-phase IIR filters designed to given absorption coefficient data. Solid line: octave-band data, dashed line: IIR filter approximation.

4.2 Air Absorption Filters

The effect of air absorption is an important factor in image-source calculations of large acoustical spaces, such as concert halls where higher-order reflections can arrive considerably delayed from the direct sound. The absorption of sound in the transmitting medium (normally air) depends mainly on the distance, temperature and humidity. There are various factors which participate in absorption of sound in air (Kuttruff, 1991). In a normal environment the most important of those is the thermal relaxation. The phenomenon is observed as increasing lowpass filtering as the distance between the listener and the sound source increases.

For real-time simulation purposes, air absorption was approximated in the following efficient manner. Analytical expressions for attenuation of sound in air as a function of temperature, humidity and distance have been published by Bass and Bauer (1972) and discussed in (Kuttruff, 1991). Equations for sound absorption in air have also been standardized (ISO, 1993). These standard formulae were used for calculation, yielding an attenuation a in dB per meter at the frequency f with variables T (temperature in Kelvin), h (molar concentration of water vapor), and p_a (ambient sound pressure amplitude):

$$\begin{aligned} a = 8.686 f^2 \Bigg(& \left[1.84 \times 10^{-11} \left(\frac{p_a}{p_r} \right)^{-1} \left(\frac{T}{T_0} \right)^{1/2} \right] + \left(\frac{T}{T_0} \right)^{-5/2} \\ & \times \Bigg(0.01275 e^{\frac{-2239.1}{T}} \left[f_{ro} + \left(\frac{f^2}{f_{ro}} \right) \right]^{-1} \\ & + 0.1068 e^{\frac{-3352.0}{T}} \left[f_{rn} + \left(\frac{f^2}{f_{rn}} \right) \right]^{-1} \Bigg) \Bigg) , \end{aligned} \tag{4.3}$$

where

$$\begin{aligned} f_{ro} &= \frac{p_a}{p_r} \left(24 + 4.04 \times 10^4 h \frac{0.02 + h}{0.391 + h} \right) \\ f_{rn} &= \frac{p_a}{p_r} \left(\frac{T}{T0} \right)^{-1/2} \left(9 + 280 h e^{-4.170 \left[\left(\frac{T}{T0} \right)^{-1/3} - 1 \right]} \right) , \end{aligned} \tag{4.4}$$

where f_{ro} is the oxygen relaxation frequency, f_{rn} is the nitrogen relaxation frequency, T_0 is the reference air temperature ($T_0 = 293.15$ K), and p_r is the reference ambient atmospheric pressure ($p_r = 101.325$ kPa).

In the following, transfer functions and filter design examples for air absorption at 20°C and 20% humidity ($h = 0.4615$) for various distances are presented. The resulting magnitude responses using Eqs. 4.3–4.4 were fitted with IIR filters (two poles: $na = 2$, one zero: $nb = 1$) designed using the technique described previously. Results of modeling for six distances from the source to the receiver are illustrated in the top plot of Fig. 4.3. As can be seen from the figure, the

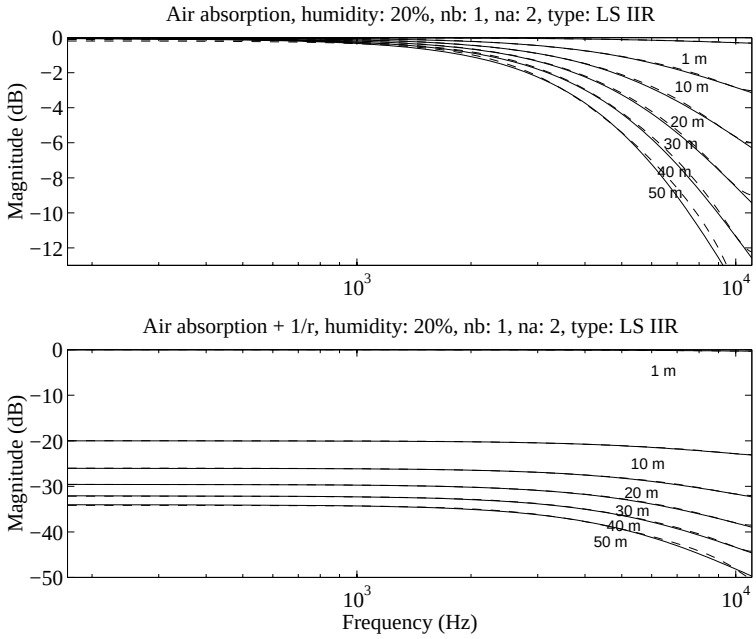


Figure 4.3: Magnitude response of air absorption filters as function of distance (1m-50m) and frequency. The continuous line represents the ideal response and the dashed is the filter response. The air humidity is chosen to be 20% and temperature 20°C. Upper plot: air absorption only. Lower plot: air absorption + 1/r distance attenuation.

lowpass characteristic is very considerable at larger distances from the source. This implies that for perception of the direct sound and first reflections the air absorption is a contributing factor.

The distance attenuation ($1/r$ law) also has a profound effect on the gain of the direct sound and the early reflections. When an $1/r$ gain is added to the filter simulations of Fig. 4.3 (top plot), we result in transfer functions depicted in the lower plot of Fig. 4.3.

4.3 Implementation of Extended Image-Source Model

The real-time implementation of the extended image-source model follows the principles depicted in Fig. 2.6 (on page 38). An exploded view of filter structures $T_k(z)$ and $\mathrm{F}_k(z) = [F_{lk}(z)F_{rk}(z)]^T$ is shown in Figs. 4.4–4.5.

The raw audio input to be auralized is fed to a delay line (as shown in Fig. 2.6). The delay line length corresponds to the propagation delay from the sound source and its image sources to the listener. In the next phase all image sources are filtered with filter blocks $T_k(z)$, which contain the following filters (cascaded or

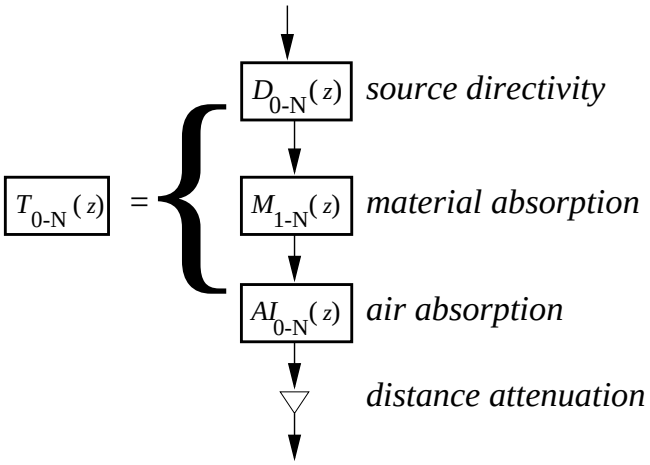


Figure 4.4: Detailed view of the real-time extended image-source modeling algorithm shown in Fig. 2.6 (page 38). The filter structure $T_k(z)$ comprises source directivity filters $D_k(z)$, material absorption filters $M_k(z)$, air absorption filter $AI_k(z)$, and a gain factor for the $1/r$ attenuation law (Savioja et al., 1999).

integrated, depending on system implementation):

- Source directivity filter
- Surface material filters, which represent the filtering occurring at corresponding reflections (not applied to the direct sound).
- The $1/r$-law distance attenuation
- Air absorption filter

The signals produced by $T_k(z)$ are then filtered with listener model head-related transfer function (HRTF) filters $F_k(z)$, which perform binaural spatialization. The recursive late reverberation module $R(z)$ (see Fig. 2.6) is summed to the outputs from the direct sound and early reflection module.

4.4 Discussion and Conclusions

In this chapter, topics in real-time geometrical room acoustics modeling were discussed. Absorption due to reflections from boundary materials and the transmitting medium was reviewed. A modeling technique for material and air absorption using low-order digital filters was presented. Implementation aspects of an extended image-source model that takes into account directivity, and material and air absorption effects were summarized.

It should be noted that the image-source method has limitations in the capability to accurately model diffusion and diffraction. On the other hand, it is

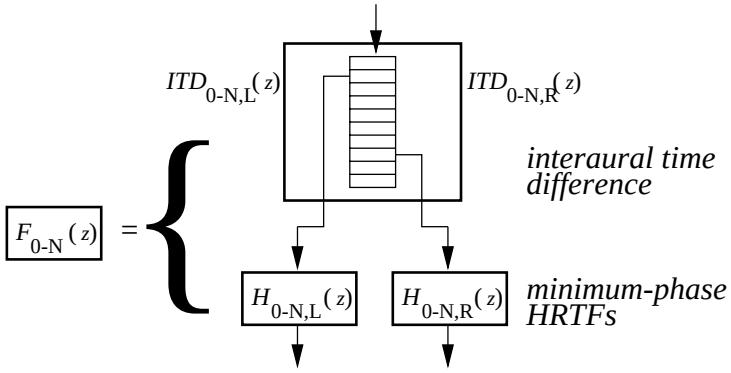


Figure 4.5: Detailed view of the real-time extended image-source modeling algorithm shown in Fig. 2.6 (page 38). The filter structure $F_k(z)$ implements the interaural time delays ITD_k and the minimum-phase HRTFs $H_k(z)$ (Savioja et al., 1999).

a practical choice for implementation of the direct sound and early reflections in a parametric room impulse rendering scheme. The perceptual importance of improved modeling of reflections and air absorption was not verified in subjective listening experiments. It is expected that the effect of these features is not significant in static reproduction, but in dynamic simulation, where both the listener and source can move in the virtual acoustic space, correct implementation of the features is important.

In the next chapter, topics in binaural modeling are covered. Application and approximation methods of head-related transfer functions (HRTFs) are explored.

Chapter 5

Binaural Modeling and Reproduction

The accuracy needed in binaural filter design and synthesis is a current issue in multimedia and virtual environment research. More and more effort is being laid on computationally efficient binaural synthesis models for headphone and loudspeaker reproduction (Huopaniemi and Smith, 1999; Jot et al., 1998b).

In this chapter, the basic theory for binaural synthesis is presented first. Manipulation of modeled or measured HRTFs to suitable form by proper equalization and smoothing techniques is reviewed. Digital filter design aspects of HRTF modeling are reviewed extensively, and novel methods are proposed that take into account the non-uniform frequency resolution of the human ear. New objective and subjective methods of binaural filter design quality estimation are presented, based on binaural auditory modeling and listening experiments.

In the literature, several review papers on 3-D sound can be found that give a detailed description and discussion on topics related to this chapter. Such examples are articles by Møller (1992); Kleiner et al. (1993); Kendall (1995) and books by Blauert (1997); Begault (1994); Gilkey and Anderson (1997).

5.1 Properties and Modeling of HRTFs

HRTFs are the output of a linear and time-invariant (LTI) system, that is, the diffraction and reflections of the human head, the outer ear, and the torso, in response to a free-field stimulus registered at a point in the ear canal. Thus, in theory, the head-related impulse responses (HRIRs), when properly processed, can directly be represented as non-recursive finite-impulse response (FIR) filters. There are often computational constraints, however, that call for HRTF approximation. This can be carried out using conventional digital filter design techniques, but it is necessary to note that the filter design problem is not a straightforward one. It should be possible to design arbitrary-shaped mixed-phase filters that

meet the set criteria both in both amplitude and phase response. The main questions of interest that the filter design expert is faced with are at this point: What is important in HRTF modeling? Are there constraints in the amplitude and phase response, and if so, how are they distributed over the frequency range of hearing? In spatial hearing, it is well known that low-frequency (below approximately 1.5 kHz) interaural time delay (ITD) cues have a dominant role in localization, whereas the interaural level difference (ILD) is the major localization cue at higher frequencies. Furthermore, mid- to high-frequency spectral cues enable elevation detection and out-of-head localization, especially in the median plane (where interaural cues vanish) and on the cone of confusion (where interaural cues are ambiguous). It is, however, not clear what the salience of these cues in localization is, and how accurately they should be modeled in binaural synthesis.

The process of binaural synthesis (see Fig. 2.2 in Section 2.2) can be formulated in terms of an LTI system. If the monophonic input signal is denoted by $x_m(n)$, the resulting binaural signals are obtained by the following convolution:

$$\begin{aligned} \mathbf{y} &= \mathbf{h} * x_m(n) \\ \mathbf{y} &= \begin{bmatrix} y_l(n) \\ y_r(n) \end{bmatrix}, \mathbf{h} = \begin{bmatrix} h_l(n) \\ h_r(n) \end{bmatrix} \end{aligned} \tag{5.1}$$

where $\mathbf{y}$ is a column vector of binaural signals and $\mathbf{h}$ is a column vector of HRIRs used for binaural synthesis. Going from discrete time-domain to a discrete frequency-domain representation, Eq. 5.1 can be rewritten as:

$$\begin{aligned} \mathbf{Y} &= \mathbf{H} X_m(e^{j\omega}) \\ \mathbf{Y} &= \begin{bmatrix} Y_l(e^{j\omega}) \\ Y_r(e^{j\omega}) \end{bmatrix}, \mathbf{H} = \begin{bmatrix} H_l(e^{j\omega}) \\ H_r(e^{j\omega}) \end{bmatrix} \end{aligned} \tag{5.2}$$

Next, any single HRTF can be noted by $H(e^{j\omega})$ (for convenience from now on the term HRTF refers both to a time-domain HRIR and a frequency-domain HRTF). The frequency response of an HRTF can be decomposed into amplitude and phase responses:

$$H(e^{j\omega}) = \left|H(e^{j\omega})\right| e^{j\arg\left\{H(e^{j\omega})\right\}} \tag{5.3}$$

Any LTI system function can be further represented as a cascade of a minimum-phase and an allpass-phase system (Oppenheim and Schafer, 1989):

$$H(e^{j\omega}) = \left|H(e^{j\omega})\right| e^{j(\arg\left\{H_{ap}(e^{j\omega})\right\} + \arg\left\{H_{mp}(e^{j\omega})\right\})} \tag{5.4}$$

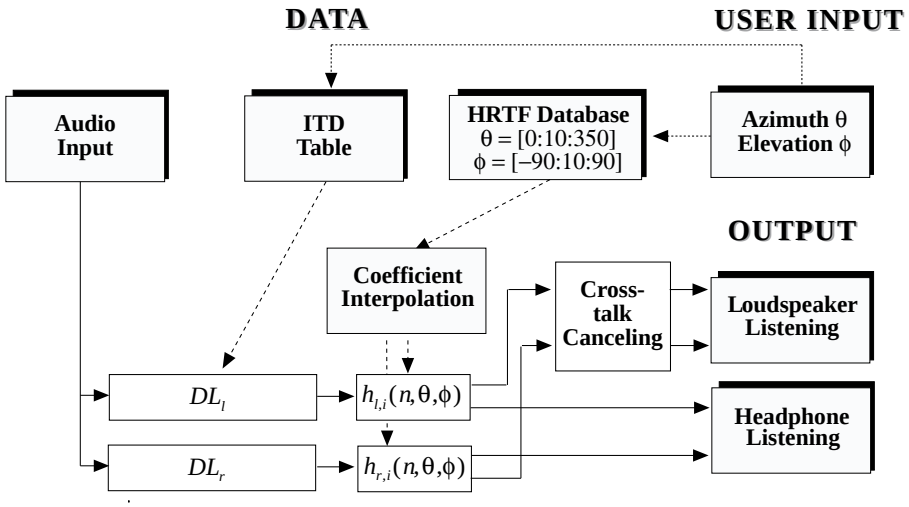


Figure 5.1: General implementation framework of dynamic 3-D sound using non-recursive filters. HRTF modeling is carried out using pure delays to represent the ITD (DL_l and DL_r), and minimum-phase reconstructed interpolated HRIRs ($h_{l,i}(n)$ and $h_{r,i}(n)$).

In Fig. 5.1, a schematic for 3-D sound processing for a single source using non-recursive filtering with dynamic interpolation is presented. The user gives desired azimuth and elevation angles θ and ϕ, after which the corresponding filter coefficients are calculated by fetching the needed ITD and HRTF (HRIR) values from a database using table lookup and coefficient interpolation. Delay lines DL_l and DL_r correspond to the ITD values given for the desired azimuth and elevation angles, and $h_{l,i}(n,\theta,\phi)$ and $h_{r,i}(n,\theta,\phi)$ are the interpolated minimum-phase HRIRs.

Minimum-Phase Reconstruction

An attractive property of head-related transfer functions is that they are nearly minimum phase (Mehrgardt and Mellert, 1977; Kistler and Wightman, 1992). Independent research studies have shown that minimum-phase reconstruction of HRTFs does not have any perceptual consequences (Kistler and Wightman, 1992; Kulkarni et al., 1995, 1999). A minimum-phase system refers to a transfer function in which all the zeros (and poles) lie inside the unit circle. It also has the least energy delay of all realizable systems (with the same magnitude response), thus suggesting benefits in filter design and in interpolation of coefficients (in time-varying implementations). A minimum-phase reconstruction of any system function can be conveniently carried out based on the property that the real and imaginary parts of $H(e^{j\omega})$ are related by the Hilbert transform (Oppenheim and Schafer, 1989). Therefore, a minimum-phase reconstructed version $\left|H_{\mathrm{mp}}(e^{j\omega})\right|$ of

a system response $|H(e^{j\omega})|$ can be found by windowing the *real cepstrum* in the following way (see Oppenheim and Schafer (1989, p. 784)). Let us denote the cepstral window function yielding a minimum-phase reconstruction by $w_{\mathrm{mp}}(n)$, where

$$w_{\mathrm{mp}}(n) = \begin{cases} 1, & n = 1 \text{ or } n = N/2 + 1, \\ 2, & n = 2, \ldots, N/2, \\ 0, & n = N/2 + 2, \ldots, N. \end{cases} \tag{5.5}$$

for even N. The cepstral windowing is accomplished by first taking the inverse fast Fourier transform (IFFT) of the logarithm of the magnitude frequency response (the definition of cepstrum):

$$c(n) = \mathrm{F}^{-1}\left\{\log\left|\mathrm{F}\left\{h(n)\right\}\right|\right\} = \mathrm{F}^{-1}\left\{\log\left|H(e^{j\omega})\right|\right\} \tag{5.6}$$

and windowing the cepstrum by the defined function $w_{\mathrm{mp}}(n)$:

$$\hat{c}(n) = c(n) w_{\mathrm{mp}}(n) \tag{5.7}$$

The minimum-phase impulse response is then reconstructed by the following operation:

$$h_{\mathrm{mp}}(n) = \mathrm{F}^{-1}\left\{\mathrm{e}^{\mathrm{F}\{\hat{c}(n)\}}\right\} \tag{5.8}$$

The decomposition of HRTFs into minimum-phase and allpass components has many advantages. With minimum-phase reconstructed HRTFs, it is possible to separate and estimate the ITD of the filter pair, and insert the delay as a separate delay line (and/or an allpass filter) in series with one of the filters in the simulation stage (an example is seen in Fig. 5.1). The same delay line may also be used for implementing a reflection path delay in the image-source method (Takala et al., 1996; Savioja et al., 1999). The frequency-independent ITD can be calculated using the methods presented in the following subsection. It should be noted, however, that the delay error τ_{error} due to rounding of the ITD to the nearest unit-delay multiple may cause artifacts in the processing stage. The maximum angular error caused by ITD rounding can be estimated using Eq. 5.11 for and solving for θ:

$$\theta_{error} = \arcsin\left(\frac{c\tau_{\mathrm{error}}}{2a}\right) \tag{5.9}$$

(the maximum ITD error is thus $\tau_{\mathrm{error}} = 1/f_s$). For example, at $f_s = 48$ kHz, the maximum angular error $\theta_{\mathrm{error}} = 2.3°$, but at $f_s = 32$ kHz, θ_{error} increases to 3.5°, when $a = 0.0875$ m and $c = 340$ m/s. According to Blauert (1997), the localization blur (localization error) of an average listener in frontal directions

is approximately 3.6°. This would indicate that ITD approximation using an integer number of samples delay is sufficient unless a low sampling rate ($f_s < 32$ kHz) is used. If more accurate control is desired or if the sampling frequency is decreased, the rounding problem can be avoided using fractional delay filtering (see (Laakso et al., 1996) for a comprehensive review on this subject).

An example of time- and frequency-domain representations of HRTFs[1] are shown in Figs. 5.2 and 5.3. The measurements were made on a Cortex MK2 dummy head with a sampling rate $f_s = 48$ kHz. The first 256 samples of the impulse response were used for both the time-domain and the fast Fourier transform (FFT) plots. It can be seen that the bulk delay is removed from the beginning of the impulse responses as a result of minimum-phase reconstruction. Furthermore, some of the fine structure of the impulse responses appears changed. This is evidence that the original HRIRs cannot be made minimum phase by a simple time shift to remove the initial time delay[2]. The magnitude responses shown in Fig. 5.3 are identical as expected. In the following sections, a closer look at the basic spatial hearing cues – ITD, ILD, and spectral information – will be taken.

5.1.1 Interaural Time and Phase Difference

The difference in the time of arrival of sound at the two ears is one of the basic cues of spatial hearing, especially at low frequencies. The ear is sensitive to phase-derived low-frequency ITD cues at frequencies below approximately 1.5 kHz, where the phase of the incident sound is uniquely determined (Blauert, 1997). The low-frequency interaural time difference cues have been found to dominate sound localization (Wightman and Kistler, 1992). The limit to accurate ITD cues is caused by the dimensions of the human head, because at higher frequencies the phase of the incident sound is ambiguous. The frequency-dependent behavior of ITDs has been discussed and modeled in (Kuhn, 1977, 1987). Computational models for ITDs in the horizontal plane were compared and the resulting approximations were as follows. For low-frequency phase-derived ITDs ($ka \ll 1$), it was shown that

$$\tau_{\text{low}} = \frac{3a}{c} \sin\theta, \quad \text{for } ka \ll 1 \tag{5.10}$$

[1] By definition an HRTF is a free-field response at a point in the ear canal divided by the response in the middle of the head with the head absent (Blauert, 1997). Deconvolution was performed to obtain the results shown in Figs. 5.2 and 5.3.

[2] To draw this conclusion, it is necessary to ensure that time aliasing is insignificant in the computation of the cepstrum, which will occur if there are any notches in the HRTF too close to the unit circle. A simple but effective method is to verify that the signal energy at the midpoint of the cepstral inverse-FFT (IFFT) buffer is small compared to the total energy in the buffer (Smith, 1983).

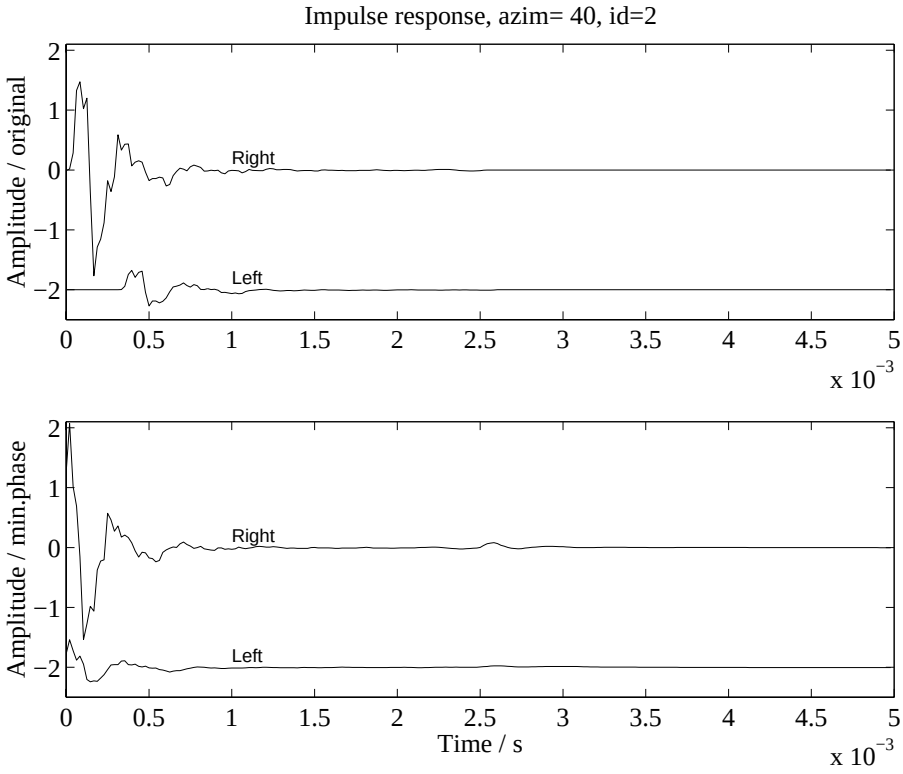


Figure 5.2: Upper plot: Mixed-phase HRIRs. Lower plot: Minimum-phase HRIRs.

where θ is the azimuth angle and $k = 2\pi f/c$ is the acoustic wave number. For high-frequency ITD ($ka >> 1$) the approximation is of the form

$$\tau_{\text{high}} = \frac{2a}{c} \sin\theta, \quad \text{for } ka \gg 1 \tag{5.11}$$

It has been shown in (Kuhn, 1977) that these low- and high-frequency ITD asymptotes match well to measured data on a dummy head. Thus, it can be summarized that the theoretical ratio of the low-frequency ITD (below 500 Hz) to the high-frequency ITD (above 2000 Hz) is approximately 3/2 (Kuhn, 1977). Below and above the asymptotes the ITD behavior is frequency-independent, but in the transition band (approximately 500-2000 Hz) the ITD has a decreasing slope. In (Woodworth and Schlosberg, 1954), another ITD approximation is presented, which is based on a physical model and simplified spherical geometry. This model is a frequency-independent, high-frequency approximation for predicting the ITD. The method of calculation is presented in Fig. 5.4. This approach yields for parallel sound incidence (far-field case):

$$\tau_{\text{group}} = \frac{a\,(\sin\theta + \theta)}{c} \tag{5.12}$$

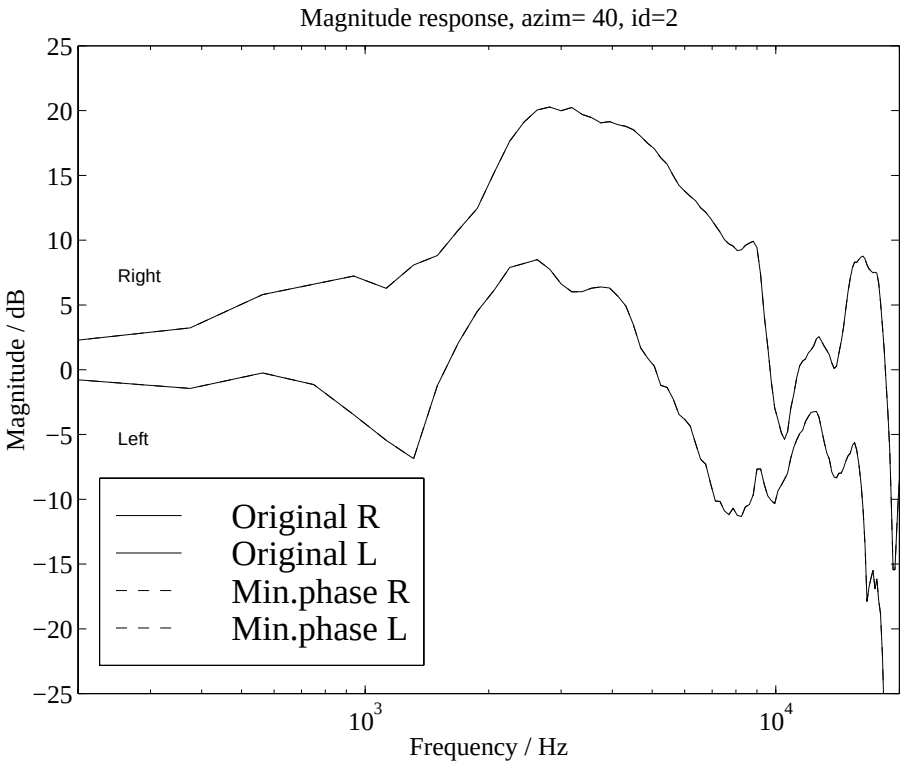


Figure 5.3: Mixed-phase and minimum-phase HRTFs (note that the mixed-phase and minimum-phase magnitude responses are equal).

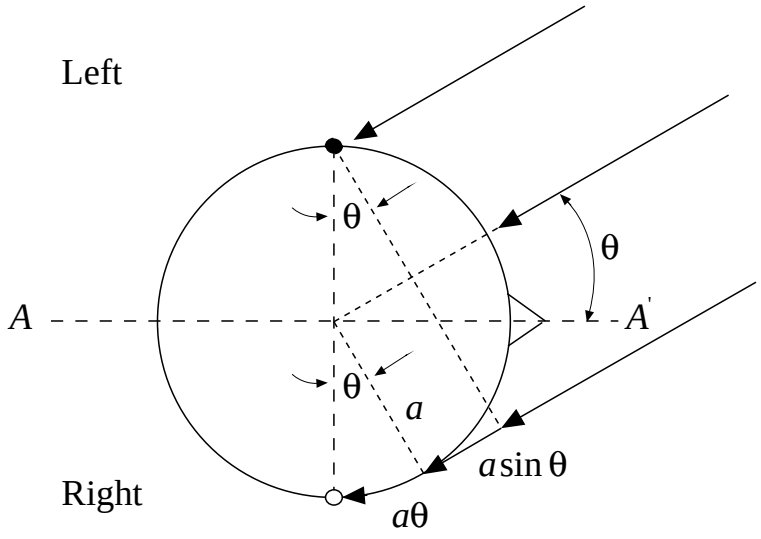


Figure 5.4: ITD approximation based on distance difference of a sound arriving at the two ears.

Estimating the ITD from Measured HRTFs

Apart from theoretical models for ITD, there is often a need to estimate the ITD from actual measurements of HRTFs. Here a distinction should be made between methods that are purely a) for estimating the ITD between two HRTFs and b) for estimating the ITD for a digital filter realization of HRTFs (the decomposition to minimum-phase and allpass sections). Three techniques are generally used for derivation of the ITD:

- Computation of the exact interaural phase difference (IPD) and derivation of a frequency-independent or frequency-dependent ITD approximation
- The cross-correlation method, where the ITD is computed as the offset of the cross-correlation maxima of the two impulse responses
- The leading-edge method, where the time difference of the start of the impulses (above a threshold) is computed. In this method it is assumed that the HRTFs are minimum-phase or nearly minimum-phase.

Of these approaches, the first one naturally gives the exact answer based on the IPD. The latter methods are sometimes used for deriving a frequency-independent ITD value. If the signals are not minimum-phase, the three above methods may give different results. For ITD approximation to be used in conjunction with a minimum-phase digital filter realization, the interaural phase characteristics of the HRTF filters should also be taken into account and compensated for. In the following, a method will be outlined that uses minimum-phase reconstruction and linear interaural excess phase approximation (discussed in, e.g., Jot et al. (1995)). This method has been used throughout the rest of this study for the approximation of ITDs for HRTF synthesis.

As discussed by Jot et al. (1995) and subsequently by Gardner (1997), the excess interaural phase response of the HRTFs (the term $\arg\{H_{ap}(e^{j\omega})\}$ in Eq. 5.4) can be modeled using linear regression over a frequency band. The desired excess phase response is first calculated by deconvolving the original HRTF with its minimum-phase counterparts. In the second step, a straight line is fitted to the interaural excess phase response over a frequency band (e.g., 500-2000 Hz). In Fig. 5.5, this minimum-phase and linear excess-phase approximation method is illustrated using the example HRTF pair. It can also be seen in the figure that the theoretical low- and high-frequency ITD models governed by Eqs. 5.10 and 5.11 fit well to measured data.

Comparison of Empirical and Analytical ITDs

When a computational ITD calculation method is compared to ITDs derived from HRTF measurements averaged from human subjects (Riederer, 1998a), it can be seen that the Woodworth model given in Eq. 5.12 matches quite well to the

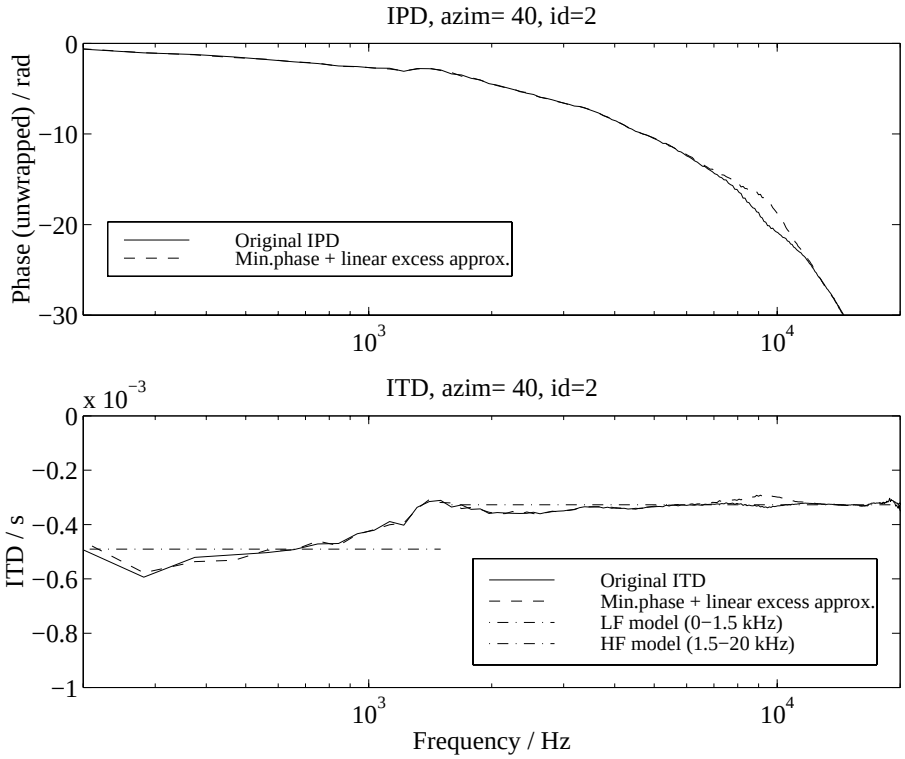


Figure 5.5: Estimation of the ITD using the minimum-phase and linear excess phase approximation method (Jot et al., 1995). Upper plot: Original IPD (solid line) and IPD approximation (dashed line). Lower plot: Original ITD (solid line), ITD approximation (dashed line), and low-frequency and high-frequency ITD models (dash-dot lines).

empirical data. The results for the horizontal plane (elevation=0°) are illustrated in Fig. 5.6. The offset of the ITDs around 90° and 270° suggests that the ears are placed a little asymmetrically, approximately at 100° and 260° azimuth, which is a commonly accepted fact (Blauert, 1997). For comparison, the results of high-frequency and low-frequency models given in Eqs. 5.10–5.11 are also depicted.

Elevation Dependency of the ITD

Elevation dependency modeling of the ITD has not been explored extensively in the literature (Larcher and Jot, 1997; Savioja et al., 1999). This has been mainly due to the lack of an appropriate model. Empirical results have shown, however, that the ITD is reduced when the elevation angle is increasing, as expected. The ITD curves have also been found to be approximately symmetrical with respect to 0° elevation (e.g., ±15° elevation ITD curves are of similar nature). The ITD measurement data has been found to fit well to a spherical head based ITD model

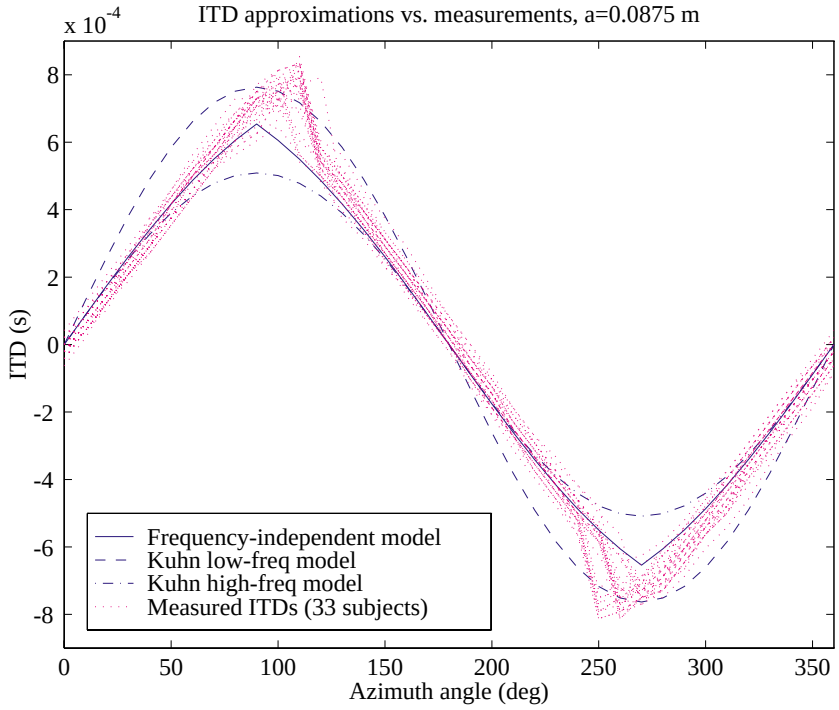


Figure 5.6: Comparison of ITD approximations to measured ITDs of 33 human subjects at 0° elevation. ITD computation for measured HRTFs was carried out using the cross-correlation method.

(given in Eq. 5.12). The elevation dependency of the ITD has been taken into account by adding a scaling term to the basic ITD equation:

$$\tau_{\text{elev}} = \frac{a(\sin\theta + \theta)}{2c}\cos\phi \tag{5.13}$$

where ϕ is the elevation angle. Another approximation method including the elevation angle has been proposed by (Larcher and Jot, 1997):

$$\tau_{\text{elev}} = \frac{a}{2c}\left(\text{asin}\left(\cos\phi\sin\theta\right) + \cos\phi\sin\theta\right) \tag{5.14}$$

A simple cosine dependency of the elevation angle was found to be accurate enough for simulation purposes. A plot illustrating the elevation dependency of ITDs for a test subject is shown in Fig. 5.7. The empirical frequency-independent ITD values were calculated using the cross-correlation method. It can be seen that a simple cosine model is able to model the ITD behavior in different elevations. An ellipsoidal head model that can be used to predict elevation ITD behavior has been presented in (Duda et al., 1999).

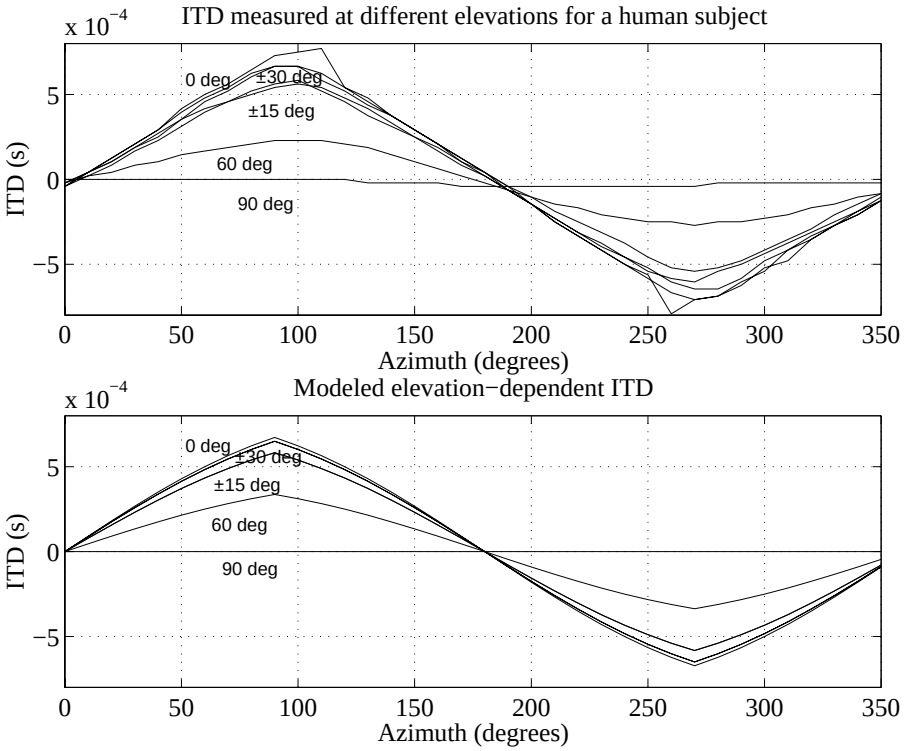


Figure 5.7: Measured and simulated ITDs at various elevation angles as a function of azimuth angle.

5.1.2 Interaural Level Difference and Spectral Cues

The interaural level difference (ILD) is the second major cue of Lord Rayleigh's duplex theory. It is mainly due to direction-dependent and frequency-dependent shadowing caused by the pinnae, the head, and the torso. The role of the ITD starts to decrease with frequencies above 2 kHz. This is due to the shortening wavelength of the incident sound and the resulting weaker phase change perception. Instead, the ILD dominates the localization at higher frequencies. The effective frequency range of the interaural pressure level difference (ILD) spans throughout the audible frequency range, although at lower frequencies the role of the ITD is dominant. Problems in the localization of sound occur mostly in the median plane, where the ITD and ILD are small or zero, and on the cone of confusion, where ITD and ILD cues are ambiguous. Although actual human heads are not spherical and the ears are asymmetrical in shape and placement, the proposed mechanisms causing localization mismatch are valid.

The behavior of ILD varies considerably as a function of incident angle and frequency. Source distance starts to effect the ILD below approximately 1m. Furthermore, since ILD cues are of individual nature, accurate modeling using mathematical principles is impossible. In the literature, models for the ILD have

been computed based on Rayleigh's formulas for diffraction around a rigid sphere (discussed in the context of source directivity modeling earlier in Chapter 3).

Interaural cues, ITD and ILD, both have the effect of causing lateral displacement to the position of the auditory event. Trading of interaural cues (also called time-intensity trading) has been applied to compensate, e.g., for asymmetrical listening positions. See, e.g., (Blauert, 1997) for a review on this issue.

The pinna cavities have an important role in adding frequency-dependent directional behavior to the sound field emitted to the ear canal. The role of spectral cues especially at high frequencies in median plane localization and elevation distinction is very important. Spectral cues are also dominant in monaural localization.

5.1.3 Other Cues for Sound Localization

In the previous section, the process of human sound localization was limited only to static case and without intersensory cues. It has, however, been shown that human sound localization may be enhanced, even exaggerated, by introducing additional cues to obtain a better performance, for example, in virtual environment simulation.

Intersensory Cues

The role of visual cues in the perception of spatial sound must not be underestimated. The visual cues often dominate over spatial hearing cues when perceiving the direction of the incident sound. The effect called ventriloquism is defined as the spatially biased perception of the auditory stimulus from the same point as the visual stimulus (Shinn-Cunningham et al., 1997, p. 620). It is very likely that ambiguous direction perception of sound sources has a great benefit of visual cues. Furthermore, visual cues may be an important factor in distance judgments of a sound source.

Room Acoustic Cues

Room acoustic phenomena have been found to influence and enhance localization. From room acoustical theory it is known that the early reflections add to spatial impression and perception of the acoustic space (Barron, 1993). In a small room, early reflections may resolve localization problems in the median plane or on the cone of confusion.

Head Movements

It is clear that the accuracy of sound localization is increased with the aid of head movements. Head movements are especially useful in solving ambiguous directional information, particularly in the "cone of confusion" where front-back and

up-down reversals most often occur. The primary cues for distinguishing front from back seem to be head movement and pinna response. A study conducted by Wightman and Kistler (1999) shows (and verifies earlier results) that user-induced head or source movements involving rotation provide significant reduction in the horizontal errors of directional localization. A similar study presented on free head movements during sound localization (Thurlow et al., 1967) showed that rotation movements of the head about the vertical axis (Z) were most commonly found alone, or in combinations with rotations about the interaural axis (Y) and the X-axis. Many subjects also showed reversal motions, that is, moving the head back and forth during sound localization. These results show that listener movement and head-tracking may be a significant factor to aid in the localization of virtual sources in headphone listening.

Super-Auditory Localization

Some applications that take advantage of 3-D sound benefit from models of exaggerated spatial hearing cues. In super-auditory localization, both the ITD and ILD cues can be modified to match an enlarged human head (Rabinowitz et al., 1993; Shinn-Cunningham et al., 1997). This results in exceeded localization performance that may be desired in virtual auditory displays designed for teleoperation and virtual environments.

5.1.4 Distance-Dependency of the HRTF

In this section the characterization of distance in HRTF models is studied – a factor that has often been neglected in spatial hearing research. In the far-field, at a distance larger than approximately 1 meter, distance effects can generally be neglected, but at closer ranges they play an important yet fairly unexplored role in localization. Related literature can be found from, e.g., (Rabinowitz et al., 1993; Brungart and Rabinowitz, 1996; Duda and Martens, 1997; Brungart et al., 1997; Calamia and Hixson, 1997; Brungart, 1998). In the experiments carried out by the author (Huopaniemi and Riederer, 1998), near-field and far-field HRTFs were measured and compared to analytical HRTF models. In the following, the experiment and model results are presented for the ITD and ILD.

The HRTF measurements were performed in the large anechoic chamber of the Laboratory of Acoustics and Audio Signal Processing at the Helsinki University of Technology (Riederer, 1998a). For the far-field measurements, Audax AT100M0 4-inch elements in plastic 190 mm spherical enclosures attached to an aluminum framework (covering seven elevation angles) were placed at a 2.0 m distance. In the near-field measurement, a smaller sound source (LPB 80 SC 3-inch element in a plastic 150 mm spherical enclosure) was at a distance of 0.7 m to the subject. For this study, measurements were carried out at 0° elevation and 10° azimuth increments. The spherical shape of the sound sources was considered as the

most optimal for a single clement loudspeaker casing with minimum diffraction. The size and shape of the element used for near-field measurements was found not to affect the measurement result at the distance of 0.7 m; at a closer range the point-source approximation would not have been valid anymore[3]. A total of eight human subjects and one dummy head (Cortex MK II) were measured in this study in both near-field and far-field setups. The test person was placed in a measurement chair that was fixed to a turntable. The ear canals of the subject were blocked with moldable silicon putty, to which two Sennheiser KE-41 l-2 miniature microphones were attached. The measurements were fully computer controlled.

Investigating the ITD as a Function of Distance

The question of whether the ITD changes as a function of distance has been addressed in several publications (Brungart and Rabinowitz, 1996; Duda and Martens, 1997). In order to verify the results presented elsewhere, a simple study was conducted using the 9 measured subjects' HRTFs and comparing those ITDs to the ones computed using a structural model (presented in Section 3.2.1). The measurement results were compared to analytical data based on a spherical head HRTF model (Rabinowitz et al., 1993; Duda and Martens, 1998). In Fig. 5.8 it can be seen that the ITD varies only slightly when the source moves from far-field to near-field, as has been stated in previous articles (Brungart et al., 1997; Duda and Martens, 1998).

Investigating the ILD as a Function of Distance

As previously shown and discussed in the literature (Brungart and Rabinowitz, 1996; Duda and Martens, 1997), the ITD cue differences as a function of distance (except for very close distances) are not perceptually important. It is, however, not the case in ILD behavior.

Results for modeling HRTFs from two observation distances in the free-field are shown in Figs. 5.9–5.10 (based on (Duda and Martens, 1998)). From these figures it can be clearly seen that the lower frequencies appear boosted in the ipsilateral (same) side and damped in the contralateral (opposite) side at a close distance when compared to far-field. At higher frequencies the tendency is similar. In these figures the angle values (on the right side of the plots) are those of the incidence from the source to the observation point (the ear) on the surface of the sphere (not to be confused with absolute azimuth angle used elsewhere). In order to study the ILD measurements as a function of distance, the following procedure for the HRTFs was carried out. For near-field and far-field measurements on

[3] Note that the near-field effects in HRTFs at the distance of 0.7 m are not expected to be significant. This has been shown by (Duda and Martens, 1998; Brungart, 1998). This distance was chosen due to practical reasons caused by the measurement setup.

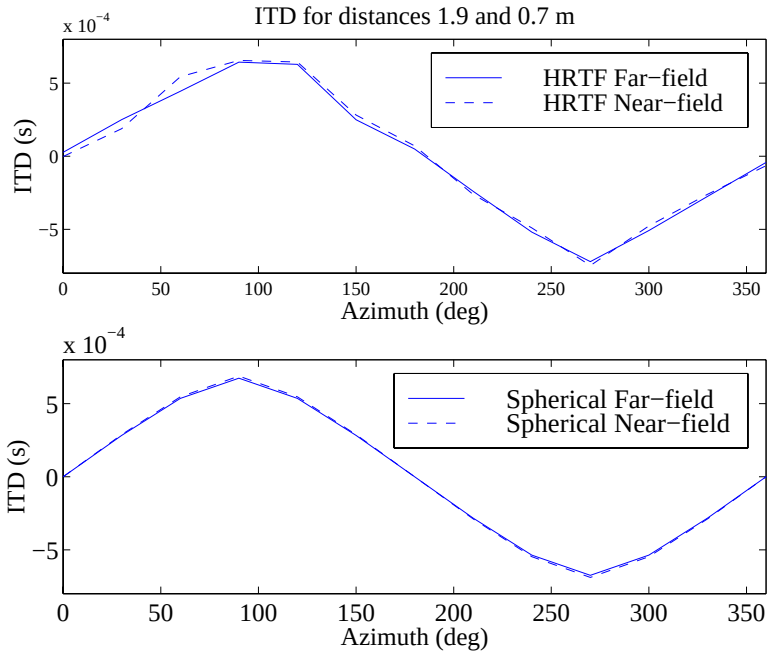


Figure 5.8: ITD plots in far-field and near-field for HRTFs measured on human subjects (top graph, mean of 9 subject ITDs) and a spherical head model (bottom graph).

human subjects, the mean of the magnitude response for both cases at 30 degree azimuth angle intervals was calculated. The ILD was thus computed using the following formula

$$ILD_{azi}(\omega) = 20\log\frac{\sum_{i=1}^{N}|\text{fft}\left[hrtf_{r,azi}(i,\omega)\right]|}{N} - 20\log\frac{\sum_{i=1}^{N}|\text{fft}\left[hrtf_{l,azi}(i,\omega)\right]|}{N} \tag{5.15}$$

where N is the number of measured subjects (in this case, $N = 9$), azi is the azimuth angle, and ω is the frequency variable. In Figs. 5.11–5.14, ILD results based on measurements are depicted and compared to those based on a Rayleigh diffraction model. In the upper plots, the ILDs for near-field and far-field for the measured HRTFs and the analytical model are plotted. In the lower plots, the change in ILD as the source moves to the near-field is plotted for the measured HRTFs and the analytical model. From the figures an increase in ILD can be seen at closer distance when the source moves to the side (in both ipsilateral and contralateral cases). The trend of ILD increase can be seen in the figures. As the source comes closer, the ILD differences increase rapidly as shown by Brungart (1998); Duda and Martens (1998). The analytical model is shown to predict the ILD increase well (when moving the source from the 2.0 m distance to 0.7 m).

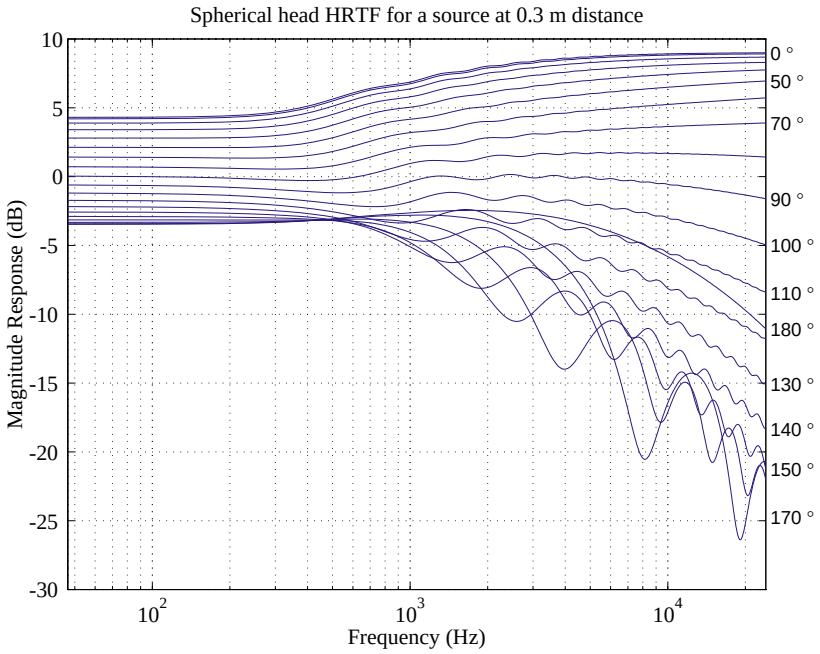


Figure 5.9: HRTFs computed from a Rayleigh diffraction model (Duda and Martens, 1998). Source distance: 0.3m.

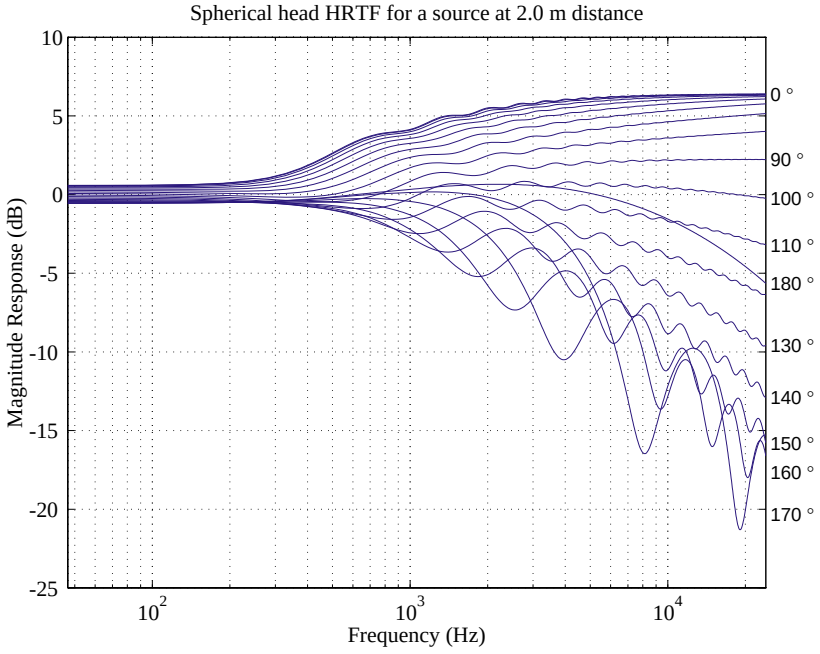


Figure 5.10: HRTFs computed from a Rayleigh diffraction model (Duda and Martens, 1998). Source distance: 2.0m.

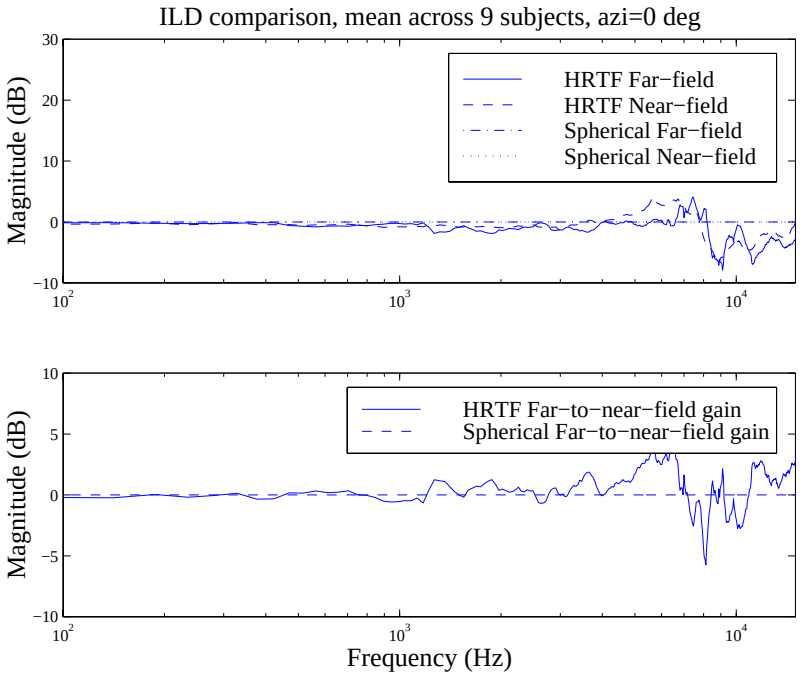


Figure 5.11: Near-field and far-field interaural level differences at 0°.

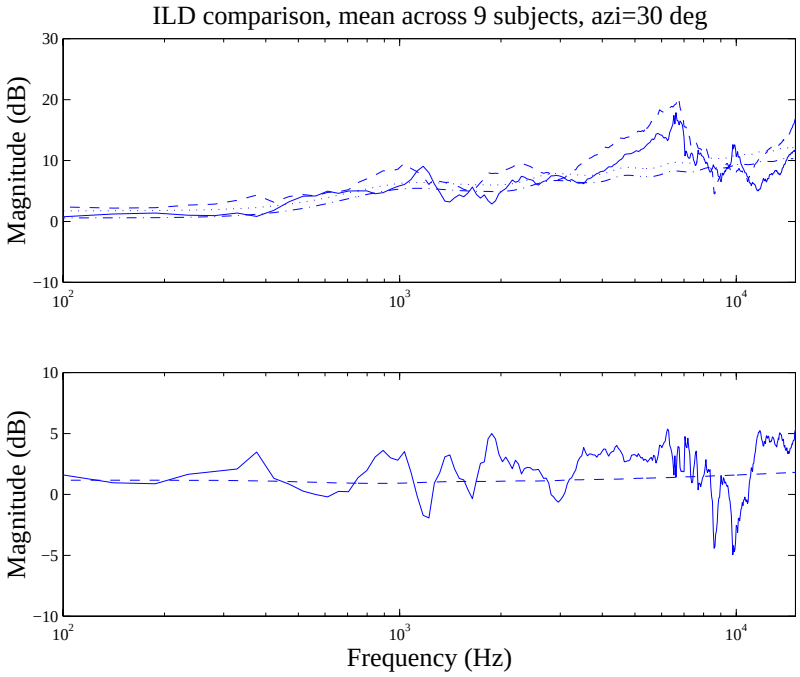


Figure 5.12: Near-field and far-field interaural level differences at 30°.

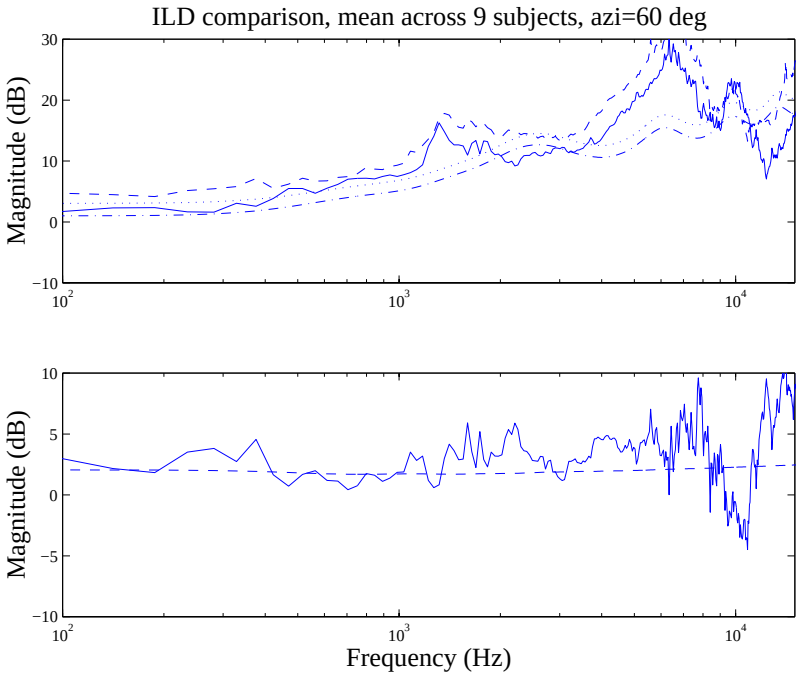


Figure 5.13: Near-field and far-field interaural level differences at 60°.

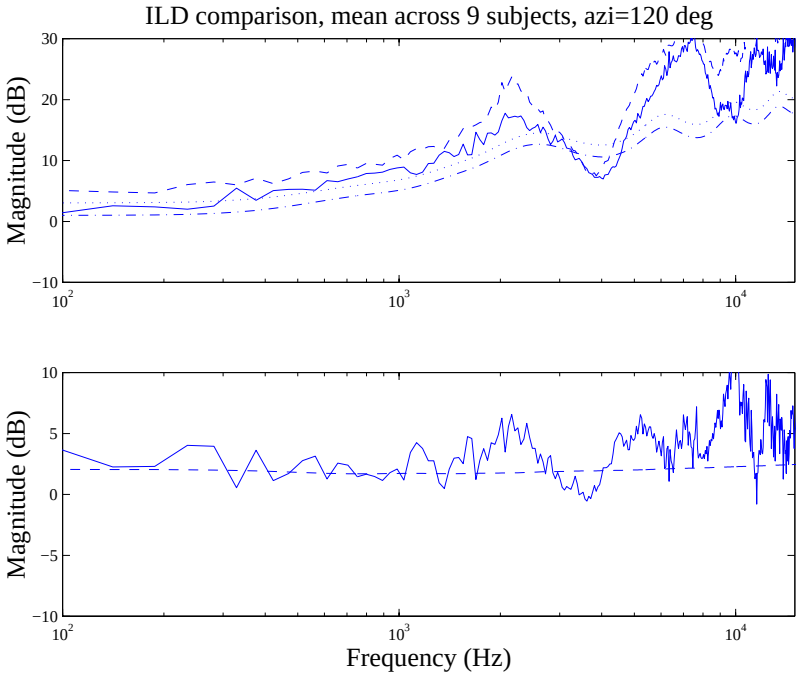


Figure 5.14: Near-field and far-field interaural level differences at 120°.

5.1.5 Functional Modeling of HRTFs

Apart from a signal-modeling approach to HRTFs, interest in functional representations of HRTFs has also risen over the past years in search of efficient auralization techniques. These methods are not directly related to specific filter design issues, but can serve as a basis for, e.g., structural smoothing and preprocessing of the data. One of the main goals in this approach is the decomposition of HRTFs to components that enable computational or memory reduction in the synthesis stage. Principal components analysis (PCA) is one of the best-known decomposition methods used for reorganizing data. In the decomposition, the data is reorganized into basis functions and weight vectors that produce an ordered model of the effective variance of the sample space. Using this decomposition, perfect resynthesis of the data can be accomplished. Remarkable data reduction can also be achieved if only the first basis functions (explaining often 90-95% of the variance) are used.

Principal components analysis (PCA) for HRTF modeling was first proposed by Martens (1987). It has been since been studied by, e.g., Kistler and Wightman (1992); Middlebrooks and Green (1992); Blommer (1996). Kistler and Wightman (1992) used PCA to approximate minimum-phase diffuse-field HRTFs (that they called directional transfer functions, DTFs). In this method the magnitude spectra of the HRTFs were approximated using five principal spectral components of the response. With this method the order of the resulting FIR filters was successfully reduced to 1/3 of the original impulse response with only a slight decrease in localization accuracy.

Chen et al. (1995) have proposed a feature extraction method, where a complex valued HRTF was represented as a weighted sum of eigentransfer functions generated using the Karhunen-Loéve expansion (KLE)[4]. The difference from the previous PCA model is that a complex HRTF transfer function including magnitude and phase information can be modeled. In the method proposed by Evans et al. (1998), functional representations for a set of HRTF measurements have been developed. The HRTFs were represented as a weighted sum of surface spherical harmonics that are a hierarchical set of basic functions orthogonal upon the surface of a sphere. The reconstruction based on surface spherical harmonics yielded complex frequency responses that model HRTF data with good accuracy. Computational efficiency when compared to the KLE has, however, been found inferior (Evans et al., 1998). An attractive method suitable for real-time implementation of virtual acoustic displays (VAD) was presented in (Abel and Foster, 1997, 1998). In this patent, a method utilizing singular value decomposition (SVD) was used to derive an orthogonal set of functions for a set of HRTFs. Thus, a linear combination of these functions (basis impulse responses, or complex HRTF basis spectra modeled by digital filters) can be used to synthesize any HRTF in the set with reasonable computational cost, in a similar manner

[4] The PCA and KLE methods are equivalent in this context (Blommer, 1996).

as proposed by previous studies (e.g., (Martens, 1987; Kistler and Wightman, 1992)). In addition to decomposition methods, structural analysis and modeling of the HRTF can yield efficient approximations suitable for real-time processing (Brown and Duda, 1998).

5.2 Auditory Analysis of HRTFs and Application to Filter Design

In this section, binaural processing is approached from the auditory perception point of view. The properties of human peripheral hearing serve as a starting point for auditorily motivated binaural signal processing and filter design. It is proposed by the author that binaural filter design should be carried out using auditory criteria. Similar conclusions have been made by Kulkarni and Colburn (1997b).

5.2.1 Properties of the Human Peripheral Hearing

From the psychoacoustical point of view, the linear frequency scale used in linear transforms such as the Fourier transform is not optimal. In psychoacoustics it has been shown experimentally that scales such as the *Bark* scale (or the critical-band rate scale) (Zwicker and Fastl, 1990) or the *ERB* (Equivalent Rectangular Bandwidth) rate scale (Moore et al., 1990) match closely the properties of human hearing. Moreover, presently the ERB scale is believed to be theoretically better motivated than the Bark scale (Moore et al., 1996). Approximation formulas for the psychoacoustic scales are the following. The ERB scale bandwidth as a function of center frequency f_c (in kHz) is given by the following equation (Moore et al., 1990):

$$\Delta f_{\mathrm{CE}} = 24.7(4.37 f_c + 1). \tag{5.16}$$

Similarly, the Bark scale can be approximated by the equation (Zwicker and Fastl, 1990):

$$\Delta f_{\mathrm{CB}} = 25 + 75\left(1 + 1.4 f_c^2\right)^{0.69}. \tag{5.17}$$

In Fig. 5.15, four resolution functions are compared. The plotted functions are linear, logarithmic, Bark, and ERB rate scales vs. log frequency (Huopaniemi and Karjalainen, 1997). As can be seen from Fig. 5.15, the log and the ERB resolution functions are relatively close to each other. The Bark resolution is similar above 500 Hz. The constant bandwidth resolution function related to the linear frequency scale is generally not acceptable when characterizing responses from the auditory point of view. Based on modern psychoacoustic theory (Moore et al., 1990), a conclusion may be drawn that the auditory resolution is best

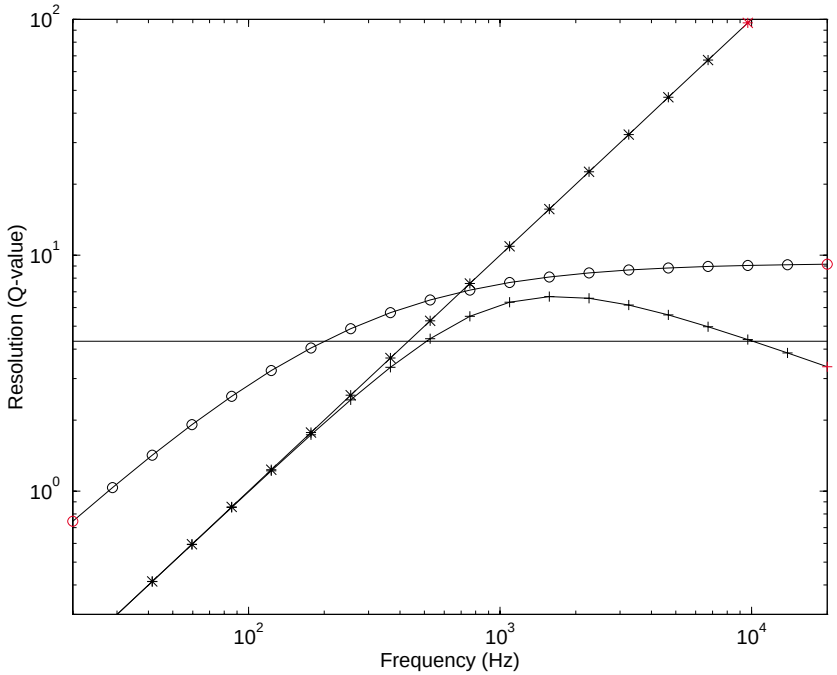


Figure 5.15: Frequency resolution (Q-value) curves as functions of frequency: Solid line = third-octave (constant Q); o-o = ERB resolution; +-+= Bark (critical band) scale; *-* = constant 100 Hz bandwidth (linear resolution)(Karjalainen et al., 1999).

represented using the ERB scale, the logarithmic and the Bark scales being useful approximations, and the linear scale being inferior. The question whether the monaural psychoacoustic principles apply in the same way to binaural hearing is an open one. Thus, care should be taken when using (monaural) psychoacoustical measures in binaural design.

There are different methods of incorporating the non-uniform resolution of hearing into a filter design. The approaches have been divided into three different categories:

- auditory smoothing of the responses prior to filter design
- use of a weighting function in filter design
- warping of the frequency axis prior or during the filter design

In the following, these three approaches will be discussed.

5.2.2 Auditory Smoothing

In many cases, the data (in this case the HRTF frequency-domain or time-domain responses) may be preprocessed using a smoothing technique prior to the filter

design (Abel, 1997). The most straightforward way to accomplish a frequency-dependent magnitude response smoothing is to use moving averaging of variable window size (Smith, 1983; Köring and Schmitz, 1993):

$$|H_S(f)| = \sqrt{\frac{1}{f_1 - f_0} \int_{f_0}^{f_1} |H(f')|^2 \, df'}, \tag{5.18}$$

where $f_1 - f_0$ is the window width at frequency f'. Furthermore, if we define $f_0 = f/\sqrt{K}$ and $f_1 = f\sqrt{K}$, it follows that for third-octave width, $K = 5/4$ and for octave width, $K = 2$ (Köring and Schmitz, 1993). The window size as a function of frequency can also be derived using auditory criteria, based on the Bark or ERB scale. In these smoothing techniques the window widths at frequencies f' are derived using Eqs. (5.16–5.17).

Another technique for HRTF smoothing that has been considered is *cepstral smoothing* (Huopaniemi et al., 1995; Kulkarni and Colburn, 1997a). This technique is widely used in speech processing (Oppenheim and Schafer, 1989) and is also closely related to minimum-phase reconstruction of impulse-responses via cepstral analysis, described in Section 5.1. Cepstral smoothing can be understood as linear smoothing of the log magnitude response. Because it does not involve a non-uniform frequency resolution, it is not strictly an "auditory smoothing" technique. In Figures 5.16–5.17, the effects of different smoothing techniques applied to an HRTF magnitude response are illustrated. The HRTF data in these figures are based on real head measurements (Riederer, 1998a).

5.2.3 Auditory Weighting

An alternative to auditory smoothing techniques is to use weighting functions in the filter design that approximate the human auditory resolution. Approximation formulas presented in Eqs. (5.16) and (5.17) have been used to calculate the weighting functions shown in Fig. 5.18. The weight of each frequency point is the inverse of the bandwidth calculated with Eqs. (5.16) and (5.17). As can be seen from Fig. 5.18, the ERB scale weighting function focuses more at low and high frequencies (below 500 and above 4000 Hz) than the corresponding Bark function, but remains very close to the Bark scale at mid-frequencies.

5.2.4 Frequency Warping

It has been proposed earlier in this section that modeling and filter design of HRTFs should be carried out using psychoacoustical criteria. One attractive possibility to approximate a non-linear frequency resolution is to use *frequency warping*. Approximations of HRTFs using frequency warping have not been extensively studied. Jot et al. (1995) have proposed a method where the HRTFs are preprocessed using auditory smoothing and the IIR filter design is carried

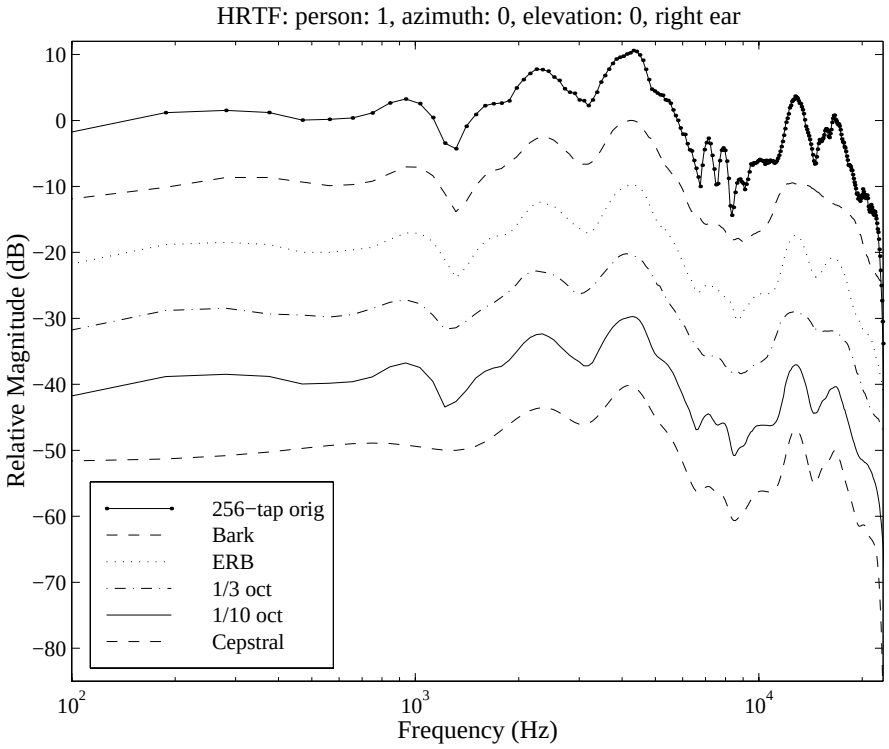


Figure 5.16: Different smoothing techniques applied to HRTF magnitude response approximation.

out using a Yule-Walker algorithm (Friedlander and Porat, 1984) in the warped frequency domain[5]. A framework for warped binaural filter design has been established by the author (Huopaniemi and Karjalainen, 1996b,a, 1997; Huopaniemi et al., 1998, 1999b). The fundamentals of warped filters are studied in the following.

Frequency scale warping is in principle applicable to any design or estimation technique. The most popular warping method is to use the bilinear conformal mapping (Strube, 1980; Smith, 1983; Smith and Abel, 1995; Karjalainen et al., 1997b; Smith and Abel, 1999). The bilinear warping is realized by substituting unit delays with first-order allpass sections:

$$z^{-1} \leftarrow D_1(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}, \tag{5.19}$$

where λ is the warping coefficient. This implies that the frequency-warped version of a filter can be implemented by such a simple replacement technique. It is easy to show that the inverse warping (*unwarping*) can be achieved with a similar substitution but using $-\lambda$ instead of λ (Smith, 1983; Jot et al., 1995). Warping curves for different values of λ are illustrated in Fig. 5.19.

[5] This procedure was originally suggested in (Smith, 1983) and applied for modeling of the violin body.

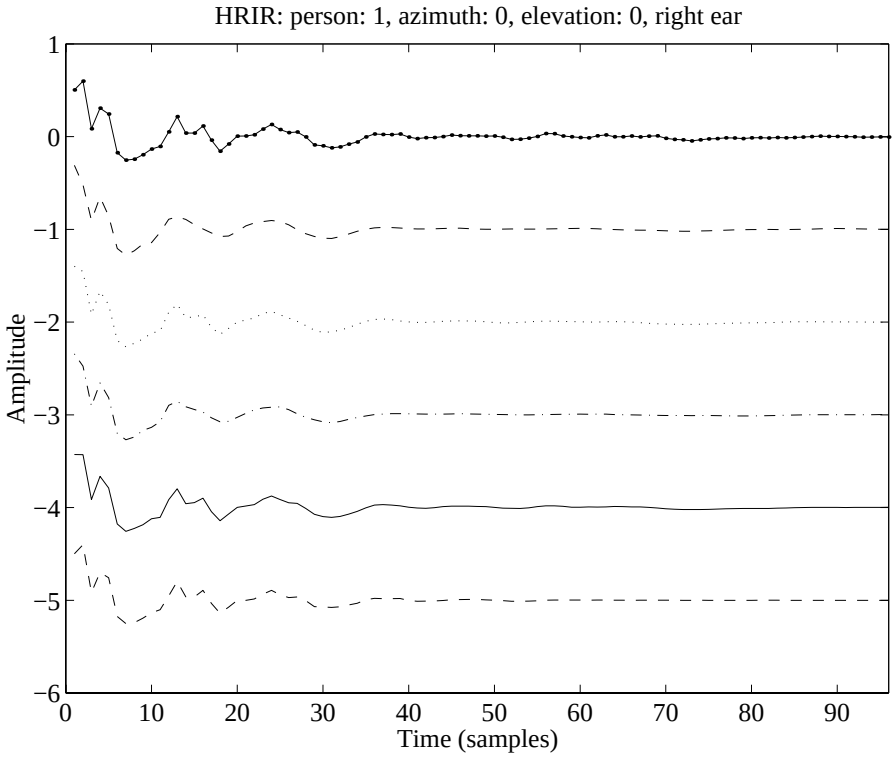


Figure 5.17: Minimum-phase time-domain HRIRs of smoothed magnitude responses shown in Fig. 5.16.

The usefulness of warping by conformal mapping comes from the fact that, given a target transfer function $H(z)$, it may be possible to find a lower order warped filter $H_w(D_1(z))$ that is a good approximation of $H(z)$ in an auditory sense. $H_w(D_1(z))$ is then designed in a warped frequency domain so that using allpass delays instead of unit delays maps the design to a desired transfer function in the ordinary frequency domain. For an appropriate value of λ, the bilinear warping can fit the psychoacoustic Bark scale, based on the critical band concept (Zwicker and Fastl, 1990), surprisingly accurately (Smith and Abel, 1995). An approximate formula for the optimum value of λ as a function of sampling rate is given in (Smith and Abel, 1995)[6]. For a sampling rate of $f_s = 48$ kHz this yields $\lambda = 0.7364$, and for $f_s = 32$ kHz $\lambda = 0.6830$ (Smith and Abel, 1999).

5.3 Digital Filter Design of HRTFs

The task of approximating an ideal HRTF response $H(z)$ by a digital filter $\hat{H}(z)$ is studied in this section. For any filter design, the goal is to minimize an error

[6]Revised version and formula has been published in (Smith and Abel, 1999).

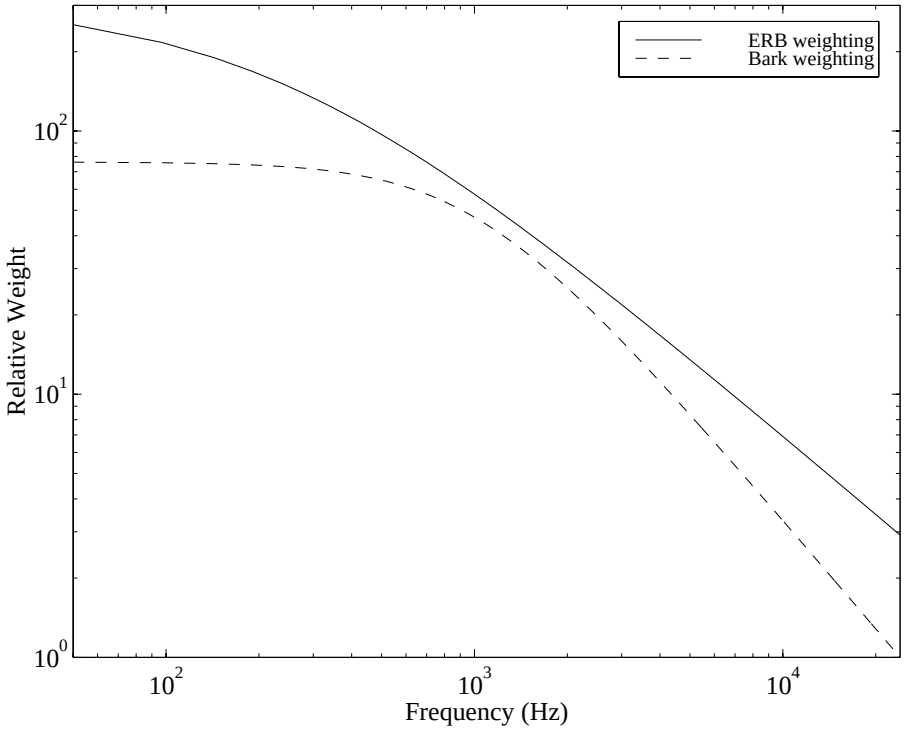


Figure 5.18: Auditory weighting functions as a function of frequency, $f_s = 48$ kHz.

function that is of the form:

$$E(e^{j\omega}) = H(e^{j\omega}) - \hat{H}(e^{j\omega}) \tag{5.20}$$

A list representing research carried out in the field of HRTF approximation by various authors is illustrated in Table 5.1. The filter order values have been collected from the corresponding publications and represent FIR filter orders (number of taps - 1) or IIR filter orders (number of poles or zeros). One can see from the table that the results from different studies vary considerably. There are many reasons for this. Some of the studies are purely theoretical, meaning that the results are formulated in the form of a spectral error measure or by visual inspection. In some of the references, the authors claim that a certain filter order appeared to be satisfactory in informal listening tests. These cases are marked in the table with a question mark. There have been very few formal listening tests in this field that also give statistically reliable results (Sandvad and Hammershøi, 1994a; Huopaniemi and Karjalainen, 1997; Kulkarni and Colburn, 1997a,b). Another question is the validity of the HRTF data used in the studies. The HRTFs may have been equalized for free-field conditions or a certain headphone type, or the data may have have been based on different HRTF measurements (dummy-head, individual/non-individual data). Different preprocessing such as minimum-phase reconstruction may also have been applied. All

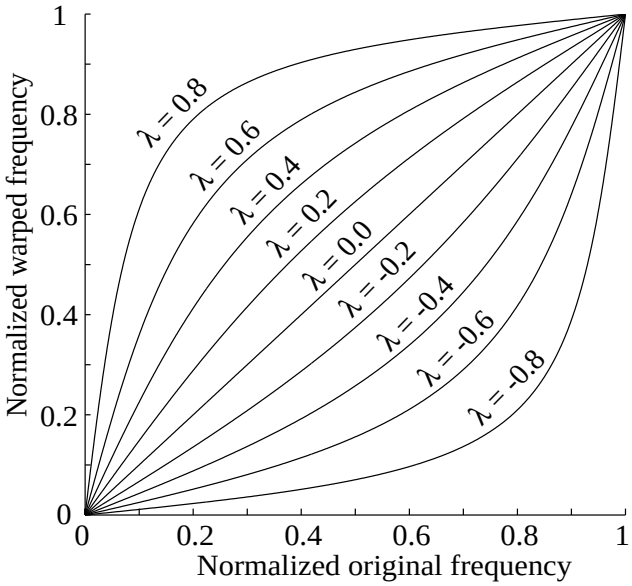


Figure 5.19: Frequency warping characteristics of the first-order allpass section $D_1(z)$ for different values of λ. Frequencies are normalized to the Nyquist rate (Karjalainen et al., 1999).

these aspects could cause the large deviation seen in the results of the studies (Table 5.1).

5.3.1 HRTF Preprocessing

The sound transmission in an HRTF measurement includes characteristics of many subsystems that are to be compensated in order to achieve the desired response. The transfer functions of the driving loudspeaker, the microphone, and the ear canal (if the measurement position were inside an open ear canal) may then have to be equalized. If, however, a more general database of HRTFs is desired, other equalization strategies should be considered like equalization with respect to a given reference direction, or diffuse-field equalization. In general, HRTF preprocessing strategies can be divided into (Huopaniemi and Smith, 1999):

- Temporal preprocessing
- Spectral preprocessing

By temporal preprocessing we normally consider minimum-phase reconstruction and separate modeling and implementation of the ITD (discussed in Section 5.1.1). Spectral preprocessing comprises auditory spectral preprocessing (smoothing, weighting, and warping as discussed in the previous section) and equalization techniques.

Research Group	**Design Type**	**Filter Order**	**Study**
Begault, 1991	Binaural / FIR	81-512	Empirical
Sandvad and Hammershoi, 1994ab	Binaural / FIR	72	Empirical
Kulkarni and Colburn, 1995, 1997	Binaural / FIR	64	Empirical
Hartung and Raab, 1996	Binaural / FIR	48	Empirical
Asano et al., 1990	Binaural / IIR	>40	Empirical
Sandvad and Hammershoi, 1994ab	Binaural / IIR	48	Empirical
Blommer and Wakefield, 1994	Binaural / IIR	14	Theoretical
Jot et al., 1995	Binaural / IIR	10-20	Empirical?
Ryan and Furlong, 1995	Binaural / IIR	24	Empirical?
Kulkarni and Colburn, 1995, 1997	Binaural / IIR	6	Empirical?
Kulkarni and Colburn, 1995, 1997	Binaural / IIR	25 (all-pole)	Empirical?
Hartung and Raab, 1996	Binaural / IIR	34/10	Empirical
Mackenzie et al., 1996	Binaural / IIR	10	Theoretical
Blommer and Wakefield, 1997	Binaural / IIR	40	Theoretical

Table 5.1: Binaural HRTF filter design data from the literature.

Equalization Methods

Equalization methods of head-related transfer functions have been summarized in (Møller, 1992; Blauert, 1997; Larcher et al., 1998). The main reason to equalize the data is to compensate for the response of the measurement or reproduction systems. For example, one could aim at designing a general database of HRTFs for diffuse-field equalized headphones. There are two main methods for HRTF equalization:

- Free-field equalization
- Diffuse-field equalization

Free-field equalization is achieved by dividing the measured HRTF by a reference measured in the same ear from a certain direction. The reference direction is typically chosen as 0° azimuth and 0° elevation, that is, from the front of the listener (Møller, 1992; Jot et al., 1995). For clarity, only the magnitude spectrum equalization is considered here; the phase is assumed to be minimum-phase reconstructible.

$$|H_{\mathrm{ff}}(\omega,\theta,\phi)| = \frac{|H(\omega,\theta,\phi)|}{|H(\omega,\theta=0^\circ,\phi=0^\circ)|} \tag{5.21}$$

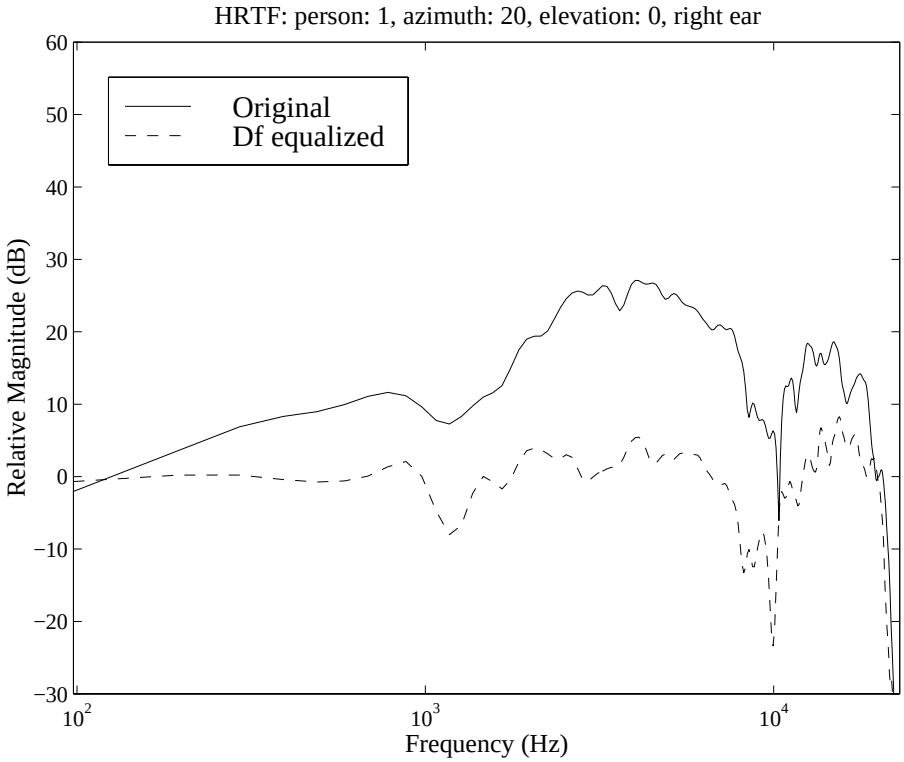


Figure 5.20: Illustration of unequalized and diffuse-field equalized HRTF magnitude responses.

where θ is the azimuth angle, ϕ is the elevation angle, and ω is the angular frequency. Diffuse-field HRTFs, on the other hand, are estimated by power-averaging all HRTFs from each ear. The equalized HRTFs are obtained by dividing the original measurement by the diffuse-field HRTF of that ear.

$$|H_{\mathrm{df}}(\omega,\theta,\phi)| = \frac{|H(\omega,\theta,\phi)|}{\sqrt{\frac{1}{M}\sum_{i=1}^{M}|H_i(\omega,\theta,\phi)|^2}} \tag{5.22}$$

where M is the total number of HRTFs in the set. In the diffuse-field normalized responses a flattening of the HRTF spectra is normally achieved. This is due to the removal of spectral factors that are not incident-angle dependent (such as the ear canal resonance, if present in the measurements). In Fig. 5.20, the difference between a normal (measurement equalized) HRTF and a diffuse-field equalization is illustrated. Flattening of the spectra and reduced dynamic range is clearly visible. This suggests that the needed filter order for perceptually correct design should also be lower, as will be seen in further sections.

5.3.2 Error Norms

Let us look at the problem of finding an optimal filter approximation to a given HRTF. A common way to quantify errors in digital filter design is to use the L_p-norms (Golub and Loan, 1996; Smith, 1983). An L_p error norm of a frequency-domain representation of an arbitrary filter $H(e^{j\omega})$ and its approximation $\hat{H}(e^{j\omega})$ is defined by the following equation:

$$\|E\|_p = \|H(e^{j\omega}) - \hat{H}(e^{j\omega})\|_p \triangleq \left(\int_{-\pi}^{\pi} \left| H(e^{j\omega}) - \hat{H}(e^{j\omega}) \right|^p \frac{d\omega}{2\pi} \right)^{\frac{1}{p}}, p \geq 1 \quad (5.23)$$

Least-Squares Norm

The L_2 norm is generally known as the *least squares* (integrated squared magnitude) norm. It is particularly appealing due to the fact that by using Parseval's relation this norm can directly be used both in the frequency and the time domains (Smith, 1983; Parks and Burrus, 1987):

$$\begin{aligned} E_2 &= \|H(e^{j\omega}) - \hat{H}(e^{j\omega})\|_2^2 \\ &\triangleq \int_{-\pi}^{\pi} \left| H(e^{j\omega}) - \hat{H}(e^{j\omega}) \right|^2 \frac{d\omega}{2\pi} \\ &= \sum_{n=0}^{\infty} \left| h(n) - \hat{h}(n) \right|^2 \triangleq \|h(n) - \hat{h}(n)\|_2^2 \end{aligned} \quad (5.24)$$

If a positive weighting function $W\left(e^{j\omega}\right)$ is further employed, a general *weighted least squares* (WLS) complex approximation problem in the frequency domain can be written as:

$$E_{w2} \triangleq \int_{-\pi}^{\pi} W\left(e^{j\omega}\right) \left| H(e^{j\omega}) - \hat{H}(e^{j\omega}) \right|^2 \frac{d\omega}{2\pi} \quad (5.25)$$

This property is desirable since it is possible to incorporate weighting functions into the filter design and optimization in order to model, e.g., the auditory frequency resolution. Another advantage of the least squares formulation is that since the error term is quadratic, there is a global minimum. Popular methods using least-squares error norm minimization include the equation-error method, Prony's method, and the Yule-Walker methods (Smith, 1983).

Chebyshev Norm

The L_∞ norm is often referred to as the *Chebyshev norm*. This norm, as p approaches infinity, aims to minimize the maximum component of the error term E (hence the term minimax):

$$\|E\|_\infty = \|H(e^{j\omega}) - \hat{H}(e^{j\omega})\|_\infty \triangleq \max_{-\pi<\omega<\pi} \left| H(e^{j\omega}) - \hat{H}(e^{j\omega}) \right| \quad (5.26)$$

The Chebyshev norm minimization may be an applicable choice for modeling HRTFs, since the magnitude responses typically exhibit prominent peaks and valleys that are important for localization. Since the minimax approach seeks to minimize the maximum error, which is often found at the transitions in the response, good performance in HRTF modeling would be expected. Another property of human hearing, the *logarithmic* amplitude resolution, can also be incorporated in a Chebyshev norm minimization algorithm (a weighted log-magnitude spectral error (Smith, 1983)). A drawback of the Chebyshev norm optimization is, however, that the error surface may not always be convex and the result can be unstable and/or the search process only locally optimal. Well-known methods for Chebyshev norm optimization are, e.g., the Remez multiple exchange algorithm (used in the Parks-McClellan algorithm) and simplex methods (Steiglitz, 1996).

Hankel Norm

Hankel norm based methods are attractive for HRTF modeling since they provide a general and stable solution to complex frequency response modeling. The Hankel norm lies between L_2 and L_∞ norms and is quantified by the spectral norm of the *Hankel matrix* of a given response. The most straightforward technique for obtaining an optimal Hankel norm approximation is to find the Hankel singular values of a Hankel matrix (by singular value decomposition) and select a desired number of the singular values to construct the final filter (Smith, 1983; Mackenzie et al., 1997). In this work, two Hankel norm filter design methods have been studied: balanced model truncation (BMT) (Mackenzie et al., 1997) and the Caratheodory-Fejer (CF) method (Gutknecht et al., 1983).

5.3.3 Finite Impulse-Response Methods

The most straightforward ways to approximate HRTF measurements are to use the window method for FIR filter design or a direct frequency sampling design technique (Parks and Burrus, 1987). If $H(e^{j\omega})$ is the desired frequency response, the direct frequency sampling method is carried out by sampling N points in the frequency domain and computing the inverse discrete Fourier transform (normally using the FFT):

$$h(n) = \frac{1}{N} \sum_{k=0}^{N-1} H(e^{j\omega_k}) e^{j\omega_k n}, \tag{5.27}$$

where $\omega_k = \frac{2\pi}{N} k$, and $k = 0, \ldots, N-1$. A time-domain equivalent to uniform frequency-sampling is windowing (truncating) the sampled impulse response $h(n)$ with a rectangular window function $w(n)$. Generally, the window method can be presented as a multiplication of the desired impulse response with some window

function $w(n)$:

$$\hat{h}(n) = h(n)w(n). \tag{5.28}$$

The use of the rectangular window is known to minimize a truncated time-domain L_2 norm (Oppenheim and Schafer, 1989). By Parseval's relation, then, the frequency response is also optimal in the unweighted least squares sense over the spectral samples ω_k. The effect of different window functions in HRTF filter design has been discussed by Sandvad and Hammershøi (1994a). They concluded that although rectangular windowing provokes the Gibbs' phenomenon, seen as ripple around amplitude response discontinuities, it is still favorable when compared to, e.g., the Hamming window. Another and perhaps more intuitive conclusion is based on the fact that HRTFs do not contain spectral discontinuities but rather broad fluctuations in the magnitude response, which can be modeled effectively using least squares fitting. A severe limitation of the windowing method is, however, the lack of a frequency-domain weighting function that could model the non-uniform frequency resolution of the ear. For an extended frequency-sampling method yielding an LS solution, it is possible, however, to introduce non-uniform frequency sampling and weighting (Parks and Burrus, 1987).

Kulkarni and Colburn (1995a, 1997a) have proposed the use of a weighted least squares technique based on log-magnitude error minimization for finite-impulse response HRTF filter design. They claim that an FIR filter of order 64 is capable of retaining most spatial information.

In a method proposed by Hartung and Raab (1996), binaural filters were optimized using auditory criteria, approximating the sensitivity of the human ear with a logarithmic magnitude weighting function. Non-uniform sampling of the frequency grid was applied in order to achieve auditory resolution. Optimization was then carried out to yield the following results: an FIR filter of order 48 "revealed no significant differences in localization performance... Minor divergence was noted with the 32nd-order FIR filter" (Hartung and Raab, 1996).

Issues in FIR filter design using auditory criteria have been discussed by Wu and Putnam (Wu and Putnam, 1997). They derived a perceptual spectral distance measure using a simplified auditory model. The technique was applied to HRTF-like magnitude responses and as a result FIR approximations of order 20 were successfully calculated.

5.3.4 Infinite Impulse-Response Methods

The earliest HRTF filter design experiments using pole-zero models were carried out by Kendall and Rodgers (1982); Kendall and Martens (1984). A comparison of FIR and IIR filter design methods has been presented in (Sandvad and Hammershøi, 1994a,b). The non-minimum-phase FIR filters based on individual HRTF measurements were designed using rectangular windowing. The IIR

filters were generated using a modified Yule-Walker algorithm that performs least-squares magnitude response error minimization. The low-order fit was enhanced a posteriori by applying a weighting function and discarding selected pole-zero pairs at high frequencies. Listening tests showed that an FIR of order 72 equivalent to a 1.5 ms impulse response was capable of retaining all of the desired localization information, whereas an IIR filter of order 48 was needed for the same localization accuracy.

In the research carried out by Blommer and Wakefield (1994), the error criteria in the auto-regressive moving average (ARMA) filter design were based on log-magnitude spectrum differences rather than magnitude or magnitude-squared spectrum differences. Furthermore, a new approximation for the log-magnitude error minimization was defined. The theoretical study concluded that it was possible to design low-order HRTF approximations (the given example used 14 poles and zeros) using the proposed method. In more recent studies (Blommer, 1996; Blommer and Wakefield, 1997), the results have been generalized, and they concluded that pole-zero models of order 40 were needed for accurate modeling of HRTFs. They also stated that a least squares (LS) algorithm was inferior in comparison to using a logarithmic error measure.

Asano et al. (1990) have investigated sound localization in the median plane. They derived IIR models of different orders (equal number of poles and zeros) from individual HRTF data using the least-squares error criterion (equation-error method). When compared to a reference, a 40th-order pole-zero approximation yielded good results in the localization tests with the exception of increased front-back confusions in frontal incident angles.

Other IIR approximation models for HRTFs have been presented by Ryan and Furlong (1995); Jenison (1995); Hartung and Raab (1996); Kulkarni and Colburn (1995b, 1997b). In the following, a novel IIR modeling technique based on balanced model truncation (Mackenzie et al., 1997) is discussed.

Balanced Model Truncation

In this section, a new low-order filter design technique for HRTFs is presented. An attractive technique for HRTF modeling has been proposed in (Mackenzie et al., 1997). By using this *balanced model truncation* (BMT) it is possible to approximate HRTF magnitude and phase response in a Hankel norm sense with low order IIR filters (down to order 10). A complex HRTF system transfer function is written as a state-space difference function (see, e.g., (Proakis and Manolakis, 1992)), which is then represented in balanced matrix form (Moore, 1981). A truncated state-space realization $F_m(z)$ corresponds to the original system transfer function $F(z)$ with a similarity to the original system which is quantified by the Hankel norm:

$$\|F(z) - F_m(z)\|_H \leq 2\mathrm{tr}(\Sigma_2), \tag{5.29}$$

where

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, \tag{5.30}$$

where $\Sigma_2 = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$ are the Hankel singular values (HSVs) of the rejected system after truncation and $\Sigma_1 = \mathrm{diag}(\sigma_{k+1}, \ldots, \sigma_n)$ are the HSVs of the truncated system.

A practical and straightforward implementation of BMT has been outlined in (Beliczynski et al., 1992) and will be presented in the following [7]. The state-space difference equations for an FIR filter $F(z) = c_0 + c_1 z^{-1} + \cdots + c_n z^{-n} = c_0 + F_1(z)$ can be written as

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k) + Du(k), \end{aligned} \tag{5.31}$$

where

$$\begin{aligned} A &= \begin{bmatrix} 0 & 0 & \ldots & 0 & 0 \\ 1 & 0 & \ldots & 0 & 0 \\ & & \ldots & & \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}, \\ B &= \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \\ C &= [c_1 \quad c_2 \quad c_3 \ \ldots \ c_n], \quad D = c_0. \end{aligned} \tag{5.32}$$

The Hankel matrix is formed from the FIR filter coefficients $F_1(z)$ in the following way[8]:

$$H = \begin{bmatrix} c_1 & c_2 & \ldots & c_n \\ c_2 & c_3 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ c_n & 0 & \ldots & 0 \end{bmatrix} \tag{5.33}$$

Since H is a symmetric matrix, the HSVs can be found and ordered using the singular value decomposition method (SVD) (Golub and Loan, 1996):

$$H = V \Lambda V^T \tag{5.34}$$

[7] A Matlab algorithm for BMT implementation is available at `http://www.acoustics.hut.fi/software/hrtf/`.

[8] There is an error in (Beliczynski et al., 1992) for the Hankel matrix formulation. The equation given in Eq. (5.33) is correct.

where Λ is a diagonal matrix containing the HSVs and $VV^T = I$ is a unit matrix. In practice, it is desirable to plot the HSVs and determine the desired order of approximation. In (Beliczynski et al., 1992), the following formulas are presented for deriving a balanced truncated system of order k from matrix V:

$$\begin{aligned} A_k &= V(2:n, 1:k)^T V(1:n-1, 1:k) \\ B_k &= V(1, 1:k)^T \\ C_k &= CV(1:n, 1:k) \\ D &= c_0, \end{aligned} \tag{5.35}$$

where ":" is interpreted as in the Matlab "colon operators" ($1 : n \triangleq 1, 2, .., n$). Now it is possible to write the truncated system in state-space representation, which can then be conveniently transposed to a traditional system transfer function form. Although the above formulation is only an approximation to the Hankel norm, it has been argued in the literature (Kale et al., 1994) that in cases where wideband fitting is applied, a BMT outperforms in terms of frequency response error an optimal Hankel-norm design (Chen et al., 1992).

In the experiments carried out in (Mackenzie et al., 1997), minimum-phase, diffuse-field equalized, auditory smoothed HRTFs (based on Kemar measurements by Gardner and Martin (1994, 1995)) were modeled by 10th-order IIR filters created using BMT. The signal-to-error power ratios (SER) were compared to IIR models designed using Prony's method (Parks and Burrus, 1987) and the Yule-Walker method (Friedlander and Porat, 1984). The average SER was found to be approximately 10 dB better in BMT models (see Fig. 2 in (Mackenzie et al., 1997)).

5.3.5 Warped Filters

In this section, issues in warped filter design are discussed[9]. The underlying theory of frequency warping was presented in Section 5.2.4.

Warped FIR Structures

A warped FIR filter (WFIR) may be interpreted as an FIR structure in which the unit delays have been replaced by first-order allpass filters. Its transfer function is given by

$$\beta_w(z) = \sum_{i=0}^{M} \beta_i \left[D_1(z)\right]^i, \tag{5.36}$$

[9] A Matlab toolbox for frequency warping and implementation of warped filters is available at `http://www.acoustics.hut.fi/software/warp/`.

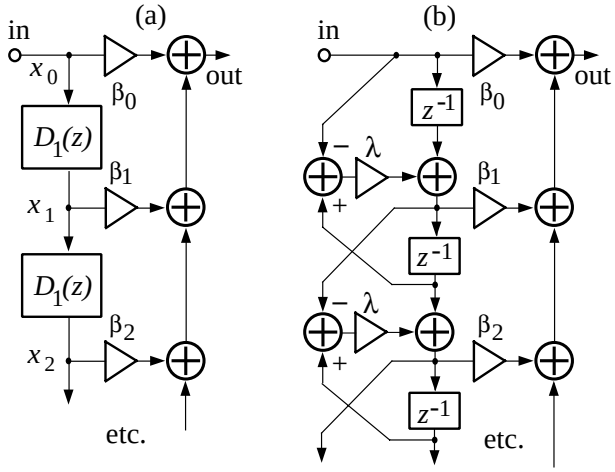


Figure 5.21: WFIR structures (Karjalainen et al., 1999). a) Principle of warped FIR filter. b) Implementation used in present study (Karjalainen et al., 1997a).

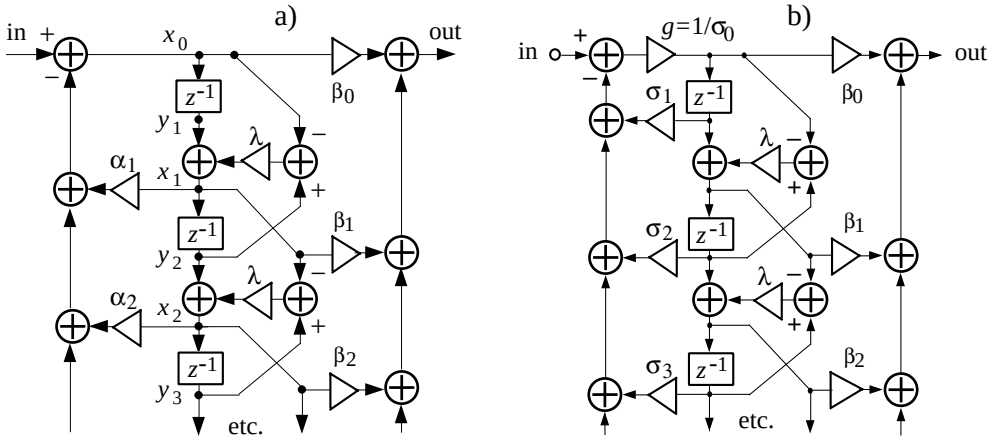


Figure 5.22: WIIR structures (Karjalainen et al., 1999). a) Unrealizable direct form of WIIR filter. b) Realizable modified implementation used in present study.

where $D_1(z)$ is as defined in Eq. (5.19). A more detailed filter structure for implementation is depicted in Fig. 5.21 (Strube, 1980). It can be seen that a warped FIR structure is actually recursive, i.e., an IIR filter with M poles at $z = 1$, where M is the order of the filter. A straightforward method to get the tap coefficients β_i for a WFIR filter is to warp the original HRTF impulse response using Eq. (5.19), and to truncate by rectangular windowing the portion that has the amplitude above a threshold of interest.

Warped IIR Structures

Having the warped HRTF response at hand, a traditional filter design method such as Prony's method (Parks and Burrus, 1987) may be applied to yield a warped pole-zero model of the form:

$$H_w(z) = \frac{\sum_{i=0}^{M} \beta_i \left[D_1(z)\right]^i}{1 + \sum_{i=1}^{R} \alpha_i \left[D_1(z)\right]^i}. \tag{5.37}$$

This cannot be implemented directly due to delay-free recursive propagation through $D_1(z)$ units (Strube, 1980). By proper mapping of α_i coefficients to new σ_i coefficients the warped IIR filter can be implemented as shown in Fig. 5.22 (Imai, 1983; Karjalainen et al., 1997a).

5.4 Binaural System Implementation

Practical binaural filter design and implementation requires knowledge of what the system will feature. When comparing different filter design and implementation strategies, one should pay attention particularly to the following questions:

- Is the system dynamic, i.e., is HRTF interpolation and commutation needed?
- Are minimum-phase and pure-delay HRTF approximations being used?
- Is specialized hardware (signal processors) for implementation being used?
- Are there memory constraints for storing HRTF data?
- Is the HRTF processing part of a room acoustics modeling scheme?

In this work, the goal has been to present and compare methods for HRTF filter design; thus, the solutions to the above questions are only outlined. In many dynamic virtual acoustic environments, minimum-phase FIR approximations have been chosen for HRTF implementation due to straightforward interpolation, relatively good spectral performance, and simplicity of implementation. Furthermore, this approach may easily be integrated into real-time geometric room acoustics modeling schemes, such as the image-source method (Savioja et al., 1999). One of the drawbacks is, however, the large memory requirement that this approach produces. A PCA-based method may be attractive when conducting HRTF spatialization, since as few as 5 principal component basis functions have been found to model human HRTFs (or DTFs, directional transfer functions) with good accuracy (Kistler and Wightman, 1992).

The following benchmarks have been calculated for a Texas Instruments TMS-320C3x floating point signal processor, but are practically similar in other processors as well. FIR implementation is efficient (N+3 instructions for N taps), and

dynamic coefficient interpolation is possible. Designs are usually straightforward (e.g., frequency sampling), but give limited performance, especially at low orders.

IIR implementations are generally slower ($2N$+3 instructions for order N direct form II implementations), especially if dynamic synthesis is required. Interpolation and commutation methods, such as cross-fading, and different transient elimination techniques, increase computation. Pole-zero models are suited for arbitrarily shaped magnitude-response designs; thus, low-order designs are possible.

A second-order section decomposition of IIR filter coefficients may be desirable in many implementations:

$$H_{sos}(z) = K_i \frac{B_i(z)}{A_i(z)} = \frac{\prod_{k=1}^{M/2} b_{0k} + b_{1k}z^{-1} + b_{2k}z^{-2}}{\prod_{k=1}^{N/2} 1 + a_{1k}z^{-1} + a_{2k}z^{-2}} \tag{5.38}$$

In Figs. 5.21–5.22, direct (warped domain) implementations for WFIR and WIIR filters were illustrated. A possible alternative for low-order implementations of warped filters (high orders may be unstable due to computational precision problems) is to *unwarp* the warped structures to traditional direct-form pole-zero filter structures. One such solution is presented in the following. If frequency warping is used prior to filter design, the warped second-order section implementation can be unwarped to regular second-order sections by the following substitutions:

$$H_{dwsos} = \frac{B_{wi}(z)}{A_{wi}(z)} = \frac{\prod_{k=1}^{M/2} \hat{b}_{0wk} + \hat{b}_{1wk}z^{-1} + \hat{b}_{2wk}z^{-2}}{\prod_{k=1}^{N/2} 1 + \hat{a}_{1wk}z^{-1} + \hat{a}_{2wk}z^{-2}}, \tag{5.39}$$

where

$$\begin{aligned}
\hat{b}_{0wk} &= \left(b_{0wk} + \lambda b_{1wk} + \lambda^2 b_{2wk}\right)/\hat{a}_{0wk} \\
\hat{b}_{1wk} &= \left(2\lambda b_{0wk} + (1+\lambda^2) b_{1wk} + 2\lambda b_{2wk}\right)/\hat{a}_{0wk} \\
\hat{b}_{2wk} &= \left(\lambda^2 b_{0wk} + \lambda b_{1wk} + b_{2wk}\right)/\hat{a}_{0wk} \\
\hat{a}_{0wk} &= 1 + \lambda a_{1wk} + \lambda^2 a_{2wk} \\
\hat{a}_{1wk} &= \left(2\lambda + (1+\lambda^2) a_{1wk} + 2\lambda a_{2wk}\right)/\hat{a}_{0wk} \\
\hat{a}_{2wk} &= \left(\lambda^2 + \lambda a_{1wk} + a_{2wk}\right)/\hat{a}_{0wk},
\end{aligned} \tag{5.40}$$

where λ is the warping coefficient.

The efficiency of warped vs. non-warped filters depends on the processor that is used. For Motorola DSP56000 series signal processors, a WFIR takes three instructions per tap instead of one for an FIR. For WIIR filters, four instructions are needed per pole instead of two for an IIR filter. In custom-designed chips, the warped structures may be optimized so that the overhead due to complexity can be minimized. The warped structures may also be expanded (unwarped) into direct form filters, which will lead to the same computational demands as with normal IIR filters.

5.4.1 Interpolation and Commutation of HRTF Filters

The task of *interpolation* and *commutation* of HRTF filters is an important issue when designing dynamic real-time virtual audio environments. In (Jot et al., 1995), the term interpolation has been defined as synthesizing an intermediate transfer function from a database of predefined filters. By commutation (changing) of the coefficients of a filter we simply mean dynamic coefficient updating. The topic has been studied in publications by, e.g., Wenzel and Foster (1993); Runkle et al. (1995); Jot et al. (1995). In the following, methods for FIR and IIR filter coefficient interpolation and commutation (for real-time applications) are overviewed.

FIR Interpolation

For non-recursive FIR filters, direct coefficient interpolation of HRIRs is possible. Minimum-phase approximations used in conjunction with a delay line implementation for ITD have been found to be perceptually indistinguishable from original (mixed-phase) HRIRs (Kistler and Wightman, 1992). However, HRIRs are not completely minimum-phase, and interpolation on mixed-phase filter coefficients may result in a comb-filtering effect in the frequency domain due to changing delay properties. Interpolation of FIR filter coefficients in dynamic auralization can be expressed both in two- and three-dimensional cases.

For 2-D interpolation the coefficients $h_c(n)$ for a desired azimuth angle are obtained from measured HRIRs using the following formula:

$$h_C(n) = \quad (1 - c_\theta)h_A(n) + c_\theta h_B(n), \tag{5.41}$$

where h_A and h_B are h_C's two neighboring data points. Similarly, in a three-dimensional case, the coefficients are obtained using 4-point bilinear interpolation from the four nearest available data points (Begault, 1994). Since the HRTFs are minimum-phase FIRs this interpolation can be carried out. The interpolation scheme for point E located at azimuth angle θ and elevation ϕ is:

$$\begin{aligned} h_E(n) = \quad & (1 - c_\theta)(1 - c_\phi)h_A(n) + c_\theta(1 - c_\phi)h_B(n) + \\ & c_\theta c_\phi h_C(n) + (1 - c_\theta)c_\phi h_D(n), \end{aligned} \tag{5.42}$$

where h_A, h_B, h_C, and h_D are h_E's four neighboring data points as illustrated in Fig. 5.23, c_θ is the azimuth interpolation coefficient $(\theta \bmod \theta_{grid})/\theta_{grid}$, and n goes from 1 to the number of taps of the FIR filter. The elevation interpolation coefficient is obtained similarly $c_\phi = (\phi \bmod \phi_{grid})/\phi_{grid}$.

The task of interpolating non-minimum-phase FIR approximations of HRIRs has been found to produce severe comb-filtering effects if the phase delays of the interpolating filters vary considerably. This phenomenon may occur if non-minimum-phase HRTFs measured at sparse azimuth intervals ($> 10^\circ - 15^\circ$) are interpolated to reproduce intermediate angles.

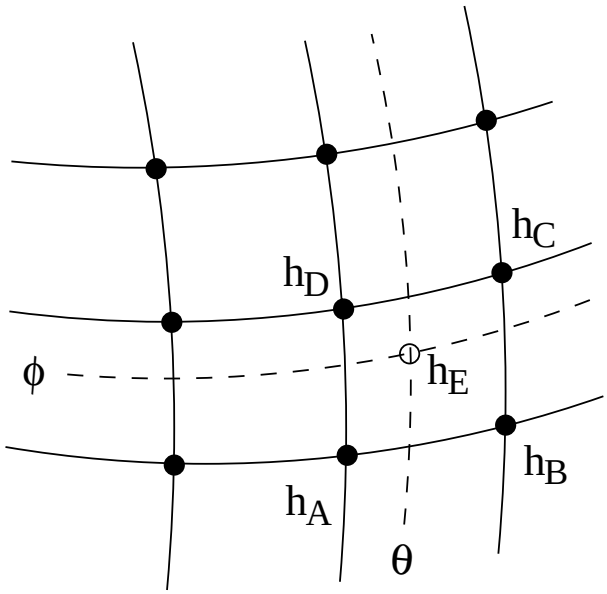


Figure 5.23: HRTF-filter coefficients corresponding to point h_E at azimuth θ and elevation ϕ are obtained by bilinear interpolation from measured data points h_A, h_B, h_C, and h_D (Begault, 1994; Savioja et al., 1999).

IIR Interpolation

Dynamic realization of recursive IIR structure interpolation is more complicated. In (Jot et al., 1995), the possibilities for realizing dynamic 3-D sound synthesis using IIR filters were discussed. Linear interpolation of two stable second-order section- or lattice-realization of IIR filters is guaranteed to be stable. In order to achieve regularity in interpolation, a special pairing and ordering algorithm applied globally to the filter database was presented in (Jot et al., 1995). In that study, four possible choices for IIR filter representation were discussed:

- direct-form coefficients of the second-order sections in cascade
- magnitudes and log-frequencies of the poles and zeros
- lattice filter coefficients k_i
- log area ratios $\log\left[(1-k_i)\,/\,(1+k_i)\right]$

The process of dynamically updating (commuting) the coefficients of recursive filters may cause transients that are audible as disturbing clicks. This problem may be dealt with using different techniques. The method of *cross-fading* uses two filters in parallel, and the output is calculated as a weighted sum (linear interpolation) of the outputs from the filters. The transition time is defined as the

time it takes to completely change the output from one filter to the other. Typically, in HRTF processing, the filters are chosen to be the azimuth and elevation states before and after the transition. This method doubles the computational cost of IIR implementation. In (Välimäki, 1995; Välimäki and Laakso, 1998), a transient elimination method has been presented, which is based on state variable updates of the filter coefficients. This method reduces the computational cost of commutation by 50 %. Various methods for transient elimination in time-varying recursive filters have been discussed in (Välimäki, 1995).

5.5 Objective and Subjective Evaluation of HRTF Filter Design Methods

This section presents methods and procedures for objective and subjective evaluation of binaural filter design. It is clear that the performance of binaural systems has to be evaluated using human subjects in formal listening experiments. In addition, it is desirable to have objective measures of performance quality that can be used for comparing different filter designs or implementation techniques. The objective measures that have been considered in this study are based on modeling of auditory perception.

5.5.1 Objective Methods

Spectral Distance Measure

There is a need to have a simple numerical measure of filter design quality that is meaningful also from the perceptual point of view. A similar approach has been considered in, e.g., (Wu and Putnam, 1997). In (Huopaniemi and Karjalainen, 1997), a spectral (magnitude) distance measure was experimentally derived in the following way.

The equalized impulse response is first FFT transformed to magnitude-square of the frequency response, sampled (by interpolation) uniformly on a logarithmic frequency scale, smoothed with about 0.2 octave resolution (this resolution value was specified somewhat arbitrarily to be not too far from the ERB resolution), and converted to dB scale. In the next step the difference of the spectrum to be analyzed and a reference spectrum is computed for the passband region of approximation. The reference spectrum may be simply the average level of the spectrum to be analyzed or some other reference. In our case it was the corresponding reference response in our listening experiments. A root-mean-square value of the difference spectrum is then computed, and this is used as an objective spectral distance measure to characterize the perceptual difference between the magnitude responses or a deviation from a reference response. Notice that the

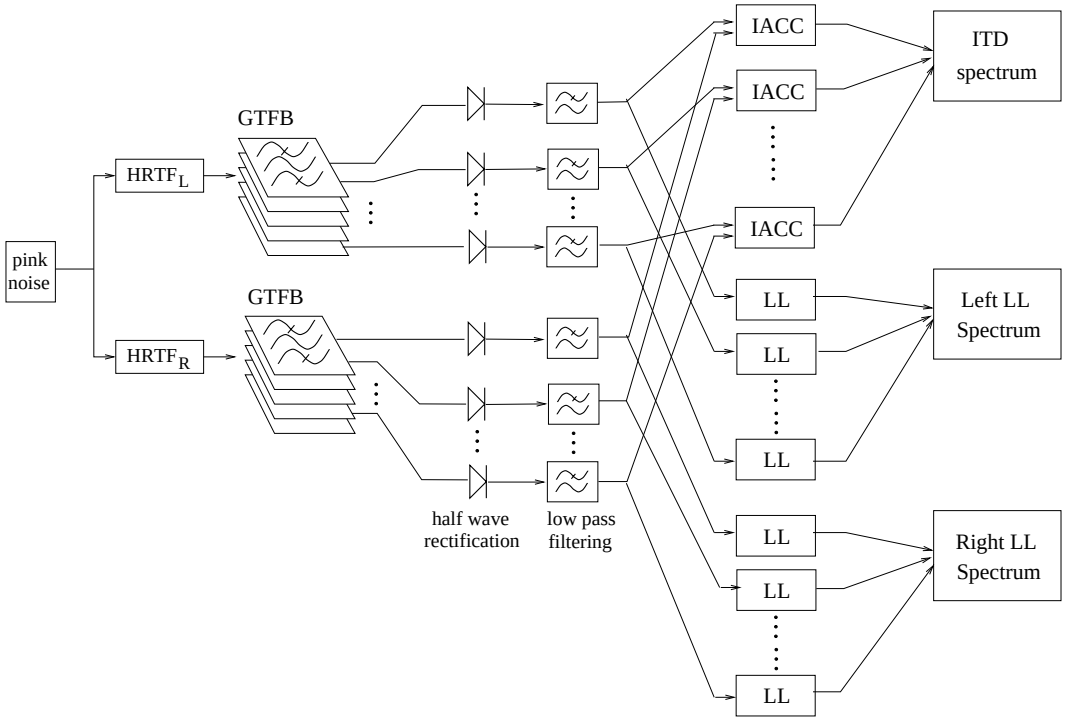


Figure 5.24: Detailed view of the binaural auditory model used in the objective evaluation of HRTF filter designs (Pulkki et al., 1999).

values of the spectral distance measure used in our study are not calibrated to be compared directly with any perceptual difference measures.

Binaural Auditory Model

A novel idea of using a binaural auditory model for virtual source quality estimation has been proposed by Pulkki et al. (1998, 1999)[10] This method has been adopted successfully for binaural filter quality estimation by the author in (Pulkki et al., 1998; Huopaniemi et al., 1998; Pulkki et al., 1999; Huopaniemi and Smith, 1999; Huopaniemi et al., 1999b). The basis of the auditory model applied in this study lies in the coincidence and cross-correlation principles introduced by Jeffress (1948). The early model has been further extended by several authors (e.g. Lindemann (1986) and Gaik (1993)). A schematic of the binaural auditory model used in this study is depicted in Fig. 5.24 (slightly modified from (Pulkki et al., 1998)). The model consists of the following steps.

- A pink noise sample convolved with the HRTFs under study is used as the excitation. Pink noise yields a spectrally balanced excitation to the

[10] Similar techniques have been presented by Macpherson (1991).

auditory system lacking major temporal attacks, thus the influence of the precedence effect is minimized.

- The HRTF-filtered pink noise samples are passed through a gammatone filterbank (GTFB) (Slaney, 1993) of N bandpass ERB (equivalent rectangular bandwidth) channels (N depends on the sampling frequency f_s).
- Half-wave rectification and lowpass filtering (cutoff frequency at 1 kHz) are used to simulate the hair cells and the auditory nerve behavior in each bandpass channel.

For the ITD model, the *interaural cross-correlation* function (IACC) is calculated for each bandpass channel pair (see Fig. 5.24). The ITD as a function of ERB channel is estimated by calculating the time delay corresponding to the position of the maximum in each bandpass IACC curve. Loudness estimates (L in sones) for the left and right ERB bandpass channels are calculated using the equation $L = \sqrt[4]{< H^2 >}$ (an approximation of Zwicker's formula, where he used an exponent of 0.23 (Zwicker and Fastl, 1990)), where $< H^2 >$ is the average power of the ERB channel output. The loudness levels L_L (in phons) for each ERB channel are computed using the formula $L_L = 40 + 10 \log_2 L$, resulting in a loudness level spectrum for left and right ear signals.

Error Criteria for Binaural Auditory Model

In order to be able to compare the binaural auditory model outputs for different filter designs, a suitable error measure had to be considered. The root-mean-square (RMS) error was found attractive for comparing the binaural auditory model outputs of different filter designs (and filter order) with a reference response. In the calculation of the distance measures, the basic phenomena found in human sound localization (Blauert, 1997) were observed: 1) the spectral and interaural level difference (ILD) cues dominate localization at frequencies above approximately 1.5 kHz, but may also contribute to localization at lower frequencies, and 2) the interaural time difference (ITD) is the dominant localization cue at frequencies below approximately 1.5 kHz. Based on these assumptions, two quality measures were derived:

- Perceived loudness level spectrum error
- Perceived loudness level spectrum + ITD modeling error

The modeling errors are calculated as an RMS difference between the outputs from the auditory models (loudness level spectra or ITDs) of the reference and the filter approximation over a passband $f_l - f_h$. In the results presented below the following limits were chosen: $f_l = 1.5$ kHz and $f_h = 16$ kHz. The loudness level spectrum error was calculated by summing the left and right ear loudness

errors[11]. The loudness level spectrum + ITD modeling error was calculated by scaling and summing the low frequency ($f < 1.5$ kHz) ITD modeling error with the high frequency loudness level spectrum modeling error. Since the role of low-frequency spectral cues in sound source localization is ambiguous, a choice was made to exclude the loudness level error at $f < 1.5$ kHz from the RMS error measures.

5.5.2 Subjective Methods

The more traditional and straightforward approach (albeit more time-consuming) for evaluating the quality of binaural systems is to use subjective listening experiments. During the course of this work, several listening experiments were conducted on the perception of HRTF filter quality degradation and localization of virtual sound sources in headphone listening.

5.6 Experiments in Binaural Filter Design and Evaluation

In the following, results from four different studies carried out by the author on binaural filter design and evaluation are presented. First, design examples for the papers (Huopaniemi and Karjalainen, 1996b,a) are overviewed. Second, the design examples and listening tests presented in (Huopaniemi and Karjalainen, 1997) are discussed. The third study (Huopaniemi et al., 1998, 1999b) presents methods and results of individualized HRTF filter design and listening tests. Finally, the fourth study (Huopaniemi and Smith, 1999) discusses the role of spectral and temporal preprocessing and equalization methods in binaural filter design. The goal in these experiments has been to investigate subjective and objective evaluation of HRTF filter design methods and to determine the validity of auditory-based filter design methods proposed by the author. Furthermore, the task has been to find benchmarks for binaural synthesis, that is, thresholds for HRTF filter approximation where no perceptual differences can be found when compared to original (raw) HRTFs.

5.6.1 Binaural Filter Design Experiment

In (Huopaniemi and Karjalainen, 1996b,a), the application of warped filter design methods to HRTF modeling was tested using measurements carried out on a dummy head (Brüel&Kjaer 4100). HRTFs were measured for both ears at 5° azimuth intervals in an anechoic chamber using the DRA Laboratories' MLSSA

[11] Summation of left and right ear loudnesses to obtain a binaural loudness estimate was used in (Moore et al., 1997).

system. Additional measurements were carried out to compensate for the measurement equipment and to obtain headphone correction filters (four headphone types were used). The internal sampling rate of the MLSSA system was 75.47 kHz and the results were bandpass filtered from 20 to 25000 Hz. In addition, HRTF measurements made on a Kemar dummy head were modeled (Gardner and Martin, 1994, 1995). The BK4100 HRTF measurements were preprocessed according to the following principles. These stages included 1) minimum-phase reconstruction and 2) smoothing of the magnitude response data. For comparison, diffuse-field equalized Kemar dummy head HRTFs were used.

Frequency warping was performed after preprocessing. Two values of λ were used (0.65 and 0.7233), the first of which is slightly lower than for approximate Bark-scale warping, and the latter being close to the optimal Bark-scale match (for $f_s = 44.1$ kHz). In the warped frequency domain, different FIR and IIR filter design methods were compared. A time-domain IIR design method, Prony's s method (Parks and Burrus, 1987) was used, because it outperformed other tested methods (i.e. the Yulewalk method and discrete least-squares fitting) especially in low-order approximations.

In Fig. 5.25, the modeling results for Kemar are illustrated. The example HRTF was measured at 0° elevation, 30° azimuth, and diffuse-field equalization was used. It can be seen that a warped Prony design (denoted WIIR2) easily outperforms a linear Prony design (denoted IIR) of equivalent order for frequencies up to 9 kHz, which is expected. In this case, the order of the IIR filters was 20, compared to the FIR designs of order 40. The frequency-sampling method was used both in the WFIR and FIR designs. The performance of a WFIR order 13 was compared to a non-warped FIR or order 40. Accuracy is slightly reduced in WFIR implementation. The performance of WIIRs when compared to the IIR are superior in both examples at lower frequencies. The tradeoff at higher frequencies is tolerable according to psychoacoustical theory. Furthermore, the BMT filter design method presented in section 5.3.4 was compared to the previous methods. It can be seen that a very low-order BMT model (order 10) retains most of the spectral features of the HRTF, although not at the same accuracy as a WIIR of same order.

In Fig. 5.26, the BK4100 HRTFs were used. The filter orders for modeling were higher than in the previous experiment. The results are similar; the efficiency of the WIIR structure at lower frequencies with a tradeoff at high frequencies can be seen clearly. If the WIIR filters are implemented by inverse warping (unwarping) the coefficients into traditional direct form representations, the WIIR appears to be the best solution. There are constraints in inverse warping caused by numerical accuracy, but in this study WIIR approximations of orders 10-32 have been successfully unwarped.

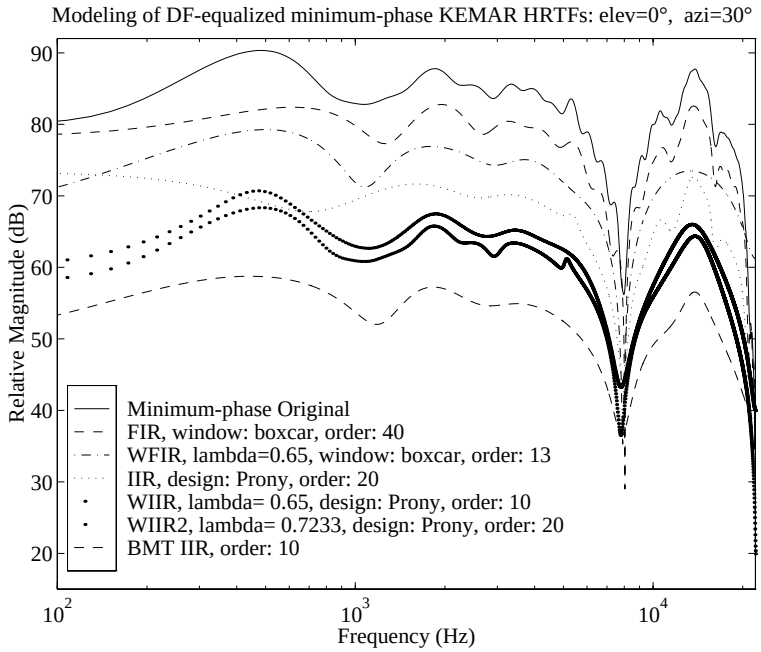


Figure 5.25: Warped filter design compared with traditional FIR and IIR designs. Kemar dummy head HRTFs were used (Huopaniemi and Karjalainen, 1996b,a).

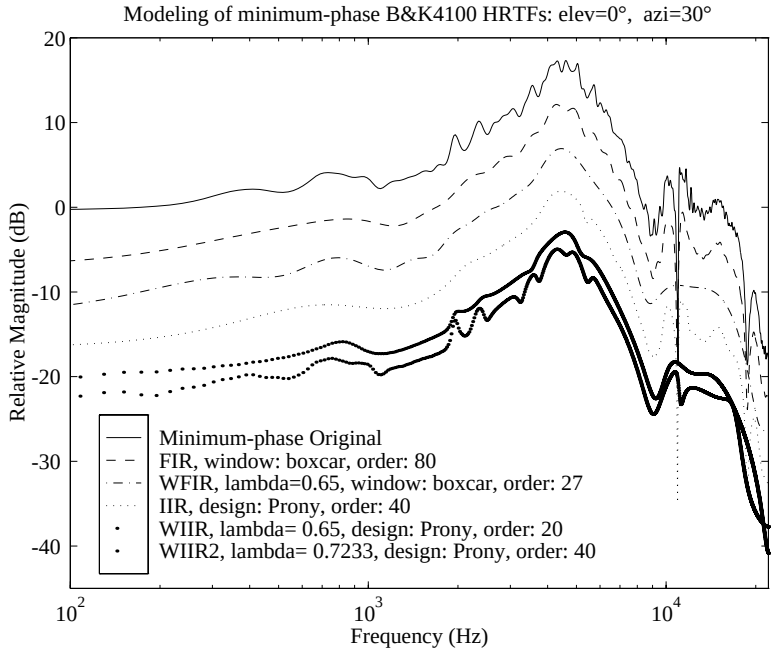


Figure 5.26: Warped filter design compared with traditional FIR and IIR designs. BK4100 dummy head HRTFs were used (Huopaniemi and Karjalainen, 1996b,a).

5.6.2 Evaluation of Non-Individualized HRTF Performance

In the second experiment (Huopaniemi and Karjalainen, 1997), the HRTFs used in the filter design examples and listening tests were measured from a Cortex MK2 dummy head in an anechoic chamber. The Cortex dummy head was equipped with Brüel&Kjaer 4190 microphones (blocked ear canal version). The transducer used in the measurements was a four-inch Audax AT100M0 loudspeaker mounted in a plastic ball. A random-phase flat amplitude spectrum pseudorandom noise signal was used as the excitation sequence. Data were played and recorded using an Apple Macintosh host computer and the QuickSig signal processing environment (Karjalainen, 1990). A National Instruments NI-2300B DSP card with Texas Instruments TMS320C30 signal processor and high-quality 16-bit AD/DA converters were used for both HRTF measurements and listening experiments.

The HRTF data were post-processed for headphone listening experiments in the following way. A compensation measurement was made to account for the measurement system by placing a microphone at the dummy head position with the head absent. Headphone transfer functions for the Sennheiser HD580 headphone were measured on the dummy head. A 300-tap FIR inverse filter for each ear was designed using least-squares approximation. The HRTF data was then convolved by the compensation response and the headphone correction filter for both ears.

The minimum-phase reconstruction was carried out using windowing in the cepstral domain (as implemented in the Matlab Signal Proc. Toolbox `rceps.m` function (Mathworks, 1994)). The cross-correlation method was used to find the ITD for each incident angle. The ITD was inserted as a delay line. Three different minimum-phase HRTF approximations were used: windowed FIR design (rectangular window), time-domain IIR design (Prony's method (Parks and Burrus, 1987)) and a warped IIR design (warped Prony's method, warping coefficient $\lambda = 0.65$. In Table 5.2, the processed filter lengths for different filter types are collected. The example HRTFs were measured at 0° and 135° azimuth (0° elevation) positions. The magnitude responses of the HRTF filter approximation using IIR and WIIR design for 135° azimuth can be seen in Figs. 5.27–5.28. Again, it can be seen that a warped Prony design has an improved low-frequency fit (up to approximately 9 kHz) when compared to a linear Prony design of equivalent order. The better fit at low frequencies when comparing WIIR approximation to windowed FIR can also be observed. The value of $\lambda = 0.65$ was used, which is slightly lower than for approximate Bark-scale warping. Figure 5.29 plots the spectral distance measures as functions of the number of filter coefficients for the three HRTF filter types used in our study: FIR, IIR, and WIIR. The spectral distance measure has been calculated using the method described in Section 5.5.1 for HRTFs from both ears at the two azimuth angles (0° and 135°) used in the listening test. The computational load from the implementation point of view

FIR (rect. wind.)	**IIR (Prony's method)**	**Warped IIR** ($\lambda = 0.65$)
256 (reference)	128	128
128	64	64
96	48	48
88	44	44
80	40	40
72	36	36
64	32	32
60	30	30
56	28	28
52	26	26
48	24	24
44	22	22
40	20	20
36	18	18
32	16	16
28	14	14
24	12	12
20	10	10
16	8	8

Table 5.2: HRTF filter types and orders used in the second listening experiment.

of these structures is comparable, provided that the WIIR filters are unwarped to traditional direct form IIRs. The reference for distance computation was the highest-order response, in order to make the results compatible to the setup used in our listening experiments.

Listening Test Procedure

In order to verify the theoretical filter design results, headphone listening experiments were carried out. The goal in the study was to detect the just noticeable difference (JND) thresholds of audibility by varying the filter order and comparing it to a reference HRTF (similarly as in (Sandvad and Hammershøi, 1994a,b)). Listening experiments were carried out for the three HRTF filter design methods described in the previous section: FIR, IIR, and WIIR. A total of 8 test subjects participated in the listening experiment, 6 male and 2 female with ages ranging between 21 and 35. The hearing of all test subjects was tested using standard audiometry. None of the subjects had reportable hearing loss that could effect the test results. It should be pointed out here that the experiment was done using *non-individualized HRTFs* (measured on a dummy head) that were equalized for a specific headphone type (Sennheiser HD580).

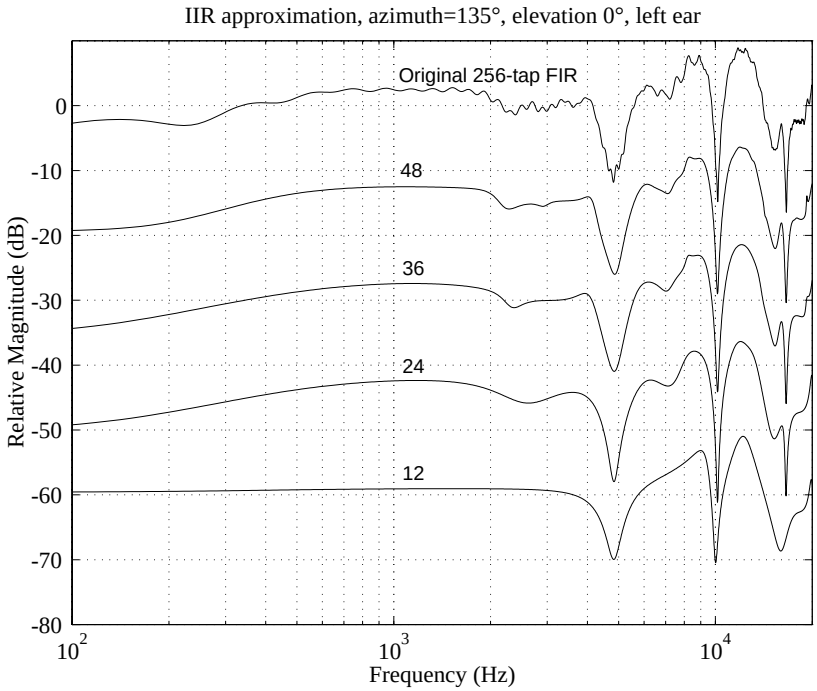


Figure 5.27: IIR filter design of Cortex dummy head HRTFs (135° azimuth). IIR filter orders: 12, 24, 36, 48 (Huopaniemi and Karjalainen, 1997).

In the listening experiment the TAFC (Two Alternatives Forced Choice) bracketing method was used. The method is described in great detail in (ISO, 1989) and widely used in, e.g., audiometric tests. In each trial, two test sequences were presented with a 0.5 s interval between the samples. The first test signal was always the reference signal, and the second signal varied according to adaptation. Each test type was played three times and only the last two were accounted for in the data analysis. The test subjects were given written and oral instructions. They were also familiarized with a test sequence that demonstrated both distinguishable and indistinguishable test signal pairs.

A total of four different stimuli were first processed for a pilot study; pink noise, male speech, and female speech, and a music sample. All samples were digitally copied and processed from the Music for Archimedes CD[12]. In the final experiment, however, only the pink noise sample was used. This was due to the fact that remarkable differences in different filter designs could clearly be heard only using wide-band test signals. A pink noise sample with a length of one second (50 ms onset and offset ramps) was used in the final experiment. The level of the stimuli was adjusted so that the peak A-weighted SPL did not exceed 70 dB at any point. This has been done in order to avoid level adaptation and the acoustical reflex (Stapedius reflex).

[12] Music for Archimedes, CD B&O 101 (1992).

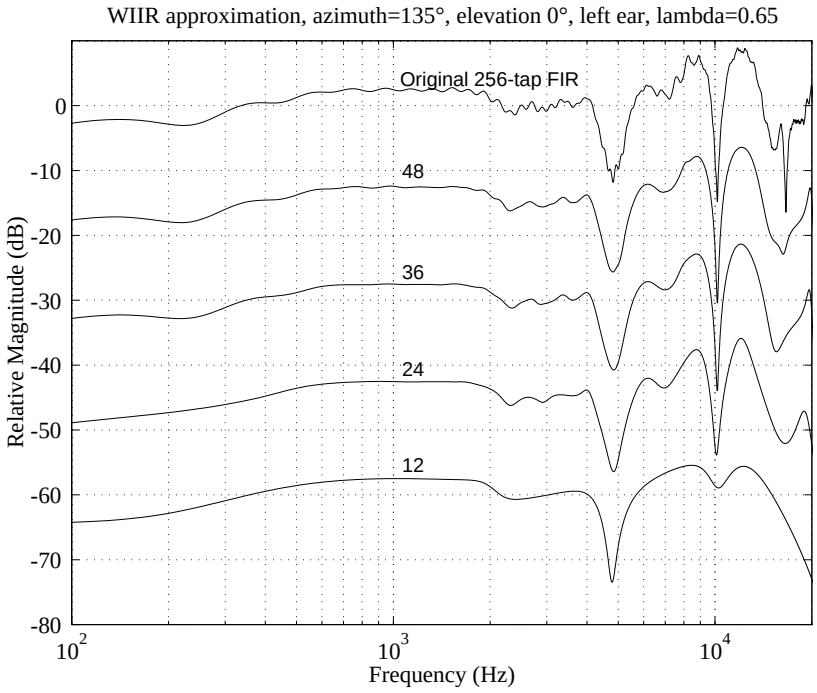


Figure 5.28: WIIR filter design of Cortex dummy head HRTFs (135° azimuth). WIIR filter orders: 12, 24, 36, 48 (Huopaniemi and Karjalainen, 1997).

The test subject was seated in a semi-anechoic chamber (anechoic chamber with a hard cardboard floor). The test stimuli were presented over headphones. A computer keyboard was placed in front of the test subject. Each test subject was individually familiarized and instructed to respond in the following way:

- Press 1 if the signals are the same
- Press 2 if the signals are different
- Press Space if you want to repeat the signal pair

As a total, three different filter approximations for two apparent source positions were used. Each alternative was played three times. The results of the listening tests were gathered automatically by a program written for the QuickSig environment (Karjalainen, 1990). The resulting data were transferred into Matlab, where analysis was performed.

Results

In Fig. 5.30, the results of the listening test are presented. This figure illustrates the distribution of just noticeable difference (JND) thresholds calculated across two tested azimuth angles, 0° and 135° with three filter types, FIR, IIR, and

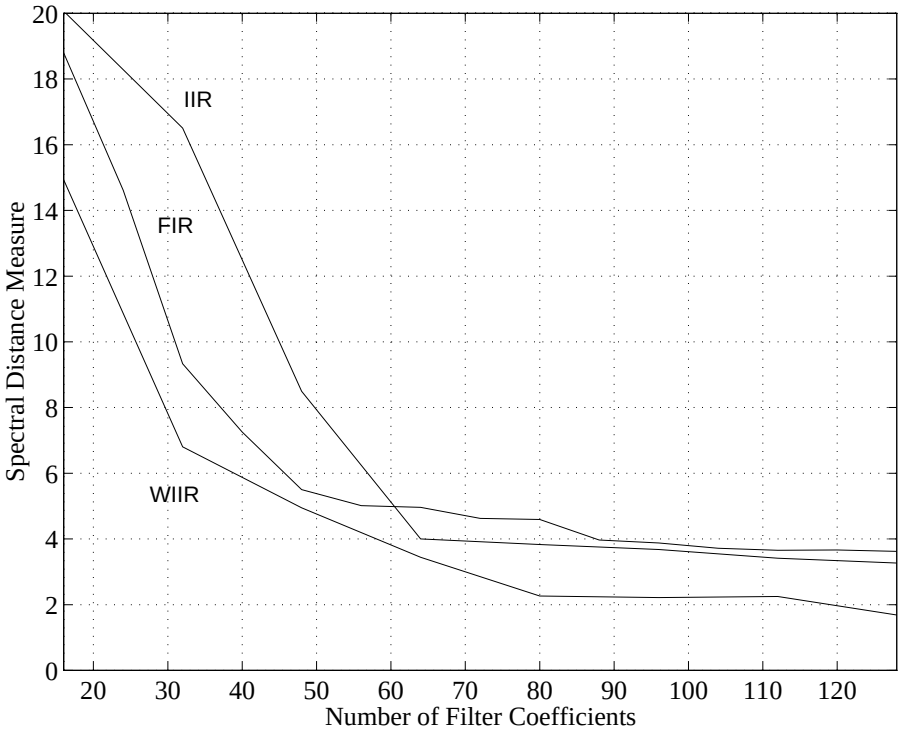


Figure 5.29: Characterization of filter design quality using a spectral distance measure as a function of the number of filter coefficients for the three filter types of the study: FIR, IIR, and WIIR (Huopaniemi and Karjalainen, 1997).

WIIR. The median value as well as the lower and upper quartile (25% and 75%) values are shown.

The results show that the distribution of results in the listening panel was relatively small, although not as well defined as a pilot study indicated. This may be a consequence of using an inhomogeneous listener panel. Some of subjects were experienced analytic listeners while some did not have prior experience in a listening panel. A longer training prior to final experiments could have made the test results more systematic (Bech, 1993).

From Fig. 5.30 it can be seen that the WIIR performance from the filter order point of view is superior when compared to FIR and IIR designs. From the computational point of view, however, the FIR and WIIR implementations appear to be approximately equal in performance. The warped IIR designs, however, easily outperforms a conventional IIR design. A useful criterion to select filter order values could be the upper quartile (75%) or even higher level of subject reactions. Using the 75% quartile results, one concludes in the following statements.

For non-individualized (dummy-head) HRTFs equalized to a specific headphone, the filter orders at the sampling rate of 48 kHz where 75% of the popula-

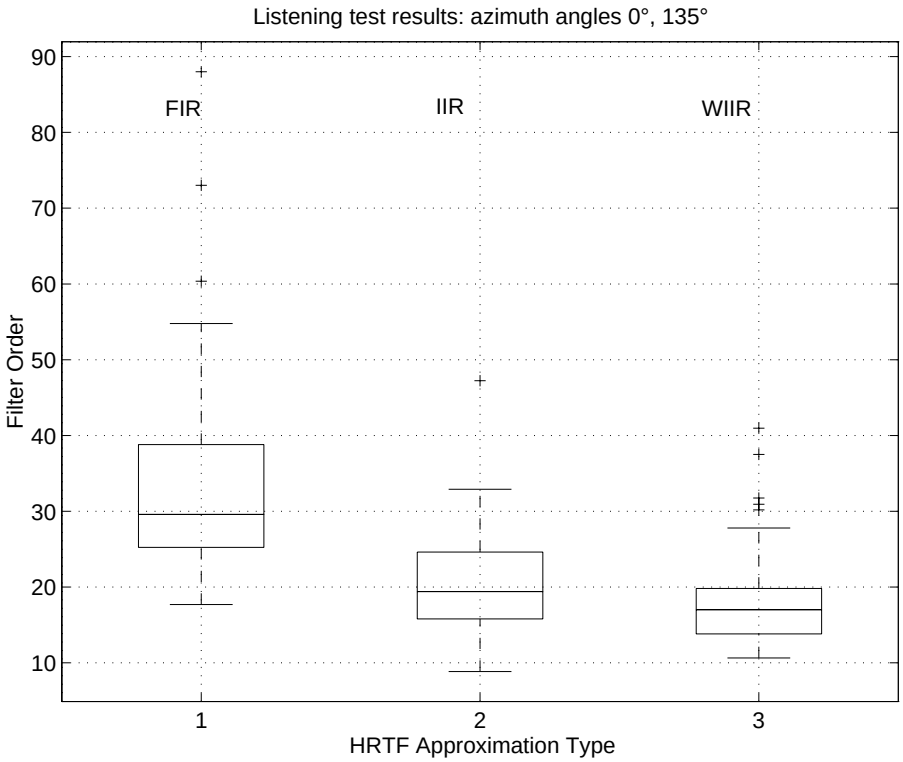


Figure 5.30: Results of the listening test. The boxplot depicts the median (straight line) and the 25%/75% percentiles (Huopaniemi and Karjalainen, 1997).

tion stated no difference when compared to the reference were approximately:

- Order 40 for a frequency-sampling FIR design
- Order 24 for a time-domain IIR design (Prony's method)
- Order 20 for a warped IIR design (Prony's method, $\lambda = 0.65$)

In comparison to the results presented in the literature, some comments can be made. The empirical study by Sandvad and Hammershøi (1994a,b) resulted in orders 72 for a FIR and 48 for an IIR filter. The difference compared to the results may be caused by the fact that Sandvad and Hammershoi used individual HRTFs and headphone calibration, and both speech and pink noise samples. However, the estimated detection probabilities (maximum likelihood estimation was used as a statistical model) for the given results were approximately 0.08, higher than in our conclusions. Moreover, the filter orders used in that study were relatively sparse (orders 24, 36, 72 and 128 for FIR, and orders 10, 20, 30 and 48 for IIR using pink noise).

Conclusion and Discussion

In the spectral distance measure of Fig. 5.29 it can be seen that from the auditory point of view the WIIR filters have the best performance. As can be seen in further analysis, The correlation with the listening test results is good. The JND values for FIR and WIIR filters according to the listening tests were approximately 40 (filter coefficients), and for the IIR structure approximately 50. It is, however, a valid question, why the FIR filters performed so well in the listening tests without having any weighting or frequency warping applied. This may have been caused by the use of non-individualized (dummy-head) HRTFs.

5.6.3 Evaluation of Individualized HRTF Performance

This section presents the results of individualized HRTF filter design in objective and subjective experiments (Huopaniemi et al., 1998, 1999b).

First, the HRTF data and the methods of filter design used in the objective and subjective analysis of this work are described. The task was to further investigate three different filter design approaches (FIR, IIR, WIIR), and the goal was to find methods and criteria for subjective and objective evaluation. This research concentrated on finding answers to two problems often found in HRTF filter design:

- What is the needed filter length for perceptually relevant HRTF synthesis?
- Is it possible to introduce an "auditory resolution" in the filter design, whereby the spectral and interaural phase details are modeled more accurately at low frequencies and considerably smoothed at high frequencies?

A set of 10 human test subjects were chosen for the experiment. A blocked ear canal HRTF measurement technique (Møller et al., 1995) was used to obtain the needed transfer functions for the experiments (the used measurement setup is discussed in greater detail in (Riederer, 1998a). An Audax AT100M0 4-inch transducer element in a plastic 190 mm enclosure was used as a sound source for the measurements. The test subjects were seated in an anechoic chamber on a rotating measurement chair. Sennheiser KE211-4 miniature microphones were used. HRTF and headphone responses (for the used headphone type: Sennheiser HD580) were measured for each test person to enable full individual HRTF simulation. Pseudorandom noise was used as the measurement sequence. For the objective and subjective tests, four azimuth angles were chosen: 0^o, 40^o, 90^o, and 210^o. These incident angles represent the median plane, frontal plane, and horizontal plane responses.

The minimum-phase reconstruction was carried out using windowing in the cepstral domain. The minimum + excess phase approximation method (Jot et al., 1995) was used to find the ITD for each incident angle. The ITD was inserted in HRTF synthesis as a delay line. Three different minimum-phase HRTF

approximations were used: Windowed FIR design (rectangular window), time-domain IIR design (Prony's method (Parks and Burrus, 1987), as implemented in the Matlab Signal Processing Toolbox (Mathworks, 1994)) and a warped IIR design (warped Prony's method, warping coefficient $\lambda = 0.65$). The reference filter order for each incident angle was chosen to be 256 (257 FIR taps). The tested filter orders were:

- FIR: 96, 64, 48, 32, 16
- IIR: 48, 32, 24, 16, 8
- WIIR: 48, 32, 24, 16, 8

As an example, in Fig. 5.31 magnitude responses of original and filter approximations from one test subject's HRTFs at 40^o azimuth (0^o elevation) are depicted. It can be seen that a WIIR design (using Prony's method) has an improved low-frequency fit (up to approximately 9 kHz) when compared to a uniform frequency resolution IIR design of equivalent order. The better fit at low frequencies when comparing WIIR approximation to windowed FIR can also be observed both in the amplitude response plots and the ITD plot. The warping value of $\lambda = 0.65$ was used, which is slightly lower than for approximate Bark-scale warping. This value was chosen by visual inspection of the magnitude responses to enhance low-frequency fit but still retain the overall high-frequency magnitude envelope. The results of using a binaural auditory model (discussed in section 5.5.1) in HRTF filter design quality estimation are depicted in Figs. 5.32–5.34. The loudness level spectrum estimates of HRTFs for a test subject at 40^o azimuth are shown in Figs. 5.32–5.33 and the perceived ITD of the corresponding filter approximations is plotted in Fig. 5.34 (note that the loudness level spectrum estimates shown in Fig. 5.33 correspond to the magnitude responses shown in Fig. 5.31). The dashed line in each subplot is the reference response, and the solid line is the current approximation result. The number of filter coefficients for the row of plots is shown to the right of the figures. Furthermore, an RMS error estimate of the corresponding design is shown in Figs. 5.32–5.33 in each of the subplots. From the plots it can seen that the spectral detail clearly visible in the original HRTF (see Fig. 5.31) is smoothed as a consequence of auditory modeling. The RMS error estimate shows that over the passband of loudness level error calculation (1.5-16 kHz) the warped IIR design method provides the best fit to the desired response. This is also true in the case of the ITD estimate, shown in Fig. 5.34.

The results of estimating HRTF filter design quality using a binaural auditory model suggest that high-frequency smoothing of the HRTFs is motivated. The results depicted in Figs. 5.35–5.36 confirm that the RMS loudness level spectrum error is lowest for the auditory WIIR filter design, which essentially provides a better low-frequency fit with a tradeoff in high-frequency accuracy. In the

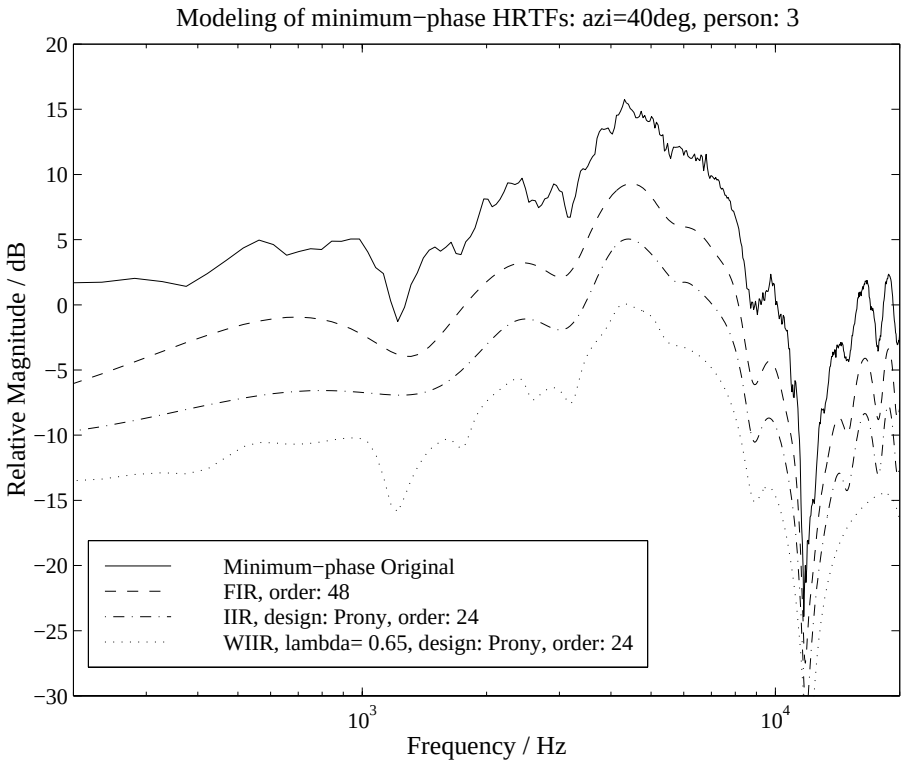


Figure 5.31: HRTF magnitude response and different filter approximations. Right ear, azimuth angle 40^o, person 3 (Huopaniemi et al., 1998, 1999b).

figures, both left and right ear HRTF approximations were modeled for 9 test subjects at 4 azimuth angles. It can be seen that an auditory scale filter (WIIR) outperforms both FIR and IIR filter design methods. The results are even more dramatic when the ITD modeling error is added to the error measure. In Fig. 5.36, these results for 9 test subjects at 4 azimuth angles are depicted. In both plots, the filter design errors start to increase as the number of coefficients is below 48. The WIIR design error is tolerable to order 16, whereas both the FIR and IIR design errors start to increase earlier.

Subjective Analysis

In order to verify the theoretical filter design results headphone listening experiments were carried out. The goal was to study the performance of individualized HRTF filter approximations using different design techniques and different filter orders. Listening experiments were carried out for the three HRTF filter design methods as described in the previous section: FIR, IIR, and WIIR. A total of 10 male test subjects participated in the listening experiment, with ages ranging between 25 and 51. The hearing of all test subjects was tested using standard audiometry. None of the subjects had reportable hearing loss that could effect

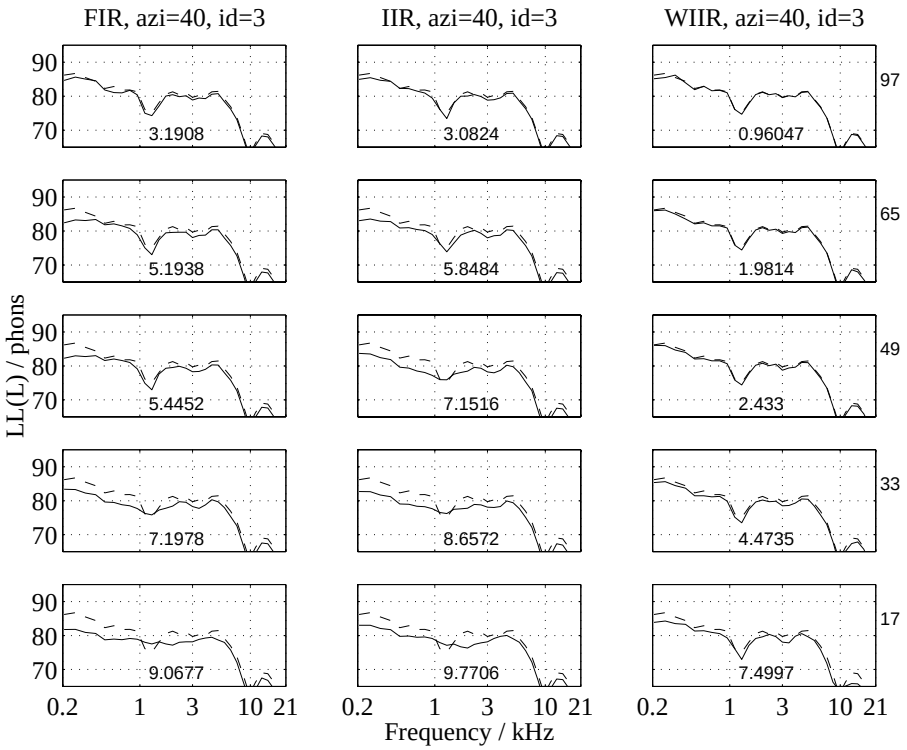


Figure 5.32: Binaural auditory model evaluation of filter design quality. Left ear, azimuth angle 40°, person 3. Solid line: filter approximation, dashed line: reference (Huopaniemi et al., 1998, 1999b). The RMS error is shown on each graph, and the corresponding number of filter coefficients is given on the right side of the figure.

the test results. The HRTF approximations were individually equalized for a specific headphone type (Sennheiser HD580).

An A/B paired comparison hidden reference paradigm was employed for the listening tests with two independent grading scales. The subjects were asked to grade localization and timbre impairment against the hidden reference on a continuous 1.0 to 5.0 scale (as proposed in ITU-R BS 1116-1 (ITU-R, 1997)). The hidden reference in each case was the 257-tap filter. In each trial, two test sequences were presented with 0.5s between sample (i.e. A/B//A/B). A full permutation set was employed and two different random orders of presentation were used to minimise bias. To obtain data regarding listener reliability the reference (the 257-tap FIR) case was also tested against itself. Each test type was repeated two times. Listeners were given written and oral instructions.

A pink noise sample with a length of one second (50 ms onset and offset ramps) was used as sound stimulus in the final experiment. The level of the stimuli was adjusted so that the peak A-weighted SPL did not exceed 70 dB at any point. This has been done in order to avoid level adaptation and the acoustical reflex

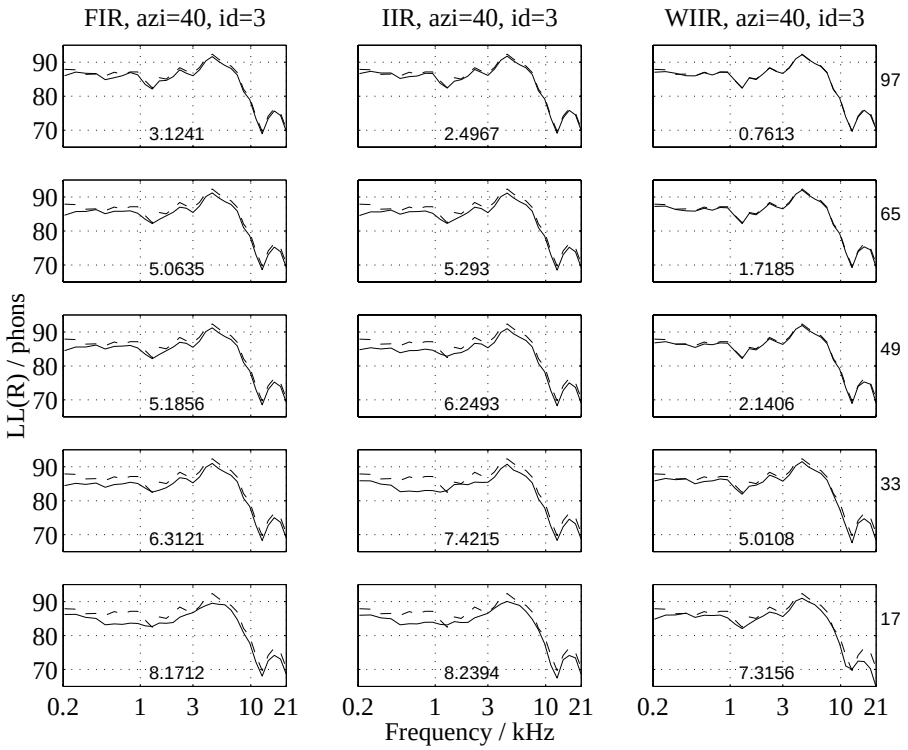


Figure 5.33: Binaural auditory model evaluation of filter design quality. Right ear, azimuth angle 40°, person 3. Solid line: filter approximation, dashed line: reference (Huopaniemi et al., 1998, 1999b).

(Stapedius reflex). No gain adjusting of the test sequences calculated for one person was carried out, since the only variability in level was (possibly) introduced by the used HRTF filters.

The test procedure in the experiment was as follows. The listener was seated in a semi-anechoic chamber (anechoic chamber with hard cardboard floor). The test stimuli were presented over headphones. A computer keyboard was placed in front of the listener. Each listener was individually familiarized and instructed to grade the localization and timbre scales for each test signal pair. An example plot of the listening test software user interface is shown in Fig. 5.37.

A total of five different filter approximations for each of the three filter types at four apparent azimuth source positions were used. Each alternative was played three times. The results of the listening tests were gathered automatically by a program written for the QuickSig environment (Karjalainen, 1990). The result data were transferred into the SPSS software (Statistical Package for Social Sciences), where analysis was performed.

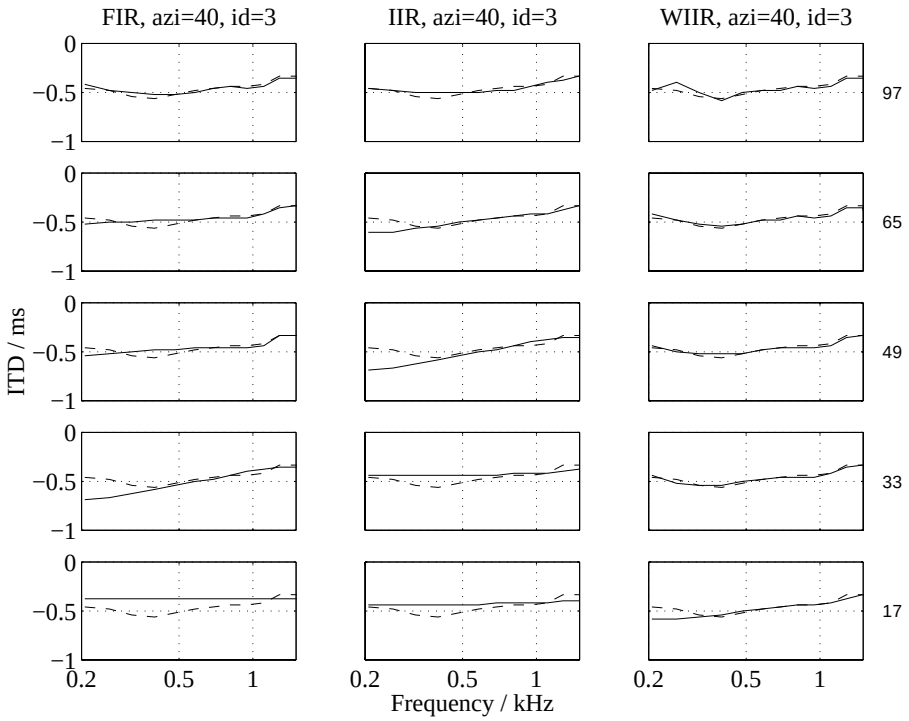


Figure 5.34: Binaural auditory model evaluation of filter design quality. ITD, azimuth angle 40°, person 3. Solid line: filter approximation, dashed line: reference (Huopaniemi et al., 1998, 1999b).

Results and Discussion

The data were initially checked to ensure equal variance across listeners. At this point it was found that one listener had very different variance than the others. Upon closer inspection this listener was found to be grading very similarly for all systems, had non-normal distribution of data and also had very low error variance and poor F-statistics, based upon an analysis of variance (ANOVA). This was an indication that this listener could not discriminate between systems and was thus eliminated from subsequent analysis.

The data were then tested for conformance with the analysis of covariance model (ANCOVA) assumption. The data were found to be normally distributed, though slightly skewed, which is typical of subjective data. However, the ANCOVA model is fairly robust to slight skewness of data. Residuals were found to be normally distributed. It should be noted that the model did not meet the requirements for homogeneity of variance and the Levene statistics were found to be significant. This is not considered problematic as in all other respects the (ANCOVA) assumptions have been met and the raw and modeled data were found to be strongly correlated.

An ANCOVA was employed for the full analysis of the data considering all

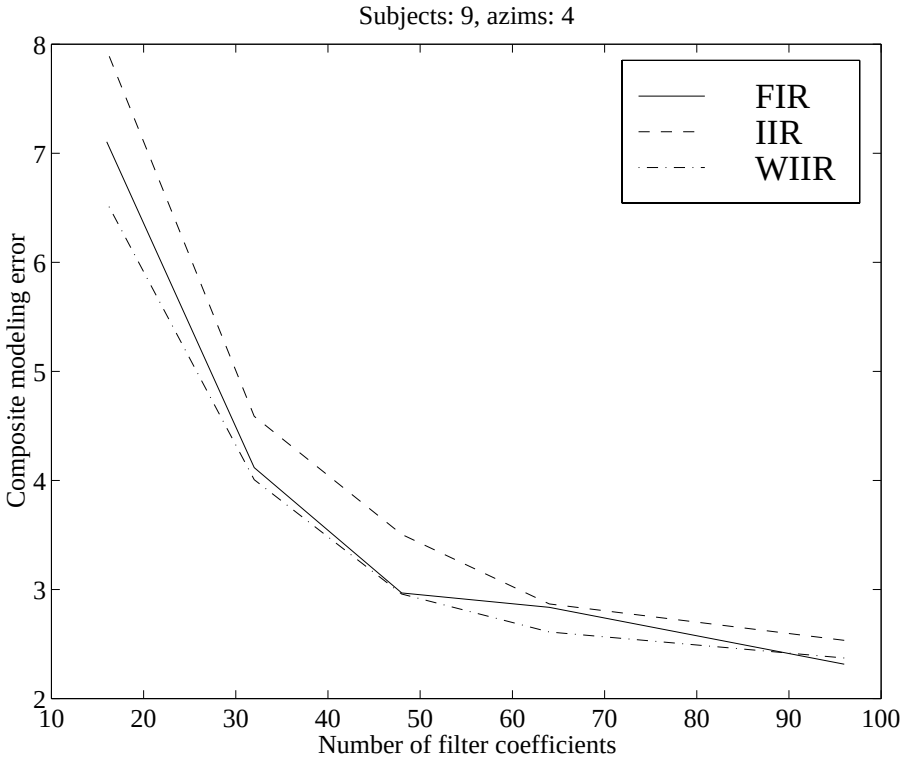


Figure 5.35: Binaural auditory model results for 9 subjects and 4 azimuth angles using three different filter design methods. The composite (L + R) loudness level spectrum error estimate was used (Huopaniemi et al., 1998, 1999b).

factors: filter order (FILTSIZ), filter type (TYPE), listener (PERSON), reproduction angle (ANGLE). A covariate (ORDER) was included to represent the order (two levels) in which the test was performed. A full factorial analysis was made for all factors and covariates, employing a type III sum of squares. This analysis was repeated for both dependent variables: Localization and Timbre. A thorough discussion of the analysis of covariance results can be found in (Huopaniemi et al., 1999b). The dominant factor for both variables was that of FILTSIZ. When considering the 257-tap FIR case, compared against the reference (i.e. itself) it can be seen that the mean grading is not 5.0, as it should be (Localization: 4.8, Timbre: 4.6). This is a common phenomenon in listening tests, as often untrained listeners are not eager to employ the extremes of the grading scale (Stone and Sidel, 1993, pp. 221-222). However, these values are very close to the reference and may be within the Just Noticeable Difference (JND) for this task. Upon inspection of means as a function or PERSON, it can be seen that listeners are grading consistently with a common trend (see Figs. 5.38–5.39).

It is clear from Figs. 5.40–5.41 that there is only marginal performance difference for both variables between FIR and IIR filters for all filter sizes. Clearly, the

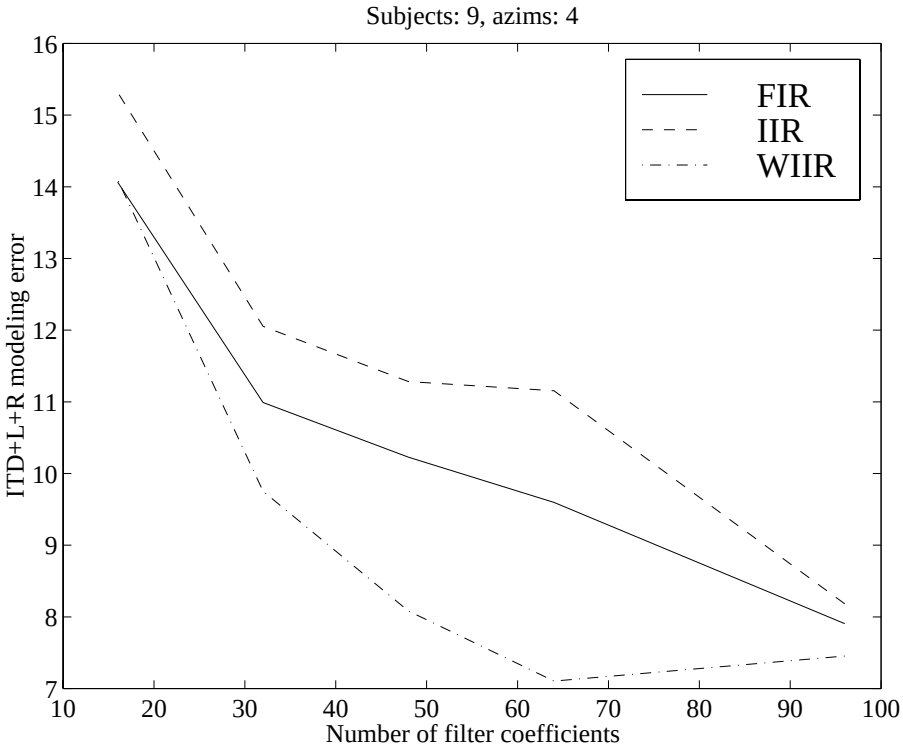


Figure 5.36: Binaural auditory model results for 9 subjects and 4 azimuth angles using three different filter design methods. The combined ITD and composite (L + R) loudness level spectrum error estimate was used (Huopaniemi et al., 1998, 1999b).

17-coefficient IIR filter, graded below 3.0 (slightly annoying), should be avoided. The WIIR design appears to reach an asymptotic level above 4.5 for 49 coefficients and above. The FIR and IIR designs only reach this quality level with the 97-coefficient filter. For these designs at least a 65-coefficient filter should be used to exceed the 4.0 (perceptible but not annoying) level. The WIIR filter type also affords quite appreciable qualities with only 33 coefficient filters, providing grades exceeding 4.0 for both variables.

Considering the second most significant factor, TYPE, in all cases the WIIR is found superior, with the FIR design in second place. Based upon the subjective data presented in Figs. 5.40–5.41, obtaining the same level of localization and timbre quality with individualised HRTFs, requires an FIR filter of approximately double the length of an equivalent quality WIIR filter. The IIR filter implementation must be even longer to reach this level of quality.

When considering the degradation as a function of ANGLE, it can be seen from Figs. 5.42–5.43 that timbre degradation is more strongly affected than localization. The highest quality for both scales occurs at 90°. This implies that it is possible to achieve the same quality level for 90° with inferior filters than at other

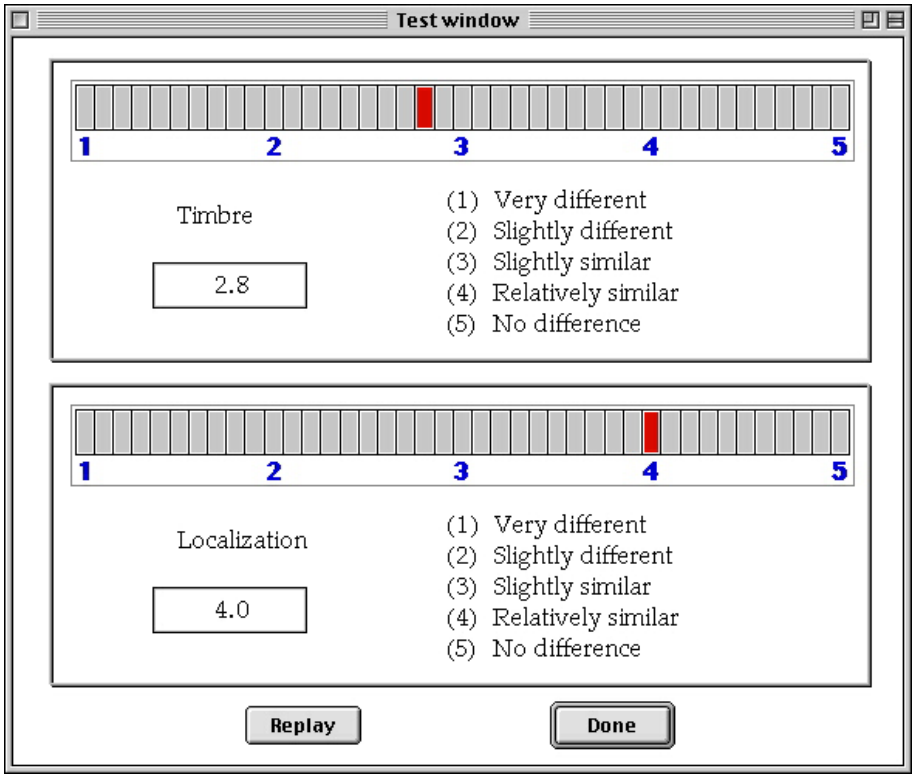


Figure 5.37: Listening test questionnaire window for HRTF filter design quality (localization and timbre grading) (Huopaniemi et al., 1998, 1999b).

angles. Blauert has discussed the human localization insensitivity (localization blur) in the horizontal plane at the sides, which may provide an explanation for this phenomenon (Blauert, 1997). It can also be considered that listeners are more sensitive to timbre, and to a lesser extent localization degradation in the frontal direction.

In this experiment (Huopaniemi et al., 1998, 1999b), the HRTF filter design problem was addressed from the objective and subjective evaluation point of view. Filter design methods taking into account the non-uniform frequency resolution of the human ear were studied and summarized. A new technique for deriving an objective HRTF filter design error was incorporated, based on a binaural auditory model. Subjective listening tests were performed to compare the theoretical model results with empirical localization performance.

The results suggest the following conclusions:

- The high-frequency spectral content present in HRTFs can be smoothed using auditory criteria without degrading localization performance.

- A binaural auditory model can be used to give a quantitative prediction of perceptual HRTF filter design performance.

- A warped IIR filter of order 16 (33 coefficients) appears to be sufficient for retaining most of the perceptual features of HRTFs in individual subjective analysis. This conclusion is supported by the objective analysis results.

In conclusion, it can be stated that filter design methods for 3-D sound can gain considerable efficiency when an auditory frequency resolution is used. The non-uniform frequency resolution can be approximated using pre-smoothing, weighting functions, or frequency warping. The binaural auditory model outputs and listening test results gave similar results in terms of detectable (perceptually audible) differences in original and approximated HRTFs. The required filter length for high-quality 3-D sound synthesis is, however, also dependent on the incident angle of the incoming sound.

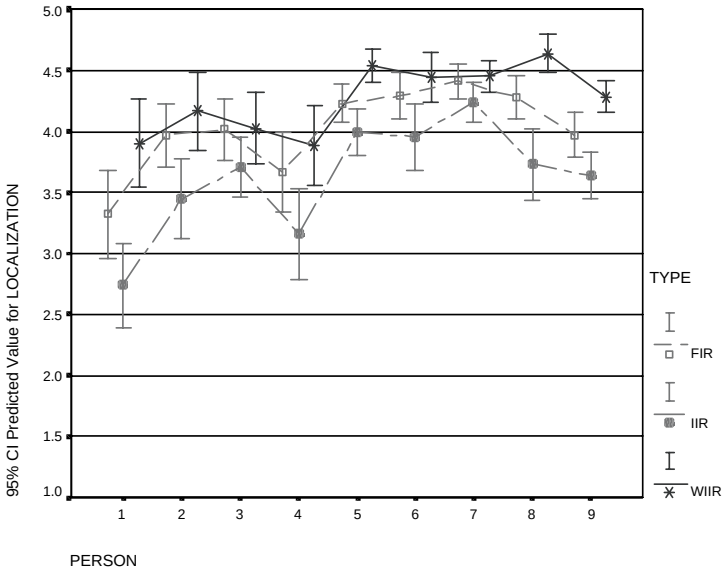


Figure 5.38: Listening test results for three filter types (FIR, IIR, WIIR). Predicted values as a function of test persons. Tested variable: localization (Huopaniemi et al., 1998, 1999b).

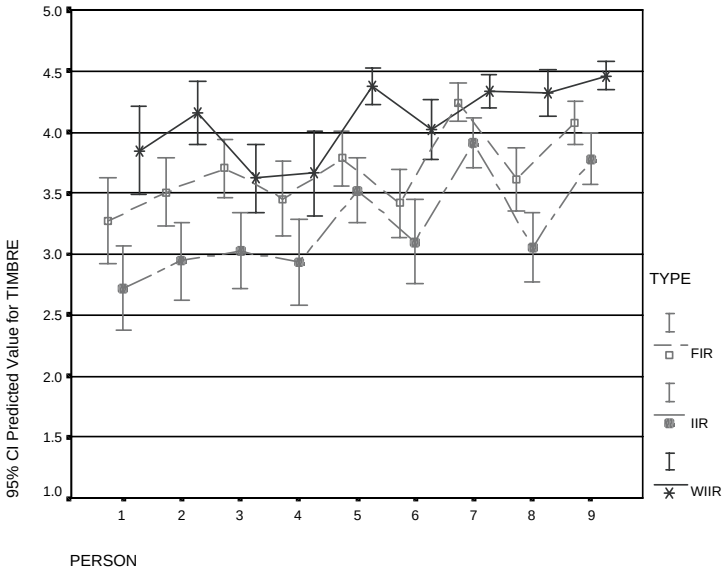


Figure 5.39: Listening test results for three filter types (FIR, IIR, WIIR). Predicted values as a function of test persons. Tested variable: timbre (Huopaniemi et al., 1998, 1999b).

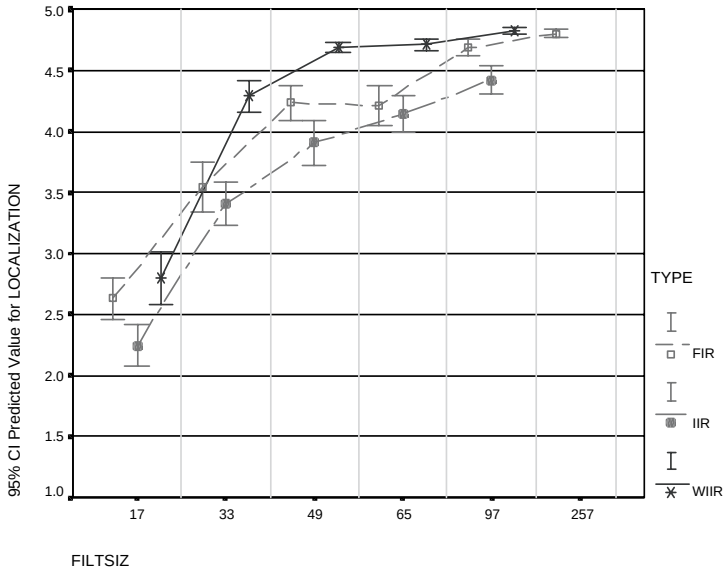


Figure 5.40: Listening test results for three filter types (FIR, IIR, WIIR). Predicted values as a function of filter type and order. Tested variable: localization (Huopaniemi et al., 1998, 1999b).

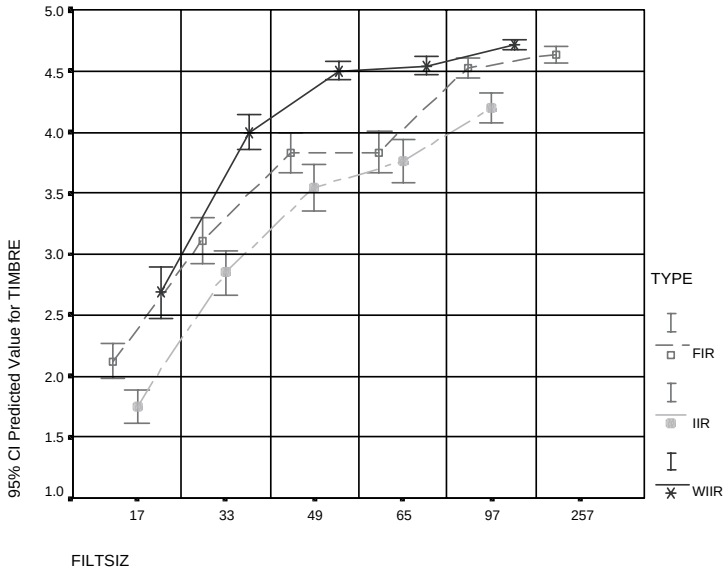


Figure 5.41: Listening test results for three filter types (FIR, IIR, WIIR). Predicted values as a function of filter type and order. Tested variable: timbre (Huopaniemi et al., 1998, 1999b).

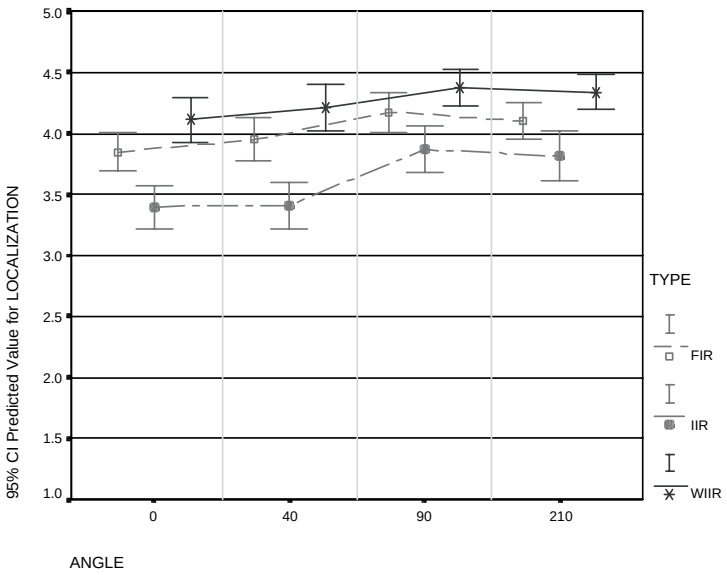


Figure 5.42: Listening test results for three filter types (FIR, IIR, WIIR). Predicted values as a function of presentation angle. Tested variable: localization (Huopaniemi et al., 1998, 1999b).

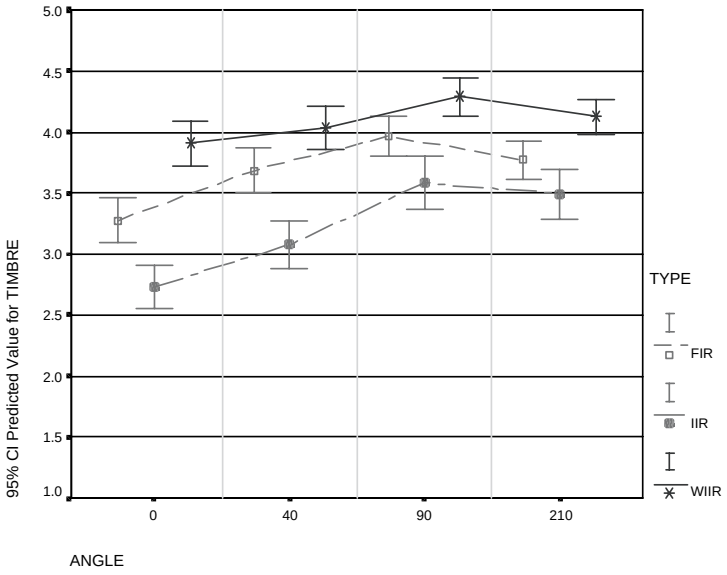


Figure 5.43: Listening test results for three filter types (FIR, IIR, WIIR). Predicted values as a function of presentation angle. Tested variable: timbre (Huopaniemi et al., 1998, 1999b).

5.6.4 Effects of HRTF Preprocessing and Equalization

This section presents the results of preprocessing and equalization in HRTF filter design in objective and subjective experiments (Huopaniemi and Smith, 1999). First the HRTF data, the methods of filter design, and the objective analysis tool shall be described. The task was to investigate the behavior of different filter designs with varying preprocessing. The goal was to determine and verify the quality of reproduction using objective evaluation.

Description of HRTF Data

A set of 5 human test subject HRTFs were chosen as data for the experiment (subjects AFW, SJX, SOS, SOU, SOW from the Wightman&Kistler HRTF database). HRTFs from the right ear of each subject at 30° azimuth intervals ($-150 : 30 : 180°$) and 0° elevation were used.

The minimum-phase reconstruction was carried out using windowing in the cepstral domain (as discussed in Section 5.1.1). The minimum + interaural excess phase approximation method was used to find the ITD for each incident angle. The ITD was inserted in HRTF synthesis as a non-fractionally addressed delay line. Two sets of HRTFs were derived: the original measurements (loudspeaker and microphone responses deconvolved) and the diffuse-field equalized versions (using Eq. 5.22).

In the pilot test six different minimum-phase HRTF approximations, six different smoothing techniques, and three different filter design orders were used:

- Filter Order: 24, 12, 8
- Filter Type: Baseline comparison (FIR windowing method[13]), Yulewalk (Mathworks, 1994), Prony (Mathworks, 1994), Invfreqz (Mathworks, 1994), Steiglitz-McBride (Mathworks, 1994), BMT (Mackenzie et al., 1997), CF (Gutknecht et al., 1983)
- Smoothing Type: No smoothing, Bark, ERB, 1/3-octave, 1/10-octave, Cepstral (see 5.2.2 for details on the smoothing algorithms)
- Equalization: No equalization, diffuse-field equalization

In the final experiment and statistical analysis of the results, all previous methods except the CF and cepstral smoothing were present.

In order to be able to compare the binaural auditory model outputs (Section 5.5.1) for different filter designs, a suitable error measure had to be defined. A "monaural loudness-level spectrum error" has been defined as the root-mean-square (RMS) difference between the loudness level (computed by the binaural

[13] The filter order for the baseline FIR filters is double to that of the IIR filters in order for the comparison to be relevant. Thus the corresponding FIR orders are 48, 24, and 16.

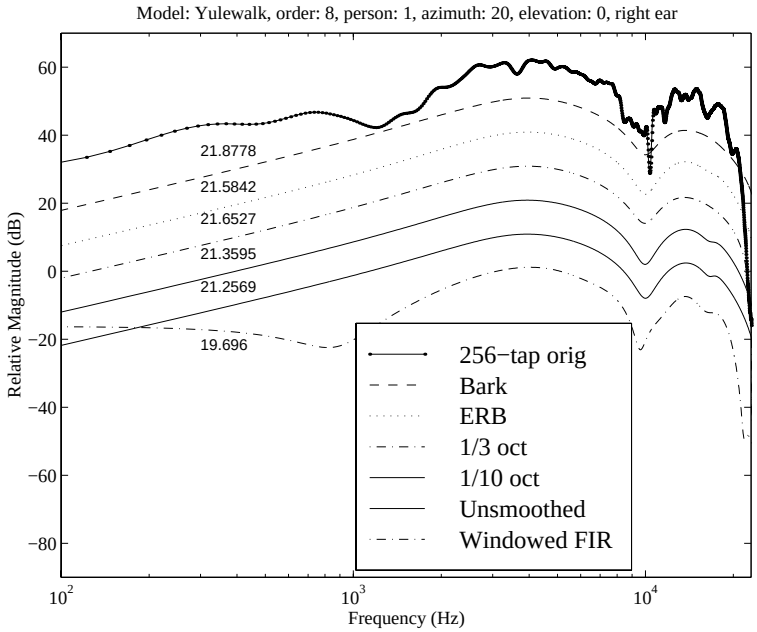


Figure 5.44: Yulewalk IIR modeling, order 8, auditory model error shown in graphs (Huopaniemi and Smith, 1999).

auditory model) of a reference HRTF (256-tap FIR) and that of an approximating filter. The designed filters varied according to spectral smoothing used, equalization, filter type, and filter order. Examples of auditory model output for two filter design methods (Yulewalk, BMT) and different smoothing techniques are illustrated in Figs. 5.44–5.45.

Results and Discussion

The filter designs and binaural auditory model simulations were carried out in Matlab (Mathworks, 1994). A total of 9360 filters were designed from the HRTF database of 5 human subjects. Statistical analysis and plotting was conducted using SPSS. The results are graphically summarized in Figs. 5.46–5.51. The results show clearly that diffuse-field equalized filters generally provide a lower perceptual error, and the error variance across HRTF designs is smaller. This is due to flattening of the spectra as discussed in previous sections. Furthermore, the baseline filter design (Windowed FIR) seems to be outperformed by other design methods, especially the Hankel norm based BMT. (The CF method performed similarly to the BMT, but not quite as good, so it is not shown in the figures.) It is notable that by choosing proper preprocessing techniques (namely, minimum-phase reconstruction, diffuse-field equalization and auditory smoothing), the choice of filter design methods becomes a non-crucial task, although clear differences can still be found between methods.

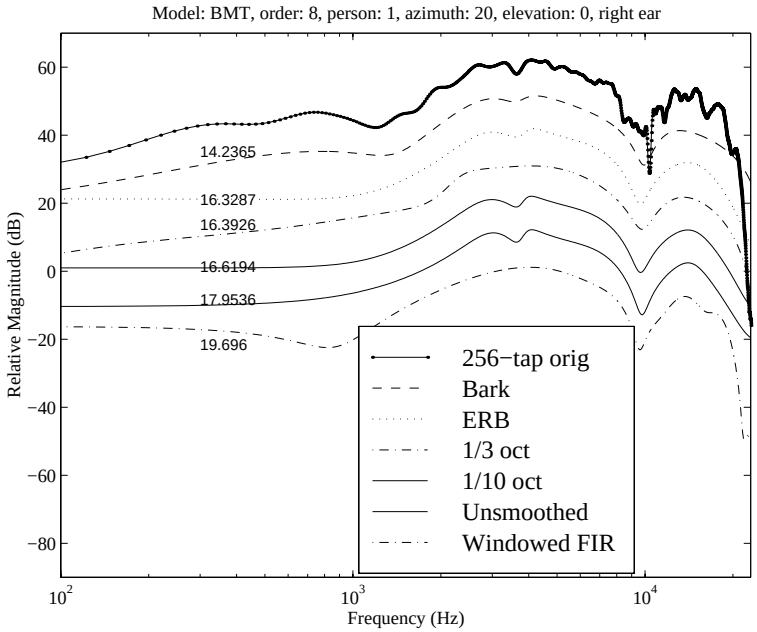


Figure 5.45: BMT IIR modeling, order 8, auditory model error shown in graphs (Huopaniemi and Smith, 1999).

In this study (Huopaniemi and Smith, 1999), methods for temporal and spectral preprocessing of binaural digital filters were presented. The perceptual quality of different preprocessing and smoothing schemes was compared using a simple binaural auditory model. The results suggest the following conclusions:

- controlled smoothing applied to HRTFs results in better (psychoacoustically motivated) magnitude fitting
- diffuse-field equalization allows for lower order filter designs
- Hankel norm optimal algorithms (BMT) perform slightly better at lower filter orders when compared to other tested methods, but in general proper choices of preprocessing diminish the difference between filter designs and error norms.

It should be noted that the objective validation of the methods was only performed on the magnitude response. Since all designed filters were minimum-phase, the phase characteristics are likely to have little variation across designs. Nevertheless, more detailed auditory analysis of combined temporal and spectral preprocessing modeling aspects will be performed in the future. Furthermore, the perceptual validity of diffuse-field equalization should be verified in subjective listening experiments. The validity of the objective binaural auditory model in comparing the different designs can be questioned, but in a previous study

(Huopaniemi et al., 1998, 1999b) the model was found to correspond well to subjective listening experiment results.

5.7 Discussion and Conclusions

In this chapter, the principles of spatial hearing and binaural synthesis were first presented. The author has contributed to binaural filter design by introducing new methods based on modeling the human auditory frequency resolution. New methods proposed by the author based on frequency warping and balanced model truncation were discussed. Furthermore, techniques and experiments on objective and subjective evaluation of binaural filter design were presented. Novel methods estimating the quality of HRTF filters have been created by the author, enabling objective and subjective evaluation of binaural systems. Discussion and results of four experiments carried out by the author were presented.

The main conclusion of this chapter is that the computational efficiency of binaural processing can be improved by taking into account the auditory resolution in the filter design stage. This finding is supported by theoretical analysis and objective metrics as well as by subjective listening experiments.

In the next chapter, aspects in crosstalk canceled binaural filter design and implementation are discussed.

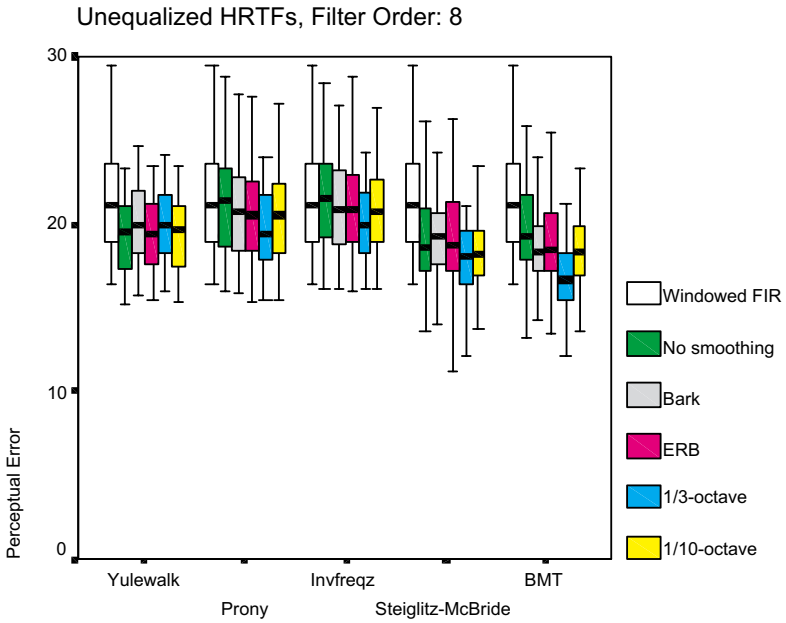


Figure 5.46: Design of unequalized HRTFs, order 8 (Huopaniemi and Smith, 1999).

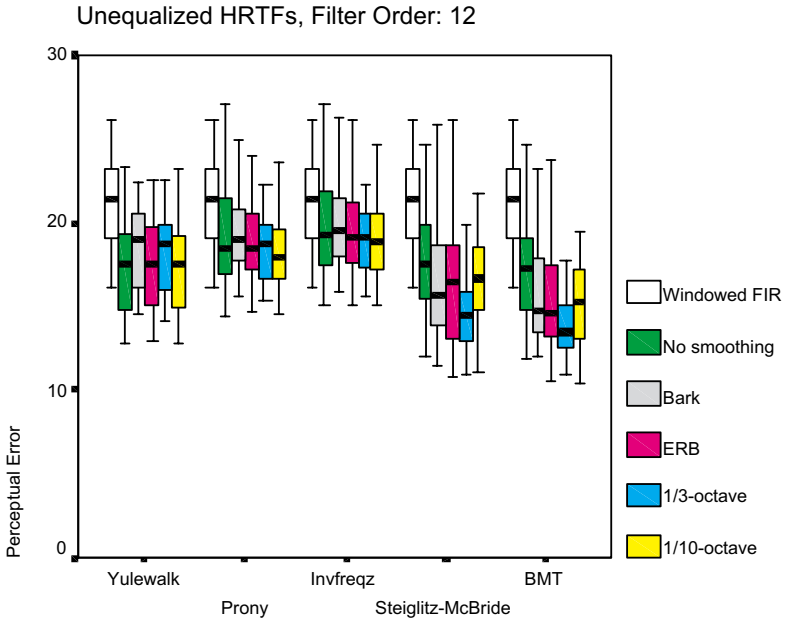


Figure 5.47: Design of unequalized HRTFs, order 12 (Huopaniemi and Smith, 1999).

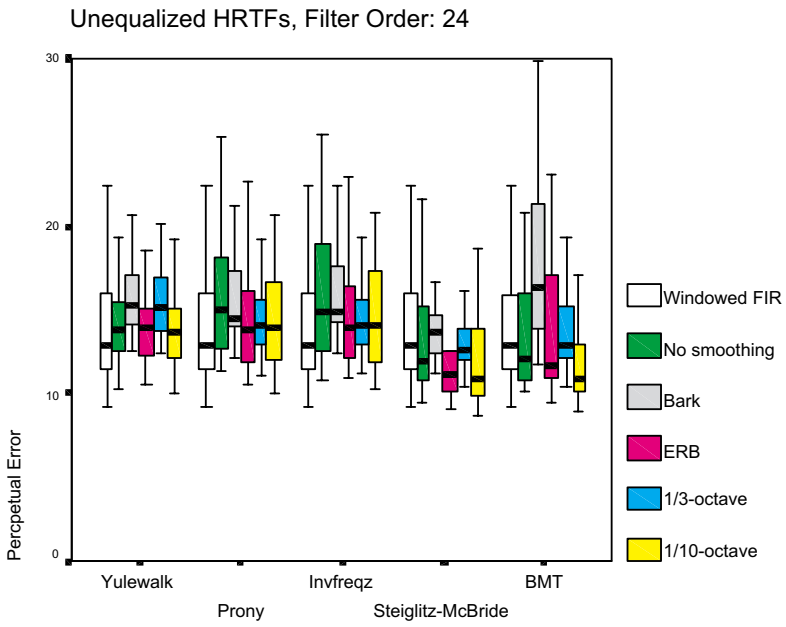


Figure 5.48: Design of unequalized HRTFs, order 24 (Huopaniemi and Smith, 1999).

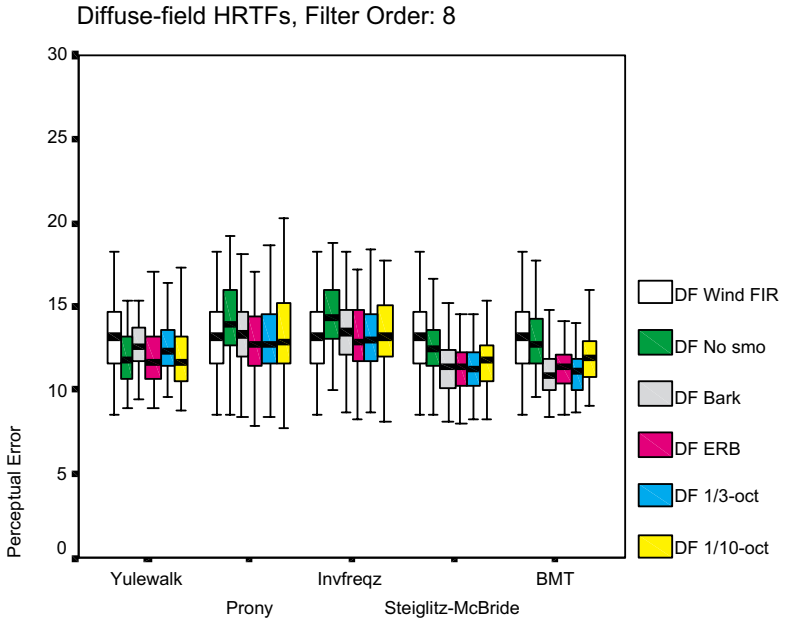


Figure 5.49: Design of DF-equalized HRTFs, order 8 (Huopaniemi and Smith, 1999).

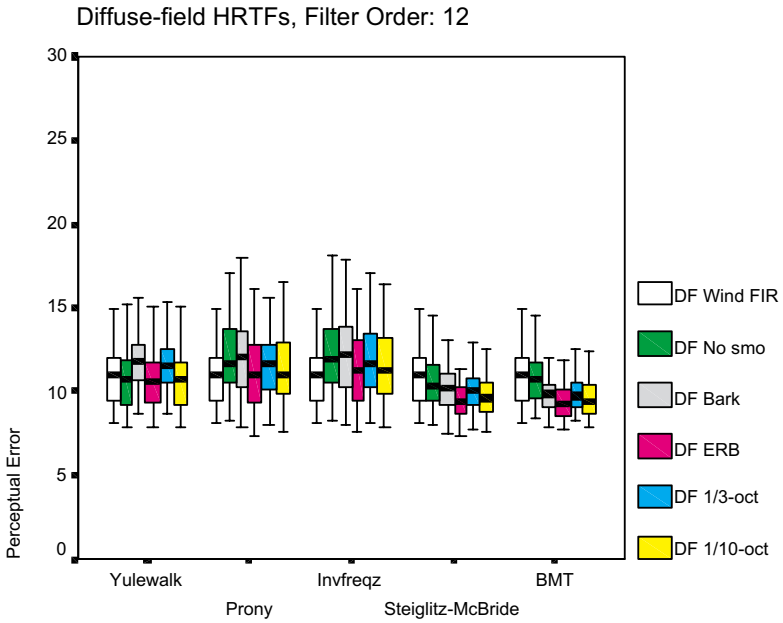


Figure 5.50: Design of DF-equalized HRTFs, order 12 (Huopaniemi and Smith, 1999).

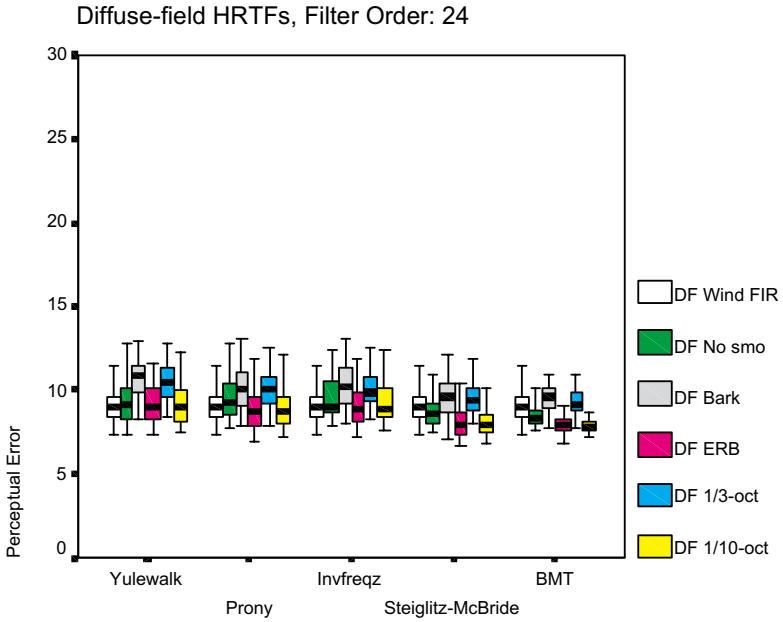


Figure 5.51: Design of DF-equalized HRTFs, order 24 (Huopaniemi and Smith, 1999).

Chapter 6

Crosstalk Canceled Binaural Reproduction

The idea of presenting crosstalk-compensated binaural information over a pair of loudspeakers was introduced almost 40 years ago (Bauer, 1961), and first formulated into practice by Schroeder and Atal (Schroeder and Atal, 1963; Atal and Schroeder, 1966). They described the use of a crosstalk cancellation filter for converting binaural recordings made in concert halls for loudspeaker listening. Their impressions of listening to loudspeaker reproduced dummy-head recordings were "nothing less than amazing". However, they observed the limitations of the listening area, the "sweet spot", which has remained an unsolved problem in loudspeaker binaural reproduction ever since.

The psychophysical and acoustical basis for manipulating stereophonic signals has been studied in, e.g., (Bauer, 1961; Blauert, 1997). Blauert's term "summing localization" corresponds to manipulation of level and/or time delay differences in stereophonic reproduction in such a way that the sound image appears shifted from the position between the loudspeakers ((Blauert, 1997, pp. 201-271)). Damaske studied loudspeaker reproduction issues and formulated the basic theories further in the TRADIS project (True Reproduction of All Directional Information by Stereophony) (Damaske, 1971). He conducted studies on sound image quality deterioration as a function of listener placement. The transaural theory was refined and to some extent revitalized by works of Cooper and Bauck (Cooper and Bauck, 1989). They created a concept of *spectral stereo* and gave a new term to crosstalk canceled binaural presentation, *transaural processing*. The transaural stereo concept originally applied simplified head models for crosstalk cancelling. These techniques have been further developed by, for example, (Kotorynski, 1990; MacCabe and Furlong, 1991; Rasmussen and Juhl, 1993; Juhl, 1993; Walsh and Furlong, 1995) to include improved head models and more sophisticated signal processing techniques. Recently, adaptive processing systems have been presented by (Nelson et al., 1992, 1995, 1996a,b) that take into account multiple point equalization and the possibility of a wider listening area.

Also, the concept of *stereo dipole* has been introduced, where closely spaced loudspeakers are used to generate virtual sources (Bauck and Cooper, 1996; Watanabe et al., 1996; Takeuchi et al., 1997; Kirkeby et al., 1997). A method for robustness analysis of crosstalk canceling using different loudspeaker spacings has been proposed by Ward and Elko (1998, 1999). Fundamental and theoretical limitations of cross-talk canceled binaural audio processing have been studied by Kyriakakis (1998); Kyriakakis et al. (1999). An excellent recent study of cross-talk canceling methods has been published by Gardner (1998a).

In this chapter, methods for crosstalk canceling and virtual loudspeaker implementation in symmetrical and asymmetrical cases will be introduced. Filter design and practical implementation issues are discussed. As a case study, a novel filter design algorithm for virtual loudspeakers based on joint minimum-phase reconstruction and warped filters is presented. The quality of the filter design has been verified in a round robin subjective test on virtual home theatre algorithms (Zacharov et al., 1999).

6.1 Theory of Crosstalk Canceling

The listening process in a two-channel loudspeaker case can be formulated in matrix notation. The basic theory overviewed here originates to works by (Schroeder and Atal, 1963), and (Cooper and Bauck, 1989), the notation follows that of (Gardner, 1995). A situation is considered, which is depicted in Fig. 6.1, where $\hat{x}_l(n)$ and $\hat{x}_r(n)$ are binaural signals delivered to the speakers, and $y_l(n)$ and $y_r(n)$ are the resulting signals registered at the listener's ears. The sound propagation in a stereophonic system can be described by the following equations:

$$y(n) = H(z)\hat{x}(n) \tag{6.1}$$

where

$$y(n) = \begin{bmatrix} y_l(n) \\ y_r(n) \end{bmatrix}, \quad \hat{x}(n) = \begin{bmatrix} \hat{x}_l(n) \\ \hat{x}_r(n) \end{bmatrix}, \quad H(z) = \begin{bmatrix} H_{ll}(z) & H_{rl}(z) \\ H_{lr}(z) & H_{rr}(z) \end{bmatrix} \tag{6.2}$$

If $x(n) = [x_l(n)x_r(n)]^T$ is the binaural signal that is to be delivered to the ears, an inverse matrix $G(z)$ must be found to the system transfer matrix $H(z)$ such that $G(z) = H(z)^{-1}$ and $\hat{x}(n) = G(z)x(n)$. The exact inverse matrix can be analytically written as

$$G(z) = \frac{1}{H_{ll}(z)H_{rr}(z) - H_{lr}(z)H_{rl}(z)} \begin{bmatrix} H_{rr}(z) & -H_{rl}(z) \\ -H_{lr}(z) & H_{ll}(z) \end{bmatrix} \tag{6.3}$$

However, an ideal inverse filter can not necessarily be computed (due to possible non-minimum phase functions in the denominator) analytically, and thus the

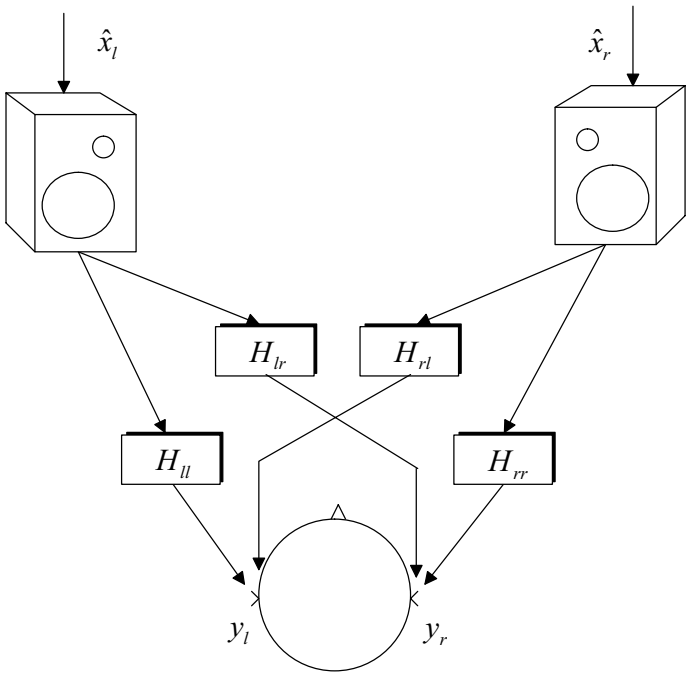


Figure 6.1: Loudspeaker-to-ear transfer functions in stereophonic listening arrangement.

equality $y(n) = H(z)\hat{x}(z) = H(z)G(z)x(n)$ should be expressed in the form

$$\begin{bmatrix} y_l(n) \\ y_r(n) \end{bmatrix} = \begin{bmatrix} H_{ll}(z) & H_{rl}(z) \\ H_{lr}(z) & H_{rr}(z) \end{bmatrix} \begin{bmatrix} G_{ll}(z) & G_{rl}(z) \\ G_{lr}(z) & G_{rr}(z) \end{bmatrix} \begin{bmatrix} x_l(n) \\ x_r(n) \end{bmatrix} \tag{6.4}$$

From this point forward, the formulations are divided into two categories: symmetrical and asymmetrical listening position.

6.1.1 Symmetrical Crosstalk Canceling

In the first case, symmetry is assumed in the listening situation and head geometry (that is, $H_{lr}(z) = H_{rl}(z)$ and $H_{ll}(z) = H_{rr}(z)$). Thus Eqs. 6.1, 6.2 and 6.4 can be simplified using ipsilateral responses ($H_i(z) = H_{ll}(z) = H_{rr}(z)$) for the HRTF to the same side and contralateral responses ($H_c(z) = H_{lr}(z) = H_{rl}(z)$ for the HRTF to the opposite side:

$$\begin{bmatrix} y_l(n) \\ y_r(n) \end{bmatrix} = \begin{bmatrix} H_i(z) & H_c(z) \\ H_c(z) & H_i(z) \end{bmatrix} \begin{bmatrix} \hat{x}_l(n) \\ \hat{x}_r(n) \end{bmatrix} \tag{6.5}$$

$$\begin{bmatrix} y_l(n) \\ y_r(n) \end{bmatrix} = \begin{bmatrix} G_i(z) & G_c(z) \\ G_c(z) & G_i(z) \end{bmatrix} \begin{bmatrix} H_i(z) & H_c(z) \\ H_c(z) & H_i(z) \end{bmatrix} \begin{bmatrix} x_l(n) \\ x_r(n) \end{bmatrix} \tag{6.6}$$

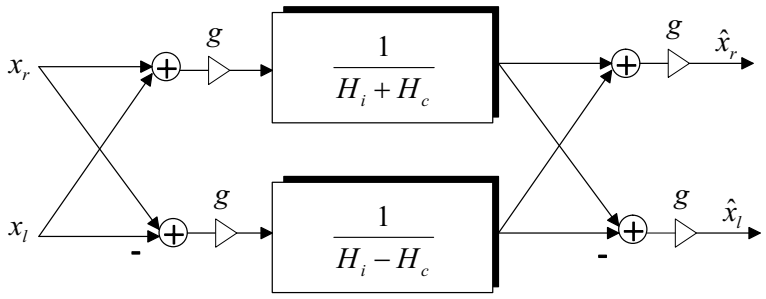


Figure 6.2: Shuffler implementation of cross-talk canceling filters in a symmetric listening arrangement.

The exact inverse filter **G** can similarly be written as

$$G(z) = \frac{1}{H_i^2(z) - H_c^2(z)} \begin{bmatrix} H_i(z) & -H_c(z) \\ -H_c(z) & H_i(z) \end{bmatrix} \tag{6.7}$$

(Cooper and Bauck, 1989) have proposed a "shuffler" structure for the realization of crosstalk canceling filters. This system involves generation of the sum and difference for input signals $x_l(n)$ and $x_r(n)$, and undoing the sum and difference after the filtering, as depicted in Fig. 6.2. The sum and difference operations are made possible by a unitary matrix **D**, which is called the shuffler matrix or MS matrix:

$$D = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{6.8}$$

It can be shown that this shuffler matrix **D** diagonalizes the matrix $G(z)$ using a similarity transformation:

$$D^{-1}G(z)D = \begin{bmatrix} \frac{1}{H_i(z)+H_c(z)} & 0 \\ 0 & \frac{1}{H_i(z)-H_c(z)} \end{bmatrix} \tag{6.9}$$

The shuffler structure is illustrated in 6.2. The normalizing gains can be commuted to a single gain of 1/2 for each channel, or just be ignored.

In this work, three different crosstalk cancellation approaches were considered. In (Cooper and Bauck, 1989), it was suggested that the use of a computational model (diffraction around a rigid sphere) would result in generalized transfer functions. These filters (Cooper approximations), although not as accurate, were claimed to give a wider and more stable listening area than with individual or dummy-head based crosstalk cancelers. In (Gardner, 1995), the following simple formulas for $H_i(z)$ and $H_c(z)$ were used[1].

$$H_i(z) = 1, \quad H_c(z) = gz^{-m}H_{lp}(z) \tag{6.10}$$

[1] Gardner (1998a) presents more advanced methods of cross-talk canceler design, but these were not applied in the current work.

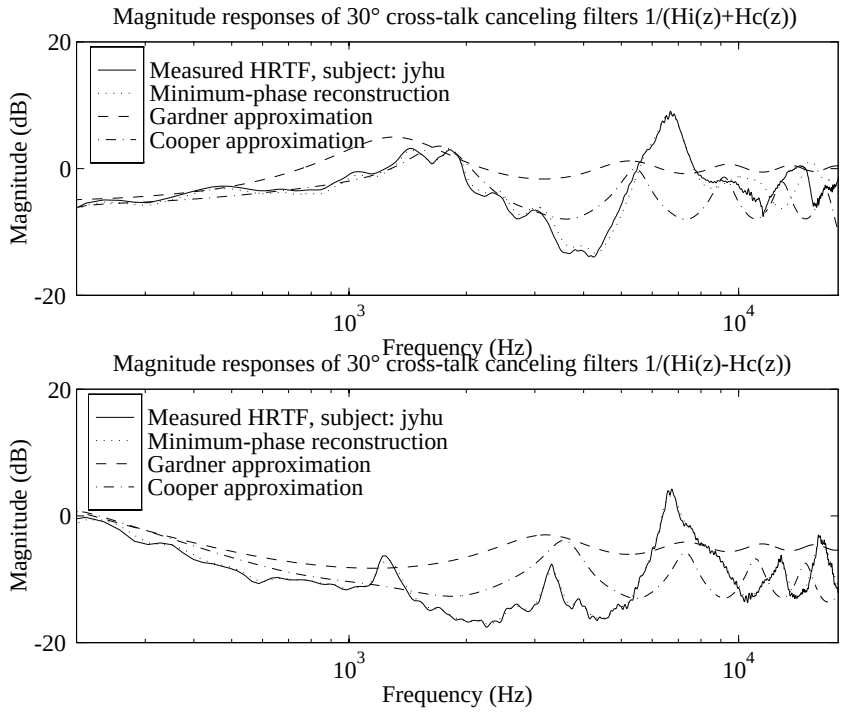


Figure 6.3: Magnitude responses of three crosstalk canceling filters for sources at 30^o azimuth.

where $g < 1$ is the interaural gain, m is the approximated frequency-independent ITD in samples, and $H_{lp}(z)$ is a one-pole lowpass filter that models the frequency-dependent head shadowing:

$$H_{lp}(z) = \frac{1-a}{1-az^{-1}} \tag{6.11}$$

where coefficient a determines the lowpass cutoff.

In Figs. 6.3 and 6.4, results for modeling crosstalk canceling with the above discussed two methods are presented. These transfer functions are compared to crosstalk canceling filters based on measured HRTFs (Riederer, 1998b). The loudspeakers in the simulation were placed at 30^o azimuth. From the magnitude response plots of Fig. 6.3 it is clearly seen that the two model-based approaches (Cooper and Gardner approximation) give predictable results only to approximately 1-2 kHz. According to Cooper and Bauck (1989, 1992), the crosstalk canceling filters, although not minimum-phase, are of *joint minimum phase*, that is, they have a common excess phase which is close to a frequency-independent delay. The delay-normalized crosstalk canceling filters are then minimum-phase. Thus the shuffling filters may be defined by their magnitude only, and the phase may be calculated, e.g., using minimum-phase reconstruction. This statement is verified in Fig. 6.4, where the measured HRTF-based crosstalk canceler and

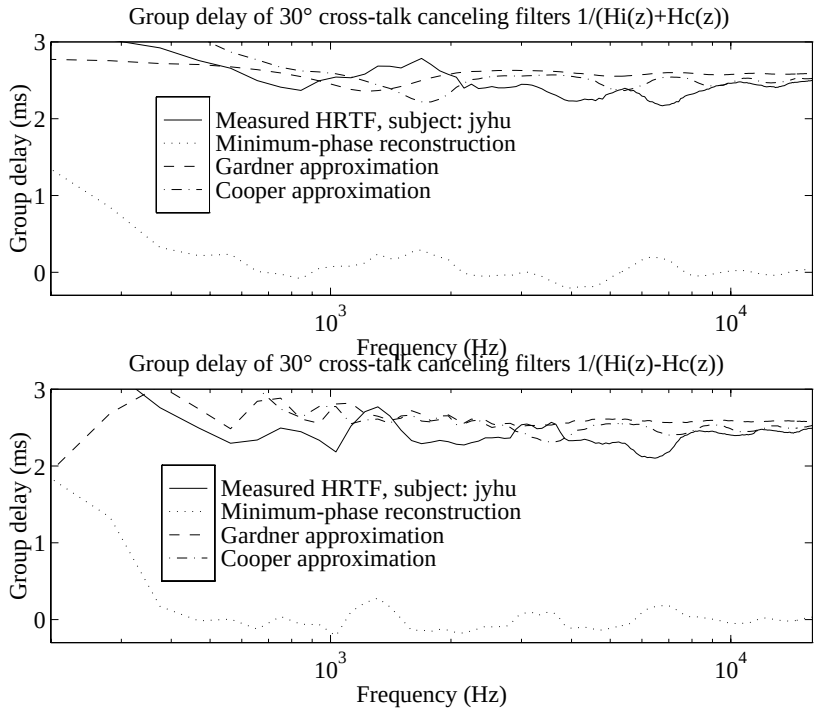


Figure 6.4: Group delay of three crosstalk canceling filters for sources at 30^o azimuth.

the minimum-phase reconstructed version differ in group delay up to 5-6 kHz approximately by a frequency-independent bulk delay. According to Kotorynski (1990), the phase distortion at higher frequencies caused by minimum-phase approximation should not be neglected, but separately modeled using an allpass filter.

6.1.2 Asymmetrical Crosstalk Canceling

In many cases the listening situation is not symmetrical. Such examples are the interior of a car, listening to music in a concert hall, or at home on a couch which is quite not symmetrically placed between the speakers. The asymmetrical listening condition leads to the fact that the shuffler structures presented in the previous section are not anymore valid. It is possible to resort to use a crosstalk canceling system illustrated in Fig. 6.1 which uses four filters. The exact inverse matrix G is given by Eq. 6.3. The problem of asymmetrical crosstalk cancellation and virtual source synthesis is discussed in Section 6.2.3.

6.1.3 Other Crosstalk Canceling Structures

The methodology for cross-talk canceling introduced by Schroeder and Atal (1963) and the shuffler structure proposed by Cooper and Bauck (1989) have been the basis for the current research. There are, however, many other crosstalk canceling structures that have been proposed in the literature. An excellent overview of existing methods is presented in (Gardner, 1998a), including feedforward and recursive crosstalk canceling for both symmetric and asymmetric listening positions.

6.2 Virtual Source Synthesis

An interesting application of cross-talk canceled binaural technology is the concept of virtual sources, first discussed in (Cooper and Bauck, 1989). The term *virtual loudspeaker* (VL, or virtual speaker, VS) and VL synthesis has been invented to mean the generation of virtual sources using cross-talk canceling techniques. This method consists of two stages: a) implementation of binaural synthesis for mono- or stereophonic source signals, and b) crosstalk cancellation for presentation with two loudspeakers. These steps may be combined to single filtering task in order to optimize computational efficiency. A topic related to virtual source synthesis is *stereo widening*, which is a general term representing methods that enhance or exaggerate the stereophonic image of sounds in two-speaker reproduction. This topic is discussed in Section 6.3.

6.2.1 Symmetrical Listening Position

In the previous section the treatment was restricted only to the problem of crosstalk canceling. In many applications of interest it is, however, desired to synthesize a virtual source or multiple virtual sources for loudspeaker listening. In other words, virtual source synthesis combines binaural processing and crosstalk cancellation. A schematic for signal transmission in virtual source synthesis is illustrated in Fig. 6.5. Signals $x_l(n)$ and $x_r(n)$ are raw input signals that are processed with binaural filters H' and cross-talk canceling filters G to deliver signals $y_l(n)$ and $y_r(n)$ at the listener's ears. In other words, the lattice structure of Fig. 6.2 can be expanded to include the transfer functions $H_i'(z)$ and $H_c'(z)$ that, for example, place the virtual source at $+90^o$ azimuth. It may be assumed that the loudspeakers are placed at $\pm 10^o$ azimuth relative to the listener, and thus those HRTFs will be used for crosstalk canceling transfer functions $H_i(z)$ and $H_c(z)$. This structure, as well as all the other shuffler structures, can be realized using two digital filters.

The structure depicted in Fig. 6.6 can also be computed using a direct approach without the lattice form. In this case the transfer functions of the mono-

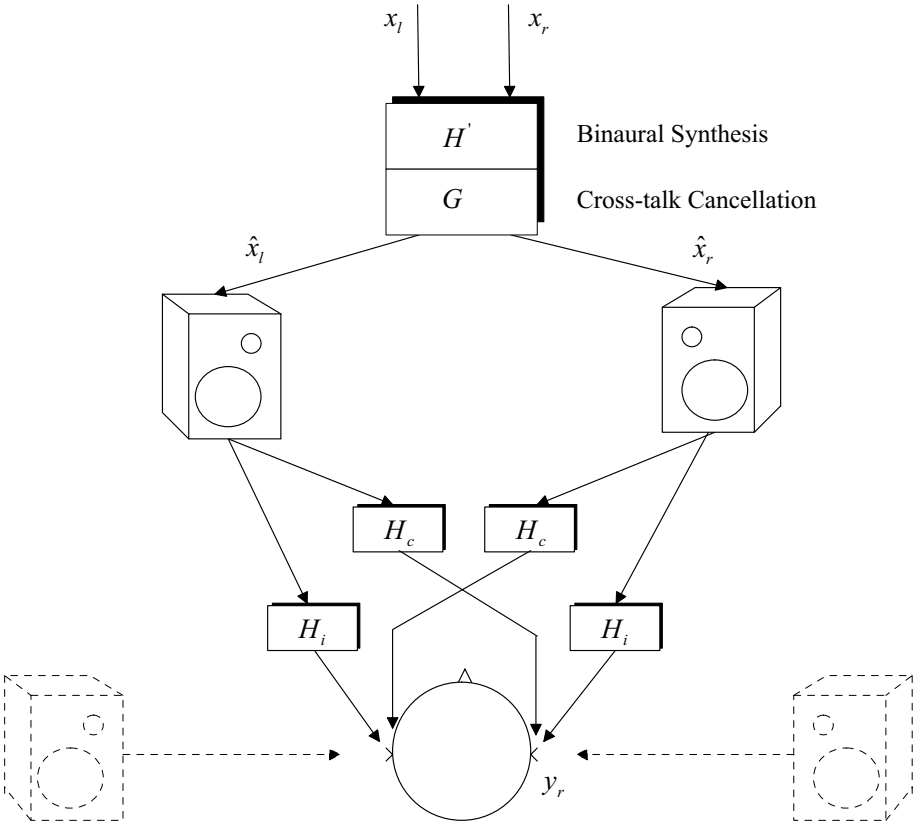


Figure 6.5: Signal transmission for cross-talk cancellation in virtual loudspeaker synthesis.

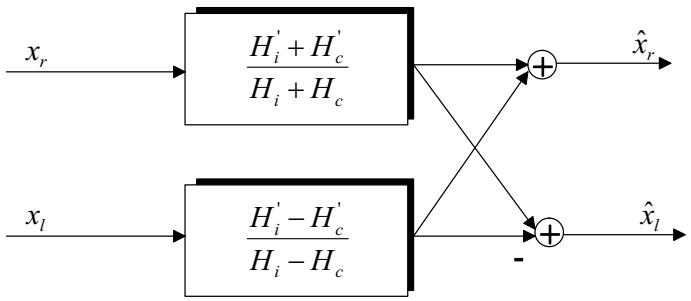


Figure 6.6: Shuffler structure for VL synthesis of one virtual source.

phonic-to-crosstalk canceled converter are of the form:

$$\hat{x}_l(n) = \frac{H_i(z)H_i'(z) - H_c(z)H_c'(z)}{H_i^2(z) - H_c^2(z)} x_l(n) \tag{6.12}$$

$$\hat{x}_r(n) = \frac{H_i(z)H_c'(z) - H_c(z)H_i'(z)}{H_i^2(z) - H_c^2(z)} x_r(n) \tag{6.13}$$

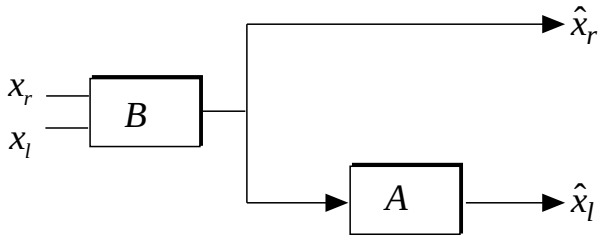


Figure 6.7: Spectral stereo shuffler for synthesis of one virtual loudspeaker.

These equations will also be used subsequently in the case of asymmetrical crosstalk cancellation. Cooper and Bauck's spectral stereo system (Cooper and Bauck, 1989; Rasmussen and Juhl, 1993; Walsh and Furlong, 1995) uses a rearrangement of equations 6.12 and 6.13 in the following way. Two filters, A and B, are defined which satisfy the relations $\hat{x}_r(n) = B(z)x_r(n)$ and $\hat{x}_l(n) = B(z)A(z)x_l(n)$ when

$$B(z) = \frac{H_i(z)H_c'(z) - H_c(z)H_i'(z)}{H_i^2(z) - H_c^2(z)} \tag{6.14}$$

$$A(z) = \frac{H_i(z)H_i'(z) - H_c(z)H_i'(z)}{H_i(z)H_c'(z) - H_c(z)H_i'(z)} \tag{6.15}$$

This structure is illustrated in Fig. 6.7. According to (Walsh and Furlong, 1995), this rearrangement of filters enables us to divide the system to sound localization part (filter A) and virtual sound image equalization part (filter B). The validity of this statement is yet to be verified although it is clear that filter B cannot contain any interaural phase information (because it is applied to both channels), and furthermore, filter A is only applied to one of the channels. The drawback in this structure is that it can not be rearranged for two-channel virtual loudspeaker implementation with two filters, as is done in the following.

The structures presented in Eqs. 6.12 and 6.13 can be directly applied to the implementation of two virtual sources. However, if it is assumed that the listening position and loudspeaker placements are symmetrical with respect to the median plane, the structures can be further simplified. It is possible to reduce the order of the filters to two with the aid of the shuffler structure of Fig. 6.6. The resulting two-filter structure for two virtual sources is illustrated in Fig. 6.8 (Kotorynski, 1990; Toshiyuki et al., 1994; Jot et al., 1995). This proposed two-filter structure is suitable for practical implementations due to efficient realization. In matrix form, the normalized underlying equations can be written using Eqs. 6.5, 6.6 and 6.9:

$$GH' = \frac{1}{4}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} \frac{H_i'(z)+H_c'(z)}{H_i(z)+H_c(z)} & 0 \\ 0 & \frac{H_i'(z)-H_c'(z)}{H_i(z)-H_c(z)} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{6.16}$$

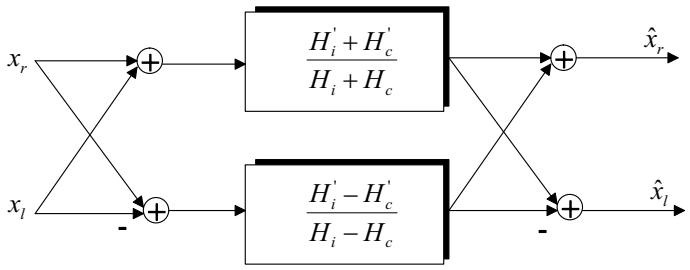


Figure 6.8: Shuffler structure for VL synthesis of two sound sources.

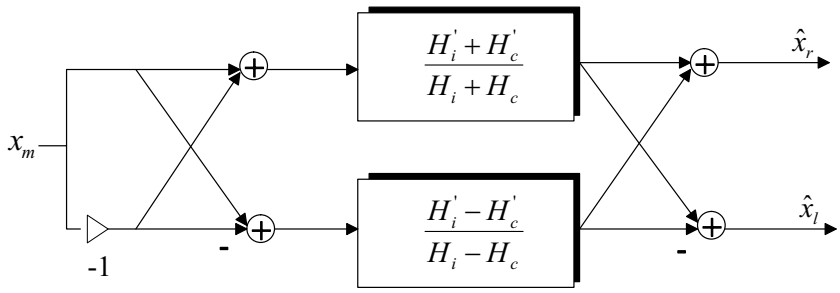


Figure 6.9: Monophonic decorrelation shuffler for synthesizing two VLs.

6.2.2 Decorrelation of Virtual Sources

In certain applications it is of interest to create two virtual sources of a monophonic input (for example, a monophonic surround channel). It is clear that creating two virtual speakers for a monophonic input results in a sound that is heard in the middle or even inside the listener's head. Thus it is necessary to decorrelate the monophonic input before it is fed to the virtual speaker filters. The most straightforward method for decorrelating the input is to invert the phase of the other input channel as shown in Figs. 6.9 and 6.10. Other possible methods are delaying the other channel, or performing more demanding decorrelation analysis. The principle of converting a monophonic signal to two virtual speakers is illustrated in Fig. 6.9 (Toshiyuki et al., 1994). Informal listening tests carried out using this structure have shown that although negation has an effect on the tonal quality of the sound, the virtual speaker image is still audible and quite substantial. The structure presented in Fig. 6.9 can be further simplified by noticing that the upper filter block actually vanishes due to the decorrelative negation. Thus, a one-filter configuration for generating two virtual speakers from a monophonic input is possible. This structure is depicted in Fig. 6.10.

Finally, it may be concluded that simple phase inverting and structures shown in Figs. 6.9–6.10 are preferable for virtual speaker synthesis of a monophonic Dolby Pro Logic surround channel.

6.2.3 Asymmetrical Listening Position

In the case of an asymmetrical listening position, the shuffling structures can not generally be used. In Fig. 6.11, a direct implementation of two virtual loudspeakers is presented in an asymmetrical listening position. This structure requires four parallel filters. The asymmetrical positions can, however, be taken into account in the filter design and optimization techniques for shuffler structures. The following equations (based on Fig. 6.11) can be used to create virtual sources described by filters $H_i'(z)$ and $H_c'(z)$ for an aymmetrical listening position.

$$\hat{x}_l(n) = \frac{H_{rr}(z)H_i'(z) - H_{rl}(z)H_c'(z)}{H_{ll}(z)H_{rr}(z) - H_{lr}(z)H_{rl}(z)}x_l(n) + \frac{H_{rr}(z)H_c'(z) - H_{rl}(z)H_i'(z)}{H_{ll}(z)H_{rr}(z) - H_{lr}(z)H_{rl}(z)}x_r(n) \tag{6.17}$$

$$\hat{x}_r(n) = \frac{H_{ll}(z)H_c'(z) - H_{lr}(z)H_i'(z)}{H_{ll}(z)H_{rr}(z) - H_{lr}(z)H_{rl}(z)}x_r(n) + \frac{H_{ll}(z)H_i'(z) - H_{lr}(z)H_c'(z)}{H_{ll}(z)H_{rr}(z) - H_{lr}(z)H_{rl}(z)}x_l(n) \tag{6.18}$$

6.2.4 Binaural and Crosstalk Canceled Binaural Conversion Structures

In Fig. 6.12, digital filter and shuffler structures (proposed by Cooper and Bauck) for converting binaural and crosstalk canceled binaural signals are presented (Jot et al., 1995). The use of shuffler structures is valid for symmetrical loudspeaker and listening arrangement, as discussed in Section 6.2.1.

6.2.5 Virtual Center Channel

In certain applications such as playback of 5.1-channel audio material using two loudspeakers, it may be desired to synthesize not only the surround channels,

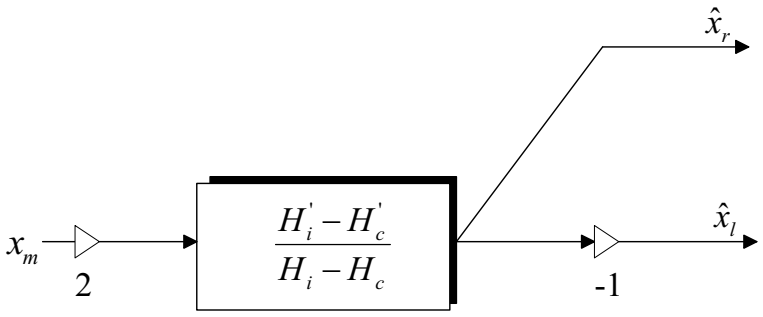


Figure 6.10: Monophonic decorrelation shuffler for synthesizing two VLs implemented with a single filter.

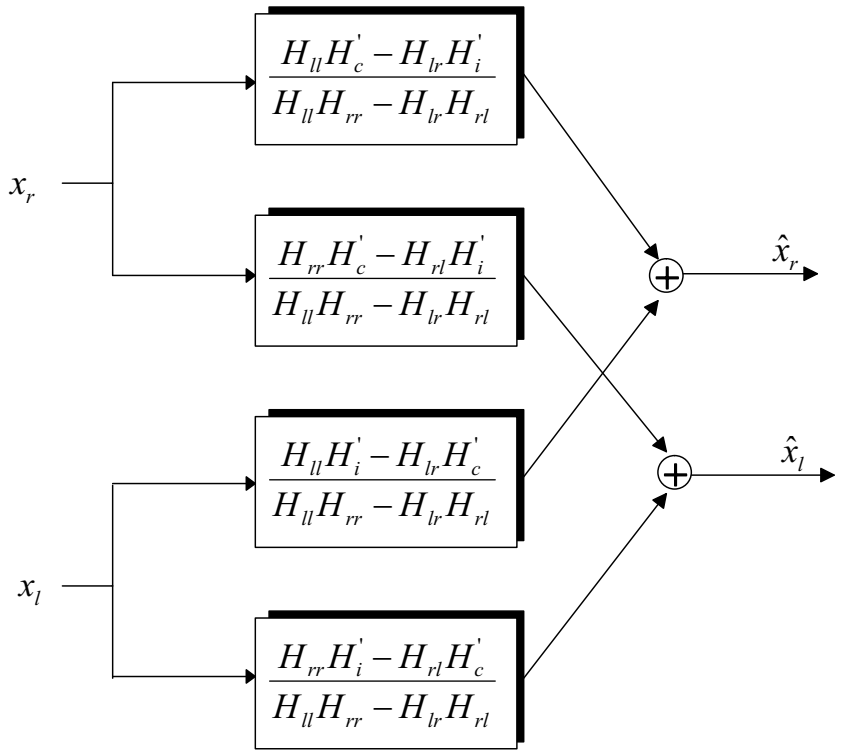


Figure 6.11: Direct structure for VL synthesis of two sound sources in asymmetrical listening.

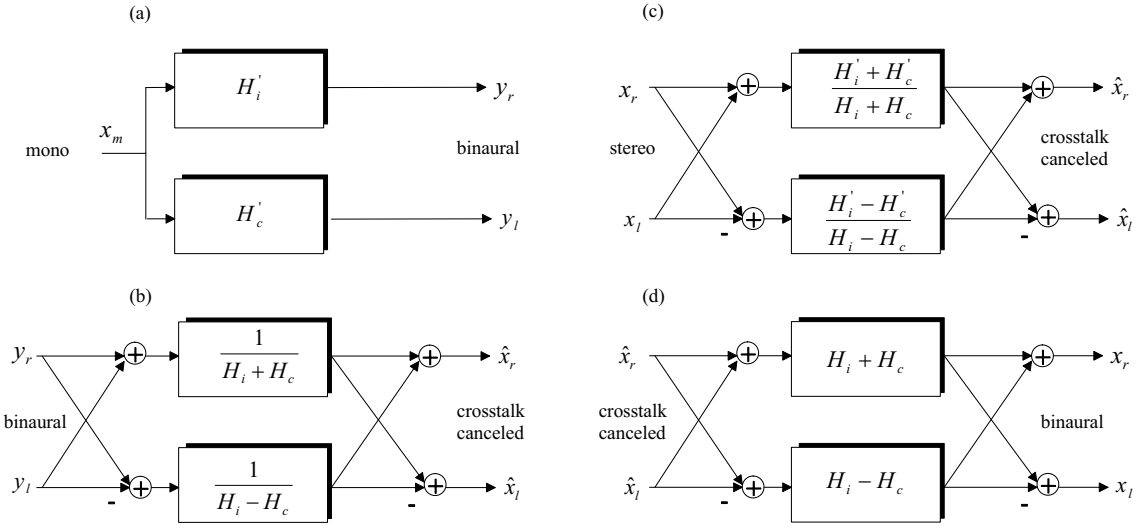


Figure 6.12: Binaural and crosstalk canceled binaural implementation and conversion structures. a) monophonic-to-binaural, b) binaural-to-crosstalk canceled, c) stereophonic-to-crosstalk canceled, and d) crosstalk canceled-to-binaural shuffler structures (Jot et al., 1995).

but also the center channel using virtual loudspeaker technology. Although the monophonic center channel may be reproduced using left and right speakers, virtual technology may be useful in stabilizing the front image both in terms of tonal color and listener position.

6.3 Stereophonic Image Widening

It is clear that HRTF-based virtual sound source processing affects the sound quality in many ways. Crosstalk canceling and virtual source synthesis are aimed at specialized applications and require a priori knowledge of the type of source signal that is fed into the system. In the case of crosstalk canceling, the input material is generally binaural, and when virtual source synthesis is performed, the inputs are multiple monophonic channels.

Methods that enhance the stereophonic image (stereo widening algorithms), on the other hand, are aimed at processing any input material, monophonic or stereophonic. An overview of artificial stereophonic image enhancement techniques is given in (Maher et al., 1996). In the following, traditional and novel methods for stereophonic image enhancement are presented.

6.3.1 Traditional Stereophonic Image Enhancement Techniques

The traditional methods for enhancement of stereo image involve processing the sum and difference signals of the inputs. It is assumed that the left and right input signals $x_l(n)$ and $x_r(n)$ consist of a monophonic portion $x_m(n)$ common to both inputs, and of signals $x_{l0}(n)$ and $x_{r0}(n)$ that are emanating from left and right channels only:

$$\begin{aligned} x_l(n) &= x_m(n) + x_{l0}(n) \\ x_r(n) &= x_m(n) + x_{r0}(n) \end{aligned} \tag{6.19}$$

The sum and difference signals can be calculated from Eq. 6.19 in the following way:

$$\begin{aligned} x_l(n) - x_r(n) &= x_{l0}(n) - x_{r0}(n) \\ x_l(n) + x_r(n) &= 2x_m(n) + x_{l0}(n) + x_{r0}(n) \end{aligned} \tag{6.20}$$

From this equation it can be seen that adding $x_l(n) - x_r(n)$ to $x_l(n)$ produces $x_m(n) + 2x_{l0}(n) - x_{r0}(n)$, which clearly boosts the left-only signal. Similarly, when $x_l(n) - x_r(n)$ is subtracted from $x_r(n)$, the right-only signals are enhanced. In this work, three algorithms have been investigated which take advantage of the theory presented above.

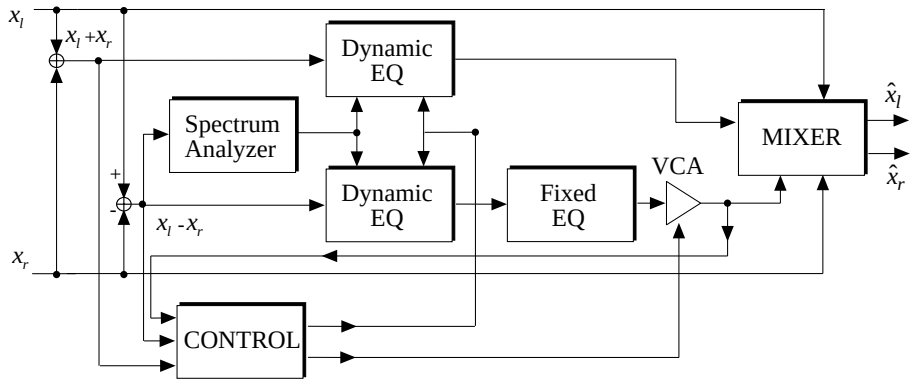


Figure 6.13: Stereo widening algorithm presented in (Klayman, 1988).

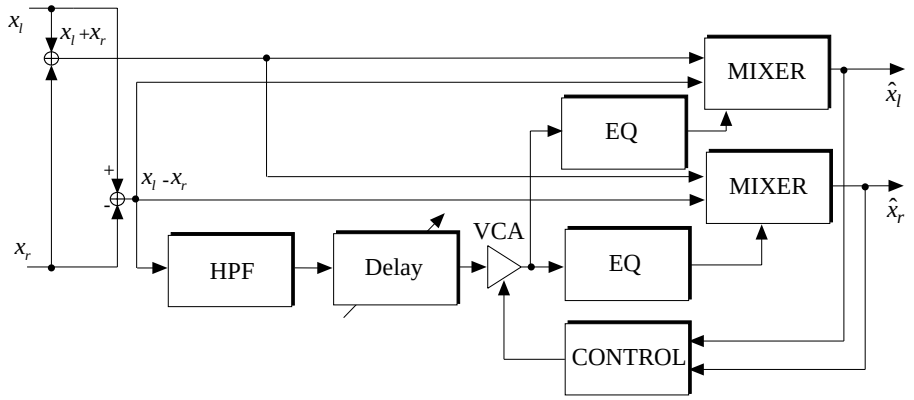


Figure 6.14: Stereo widening algorithm presented in (Desper, 1995).

In (Klayman, 1988), the sum and difference signals of the stereophonic input are processed as illustrated in Fig. 6.13. The enhanced (not binaural in a strict sense) output signals $\hat{x}_l(n)$ and $\hat{x}_r(n)$ are constructed by adding the unprocessed input signals and the processed sum and difference signals in the mixing stage. Adaptive frequency-dependent equalization of the $x_l(n)+x_r(n)$ and $x_l(n)-x_r(n)$ signals is carried out to correct for tone coloring.

Another similar algorithm is presented in the patent by (Desper, 1995). The structure is illustrated in Fig. 6.14. The main difference compared to the previous design is the fact that only the difference signal $x_l(n) - x_r(n)$ of the stereophonic input is dynamically controlled and processed, the sum signal $x_l(n) + x_r(n)$ is mixed into the enhanced output without dynamic processing.

The technique presented in (Maher et al., 1996) is also based on processing of the difference signals of the stereophonic input as in the two previous examples. This structure is shown in Fig. 6.15. In addition to previous algorithms, the use of two adaptive FIR filters $H_l(z)$ and $H_r(z)$ is suggested. An adaptive least-mean square (LMS) algorithm (Widrow and Stearns, 1985) is used to minimize

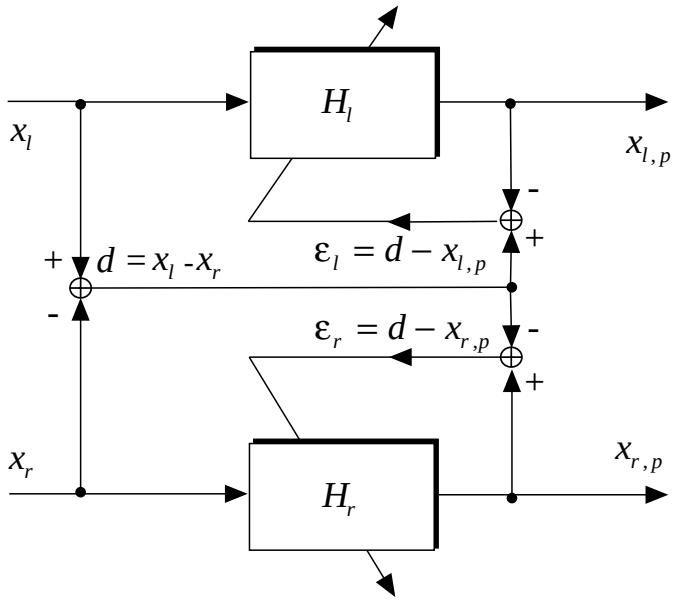


Figure 6.15: Stereo widening algorithm presented in (Maher et al., 1996).

the error terms ϵ_r and ϵ_l and between the desired output $(x_l(n) - x_r(n))$ and the processed outputs $x_{l,p}(n)$ and $x_{r,p}(n)$.

6.3.2 Stereo Widening Based on 3-D Sound Processing

When 3-D sound processing algorithms are used for stereo widening, the *tone quality* of the reproduced stereophonic sound becomes a more and more critical issue. In the literature, a system that performs stereophonic image enhancement with "placement filters" is described (Lowe et al., 1995). In this system (illustrated in Fig. 6.16) it is assumed that the inputs $x_{l0}(n)$ and $x_{r0}(n)$ have been processed so that the monophonic portion $x_m(n)$ common to both channels has been extracted. The algorithm resembles a virtual loudspeaker synthesis scheme of Fig. 6.8 with an additional delay unit for mixing the unprocessed left and right signals with the output. A novel algorithm for stereo widening based on crosstalk canceled HRTF processing was presented in (Valio and Huopaniemi, 1996). The proposed structure enhances the stereophonic image using virtual speaker processing, but retains the original tone color for monophonic sounds. This is achieved by extracting a sum signal of the inputs and applying a tone correction filter to that portion of sound while the left and right channels are separately processed as in virtual speaker reproduction. The system has been further modified and optimized for computational requirements. A schematic of the new system (modification of (Valio and Huopaniemi, 1996)) is illustrated in Fig. 6.17. In this structure the crosstalk canceling filters $H_i(z)$ and $H_c(z)$ and the virtual speaker placement filters $H_i'(z)$ and $H_c'(z)$ are calculated using methods

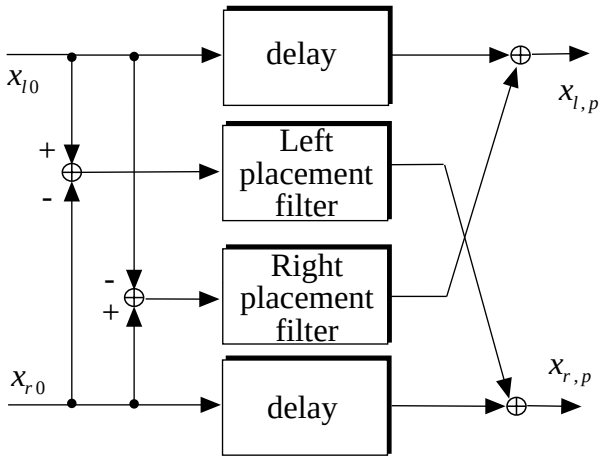


Figure 6.16: Stereo widening algorithm presented in (Lowe et al., 1995).

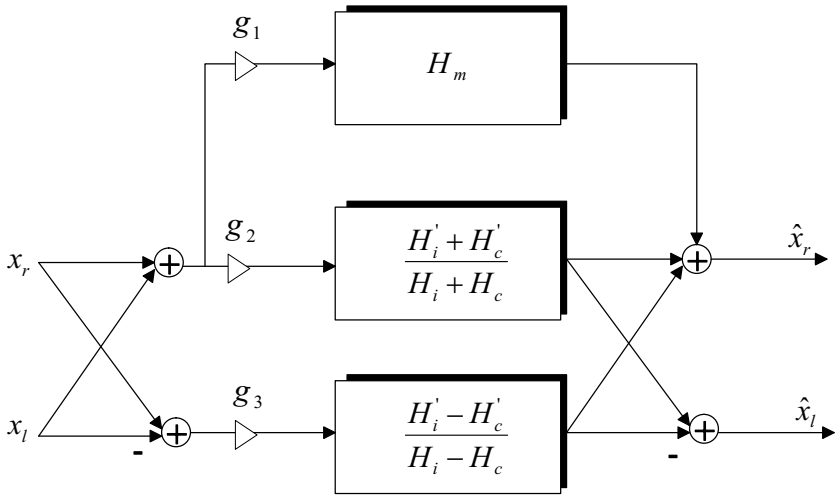


Figure 6.17: Stereo widening algorithm based on virtual loudspeaker positioning and control of the monophonic sound.

presented in Section 6.2. The filter $H_m(z)$ is used to compensate for coloration and delay caused by virtual speaker processing. By choosing an appropriate distribution for the gains g_1 and g_2 for the pseudo-monophonic sum-signal the tonal quality of the overall enhanced stereo image can be controlled.

6.4 Virtual Loudspeaker Filter Design

There are two major difficulties in the design and implementation of virtual loudspeaker or crosstalk canceling structures:

- limited listening area ("sweet spot"), where the loudspeaker binaural system

Research Group	**Design Type**	**Filter Order**	**Study**
MacCabe and Furlong, 1991	CT-canceled binaural / IIR	20	Empirical?
Cooper and Bauck, 1993	CT-canceled binaural / IIR	12	Empirical?
Kotorynski, 1995	CT-canceled binaural / FIR	64	Empirical?
Kotorynski, 1995	CT-canceled binaural / IIR	32	Empirical?
Jot *et al.*, 1995	CT-canceled binaural / IIR	12	Empirical?
Gardner, 1998	CT-canceled binaural / IIR	8	Empirical

Table 6.1: crosstalk canceled binaural HRTF filter design data from the literature.

functions well

- undesired coloration caused by VL or crosstalk filtering

The concept of joint minimum phase in crosstalk canceling and VL design (discussed in Section 6.1) helps us to understand the general problematics of crosstalk canceled binaural reproduction. The phase information in shuffler filter structures may be considered redundant, i.e., a minimum-phase component for the sum and difference signals can be reconstructed using the Hilbert transform (Oppenheim and Schafer, 1989). Although this assumption is only valid in symmetrical listening and may cause sound image degradation (Kotorynski, 1990), it will help in the generalization of VL filters to work on a larger listening area with less coloration.

The task of crosstalk canceling binaural filter design differs somewhat from binaural filter design issues that were discussed in the previous Chapter. Crosstalk canceling and the problem of limited listening area introduce complications and constraints that require advanced filter design techniques. In Table 6.1, results from studies published in the literature are listed. As can be seen, crosstalk canceled binaural filter design issues have not been widely explored. By nature, the form of crosstalk canceling filters is recursive, because the canceling process is infinite. This would suggest that recursive pole-zero models can be used for modeling crosstalk or virtual loudspeaker synthesis. The design of joint minimum-phase cross-talk canceling filters can be carried out using similar techniques as in binaural filter design. In the following, methods for crosstalk canceled binaural filter design are presented. The techniques are equally applicable to the design of both crosstalk filters, and combined crosstalk and virtual loudspeaker filters.

6.4.1 Finite Impulse-Response Methods

The windowing method gives an optimal fit in the least-squares sense to the given frequency response. In VL design, however, the problem of solving multiple transfer functions simultaneously, as is the case in the shuffler structures, arises. An optimal solution to the problem in the least-squares sense can be found by solving for all the filters in the inverse matrix formulation simultaneously so that the estimation errors are matched and optimized. This type of approach using adaptive LS optimization has been taken by (Nelson et al., 1992, 1996a,b; Kirkeby et al., 1998; Kirkeby and Nelson, 1998).

6.4.2 Infinite Impulse-Response Methods

The principle of warped filter design discussed in Chapter 5 can be extended to crosstalk canceled binaural processing. When the shuffler filters can be specified as minimum-phase systems, warped structures are an attractive solution for efficient auditory-based filter design and implementation. An outline for a *warped shuffler filter* (WSF) design method is as follows:

- Compute the shuffler transfer functions using Eq. 6.16
- Extract the joint minimum-phase part of the transfer functions
- Warp the frequency axis using Eq. 5.19
- Design the filters using a traditional IIR filter design method (e.g., Prony's method (Parks and Burrus, 1987)
- Unwarp the designed warped filter to second-order section implementation using Eqs. 5.39–5.40.

In Fig. 6.18, results for designing a minimum-phase VL shuffler filter for $10°--90°$ are presented. It can be seen that a WIIR of order 15 is capable of retaining most of the spectral features up to 15 kHz with an enhanced low frequency fit. The validity and performance of the WSF filters has been verified in a round-robin experiment of virtual home theater (VHT) systems, discussed in Section 6.6.

6.4.3 Conclusion and Discussion

Filter design for crosstalk canceled binaural systems demands special care when specifying the target response and the goal of optimization. Basically, the crosstalk canceling or virtual loudspeaker filters should satisfy two goals:

- optimization for a large listening area versus the stability of the image
- optimization for a generalized set of filters that minimize sound coloration

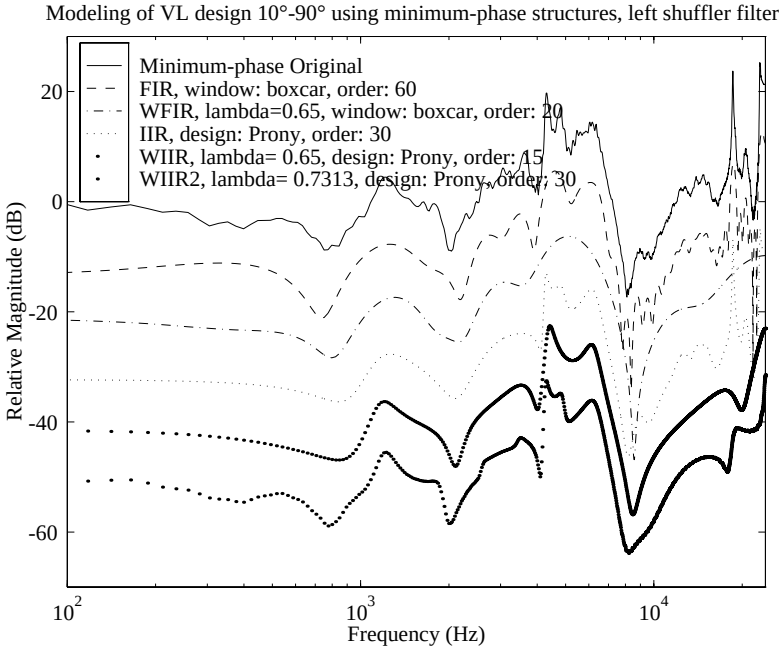


Figure 6.18: Virtual loudspeaker filter design for $10° - -90°$ using FIR, IIR, WFIR, and WIIR filters.

To meet these specifications, experiments have been carried out in listening area widening and different filter design methods. As a conclusion, according to informal listening tests, FIR filters of order 40-50 and WIIR filters of order 15-20 based on the minimum-phase shuffler structure proposed by Cooper and Bauck (1989) would suffice in realization of crosstalk canceling and virtual loudspeaker implementation.

6.5 System Analysis

6.5.1 Widening the Listening Area

In this section, techniques for listening area widening and the control of tone coloration are discussed. The "sweet spot" for listening remains a severe problem in loudspeaker binaural synthesis of audio signals. There are, however, several approaches to study and partly solve this problem. Intuitively, two-channel stereophony is very limited in giving possibility to listening area widening due to the precedence effect (Yost and Gourevitch, 1987). A related topic is the stability and coloration of the image, i.e. the virtual sound image quality deterioration in an asymmetrical listening position.

Head-Tracked Systems

One of the most prominent techniques for expanding the sweet spot is to use tracking of the listener position (Gardner, 1998a; Kyriakakis and Holman, 1998). This partly overcomes the problems of robust crosstalk canceling due to the fact that the canceling filters can be steered according to listener movements. Methods based on magnetic tracking (Gardner, 1998a) or visual tracking (Kyriakakis and Holman, 1998) of the listener's position have been proposed. This technique is suitable for one listener only, which is also true for most loudspeaker based binaural reproduction systems.

Stereo Dipole Technique

A principle called "stereo dipole" has been recently introduced that uses two closely spaced loudspeakers for virtual source synthesis in loudspeaker listening (Watanabe et al., 1996; Takeuchi et al., 1997; Kirkeby et al., 1998; Kirkeby and Nelson, 1998). Based on theoretical formulations, the conclusions were that bringing the loudspeakers closer to each other enhances and stabilizes the virtual speaker image, and widens the listening area. This has also been verified by robustness analysis of crosstalk canceling for different loudspeaker placings (Ward and Elko, 1998, 1999). One of the inherent problems of the stereo dipole technique, however, is the computational complexity that results from the fast deconvolution and frequency-dependent regularization (Kirkeby and Nelson, 1998) algorithms. This is due to the goal of optimal inversion of the two-by-two system of two-loudspeaker listening, not taking into account the joint minimum-phase properties of the system. Warped FIR structures have been proposed for the stereo dipole technique (Kirkeby et al., 1999) to enhance computational efficiency, resulting in a two-by-two matrix of WFIR filters each containing 32 coefficients. This presents a clear reduction in computation, but is still outside of the scope of efficient real-time systems (such as the WSF method proposed by the author in Section 6.4.2).

Presenting Binaural Material over Multiple Loudspeakers

In a method proposed by (Bauck and Cooper, 1992, 1996), the center image can be stabilized by adding a third loudspeaker, a center channel. This addition also introduces natural widening of the listening area by adding a second "sweet spot" to the listening area. An exact solution can be found, which uses four filters in realization, but the other listener gets a reversed image. According to (Bauck and Cooper, 1996), a nonreversed image can also be computed, but the result is not exact due to an overdetermined solution.

It is also possible to use pairwise binaural presentation with crosstalk cancellation in multiple speaker systems. For example, using two frontal and two rear speakers with pairwise crosstalk cancellation it is possible to create convincing

front-back discrimination and a more stable 360° panning of virtual sources. Furthermore, using intermediate formats such as Ambisonics (Malham and Myatt, 1995) for binaural rendering[2] it is possible to apply spatial transcoding of signals for different reproduction formats, and to reduce the cost of binaural synthesis of multiple sources.

6.5.2 Analysis of Crosstalk Canceled Binaural Designs

In this experiment, different methods for crosstalk canceled binaural filter design were compared. The subjective and objective quality of two crosstalk canceling methods based on computational models was compared to HRTF-based VL and crosstalk designs. The computational models were based on Rayleigh's diffraction formula, and the approximation proposed by Gardner (Eqs. 6.10 and 6.11). The HRTF database consisted of measurements made on 33 human subjects (Riederer, 1998b). The design goal was to create virtual loudspeakers at $\pm 90^o$ azimuth when the physical speaker placement was $\pm 10^o$.

As discussed earlier, in the symmetrical case the crosstalk canceling shuffler structures are approximately of joint minimum phase, that is, the phase response of $H_i(z) + H_c(z)$ and $H_i(z) - H_c(z)$ can be calculated using minimum-phase reconstruction. Thus the transfer functions can be derived using magnitude only approximation without phase constraints. In Fig. 6.19, crosstalk canceling transfer functions for 33 subject HRTFs are presented. A mean of the magnitude values was also calculated, which is shown as the thick line in Fig. 6.19. Although it is clear that some of the idiosyncratic information (above 4-5 kHz) present in HRTFs is smoothed when the mean value is taken on a frequency-point basis (see, e.g., (Møller et al., 1995) for discussion), in this context the use of a generalized magnitude response is motivated. It is further possible to smooth the mean responses by, e.g., 1/3 octave or auditory smoothing in order to remove such magnitude fluctuations that are not audible according to psychoacoustic theory. In theory, also the VL positioning filters functions $H_i'(z)$ and $H_c'(z)$ should be of nearly joint minimum phase, so the same technique could be applied to yield minimum-phase realizations of VL shuffler filters.

In Fig. 6.20, the results for VL design based on four techniques are illustrated. The empirically based VL structures used data measured from one test person. The solid line of Fig. 6.20 represents the VL design for the test person's HRTFs. The dashed line in the figure represents a VL design where computed magnitude means from the population of HRTFs have been used both for virtual loudspeaker placement filtering (the transfer functions $H_i'(z)$ and $H_c'(z)$ as depicted in Figs. 5.6 and 5.8), and for crosstalk canceling. This results in a minimum-phase system. The dash-dotted and dotted lines represent the Cooper and Gardner VL approximations, respectively. It can be clearly seen that the computational mod-

[2] This approach is called the "binaural B format" (Jot et al., 1998b).

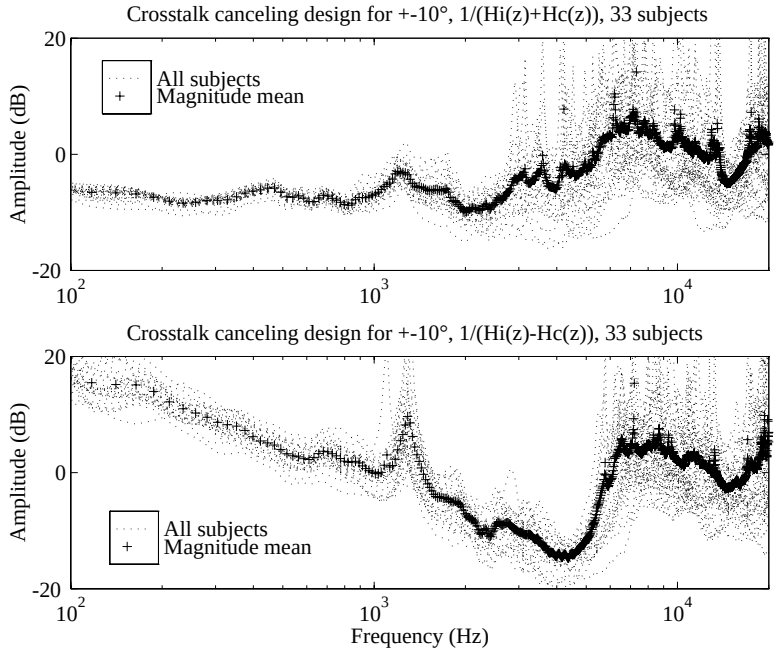


Figure 6.19: Crosstalk canceling magnitude responses for loudspeakers placed at $\pm 10^o$ azimuth based on 33 human subject HRTFs (shown as dotted lines). The line marked with (+) represents a calculated mean of the magnitude responses.

els resemble the empirical results only up to approximately 2 kHz. In informal listening tests, the VL structure based on a mean minimum-phase approximation performed in a satisfactory manner when compared to individualized VL design. Coloration caused by sharp peaks in the individualized VL design was reduced (seen in Fig. 6.20), but the listening area remained stable. A minimum-phase realization of the individualized VL filters was also tested, and no audible difference to the non-minimum-phase version was found. The performance of VL and crosstalk designs based on computational HRTFs, however, was found unsatisfactory. The virtual loudspeaker image was unstable and often the test signal was perceived as emanating from one of the speakers.

These results suggest that generalization and smoothing of symmetrical crosstalk canceling structures could reduce tone coloration and expand the listening area. More thorough listening tests should be carried out, however, to fully support these conclusions.

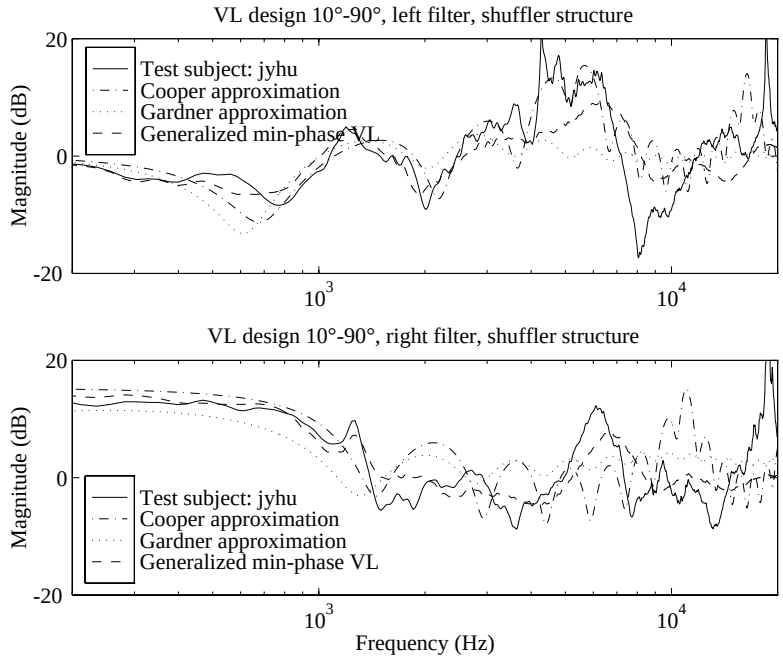


Figure 6.20: VL design magnitude responses for loudspeakers placed at $\pm 10^o$ azimuth and virtual speaker placement of $\pm 90^o$ azimuth. The shuffler structure was used.

6.6 Subjective Evaluation of Virtual Surround Systems

In this section, overview and results of a round robin subjective evaluation experiment on virtual surround systems are provided (Zacharov et al., 1999; Zacharov and Huopaniemi, 1999). The goal of the experiment was to compare discrete 5 channel reproduction to reproduction of 5 channel material over two loudspeakers placed at $\pm 30°$ or at$\pm 5°$. Furthermore, the aim was to compare the quality and performance of different virtual surround algorithms provided by industry and academia (Zacharov et al., 1999).

6.6.1 Background

As the development of high quality multichannel audio encoding, storage media and broadcast techniques evolves, the availability of multichannel audio has become widespread. While matrixed multichannel methods (e.g. Dolby Surround) have been widely available for some time, the benefits of discrete multichannel reproduction are finding increasing favor in professional and consumer markets, for example, with new audiovisual storage formats such as DVD.

The so-called *virtual surround* or *virtual home theater (VHT)* (Olive, 1998)

aim to faithfully reproduce the spatial sound qualities of 5.1 channel systems using only two loudspeakers. Virtual loudspeaker technology enables the virtualization of surround left and right channel information. Thus the surround channels appear to emanate from virtual sources placed outside the physical stereo setup of the two loudspeakers. Virtualization can also be created for the left and right signals in order to, for example, expand the stereo base. Furthermore, virtualization may be applied for the center channel, or it can be left unprocessed (the traditional phantom image)[3].

6.6.2 Experimental Procedure and Setup

This section summarizes the experimental procedure and setup for the virtual surround experiment performed in the two sites. At each site a different listening room and listening panel were employed as described below. In all other respects the experiments performed at each site were identical. For a full description of the design, the interested reader is referred to (Zacharov et al., 1999).

The first of the two experiments was performed in the new Nokia Research Center (NRC), Speech and Audio Systems Laboratory listening room, which is fully conformant with ITU-R BS.1116-1 (ITU-R, 1994a) as illustrated in Fig. 6.21. The second experiment was performed at the AES 16th International Conference in a smaller meeting room. For reproduction, Genelec 1030A speakers were employed for all of the experiments consisting of a two-way speaker with a frequency response of 52 Hz – 20 kHz (±3 dB).

The setup for the 5-channel reproduction was in accordance with ITU-R BS.775-1 (ITU-R, 1994b) as far as possible and thus speakers were placed at the normal 0°, ±30°, ±110° angles, with the speaker's axis at average ear height. In addition, a pair of speakers were set up at an angle of ±5° to enable virtual surround reproduction using closely spaced loudspeakers. Loudspeakers were placed behind an acoustically transparent screen to limit any visual bias effects. As only discrete 5-channel material was employed for this experiment, no low frequency energy (LFE) channel or loudspeaker was available.

For the experiment, a rank order procedure was chosen due to its efficiency and the simplicity of the rank ordering task when employing untrained listeners (Zacharov et al., 1999). Furthermore, as no fixed reference was employed, the rank order procedure enabled the comparison of the perceived quality difference between a discrete 5-channel system and *virtual home theatre* systems, without the assumption that the 5-channel is either inferior or superior in performance.

[3] It should be noted that all virtual surround or virtual home theater concepts discussed in this experiment refer to 2-loudspeaker reproduction of 5.1 audio material in such a way that at least the surround (and possibly also the center, left and right) audio channels are processed virtually.

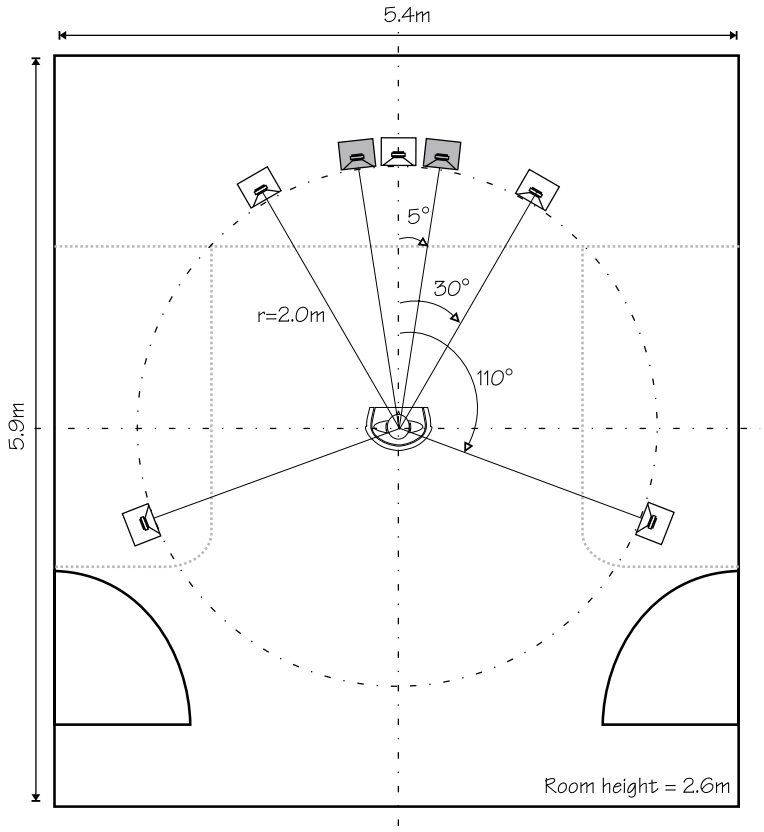


Figure 6.21: Layout of the new NRC ITU-R BS.1116-1 (ITU-R, 1997) compliant listening room and loudspeaker setup (Zacharov et al., 1999).

Tested Systems

A total of six VHT systems were compared to a reference discrete 5 channel reproduction. A complete list of the proponents and the details of the systems can be found in Table 6.2. For the VHT experiment, a set of warped shuffler filters (WSF), described in detail in Section 6.4.2 were designed and employed using the following specifications. This set of filters designed by the author was called PBVS (Polar Bear Virtual Surround) in the VHT experiment (Zacharov et al., 1999; Zacharov and Huopaniemi, 1999).

- Shuffler filter structure was used for $\pm 110°$ virtual surround channel generation.
- Two minimum-phase warped filters of order 20 were designed[4]
- The filters were implemented by unwarping the responses to second-order sections.

[4] Note that there is an error in (Zacharov et al., 1999). The order of the WIIR filters used in the experiment was 20.

Proponent	System	Reproduction angle	Center channel	MIPS @ 48 kHz	System description
Harman Multimedia	VMAx 3D Virtual Theatre	5°	Virtual	6	85 filter coefficients, memory 400×(word size)
Aureal Semiconductor	A3DS	30°	Phantom	20	Virtual front channels and decorrelation processing, coefficient memory: 150 words, delay memory: 1k words (inc. decorrelator)
SRS Labs, Inc.	TruSurround	30°	Phantom	4.7	IIR filters with 18 filter coefficients / Total ROM code + filter coefficients = 306, RAM=36
Helsinki University of Technology	PBVS	30°	Phantom	6.4	Warped IIR filters of order 20
Sensaura Ltd	Sensaura Virtual Surround	30°	Phantom	5.2	25 coefficients per filter, no additional memory required
Dolby Laboratories	Dolby Virtual Surround	5°	Phantom	~ 5	10 filters and 10 delays

Table 6.2: Summary of proponent systems, based upon the information provided by manufacturers (Zacharov et al., 1999).

6.6.3 Test Items and Grading

Four program items were selected for the experiment. The items were selected to provide a range of spatial sound cues and different types of programme.

- Blue bell railway, BBC. Scene consisting of steam training pulling away from station and approaching bridge. Contains country atmosphere with directional cues, and panning effects.
- Tipsy Gypsy, BBC. A concert at the Albert hall, consisting of audience cheering, applause and the conductor talking.
- Rain storm. Sample consisting of a thunder roll followed by the sound of hard rain.
- Felix Mendelssohn-Bartoldy, Symphony No. 4. Live recording at the Neues Gewandhaus, Leipzig, Deutsche Telekom, 1993.

Two grading scales were employed to evaluate the perceived spatial and timbral quality of reproduction of the seven systems under test. The following questions were posed to each listener.

- Rank order the samples by spatial sound quality (1 = lowest rank, 7 = highest rank)

When evaluating the spatial sound quality, please consider all aspects of spatial sound reproduction. This might include the locatedness or localisation of the sound, how enveloping it is, it's naturalness and depth.

- Rank order the samples by timbral quality (1 = lowest rank, 7 = highest rank)

  When considering aspects of the timbre quality, please consider timbre as a measure of the tone colour. Timbre can be considered as the sensory attribute which allows two similar sounds, of the same pitch and loudness, to be different, for example a clarinet and a cello. Any audible distortions can also be considered as an aspect of the timbral quality.

6.6.4 Results and Discussion

The results of the VHT experiment are shown in Figs. 6.22–6.23 (Zacharov and Huopaniemi, 1999). The PBVS algorithm using WSF filters is marked "5" in the NRC test and "25" in the AES16 test. From the results it can be seen that the performance of the warped shuffler filters is very good, above average when compared to other tested methods (resulting second and third spatially, and first and second timbrally of the VHT algorithms). As a summary, the following conclusions from the test can be made:

- The WSF structure performed very well in the test compared to commercial algorithms and implementations. The WSF algorithm is computationally efficient and provides good virtual source imaging with little coloration.
- Results are statistically significant and considered meaningful.
- Results are similar between both sites.
- Discrete 5 channel is significantly superior.
- There are strong correlations between the ranking for both spatial and timbre quality for all programme items.

6.7 Discussion and Conclusions

In this chapter, concepts and methods for crosstalk canceled binaural reproduction were discussed. Theory and formulas for virtual source synthesis in symmetrical and asymmetrical listening setups were reviewed. The validity of joint minimum-phase reconstruction was verified. A new virtual loudspeaker filter design method based on the shuffler structure using warped filters was introduced. The performance of the warped shuffler filter was verified in subjective experiments.

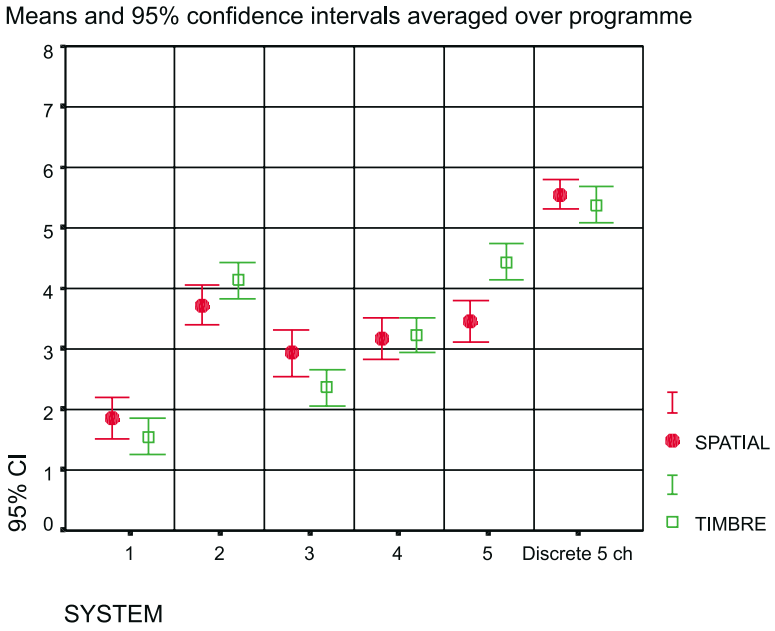


Figure 6.22: Subjective experiment result from the NRC experiment (Zacharov and Huopaniemi, 1999).

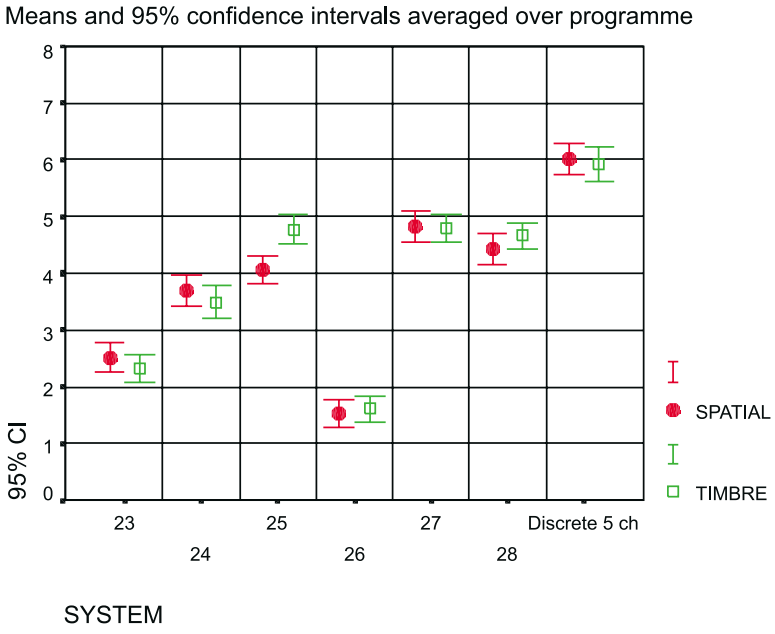


Figure 6.23: Subjective experiment result from the AES16 experiment (Zacharov and Huopaniemi, 1999).

Chapter 7

Conclusions

In this work, issues in 3-D sound and virtual acoustics in the field of multimedia were treated. New concepts, methods, algorithms and experimental results were presented, and their importance to virtual acoustics modeling was discussed. The topic is both up-to-date and innovative from the research point of view, because the integration of different media such as telecommunications, multimedia and virtual reality has placed new requirements and standards for audio and audio signal processing. The concept of virtual acoustics was treated, which consists of source, environment and receiver modeling. In this work, new methods were introduced for parametrization, filter design and real-time rendering of virtual acoustics, allowing for more realistic 3-D audio presentation with efficient computation.

In Chapter 2, an overview of virtual acoustic technology was given. The author's contribution to the DIVA system and to MPEG-4 standardization were presented, and overall design and implementation methodology for virtual acoustics was discussed.

In Chapter 3, topics in sound source parametrization and modeling were studied. The author has contributed to source directivity modeling by introducing two novel techniques: a) Directional filtering, and b) Set of elementary sources. Furthermore, the principle of reciprocity was applied to modeling the sound radiation from the human mouth. This novel technique enables fast and reliable approximation of the directivity characteristics of human speech for virtual reality modeling purposes.

Chapter 4 dealt with real-time room acoustical issues, concentrating on filter design aspects for the phenomena of absorption due to air and material reflections. The author proposed new methods for efficient low-order filter design for air absorption and acoustical material properties.

Chapter 5 concentrated on HRTF approximation and filter design. New design methods for binaural filters based on balanced model truncation and frequency warping were proposed by the author. Methods for subjective and objective evaluation of binaural filter design were introduced. The aim in applying a binaural

auditory model for objective evaluation of binaural filter design quality is to be able to predict the subjective preferences, i.e., what is relevant in binaural processing for perception. The results provide guidelines and recommended practice for binaural filter design and are useful for both research and applications of binaural technology.

In Chapter 6, binaural techniques were extended for loudspeaker reproduction by crosstalk canceling mechanisms and virtual loudspeaker technology. The author introduced a warped shuffler filter (WSF) design method, which enables low-order designs of minimum-phase crosstalk cancelers and virtual loudspeakers. The performance of the WSF method was verified in a round-robin subjective listening experiment.

The main results of this thesis can be summarized as follows:

- Sound source directivity can be incorporated in virtual reality systems using two methods: a) Directional filtering, b) Set of elementary sources.
- A reciprocity method can successfully be applied to estimation of directional sound radiation from the human mouth.
- Air absorption and boundary material reflection effects can be efficiently modeled using low-order digital filters
- The high-frequency spectral content present in HRTFs can be smoothed using auditory criteria without degrading localization performance.
- A binaural auditory model can be used to give a quantitative prediction of perceptual HRTF filter design performance.
- Auditory filter design principles using, for example, frequency warping or balanced model truncation result in more efficient and more perceptually relevant binaural processing.
- Virtual loudspeaker filters designed using a joint minimum-phase shuffler structure and auditory design principles provide high-quality virtual source imaging for virtual surround applications.

The author has contributed to source modeling, environment modeling, and listener modeling in virtual acoustics. The results presented in this thesis may directly be applied to digital audio signal processing in general, and more specifically the design and implementation of virtual acoustic environments. The theory and applications of 3-D sound and virtual acoustics is still a rapidly evolving area, and both commercial and scientific interest in binaural signal processing has increased. Interesting topics such as binaural models of hearing, auralization based on auditory optimization, and spatial transcoding of audio formats remain as future challenges for the author.

Bibliography

Abel, J. 1997. Three-dimensional virtual audio display employing reduced complexity imaging filters, U.S. patent no. 5,659,619.

Abel, J. and Foster, S. 1997. Method and apparatus for efficient presentation of high-quality three-dimensional audio, U.S. patent no. 5,596,644.

Abel, J. and Foster, S. 1998. Method and apparatus for efficient presentation of high-quality three-dimensional audio including ambient effects, U.S. patent no. 5,802,180.

Allen, J. B. and Berkley, D. A. 1979. Image method for efficiently simulating small-room acoustics, *Journal of the Acoustical Society of America* **65**(4): 943–950.

Asano, F., Suzuki, Y. and Sone, T. 1990. Role of spectral cues in median plane localization, *Journal of the Acoustical Society of America* **88**(1): 159–168.

Atal, B. S. and Schroeder, M. R. 1966. Apparent sound source translator, U.S. patent no. 3,236,949.

Barron, M. 1993. *Auditorium Acoustics and Architectural Design*, E & FN Spon.

Bass, H. and Bauer, H.-J. 1972. Atmospheric absorption of sound: Analytical expressions, *Journal of the Acoustical Society of America* **52**(3): 821–825.

Bauck, J. L. and Cooper, D. H. 1992. Generalized transaural stereo, *Presented at the 93*rd *Convention of the Audio Engineering Society*, preprint 3401, San Francisco, CA, USA.

Bauck, J. L. and Cooper, D. H. 1996. Generalized transaural stereo and applications, *Journal of the Audio Engineering Society* **44**(9): 683–705.

Bauer, B. B. 1961. Stereophonic earphones and binaural loudspeakers, *Journal of the Audio Engineering Society* **9**: 148–151.

Bech, S. 1993. Training of subjects for auditory experiments, *Acta Acustica* **1**: 89–99.

Begault, D. R. 1994. *3-D Sound for Virtual Reality and Multimedia*, Academic Press, Cambridge, MA, USA, p. 293.

Beliczynski, B., Kale, I. and Cain, G. D. 1992. Approximation of FIR by IIR digital filters: An algorithm based on balanced model reduction, *IEEE Transactions on Signal Processing* **40**(31): 532–542.

Blauert, J. 1997. *Spatial Hearing. The Psychophysics of Human Sound Localization*, MIT Press, Cambridge, MA, USA, p. 494.

Blommer, M. A. 1996. *Pole-zero Modeling and Principal Component Analysis of Head-Related Transfer Functions*, PhD thesis, University of Michigan, p. 165.

Blommer, M. A. and Wakefield, G. H. 1994. On the design of pole-zero approximations using a logarithmic error measure, *IEEE Transactions on Signal Processing* **42**(11): 3245–3248.

Blommer, M. A. and Wakefield, G. H. 1997. Pole-zero approximations for head-related transfer functions using a logarithmic error criterion, *IEEE Transactions on Speech and Audio Processing* **5**(3): 278–287.

Borish, J. 1984. Extension of the image model to arbitrary polyhedra, *Journal of the Acoustical Society of America* **75**(6): 1827–1836.

Botteldooren, D. 1995. Finite-difference time-domain simulation of low-frequency room acoustic problems, *Journal of the Acoustical Society of America* **98**(6): 3302–3308.

Brandenburg, K. and Bosi, M. 1997. Overview of MPEG audio: Current and future standards for low-bit-rate audio coding, *Journal of the Audio Engineering Society* **45**(1/2): 4–20.

Brown, C. P. and Duda, R. O. 1998. A structural model for binaural sound synthesis, *IEEE Transactions on Speech and Audio Processing* **6**: 476–488.

Brungart, D. 1998. *Near-field Auditory Localization*, PhD thesis, Massachusetts Institute of Technology, p. 253.

Brungart, D. and Rabinowitz, W. 1996. Auditory localization in the near-field, *Proceedings of the International Conference on Auditory Display*, Palo Alto, California, USA.

Brungart, D. S., Durlach, N. I. and Rabinowitz, W. R. 1997. Three-dimensional auditory localization of nearby sources, *Journal of the Acoustical Society of America* **102**(5): 3140.

Calamia, P. and Hixson, E. 1997. Measurement of the head-related transfer function at close range, *Journal of the Acoustical Society of America* **102**(5): 3117.

Chen, B.-S., Peng, S.-C. and Chiou, B.-W. 1992. IIR filter design via optimal Hankel norm approximation, *IEE Proc. G: Circuits, Devices Syst.* **139**(5): 586–590.

Chen, J., Veen, B. V. and Hecox, K. E. 1995. A spatial feature extraction and regularization model for the head-related transfer function, *Journal of the Acoustical Society of America* **97**(1): 439–452.

Chung, J. Y. and Blaser, D. A. 1980. Transfer function method of measuring the in-duct acoustic properties. Part I: Theory, and Part II: Experiment, *Journal of the Acoustical Society of America*.

Cook, P. R. and Trueman, D. 1998. A database of measured musical instrument body radiation impulse responses, and computer applicatoins for exploring and utilizing the measured filter functions, *Proceeding of the International Symposium on Musical Acoustics*, Leavenworth, WA, USA.

Cooper, D. H. 1982. Calculator program for head-related transfer function, *Journal of the Audio Engineering Society* **30**(1/2): 34–38.

Cooper, D. H. and Bauck, J. L. 1989. Prospects for transaural recording, *Journal of the Audio Engineering Society* **37**(1/2): 3–19.

Cooper, D. H. and Bauck, J. L. 1992. Head diffraction compensated stereo system, U.S. patent no. 5,136,651.

Dalenbäck, B.-I. 1995. *A New Model for Room Acoustic Prediction and Auralization*, PhD thesis, Chalmers University of Technology, Gothenburg, Sweden.

Damaske, P. 1971. Head-related two-channel stereophony with loudspeaker reproduction, *Journal of the Acoustical Society of America* **50**(4, pt. 2): 1109–1115.

Desper, S. 1995. Automatic stereophonic manipulation system and apparatus for image enhancement, U.S. patent no. 5,412,731.

Duda, R. and Martens, W. 1997. Range-dependence of the HRTF for a spherical head, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York.

Duda, R. and Martens, W. 1998. Range-dependence of the HRTF of a spherical head, *Journal of the Acoustical Society of America* **104**(5): 3048–3058.

Duda, R. O., Avendano, C. and Algazi, V. R. 1999. An adaptable ellipsoidal head model for the interaural time difference, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, Phoenix, AZ, USA, pp. 965–968.

Evans, M. J., Angus, J. A. S. and Tew, A. I. 1998. Analyzing head-related transfer functions using surface spherical harmonics, *Journal of the Acoustical Society of America* **104**(4): 2400–2411.

Fahy, F. J. 1995. *Sound Intensity*, 2nd edn, E & FN SPON.

Flanagan, J. L. 1960. Analog measurements of sound radiation from the mouth, *Journal of the Acoustical Society of America* **32**(12): 1613–1620.

Flanagan, J. L. 1972. *Speech Analysis, Synthesis and Perception*, 2nd edn, Springer-Verlag, Berlin. pp. 38-41.

Fletcher, N. H. and Rossing, T. D. 1991. *The Physics of Musical Instruments*, Springer-Verlag, New York, USA.

Foster, S., Wenzel, E. and Taylor, R. 1991. Real-time synthesis of complex acoustic environments, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, New York.

Friedlander, B. and Porat, B. 1984. The modified Yule-Walker method of ARMA spectral estimation, *IEEE Trans. Aerospace Electronic Syst.* **AES-20**(2): 158–173.

Gaik, W. 1993. Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling, *Journal of the Acoustical Society of America* **94**(1): 98–110.

Gardner, W. G. 1995. *Transaural 3-D audio*, MIT Media Lab Perceptual Computing, technical report 342.

Gardner, W. G. 1997. *3-D Audio Using Loudspeakers*, PhD thesis, MIT Media Lab, p. 153.

Gardner, W. G. 1998a. *3-D Audio Using Loudspeakers*, Kluwer Academic Publishers, Boston.

Gardner, W. G. 1998b. Reverberation algorithms, *in:* M. Kahrs and K. Brandenburg (eds), *Applications of Digital Signal Processing to Audio and Acoustics*, Kluwer Academic Publishers, Norwell, MA, USA, chapter 3, pp. 85–131.

Gardner, W. G. and Martin, K. D. 1994. *HRTF measurements of a KEMAR dummy-head microphone*, MIT Media Lab Perceptual Computing, technical report 280.

Gardner, W. G. and Martin, K. D. 1995. HRTF measurements of a KEMAR, *Journal of the Acoustical Society of America* **97**(6): 3907–3908.

Gerzon, M. 1973. Periphony: with-height sound reproduction, *Journal of the Audio Engineering Society* **21**(1/2): 2–10.

Gerzon, M. 1992. Panpot laws for multispeaker stereo, *the 92nd Convention of the Audio Engineering Society*, preprint 3309, Vienna, Austria.

Gilkey, R. and Anderson, T. 1997. *Binaural and Spatial Hearing in Real and Virtual Environments*, Lawrence Erlbaum Associates, Mahwah, New Jersey, p. 795.

Golub, G. H. and Loan, C. F. V. 1996. *Matrix Computations*, 3 edn, Princeton University Press, The Johns Hopkins University Press, p. 694.

Gutknecht, M., Smith, J. O. and Trefethen, L. N. 1983. The Caratheodory-Fejer (CF) method for recursive digital filter design, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **31**(6): 1417–1426.

Hänninen, R. 1999. *LibR – An Object-Oriented Software Architecture for Realtime Sound and Kinematics*, Helsinki University of Technology. Lic.Tech. Thesis.

Hartung, K. and Raab, A. 1996. Efficient modeling of head-related transfer functions, *Acta Acustica* **82**(suppl. 1): S88.

Hiipakka, D. G. J., Hänninen, R., Ilmonen, T., Napari, H., Lokki, T., Savioja, L., Huopaniemi, J., Karjalainen, M., Tolonen, T., Välimäki, V., Välimäki, S. and Takala, T. 1997. Virtual orchestra performance, *Visual Proceedings of SIGGRAPH'97*, Los Angeles, p. 81.

Huopaniemi, J. 1997. *Modeling of Human Spatial Hearing in the Context of Digital Audio and Virtual Environments*, Helsinki University of Technology, Department of Electrical and Communications Engineering, Laboratory of Acoustics an d Audio Signal Processing, p. 101. Lic.Tech. Thesis.

Huopaniemi, J. and Karjalainen, M. 1996a. Comparison of digital filter design methods for 3-D sound, *Proc. IEEE Nordic Signal Processing Symposium*, Espoo, Finland, pp. 131–134.

Huopaniemi, J. and Karjalainen, M. 1996b. HRTF filter design based on auditory criteria, *Proc. Nordic Acoustical Meeting (NAM'96)*, Helsinki, Finland, pp. 323–330.

Huopaniemi, J. and Karjalainen, M. 1997. Review of digital filter design and implementation methods for 3-D sound, *Presented at the 102nd Convention of the Audio Engineering Society*, preprint 4461, Munich, Germany.

Huopaniemi, J. and Riederer, K. 1998. Measuring and modeling the effect of distance in head-related transfer functions, *Proceedings of the joint meeting of the Acoustical Society of America and the International Congress on Acoustics*, Seattle, WA, USA, pp. 2083–2084.

Huopaniemi, J. and Smith, J. O. 1999. Spectral and time-domain preprocessing and the choice of modeling error criteria for binaural digital filters, *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*, Rovaniemi, Finland, pp. 301–312.

Huopaniemi, J., Karjalainen, M. and Välimäki, V. 1995. Physical models of musical instruments in real-time binaural room simulation, *Proceedings of the International Congress on Acoustics*, Trondheim, Norway, pp. 447–450.

Huopaniemi, J., Karjalainen, M., Välimäki, V. and Huotilainen, T. 1994. Virtual instruments in virtual rooms – a real-time binaural room simulation environment for physical modeling of musical instruments, *Proceedings of the International Computer Music Conference*, Aarhus, Denmark, pp. 455–462.

Huopaniemi, J., Kettunen, K. and Rahkonen, J. 1999a. Measurement and modeling techniques for directional sound radiation from the mouth, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, New York.

Huopaniemi, J., Savioja, L. and Karjalainen, M. 1997. Modeling of reflections and air absorption in acoustical spaces: a digital filter design approach, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, New York.

Huopaniemi, J., Savioja, L. and Takala, T. 1996. DIVA virtual audio reality system, *Proceedings of the International Conference on Auditory Display*, Palo Alto, California, USA.

Huopaniemi, J., Zacharov, N. and Karjalainen, M. 1998. Objective and subjective evaluation of head-related transfer function filter design, *Presented at the 105th Audio Engineering Society Convention*, preprint 4805 (invited paper), San Francisco, CA, USA.

Huopaniemi, J., Zacharov, N. and Karjalainen, M. 1999b. Objective and subjective evaluation of head-related transfer function filter design, *Journal of the Audio Engineering Society* **47**(4): 218–239.

Imai, S. 1983. Cepstral analysis synthesis on the Mel frequency scale, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Boston, MA, USA, pp. 93–96.

ISO 1985. ISO 354. Acoustics — Measurement of sound absorption in a reverberation room.

ISO 1989. ISO 8253-1. Acoustics — Audiometric test methods — Part 1: Basic pure tone air and bone conduction threshold audiometry, pp. 587–599.

ISO 1993. ISO 9613-1. Acoustics — Attenuation of sound during propagation outdoors — Part 1: Calculation of the absorption of sound by the atmosphere, pp. 377–385.

ISO 1996. ISO 10534-1. Acoustics — Determination of sound absorption coefficient and impedance in impedance tubes — Part 1: Method using standing wave ratio.

ISO/IEC 1997. ISO/IEC JTC/SC24 IS 14772-1. Information technology – Computer graphics and image processing - The Virtual Reality Modeling Language (VRML97).

ISO/IEC 1999. ISO/IEC JTC1/SC29/WG11 IS 14496 (MPEG-4). Information technology – Coding of audiovisual objects.

ITU-R 1994a. Recommendation BS.1116. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. Geneva.

ITU-R 1994b. Recommendation BS.775-1. Multichannel stereophonic sound system with and without accompanying picture. International Telecommunications Union Radiocommunication Assembly, Geneva.

ITU-R 1997. Recommendation BS.1116-1. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. International Telecommunications Union Radiocommunication Assembly, Geneva.

Jeffress, L. A. 1948. A place theory of sound localization, *J. Comp. Physiol. Psych.* **61**: 468–486.

Jenison, R. L. 1995. A spherical basis function neural network for pole-zero modeling of head-related transfer functions, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York.

Jot, J.-M. 1999. Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces, *Multimedia Systems* **7**: 55–69.

Jot, J.-M. and Chaigne, A. 1996. Method and system for artificial spatialisation of digital audio signals, U.S. patent no. 5,491,754.

Jot, J.-M., Jullien, J.-P. and Warusfel, O. 1998a. Method to simulate the acoustical quality of a room and associated audio-digital processor, U.S. patent no. 5,812,674.

Jot, J.-M., Larcher, V. and Warusfel, O. 1995. Digital signal processing issues in the context of binaural and transaural stereophony, *Presented at the 98th Convention of the Audio Engineering Society*, preprint 3980, Paris, France.

Jot, J.-M., Wardle, S. and Larcher, V. 1998b. Approaches to binaural synthesis, *Presented at the 105th Audio Engineering Society Convention*, preprint 4861 (invited paper), San Francisco, CA, USA.

Juhl, P. M. 1993. *The Boundary Element Method for Sound Field Calculations*, Ph.D. thesis, The Acoustics Laboratory, Technical University of Denmark, report 55.

Kale, I., Gryka, J., Cain, G. D. and Beliczynski, B. 1994. FIR filter order reduction: balanced model truncation and Hankel-norm optimal approximation, *IEE Proc.-Vis. Image Signal Process.* **141**(3): 168–174.

Karjalainen, M. 1990. DSP software integration by object-oriented programming - a case study of QuickSig, *IEEE ASSP Magazine* pp. 21–31.

Karjalainen, M., Härmä, A. and Laine, U. K. 1997a. Realizable warped IIR filters and their properties, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany.

Karjalainen, M., Härmä, A., Laine, U. and Huopaniemi, J. 1997b. Warped filters and their audio applications, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, New York.

Karjalainen, M., Huopaniemi, J. and Välimäki, V. 1995. Direction-dependent physical modeling of musical instruments, *Proceedings of the International Congress on Acoustics*, Vol. 3, Trondheim, Norway, pp. 451–454.

Karjalainen, M., Piirilä, E., Järvinen, A. and Huopaniemi, J. 1999. Comparison of loudspeaker equalization methods based on DSP techniques, *Journal of the Audio Engineering Society* **47**(1/2): 14–31.

Kendall, G. 1995. A 3-D sound primer: directional hearing and stereo reproduction, *Computer Music Journal* **19**(4): 23–46.

Kendall, G. S. and Martens, W. L. 1984. Simulating the cues of spatial hearing in natural environments, *Proceedings of the International Computer Music Conference*, Paris, France, pp. 111–125.

Kendall, G. S. and Rodgers, C. A. P. 1982. The simulation of three-dimensional localization cues for headphone listening, *Proceedings of the International Computer Music Conference*.

Kirkeby, O. and Nelson, P. A. 1998. Digital filter design for virtual source imaging systems, *Presented at the 104th Convention of the Audio Engineering Society*, preprint 4688, Amsterdam, The Netherlands.

Kirkeby, O., Nelson, P. and Hamada, H. 1997. The stereo dipole - binaural sound reproduction using two closely spaced loudspeakers, *Presented at the 102nd Convention of the Audio Engineering Society*, preprint 4463, Munich, Germany.

Kirkeby, O., Nelson, P. and Hamada, H. 1998. The stereo dipole - binaural sound reproduction using two closely spaced loudspeakers, *Journal of the Audio Engineering Society* **46**(5): 387–395.

Kirkeby, O., Rubak, P., Johansen, L. G. and Nelson, P. A. 1999. Implementation of cross-talk cancellation networks using warped FIR filters, *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*, Rovaniemi, Finland, pp. 358–365.

Kistler, D. J. and Wightman, F. L. 1992. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction, *Journal of the Acoustical Society of America* **91**(3): 1637–1647.

Klayman, A. 1988. Stereo enhancement system, U.S. patent no. 4,748,669.

Kleiner, M., Dalenbäck, B.-I. and Svensson, P. 1993. Auralization – an overview, *Journal of the Audio Engineering Society* **41**(11): 861–875.

Koenen, R. 1999. MPEG-4: Multimedia for our time, *IEEE Spectrum* pp. 26–33.

Köring, J. and Schmitz, A. 1993. Simplifying cancellation of cross-talk for playback of head-related recordings in a two-speaker system, *Acustica* **179**: 221–232.

Kotorynski, K. 1990. Digital binaural/stereo conversion and crosstalk canceling, *Proc. 89th Convention of the Audio Engineering Society*, preprint 2949, Los Angeles, USA.

Krokstad, A., Strom, S. and Sorsdal, S. 1968. Calculating the acoustical room response by the use of a ray tracing technique, *Journal of Sound and Vibration* **8**(1): 118–125.

Kuhn, G. F. 1977. Model for the interaural time differences in the azimuthal plane, *Journal of the Acoustical Society of America* **62**(1): 157–167.

Kuhn, G. F. 1987. Physical acoustics and measurements pertaining to directional hearing, *in:* W. A. Yost and G. Gourevitch (eds), *Directional Hearing*, Springer-Verlag, pp. 3–25.

Kulkarni, A. and Colburn, H. S. 1995a. Efficient finite-impulse-response filter models of the head-related transfer function, *Journal of the Acoustical Society of America* **97**(5): 3278.

Kulkarni, A. and Colburn, H. S. 1995b. Infinite-impulse-response filter models of the head-related transfer function, *Journal of the Acoustical Society of America* **97**(5): 3278.

Kulkarni, A. and Colburn, H. S. 1997a. Finite-impulse-response models of the head-related transfer-function. Submitted to J. Acoust. Soc. Am.

Kulkarni, A. and Colburn, H. S. 1997b. Infinite-impulse-response models of the head-related transfer-function. Submitted to J. Acoust. Soc. Am.

Kulkarni, A., Isabelle, S. K. and Colburn, H. S. 1995. On the minimum-phase approximation of head-related transfer functions, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York.

Kulkarni, A., Isabelle, S. K. and Colburn, H. S. 1999. Sensitivity of human subjects to head-related transfer-function phase spectra, *Journal of the Acoustical Society of America* **105**(5): 2821–2840.

Kulowski, A. 1985. Algorithmic representation of the ray tracing technique, *Applied Acoustics* **18**(6): 449–469.

Kuttruff, H. 1995. Sound field prediction in rooms, *Proc. 15th Int. Congr. Acoust.*, Vol. 2, Trondheim, Norway, pp. 545–552.

Kuttruff, K. H. 1991. *Room Acoustics*, 3rd edn, Elsevier Science Publishers, Essex, England.

Kuttruff, K. H. 1993. Auralization of impulse responses modeled on the basis of ray-tracing results, *Journal of the Audio Engineering Society* **41**(11): 876–880.

Kyriakakis, C. 1998. Fundamental and technological limitations of immersive audio systems, *Proceedings of the IEEE* **86**(5): 941–951.

Kyriakakis, C. and Holman, T. 1998. Video-based head tracking for improvements in multichannel loudspeaker audio, *Presented at the 105th Convention of the Audio Engineering Society*, preprint 4845, San Francisco, CA, USA.

Kyriakakis, C., Tsakalides, P. and Holman, T. 1999. Surrounded by Sound, *IEEE Signal Processing Magazine* **16**(1): 55–66.

Laakso, T. I., Välimäki, V., Karjalainen, M. and Laine, U. K. 1996. Splitting the unit delay – tools for fractional delay filter design, *IEEE Signal Processing Magazine* **13**(1): 30–60.

Lahti, T. and Möller, H. 1996. The Sigyn Hall, Turku —- a concert hall of glass, *Proc. 1996 Nordic Acoustical Meeting*, Helsinki, Finland, pp. 43–48.

Larcher, V. and Jot, J.-M. 1997. Techniques d'interpolation de filtres audio-numériques, Application á la reproduction spatiale des sons sur écouteurs, *Proc. 4th Congress on Acoustics*, pp. 97–100.

Larcher, V., Jot, J.-M. and Vandernoot, G. 1998. Equalization methods in binaural technology, *Presented at the 105th Convention of the Audio Engineering Society*, preprint 4858, San Francisco, CA, USA.

Lehnert, H. and Blauert, J. 1992. Principles of binaural room simulation, *Applied Acoustics* **36**(3-4): 259–291.

Lindemann, W. 1986. Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals., *Journal of the Acoustical Society of America* **80**(6): 1608–1622.

Lokki, T., Hiipakka, J., Hänninen, J., Ilmonen, T., Savioja, L. and Takala, T. 1999. Real-time audiovisual rendering and contemporary audiovisual art, *Organized Sound*.

Lowe, D., Willing, S., Gonnason, W. and Williams, S. 1995. Stereo enhancement system, U.S. patent no. 5,440,638.

Lyon, R. and DeJong, R. 1995. *Theory and Application of Statistical Energy Analysis*, 2nd edn, Butterworth-Heinemann, 313 Washington Street, Newton, MA 02158.

MacCabe, C. J. and Furlong, D. J. 1991. Spectral stereo surround pan-pot, *Presented at the 90th Convention of the Audio Engineering Society*, preprint 3067, Paris, France.

Mackenzie, J., Huopaniemi, J., Välimäki, V. and Kale, I. 1997. Low-order modelling of head-related transfer functions using balanced model truncation, *IEEE Signal Processing Letters* **4**(2): 39–41.

Macpherson, E. A. 1991. A computer model of binaural localization for stereo imaging measurement, *Journal of the Audio Engineering Society* **39**(9): 604–622.

Maher, R., Lindemann, E. and Barish, J. 1996. Old and new techniques for artificial stereophonic image enhancement, *Presented at the 101st Convention of the Audio Engineering Society*, preprint 4371, Los Angeles, CA, USA.

Malham, D. and Myatt, A. 1995. 3-D sound spatialization using ambisonic techniques, *Computer Music Journal* **19**(4): 58–70.

Marshall, A. H. and Meyer, J. 1985. The directivity and auditory impressions of singers, *Acustica* **58**: 130–140.

Martens, W. L. 1987. Principal components analysis and resynthesis of spectral cues to perceived direction, *Proceedings of the International Computer Music Conference*, pp. 274–281.

Martin, J., Maercke, D. V. and Vian, J.-P. 1993. Binaural simulation of concert halls: a new approach for the binaural reverberation process, *Journal of the Acoustical Society of America* **94**(6): 3255–3264.

Mathworks 1994. *MATLAB Signal Processing Toolbox. User's Manual.*

McGrath, D. 1996. Method and apparatus for filtering an electronic environment with improved accuracy and efficiency and short flow-through delay, U.S. patent no. 5,502,747.

Mehrgardt, S. and Mellert, V. 1977. Transformation Characteristics of the External Human Ear, *Journal of the Acoustical Society of America* **61**(6): 1567–1576.

Meyer, J. 1978. *Acoustics and the Performance of Music*, Verlag das Musikinstrument, Frankfurt/Main.

Middlebrooks, J. C. and Green, D. M. 1992. Observations on a principal components analysis of head-related transfer functions, *Journal of the Acoustical Society of America* **92**(1): 597–599. Letters to the Editor.

Møller, H. 1992. Fundamentals of binaural technology, *Applied Acoustics* **36**(3-4): 171–218.

Møller, H., Sørensen, M., Hammershøi, D. and Jensen, C. 1995. Head-related transfer functions of human subjects, *Journal of the Audio Engineering Society* **43**(5): 300–321.

Møller, H., Sørensen, M., Jensen, C. and Hammershøi, D. 1996. Binaural technique: Do we need individual recordings?, *Journal of the Audio Engineering Society* **44**(6): 451–469.

Moore, B. C. 1981. Principal component analysis in linear systems: Controllability, observability, and model reduction, *IEEE Trans. Automat. Contr.*

Moore, B. C. J., Glasberg, B. R. and Baer, T. 1997. A model for the prediction of thresholds, loudness, and partial loudness, *Journal of the Audio Engineering Society* **45**(4): 224–240.

Moore, B. C. J., Peters, R. W. and Glasberg, B. R. 1990. Auditory filter shapes at low center frequencies, *Journal of the Acoustical Society of America* **88**: 132–140.

Moore, B. C. J., Peters, R. W. and Glasberg, B. R. 1996. A revision of Zwicker's loudness model, *Acta Acustica* **82**: 335–345.

Moorer, J. A. 1979. About this reverberation business, *Computer Music Journal* **3**(2): 13–28.

Morse, P. M. and Ingard, U. K. 1968. *Theoretical Acoustics*, Princeton University Press, Princeton, New Jersey.

Naylor, G. and Rindel, J. 1994. *Odeon Room Acoustics Program, Version 2.5, User Manual*, Technical Univ. of Denmark, Acoustics Laboratory, Publication 49, Denmark.

Nelson, P. A., Hamada, H. and Elliott, S. 1992. Adaptive inverse filters for stereophonic sound reproduction, *IEEE Transactions on Signal Processing* **40**(7): 1621–1632.

Nelson, P. A., Orduña-Bustamante, F. and Hamada, H. 1995. Inverse filter design and equalization zones in multichannel sound reproduction, *IEEE Transactions on Speech and Audio Processing* **3**(3): 185–192.

Nelson, P. A., Orduña-Bustamante, F. and Hamada, H. 1996a. Multichannel signal processing techniques in the reproduction of sound, *Journal of the Audio Engineering Society* **44**(11): 973–989.

Nelson, P. A., Orduña-Bustamante, F., Engler, D. and Hamada, H. 1996b. Experiments on a system for the synthesis of virtual acoustic sources, *Journal of the Audio Engineering Society* **44**(11): 990–1007.

Olive, S. 1998. Subjective evaluation of 3-D sound based on two loudspeakers, *Presented at the 105th Audio Engineering Society Convention*, San Francisco, CA, USA.

Oppenheim, A. V. and Schafer, R. W. 1989. *Discrete-Time Signal Processing*, Prentice Hall, New Jersey.

Parks, T. and Burrus, C. 1987. *Digital Filter Design*, John Wiley&Sons, New York.

Proakis, J. G. and Manolakis, D. G. 1992. *Digital Signal Processing: Principles, Algorithms, and Applications*, 2nd. edn, Macmillan, p. 969.

Pulkki, V. 1997. Virtual sound source positioning using vector base amplitude panning, *Journal of the Audio Engineering Society* **45**(6): 456–466.

Pulkki, V., Karjalainen, M. and Huopaniemi, J. 1998. Analyzing virtual sound sources using a binaural auditory model, *Presented at the 104th Convention of the Audio Engineering Society*, preprint 4697, Amsterdam, The Netherlands.

Pulkki, V., Karjalainen, M. and Huopaniemi, J. 1999. Analyzing virtual sound sources using a binaural auditory model, *Journal of the Audio Engineering Society* **47**(4): 204–217.

Rabinowitz, W. M., Maxwell, J., Shao, Y. and Wei, M. 1993. Sound localization cues for a magnified head: implications from sound diffraction about a rigid sphere, *Presence* **2**(2): 125–129.

Rasmussen, K. and Juhl, P. 1993. The effect of head shape on spectral stereo theory, *Journal of the Audio Engineering Society* **41**(3): 135–142.

Rayleigh, L. 1904. On the acoustic shadow of a sphere, *Phil. Transact. Roy. Soc. London* **203A**: 87–99.

Rayleigh, L. 1907. On our perception of sound direction, *Philos. Mag.* **13**: 214–232.

Riederer, K. 1998a. Repeatability analysis of HRTF measurements, *Presented at the 105th Convention of the Audio Engineering Society*, preprint 4846, San Francisco, CA, USA.

Riederer, K. A. J. 1998b. *HRTF Measurements*, Master's thesis, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, p. 134.

Roads, C. 1995. *The Computer Music Tutorial*, The MIT Press, Cambridge, Massachusetts, USA.

Runkle, P. R., Blommer, M. A. and Wakefield, G. H. 1995. A comparison of head related transfer function interpolation methods, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Institute of the Electrical and Electronic Engineers, New Paltz, New York, pp. 88–91.

Ryan, C. and Furlong, D. 1995. Effects of headphone placement on headphone equalization for binaural reproduction, *Presented at the 98th Convention of the Audio Engineering Society*, preprint 4009, Paris, France.

Sandvad, J. and Hammershøi, D. 1994a. Binaural auralization. Comparison of FIR and IIR filter representation of HIRs, *Presented at the 96th Convention of the Audio Engineering Society*, preprint 3862, Amsterdam, The Netherlands.

Sandvad, J. and Hammershøi, D. 1994b. What is the most efficient way of representing HTF filters?, *Proc. NORSIG'94*, pp. 174–178.

Savioja, L. 1999. *Modeling Techniques for Virtual Acoustics*, PhD thesis, Helsinki University of Technology, Laboratory of Telecommunications Software and Multimedia, Espoo, Finland.

Savioja, L., Backman, J., Järvinen, A. and Takala, T. 1995. Waveguide mesh method for low-frequency simulation of room acoustics, *Proceedings of the International Congress on Acoustics*, Vol. 2, Trondheim, Norway, pp. 637–640.

Savioja, L., Huopaniemi, J., Lokki, T. and Väänänen, R. 1997. Virtual environment simulation - Advances in the DIVA project, *Proceedings of the International Conference on Auditory Display*, Palo Alto, California, USA.

Savioja, L., Huopaniemi, J., Lokki, T. and Väänänen, R. 1999. Creating interactive virtual acoustic environments, *Journal of the Audio Engineering Society* **47**(9): 675–705.

Scheirer, E. D., Väänänen, R. and Huopaniemi, J. 1999. AudioBIFS: Describing audio scenes with the MPEG-4 multimedia standard, *IEEE Trans. Multimedia* **1**(3): 237–250.

Schroeder, M. and Atal, B. 1963. Computer simulation of sound transmission in rooms, *IEEE Conv. Record, pt. 7* pp. 150–155.

Schroeder, M. R. 1962. Natural-sounding artificial reverberation, *Journal of the Audio Engineering Society* **10**(3): 219–223.

Schroeder, M. R. 1970. Digital simulation of sound transmission in reverberant spaces, *Journal of the Acoustical Society of America* **47**(2 (Part 1)): 424–431.

Schroeder, M. R. 1973. Computer models for concert hall acoustics, *Am. J. Physics* **41**: 461–471.

Semidor, C. and Couthon, L. 1998. Sound sources and simulation softwares in room acoustics, *Proceedings of the joint meeting of the Acoustical Society of America and the International Congress on Acoustics*, Seattle, WA, USA, pp. 363–364.

Shinn-Cunningham, B., Lehnert, H., Kramer, G., Wenzel, E. and Durlach, N. 1997. Auditory displays, *in:* R. Gilkey and T. Anderson (eds), *Binaural and spatial hearing in real and virtual environments*, Lawrence Erlbaum Associates, Dahwah, New Jersey, pp. 611–663.

Slaney, M. 1993. An efficient implementation of the Patterson-Holdsworth filter bank, *Apple Technical Report 35*. Apple Computer, Inc.

Smith, J. O. 1983. *Techniques for Digital Filter Design and System Identification with Application to the Violin*, PhD thesis, Electrical Engineering Dept., Stanford University (CCRMA). Available as CCRMA Technical Report STAN-M-14.

Smith, J. O. 1991. Viewpoints on the history of digital synthesis, *Proceedings of the International Computer Music Conference*, Montreal, Canada, pp. 1–10.

Smith, J. O. 1997. Acoustic modeling using digital waveguides, *in:* C. Roads, S. T. Pope, A. Piccialli and G. De Poli (eds), *Musical Signal Processing*, Swets & Zeitlinger, Lisse, the Netherlands, chapter 7, pp. 221–264.

Smith, J. O. 1998. Principles of digital waveguide models of musical instruments, *in:* M. Kahrs and K. Brandenburg (eds), *Applications of Digital Signal Processing to Audio and Acoustics*, Kluwer Academic Publishers, Norwell, MA, USA, chapter 10, pp. 417–466.

Smith, J. O. and Abel, J. S. 1995. The Bark bilinear transform, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York. Available online at http://www-ccrma.stanford.edu/~jos/.

Smith, J. O. and Abel, J. S. 1999. The Bark and ERB bilinear transforms, *IEEE Transactions on Speech and Audio Processing*. Accepted for publication. Available online at http://www-ccrma.stanford.edu/~jos/.

Spandöck, F. 1934. *Ann. Physik V* **20**: 345.

Steiglitz, K. 1996. *A Digital Signal Processing Primer*, Addison-Wesley, Menlo Park, CA.

Stone, H. and Sidel, J. L. 1993. *Sensory Evaluation Practices*, 2nd edn, Academic Press, pp. 221–222.

Strube, H. W. 1980. Linear prediction on a warped frequency scale, *Journal of the Acoustical Society of America* pp. 1071–1076.

Sugiyama, K. and Irii, H. 1991. Comparison of the sound pressure radiation from a prolate spheroid and the human mouth, *Acustica* **73**: 271–276.

Svensson, U. P., Andersson, R. and Vanderkooy, J. 1997. A new interpretation of the Biot-Tolstoy edge diffraction model. Submitted to Journal of the Acoustical Society of America.

Swaminathan, V., Väänänen, R., Fernando, G., Singer, D. and Belknap, W. 1999. MPEG-4 version 2 Committee Draft 14496-1 (Systems) ISO/IEC JTC1/SC29/WG11, Seoul, Doc. W2739.

Takala, T., Hänninen, R., Välimäki, V., Savioja, L., Huopaniemi, J., Huotilainen, T. and Karjalainen, M. 1996. An integrated system for virtual audio reality, *Presented at the 100th Convention of the Audio Engineering Society*, preprint 4229, Copenhagen, Denmark.

Takeuchi, T., Nelson, P. and Kirkeby, O. 1997. Robustness of the performance of the stereo dipole to misaligment of head position, *Presented at the 102nd Convention of the Audio Engineering Society*, preprint 4464, Munich, Germany.

Theile, G. 1991. On the naturalness of two-channel stereo sound, *Journal of the Audio Engineering Society* **39**(10): 761–767.

Thurlow, W. R., Mangels, J. W. and Runge, P. S. 1967. Head movements during sound localization, *Journal of the Acoustical Society of America* **42**(2): 489–493.

Tolonen, T., Välimäki, V. and Karjalainen, M. 1998. *Evaluation of Modern Sound Synthesis Methods*, Helsinki University of Technology, Lab. of Acoustics and Audio Signal Processing, Report 47, p. 114.

Toshiyuki, T., Tomohiro, M. and Yasuhisa, O. 1994. Surround signal processing apparatus, European patent application no. 94305664.8.

Väänänen, R. and Huopaniemi, J. 1999. Spatial presentation of sound in scene description languages, *Presented at the 106th Convention of the Audio Engineering Society*, preprint 4921, Munich, Germany.

Väänänen, R., Välimäki, V., Huopaniemi, J. and Karjalainen, M. 1997. Efficient and parametric reverberator for room acoustics modeling, *Proceedings of the International Computer Music Conference*, Thessaloniki, Greece, pp. 200–203.

Välimäki, V. 1995. *Discrete-Time Modeling of Acoustic Tubes Using Fractional Delay Filters*, Dr. Tech. thesis, Helsinki University of Technology, Lab. of Acoustics and Audio Signal Processing, Report 37.

Välimäki, V. and Laakso, T. I. 1998. Suppression of transients in variable recursive digital filters with a novel and efficient cancellation method, *IEEE Transactions on Signal Processing* **46**: 3408–3414.

Välimäki, V., Huopaniemi, J., Karjalainen, M. and Jánosy, Z. 1996. Physical modeling of plucked string instruments with application to real-time sound synthesis, *Journal of the Audio Engineering Society* **44**(5): 331–353.

Valio, H. and Huopaniemi, J. 1996. Improved stereo enhancement, Finnish patent no. FI962181.

Vercoe, B. L., Gardner, W. G. and Scheirer, E. D. 1998. Structured audio: Creation, transmission, and rendering of parametric sound representations, *Proceedings of the IEEE* **86**(5): 922–940.

Vian, J.-P. and Martin, J. 1992. Binaural room acoustics simulation: practical uses and applications, *Applied Acoustics* **36**(3-4): 293–305.

Walsh, M. J. and Furlong, D. J. 1995. Improved spectral stereo head model, *Presented at the 99th Convention of the Audio Engineering Society*, preprint 4128, New York, NY, USA.

Ward, D. and Elko, G. 1998. Optimum loudspeaker spacing for robust crosstalk cancellation, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, Seattle, WA, USA, pp. 3541–3544.

Ward, D. and Elko, G. 1999. Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation, *IEEE Signal Processing Letters* **6**(5): 106–108.

Watanabe, Y., Tokunou, H. and Hamada, H. 1996. Subjective investigation of a new sound reproduction system (stereo dipole), *Journal of the Acoustical Society of America* **100**(4, pt.2): 2809.

Wenzel, E. and Foster, S. 1993. Perceptual consequences of interpolating head-related transfer functions during spatial synthesis, *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York.

Wenzel, E. M. 1992. Spatial sound and sonification. Presented in the International Conference on Auditory Display (ICAD'92). Also published in "Auditory Display: Sonification, Audification, and Auditory Interface, SFI Studies in the Sciences of Complexity", Proc. XVIII, edited by G. Kramer, Addison-Wesley, 1994).

Wenzel, E. M., Arruda, M., Kistler, D. J. and Wightman, F. L. 1993. Localization using nonindividualized head-related transfer functions., *Journal of the Acoustical Society of America* **94**(1): 111–123.

Widrow, B. and Stearns, S. 1985. *Adaptive Signal Processing*, Prentice-Hall, p. 474.

Wightman, F. L. and Kistler, D. J. 1989. Headphone simulation of free-field listening. I: stimulus synthesis, *Journal of the Acoustical Society of America* **85**(2): 858–867.

Wightman, F. L. and Kistler, D. J. 1992. The dominant role of low-frequency interaural time differences in sound localization, *Journal of the Acoustical Society of America* **91**: 1648–1661.

Wightman, F. L. and Kistler, D. J. 1999. Resolution of front-back ambiguity in spatial hearing by listener and source movement, *Journal of the Acoustical Society of America* **105**(5): 2841–2853.

Woodworth, R. S. and Schlosberg, R. 1954. *Experimental Psychology*, Holt, Rinehard and Winston, pp. 349–361.

Wu, S. and Putnam, W. 1997. Minimum perceptual spectral distance FIR filter design, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, Munich, Germany, pp. 447–450.

Yost, W. A. and Gourevitch, G. 1987. *Directional Hearing*, Springer-Verlag, p. 305.

Zacharov, N. and Huopaniemi, J. 1999. Results of a round robin subjective evaluation of virtual home theatre systems, *Presented at the 107th Convention of the Audio Engineering Society*, preprint 5067, New York, USA.

Zacharov, N., Huopaniemi, J. and Hämäläinen, M. 1999. Round robin subjective evaluation of virtual home theatre systems at the AES 16th international conference, *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*, Rovaniemi, Finland, pp. 544–556.

Zwicker, E. and Fastl, H. 1990. *Psychoacoustics: Facts and Models*, Springer-Verlag, Heidelberg, Germany, p. 354.

Index