UNIVERSITY OF SOUTHAMPTON
FACULTY OF SCIENCE, ENGINEERING AND MATHEMATICS
INSTITUTE OF SOUND AND VIBRATION RESEARCH

Visually Adaptive Virtual Sound Imaging using Loudspeakers

by

P. V. H. Mannerheim

Doctor of Philosophy

Faculty of Science, Engineering and Mathematics
Institute of Sound and Vibration Research

February 2008

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SCIENCE, ENGINEERING AND MATHEMATICS
INSTITUTE OF SOUND AND VIBRATION RESEARCH

Ph.D

Visually Adaptive Virtual Sound Imaging using Loudspeakers

by P. V. H. Mannerheim

Advances in computer technology and low cost cameras open up new possibilities for three dimensional (3D) sound reproduction. The problem is to update the audio signal processing scheme for a moving listener, so that the listener perceives only the intended virtual sound image. The performance of the audio signal processing scheme is limited by the condition number of the associated inversion problem. The condition number as a function of frequency for different listener positions and rotation is examined using an analytical model. The resulting size of the "operational area" with listener head tracking is illustrated for different geometries of loudspeaker configurations together with related cross-over design techniques. An objective evaluation of cross-talk cancellation effectiveness is presented for different filter lengths and for asymmetric and symmetric listener positions. The benefit of using an adaptive system compared to a static system is also illustrated. The measurement of arguably the most comprehensive KEMAR database of head related transfer functions yet available is presented. A complete database of head related transfer functions measured without the pinna is also presented. This was performed to provide a starting point for future modelling of pinna responses. The update of the audio signal processing scheme is initiated by a visual tracking system that performs head tracking without the need for the listener to wear any sensors. The solution to the problem of updating the filters without any audible change is solved by using either a very fine mesh for the inverse filters or by using commutation techniques. The filter update techniques are evaluated with subjective experiments and have proven to be effective both in an anechoic chamber and in a listening room, which supports the implementation of virtual sound imaging systems under realistic conditions. The design and implementation of a visually adaptive virtual sound imaging system is carried out. The system is evaluated with respect to filter update rates and cross-talk cancellation effectiveness.

Contents

Nomenclature

2θ	Source span
β	Regularisation factor
$\kappa(\mathbf{C})$	Condition number
$\mathcal{S}$	Shape space
$\mathcal{Y}_t$	State of the modelled object up to frame t
$\mathcal{Z}_t$	Measurements acquired up to frame t
ϕ	Azimuth angle
ψ	Elevation angle
ζ	Percentage mean square error
$S(k)$	Channel separation
X	Lateral plane
Y	Vertical plane
Z	Fore and aft plane
$\mathbf{\Pi}_t(k)$	Matrix of generalised inverse transfer functions that varies with time t
$\mathbf{\Psi}_t(k)$	Matrix of generalised transfer functions that varies with time t
$\mathbf{A}(k)$	Matrix of binaural filters
$\mathbf{C}(k)$	Matrix of transfer functions for the SD
$\mathbf{D}(k)$	Matrix of transfer functions for the 2-way OSD
$\mathbf{d}(k)$	Desired signals
$\mathbf{E}(k)$	Matrix of transfer functions for the 3-way OSD
$\mathbf{F}$	Feature map

$\mathbf{H}(k)$	Matrix of inverse transfer functions for the SD
$\mathbf{J}(k)$	Matrix of inverse transfer functions for the 2-way OSD
$\mathbf{K}(k)$	Matrix of inverse transfer functions for the 3-way OSD
$\mathbf{L}_{\mathcal{B}}^{'}$	Image of the probability for background given the data ("posterior probabilities")
$\mathbf{L}_{\mathcal{F}}^{'}$	Image of the probability for foreground given the data ("posterior probabilities")
$\mathbf{P}(k)$	Control performance matrix
$\mathbf{Q}$	Spline matrix
$\mathbf{r}(s)$	Spline curve
$\mathbf{u}(k)$	Binaural signals
$\mathbf{v}(k)$	Loudspeaker input signals
$\mathbf{W}$	Shape matrix
$\mathbf{w}(k)$	Signals produced at the listener's ears
$\mathbf{X}$	Image data that holds three vectors of pixel indices of an RGB image
$\mathbf{x}$	Pixel data
$\mathbf{y}$	Shape space vector

Acknowledgements

I would like to thank Samsung Advanced Institute of Technology (SAIT) who supported this work financially. The Institute of Sound and Vibration Research (ISVR) for access to the anechoic chamber and ISVR Consulting for the loan of measurement instruments. Especially, I would like to thank my supervisors Professor Philip Nelson and Professor Mark Nixon for having the confidence in my abilities to allow me to pursue the topics that interest me. Professor Philip Nelson who provided many ideas and much insight especially with respect to virtual sound. Professor Mark Nixon at the University of Southampton image processing and computer vision group who provided specialist advice in image processing. Professor Paul White at the ISVR signal processing group who contributed with advice regarding image and signal processing issues. I also would like to thank Dr Timos Papadopoulos and Mun Hum Park who participated in the measurement of a database of head related transfer functions. Finally, I would like to thank Damien Tavan and Nicolas Reboul for the assistance in software development.

Chapter 1

Introduction

1.1 Background

Binaural technology is often used for the reproduction of virtual sound images. The principle of binaural technology is to control the sound field at the listener's ears so that the reproduced sound field coincides with the desired real sound field. For the implementation of binaural technology over loudspeakers, it is necessary to cancel the cross-talk that prevents a signal meant for one ear from being heard at the other. However, such cross-talk cancellation, normally realized by time-invariant filters, works only for a specific listening location and the sound field can only be controlled in a limited area refereed to as the "sweet-spot". If the listener moves away from the optimal listening location, it is required that the inverse filters are updated so that the sweet-spot is steered to the listener's new location. The issues related to filter updates have been investigated intensively in this work. The aim of this project is to find a way to improve filter update techniques as well as to determine the filter update rate necessary to stabilize an acoustic image regardless of listener movement. This work is based on the assumption that the location of the listener is known from a visual head tracking device.

The effectiveness of cross-talk cancellation depends on the geometry of the system and in theory each frequency band can be reproduced from a loudspeaker pair with an optimal source span. Therefore the concept of Frequency Distributed Loudspeakers (FDLs) has been studied, and the idea is to reproduce each frequency from an optimal source angle within a certain listening area. The area that the listener can move within when the filters are updated can be determined by introducing the concept of "operational area". Hence, the operational area represents the region where the "sweet-spot" can be moved within using an adaptive virtual sound system. The extent of the operational area depends on performance criteria and is investigated thoroughly.

There is a strong demand for a head tracking algorithm within the field of virtual sound, because of the relatively small "sweet-spot" of a static virtual sound imaging system.

Adding access to a video camera for the audio system gives the possibility to track head movements and update the inverse filters accordingly. The increasing interest in visual tracking is due in part to the falling cost of computing power, video cameras and memory. A sequence of images grabbed at or near video rate typically does not change radically from frame to frame, and this redundancy of information over multiple images can be extremely helpful for analysing the input, in order to track individual objects.

1.2 Thesis outline

The topics covered in this thesis can be considered interdisciplinary and include spatial hearing, audio signal processing, acoustics and image processing. The aim of the thesis is to provide guidelines for ways to improve filter update techniques as well as to determine optimal filter update rates for stable adaptive virtual sound imaging. The aim of the thesis is also to provide suitable image processing algorithms for listener head tracking, and to determine the extent of the area for successful operation within which the listener is allowed to move as well as to evaluate related cross-over design techniques.

The thesis is divided into the following chapters. Chapter 1 gives an introduction to the thesis. Chapter 2 presents the theoretical basis for the investigated virtual sound imaging systems. The effect of geometry on system performance is described as is the design of inverse filters for virtual sound. The system performance at asymmetric listener positions is evaluated by presenting the condition number of the associated plant matrix as a function of frequency and listener position/rotation. Cross-over design techniques are evaluated using fixed and adaptive cross-over frequencies.

Chapter 3 presents the measurement of a comprehensive database of Head Related Transfer Functions (HRTFs). The following three arrangements of pinna models, “large pinna”, “small pinna” and “no pinna” were combined with the open and blocked ear canal cases to give a total of six full database measurements.

Chapter 4 presents interpolation schemes for Head Related Impulse Responses (HRIRs) and guidelines for how to reduce the size of the HRIR database. The term HRIR is used for time domain response and the term HRTF is used for the frequency domain throughout this thesis.

Chapter 5 presents an extensive objective evaluation of inverse filter design and cross-talk cancellation effectiveness. Three different virtual sound imaging systems are investigated namely: the Stereo Dipole (SD) (Nelson [74]), the 2-way Optimal Source Distribution (OSD) (Takeuchi [97]) and the 3-way OSD. The objective evaluation is carried out by evaluating cross-talk cancellation effectiveness for different filter lengths and for asymmetric and symmetric listener positions. The comparison is made of cross-talk cancellation effectiveness between a static approach and an adaptive approach.

Chapter 6 presents filter update techniques and filter update criteria that are evaluated by carrying out subjective experiments. The subjective experiments are carried out both under anechoic conditions and under realistic conditions in a listening room.

Chapter 7 focuses on image processing algorithms for listener head-tracking. Colour tracking by using multivariate Gaussian probability distributions is investigated. The presented colour tracking algorithm finds the listeners translation in the horizontal plane and in the vertical plane. Stereo area correlation tracking is investigated for finding the distance information of the listeners position. The colour tracking algorithm is combined with stereo area correlation tracking in order to track the listener with three degrees of freedom (DOF). Contour tracking for listener head tracking is also investigated for improved tracking of objects in cluttered environments. The contour tracker can also be extended to handle listener rotation.

Chapter 8 introduces a visually adaptive SD system that combines visual head tracking and binaural sound reproduction. The system carries out listener head tracking in the lateral plane and in the fore and aft plane while dynamically updating the cross-talk cancellation filters. Chapter 9 presents the conclusions together with further discussions, as well as an outline of promising future research.

1.3 Contributions of the thesis

The author has contributed to the field of binaural sound reproduction by increasing the understanding of the design of visually adaptive virtual sound imaging systems and the parameters that affect the system performance. Filter update rates for adaptive binaural sound reproduction have been determined and related alternative filter update techniques have been presented. Furthermore, efficient image processing algorithms have been developed that can be used to track the position of the listener for applications in virtual sound. A comprehensive evaluation of inverse filter and cross-over design techniques has been presented. The contributions of this thesis have increased the knowledge concerning virtual sound reproduction by studies within the young and promising field of interdisciplinary research between acoustics and image processing.

In virtual sound (Chapter 2), the performance of adaptive virtual sound imaging has been evaluated by showing the condition number of the plant matrix to be inverted as a function of frequency and listener position/rotation. The limitations of the “operational area” in which the listener can move has been depicted for two virtual sound imaging systems (SD and 3-way OSD). Crossover design techniques using fixed and adaptive approaches have also been presented. The contribution made here is to show the area where the listener can move under certain performance criteria and to introduce related cross-over design techniques.

The measurement of a comprehensive database of head related transfer functions (Chapter 3) has been carried out. The database contains six different measurement cases where the head related transfer function has been measured with and without pinnae and for open ear canal and blocked ear canal. Important features, such as, HRTF measured without the pinna and the non-directional character of the propagation along the ear canal have been clearly illustrated. The database is considered to be a valuable source, which can be used for future modelling of pinna responses.

In interpolation of head related impulse response functions (Chapter 4), a time domain interpolation technique that uses up-sampling and thresholding has been developed. The presented interpolation technique has been compared to previous approaches. The presented interpolation technique, is a useful tool for increasing the angular resolution of a measured set of HRIRs and can be applied with the filter update techniques presented in Chapter 6.

Filter design and cross-talk cancellation effectiveness (Chapter 5) for three virtual sound imaging systems have been objectively evaluated. The effectiveness of cross-talk cancellation for symmetric and asymmetric listener positions has been investigated by using simulations. Likewise, the effect of different numbers of filter coefficients has been investigated. The benefit of using an adaptive compared to a static virtual sound imaging system has been illustrated by a comprehensive set of simulations. The “sweet-spot” size is very limited for a static system and it has been shown how to steer it within a larger area by using an adaptive approach. The investigation gives guidelines for how to design inverse filters for adaptive systems in an optimal way.

Filter update techniques for adaptive virtual sound imaging has been investigated (Chapter 6). Two alternative filter update techniques have been proposed and matching filter update criteria have been determined by subjective experiments. The filter update criteria were named Just Noticable Difference (JND) and Just Noticable Change (JNC). The filter update criteria were determined both under anechoic and under normal listening conditions. This chapter gives guidelines for the design of filter update algorithms that can be used in various applications related to adaptive audio systems.

Image processing algorithms have been evaluated for the purpose of tracking the head of the listener (Chapter 7). Two algorithms that can track translation and one that can track distance of objects have been presented in Chapter 7. The contribution here is the evaluation of image processing algorithms that are suitable for head tracking in adaptive virtual sound.

Finally, a prototype of a visually adaptive binaural sound reproduction system has been developed. The system combines visual head tracking and audio signal processing in order to reproduce virtual sound using a software only approach. The listener is allowed to move in the lateral plane and in the fore and aft plane without wearing any sensors while the cross-talk cancellation filters are updated in real-time. An innovative filter update technique has been introduced that results in very smooth real-time filter updates. The implemented filter update technique has been objectively evaluated by carrying out measurements in an anechoic chamber.

1.4 Related publications and reports

The following is a list of publications and reports related to the work described in this thesis.

P. Mannerheim, M. Park and P. A. Nelson. Visually Adaptive Sound Reproduction System. Technical report No 06/02, ISVR, 2006.

P. Mannerheim, P. A. Nelson and Y. Kim. Filter update techniques for adaptive virtual sound imaging. *Audio Engineering Society 120th Convention.* 2006.

P. Mannerheim, P. A. Nelson and Y. Kim. Image Processing Algorithms for Listener Head Tracking in Virtual Acoustics. *Proceedings of the Institute of Acoustics*, Vol. 28. Pt. 1. 2006.

P. Mannerheim and P. A. Nelson. A preliminary evaluation of image processing algorithms for listener head tracking. Technical report No 05/03, ISVR, 2005.

P. Mannerheim and P. A. Nelson. Adaptive Loudspeaker Based Virtual Acoustic Imaging System. Technical report No 05/07, ISVR, 2005.

Y. Cho, P. Mannerheim, and P. A. Nelson. Measurement of a Near Field Head Related Transfer Function Database. Technical report No 05/10, ISVR, 2005.

P. Mannerheim, M. Park, T. Papadopoulos and P. A. Nelson. The measurement of a database of head related transfer functions. Technical report No 04/07, ISVR, 2004.

P. A. Nelson, T. Takeuchi, J. Rose, T. Papadopoulos, and P. Mannerheim. Recent developments in virtual sound imaging systems. Technical report, No 04/01, ISVR, 2004.

P. Mannerheim, *Image processing algorithms for virtual sound*, MSc thesis, ISVR, 2003.

Chapter 2

Virtual sound

The performance of virtual sound imaging systems is affected by the geometry of the system and the design of the cross-talk cancellation filters. The system performance at asymmetric and symmetric listener positions is analysed by showing the condition number of the plant matrix as a function of frequency and listener orientation. The loudspeaker cross-over design is affected by an adaptive system approach and the problem is that the optimal cross-over frequencies for a certain source span change when the listener is moving. The cross-over design problem is addressed by exploiting fixed cross-over frequencies and adaptive cross-over frequencies for the systems under investigation. The optimisation of the geometry for the loudspeakers when using adaptive cross-talk cancellation is also discussed.

It is possible to reproduce the sound pressures at the ears of a listener that replicate accurately a pair of sound pressure time histories. The sound pressures to be reproduced can be those produced by a source of sound located at a specified spatial position relative to the listener. This approach can be used to create a convincing illusion for the listener of a virtual sound source at the specified spatial location. The position of the virtual sound source can in theory be placed at any spatial location. Hence, the binaural sound system can reproduce sound in three dimensions. This approach is based on cross-talk cancellation using loudspeakers, and is generally attributed to Atal and Schroeder [4], although Bauer [10] had previously investigated another method for the reproduction of binaural recordings. The digital cross-talk cancellation technique has been further investigated by several other authors, Damaske [24], Hamada [37], Neu [77], Cooper [19], [20], [21], Bauck [9], Nelson [73], [75], [74], [76], Gardner [30], Ward [100] and Takeuchi [97], and requires the design of a matrix of filters that operates on a binaural recording (or a pair of synthesised binaural signals) to derive the inputs of the two loudspeakers. The cross-talk cancellation matrix effectively inverts the matrix of transfer functions relating the loudspeaker input signals to the listener's ears signals, in order to reproduce the binaurally recorded signals at the ears of the listener.

Nelson and Kirkeby [74] found that the illusion in the listener was especially convincing when cross-talk cancellation was applied using two closely spaced loudspeakers (typically 10°). This sound reproduction system, named the Stereo Dipole, was found to have advantages over the traditional 60° Stereo configuration, especially with regard to robustness of system performance with respect to listener head movement. It was pointed out by Ward and Elko [100] that the transfer function matrix (between the loudspeakers and the listener) to be inverted became ill-conditioned when the path-length difference between one of the loudspeakers and two of the ears of the listener became equal to one half of the acoustic wavelength. The Stereo Dipole ensured a well-conditioned inversion problem over a particularly useful frequency range. This concept was extended by Bauck [9] and by Takeuchi [97], [95], [96], the latter introducing the concept of the Optimal Source Distribution (OSD) by demonstrating that the inverse matrix of transfer functions could be made well-conditioned over a wide frequency bandwidth by ensuring that the angular span of the loudspeakers vary with frequency. The idea is to reproduce each frequency from the loudspeaker span where the inverse problem is optimally well-conditioned.

An analytical investigation by Nelson [71] of the inversion problem showed clearly how the time domain response of the inverse filters was highly undesirable at the ill-conditioned frequencies, resulting in a sound field with a long duration and a complex wave field which would give a deterioration in the cross-talk cancellation performance for small head movements by the listener. The equivalent analysis in the frequency domain also showed that the cross-talk cancellation performance was dramatically reduced when the inversion problem is ill-conditioned. A free field model was used for the presented analytical investigation. A similar analytical investigation was presented by Nelson in [76], that extended the previous analysis by using a model of scattering of sound by the head of the listener based on Lord Rayleigh's analysis of sound interacting with a rigid sphere.

Sound localisation by the human auditory system was investigated by Lord Rayleigh [85] in the development of his Duplex Theory by using spherical scattering at a single frequency. It was concluded that the influence of the head at low frequencies (where the wavelength of sound is much larger than the diameter of the head) does not affect the relative amplitude of the sound at the two ears much, therefore the available cues for localisation must be the inter-aural time difference (ITD) between the signals arriving at the two ears. At higher frequencies (where the wavelength of sound is smaller or comparable to the diameter of the head) Lord Rayleigh concluded that the inter-aural level difference (ILD) must be the dominant cue for localisation. The Duplex Theory has been reviewed by Hafter and Trahiotis [36], and they conclude that ITDs are indeed significant at high frequencies, at least when the time difference envelopes of high-frequency carrier signals are detectable by the auditory system. The binaural approach taken here includes the ITD cue also at high frequencies.

The cross-talk cancellation scheme is normally realized by time-invariant filters that only works a specific location of listener with a relatively small “sweet-spot”. Due to these limitations, adaptive cross-talk cancellation schemes have been investigated by several researchers, Kyriakakos [56], Gardner [30], Rose [86] and Lentz [58]. Kyriakakos [56] developed such a system that tracks listener movement and then modifies the loudspeaker output based on the listener’s location. The approach was to use a simple time delay adjusted to take account for the head’s location. Although time delay is an important sound localisation cue for detecting the horizontal location of sounds containing low frequencies, the system is limited by not adjusting for other important sound localisation cues, such as ILD and spectral cues. The ILD cue especially important in localisation of middle and high frequency sounds and spectral cues are important in determining vertical location. The use of cross-talk cancellation was suggested as an improvement of the presented system.

Gardner [30], found that a system that use cross-talk cancellation and HRTFs and steering the “sweet-spot” greatly improves horizontal sound source localization performance when the listener’s head is laterally displaced or rotated with respect to the ideal position. It was also found that the head tracking scheme also enables dynamic localization cues that are useful for resolving front-back reversals. The results from this investigation also suggest that it is difficult to synthesize consistent images on one side of the head when both loudspeakers are on the opposite side, due to the problem of inverting the high frequency transmission paths.

The work presented by Rose [86] is concerned with the development of a visually adaptive virtual sound imaging system that utilises two loudspeakers (the SD configuration). The system adjusts for lateral head motion and use the head location information to select appropriate pre-designed virtual sound imaging filters that correspond to the listener’s head location. This investigation shows simulations of the performance of cross-talk cancellation for asymmetric listener positions in the lateral plane for the SD loudspeaker system.

Lentz [58] describes binaural synthesis and reproduction over loudspeakers with a dynamic (tracked) cross-talk cancellation scheme that only needs three to four loudspeakers to cover all listening positions. This system is developed to be used in virtual reality applications, such as the CAVE system at RWTH Aachen University. A performance criterion is given for stable cross-talk cancellation and listener rotation, which suggests that a ±45° speaker configuration allows the listener to rotate ±40° and that a ±90° speaker configuration allows the listener to rotate ±75°.

The work in this chapter describes the possibilities and limitations of the adaptive cross-talk cancellation scheme by laying out a solid theoretical framework. The effect of loudspeaker geometry on cross-talk cancellation is thoroughly investigated and performance criteria are presented for the SD and a FDL system (the 3-way OSD). It is shown how to optimise the loudspeaker geometry for adaptive cross-talk cancellation systems and how to design the cross-overs. The investigation takes into account for listener movement in the lateral direction, fore and aft direction and for listener rotation around the azimuth axis.

2.1 Binaural models

2.1.1 Free field model

The free field model is the simplest approximation of the transfer functions relating the loudspeaker input to the listener's ears. The assumption that is made is to not include the head of the listener and replace it with two receivers at the positions of the listener's ears. The sound sources are of point monopole type and the environment is assumed to be anechoic. The advantage of this model is that the analytical solution becomes simple. The acoustic complex sound pressure P produced by a point monopole source at a distance r is given by

$$P(r) = \frac{j\omega\rho_0 Q e^{-jkr}}{4\pi r} \tag{2.1}$$

where k is the wave number ($k = \omega/c_0$), ρ_0 is the density of the medium and, c_0 is the speed of sound, ω is the angular frequency and Q (volume velocity) is the effective complex source strength. The free-field frequency response function $C_{ff}(j\omega)$ of the path from a monopole source to a position in space at a distance r is found by assigning the acoustic pressure P as the output and the complex source volume acceleration $j\omega Q$ as the input.

$$C_{ff}(j\omega) = \frac{\rho_0 e^{-jkr}}{4\pi r} \tag{2.2}$$

The time domain impulse response is a scaled delta function that represents the delay produced by the sound propagation time (r/c_0). The simulations that are presented for free field conditions all use Equation 2.2. The free field model can provide useful results for the effects of basic geometry on the sound field.

2.1.2 Spherical head model

The classical solution for the scattering of sound from a rigid sphere can be used as a reasonable first approximation to the HRTF of the listener. The sound field of a point monopole source having a complex volume velocity Q is given in Equation 2.1

and the free-field frequency response function is given in Equation 2.2. The sound field is assumed to be radiated by a point source situated relative to a rigid sphere. The method for calculating the scattered sound field is described by Kirkeby [47]. The expression for $C_{ff}(j\omega)$ can be expanded in terms of an infinite series by using series expansions (Abramowitz [1]) for $\cos(kr)/kr$ and $\sin(kr)/kr$. The equation for the free-field frequency response function is given by

$$C_{ff}(j\omega) = -\frac{j\rho_0 k}{4\pi}\sum_{m=0}^{\infty}(2m+1)j_m(kr)\left[j_m(kr) - jn_m(kr)\right]P_m(\cos(\phi)) \tag{2.3}$$

where the distance r and the angle ϕ are defined in Figure 2.1. The functions j_m and n_m are respectively the m^{th}-order spherical Bessel and Neumann functions and P_m represents the Legendre polynomial of m^{th}-order. The frequency response function relating the scattered field pressure to the volume acceleration of the point monopole can also be expressed in series form as follows

$$C_s(j\omega) = -\frac{\rho_0 k}{4\pi}\sum_{m=0}^{\infty} b_m\left[j_m(ka) - jn_m(ka)\right]P_m(\cos(\phi)) \tag{2.4}$$

where b_m are are coefficients to be determined, a denotes the radius of the sphere and only outward going waves are assumed. The coefficients b_m are found by the application of zero normal pressure gradient on the surface of the sphere and are given by Nelson [74]

$$b_m = j(2m+1)j_m^{'}(ka)\frac{j_m(kr) - jn_m(kr)}{j_m^{'}(ka) - jn_m^{'}(ka)} \tag{2.5}$$

where the prime denotes differentiation with respect to the argument of the function. The total frequency response function $C_t(j\omega)$ is found by adding $C_s(j\omega)$ and $C_{ff}(j\omega)$ such that

$$C_t(j\omega) = C_s(j\omega) + C_{ff}(j\omega) \tag{2.6}$$

The frequency response function may be transformed into equivalent discrete time impulse responses by first windowing (for example, using a Hanning window) this continuous function in the frequency domain and then sampling the frequency response function at N points in the range from $\omega = 0$ to ω_s, where the latter denotes an equivalent discrete time sampling frequency. The discrete time impulse response is computed from the inverse discrete Fourier transform given by

$$c(n) = \frac{1}{N}\sum_{k=0}^{N-1} C(k)e^{j(2\pi nk)/N} \tag{2.7}$$

where k is the discrete frequency variable (not to be confused with the acoustic wavenumber k) and n denotes the discrete time variable.

2.1.3 Head related transfer function model

A database of HRTFs can be used as an accurate model for the acoustical properties of the listener. An HRTF describes, for a certain angle of incidence, the sound transmission from a free field to a point in the ear canal of the subject (usually a dummy head). HRTFs are essential in the synthesis of binaural signals used in virtual sound imaging systems. The HRTF takes into account the reflections and diffractions from the human torso, head and pinna (Moller [69]). The previously presented models do not include the effect of the listener's pinnae, torso and neck. The complexities of pinnae, torso and neck makes it sometimes convenient to measure the HRTFs instead of creating either an analytical or numerical model of the listener.

The HRTF database is usually measured in an anechoic chamber using a dummy head, as described in Chapter 3. In some cases individual head related transfer functions are used, which are found by placing microphones in the ears of the listener and measuring the transfer functions. The ISVR-Samsung KEMAR database was measured in an anechoic chamber at a sampling frequency of 48 kHz. This database is used for the simulations and subjective experiments carried out during this project.

In order to find the angle of interest some interpolation of the data is necessary. The database of HRTFs employs a discrete sampling of a continuous space of spatial locations. The interpolation task is far from obvious and there are many approaches possible. A spherical interpolation scheme is described by Takeuchi [95] that uses bilinear interpolation. This approach first decomposes the HRTFs into magnitude and phase and then carries out the interpolation on these elements separately. A commonly used interpolation process is carried out by interpolating directly on the complex valued frequency response (Middlebrooks [67]). The approach throughout this thesis is to perform interpolation on an ITD-equalised database whenever it is possible. The exception is when commutation is applied on inverse filters, then the interpolation is performed with the delay information included. However, when commutation is carried out then the distance between the neighboring HRTFs is very small as is the delay. The interpolation of the ITD-equalised database can be carried out either in the time domain on the coefficients directly or in the frequency domain on the complex valued frequency response with similar results. The interpolation algorithms that are used in this project are further described in Chapter 4.

The HRIRs are sampled in the time domain and in some cases it is desirable to correct the propagation delay with a finer resolution than one unit of sample delay. When the amount of delay corresponds to an integer number of samples then the HRIR can simply be shifted in time by the given number of samples. If the propagation time delay is not equal to an integer number of samples, one can first reconstruct the continuous signal from the sampled signal and then shift by the appropriate amount of delay and then re-sample the signal. In practice one can apply up-sampling in order to create a finer resolution in the time domain and then shift the HRIR by an appropriate number of samples and then down-sample the signal. This is the procedure that is used for the ITD-equalisation technique presented in Chapter 4. An alternative method is to filter the signal with a linear phase all-pass filter with unity magnitude and with constant group delay that is equal to the propagation time (Laakso [57]). An often used fractional delay filter is a sampled sinc function given by

$$h_D(n) = \frac{\sin\left(\pi(n - T_D)\right)}{\pi(n - T_D)} \tag{2.8}$$

where m is the discrete time index and T_D is the delay time in samples. The delay time T_D can either be an integer or non-integer number. Fractional delay filters are used in many of the simulations in Chapter 5 to compensate for distances of non-integer delay. Re-sampling and time shifting of the HRIRs are used in the filter design for the subjective experiments presented in Chapter 6.

2.2 Matrix inversion for cross-talk cancellation filters

The signal processing block diagram associated with the inversion problem is illustrated in Figure 2.2, where $\mathbf{C}(k)$ is the matrix of transfer functions in the frequency domain that relates the vector of loudspeaker input signals $\mathbf{v}(k)$ to the vector of output signals $\mathbf{w}(k)$ produced at the listener's ears. This makes the signals produced the listener's ears $\mathbf{w}(k) = \mathbf{C}(k)\mathbf{v}(k)$. In order to create the vector $\mathbf{v}(k)$ of loudspeaker input signals, the matrix of cross-talk cancellation filters $\mathbf{H}(k)$ are multiplied with the vector of binaural signals $\mathbf{u}(k)$ such that $\mathbf{v}(k) = \mathbf{H}(k)\mathbf{u}(k)$. The binaural signals $\mathbf{u}(k)$ can be recorded or alternatively synthesised by convolving them with a pair of filters representing the transfer functions between the ears of the listener and the virtual source. The filters representing the transfer functions for the virtual sound source can be derived from a measured HRTF database. The cross-talk cancellation matrix should ideally make certain that the reproduced signals $\mathbf{w}(k)$ are a delayed version of $\mathbf{u}(k)$. The reproduced signals $\mathbf{w}(k)$ should be made equal to the desired signals $\mathbf{d}(k)$ at the listeners ears, where $\mathbf{d}(k) = \mathbf{u}(k)e^{-j\omega\Delta}$ (Δ represents the number of samples delay). It follows that

$$\mathbf{C}(k)\mathbf{H}(k) \approx e^{-j\omega\Delta}\mathbf{I} \tag{2.9}$$

where $\mathbf{I}$ is the identity matrix. The solution at each discrete frequency k for the cross-talk cancellation matrix is, in principle, given by

$$\mathbf{H}(k) \approx \mathbf{C}^{-1}(k)e^{-j\omega\Delta} \tag{2.10}$$

The transfer function matrix $\mathbf{C}(k)$ for a symmetric arrangement of two loudspeakers and the listener is given by

$$\mathbf{C}(k) = \begin{bmatrix} C_{11}(k) & C_{12}(k) \\ C_{21}(k) & C_{22}(k) \end{bmatrix} \tag{2.11}$$

where $C_{11}(k) = C_{22}(k)$ and $C_{21}(k) = C_{12}(k)$ are respectively, the frequency responses at the direct and the cross-talk paths. The inverse matrix of $\mathbf{C}(k)$ is given by

$$\mathbf{C}^{-1}(k) = \frac{1}{C_{11}^2(k) - C_{21}^2(k)} \begin{bmatrix} C_{11}(k) & -C_{21}(k) \\ -C_{21}(k) & C_{11}(k) \end{bmatrix} \tag{2.12}$$

writing the ratio of the cross-talk to the direct paths as $R(k) = C_{11}(k)/C_{21}(k)$ results in the following expression

$$\mathbf{C}^{-1}(k) = \frac{1}{C_{11}(k)\{(1 - R(k))(1 + R(k))\}} \begin{bmatrix} 1 & -R(k) \\ -R(k) & 1 \end{bmatrix} \tag{2.13}$$

Assuming that the two transmission paths are governed only by the propagation delay and amplitude reduction associated with the spherical spreading of sound from a point monopole source as defined by Equation 2.1, such that

$$C_{11}(k) = \frac{\rho_0 e^{-j\omega r_{11}/c_0}}{4\pi r_{11}} \tag{2.14}$$

$$C_{21}(k) = \frac{\rho_0 e^{-j\omega r_{21}/c_0}}{4\pi r_{21}} \tag{2.15}$$

then we may write $R(k) = ge^{-j\omega\tau}$ where $g = r_{21}/r_{11}$ is the ratio of the two path lengths and $\tau = (r_{21} - r_{11})/c_0$ is the difference between the acoustic travel times from one of the loudspeakers to the furthest and nearest ears of the listener.

This shows that it is in principle possible to realise the matrix $\mathbf{H}(k)$ of cross-talk cancellation filters. The only component that can not be realisable in Equation 2.13 is the term $1/C_{11}(k)$. Hence, the need for a modelling delay, which is introduced into the numerator of the solution for $\mathbf{H}(k)$ by the term $e^{-j\omega\Delta}$ and provided the delay Δ exceeds r_{11}/c_0 then there is no need to implement a time advance. In the inverse filters, the terms

$R(k)$ appearing in the numerator of Equation 2.13 are realisable since they represent pure delays and the terms $1/(1+R(k))$ and $1/(1-R(k))$ could also be realisable in discrete time as recursive filters. However, there is a potential difficulty with Equation 2.13 caused by the modulus squared of the filters appearing in the denominator i.e.

$$|1-R(k)|^2 = (1+g-2\cos(\omega\tau)), \quad |1+R(k)|^2 = (1+g+2\cos(\omega\tau)) \tag{2.16}$$

The ratio of g will be close to unity and these terms will become small as the frequency ω tends to zero and at frequencies where $\cos(\omega\tau) = 1$ or -1 respectively. This occurs when

$$\omega\tau = \frac{2\pi f \Delta r}{c_0} = n\pi \tag{2.17}$$

where n is an integer. This integer number should not be confused with the discrete time sample n. At these frequencies the response of the filters in Equation 2.13 becomes large and the frequency at which $\omega\tau = \pi$ or $f = 1/2\tau$ being the "ringing frequency" identified by Kirkeby [49]. The ringing frequency is associated with an undesirable response in the time domain.

Under the condition that Δh is small compared to the distance r ($r >> \Delta h$), the path-length difference Δr is given by,

$$\Delta r = r_{11} - r_{21} = 4\Delta h \sin(\theta) \tag{2.18}$$

Note that at asymmetric listener positions, the path-length difference becomes

$$\Delta r = \frac{1}{2}(r_{12} + r_{21} - r_{11} - r_{22}) \tag{2.19}$$

The ill-conditioned and well-conditioned frequencies can be written as a function of source span 2θ and integer n by combining Equation 2.17 and 2.18. When n is an odd number represents well-conditioned frequencies and n even number represents ill-conditioned frequencies such that

$$f = \frac{nc_0}{8\Delta h \sin(\theta)} \tag{2.20}$$

Figure 2.9 illustrates the relationship between source span 2θ and frequency for different odd integer numbers n.

The condition number of the plant matrix in the free field case can be compared to the condition number of the plant matrix in the HRTF case. It can be shown that by varying the receiver distance parameter Δh, the free field model can be adjusted to a better approximation of the HRTF model (Takeuchi [97]). The receiver distance corresponds to the shortest distance between the entrances of the ear canals of the KEMAR dummy head. The plant matrix of the HRTF model is similar to that of the free field model with a receiver distance of $2\Delta h = 0.13$ m where the incidence angle θ is small, which corresponds to the shortest distance between the ear canals of the KEMAR dummy head. Likewise, the plant matrix of the HRTF model is similar to that of the free field model with a receiver distance of $2\Delta h = 0.25$ m where the incidence angle θ is large. This is a significantly larger distance than the shortest distance between the entrances of the ear canals of the KEMAR dummy head and is likely to be caused by diffraction around the head. The receiver distance can for example be modeled to be a linear function of incident angle as in the simulations in Section 2.4.1.

2.2.1 Matrix inversion, SVD and regularisation

It is common practice to use regularisation for dealing with ill-conditioned inversion problems (Kirkeby [51], Nelson [76]). An optimal solution for the vector $\mathbf{v}(k)$ is the one that minimizes the sum of the squares of the errors $\mathbf{e}(k) = (\mathbf{d}(k) - \mathbf{w}(k))$ between the desired signals $\mathbf{d}(k)$ and the reproduced signals $\mathbf{w}(k)$. The solution to this optimisation problem (Nelson and Elliott [72]) is given by the optimal loudspeaker input signals as follows

$$\mathbf{v}_{opt}(k) = \left[\mathbf{C}^{\mathrm{H}}(k)\mathbf{C}(k) + \beta\mathbf{I}\right]^{-1}\mathbf{C}^{\mathrm{H}}(k)\mathbf{d}(k) \tag{2.21}$$

where β is the regularisation parameter. Given that $\mathbf{d}(k) = \mathbf{u}(k)e^{-j\omega\Delta}$ the cross-talk cancellation matrix in Equation 2.9 can be written in terms of the pseudo-inverse matrix $\left[\mathbf{C}^{\mathrm{H}}(k)\mathbf{C}(k) + \beta\mathbf{I}\right]^{-1}\mathbf{C}^{\mathrm{H}}(k)$, such that,

$$\mathbf{H}_R(k) = \left[\mathbf{C}^{\mathrm{H}}(k)\mathbf{C}(k) + \beta\mathbf{I}\right]^{-1}\mathbf{C}^{\mathrm{H}}(k)e^{-j\omega\Delta} \tag{2.22}$$

Using the singular value decomposition (SVD) of the matrix $\mathbf{C}(k)$ which can be written as

$$\mathbf{C}(k) = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{H}} \tag{2.23}$$

where $\mathbf{U}$ and $\mathbf{V}^{\mathrm{H}}$ are the unitary matrices of the left and right singular vectors, respectively, the superscript H denotes the Hermitian (complex conjugate) transpose and $\boldsymbol{\Sigma}$ is the diagonal matrix of singular values. Substituting into this solution gives the following solution

$$\mathbf{H}_R(k) = \mathbf{V}\left[\mathbf{\Sigma}^{\mathrm{H}}\mathbf{\Sigma} + \beta\mathbf{I}\right]^{-1}\mathbf{\Sigma}^{\mathrm{H}}\mathbf{U}^{\mathrm{H}}e^{-j\omega\Delta} \tag{2.24}$$

Using the free-field two-source two-field point model described above results in

$$\left[\mathbf{\Sigma}^{\mathrm{H}}\mathbf{\Sigma} + \beta\mathbf{I}\right]^{-1}\mathbf{\Sigma}^{\mathrm{H}} = \begin{bmatrix} \frac{|1+R(k)|}{|1+R(k)|^2+\beta} & 0 \\ 0 & \frac{|1-R(k)|}{|1-R(k)|^2+\beta} \end{bmatrix} \tag{2.25}$$

and this shows that the regularisation parameter limits the magnitude of the cross-talk cancellation filters at the particular frequencies where the terms $|1 + R(k)|$ or $|1 - R(k)|$ become small.

2.2.2 Realisability of inverse filters

The transfer function to be inverted $C(k)$ is naturally non-minimum phase where zeros appear outside the unit circle in the complex z-plane. The term $1/C(k)$ will then be unstable since zeros outside the unit circle become poles of the inverse filter. However, the poles outside the unit circle can also be interpreted as contributing to a stable but anti-causal component of the impulse response of the inverse filter [76]. The impulse response $c(n)$ associated with the transfer function $C(k)$ (using for example the spherical head model described in Section 2.1.2) can be described in terms of the z-transform

$$C(z) = z^{-q}B(z) \tag{2.26}$$

$$B(z) = b_0 + b_1z^{-1} + b_2z^{-2} + b_3z^{-3} + \ldots + b_{N-1}z^{-(N-1)} \tag{2.27}$$

where $B(z)$ is a polynomial q and denotes the number of samples delay. The number of coefficients used to represent the impulse response is denoted N. The terms b_n are non-zero values of the impulse response $c(n)$. Now the transfer function $B(z)$ can be factored into a product as

$$B(z) = b_0(1 - z_1z^{-1})(1 - z_2z^{-2})\cdots(1 - z_Nz^{-N}) \tag{2.28}$$

where z_i represent the zeros of the polynomial $B(z)$. The inverse of $B(z)$ can be written as a partial fraction expansion such as

$$B^{-1}(z) = \frac{A_1}{(1 - z_1z^{-1})} + \frac{A_2}{(1 - z_2z^{-2})} + \cdots + \frac{A_N}{(1 - z_Nz^{-N})} \tag{2.29}$$

The poles of this expression are assumed to be distinct and it can be shown (Proakis [84]) that when the inverse z-transform is evaluated, then the time domain sequence will be determined by the position of each relevant zero z_i relative to the unit circle $|z| = 1$. When zeros are placed inside the unit circle, the inverse z-transform results in a causal sequence that decays exponentially in forward time. When zeros are placed outside the unit-circle, it can be argued that the inverse z-transform results in an anti-causal sequence of infinite duration in backward time. When zeros lie on the unit circle, the resulting sequence will be of infinite duration in either forward or backward time. Hence, the closer the zero is to the unit circle, the slower the rate of decay of the impulse response. To overcome this it was shown by Kirkeby [50] that a regularised solution for the cross-talk cancellation matrix $\mathbf{H}_R(k)$ can deal with inverting a non-minimum phase system. The rate of decay of the impulse responses of the inverse filters can successfully be controlled by the regularisation parameter β that replaces each zero with a pair of zeros that are each further away from the unit circle in the z-plane (Nelson [76]). This will ensure that the response is sufficiently short compared to the length of the discrete Fourier transform used, hence the effects of "wrap around" errors are also minimized.

The effectiveness of cross-talk cancellation can be evaluated using the "control performance" matrix given by the product $\mathbf{C}(k)\mathbf{H}_R(k)$.

$$\mathbf{P}(k) = \mathbf{C}(k)\mathbf{H}_R(k) = \begin{bmatrix} P_{11}(k) & P_{12}(k) \\ P_{21}(k) & P_{22}(k) \end{bmatrix} \tag{2.30}$$

Perfect cross-talk cancellation would result in unit values of $P_{11}(k)$ and $P_{22}(k)$ and zero values of $P_{12}(k)$ and $P_{21}(k)$. The impulse responses $\mathbf{p}(n)$ is found by taking the inverse Fourier transform of the elements of $\mathbf{P}(k)$, and perfect cross-talk cancellation would then result in a Dirac impulse for $p_{11}(n)$ and $p_{22}(n)$ and zero values for $p_{12}(n)$ and $p_{21}(n)$.

$$\mathbf{p}(n) = \begin{bmatrix} p_{11}(n) & p_{12}(n) \\ p_{21}(n) & p_{22}(n) \end{bmatrix} \tag{2.31}$$

It can be demonstrated that the reduction in gain of the cross-talk cancellation filters that is produced by increasing the regularisation factor β results in a deterioration in performance as one might expect. There is a trade off between "control performance" and "control effort" that can be adjusted through the choice of regularisation factor.

2.3 Principles of the Stereo Dipole and the Optimal Source Distribution

The performance of the virtual sound imaging system depends partly on the source span as a function of frequency as described by Equation 2.20. The best control performance is achieved when n is an odd integer number as demonstrated Figure 2.9 (Takeuchi [97]). Ideally the source span varies continuously as a function of frequency in order to satisfy the requirement for n to be an odd integer number as in Equation 2.20. The choice of value for n is usually $n = 1$ since this gives the best low frequency control performance and also results in control over the sound field up to frequencies above 20 kHz (with a $\Delta h = 0.13$ m as for the KEMAR dummy head). The smallest source span is approximately 4° and the widest source span is 180°. The low frequency limit for optimal control of the sound field is about 300-400 Hz ($n = 1$) when the distance between the ears is 0.13 m.

The idea of using a loudspeaker with a source span that varies continuously as a function frequency is not feasible to implement in practice. A feasible solution is to discretise the source span. The plant matrix is well conditioned in a relatively wide frequency region around the optimal frequency. Therefore, one can allow n to have some width, for example $\pm\nu, 0 < \nu < 1$, which results only in a small reduction of control performance. This can be interpreted as using the well-conditioned frequencies only and excluding ill-conditioned frequencies by limiting the frequency range to be used for a certain source span. It is possible to build a practical system that covers most of the audible frequency range with a few sets of sources with different source spans.

An example of a 3-way OSD system is illustrated in Figure 2.5. The aim with the system is to ensure a condition number that is as small as possible over a frequency range that is as wide as possible. Therefore, the source spans were chosen at the extreme positions for the high frequency and low frequency limits. The high frequency limit is set to 20 kHz and the low frequency limit is depicted by the maximum source span of 180°. This gives $\nu = 0.7$ and the source spans becomes 6.2° for the high frequency units, 32° for the mid frequency units and 180° for the low frequency units as presented in Figure 2.6. The high frequency units (6.2°) covers the frequency range up to 20 kHz for $n+\nu = 1.7$. The low frequency limit of the low frequency units is about 110 Hz for $n-\nu = 0.3$. The cross-over frequencies are given by $n - \nu = 0.3$ and $n + \nu = 1.7$ for each pair of units and are located at 600 Hz and 4 kHz.

An example of a 2-way OSD system is illustrated in Figure 2.7. The aim with this system is again to ensure a condition number that is as small as possible over a frequency range as wide as possible. The source spans are chosen to be 6.9° and 120°, which gives $\nu = 0.9$. The mid-high frequency units (6.9°) covers the frequency range up to 20 kHz ($n + \nu = 1.9$). The low-mid frequency units covers the frequency range from about 40

Hz ($n - \nu = 0.1$) up to the cross-over frequency at 900 Hz. The cross-over frequency is given by $n - \nu = 0.1$ and $n + \nu = 1.9$ for each pair of units. The condition number of the 2-way OSD system is illustrated in Figure 2.8.

By using a larger number of sources, i.e. a 4-way OSD system or 5-way OSD system for example, the smaller the width of $n(\pm\nu)$ becomes. In this evaluation a value of $0.1 < n\pm\nu < 1.9$ is chosen, with the motivation that by allowing the listener to move and rotate (azimuth angle), degradation in performance must be allowed and it is reasonable to accept a degradation from a 3-way OSD system (with $0.3 < n \pm \nu < 1.7$) down to a 2-way OSD system ($0.1 < n \pm \nu < 1.9$). Hence, at the worst listener position, the 3-way OSD system will perform as the 2-way OSD system at its optimal position (on-axis and with azimuth angle $\phi = \mathit{0}$) in terms of the condition number of the transfer function matrix. The SD system investigated below is also using the interval of $0.1 < n \pm \nu < 1.9$ for comparative purposes.

2.4 Condition number for the inversion problem

The analytical solution of the two source-two field point inversion problem is presented by Nelson [71], [76], for the free-field model and spherical head model respectively. The “ringing frequency” is associated with ill-conditioning of the frequency response function matrix to be inverted and results in a complex sound field at the listeners ears in the time domain and a reduction in the size of the “sweet-spot”. The condition number of the matrix $\mathbf{C}(k)$ is defined in terms of the singular value decomposition (SVD) of the matrix, as described in Equation 2.23. The condition number $\kappa(\mathbf{C})$ of the matrix $\mathbf{C}(k)$ is given by the ratio of the maximum to minimum singular values that comprise the elements of the diagonal of the matrix $\mathbf{\Sigma}$. The condition number is a well-known parameter in dealing with matrix inversion problems (Golub [32]). If the loudspeaker input signals $\mathbf{v}(k)$ are determined from the solution for the cross-talk cancellation matrix in Equation 2.10, then it follows that,

$$\mathbf{v}(k) = \mathbf{C}^{-1}(k)\mathbf{d}(k) = \mathbf{C}^{-1}(k)\mathbf{u}(k)e^{-j\omega\Delta} \tag{2.32}$$

It can be shown that [32] the errors $\delta\mathbf{v}(k)$ in the solution for $\mathbf{v}(k)$ are related to the errors $\delta\mathbf{C}(k)$ in the specification of the matrix $\mathbf{C}(k)$ and the errors $\delta\mathbf{d}(k)$ in the specification of the desired signals $\mathbf{d}(k)$ by the inequality

$$\frac{\|\delta\mathbf{v}(k)\|}{\|\mathbf{v}(k)\|} \leq \kappa(\mathbf{C}) \left[\frac{\|\delta\mathbf{C}(k)\|}{\|\mathbf{C}(k)\|} + \frac{\|\delta\mathbf{d}(k)\|}{\|\mathbf{d}(k)\|} \right] \tag{2.33}$$

In Equation 2.33, the symbol $\|\|$ denotes the 2-norm, which is the sum of the squared elements of a vector or the square root of the largest singular value of a matrix. The error in the solution for the loudspeaker inputs $\mathbf{v}(k)$ and from other error sources (in recording or synthesis of binaural signals or in measuring the transfer function matrix $\mathbf{C}(k)$) can be amplified by the condition number of the matrix to be inverted. A large condition number leads to large errors in the solution. For the case with two loudspeakers placed symmetrically relative to the listener and assuming free-field transfer functions as in Equation 2.1 above then the SVD of the matrix $\mathbf{C}(k)$ results in the following,

$$\mathbf{\Sigma} = C_{11}(k) \begin{bmatrix} |1 + R(k)| & 0 \\ 0 & |1 - R(k)| \end{bmatrix} \tag{2.34}$$

$$\mathbf{U} = \frac{1}{\sqrt{2}} \begin{bmatrix} \sqrt{\frac{|1+R(k)|}{|1+R*(k)|}} & \sqrt{\frac{|1-R(k)|}{|1-R*(k)|}} \\ \sqrt{\frac{|1+R(k)|}{|1+R*(k)|}} & -\sqrt{\frac{|1-R(k)|}{|1-R*(k)|}} \end{bmatrix} \tag{2.35}$$

$$V = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{2.36}$$

The condition number is effectively given by the ratio $|1 + R(k)| \, / \, |1 - R(k)|$ and as described by Nelson [71], the peaks in the condition number occur at frequencies where the path-length difference either tends to zero or is equal to integer multiples of one half of the acoustic wavelength. The first maximum in the condition corresponding to a path-length difference of one half acoustic wavelength and defines the “ringing frequency”. The effect of narrowing the source span is that the ringing frequency increases and oppositely when the source span is increasing, then the ringing frequency is decreases.

Using the geometry of Figure 2.3 allows the condition number $\kappa(\mathbf{C})$ to be plotted as a function of source span 2θ for a fixed distance Δh. The distance between the ears Δh is chosen to be the average value between the ears of a listener (Δh=0.15 m corresponds to the average distance between a listener’s ears (Burkard [16])). Figure 2.10 illustrates the condition number $\kappa(\mathbf{C})$ as a function of source span 2θ. This illustrates $\kappa(\mathbf{C})$ plotted on a logarithmic greyscale, where the lighter bands show the regions of small condition number whilst the darker bands show the regions of high condition number.

2.4.1 Asymmetric and symmetric listener positions

The condition number $\kappa(\mathbf{C})$ for asymmetric and symmetric listener positions is investigated here. The free field condition number plot as a function of frequency and source span can be adjusted to fit an HRTF condition number plot, to represent a more realistic case. One way to adapt the free field condition number plot to the HRTF condition number plot is to vary the receiver distance. The most straightforward method is to adapt the start and end positions of the condition number plots, by taking the receiver distance at the minimum source span and at the maximum source span and then interpolating linearly in between. The receiver distance Δh varies here from $\Delta h = 0.13/2$ to $\Delta h = 0.25/2$ and is chosen to be a linear function of incidence angle θ that spans from $0° - 90°$. This aims to represent the typical value of the distance between the ears of a human head as function of incident angle θ. The condition number plotted on a logarithmic greyscale, where the lighter bands show the regions of small condition number whilst the darker bands show the regions of high condition number. The condition number is plotted as a function of frequency/lateral displacement (from -1 m to 1 m in the X-direction at for $Z = 0$), frequency/fore and aft displacement (from -1 m to 1 m in the Z-direction, on-axis $X = 0$) and frequency/azimuth angle (from 0° to 90° for azimuth angle ϕ). The coordinate system for listener positions (using the free field model) is presented in Figure 2.4. The investigation is carried out for two virtual sound imaging systems namely the SD and the 3-way OSD.

Figure 2.11 - Figure 2.14 illustrates the condition number $\kappa(\mathbf{C})$ as a function of frequency and lateral displacement for the following source spans: 10° (SD), 6.2° (3-way OSD high frequency unit), 32° (3-way OSD mid frequency unit) and 180° (3-way OSD low frequency unit). It can be seen that the well-conditioned region increases in frequency as the listener moves from the on-axis position out to -1 m and 1 m in the X-direction. This is caused by the decrease in path-length difference that occurs when the listener moves away from the optimal position in the X-direction.

Figure 2.15 - Figure 2.18 illustrates the condition number $\kappa(\mathbf{C})$ as a function of frequency and fore and aft displacement for different source spans. The well-conditioned region increases in frequency as the listener moves away from the sources (in the Z-direction from 0 m to 1 m) and the well-conditioned region decreases in frequency as the listener moves towards the sources (in the Z-direction from 0 m to -1 m). Figure 2.18 illustrates that the well-conditioned region increases in frequency as the listener moves both forward and backwards relative to the sources (in the Z-direction). This is because of the fact that the path-length difference is decreasing both when the listener is moving away and towards the sources (the maximum path-length difference occurs when the source span is 180 and the listener is located at the optimal position $X = 0$, $Z = 0$).

Figure 2.19 - Figure 2.22 illustrates the condition number $\kappa(\mathbf{C})$ as a function of frequency and azimuth angle for the different source spans. The well-conditioned region increases in frequency for smaller azimuth angles. It should also be noted that as the azimuth angle approaches 90°, all frequencies become ill-conditioned. This is due to the fact the path-length difference becomes 0 at an azimuth angle of 90°, hence in theory the cross-talk cancellation fails at 90°.

2.5 Operational area and cross-over design techniques

The performance of virtual sound imaging systems is affected by the geometry of the system and the position/rotation of the listener. When the listener is moving away from the optimal position then the condition number of the transfer function matrix will change and the system performance will decrease. This is illustrated by showing the frequency bands ($n \pm \nu$) as a function of listener position/rotation for the SD and the 3-way OSD. The area of successful operation is named the "operational area" and is a function of: frequency, listener position/rotation and $n \pm \nu$. The ν value is a performance criterion and can be chosen depending on the requirements of the application. The value of ν is here chosen to be 0.9 in order to represent the optimal performance of a static 2-way OSD system. Hence, in principle the performance of the adaptive 3-way OSD within the operational area will be as good or better (depending on position/rotation) than the performance of the 2-way OSD at its optimal listening position. The implication of the choosen performance criterion is that a low value of ν increases the robustness of the system with respect to head misalignment, which leads to improved sound source localisation performance (Takeuchi [96]). The aim is to show the size of the operational area and to find out whether the cross-overs need to be updated as a function of listener position and rotation or if they can stay fixed.

The relationship between frequency and lateral listener displacement together with fixed cross-over frequencies for the SD and the 3-way OSD are plotted in Figures 2.23 and 2.24. The "operational area" is here a function of frequency range and listener displacement in the lateral plane. The fixed upper cross-over frequencies are indicated by circles of the same colour as the plotted lines that represents the source spans and corresponds to n and $n \pm \nu$. Likewise, the fixed lower cross-over points are indicated by squares of the same colour as their associated source span. The operational area for the SD covers the frequency band 1000-13700 Hz (compared with on-axis performance of 720-13700 Hz) for displacement in the lateral plane between $-1 \leq X \leq 1$ m. The operational area for the 3-way OSD covers the frequency band 34-23200 Hz (compared with on-axis performance of 34-23200 Hz) for displacement in the lateral plane between $-1 \leq X \leq 1$ m. The fixed cross-over points for the SD and the 3-way OSD are presented in Table 2.2. It can be seen that the cross-over points for the source spans of the 3-way OSD are overlapping each other and there is not much benefit of updating the cross-overs. Within the overlapping sections, the cross-over points can be chosen depending on, for example, the transducer characteristics.

The relationship between frequency and fore and aft displacement of the listener together with fixed cross-over points for the SD and the 3-way OSD are plotted in Figure 2.25 and 2.26. The "operational area" in this case, is a function of frequency range and listener displacement in the fore and aft plane. The fixed cross-over points are again indicated by circles and squares of the same colour as the plotted lines that represents the source

System	Bandwidth, lateral X $(-1$ to $1)$ m [Hz]	Bandwidth, fore and aft Z $(-1$ to $1)$ m [Hz]	Bandwidth, azimuth angle ϕ $(-71^\circ$ to $71^\circ)$ [Hz]	Bandwidth, optimal position $(X = 0,\ Z = 0)$ [Hz]
SD	1000-13700	1100-6900	2300-13700	720-13700
3-way OSD	34-23200	38-11700	110-23200	34-23200

TABLE 2.1: **Operational bandwidth**. *Operational bandwidth for the SD and 3-way OSD at asymmetric and symmetric listener positions.*

spans and corresponds to n and $n \pm \nu$. The operational area for the SD covers the frequency band 1100-6900 Hz (compared with on-axis performance of 720-13400 Hz) for displacement in the fore and aft plane between $-1 \leq Z \leq 1$ m. The performance of the SD degrades at the high frequency end due to an increase in path-length difference Δr as the listener moves closer to the sources. The operational area for the 3-way OSD covers the frequency band 38-11700 Hz (compared with on-axis performance of 34-23200 Hz) for displacement in the fore and aft plane between $-1 \leq Z \leq 1$ m. As for the SD, the performance of the 3-way OSD degrades at the high frequency end due to an increase in path-length difference Δr as the listener moves closer to the sources. As in the previous case of lateral displacement, it can be seen that the cross-over points for the source spans of 3-way OSD are overlapping each other and there is not much benefit in updating the cross-overs.

The relationship between frequency and azimuth angle together with fixed cross-over frequencies for the SD and OSD are plotted in Figure 2.27 and Figure 2.28. The “operational area” is here a function of frequency range and azimuth angles. The operational area for the SD covers the frequency band 2300-13700 Hz for azimuth angles of $0^\circ \pm 71^\circ$, which can be compared to the on-axis performance of 720-13700 Hz. The operational area for the 3-way OSD covers the frequency range 110-23200 Hz for azimuth angles of $0^\circ \pm 71^\circ$, which can be compared to the on-axis performance of 34-23200 Hz. The fixed cross-over frequencies for the 3-way OSD are chosen so that the lower fixed cross-over frequency of the 6.2° source span coincides with the upper fixed cross-over frequency of 32°. Likewise, the lower fixed cross-over frequency of the 32° source span coincides with the upper fixed cross-over frequency of 180°. The result is a maximum listener rotation angle of ±71°. The maximum azimuth rotation angle of ±71° is imposed on the SD so that the results from 3-way OSD and SD can be compared with each other. The results from the system evaluation of asymmetric and symmetric listener positions with respect to operational bandwidth are summarised in Table 2.1.

It would be convenient and cost saving to use fixed cross-overs from a system point of view. Fixed cross-overs can be implemented using passive cross-overs, while adaptive cross-overs would probably need to be implemented in the digital domain. The adaptive cross-overs would result in a multi-channel output low level signal from a DSP and each channel would then require a digital to analogue converter and a power amplifier. The

System	Fixed cross-over, lateral $-1 \leq X \leq 1$ m [Hz]	Fixed cross-over, fore and aft $-1 \leq Z \leq 1$ m [Hz]	Fixed cross-over, azimuth angle ϕ ($-71°$ to $71°$) [Hz]
SD high-pass	1000	1100	23000
SD low-pass	13700	6900	13700
6.2° high-pass	1700	1800	3900
6.2° low-pass	23200	11700	23200
32° high-pass	280	300	650
32° low-pass	3900	2200	3900
180° high-pass	34	38	110
180° low-pass	650	650	650

Table 2.2: **Fixed cross-over frequencies**. *Fixed cross-overs for the SD and 3-way OSD. The presented cross-over frequencies are -3 dB/Octave points.*

cross-over frequencies for fixed cross-overs are presented in Table 2.2. It can be seen that the optimal cross-over frequencies are different depending on the movement of the listener.

Ideally, the cross-over should be able to handle all three types of movements without updating the filter. If the cross-over frequencies for the azimuth rotation are chosen, then the lateral movement will not be a problem. However, the fore and aft movement will become a problem for the 32° source span. The cross-over frequency for the 32° source span in the fore and aft movement is 2200 Hz compared to the cross-over frequency of 3900 Hz for the azimuth rotation. One way to overcome this without changing the system geometry, is to limit the azimuth rotation to the angle where the lower frequency of the 6.2° source span is equal to 2200 Hz, which occurs at an azimuth rotation angle of 55°. By lowering the low-pass cross-over frequency of the 32° source span and the high-pass cross-over frequency for the 6.2° source span to 2200 Hz and limiting the azimuth rotation to 55° all three types of movements can be allowed without updating the filters. The fixed cross-over frequencies for the 3-way OSD are then at 650 Hz and 2200 Hz. The high-pass filter for the 6.2° source span is set to 2200 Hz, as is the low-pass filter for the 32° source span and the high-pass filter of the 32° source span is set to 650 Hz as is the low-pass filter for the 180° source span. The upper frequency limit of the 3-way OSD system using fixed cross-overs is 11700 Hz and occurs during fore and aft movement. The lower frequency limit of the system 3-way OSD system using fixed cross-overs is 60 Hz and occurs at the maximum azimuth rotation of ±55°. The upper limit for the SD is 6900 Hz and the lower limit is 2300 Hz when all three investigated movements are allowed with an azimuth rotation angle of ±71°. If the azimuth rotation angle is reduced to 55° then the lower frequency limit becomes 1300 Hz.

The cross-over frequencies for adaptive cross-overs are presented in Table 2.3. When the cross-overs are allowed to be updated depending on the movement of the listener, then especially the azimuth rotation angle can be greater than in the case with fixed

System	Adaptive cross-over, lateral $-1 \leq X \leq 1$ m [Hz]	Adaptive cross-over, fore and aft $-1 \leq Z \leq 1$ m [Hz]	Adaptive cross-over, azimuth angle ϕ ($-71°$ to $71°$) [Hz]
SD high-pass	720-1000	360-1100	720-2300
SD low-pass	13700-19000	6900-20500	13700-43400
6.2° high-pass	1200-1700	620-1800	1200-3900
6.2° low-pass	23200-32400	11700-34800	23200-73500
32° high-pass	210-280	110-300	210-650
32° low-pass	3900-5300	2200-5700	3900-12400
180° high-pass	34-34	34-38	34-110
180° low-pass	650-650	650-730	650-2000

Table 2.3: **Adaptive cross-over frequencies**. *Adaptive cross-overs for the SD and 3-way OSD. The presented cross-over frequencies are -3 dB/Octave points.*

cross-overs. If the listener is allowed to perform all three types of described movements and to rotate ±71°, then the upper frequency limit of the 3-way OSD system is 11700 Hz and the lower frequency limit is 110 Hz. The azimuth rotation angle can be extended further by sacrificing low frequency performance. The presented cross-over frequencies are -3 dB/Octave points.

2.6 Discussion

The optimal geometry of the loudspeakers for adaptive cross-talk cancellation is different from the static case. There is an advantage in placing the loudspeakers far away from the listener since the angles between the listener and the loudspeakers will then change less when the listener is moving. For example, the 3-way OSD has low frequency units that span 180° and when the listener moves in the fore and aft direction, the angle of this source span is changing significantly. Placing the low frequency units in line with the high frequency units will result in a more robust system with respect to fore and aft movement. Another advantage is that it can be more convenient to place all the sources in a straight line in front of the listener with respect to practical installation issues. Also the low frequency performance does not change significantly when the sources are moved from for example 180° to 90°. In this case the lowest optimally reproduced frequency goes up to about 500 Hz from 350 Hz, which is illustrated in Figure 2.9. The high frequency performance can be extended by designing a system that goes up to the highest desirable frequency (often 20000 Hz) when the listener is closest to the loudspeaker.

2.7 Conclusion

It has been demonstrated that there is a relationship between cross-talk cancellation performance for virtual sound imaging systems and the condition number of the matrix of transfer functions that needs to be inverted. The dependency of condition number on frequency has been illustrated for the investigated asymmetric and symmetric listener positions. An analytical model that estimates the “operational area” where the “sweet-spot” can be moved within for the adaptive SD and 3-way OSD system has been presented. The issue of using fixed or adaptive cross-overs has been investigated and recommendations have been made. The use of fixed cross-overs will limit the azimuth rotation to 55°, which can be compared to using adaptive cross-overs that allows for 71° azimuth rotation. The adaptive cross-over approach has not been used in this project since they create additional computational complexity in the signal processing scheme. However, they can improve the overall system performance and should be considered for implementation in future projects. The analytical evaluation presented here is complementary to the objective evaluation of the system performance in Chapter 5. The next chapter presents the measurement of a database of HRTFs, and the acquired data is used for the objective evaluations in Chapter 4 and Chapter 5.

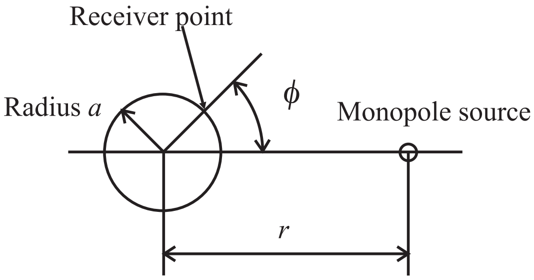


Figure 2.1: **Coordinate system for the spherical head model**. *The coordinate system with definitions that describes scattering of sound by a rigid sphere.*

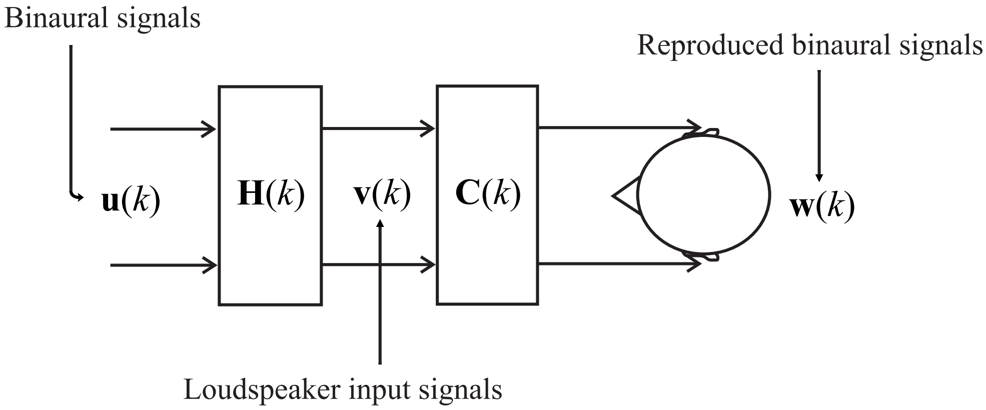


Figure 2.2: **Signal processing block diagram**. *The signal processing block diagram associated with the inversion problem.*

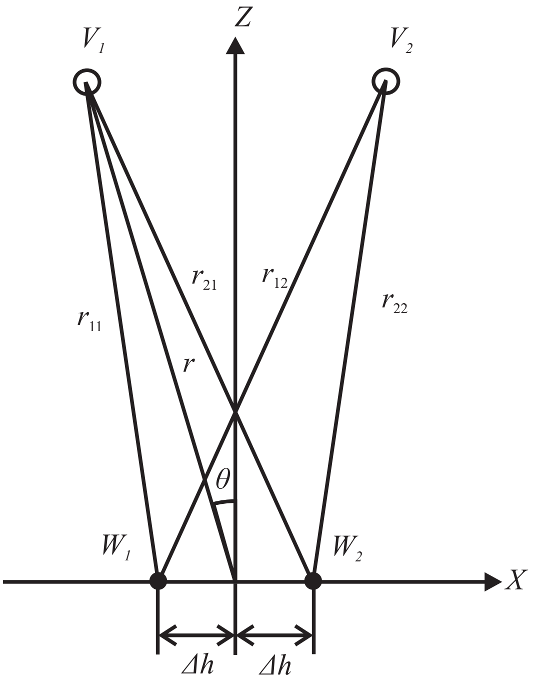


FIGURE 2.3: **Free field model**. *The geometry of the free field model with a two-source (loudspeakers) two-receiver (ears) system.*

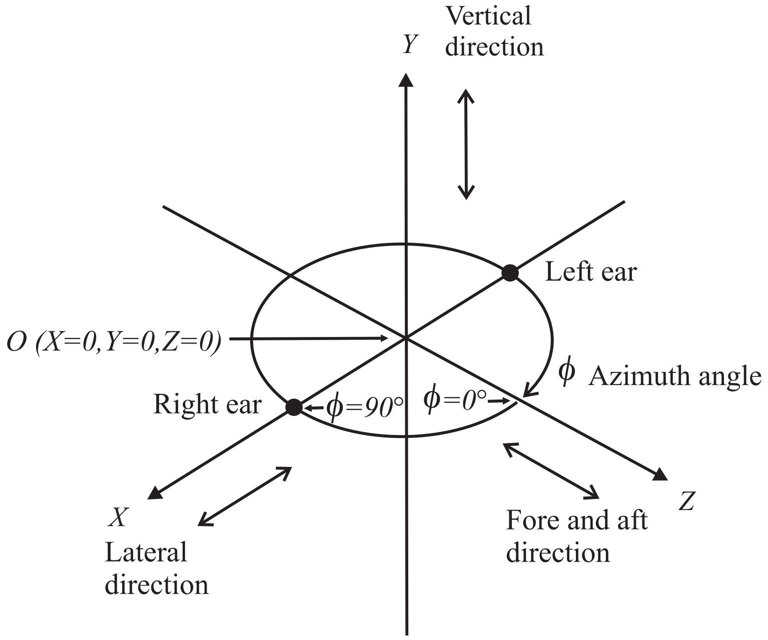


FIGURE 2.4: **The coordinate system of the free field model**.

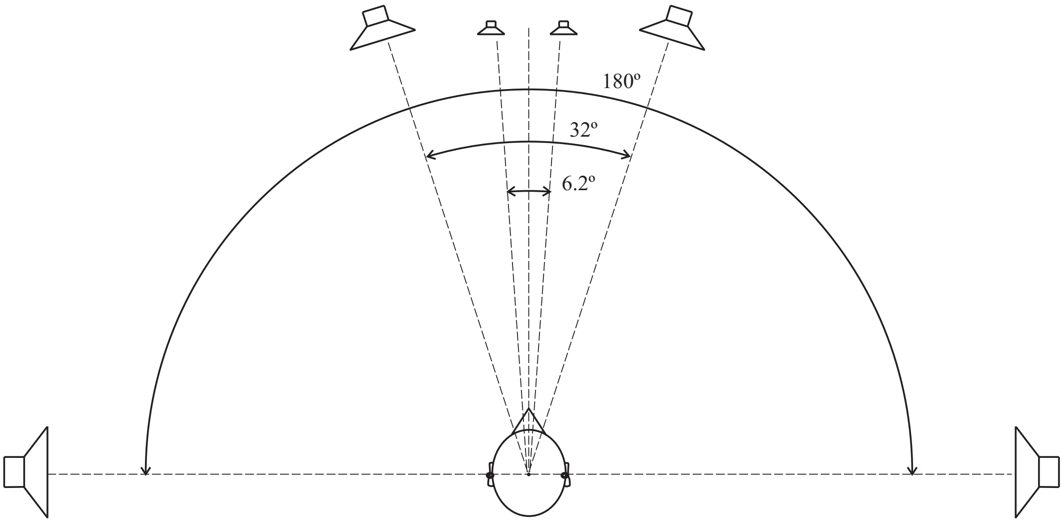


Figure 2.5: **Source spans of 3-way OSD**. *The source spans (2θ) for a 3-way OSD system with $n = 1$ and $\nu = 0.7$. The source spans are as follows:* 6.2° *for the high frequency unit,* 32° *for the mid frequency unit and* 180° *for the low frequency unit.*

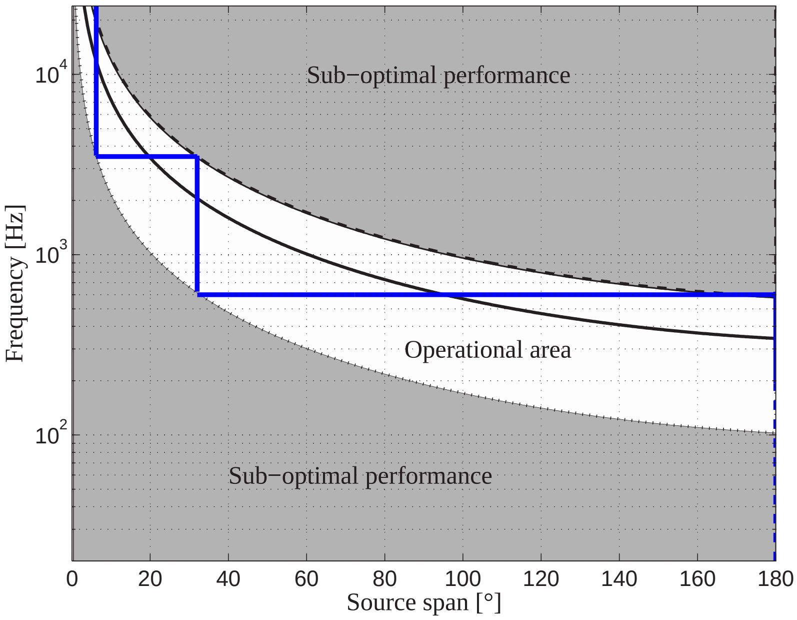


Figure 2.6: **Operational area of the 3-way OSD**. *The "operational area" for an OSD system with $n = 1$ and $\nu = 0.7$. This operational area is suitable for a 3-way OSD system. The source span 2θ is plotted as a function of frequency. The upper border of the operational area is indicated by a dashed line that represents $n + \nu$ and the lower border is indicated by a dotted line that represents $n - \nu$. The cross-over frequencies are at around 600 Hz and 4000 Hz respectively for the different source spans.*

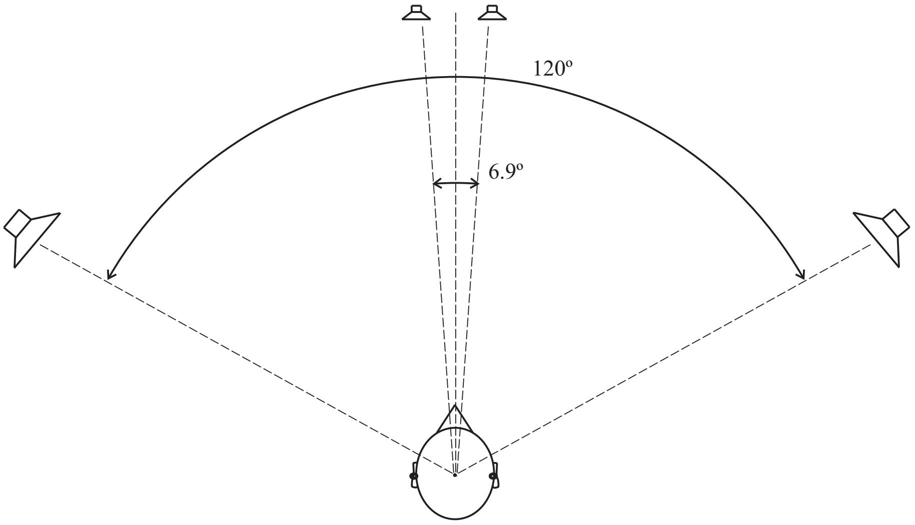


Figure 2.7: **Source spans of 2-way OSD**. *The source spans (2θ) for a 2-way OSD system with $n = 1$ and $\nu = 0.9$. The source spans are $6.9°$ for the high-mid frequency unit and $120°$ for the mid-low frequency unit.*

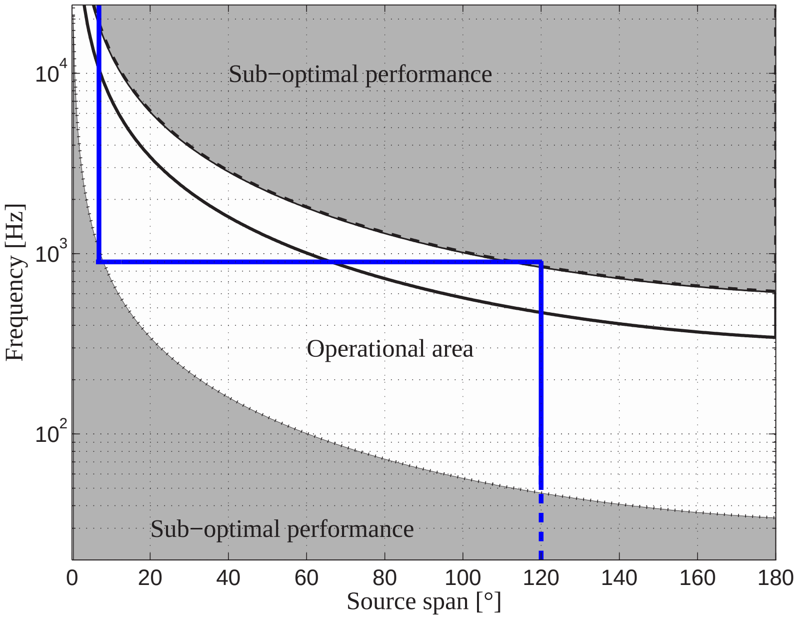


Figure 2.8: **Operational area of the 2-way OSD**. *The "operational area" for an OSD system with $n = 1$ and $\nu = 0.9$. This operational area is suitable for a 2-way OSD system. The source span 2θ is plotted as a function of frequency. The upper border of the operational area is indicated by a dashed line that represents $n + \nu$ and the lower border is indicated by a dotted line that represents $n - \nu$. The cross-over frequency is at around 900 Hz for the two source spans.*

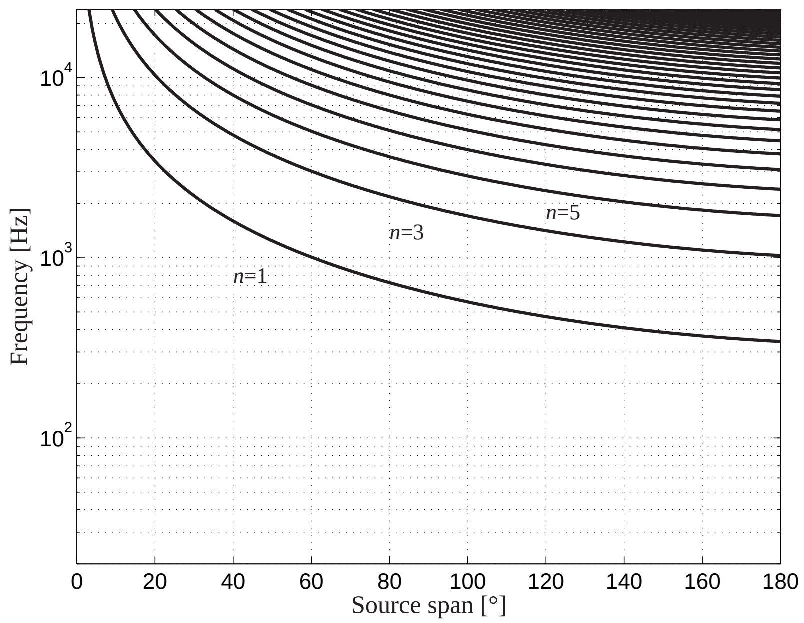


FIGURE 2.9: **Well-conditioned frequencies**. *The relationship between source span 2θ and well-conditioned frequencies for different odd integer number n.*

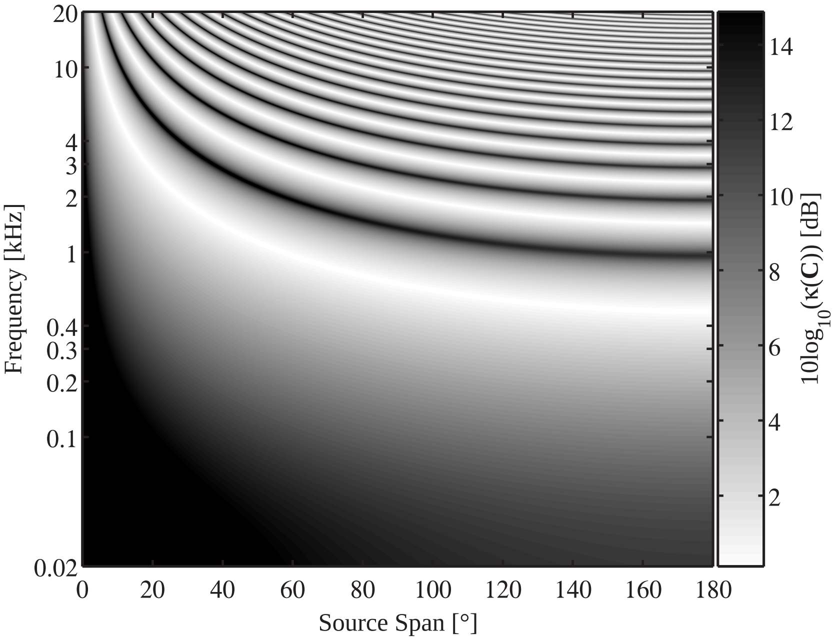


FIGURE 2.10: **Condition number as a function of source span and frequency**. *The relationship between condition number, source span 2θ and frequency is illustrated.*

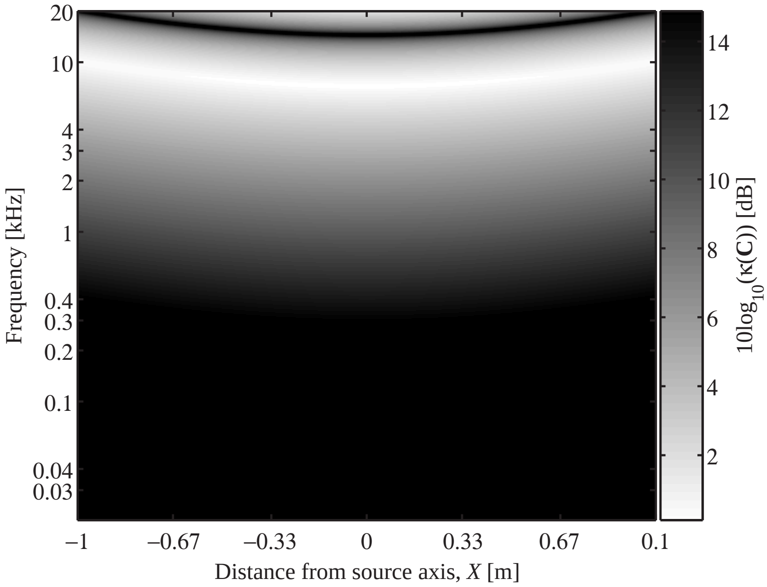


FIGURE 2.11: *The condition number as a function of frequency and lateral position for the* 10° *source span.*

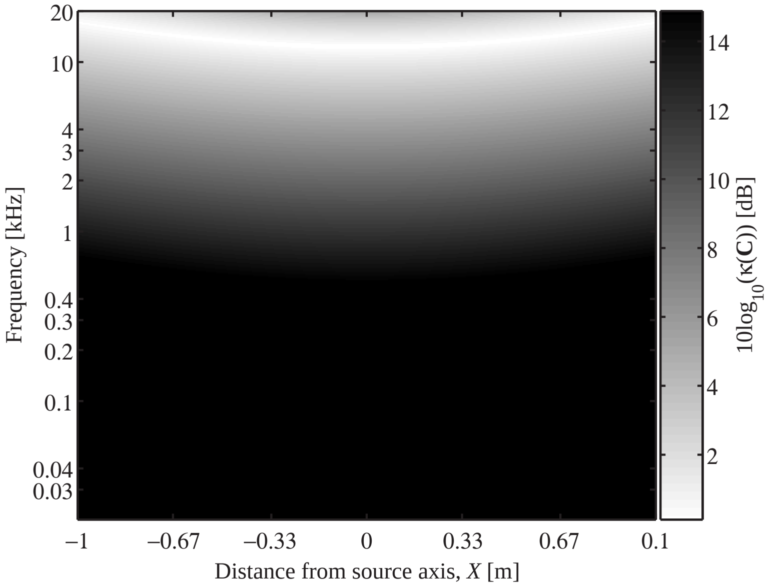


FIGURE 2.12: *The condition number as a function of frequency and lateral position for the* 6.2° *source span.*

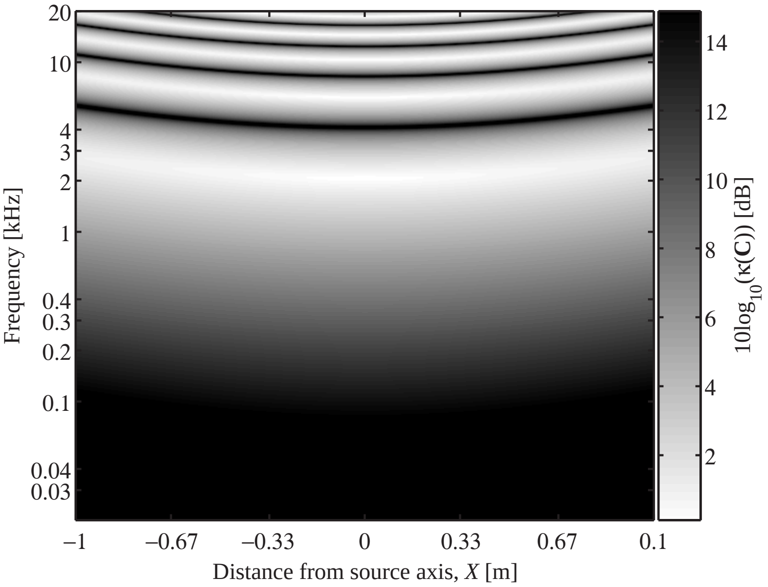


FIGURE 2.13: *The condition number as a function of frequency and lateral position for the* 32° *source span.*

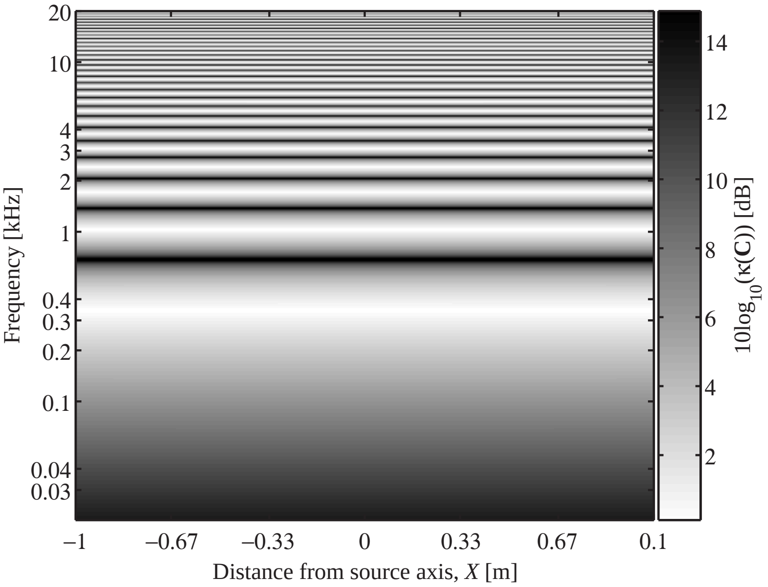


FIGURE 2.14: *The condition number as a function of frequency and lateral position for the* 180° *source span.*

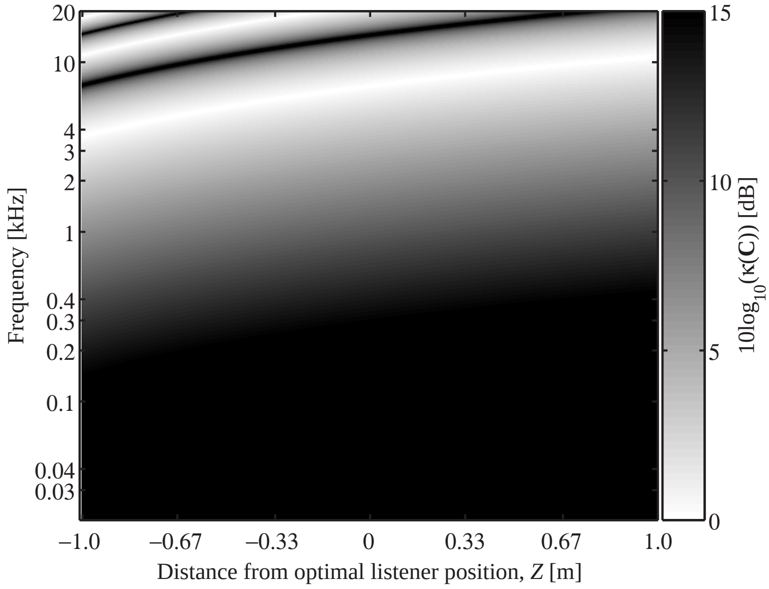


FIGURE 2.15: *The condition number as a function of frequency and fore and aft position for the* 10° *source span.*

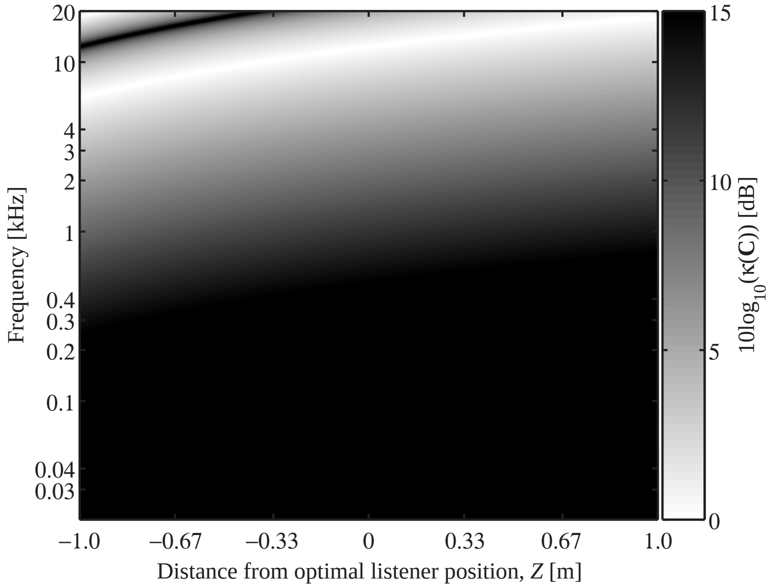


FIGURE 2.16: *The condition number as a function of frequency and fore and aft position for the* 6.2° *source span.*

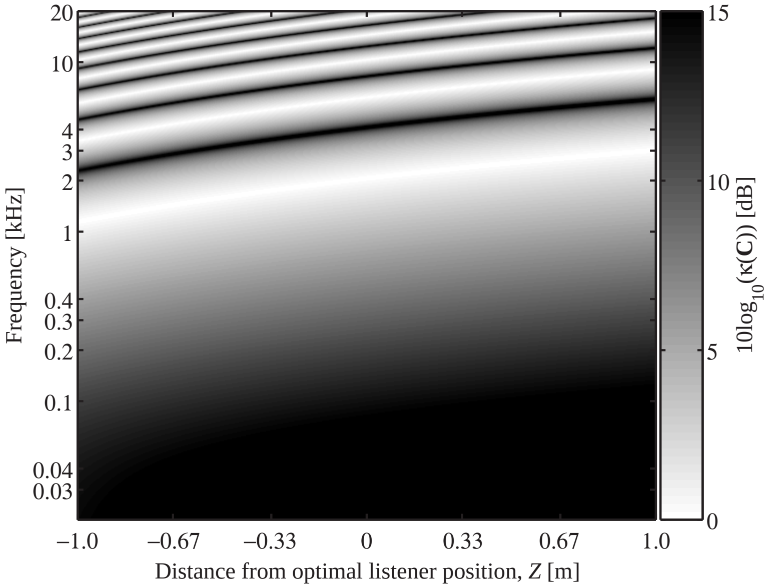


FIGURE 2.17: *The condition number as a function of frequency and fore and aft position for the* 32° *source span.*

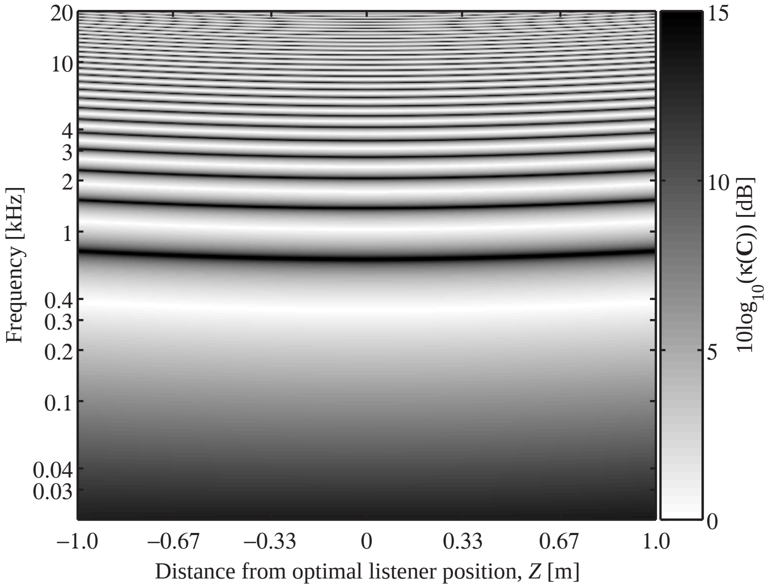


FIGURE 2.18: *The condition number as a function of frequency and fore and aft position for the* 180° *source span.*

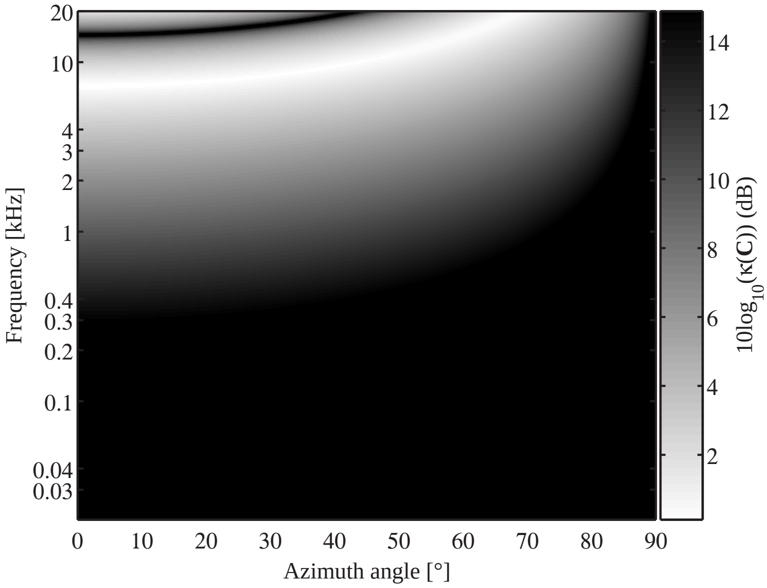


Figure 2.19: *Condition number as a function of frequency and azimuth angle ϕ for the* 10° *source span.*

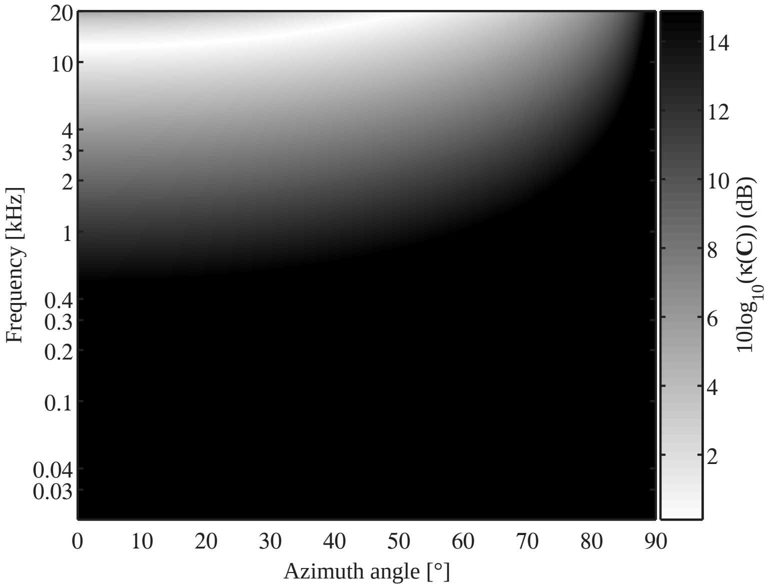


Figure 2.20: *Condition number as a function of frequency and azimuth angle ϕ for the* 6.2° *source span.*

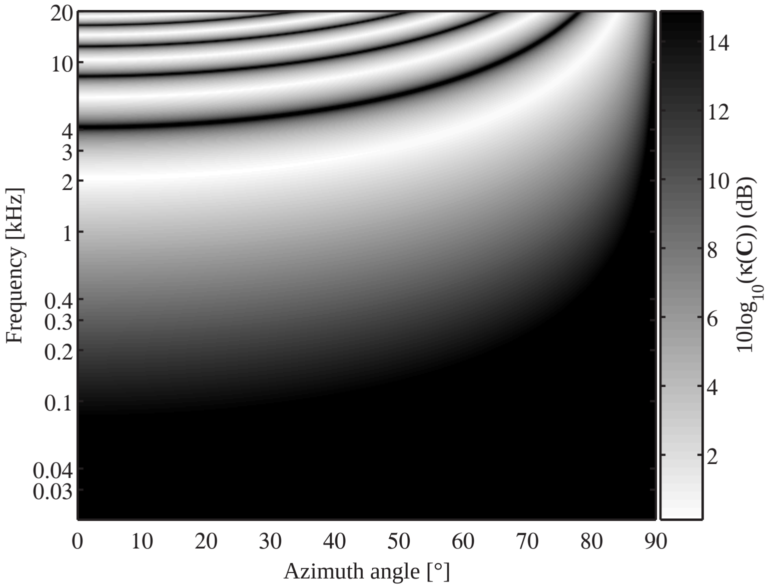


FIGURE 2.21: *Condition number as a function of frequency and azimuth angle ϕ for the* 32° *source span.*

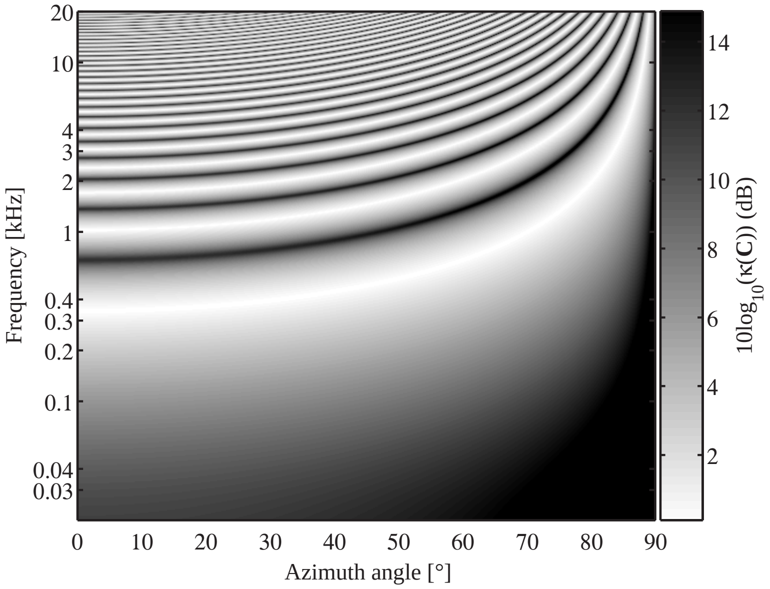


FIGURE 2.22: *Condition number as a function of frequency and azimuth angle ϕ for the* 180° *source span.*

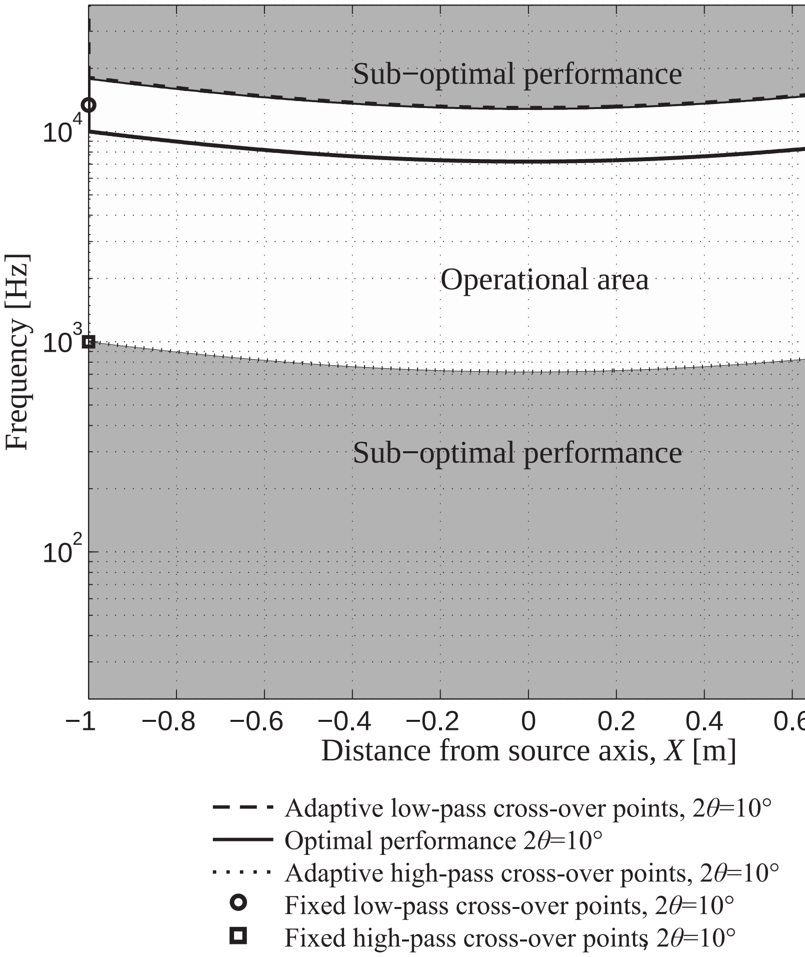


FIGURE 2.23: **Operational area of the SD in the lateral plane**. *The "operational area" (indicated as white area) for bandwidth/lateral displacement together with fixed cross-over frequencies for the SD. The grey area represents regions of sub-optimal performance. The system is based on $n = 1$ and $\nu = 0.9$, which corresponds to the performance of a 2-way OSD system on-axis. The black line represents $n = 1$, the black dotted line represents $n - \nu$ and the black dashed line represents. The upper fixed cross-over frequency is indicated by the black circle and the lower fixed cross-over frequency is indicated by the black square.*

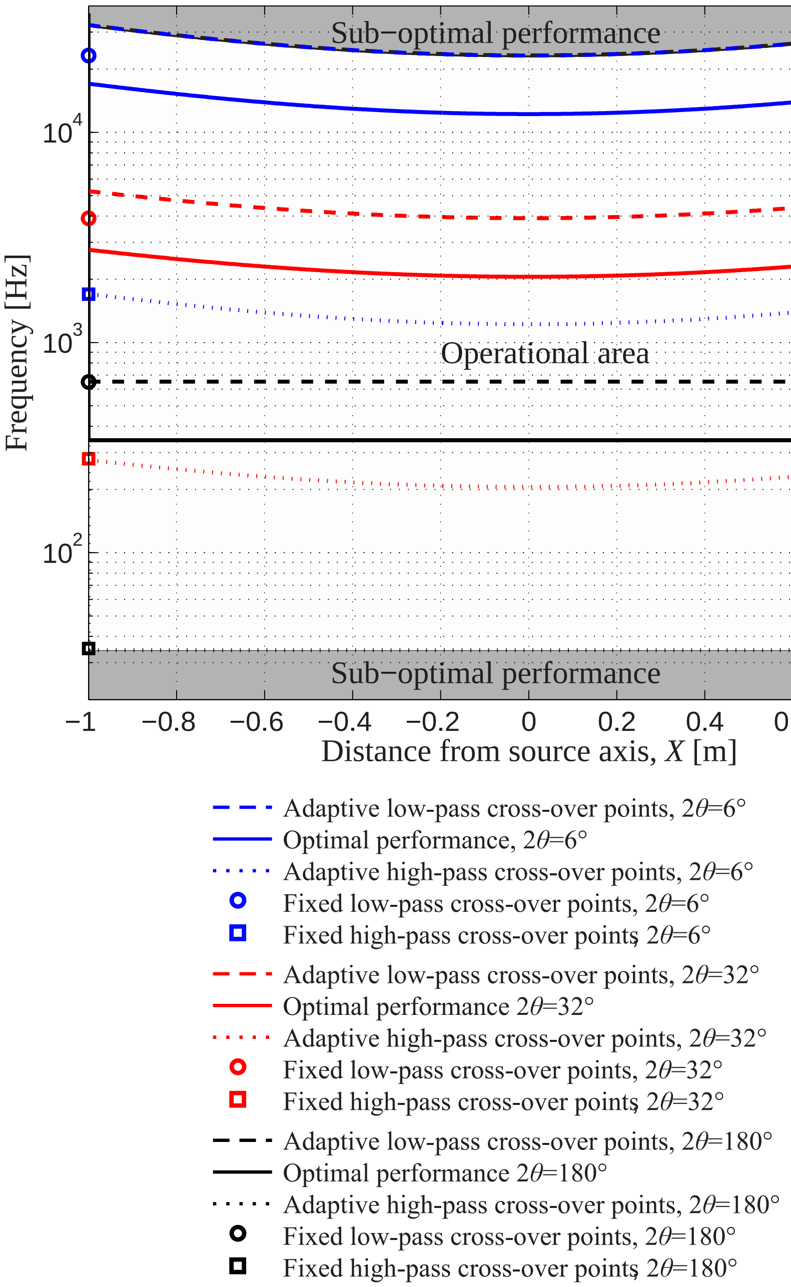


FIGURE 2.24: **Operational area of the 3-way OSD in the lateral plane**. *The "operational area" (indicated as white area) for bandwidth/lateral displacement together with fixed cross-over frequencies for the 3-way OSD are presented. The grey area represents regions of sub-optimal performance. The system is based on $n = 1$ and $v = 0.9$, which corresponds to the performance of a 2-way OSD system on-axis. The blue lines are for the 6.2° source span, the red lines are for the 32° source span, the black lines are for the 180° source span. The solid lines are for $n = 1$, dotted lines are for $n - \nu$ and the dashed lines are for $n + \nu$. The upper fixed cross-over frequencies are indicated by circles and the lower fixed cross-over frequencies are indicated by squares, both in colours that represents their respective source spans.*

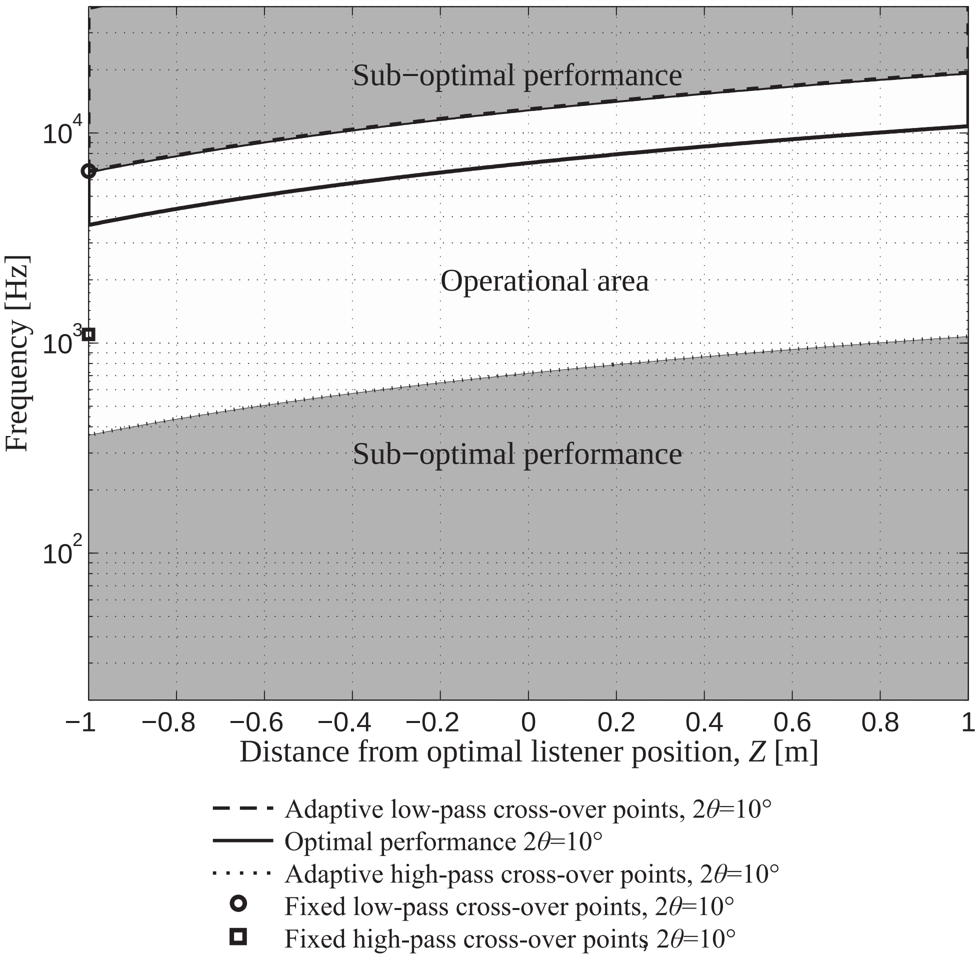


Figure 2.25: **Operational area of the SD in the fore and aft plane**. *The "operational area" (indicated as white area) for bandwidth/fore and aft displacement together with fixed cross-over frequencies for the SD. The grey area represents regions of sub-optimal performance. The system is based on $n = 1$ and $\nu = 0.9$, which corresponds to the performance of a 2-way OSD system on-axis. The black line represents $n = 1$, the black dotted line represents $n - \nu$ and the black dashed line represents. The upper fixed cross-over frequency is indicated by the black circle and the lower fixed cross-over frequency is indicated by the black square.*

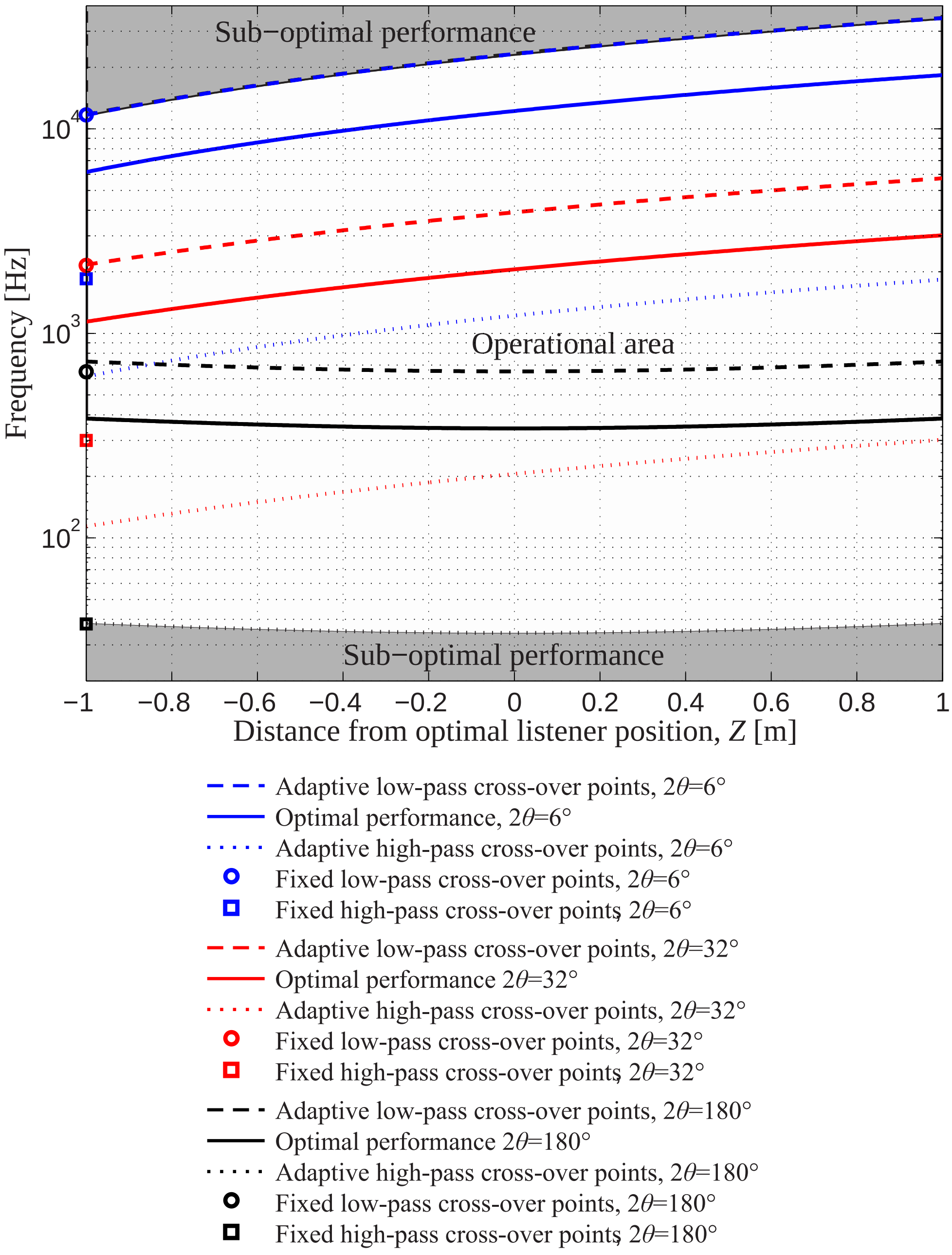


FIGURE 2.26: **Operational area of the 3-way OSD in the fore and aft plane**. *The "operational area" (indicated as white area) for bandwidth/fore and aft displacement together with fixed cross-over frequencies for the 3-way OSD are presented. The grey area represents regions of sub-optimal performance. The system is based on $n = 1$ and $v = 0.9$, which corresponds to the performance of a 2-way OSD system on-axis. The blue lines are for the 6.2° source span, the red lines are for the 32° source span, the black lines are for the 180° source span. The solid lines are for $n = 1$, dotted lines are for $n - \nu$ and the dashed lines are for $n + \nu$. The upper fixed cross-over frequencies are indicated by circles and the lower fixed cross-over frequencies are indicated by squares, both in colours that represents their respective source spans.*

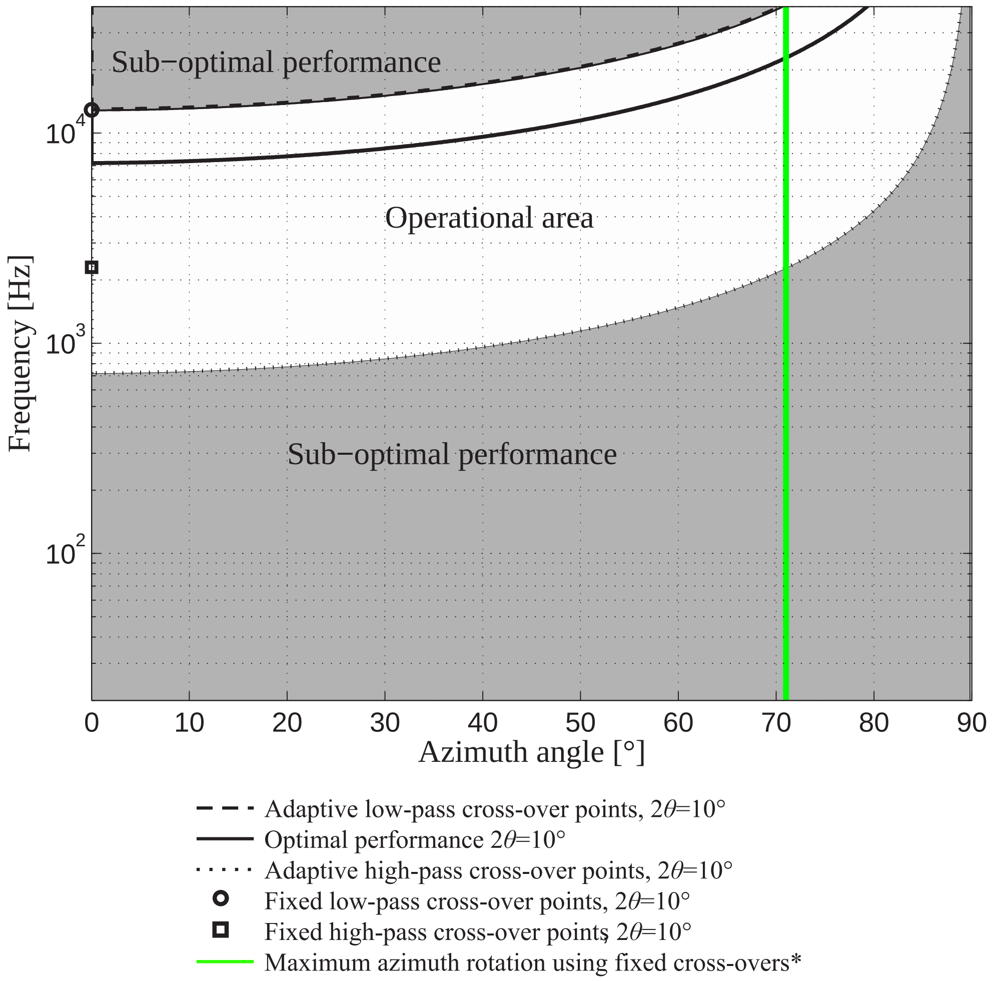


FIGURE 2.27: **Operational area of the SD for azimuth rotation**. *The "operational area" (indicated as white area) for bandwidth/azimuth rotation together with fixed cross-over frequencies for the SD. The grey area represents regions of sub-optimal performance. The system is based on $n = 1$ and $\nu = 0.9$, which corresponds to the performance of a 2-way OSD system on-axis. The black line represents $n = 1$, the black dotted line represents $n - \nu$ and the black dashed line represents. The fixed upper cross-over frequency is indicated by the black circle and the lower fixed cross-over frequency is indicated by the black square. *The thick green line is fitted to give the same maximum azimuth angle (71°) as for the 3-way OSD system when using fixed cross-overs for comparative purposes, see Figure 2.28.*

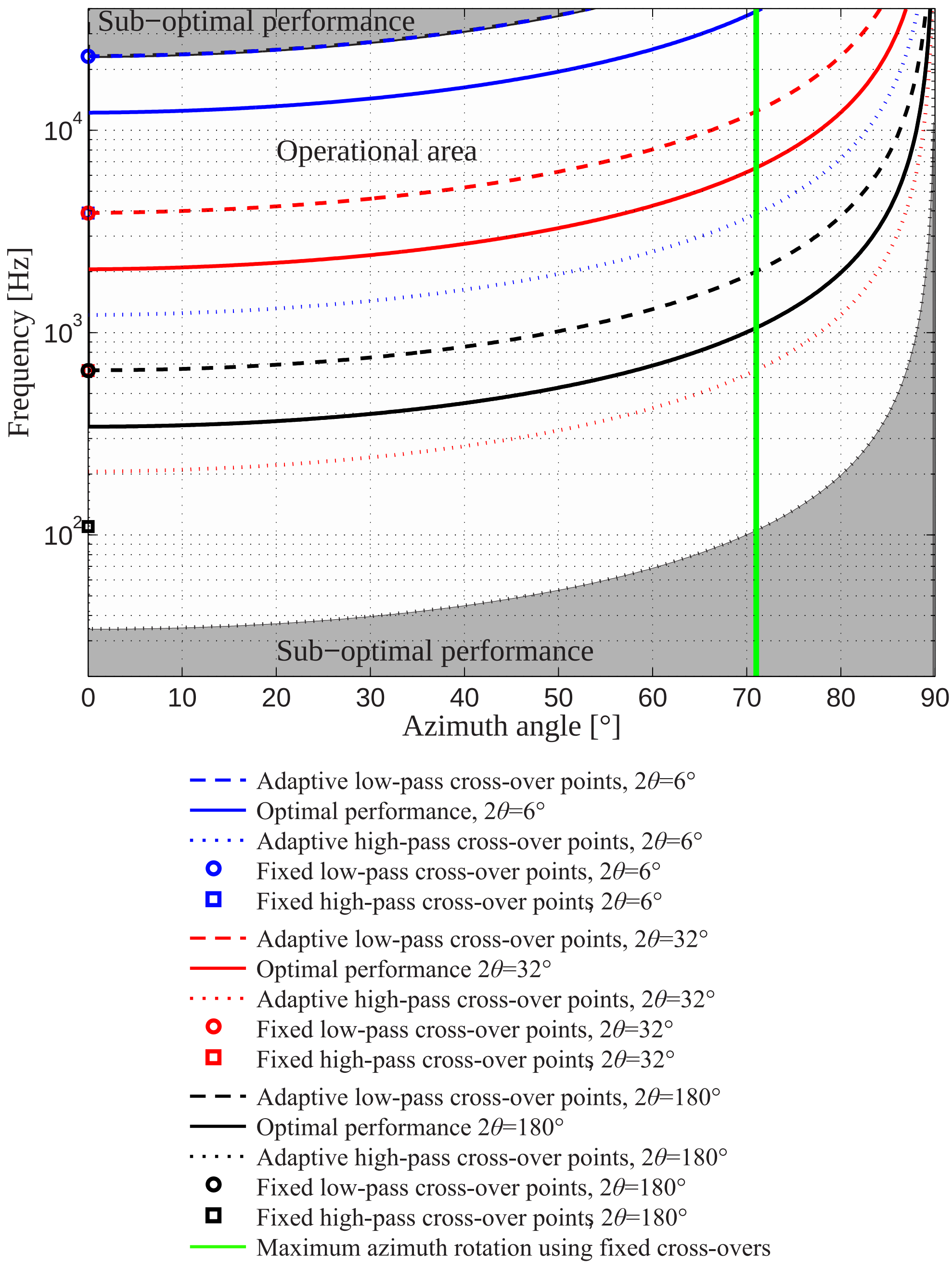


FIGURE 2.28: **Operational area of the 3-way OSD for azimuth rotation**. *The "operational area" (indicated as white area) for bandwidth/azimuth rotation together with fixed cross-over frequencies for the 3-way OSD are presented. The grey area regions represents sub-optimal performance. The system is based on $n = 1$ and $v = 0.9$, which corresponds to the performance of a 2-way OSD system on-axis. The blue lines are for the 6.2° source span, the red lines are for the 32° source span, the black lines are for the 180° source span. The solid lines are for $n = 1$, dotted lines are for $n - \nu$ and the dashed lines are for $n+\nu$. The upper fixed cross-over frequencies are indicated by circles and the lower fixed cross-over frequencies are indicated by squares, both in colours that represents their respective source spans. The thick green line is fitted to give maximum azimuth rotation angle (71°) when using fixed cross-overs.*

Chapter 3

The measurement of a database of head related transfer functions

Knowledge of HRTFs is essential in the synthesis of binaural signals, such as those used in virtual reality applications. The measurement of arguably the most comprehensive database to date of HRTFs of the KEMAR dummy head is presented here. The features of HRTFs are illustrated, such as the non-directional character of the propagation along the ear canal and the HRTF measured without the pinna. The measurement was conducted by the ISVR Samsung Joint Lab. An HRTF describes, for a certain angle of incidence, the sound transmission from a free field to a point in the ear canal of the subject. The subject in this project is a KEMAR dummy head, which is used to model an average human subject. The idea behind the binaural technique can be summarised as follows. Spatial hearing relies on the two signals that comprise the sound pressure at the each of the ear drums. If these are recorded at the ears of the listener and are later reproduced exactly, then the exact audio experience is recreated. The HRTF characterises the transfer function between source and the two ear signals and takes into account the reflections and diffractions from the human torso, head and pinna.

The ISVR-Samsung measurement was performed on a KEMAR dummy head at 72 different azimuth and 14 different angles of elevation giving a total of 1008 directions. The azimuth angle resolution is 5° and the elevation angle resolution is 10°. The sampling frequency used is 48 kHz and the quantization depth is 24 bits. Pink noise was used to obtain the impulse responses. The data was saved as both equalized and raw data, where the length of the equalised HRIRs is 512 samples and the length of the raw HRIRs is 4096 samples. The KEMAR dummy head measurements where performed for a “small pinna”, “large pinna” and “no pinna”. The “small pinna” represents typical ears of American and European females as well as Japanese males and females. The “large pinna” represents typical ears of American and European males. The “no pinna” measurement is useful for modelling of HRTFs. The measurements were taken with both blocked and open ear canals for all the six cases and on both the left and the right ear. The measurement set-up for the ISVR Samsung data where performed in the anechoic chamber of the ISVR at a distance of 2 m between the loudspeakers and the centre of the listeners head. The aim of the measurement was also to provide a database with high sound quality by using 24 bit 48 kHz sampling frequency and high signal to noise ratio (SNR).

Previous measurements have been published in the literature such as those reported by, Moller [69], Gardner [31] at MIT, and Algazi et al [2] at CIPIC. These measurements are briefly described here. The Moller [69] measurement was performed on 40 human subjects for 97 directions of sound incidence, covering the entire sphere. Individual HRTF data for the median, horizontal, and frontal planes are presented in the frequency domain. The measurements were made synchronously at both ears. The measurements were made at the entrance to a blocked ear canal and also at the entrance to an open ear canal (by means of a probe microphone).

The MIT measurements consist of the left and right ear impulse responses from a loudspeaker mounted 1.4 m from the KEMAR dummy head. Maximum length pseudo random binary sequences were used to obtain the impulse responses at a sampling rate of 44.1 kHz. In total 7200 different positions were sampled. The impulse response of the speaker in a free field and several headphones placed on the KEMAR were also measured. The length of the HRIRs is 512 samples in the full database. Left-Right symmetry was assumed for KEMAR and hence all the right-ear responses were obtained using the “small pinna” model and all the left-ear responses with the “large pinna” model.

The CIPIC database measurement was performed on 45 subjects at 25 different azimuths and 50 different elevations (1250 directions) at approximately 5° angular increments. The database also contains anthropometric parameters. The KEMAR dummy head is included in the 45 subjects for whom they have measured HRTFs. The sampling frequency they used is 44.1 kHz and the length of the HRIRs is 200 samples. The KEMAR measurements where performed on a blocked ear canal for the “small pinna” and “large pinna” on both the left and the right ear. The measurement set-up for the CIPIC data where performed in the CIPIC Interface laboratory at a distance of 1 m between the speaker and the centre of the listeners head.

3.1 Measurement arrangement and procedures

The equipment and procedures necessary for the HRTF measurements are presented here. The mechanism of the measurement rig that holds a set of 14 loudspeakers and rotates the KEMAR in the azimuth direction is described in depth.

Elevation(degrees)	-40	-30	-20	-10	0	10	20
Distance(mm)	2048	2056	2051	2047	2039	2035	2032
Elevation(degrees)	30	40	50	60	70	80	90
Distance(mm)	2033	2033	2034	2045	2055	2068	2081

Table 3.1: **Loudspeaker distances**. *Distance from the loudspeaker frontal surface to the centre of measurement rig. Mean distance: 2047 mm, Standard deviation: 14 mm.*

3.1.1 Specifications of the measurement

3.1.1.1 Rig

A new rotating rig was built and installed in the ISVR anechoic chamber. The chamber measures $9.15 \times 9.15 \times 7.32$ and all the equipment was placed on the meshed metal floor. The inner radius of the new half-circular metal frame was designed to be 2 m from the centre to the frontal surface of loudspeakers, while the actual size of the circular frame was greater in order to accommodate the width of the loudspeakers. Fourteen loudspeakers were arranged on the half circle from $-40°$ below the horizontal plane with respect to the centre up to 90° directly overhead at every 10°. (See Figure 3.2 for a photo). This circular frame was supported from above by a crane so that the two ends could be vertically aligned and made immovable by tying it to the side walls. Instead of spinning the heavy circular rig as practiced in the preliminary measurements, the KEMAR was rotated through 360°, while maintaining the rig in position. Compared to the rotating mechanism of the old rig, the current system has been proved to be more reliable, greatly reducing the distortion of the frame, whilst it allowed much safer operation of the rig. Despite such improvements, the half circle was observed to be slightly deformed downwards due to the considerable weight of the rig and the loudspeakers, which could not be fully restored to the design specification despite repeated adjustments using the crane and ropes.

For the purpose of the necessary corrections, Table 3.1 gives the distances measured from each loudspeaker to the centre of the KEMAR head. As listed in this table, the range of error is less than 1% about the mean of 2047 mm. It was regarded as reasonable to carry out the measurements assuming the loudspeakers are equidistant from the centre. The following angle measurements (Table 3.1) also showed that the individual loudspeakers were 10° apart from each other with minor errors, while the loudspeaker at 0° elevation was horizontally aligned with the centre of head.

The KEMAR mounted on the rotating shaft was aligned in such a way that the centre of the head (the crossing of inter-aural axis and the median plane) coincided with the centre of the circular frame. Since the centre of the KEMAR dummy head was also carefully aligned with the rotating axis, there was little displacement of the centre point during and after rotations.

Measurement	Pinna	Open or blocked
1	Small	Open
2	Large	Open
3	No	Open
4	Small	Blocked
5	Large	Blocked
6	No	Blocked

Table 3.2: **The six HRTF database measurement cases**.

Three arrangements of pinna model ("large pinna", "small pinna" and "no pinna") were combined with the open and blocked ear canal cases to give a total of 6 sets of measurements as required, these are listed in Table 3.2. The right "large pinna" is of model DB65 and the left "large pinna" is of model DB66. The right "small pinna" is of model DB60 and the left "small pinna" is of model DB61. For each type of pinna and ear canals, see Figure 3.3. This is in contrast to some of the previous work where two different external ears were used in a single set of measurements to save time. In these measurements an identical pinna model has been applied to both ears in each case. It is noteworthy that, in case of blocked ear canal, the ear on each side was measured separately due to the length of the microphone preamplifier that made it possible to fit two preamplifiers at the same time.

3.1.1.2 Signal processing tool - sampling rate and quantization depth

The signal processing system used throughout the current HRTF measurement was the Huron digital audio convolution workstation (hardware version 2.0, software version 3.0 manufactured by Lake Technology Ltd.) controllable through Matlab (version 5.3). As required, the Huron was capable of sampling signals at 48 kHz with a 24 bit quantization depth, which was applied to all the measurements.

3.1.1.3 Stimulus

The input signal used throughout the HRTF measurements was pink noise of 32k points equivalent to a duration of 667 ms. The pink noise was internally generated by the Huron measurement system and controlled by Matlab.

3.1.1.4 Averaging

For each HRTF, the responses were averaged ten times to reduce the variability of data. Consequently, the total time of measurement for a position was approximately 10 seconds including processing time.

3.1.2 Rig operation

3.1.2.1 Motorised turntable (azimuth)

As briefly described in Section 3.1.1.1, the half-circular metal frame was fixed and the KEMAR rotated about the vertical axis to which the two ends of the frame were aligned. This movement in azimuth was enabled by a motorized turntable engaged with a step motor by a chain, and the motor was operated by Matlab through a microcontroller.

3.1.2.2 Switch box (elevation)

Once the KEMAR was placed at a certain azimuth angle, HRTFs at each elevation were measured by switching the input signal to the loudspeaker corresponding to each elevation angle, instead of actually moving a loudspeaker along the circumference. The switch box that alternated the connections between the power amplifier and all the 14 loudspeakers was controlled, again, by Matlab through the microcontroller.

3.1.3 Measurement equipment

3.1.3.1 Microphones

There were two kinds of microphones used throughout this measurement. One is B & K type 4134, and the other type 4136. The former was mounted on the Zwislocki coupler and used for the three cases of open ear canal, and the latter was used for the other three cases of blocked ear canal and the free field measurement. This is because the type 4134 of 0.5 inch diameter was too big to be fitted into the rubber pinna (see Figure 3.3). Two 4134- and one 4136-microphones were calibrated using B & K 4230 calibrator, and two measuring amplifiers used for each of the left and right channels were tuned to indicate identical sound level. The sinusoidal signals captured during the calibrations were also recorded for further corrections. The gain difference between two 4134 microphones was about 0.05 dB, which is discountable on the dB scale, whilst the 4136 microphone appeared about 17 dB less sensitive than one of the 4134 microphones. These gain differences were fully taken into account during the data processing later on.

3.1.3.2 Loudspeakers

The 14 loudspeakers (which were 4 inch paper cone drivers manufactured by Fostex) used in the measurement were mounted in cabinets manufactured in the ISVR departmental workshop. There has been a concern about the size of these speakers that they might have higher cut-on frequencies, but the calibrations showed that they are reasonably reliable at low (down to 100 Hz) as well as at high frequencies. This can be regarded as of less importance since, for the final HRTF database, the responses from the loudspeakers to the microphone on/in the ears will be deconvolved with the free field response of the microphone and each loudspeaker. The characteristic response of the 14 loudspeakers was reasonably identical up to 10-15 kHz and with a deviation of less than 2 dB.

3.1.3.3 Measuring amplifier and power amplifier

Two B & K 2636 measuring amplifiers have been used for each of the left and right channels throughout the measurement, and the left channel of a YAMAHA H5000 power amplifier has driven the loudspeakers. Input/output gains were carefully chosen to avoid any clipping errors while maintaining acceptable signal levels.

3.1.3.4 SNR estimates

In advance of the full processing of the acquired data, a few responses were randomly chosen to investigate the SNR. This was evaluated by looking at the difference between one of the typical measurements and the measurement performed with all settings identical, but with the loudspeaker replaced by an 8-ohm resistor (Moller [69]). The SNR measured for the left ear of the KEMAR at ($\phi = 0°$, $\psi = 0°$) is in the range of 50-70 dB except for the region of "pinna notch". This SNR is found to be comparable to the previous HRTF measurements introduced in the beginning of the chapter.

3.1.4 Measurement Procedure

All the measurement procedures and processes were controlled via Matlab commands that were written in a set of script files. Once the measurement equipment was calibrated and ready in the closed anechoic chamber, the rotating turntable and the switch box were initialized so that the KEMAR faced straight towards the array of loudspeakers ($\phi = 0°$), and the pink noise could be played from the loudspeaker at the lowest elevation ($-40°$ elevation).

On completion of Huron recording the microphone output for the first angle, the pink noise signal was switched to the next loudspeaker (elevation), and, after all the 14 elevations were measured, the turntable rotated clockwise for the KEMAR to be positioned at 355° azimuth. In this way, 72 azimuth angles from 0° to 5° in $-5°$ decrements (0°, 355°, 350°,, 10°, 5°) were measured for about 3 hours duration for each of the pinna models, and the recorded data were saved in the Matlab data format of a 'mat' file. Finally, the free field transfer functions were measured for each loudspeaker (elevation) with the 4136 microphone held normal to the circular frame at its centre. These transfer functions will be used in the deconvolution process in order to obtain the response functions resulting only from the existence of KEMAR.

3.2 Data processing

The data processing of the measurement conducted in July 2004 at the ISVR is presented here. Section 3.2.1 presents the coordinate system for how the data is presented. Section 3.2.2 describes the equalization procedure used with the data and also how the data was processed and stored. Section 3.2.3 discusses data processing operations and filter design issues.

3.2.1 Coordinate system

More than one convention exists in the literature as regards to the coordinate system used for the description of the existing HRTF databases. In this thesis we use the same convention as that in the MIT database, in which the azimuth angles range from 0° to 355° and the elevation angles from $-90°$ to 90° degrees. In this sense: ($\phi = 0°$, $\psi = 0°$) corresponds to the point directly ahead, ($\phi = 90°$, $\psi = 0°$) corresponds to the point directly to the right, ($\phi = 270°$, $\psi = 0°$) corresponds to the point directly to the left, ($\phi = 0°$, $\psi = 90°$) corresponds to the point directly on top and ($\phi = 180°$, $\phi = 0°$) corresponds to the point directly behind. The coordinate system is illustrated in Figure 3.4. The figure shows the coordinate systems for the horizontal plane, the median plane and the frontal plane. The horizontal plane goes clockwise from 0° to 355° in 5° steps, which results in 72 azimuth angles. The median plane goes anti clockwise from $-40°$ to 90° ($\phi = 0°$) in 10° steps and from 90° to $-40°$ ($\phi = 180°$), which results in 36 elevation angles. The frontal plane goes anti clockwise from $-40°$ to 90° ($\phi = 90°$) in 10° steps and from $-40°$ to 90° ($\phi = 270°$), which results in 36 elevation angles.

3.2.2 Equalization procedure

The responses from the loudspeakers that were mounted on the measurement rig were all measured and in the data processing procedure the loudspeaker responses were deconvolved from the measured HRTFs. The responses from the loudspeakers were measured with a B & K type 4136 microphone, as described in section 3.1.3. The equalization was undertaken by dividing the HRTF at each angle with the free field transfer function at the respective angle (obtained by taking an FFT of the free field response). An inverse FFT was then applied in order to get back to the time domain. Finally an FFT shift was applied, so that the equalised HRIR could be obtained.

The measurement on the blocked ear canal was made with the same microphone as the free field measurement, and hence the equalization procedure takes account for the response of the measurement system in the blocked ear canal data. The equalization of the open ear canal does not take into account the exact response of the system, since another type of microphone was used, namely the B & K type 4134. However the frequency response of the microphones is similar and the equalization of the open ear canal with free field measurement is here considered as a good approximation.

The equalised HRTF data was saved in $72 \times 14 \times 2 \times 512$ multi-dimensional arrays. The equalised data was processed from the raw data arrays ($72 \times 14 \times 2 \times 4096$). The raw data holds the 4096 first points of the entire measurement series. Firstly the raw data was windowed to 256 points with a rectangular window (points 301 - 556), in order to capture the most essential information in the impulse response. The same windowing (points 301 - 556) was applied to the free field impulse response. The two truncated impulse responses are Fourier transformed with a 2048 point FFT and their ratio computed. This ratio was transformed back to the time-domain with an inverse FFT and a circular shift of the result which was then multiplied with a 2048 point Hanning window. 512 points (974 -1486) of the obtained impulse response was kept.

3.2.3 Discussion

3.2.3.1 Specification of the data

According to the project specifications the acquired data is quantised to 24 bits and sampled at 48 kHz, which is a high definition audio format. Depending on the specific application in which it will be used, the database can be appropriately modified. A brief list of potential modifications follows:

The size of the equalized data is 512 points and it may be desirable to change the size of the data to, say, 256, 128, 64 or 32 points. The desired size of the impulse response is application dependent, where for example an MP3 player may have very limited processing power and can only handle short impulse responses. The following references Jot [44], [43], Algazi [3], Duda [27] and Blommer [14] give further details on efficient implementations of HRTFs. A comparison of digital filter design methods is presented by Huopaniemi [39]. Filter design issues are also discussed in Chapter 5.

It is possible that the database has to be used for applications designed for different sample rates. Data resampling (Proakis [84]) can be practically implemented with the standard Matlab functions (e.g. resample.m Matlab [63]). The degree to which such processing (especially down sampling) degrades the audio quality of the specific application should be verified by appropriate simulations and listening tests. If used for applications designed in shorter word-lengths (e.g. 16-bit) the appropriate use of dithering should be considered.

As is typical with measured HRTF databases, the responses in the ISVR Samsung HRTF database are not of minimum phase. In case of a real-time filter application with HRTFs, the system designer needs to take account of the non-minimum phase characteristics. This will be particularly relevant in the case of designing inverse filters for cross-talk cancellation as described in Chapter 2.

3.2.3.2 Equalization comparison: MIT and ISVR

The equalization procedure in the ISVR database is similar to that used by MIT. The main difference between the ISVR and the MIT database is that the MIT database is compensated with the minimum phase representation of the free field impulse response while the ISVR databse was compensated with the mixed-phase free field impulse response of the loudspeakers. The MIT equalization procedure by Gardner [31] was undertaken in the following manner: Firstly the first 512 points of the raw response in the time domain were taken and padded with zeros to 8192 points. The response was normalised to ±1 and then the 8192 point FFT was performed and the magnitude and phase were saved as two separate vectors. All magnitudes smaller than -20 dB were clipped to -20 dB. The response was inverted, and above 15 kHz, the magnitude of the inverted response was set to the magnitude at 15 kHz. The inverse FFT was performed on this inverted response and the result was shifted so that the maximum value occurred at index 4096. The result was multiplied by a 2048 point raised cosine window (Hanning window) centred at index 4096 and saved in 16-bit format. The rceps (or real cepstrum) command in MATLAB was used to obtain a minimum phase signal. This was truncated to 2048 points. The compact MIT database has 128 point impulse response. Each 128 point response was obtained by convolving the appropriate 512 point impulse responses with the minimum phase inverse filter described here. Finally the resulting impulse responses were cropped by retaining 128 samples.

3.3 Measurement data

A large amount of data was collected in this measurement and a summary of the data is presented in this section. Section 3.3.1 presents the validation of the measurements, which shows time domain responses for different azimuth angles and compares the ISVR measurements with measurements from the MIT and the CIPIC databases. Section 3.3.2 discusses the asymmetry in the KEMAR measurements. Section 3.3.3 shows basic features of HRTFs and HRIRs, such as: ITDs and ILDs (horizontal, median and frontal plane), “small pinna” and “large pinna” comparison, “no pinna” comparison and the non-directional character of the propagation along the ear canal. The results that are presented in this section are all with deconvolved loudspeaker responses.

3.3.1 Validation of results

The measurement procedure is validated by identifying a number of basic characteristics. The time responses for the left and the right ear for three different azimuth angles are presented below in Figure 3.5. The data from the ISVR measurements are compared to the measurements from MIT and CIPIC.

Figure 3.5 shows an example of the HRIR for the two ears of the KEMAR dummy head. Figure 3.5 (a) shows sound that comes directly in front of the listener in the horizontal plane ($\phi = 0°$, $\psi = 0°$). As would be expected, the sound level at both of the ears is approximately the same and so is the arrival time. Figure 3.5 (b) shows sound that comes from the left side in the horizontal plane ($\phi = 270°$, $\psi = 0°$). The attenuation due to the head shadowing effect is clearly seen at both of the two ears. It is also observed that the signal at the left ear arrives slightly before zero, which is due to the fact that the left ear is closer to the sound source than the middle of the head (the reference point is set to be in the middle of the head where the free-field measurement is made). The right ear signal arrives later than zero, and can be measured to be approximately 0.7 ms after the signal arrived at the left ear (0.7 ms corresponds to approximately half the perimeter of the KEMAR dummy head). Figure 3.5 (c) shows sound that comes from the right side in the horizontal plane ($\phi = 90°$, $\psi = 0°$). The sound level difference at both and time arrival difference of the ears is clearly seen as in the previous case (middle graph).

Figure 3.6 shows an example of the HRTF for the two ears of the KEMAR dummy head. Figure 3.6 (a) shows sound that comes directly from in front of the listener in the horizontal plane ($\phi = 0°$, $\psi = 0°$). As would be expected, the frequency response at both of the ears is approximately the same, however an asymmetry can also be seen, which will be discussed in Section 3.3.2. Figure 3.6 (b) shows sound that comes from the left side in the horizontal plane ($\phi = 270°$, $\psi = 0°$). The frequency response difference at both of the ears is clearly seen. It is observed that the response at the left ear is higher than at the right ear, due to the shadowing effect of the head. The response at the left ear has a gain above 0 dB at most frequencies, which indicates a pressure build-up due to reflections from the head and the pinna. The "pinna notch" is also clearly visible at approximately 9 kHz for the left ear. The gain of the HRTF at 0 Hz should be 0 dB in theory, where the human body does not disturb the sound field. However, still at 100 Hz a gain of approximately 1 dB at the left ear can be observed. This can partly be explained by the fact that the left ear is closer to the sound source than the middle of the ear. Figure 3.6 (c) shows sound that comes from the right side in the horizontal plane ($\phi = 90°$, $\psi = 0°$). The frequency response difference at both of the ears is clearly seen as in the previous case.

3.3.1.1 Comparison with MIT data

The MIT measurements were performed on an open ear canal for the left “small pinna” and the right “large pinna”. The measurement set-up for the MIT data where performed in MIT’s anechoic chamber at a distance of 1.4 m between the speaker and the centre of the KEMAR dummy head. In the MIT measurements, symmetry over the median plane was assumed for KEMAR and that measurements were taken only for the “small pinna” on the left and the “large pinna” on the right. Figure 3.7 shows the open ear canal measurement of the left and right “small pinna” and “large pinna” from MIT and ISVR. The DC component of the ISVR measurement is set to be the reference and the MIT curve is fitted to the DC component of the reference. The MIT curve is shifted up to fit the ISVR curve for ease of comparison. The results from the “small pinna” measurements agree relatively well, there is a level difference of approximately 6 dB at frequencies below 2 kHz. The results from the “large pinna” measurements also agree relatively well, though there is a level difference of approximately 20 dB at frequencies below 5 kHz. The level difference may be due to an error in the MIT database, which is confirmed by the fact that the ISVR measurements match very well with the CIPIC measurements, as discussed in Section 3.3.1.2.

3.3.1.2 Comparison with CIPIC data

The CIPIC measurements where performed on a blocked ear canal for the “small pinna” and “large pinna” on both the left and the right ear. The measurement set-up for the CIPIC data where performed in the CIPIC Interface laboratory at a distance of 1 m between the speaker and the centre of the listeners head. Figure 3.8 shows the blocked ear canal measurement of the left and right “small pinna” from CIPIC and ISVR. The DC component of the ISVR measurement is set to be the reference and the CIPIC curve is fitted to the DC component of the reference (The CIPIC curve is shifted up to fit the ISVR curve for a better comparison). The results agree very well for both the left and the right responses.

3.3.2 Asymmetry in KEMAR

There is asymmetry present between the left and right ear frequency responses. The asymmetry can be seen in Figure 3.9, Figure 3.10 and Figure 3.11. This asymmetry is present in the CIPIC measurements but not in the MIT measurements. The reason for this is that the measurements at MIT were taken only for the “small pinna” on the left and the “large pinna” on the right, hence the asymmetry can not be observed.

It is seen in Figure 3.9, Figure 3.10 and Figure 3.11 that the left “small pinna” deviates from the rest of the measurements, while the “large pinna” and “no pinna” measurements look symmetric. We can thus infer that the asymmetry is directly related to the "small pinna" model and not to the overall shape of KEMAR mannequin. The reason for the deviation can be due to the manufacturing process of the pinna or if it is introduced on purpose in order to simulate the asymmetry that the human will show. It is seen that the blocked ear canal measurements do not include the distinctive broad peak at about 2-3 kHz due to the ear canal resonance. The expected absence of the most prominent pinna-related effect, namely the "pinna notch" at about 9 kHz, can be seen in the "no pinna" measurements above.

3.3.3 Basic features

3.3.3.1 ITDs and ILDs in the horizontal plane

An 2D image representation of the measured HRIR in the horizontal plane is illustrated in Figure 3.12 - Figure 3.17. The images illustrate the impulse response of the right ear as a function of azimuth angle and time, where the strength of the response is represented by brightness. It is seen that the sound is strongest and arrives soonest when it is coming from the right side ($\phi = 90°$) and similarly it is weakest and arrives latest when it is coming from the left side ($\phi = 270°$). The difference between the shortest and longest arrival time is about 0.7 ms.

It is possible to see a couple of features in the images, such as the initial sequence of fluctuations due to pinna reflections and the larger peak that arrives 0.4 ms after the initial peak, which is due to a shoulder reflection. The response when the source is in front is similar to the response when the source is at the back. In the open ear canal measurements the ear canal resonance is present. It is possible to see that the period of the ear canal resonance is approximately 0.5 ms, which corresponds to 2 kHz. The open "no pinna" measurement in Figure 3.14 shows the absence of the pinna reflections, while the ear canal resonance is still present. The blocked ear canal measurements give similar results for the pinna reflections and show the absence of the ear canal resonance. The blocked "no pinna" measurement in Figure 3.17 shows the absence of both pinna reflection and ear canal resonance. The feature that can be seen in Figure 3.17 is probably due to a shoulder reflection where the delay of the shoulder reflection changes with azimuth angle.

The images in Figure 3.18 - Figure 3.23 show the frequency response for the right ear of the KEMAR. The frequency response is shown for all measured azimuths in the horizontal plane. As expected, the response is greatest when the source is at 90° and directed into to the right ear. The response is weakest when the source is at 270° on the opposite side of the head.

As seen in the HRIR case in Figure 3.12 - Figure 3.17, front/back response at 0°/180° are quite similar. The peak around 2-3 kHz is due to the ear canal resonance and the notch around 9 kHz is the “pinna notch” whose frequency mainly changes with elevation. The frequency of the “pinna notch” is fairly constant in the horizontal plane. The “no pinna” measurements in the horizontal plane show the absence of the “pinna notch”.

3.3.3.2 ITDs and ILDs in the median plane

The image representation of the measured HRIR in the median plane is showed in Figure 3.24 - Figure 3.29. The image shows the impulse response of the right ear as a function of elevation angle and time, where the strength of the response is represented by brightness. It is seen that the arrival time is approximately the same for all elevations as expected. The difference in arrival time that can be seen is due to the fact that the radius of the rig was not perfect. Hence the distances between the loudspeaker and the KEMAR were not equal for all the elevation angles, see Table 3.1. The main feature that can be seen is that the arrival times and strength of the pinna reflections changes with elevation angle. The pinna is a more effective reflector for sounds coming from the front than for sounds from above, which results in the notch being much more pronounced for sources in front than for sources above. Also, the path length difference changes with elevation angle, so the frequency of the “pinna notch” moves with elevation. As in the horizontal plane HRIR data, for the open ear canal measurements the ear canal resonance is present. One can see that the period of the ear canal resonance is approximately 0.5 ms (2 kHz).

The shoulder reflection is also present which is seen in 3.24 - Figure 3.29 as a “bow” bending to the right, with a radius of approximately 1.2 ms. At elevation 90° the delay of the shoulder refection is approximately 1.2 ms, which corresponds to a distance of some 41 cm. The variation of the delay of the shoulder reflection can be explained by the fact that when the source is 90° above the head, the distance for sound to travel is approximately twice the distance compared to when the source is at 0°. When the source is at -40° below the head, the distance for sound to travel is shorter than when the source is at 0°, hence the shoulder reflection arrives earlier at the ear. It can be seen that the ear canal resonance is present in the shoulder reflection. The blocked “no pinna” measurement shows the absence of pinna reflections and ear canal resonance, but the shoulder reflection is still visible.

The image plots in Figure 3.30 - Figure 3.35 show the frequency response for the right ear of the KEMAR as a function of elevation. The frequency response is illustrated for all measured elevations in the median plane. The ear canal resonance around 2-3 kHz is clearly visible and it is also seen that the ear canal resonance is constant for all elevations. The frequency of the “pinna notch” changes significantly with elevation and it goes from approximately 6 kHz at low elevations up to 10 kHz when the source moves over the head. When the source is directly above the head, the “pinna notch” is very weak and the frequency response is quite flat. Similar behaviour is seen for elevations behind the head as for elevations in front of the head. The “no pinna” measurements in the median plane show the absence of the “pinna notch”, as was seen in the horizontal plane.

3.3.3.3 ITDs and ILDs in the frontal plane

The image representation of the measured HRIR in the frontal plane is showed in Figure 3.36 - Figure 3.41. The images show the impulse response of the right ear as a function of elevation angle and time, where the strength of the response is represented by brightness. It is seen that the sound is strongest and arrives soonest when it is coming from the right side ($\phi = 90°$, $\psi = 0°$) and similarly it is weakest and arrives latest when it is coming from the left side ($\phi = 270°$, $\psi = 0°$).

The path length difference between the source and the ear changes with elevation angle, which can be seen in the difference in arrival time in Figures 3.36- 3.41. As in the horizontal plane and median plane HRIR data, for the open ear canal measurements the ear canal resonance is present. Again one can see that the period of the ear canal resonance is approximately 0.5 ms (2 kHz). A shoulder reflection is also present in the frontal plane, which is seen in as a straight line that leans with an angle of $-45°$. The shoulder reflection behaves in a similar way as in the median plane, with the difference that, when the head is shadowing the source the shoulder reflection is too weak to be seen. The blocked “no pinna” measurement in the frontal plane shows the absence of pinna reflections and ear canal resonance, but the shoulder reflection is still visible.

3.3.4 Comparison between pinnae

3.3.4.1 Small pinna - large pinna comparison

The measurements from the “small pinna“ and the “large pinna” are being compared, both in the frequency domain and in the time domain. In Figure 3.42 and Figure 3.43 the open ear canal measurements for small pinna and large pinna are compared. It is seen that the frequency response is very similar up to approximately 7 kHz and that it deviates above 7 kHz. The differentiation of the two different pinna models becomes significant above 7 kHz for the left side and above the 10 kHz for the right side. That is, as expected, the pinna is only associated with high frequency features in the HRTF. Again, the left small pinna (dashed line in Figure 3.43) shows individual characteristics compared with all other cases.

3.3.4.2 Small pinna - no pinna comparison

In this subsection the measurements from the “small pinna” and the “no pinna” are compared with open and blocked ear canal, both in the frequency domain and in the time domain. In Figure 3.44 the open ear canal measurement for small pinna is compared to the no pinna measurement. It is seen that the frequency response is very similar up to approximately 1 kHz and that it deviates significantly above that frequency. It is also seen that the pinna results in an amplification of the sound level in the region from 1 kHz to 8 kHz. The no pinna measurement shows the absence of the "pinna notch". In the time domain it is seen that the impulse response without the pinna is shorter than with the small pinna and thus the secondary wave of reflections from 0.2 ms to 0.6 ms should be related to the presence of the pinna.

In Figure 3.45 the blocked ear canal measurement for the "small pinna" is compared to the "no pinna" measurement. It is seen that the frequency response is very similar up to approximately 1 kHz, as in the open ear canal comparison in Figure 3.44, and that it deviates significantly above that frequency. It is also seen that the small pinna results in an amplification of the sound level. The no pinna measurement shows again the absence of the "pinna notch". In the time domain it is seen that the impulse response without pinna is shorter than with the small pinna and that the blocked ear canal also results in a shorter impulse response.

3.3.5 Non-directional character of the propagation along the ear canal

In this section, the non-directional character of sound propagation along the ear canal is verified. The results are presented for the "large pinna", "small pinna" and "no pinna" measurements. The results are illustrated for two elevations (0° and 30°) and for all azimuth angles.

The pressure division for the "large pinna" is presented in Figure 3.46 and Figure 3.47. The definition of the pressure division for one azimuth angle, is given by P_2/P_1 where P_2 is the sound pressure at the input of the ear canal (the open ear canal measurement) and P_1 is the sound pressure at the entrance to the blocked ear canal at the same frequency (the blocked ear canal measurement).

The result at 0° elevation in Figure 3.46 shows that the pressure division deviation is approximately 2 dB up to 6 kHz. The results at 30° elevation in Figure 3.47 show that the pressure division deviation is approximately 2 dB up to 9 kHz. This indicates that the non-directional character of the propagation along the ear canal goes up to a higher frequency at higher elevation angles.

The pressure division for the "small pinna" case is presented in Figure 3.48 and Figure 3.49. The result at 0° elevation in Figure 3.48 shows that the pressure division deviation is approximately 2 dB up to 8 kHz. The results at 30° elevation in Figure 3.49 show that the pressure division deviation is approximately 2 dB up to 10 kHz. This indicates that the non-directional character of the propagation along the ear canal goes up to a higher frequency at higher elevation angles and also that the small pinna results in a larger bandwidth than the large pinna.

The pressure division for the “no pinna” case is presented in Figure 3.50 and Figure 3.51. The result at 0° elevation in Figure 3.50 shows that the pressure division deviation is approximately 2 dB up to 12 kHz. The results at 30° elevation in Figure 3.51 show that the pressure division deviation is approximately 2 dB up to 12 kHz. This indicates that the non-directional character of the propagation along the ear canal is not that dependent on elevation angle as for the “large pinna” “small pinna” cases. The no pinna pressure divisions results in larger bandwidth than the large pinna and small pinna pressure divisions. Note that the deviations that are seen are probably due to low signal to noise ratio at azimuths where the ear is shadowed by the head.

3.4 Conclusion

The HRTF database measurement was conducted as part of the ISVR Samsung Joint Lab programme and took place in the anechoic chamber at the ISVR. The basic specifications of the measurement were as follows: a spatial resolution of 10° in elevation angles (ranging from $-40°$ to 90°) and uniform spatial resolution of 5° in azimuth angles, the distance from the measurement sources to the centre of the head was 2 m, the sampling frequency was 48 kHz and the quantisation depth was 24-bits. Three arrangements of pinna model (“large pinna”, “small pinna” and “no pinna”) were combined with the open and blocked ear canal cases to give a total of six full database measurements. The following results are presented: comparison of ISVR data with CIPIC and MIT data, asymmetry in dummy head (KEMAR), ITDs and ILDs in the horizontal, median and frontal plane and pressure divisions. The comparison of the data with the CIPIC and MIT database show good agreement. The CIPIC results are very close to the ISVR results and the MIT results also show good agreement but with some level differences, which are probably due to the different measurement arrangements that were used. The asymmetry in the dummy head is shown to be caused by differences between the left and the right small pinna. The non-directional character of the propagation along the ear canal has been illustrated. The results indicates that the non-directional character of the propagation along the ear canal goes up to a higher frequency at higher elevation angles. The no pinna is not as dependent on elevation angle as for the large pinna and the small pinna. Hence, the no pinna pressure division results in larger bandwidth than the pinnae pressure divisions.

The rationale behind the “no pinna” measurements is threefold. Firstly, it will enable the clear identification of the effect of the pinna on the HRTF. Secondly, it is anticipated that this data would facilitate the development of a database in which the pinna component itself can be modelled numerically. Thirdly, it is possible that “no pinna” models may be useful in the design of virtual sound imaging systems. This use, however, may vary between loudspeaker and headphone based systems. To conclude, the outcome of this measurement have resulted in the most extensive HRTF KEMAR database yet available. The data that has been collected show good agreement with other databases, such as the CIPIC and the MIT database. The combination of the six different measurement cases aims to be a starting point for future modelling of pinna responses. The database has been used for the objective evaluation of interpolation techniques in the following chapter and for the simulations of cross-talk cancellation effectiveness in Chapter 5

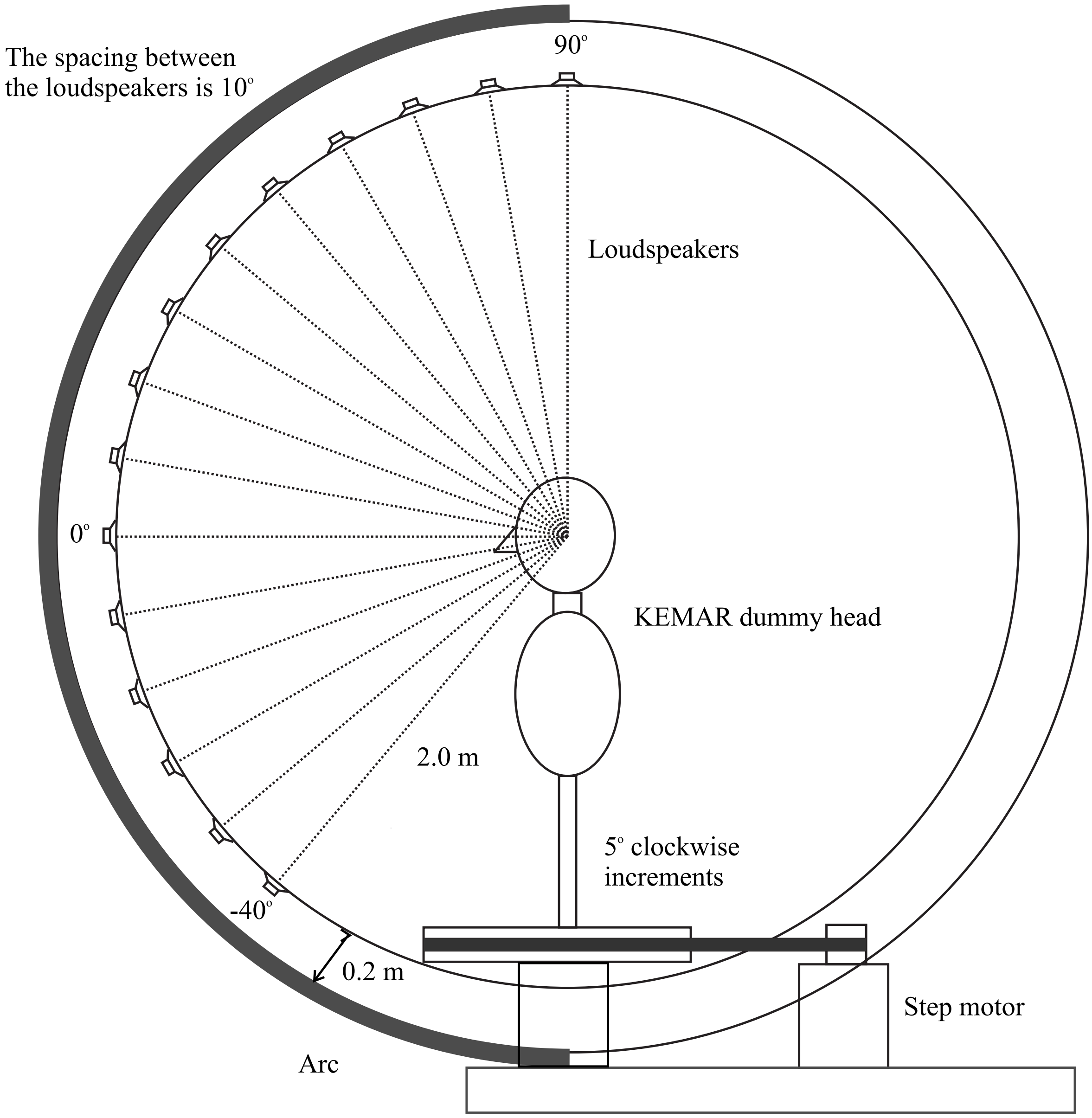


Figure 3.1: **The measurement rig for the HRTF database measurement**. *The arc holds 14 loudspeakers spaced by* 10° *starting at an elevation angle of* −40° *and goes up to an elevation angle of* 90°*. The KEMAR dummy head was rotated clockwise in the horisontal in steps of* 5° *using an electrical stepper motor.*

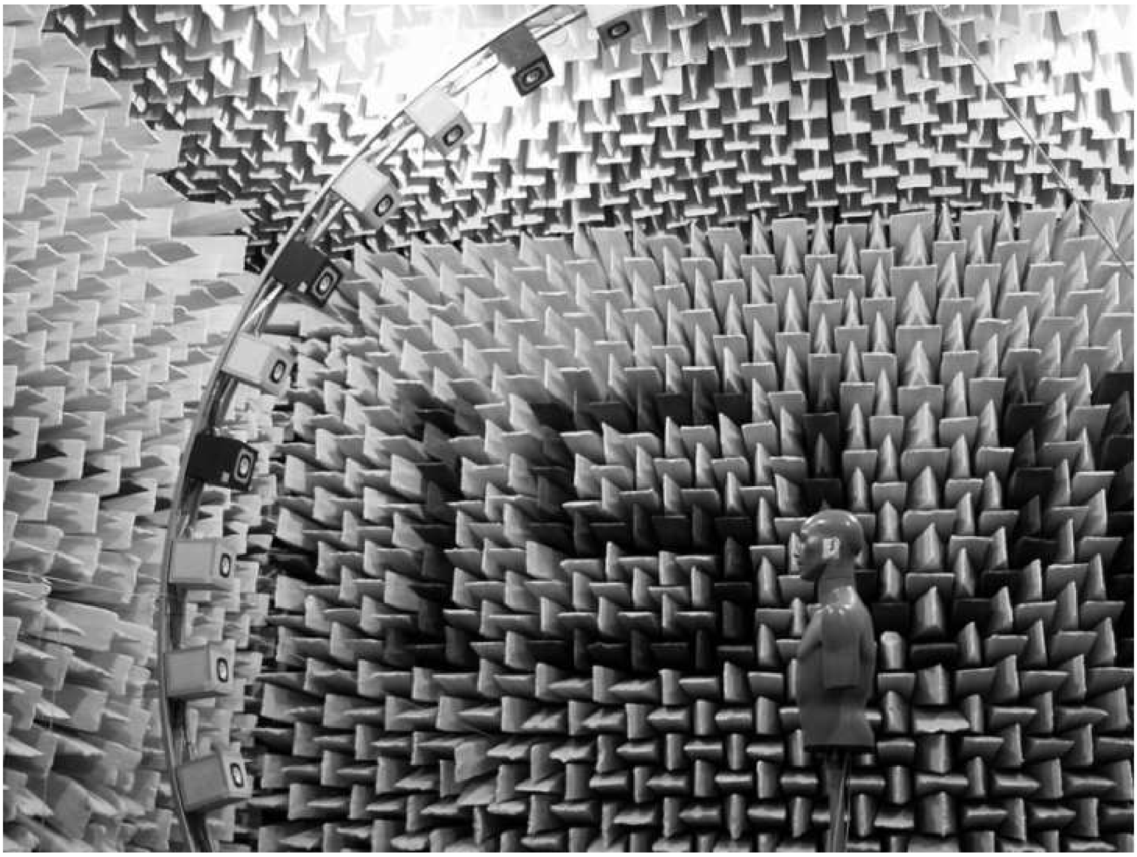
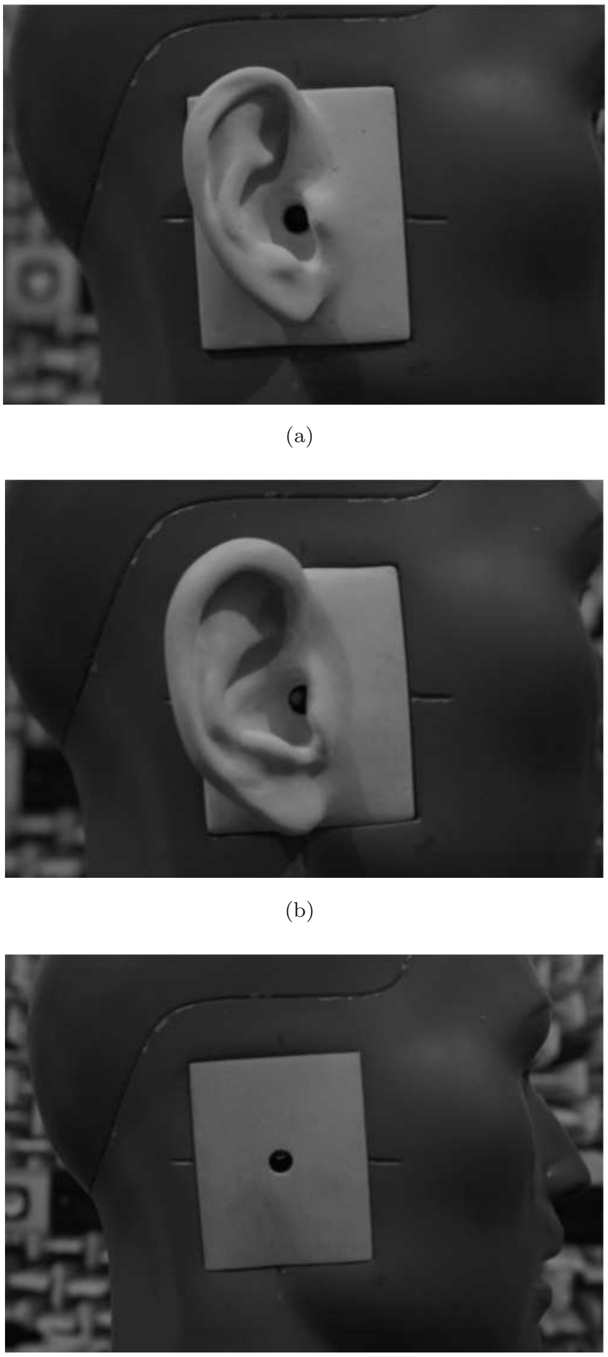


Figure 3.2: **Measurement rig**. *A new rig installed in the ISVR anechoic chamber for HRTF measurements.*



(a)

(b)

(c)

FIGURE 3.3: **Types of pinnae**. *The three different measurement configurations of pinnae used in the HRTF database measurement. The types of pinnae are shown for the open ear canal case. (a) "Small pinna". (b) "Large pinna". (c) "No pinna".*

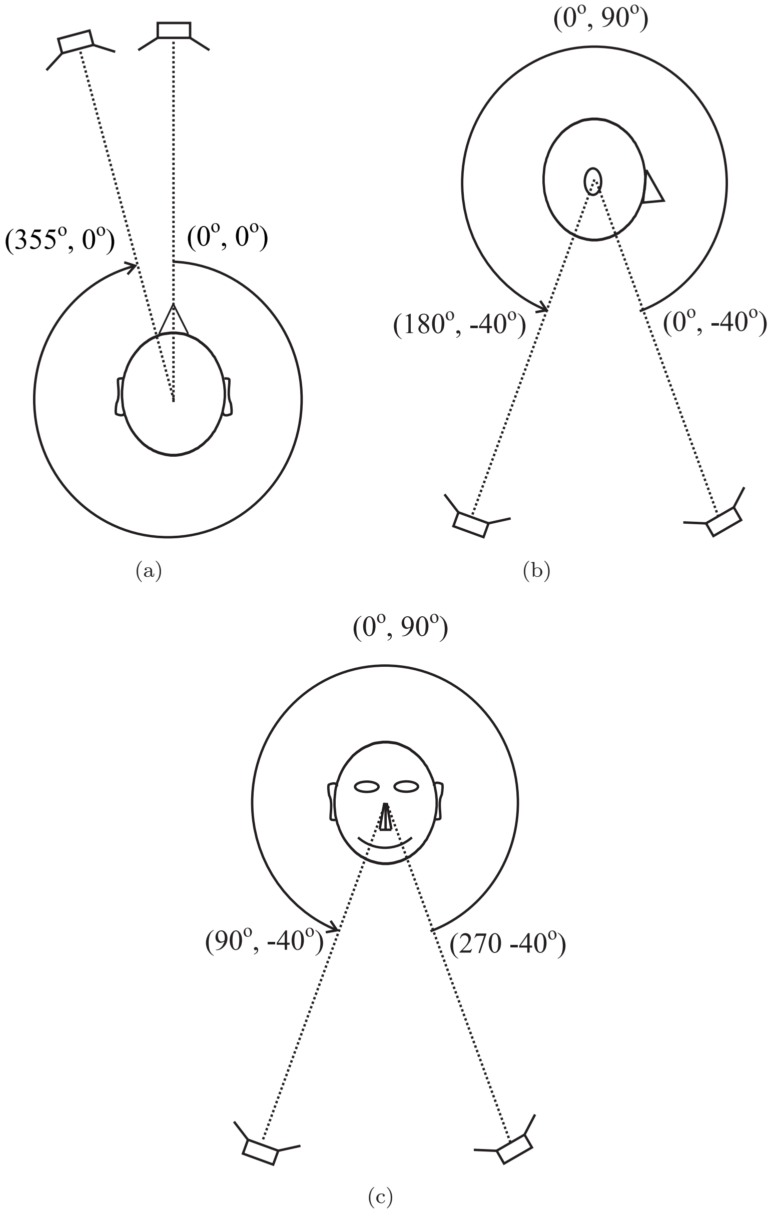
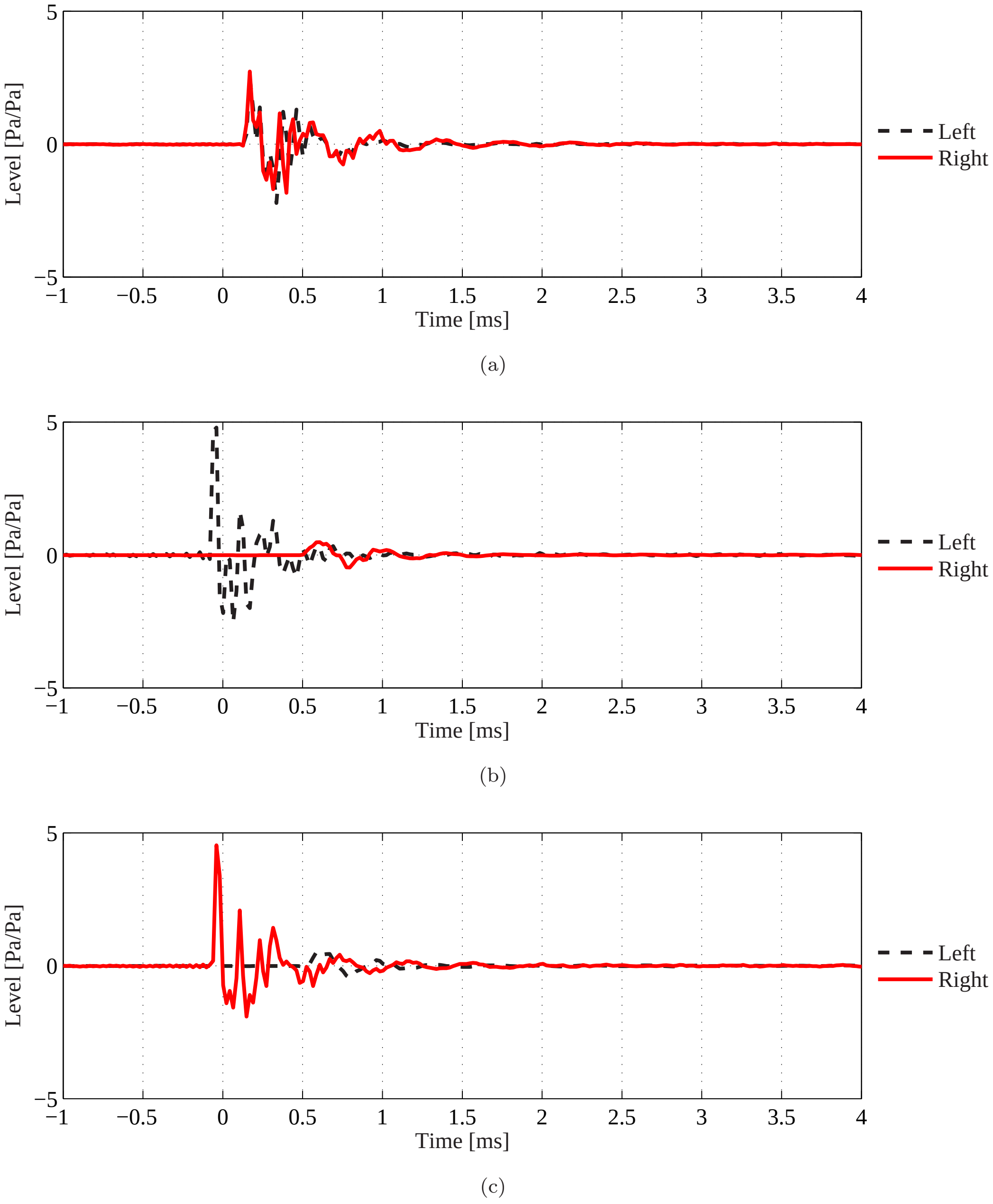


FIGURE 3.4: **Coordinate system for the HRTF database**. *The coordinate system for the HRTF database with azimuth angle ϕ and elevation angle ψ indicated as: (ϕ, ψ). (a) Horizontal plane. (b) Median plane. (c) Frontal plane.*

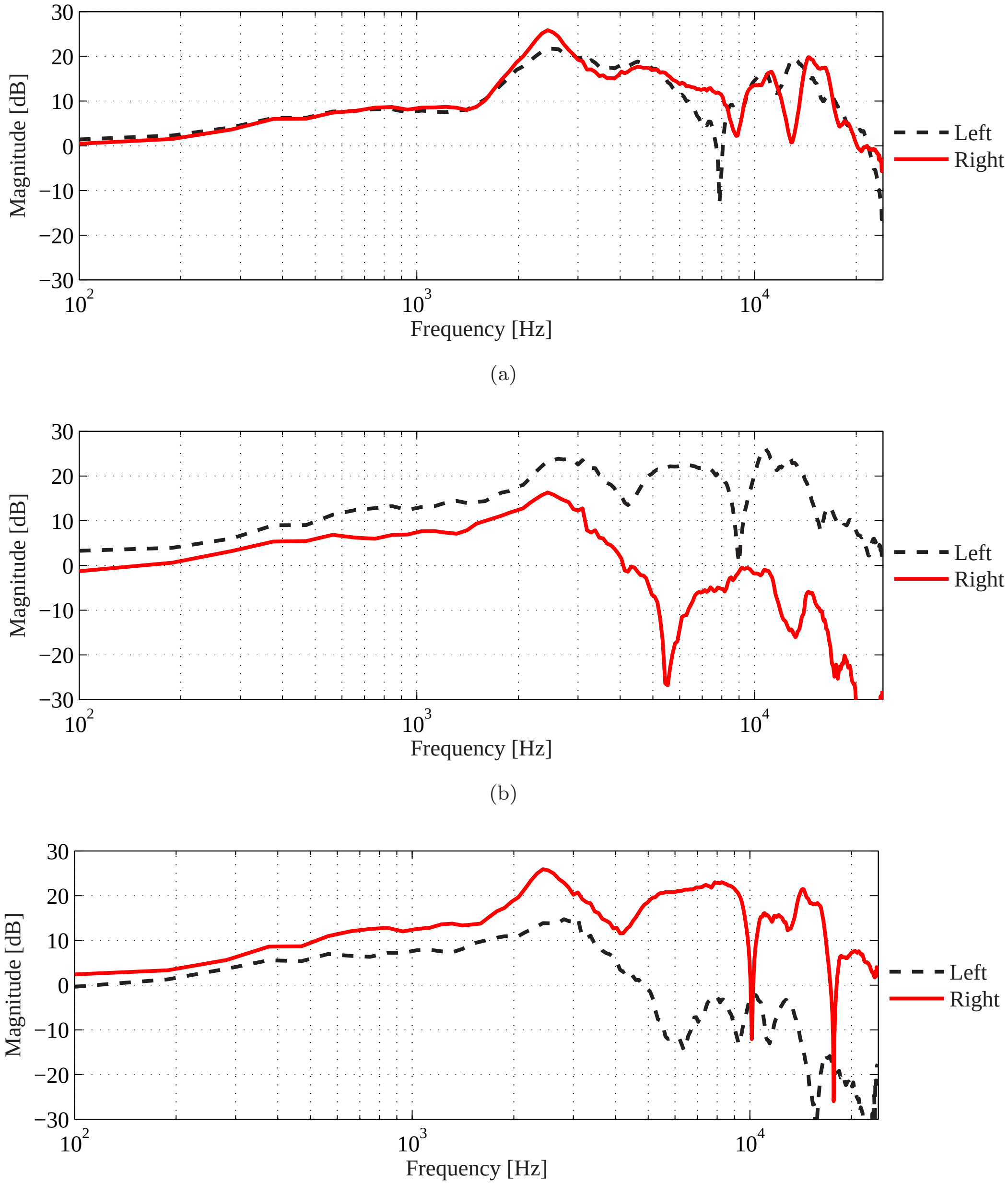


(a)

(b)

(c)

Figure 3.5: **Time domain response for three azimuth angles at $\psi = 0°$ elevation**. *The following different azimuth angles are presented: (a) $\phi = 0°$, (b) $\phi = 90°$ and (c) $\phi = 270°$. The measurement is undertaken on the "small pinna" with an open ear canal at the ISVR.*

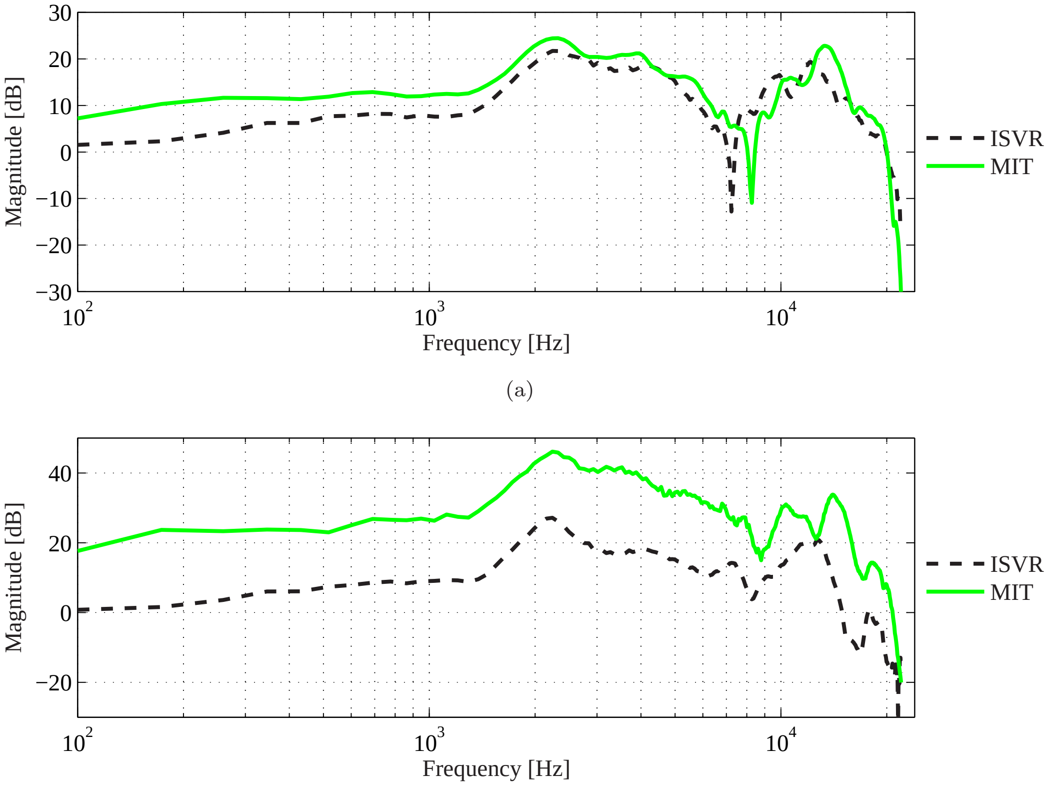


(a)

(b)

(c)

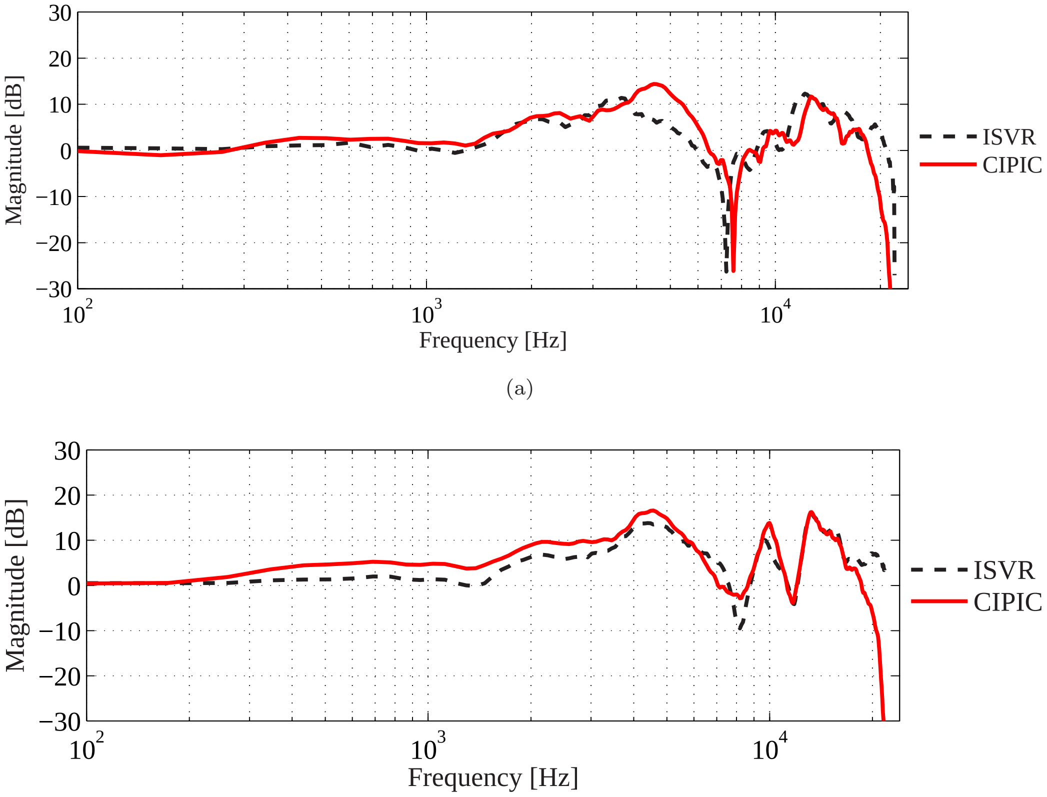
FIGURE 3.6: **Frequency response for three different azimuth angles at $\psi = 0°$ elevation**. *The following different azimuth angles are presented: $\phi = 0°$ (a) $\phi = 90°$ (b) and $\phi = 270°$ (c). The measurement is undertaken on the "small pinna" with open ear canal at the ISVR.*



(a)

(b)

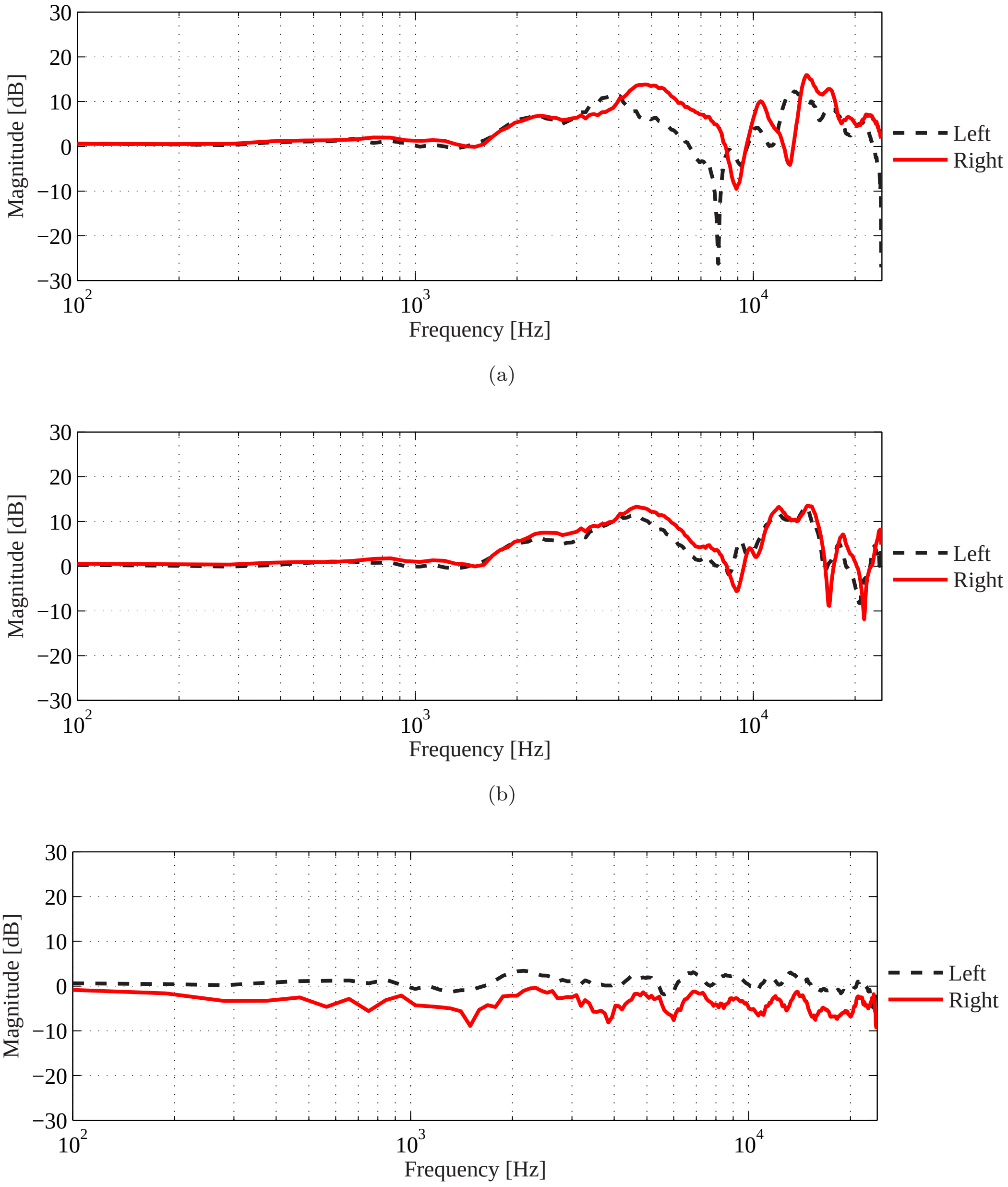
FIGURE 3.7: **Frequency response for the open ear canal comparison of data from MIT and ISVR**. *The ISVR curve is set to be the reference and the MIT curve is fitted to the DC component (0 Hz) of the reference. (a) "Small pinna". (b) "Large pinna".*



(a)

(b)

FIGURE 3.8: **Frequency response for the "small pinna" with blocked ear canal comparison of data from CIPIC and ISVR**. *The ISVR curve is set to be the reference and the CIPIC curve is fitted to the DC component (0 Hz) of the reference. (a) Left ear. (b) Right ear.*

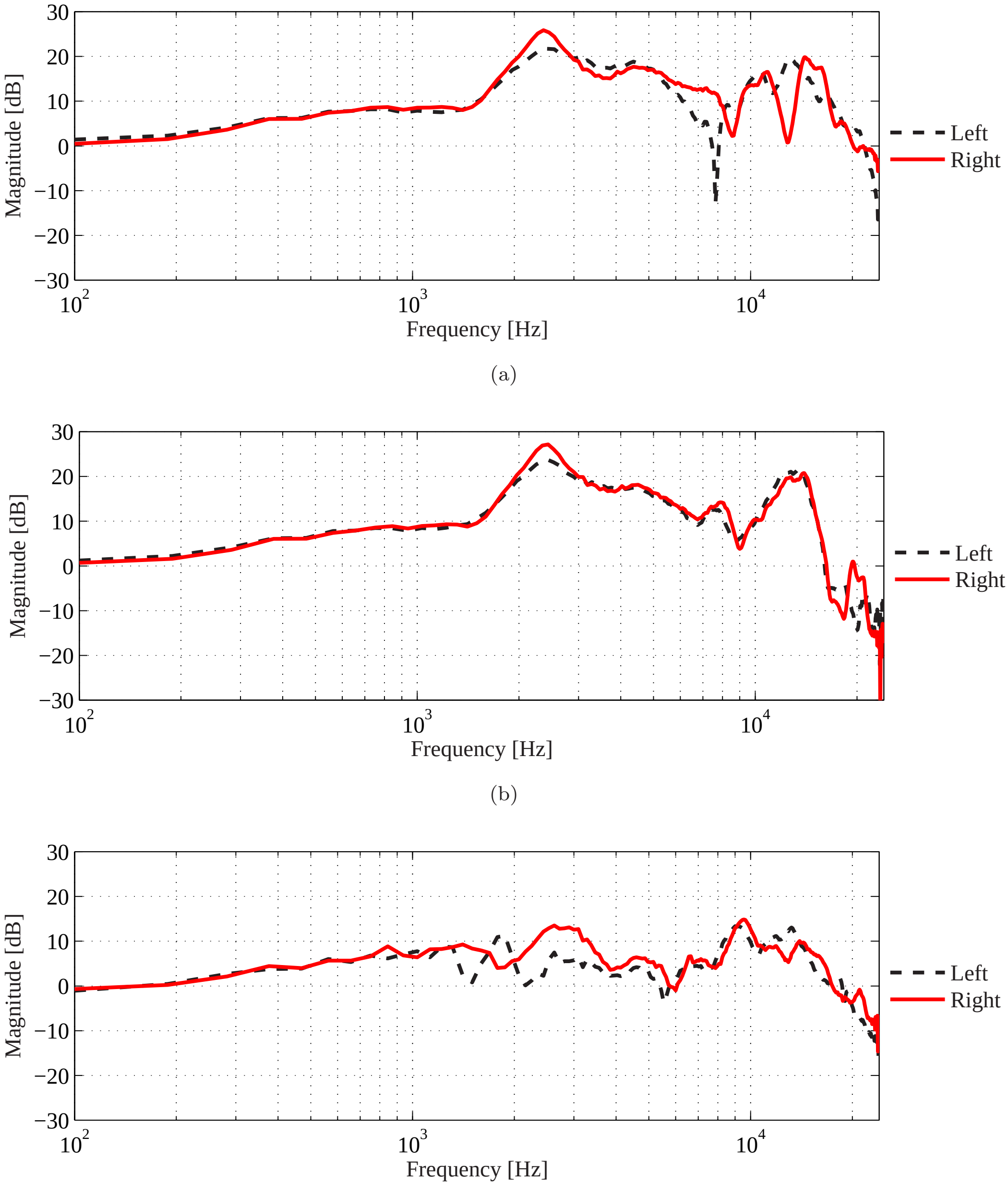


(a)

(b)

(c)

FIGURE 3.9: **Frequency response comparison of different pinnae for the blocked ear canal ($\phi = 0°$, $\psi = 0°$).** *The comparison is carried out using data from the ISVR HRTF database. (a) ”small pinna”, (b) “large pinna” and (c) “no pinna”.*

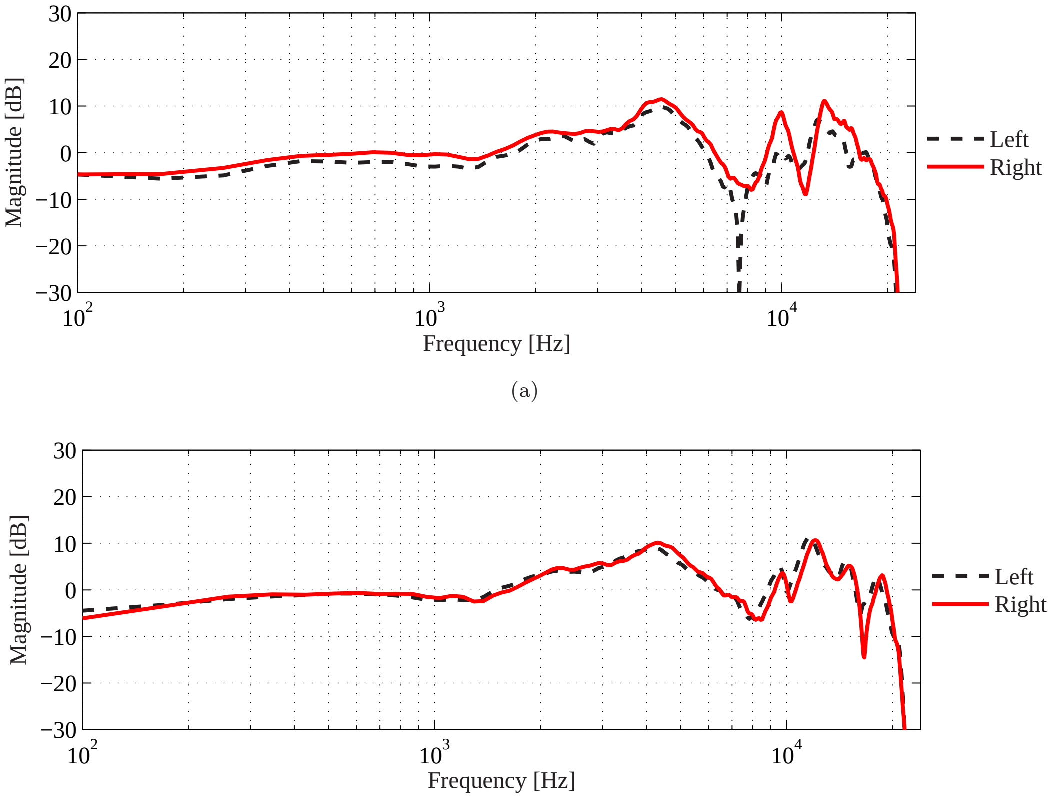


(a)

(b)

(c)

Figure 3.10: **Frequency response comparison of different pinnae for the open ear canal ($\phi = 0°$,$\psi = 0°$).** *The comparison is carried out using data from the ISVR HRTF database. (a) “small pinna”, (b) “large pinna” and (c) “no pinna”.*



(a)

(b)

FIGURE 3.11: **Frequency response comparison of different pinnae for the blocked ear canal ($\phi = 0°$, $\psi = 0°$).** *The comparison is carried out using data from the CIPIC HRTF database. (a) "small pinna", (b) "large pinna" and (c) "no pinna".*

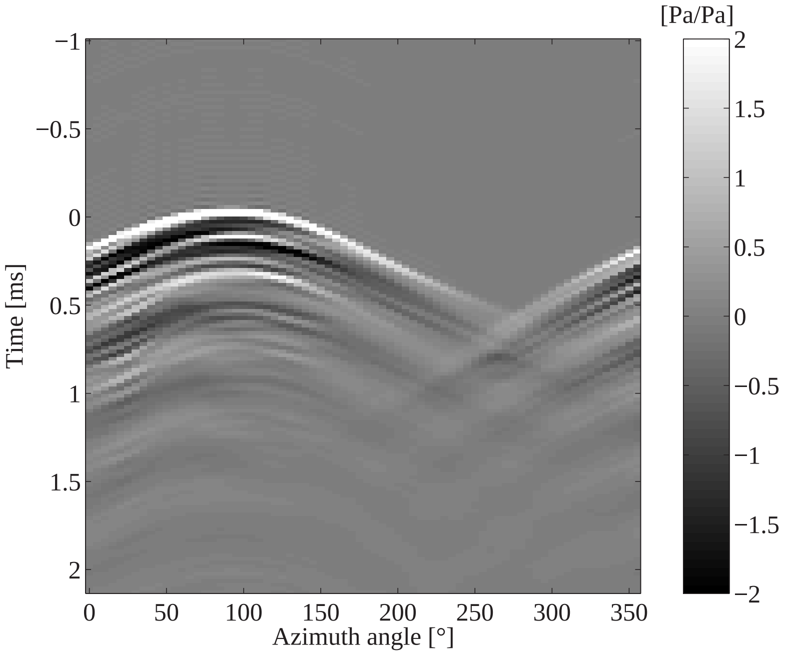


Figure 3.12: **HRIR in the horizontal plane for the right open ear canal "small pinna"**. *The time domain response is plotted as a function of azimuth angle ϕ.*

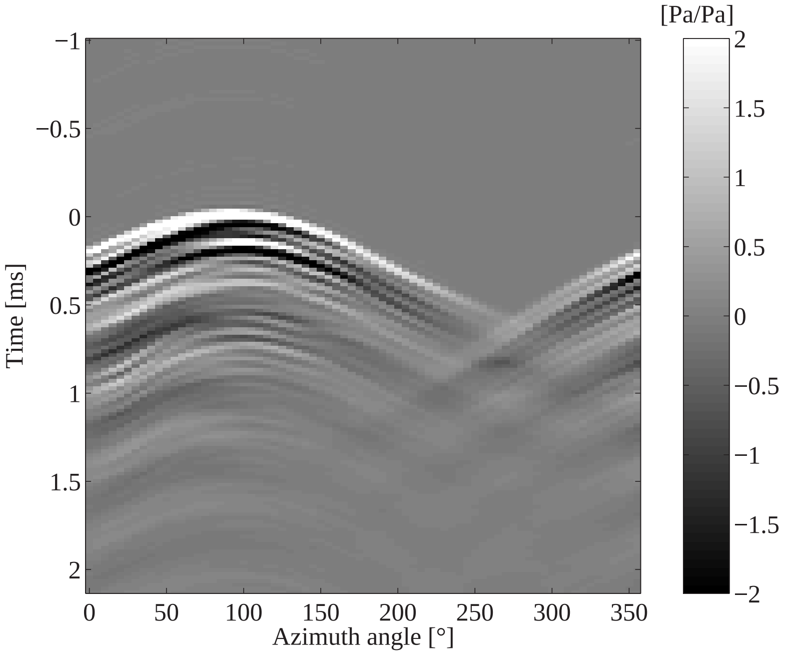


Figure 3.13: **HRIR in the horizontal plane for the right open ear canal "large pinna"**. *The time domain response is plotted as a function of azimuth angle ϕ.*

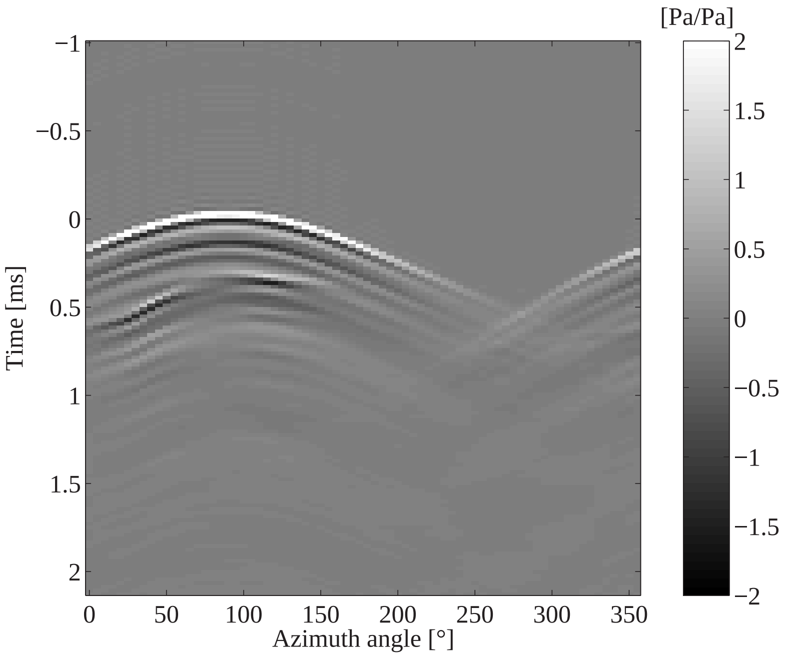


Figure 3.14: **HRIR in the horizontal plane for the right open ear canal "no pinna"**. *The time domain response is plotted as a function of azimuth angle ϕ.*

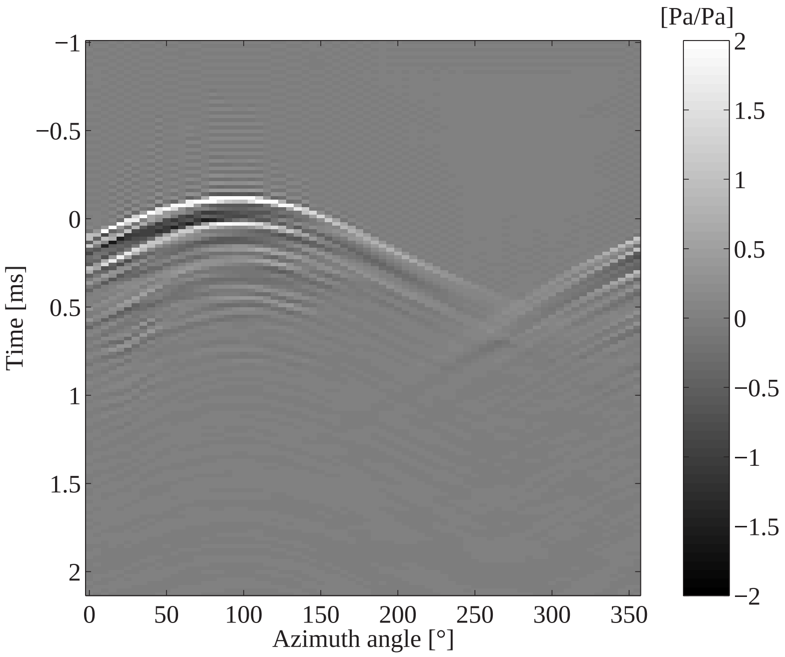


Figure 3.15: **HRIR in the horizontal plane for the right blocked ear canal "small pinna"**. *The time domain response is plotted as a function of azimuth angle ϕ.*

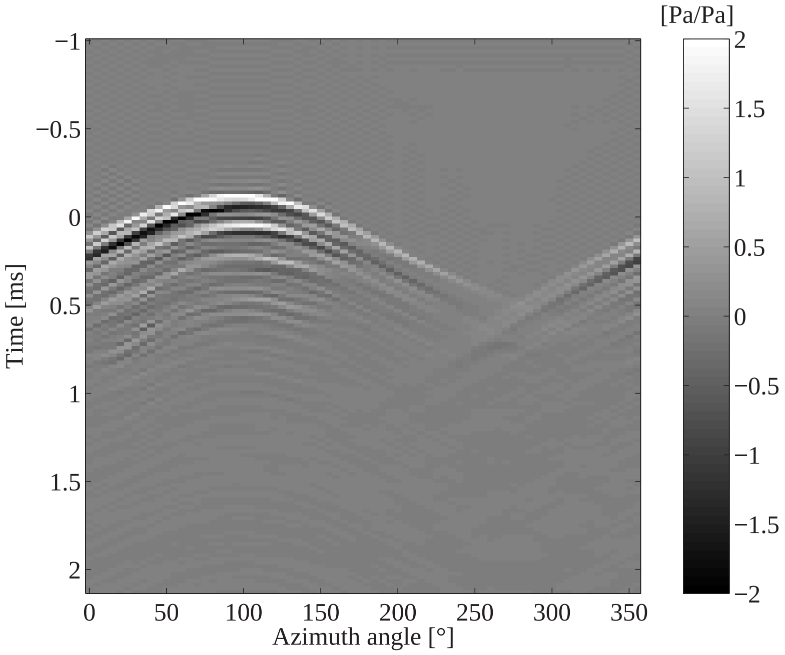


Figure 3.16: **HRIR in the horizontal plane for the right blocked ear canal "large pinna"**. *The time domain response is plotted as a function of azimuth angle ϕ.*

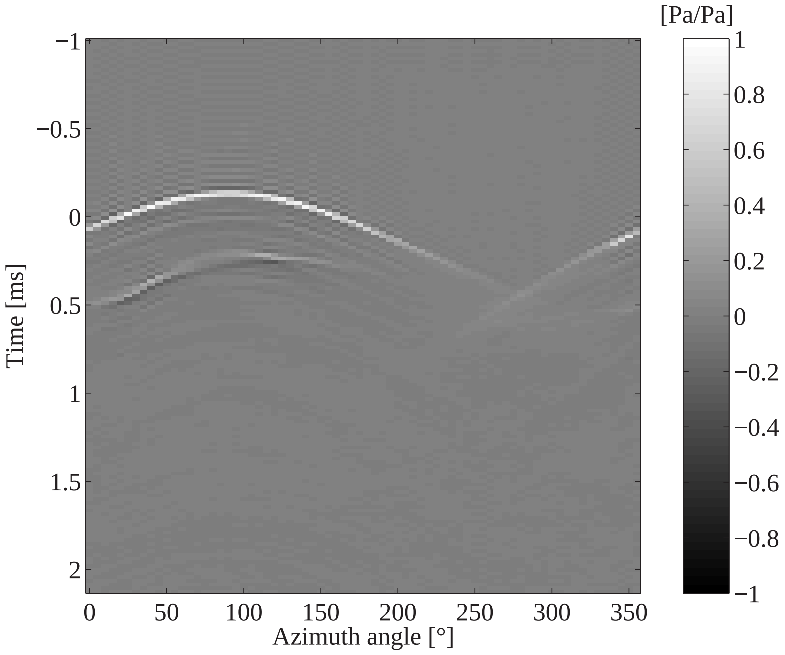


Figure 3.17: **HRIR in the horizontal plane for the right blocked ear canal "no pinna"**. *The time domain response is plotted as a function of azimuth angle ϕ.*

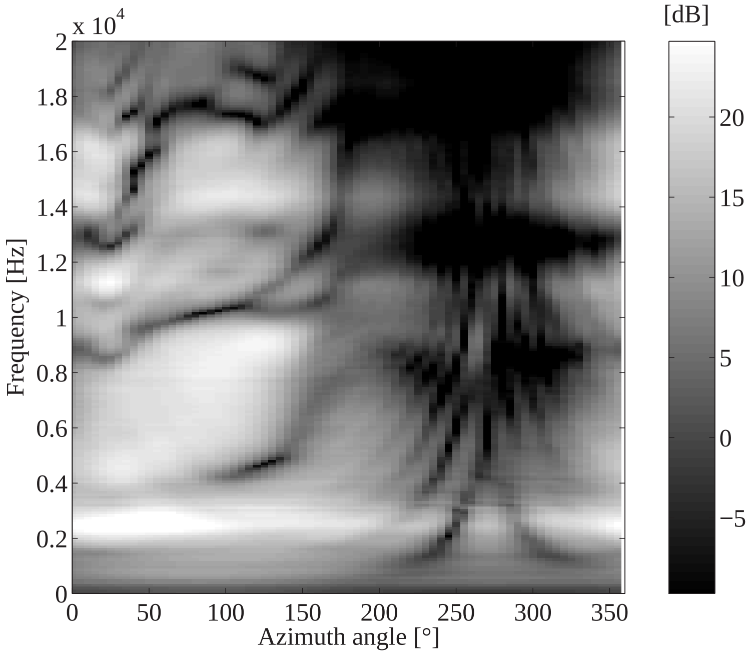


FIGURE 3.18: **HRTF in the horizontal plane for the right open ear canal "small pinna"**. *The frequency domain response is plotted as a function of azimuth angle ϕ.*

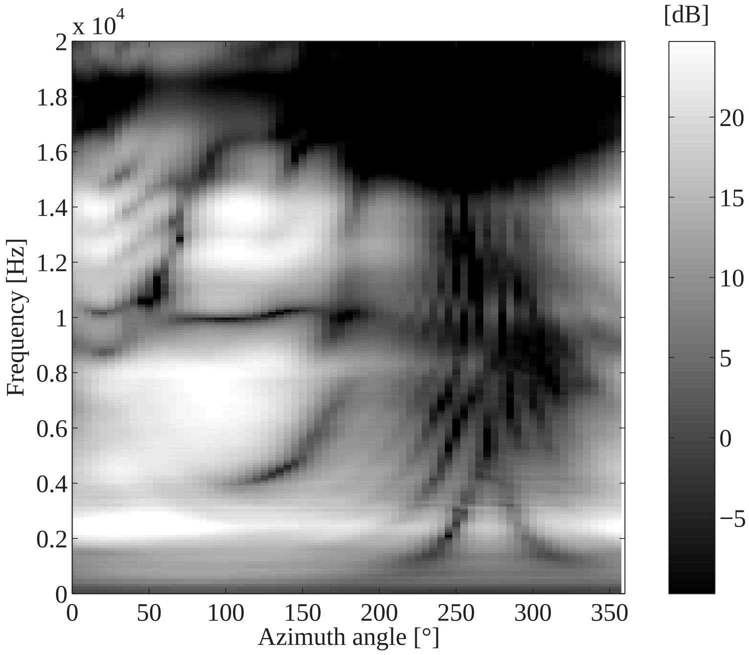


FIGURE 3.19: **HRTF in the horizontal plane for the right open ear canal "large pinna"**. *The frequency domain response is plotted as a function of azimuth angle ϕ.*

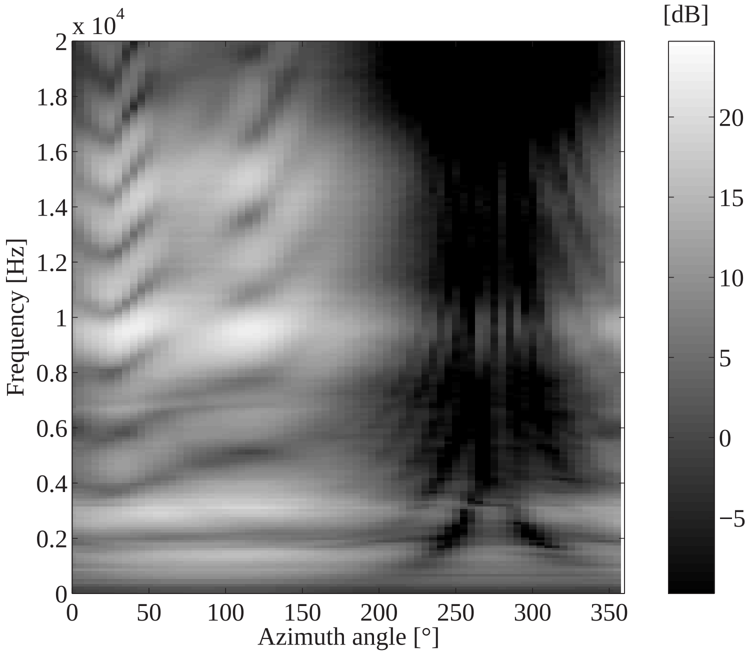


FIGURE 3.20: **HRTF in the horizontal plane for the right open ear canal "no pinna"**. *The frequency domain response is plotted as a function of azimuth angle ϕ.*

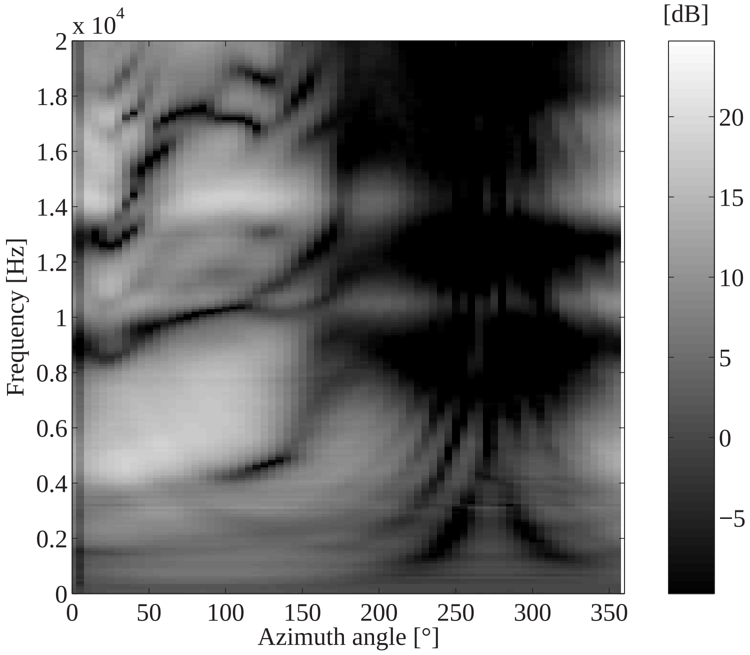


FIGURE 3.21: **HRTF in the horizontal plane for the right blocked ear canal "small pinna"**. *The frequency domain response is plotted as a function of azimuth angle ϕ.*

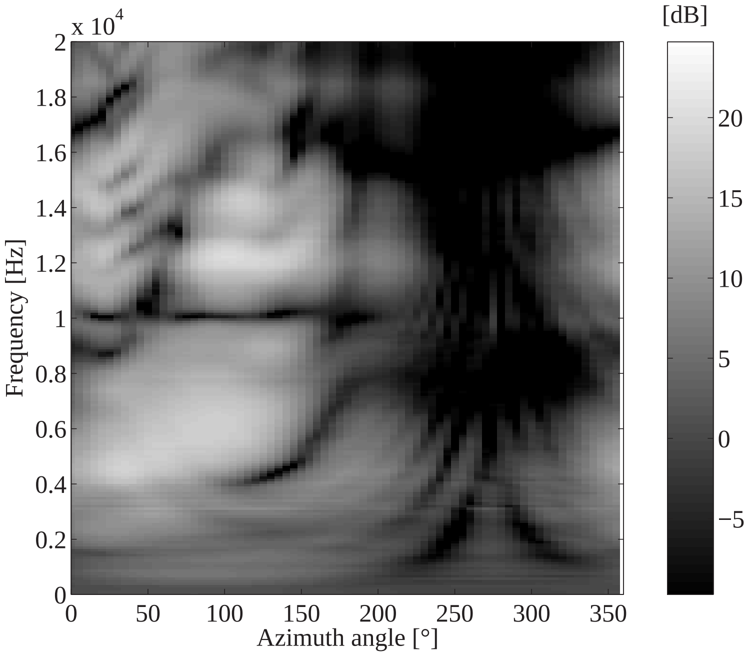


Figure 3.22: **HRTF in the horizontal plane for the right blocked ear canal "large pinna".** *The frequency domain response is plotted as a function of azimuth angle ϕ.*

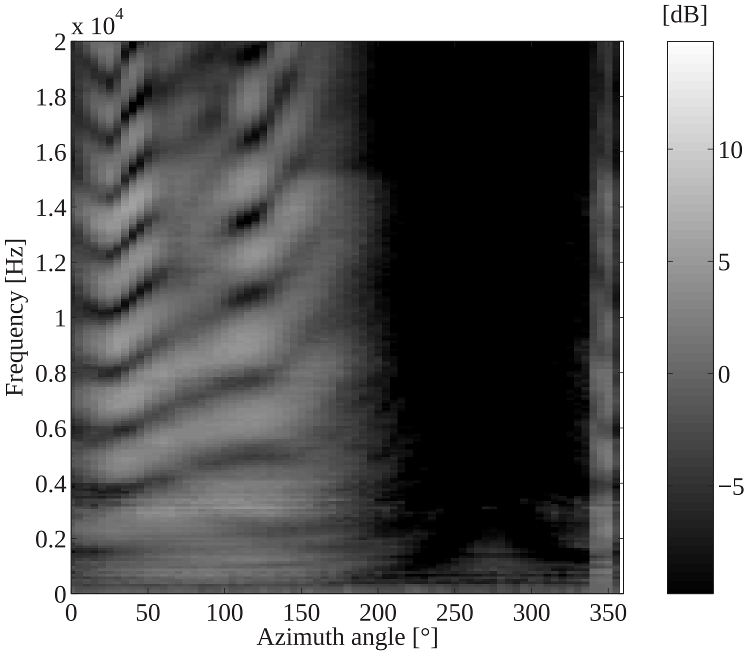


Figure 3.23: **HRTF in the horizontal plane for the right blocked ear canal "no pinna".** *The frequency domain response is plotted as a function of azimuth angle ϕ.*

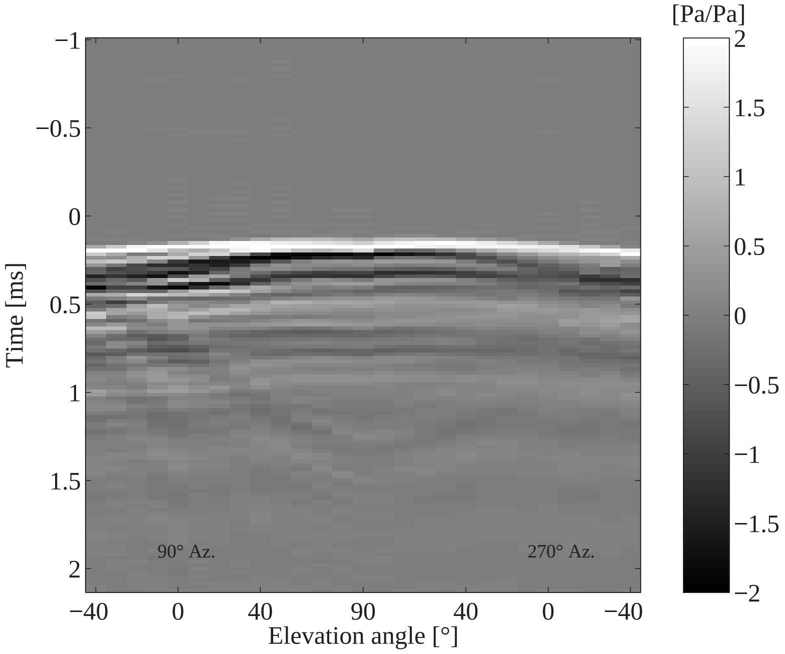


Figure 3.24: **HRIR in the median plane for the right open ear canal with "small pinna"**. *The time domain response is plotted as a function of elevation angle* ψ.

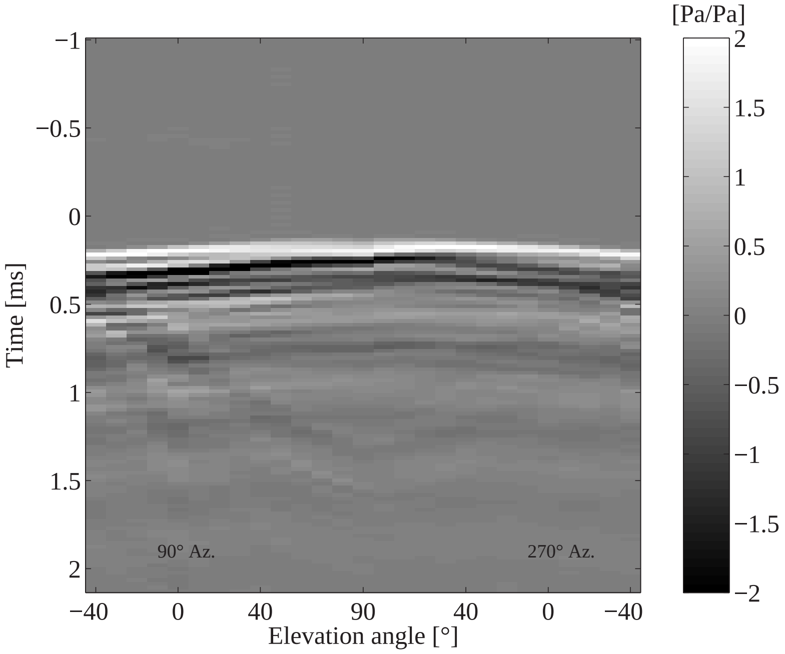


Figure 3.25: **HRIR in the median plane for the right open ear canal with "large pinna"**. *The time domain response is plotted as a function of elevation angle* ψ.

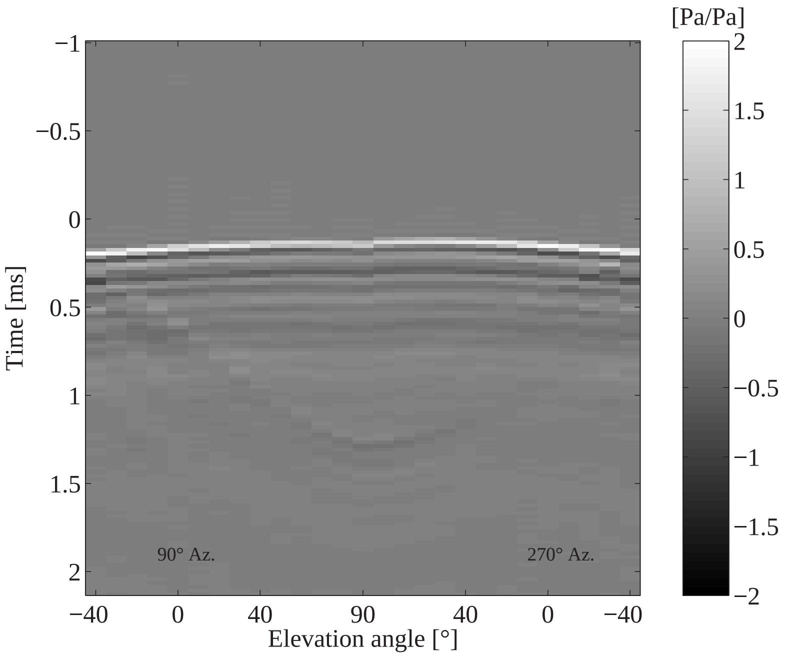


FIGURE 3.26: **HRIR in the median plane for the right open ear canal with "no pinna"**. *The time domain response is plotted as a function of elevation angle ψ.*

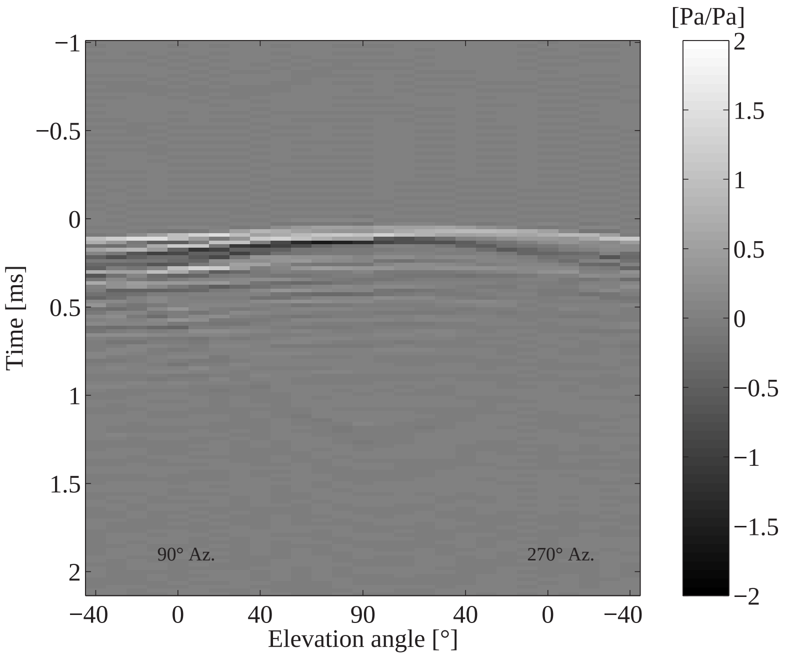


FIGURE 3.27: **HRIR in the median plane for the right blocked ear canal with "small pinna"**. *The time domain response is plotted as a function of elevation angle ψ.*

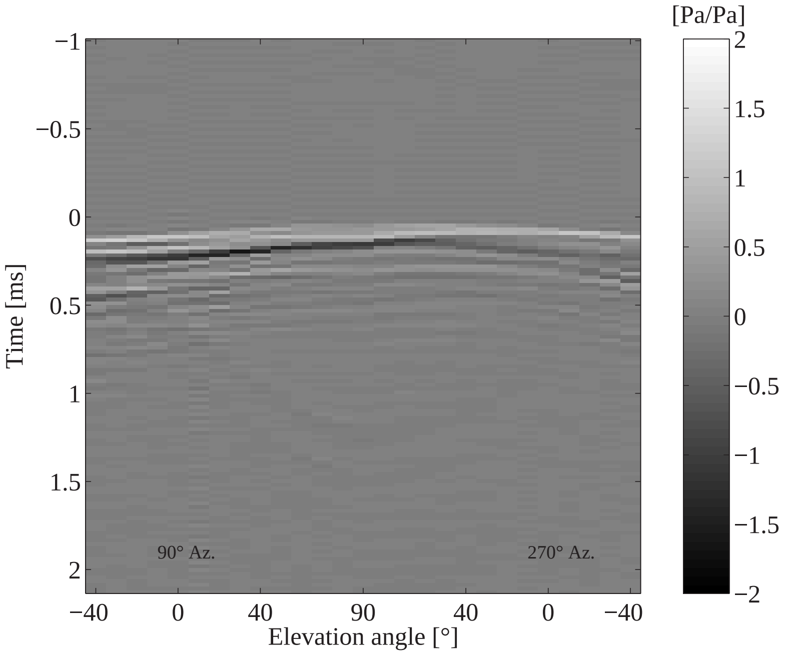


Figure 3.28: **HRIR in the median plane for the right blocked ear canal with "large pinna"**. *The time domain response is plotted as a function of elevation angle ψ.*

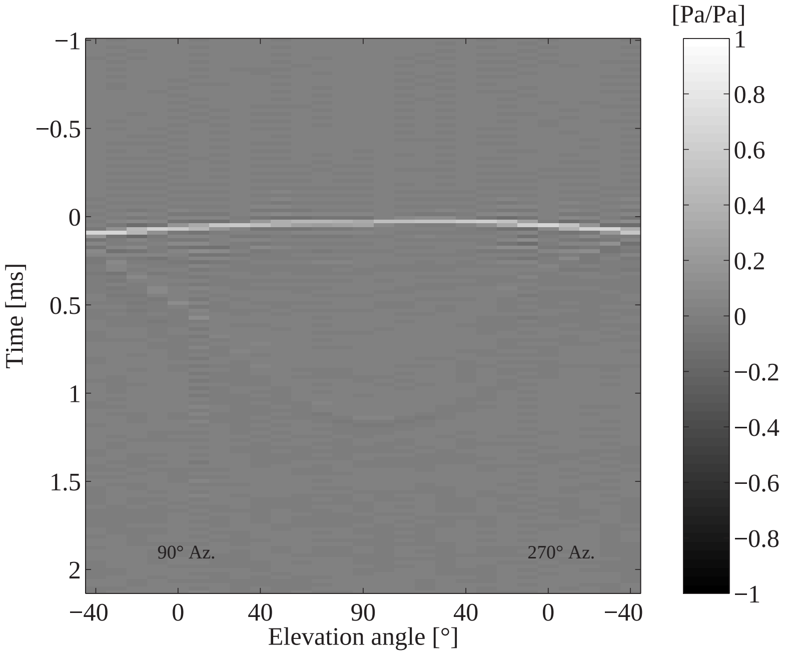


Figure 3.29: **HRIR in the median plane for the right blocked ear canal with "no pinna"**. *The time domain response is plotted as a function of elevation angle ψ.*

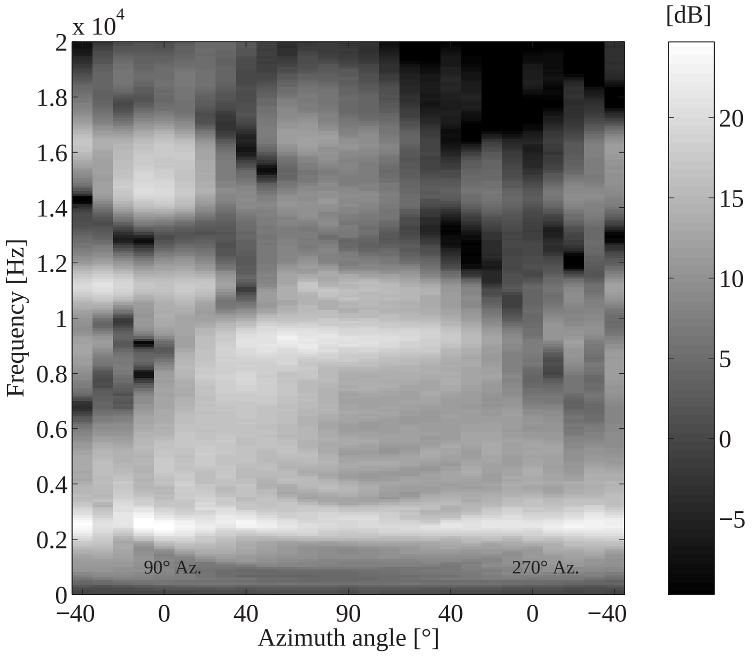


FIGURE 3.30: **HRTF in the median plane for the right open ear canal with "small pinna"**. *The frequency domain response is plotted as a function of elevation angle ψ.*

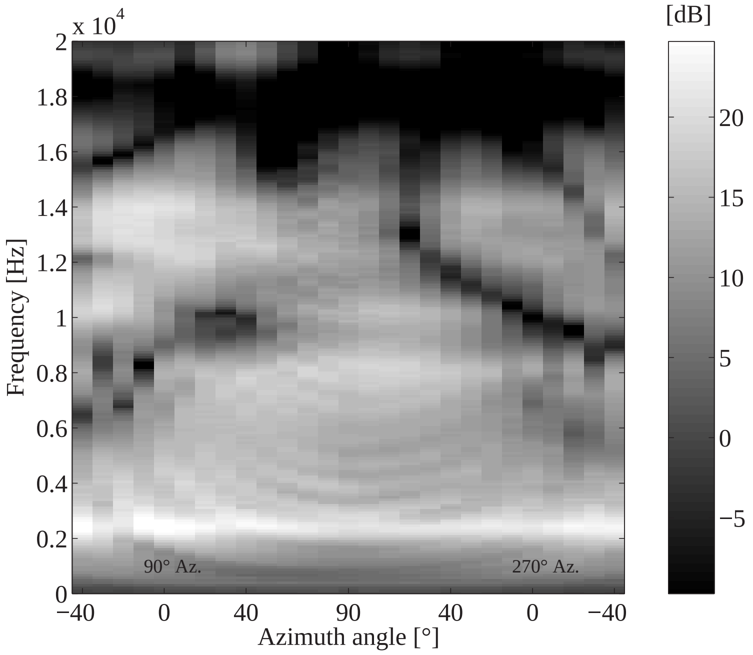


FIGURE 3.31: **HRTF in the median plane for the right open ear canal with "large pinna"**. *The frequency domain response is plotted as a function of elevation angle ψ.*

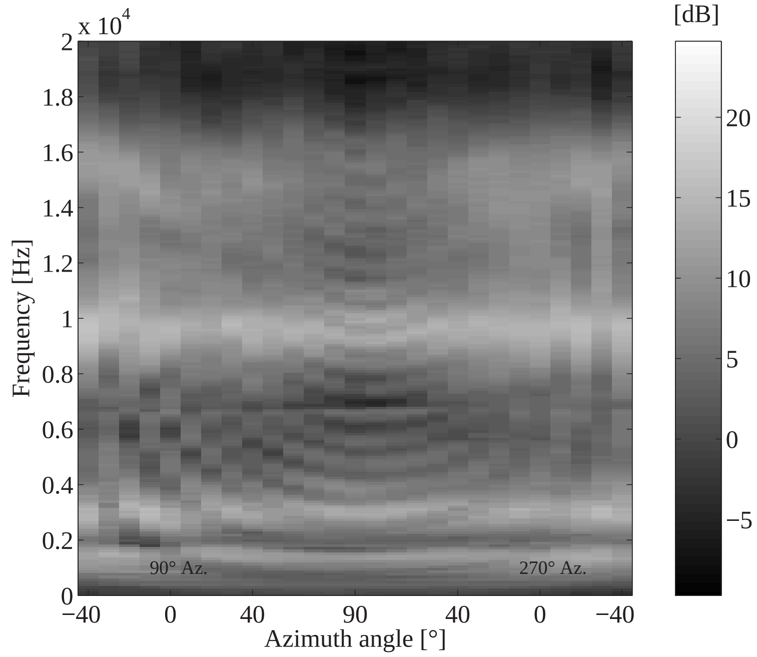


FIGURE 3.32: **HRTF in the median plane for the right open ear canal with "no pinna"**. *The frequency domain response is plotted as a function of elevation angle ψ.*

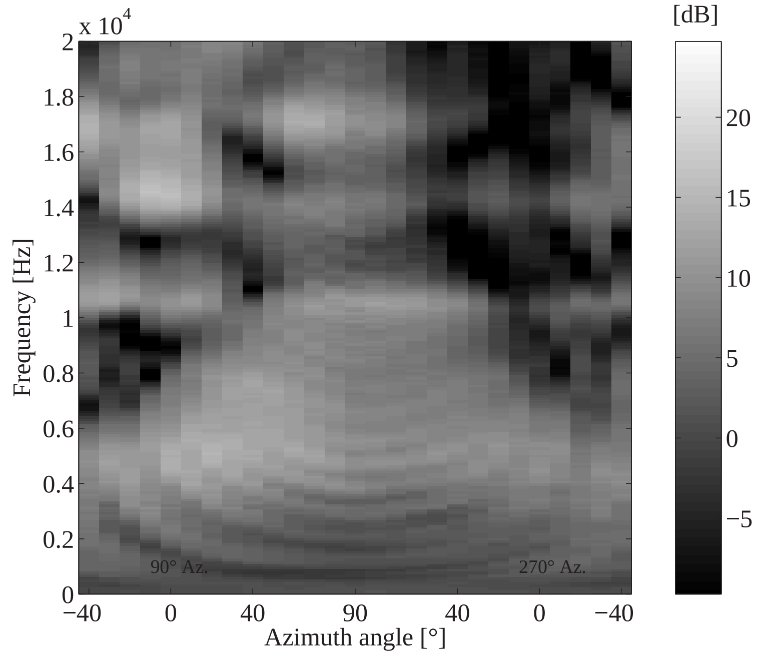


FIGURE 3.33: **HRTF in the median plane for the right blocked ear canal with "small pinna"**. *The frequency domain response is plotted as a function of elevation angle ψ.*

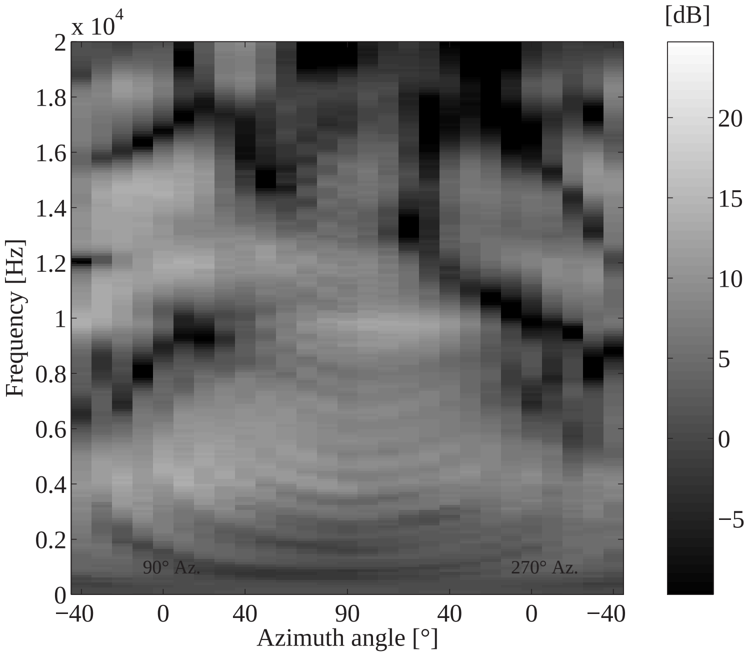


Figure 3.34: **HRTF in the median plane for the right blocked ear canal with "large pinna"**. *The frequency domain response is plotted as a function of elevation angle ψ.*

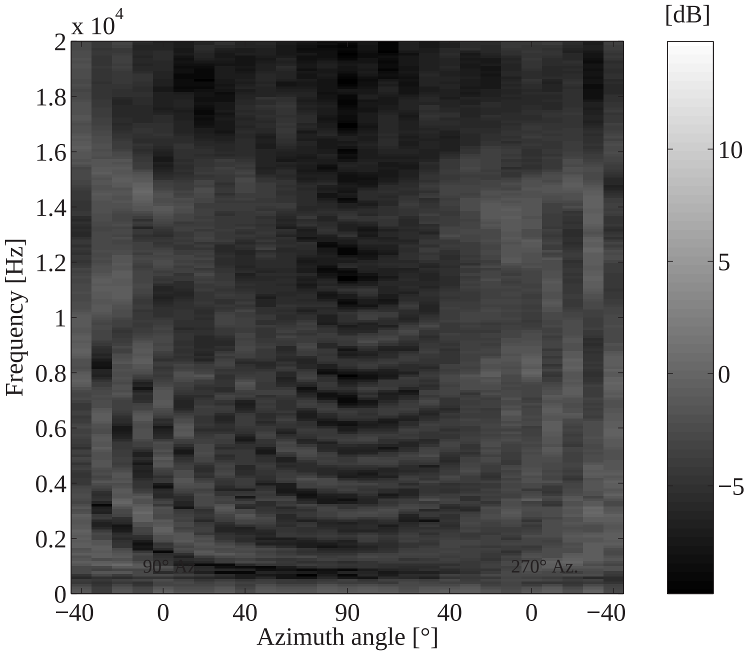


Figure 3.35: **HRTF in the median plane for the right blocked ear canal with "no pinna"**. *The frequency domain response is plotted as a function of elevation angle ψ.*

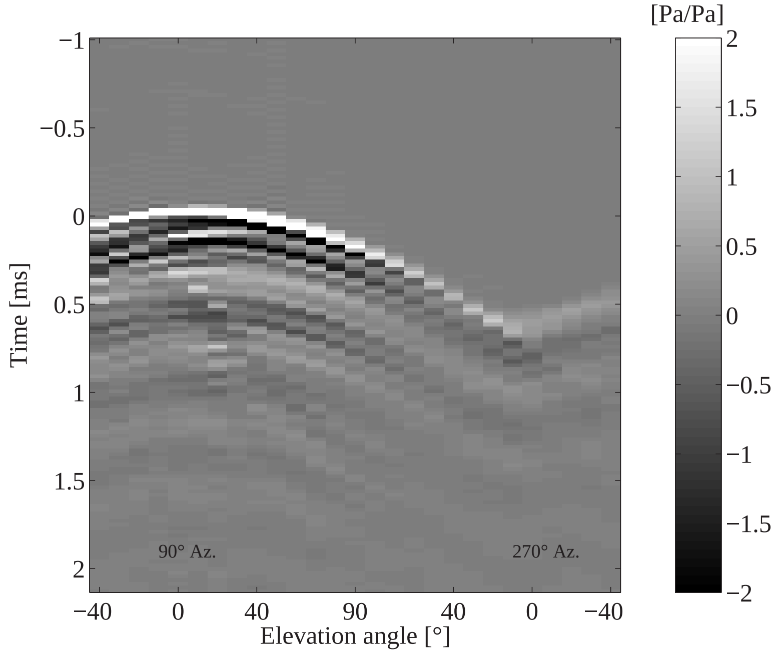


FIGURE 3.36: **HRIR in the frontal plane for the right open ear canal with "small pinna"**. *The time domain response is plotted as a function of elevation angle* ψ.

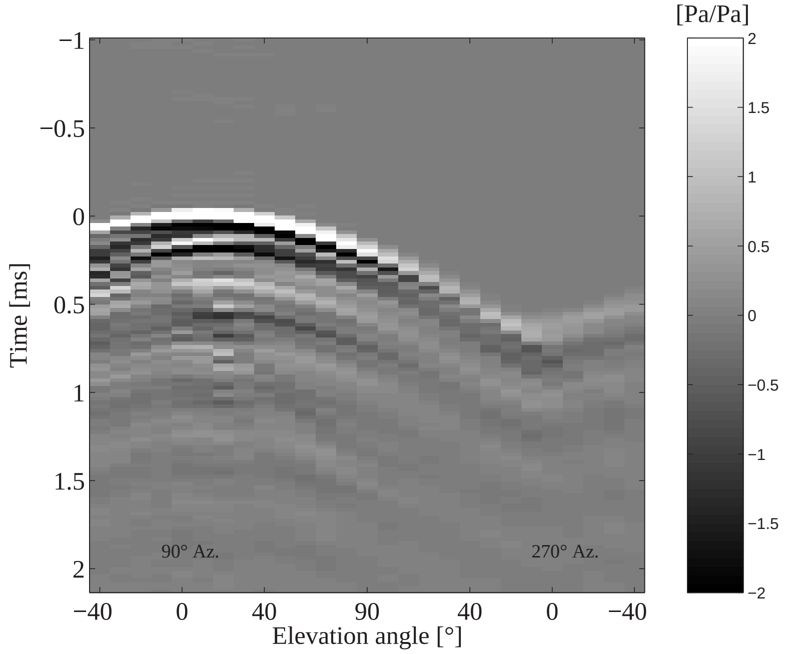


FIGURE 3.37: **HRIR in the frontal plane for the right open ear canal with "large pinna"**. *The time domain response is plotted as a function of elevation angle* ψ.

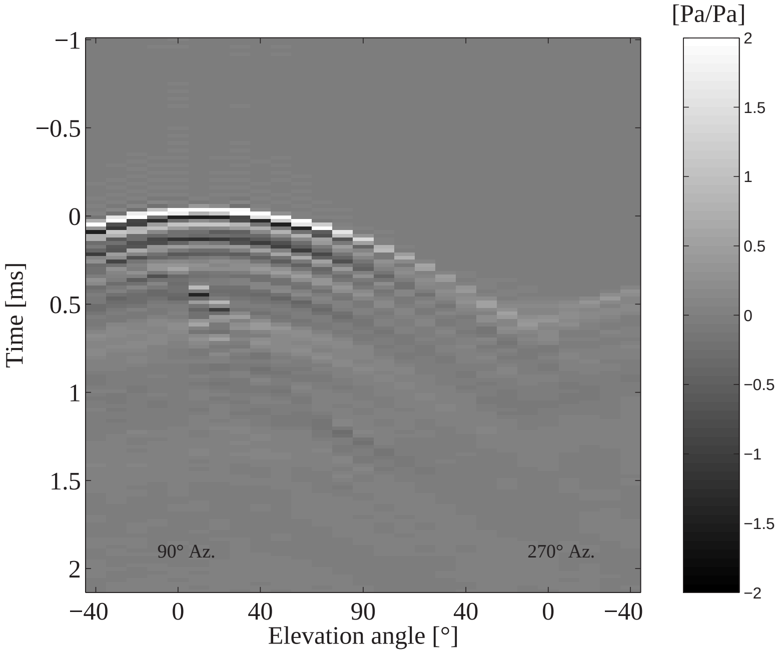


Figure 3.38: **HRIR in the frontal plane for the right open ear canal with "no pinna"**. *The time domain response is plotted as a function of elevation angle ψ.*

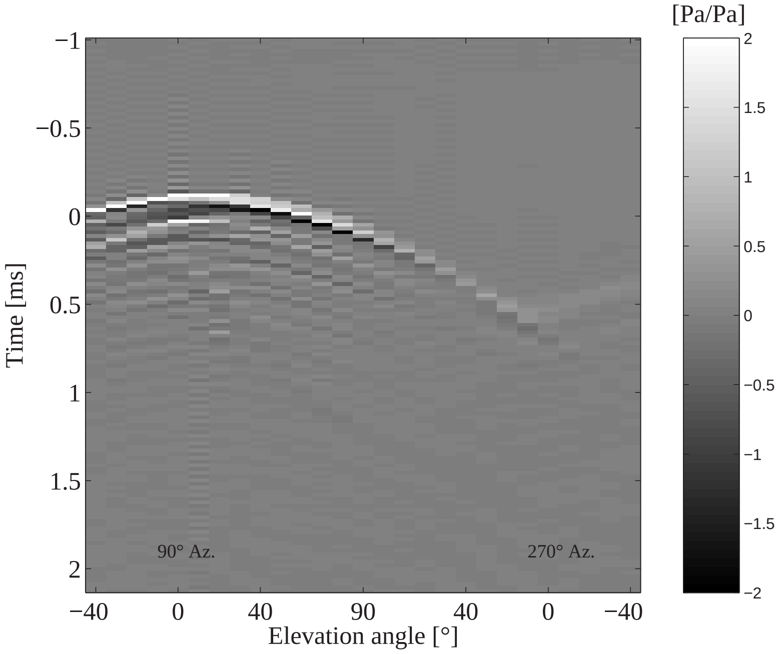


Figure 3.39: **HRIR in the frontal plane for the right blocked ear canal with "small pinna"**. *The time domain response is plotted as a function of elevation angle ψ.*

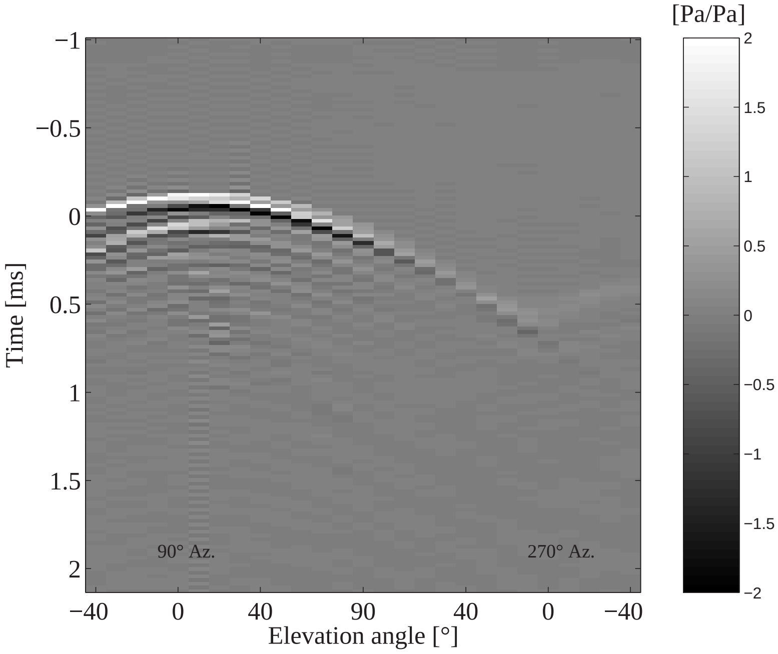


FIGURE 3.40: **HRIR in the frontal plane for the right blocked ear canal with "large pinna"**. *The time domain response is plotted as a function of elevation angle* ψ.

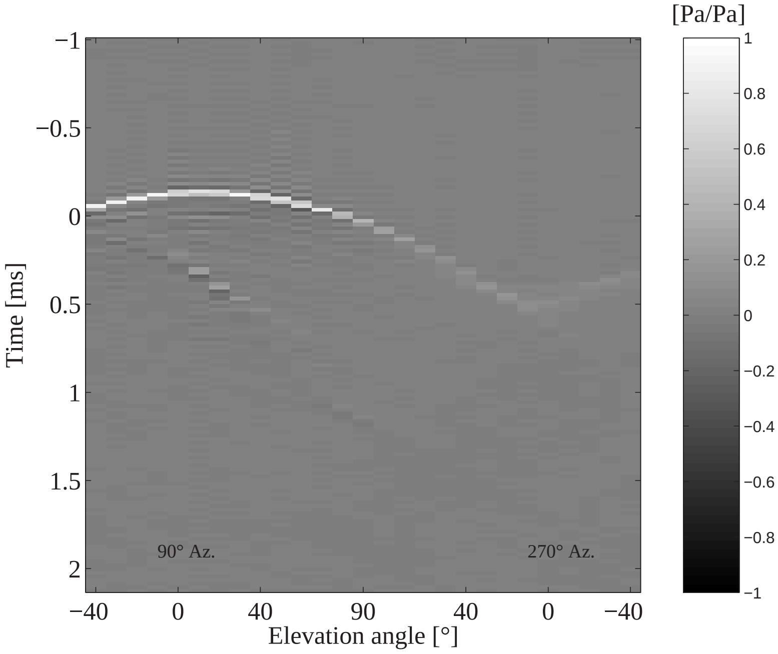
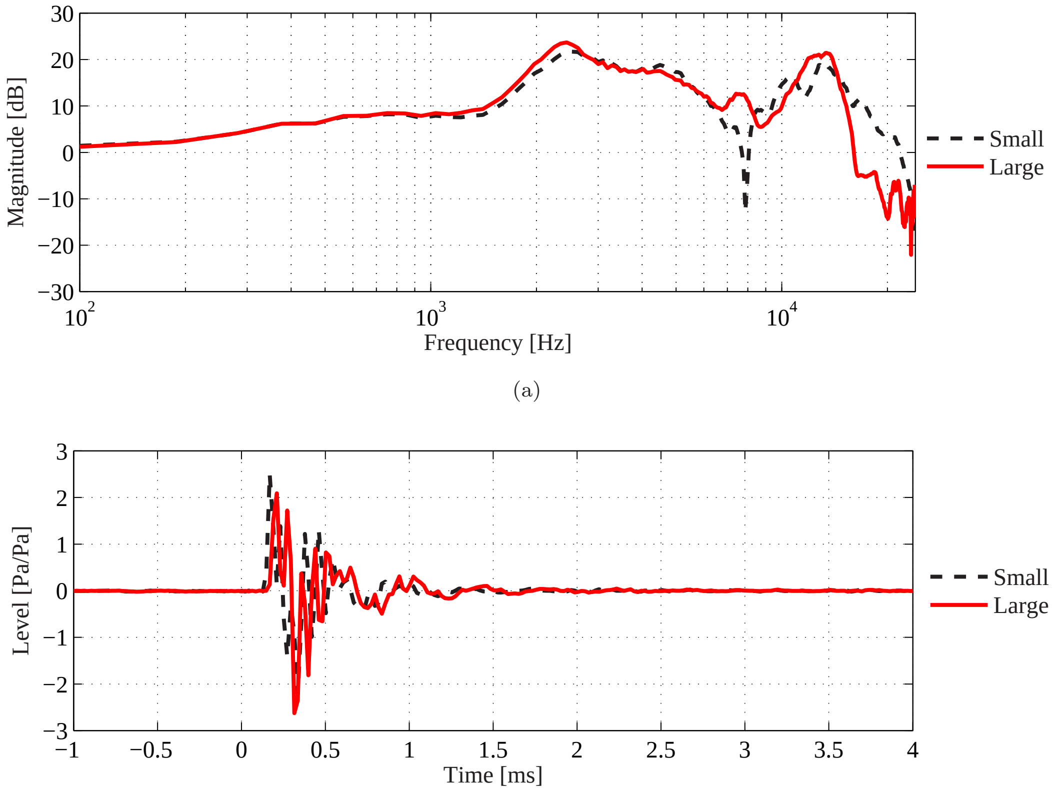
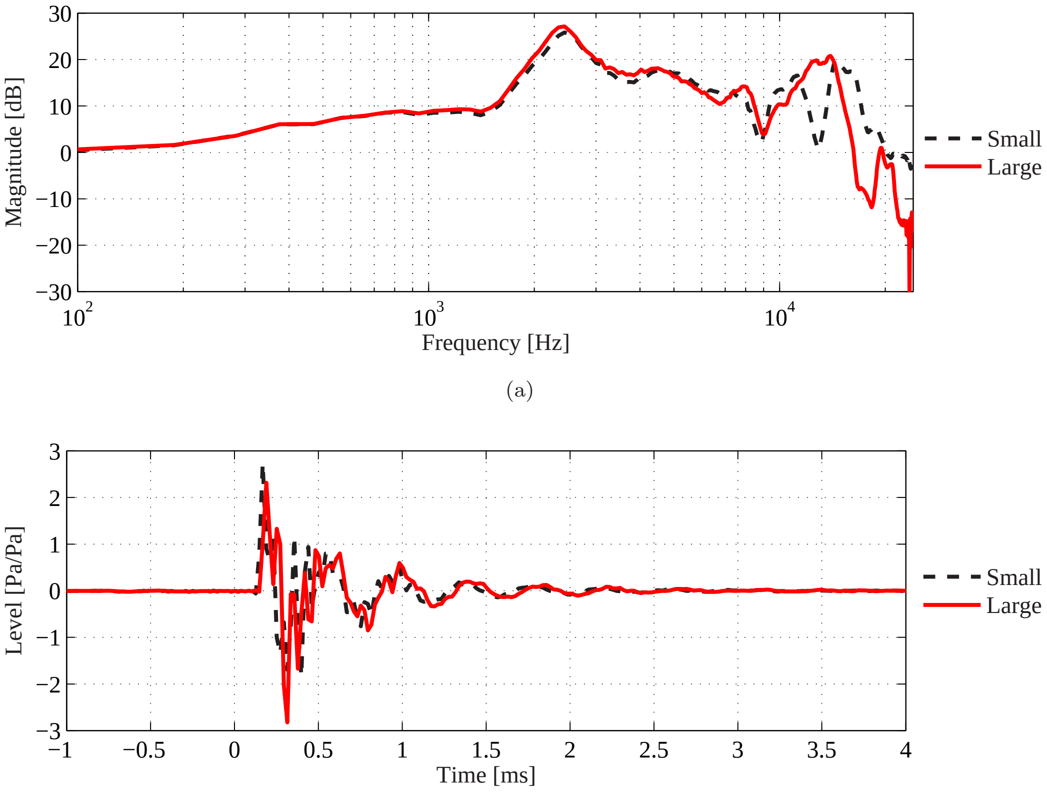


FIGURE 3.41: **HRIR in the frontal plane for the right blocked ear canal with "no pinna"**. *The time domain response is plotted as a function of elevation angle* ψ.



(b)

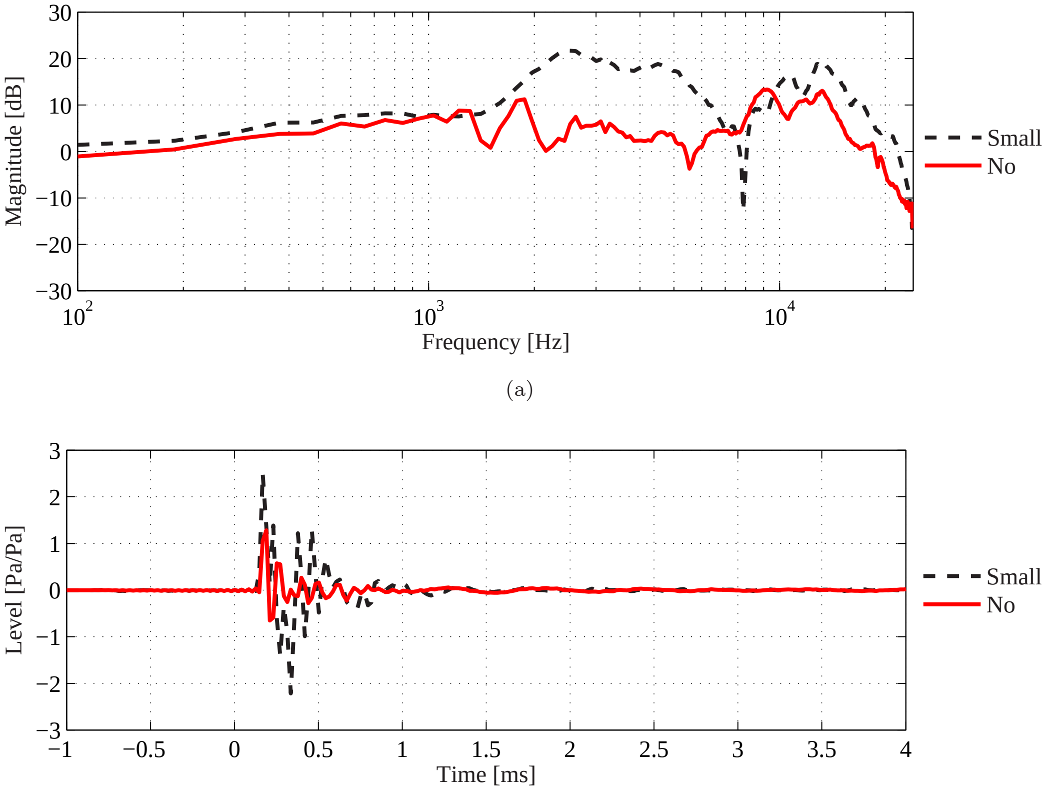
FIGURE 3.42: **Comparison of pinnae for the left ear**. *HRTF (a) and HRIR (b) the left open ear canal "small pinna" and "large pinna".*



(a)

(b)

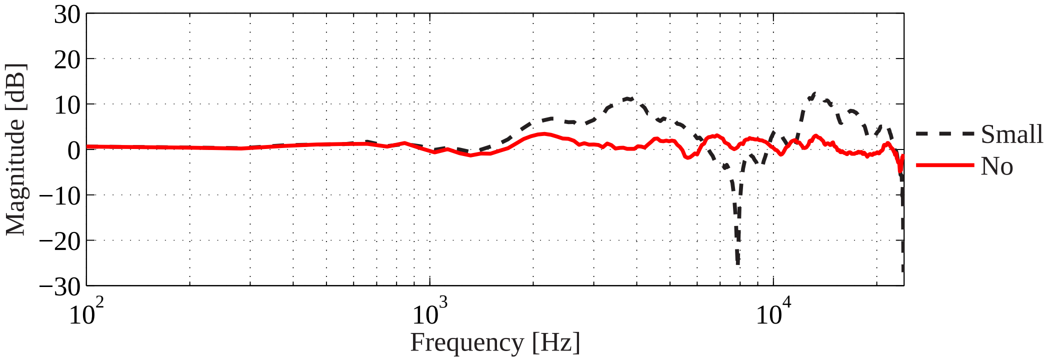
FIGURE 3.43: **Comparison of pinnae for the right ear**. *HRTF (a) and HRIR (b) for the right open ear canal "small pinna" and "large pinna".*



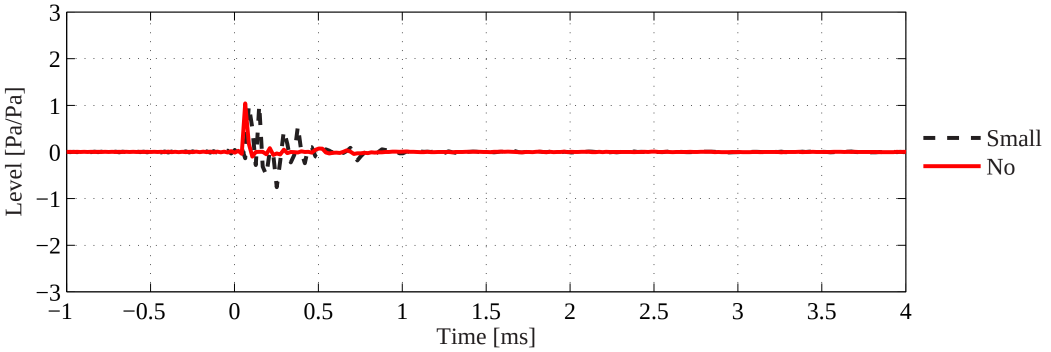
(a)

(b)

FIGURE 3.44: **Comparison of small pinna and no pinna for the left ear**. *HRTF (a) and HRIR (b) for the left open ear canal "small pinna" and "no pinna".*

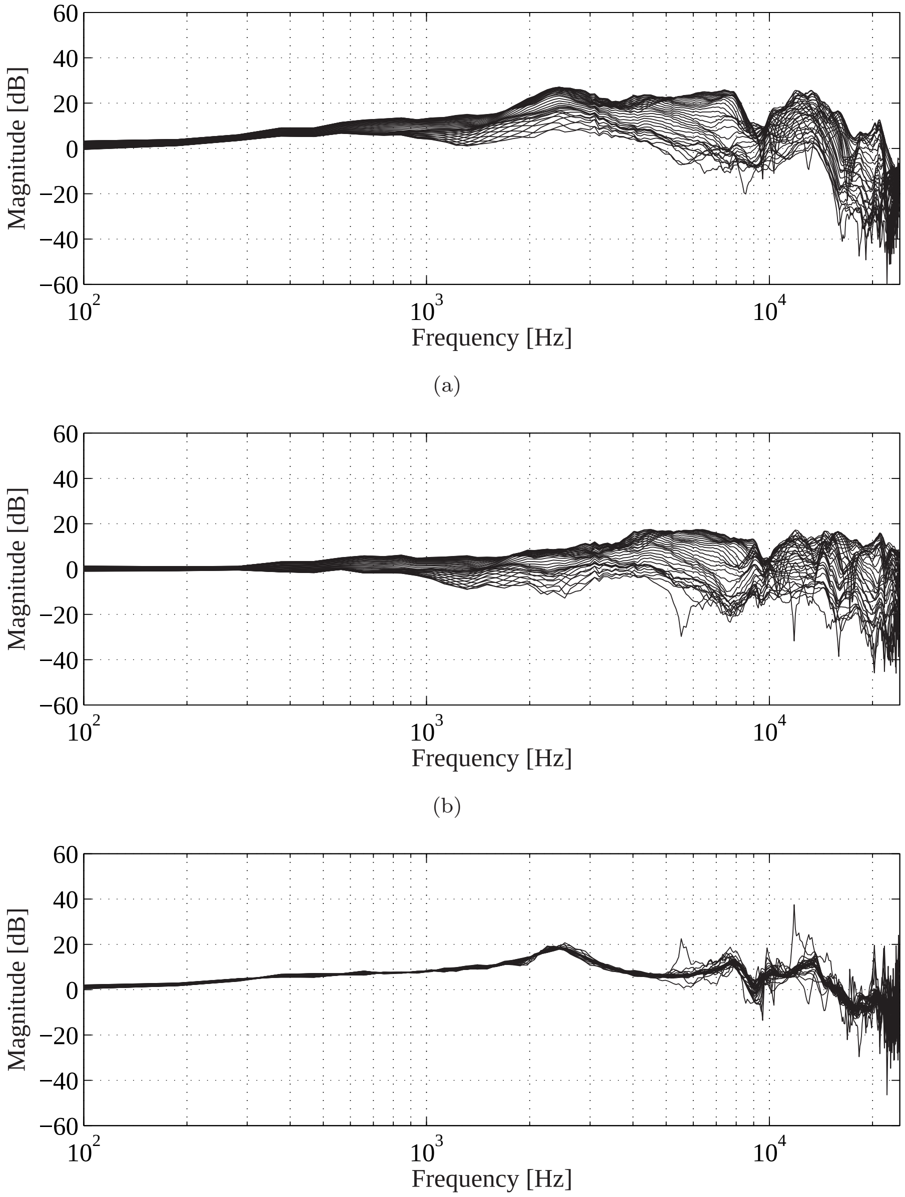


(a)



(b)

FIGURE 3.45: **Comparison of small pinna and no pinna for the left side with blocked ear canal**. *HRTF (a) and HRIR (b) for the left blocked ear canal "small pinna" and "no pinna".*

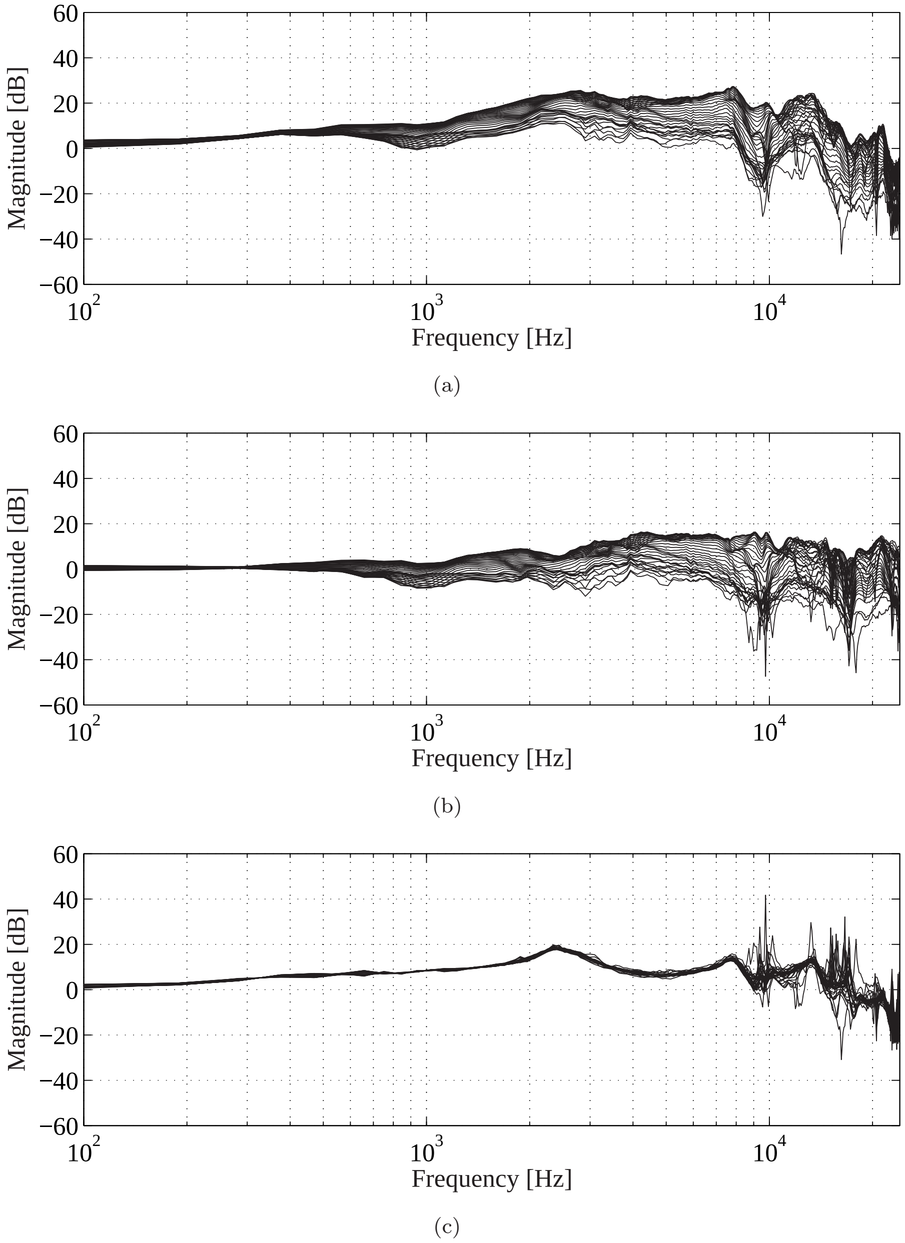


(a)

(b)

(c)

FIGURE 3.46: **Pressure divisions of HRTFs for the "large pinna" with the elevation angle** $\psi = 0°$. *(a) Open ear canal. (b) Blocked ear canal. (c) Pressure division between the open and blocked ear canals.*

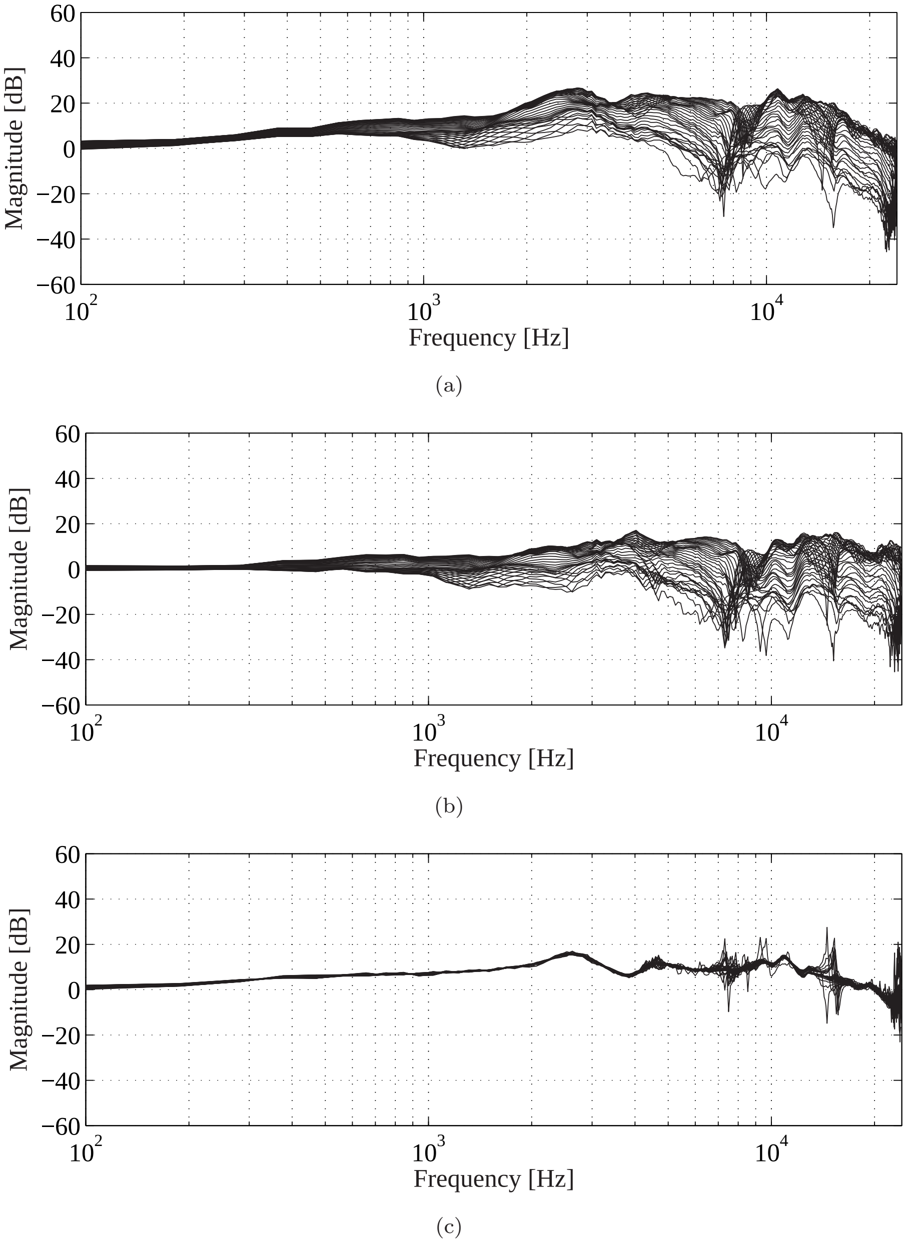


(a)

(b)

(c)

FIGURE 3.47: **Pressure divisions of HRTFs for the "large pinna" with the elevation angle** $\psi = 30°$. *(a) Open ear canal. (b) Blocked ear canal. (c) Pressure division between the open and blocked ear canals.*

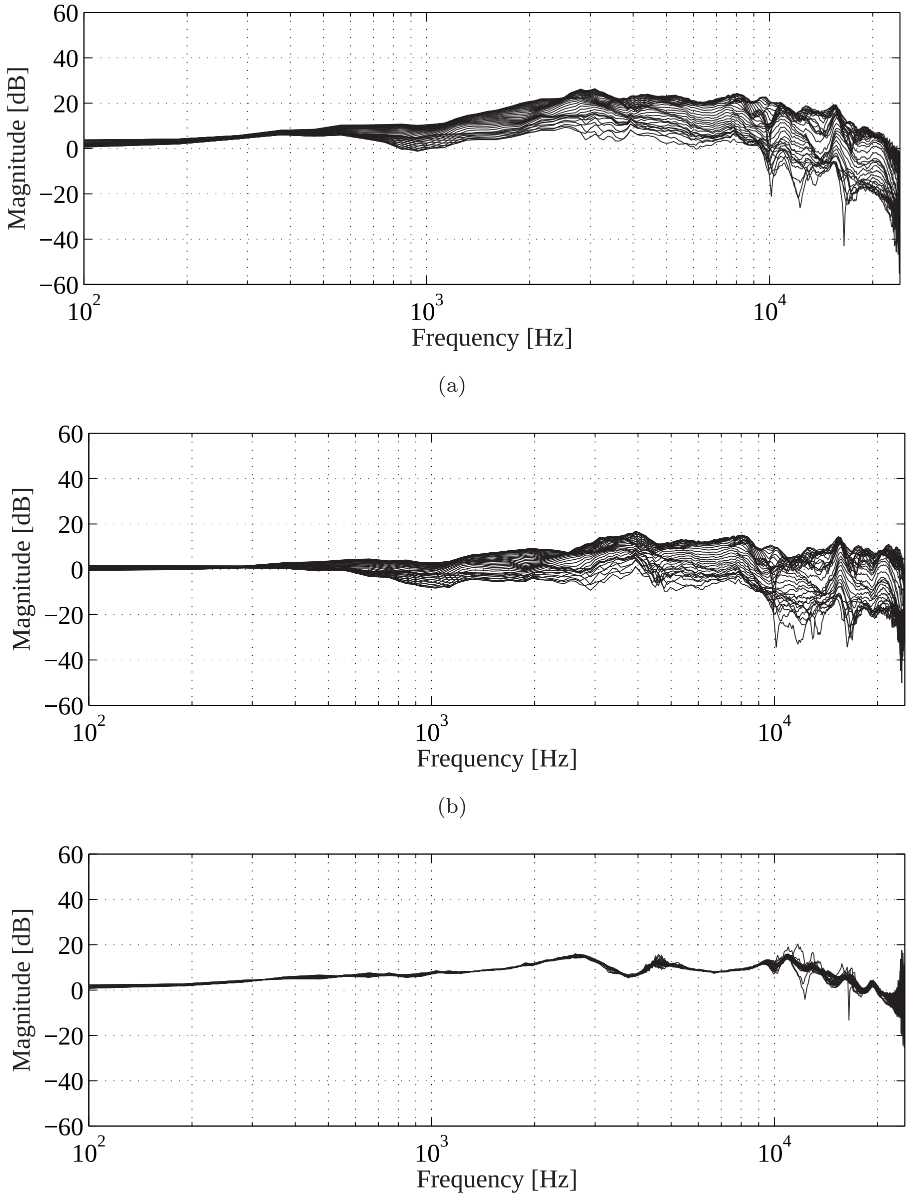


(a)

(b)

(c)

FIGURE 3.48: **Pressure divisions of HRTFs for the "small pinna" with the elevation angle** $\psi = 0°$. *(a) Open ear canal. (b) Blocked ear canal. (c) Pressure division between the open and blocked ear canals.*

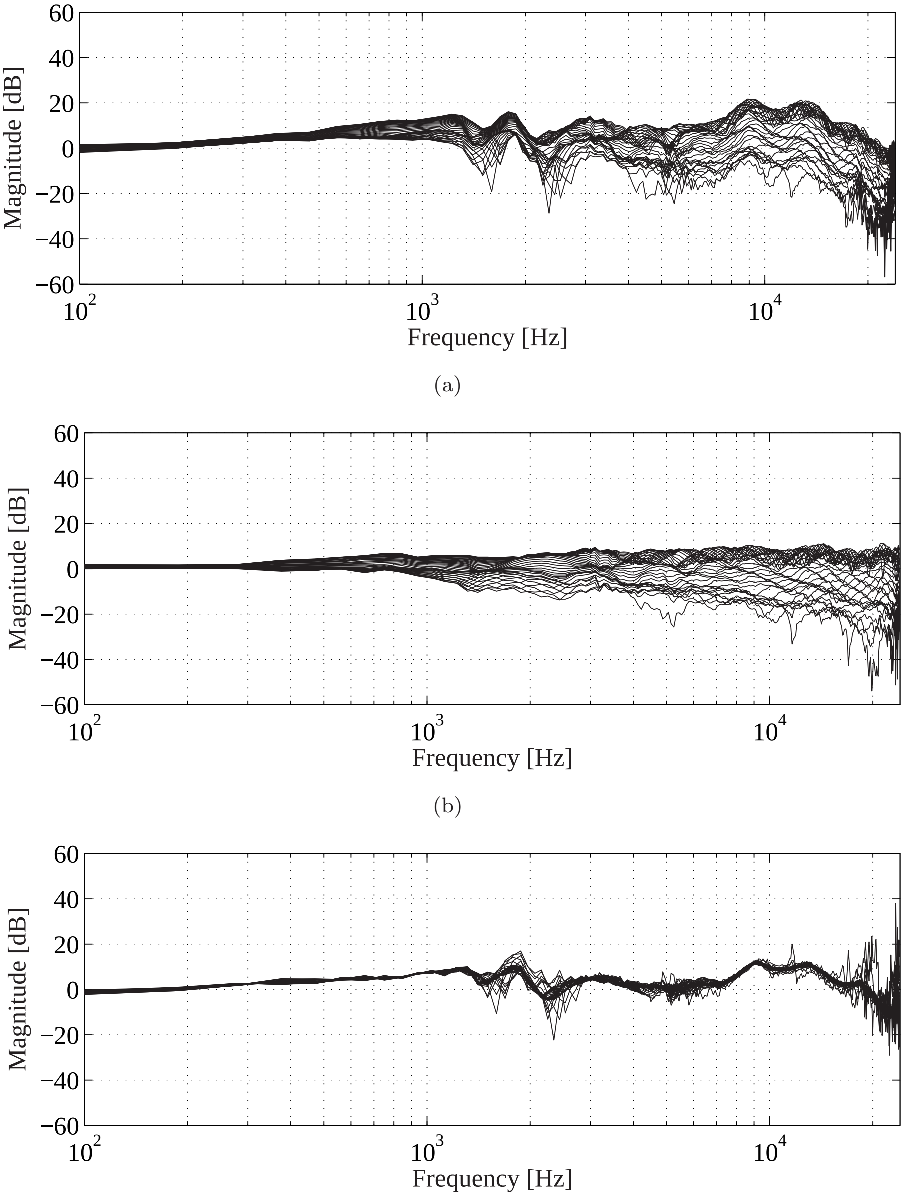


(a)

(b)

(c)

FIGURE 3.49: **Pressure divisions of HRTFs for the "small pinna" with the elevation angle** $\psi = 30°$. *(a) Open ear canal. (b) Blocked ear canal. (c) Pressure division between the open and blocked ear canals.*

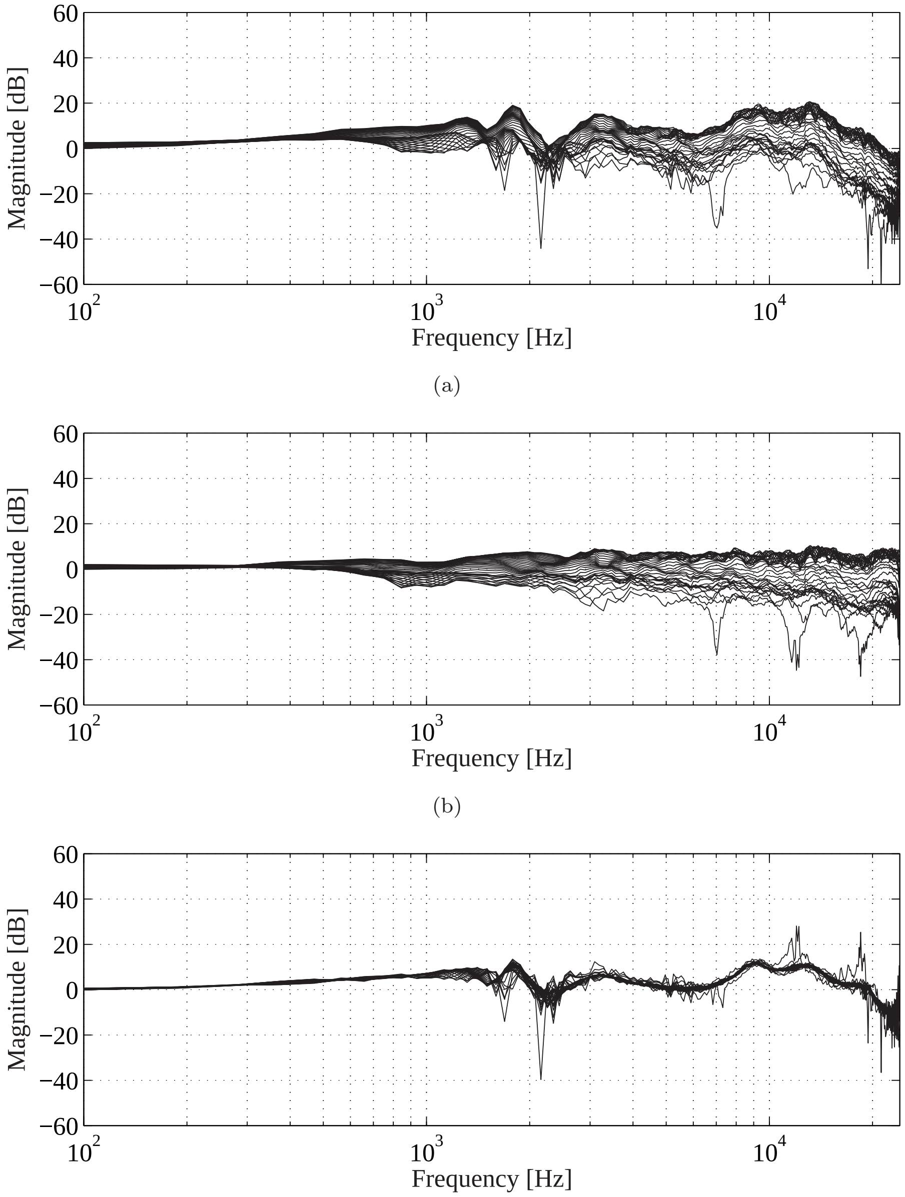


(a)

(b)

(c)

FIGURE 3.50: **Pressure divisions of HRTFs for the "no pinna" with the elevation angle** $\psi = 0°$. *(a) Open ear canal. (b) Blocked ear canal. (c) Pressure division between the open and blocked ear canals.*



(a)

(b)

(c)

FIGURE 3.51: **Pressure divisions of HRTFs for the "no pinna" with the elevation angle $\psi = 30°$.** *(a) Open ear canal. (b) Blocked ear canal. (c) Pressure division between the open and blocked ear canals.*

Chapter 4

Interpolation of head related impulse response functions

The filter update problem requires interpolation between HRIRs. A technique for linear interpolation between mixed-phase HRIRs is presented in this chapter. The presented interpolation scheme uses ITD-equalisation and linear interpolation. The HRIR interpolation is here performed in the time domain between neighbouring angular locations. The presented technique uses thresholding for the ITD-equalisation and is compared to results presented by previous researchers. The ISVR-Samsung database that was presented in the previous chapter is used in the objective evaluation carried out here.

Previous researchers have examined a number of interpolation techniques for HRTFs. Chen [18] proposes a feature extraction method, where the HRTF was represented as a weighted sum of Eigen transfer functions (EFs) generated by a Karhunen-Loeve expansion (KLE). This approach, results in an error of less than 1% for interpolation of mixed-phase HRTFs. Evan [28] presents a model where the HRTFs are represented as a weighted sum of surface spherical harmonics (SSHs). The interpolation performance is similar to the EF technique presented Chen [18]. However, the computational efficiency of the SSH technique is inferior compared to the EF technique. Various interpolation techniques including minimum-phase approximations are presented by Jot [44]. Takeuchi [95] suggests a frequency domain method for interpolating HRTFs using magnitude and phase interpolation. The last two mentioned studies do not include any measure of the percentage error that their techniques imposes. The technique presented by Takeuchi [95] is promising but there is no algorithm that wraps the phase correctly as required included in this investigation. The interpolation technique introduced here that uses thresholding is a robust and simple way to increase the resolution of a HRIR database. The main advantage of this technique compared to Chen [18] is the inherent simplicity.

There are three general techniques for extracting ITD information (Huopaniemi [38]) from HRIRs namely cross-correlation, thresholding (also known as leading-edge) and interaural phase difference (IPD). The cross-correlation and thresholding techniques are useful for deriving frequency independent ITD, while the IPD technique can include frequency dependent ITD values. The thresholding technique is evaluated in this Chapter.

4.1 Onset detection by thresholding

An interpolation algorithm that uses onset detection by thresholding for ITD-equalisation is introduced here. The ITD in the database is extracted by first using up-sampling followed by onset detection using thresholding. Ten times up-sampling has proven to be a good compromise between performance and computational effort. The up-sampling is applied to generate a higher resolution HRIR, which increases the precision in the time-alignment step. The thresholding step is performed in order to time-align the HRIRs. The thresholding is performed by normalising all the HRIRs and detecting the starting point of the HRIR at a threshold level of 0.1 of the maximum response. Now linear interpolation is applied in between the discrete measurement points for the time aligned HRIR database. The interpolation can either be applied in the time domain or in the frequency domain on the complex valued frequency response with equivalent results. The delay that was removed in the thresholding step is inserted back again and down-sampling is carried out to retrieve the database to its original size. It should be noted that the interpolation step can also be applied after down-sampling as long as the database is still time-aligned.

The sample rate conversion (up-sampling and down-sampling) algorithms that have been used here are described by Crochiere [23]. The original HRIR is denoted $c(n)$ and the up-sampled HRIR is denoted $y(m)$. The sampling rate is increased with an integer factor L and the new sampling period $T^{'}$ is given by

$$\frac{T^{'}}{T} = \frac{1}{L} \tag{4.1}$$

and the new sampling rate $F^{'}$ is given by

$$F^{'} = LF \tag{4.2}$$

The up-sampling of the impulse response $c(n)$ by L implies that we must interpolate $L-1$ new sample values between each pair of sample values of $c(n)$. Figure 4.1 shows a diagram of the up-sampling process. The impulse response $c(n)$ is zero padded with $L-1$ zero valued samples between each pair of samples of $c(n)$, resulting in the following signal

$$w(m) = \begin{cases} c\left(\frac{m}{L}\right), & m = 0, \quad \pm L \quad \pm 2L \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

Now the spectrum of $w(m)$ contains not only the baseband frequencies of interest ($-\pi/L$ to π/L) but also the images of baseband centered at harmonics of the original sampling frequency $\pm 2\pi/L, \pm 4\pi/L, ...$ To recover the baseband signal of interest and eliminate unwanted image components it is necessary to filter the signal $w(m)$ with a digital low-pass filter

$$H(e^{j\omega}) = \begin{cases} G & |\omega^{'}| \leq \frac{2\pi FT^{'}}{2} = \frac{\pi}{L} \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

In order to ensure that the amplitude of $y(m)$ is correct, the gain of the filter G, must be L in the pass-band. Thus critically the impulse response of the filter $H(e^{j\omega})$ as $h(k)$ shows that $y(m)$ is given by

$$y(m) = \sum_{k=-\infty}^{\infty} h(m-k)w(k) \tag{4.5}$$

Combining Equation 4.3 and Equation 4.5 gives the time domain input to output relationship for the up-sampling algorithm

$$y(m) = \sum_{k=-\infty}^{\infty} h(m-k)c\left(\tfrac{k}{L}\right), \quad k/L \quad \text{an} \quad \text{integer} \tag{4.6}$$

The up-sampled filter $y(m)$ is normalised so that the maximum value of the filter is equal to one.

$$y_n(m) = y(m)/y^{max}(m) \tag{4.7}$$

This is performed so that onset detection can be performed, using the same threshold for all filters. The onset is the first peak that exceeds the threshold $y(m) > \delta$ where δ is a positive constant and $y(m)$ is the up-sampled impulse response of the filter. A threshold level of $\delta = 0.1$ has proven to work well with the ISVR-Samsung HRTF database. The onset point of the filter can be reduced with a suitable number of samples in order to compensate for the small delay that the thresholding method imposes. In this case with $L = 10$ the onset of the filter was reduced with 20 samples of $y(m)$.

The equations above are applied to all HRIRs in the database in order to time-align the database. Then the interpolation stage is applied. The interpolation method used for the simulations in this chapter is linear interpolation in the time domain. The equation for interpolation in azimuth angle is given below

$$y_c(m) = (1 - k_\phi)y_a(m) + k_\phi y_b(m) \tag{4.8}$$

where $y_a(m)$ and $y_b(m)$ are $y_c(m)$'s two neighboring data points. The azimuth interpolation coefficient k_ϕ is given by $(\phi \bmod \phi_{re})/\phi_{re}$ where ϕ is the desired azimuth angle and ϕ_{re} is the sampling resolution of the database. The interpolation coefficient should not be confused with the discrete sample k or wavenumber k. The interpolation process can be extended to the three dimensional case by 4-point bilinear interpolation (Begault [11], Huopaniemi [38]).

$$y_f(m) = (1 - k_\phi)(1 - k_\psi)y_a(m) + k_\psi y_b(m) + k_\phi d_\psi h_c(m) + (1 - k_\phi)k_\psi h_e(m) \tag{4.9}$$

where $h_a(m)$, $h_b(m)$, $h_d(m)$ and $h_e(m)$ are $y_f(m)$ four neighbouring data points. The elevation interpolation coefficient k_ψ is given by $(\psi \bmod \psi_{re})/\psi_{re}$ where ϕ is the desired azimuth angle and ϕ_{re} is the sampling resolution of the database with respect to elevation.

The delay that was removed in the thresholding step is inserted back if the application requires the delay to be included in the database. Finally the down-sampling step is performed. The process of reducing the filter $y(m)$ by integer factor M is

$$\frac{T}{T'} = \frac{M}{1} \tag{4.10}$$

The new sampling rate is

$$F = \frac{1}{T} = \frac{F'}{M} \tag{4.11}$$

In order to lower the sampling rate and to avoid spatial aliasing at this lower rate, it is necessary to filter the signal $y(m)$ with digital a low pass filter having the frequency response function

$$H(e^{j\omega}) = \begin{cases} 1, & |\omega| \leq \frac{2\pi FT'}{2} = \frac{\pi}{M} \\ 0, & \text{otherwise} \end{cases} \tag{4.12}$$

The sampling rate reduction is achieved by forming the sequence $c'(n)$ by saving only every M^{th} sample of the filtered output. Thus

$$s(m) = \sum_{k=-\infty}^{\infty} h^{'}(k)y(m-k) \tag{4.13}$$

where $s(m)$ is the filtered output as illustrated in Figure 4.2 and the final output $c^{'}(n)$ is

$$c^{'}(n) = s(Mn) \tag{4.14}$$

An overview of interpolation process is illustrated by a block diagram in Figure 4.3.

4.2 Objective evaluation

The presented interpolation techniques have been investigated in the horizontal plane and in the median plane. The same coordinate system that was used in the previous chapter in Figure 3.4 also applies here. The angular resolution of the ISVR database is $\phi_{re} = 5°$ in the horizontal plane and $\psi_{re} = 10°$ in the median plane. Therefore it was possible to compare interpolated azimuth angles ϕ of 5° resolution and elevation angles ψ of 10° resolution. Then three databases of the horizontal plane with a reduced angular spacing were created with the following resolutions, 40°, 20° and 10°. The reduced databases with a grid of 40°, 20° and 10° were interpolated down to the finest grid of the original measured database of 5° resolution. The performance of the interpolated HRIRs in the horizontal plane could then be compared to the original HRIRs. In a similar manner, the interpolation was evaluated in the median plane for elevation angle spacings of 40° and 20° and the end result was compared to the original 10° elevation angle resolution.

The performance of the interpolation technique is evaluated by percentage mean square error. The mean square error in percentage is defined as

$$\zeta = \frac{\sum_n (c(n) - c^{'}(n))^2}{\sum_n c(n)^2} \times 100 \tag{4.15}$$

where $c(n)$ and $c^{'}(n)$ denote the original and the interpolated HRIRs.

4.2.1 Results

The results from the onset detection by thresholding technique are presented in Figure 4.4 - Figure 4.5 and in Figure 4.6. The interpolation results from the reduced databases with an angular resolution in azimuth of 10°, 20° and 40° are compared to the original database of 5° azimuth angle resolution, and are presented in Figure 4.4 and Figure 4.5. Again a single HRTF angle of azimuth $\phi = 5°$ and elevation $\psi = 0°$ is presented. The error in magnitude response is less than ±1 dB up to 20000 Hz for the reduced database with an azimuth angle resolution of 10°. The error in magnitude response is less than ±1 dB up to 7000 Hz for the reduced database with an azimuth angle resolution of 20°. Likewise, the error in magnitude response is less than ±1 dB up to 7000 Hz for the reduced database with an azimuth angle resolution of 40°. Figure 4.6 shows a comparison of mean square error for interpolated data from the reduced databases compared to the original database.

The average mean square error that results from interpolation in the horizontal plane of the reduced databases are presented in Table 4.1. The mean square error increases with a coarser grid for the databases as can be expected. Head shadowing also reduces the interpolation performance significantly, which may be associated with relatively low signal to noise ratio and difficulties in detecting the start of the impulse responses.

The results from HRTF interpolation methods presented by two previous researchers Chen [18] and Evans [28] are presented in Table 4.2. These results can partly be compared to the results found here. The main difference in data presentation is that in Table 4.1, the data comes from the horizontal plane with $\psi = 0°$ while Chen [18] and Evans [28] presents data from the horizontal plane with $0° \leq \psi \leq 90°$. The interpolation performance of the thresholding technique compared to the results in Table 4.2 is similar for frontal angles and even better for the unshadowed side and the back, while for the shadowed side the performance is significantly lower.

The average mean square error in the median plane from the interpolation techniques of the reduced databases are presented in Table 4.3.

Technique	Front	Shadowed side	Back	Unshadowed side
	$\phi \leq 45°; \phi \geq 315°$ [%]	$45° \leq \phi \leq 135°$ [%]	$135° \leq \phi \leq 225°$ [%]	$225° \leq \phi \leq 315°$ [%]
Th				
10° to 5°	0.52	7.52	0.34	0.25
20° to 5°	1.74	22.39	2.39	1.49
40° to 5°	8.82	51.09	11.58	9.77

Table 4.1: **Evaluation of the interpolation technique in the horizontal plane**. *Average Mean Square Error (MSE) for the interpolation technique named, onset detection by thresholding (Th). The results from the Th technique is presented in the horizontal plane with* $\psi = 0°$. *The results are presented for three different cases when the horizontal plane resolution of the original database is* 10°, 20° *and* 40°, *in all cases the interpolation is performed down to* 5° *resolution.*

Technique	Front $\phi \leq 45°; \phi \geq 315°$ [%]	Shadowed side $45° \leq \phi \leq 135°$ [%]	Back $135° \leq \phi \leq 225°$ [%]	Unshadowed side $225° \leq \phi \leq 315°$ [%]
Chen et al (1995)	0.52	2.06	0.55	0.56
Evans et al (1998)	0.47	0.93	0.78	0.35

Table 4.2: **Comparison of previously investigated interpolation techniques**. *Average MSE for two interpolation techniques namely, the EF-model by Chen [18] and the SSH-model by Evans [28]. The results from the EF-model and the SSH-model are above the horizontal plane* $0° \leq \psi \leq 90°$*. The data is derived from Chen et al. [18] Table III and from Evans et al. [28] Table VI.*

Technique	Above horizontal $0° \leq \psi \leq 90°$ [%]	Below horizontal $-40° \leq \psi \leq 0°$ [%]
Th		
20° to 5°	1.01	3.51
40° to 5°	5.11	8.51

Table 4.3: **Evaluation of the interpolation technique in the median plane**. *Average MSE for two interpolation technique named, onset detection by thresholding (Th). The results are from the median plane with* $\phi = 0°$*. The results from the Th is presented for two different cases when the median plane resolution of the original database is* 20°*, and* 40°*, in both cases the interpolation is performed down to* 10° *resolution.*

4.3 Conclusion

A technique for interpolation between HRIRs have been evaluated and compared to previously presented techniques. The onset detection by thresholding technique uses thresholding for ITD-equalisation. The average percentage mean square error for the onset detection by thresholding algorithm is 0.52% (for the ISVR-Samsung HRTF database) without head shadowing, when head shadowing occurs then the error increases to about 7.52%. This interpolation algorithm has proven to be a useful tool for the inverse filter design carried out in the next chapter and also for the filter update algorithms and subjective experiments presented in Chapter 6.

The spacing of the HRIR database can be reduced to 20° in the horizontal plane when using time-aligned database and when an average mean square error of 2.39% is acceptable for non shadowed angles in the horizontal plane with 0° elevation angle. For a full assessment on the effectiveness of the investigated techniques, psychoacoustic validation by listening tests is required. Also it would be interesting to compare the computational complexity and memory requirement of different interpolation methods.

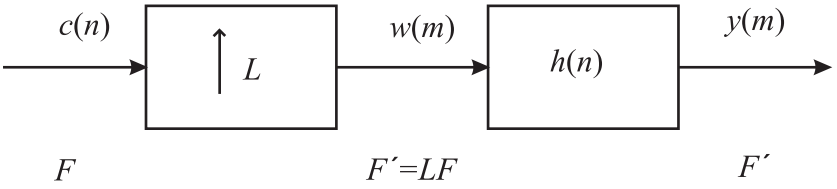


Figure 4.1: **Block diagram of the up-sampling algorithm**.

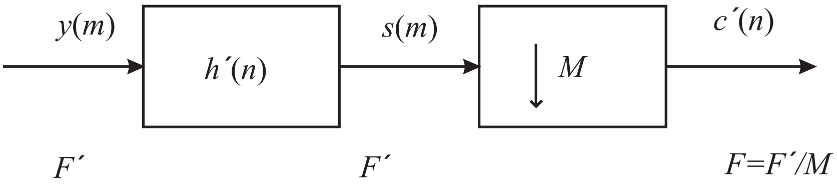


Figure 4.2: **Block diagram of the down-sampling algorithm**.

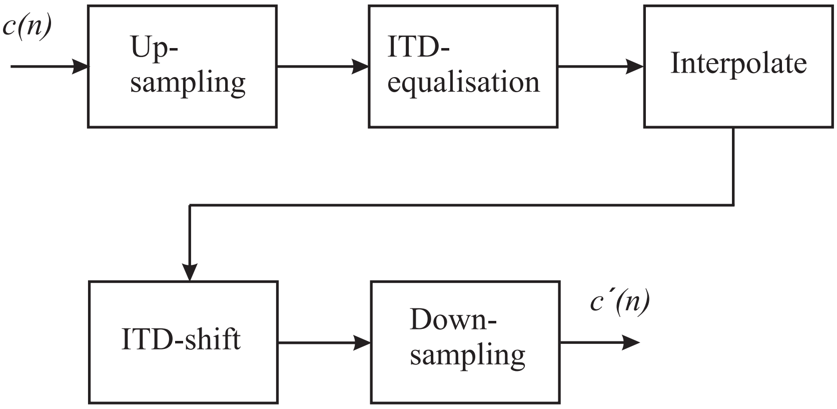
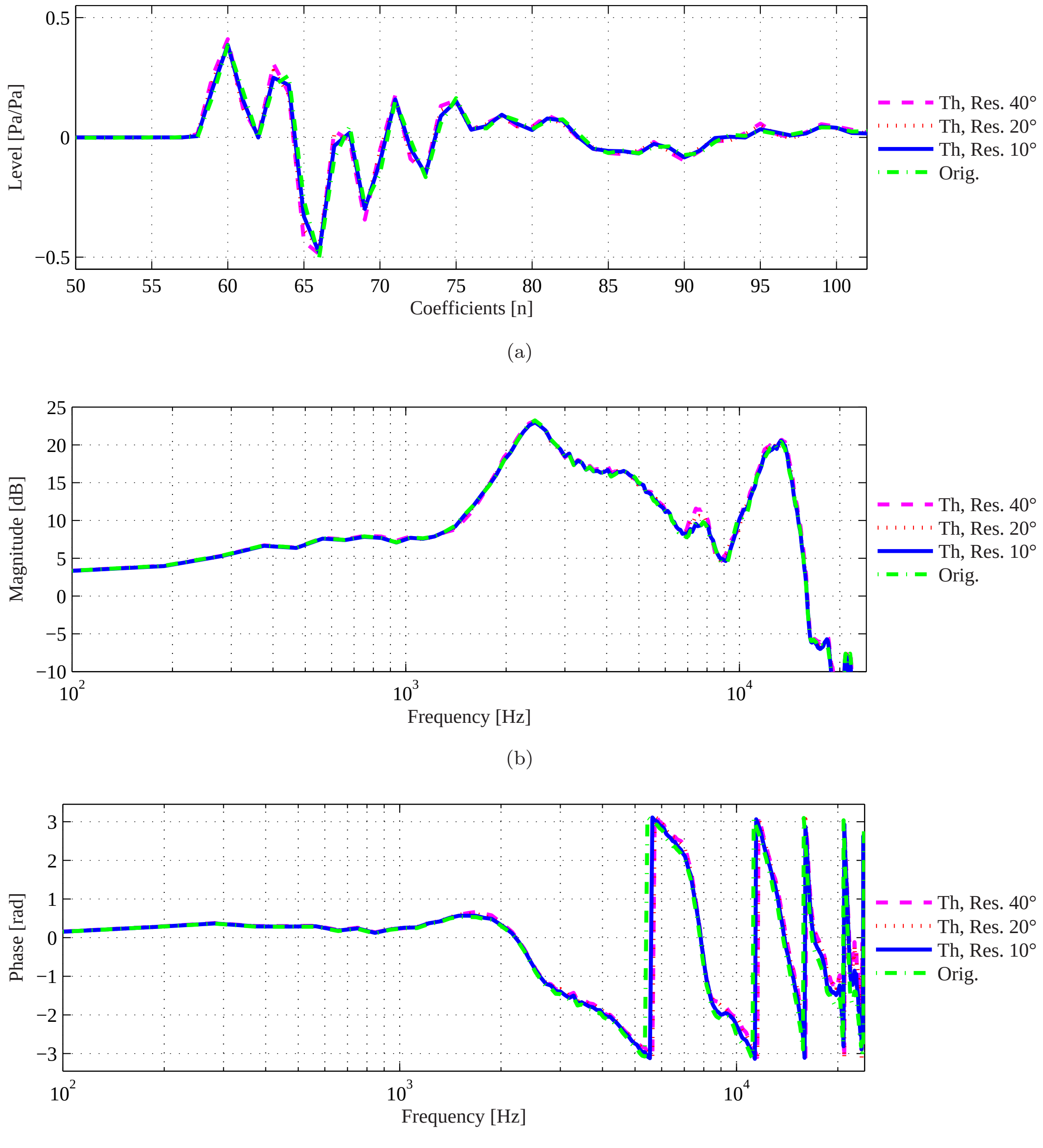


Figure 4.3: **Block diagram of the interpolation process**. *The only difference between the thresholding and cross-correlation algorithms is in the ITD-equalisation step.*

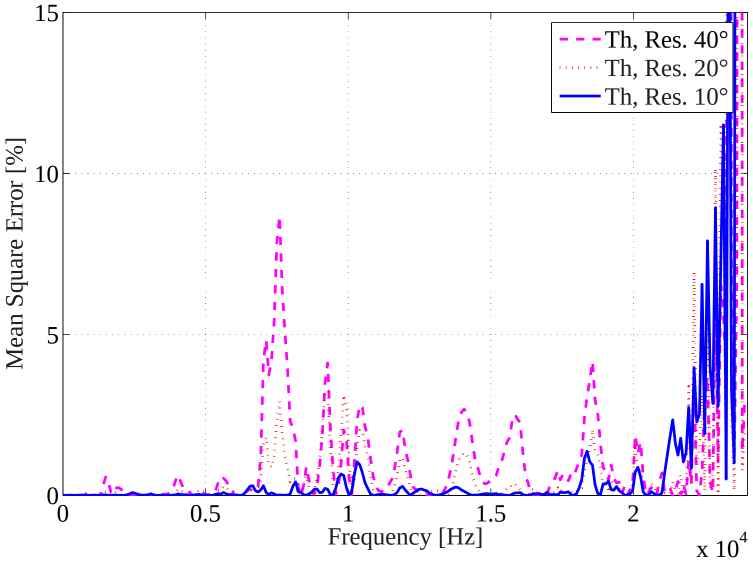


(a)

(b)

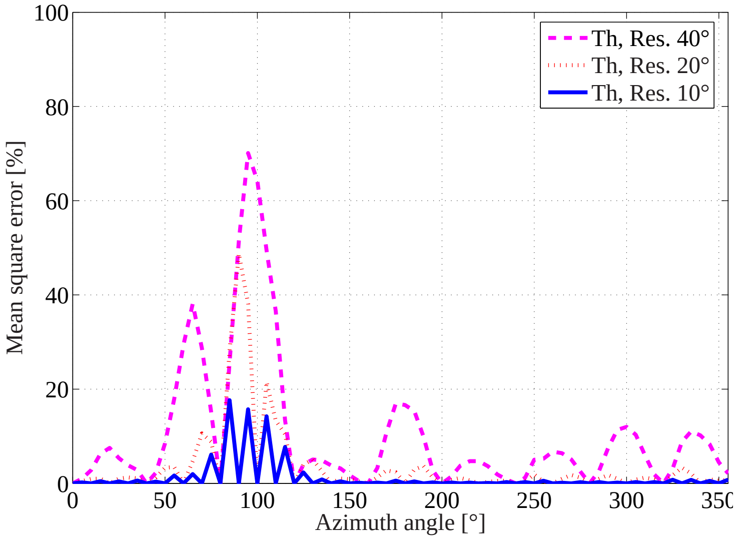
(c)

FIGURE 4.4: **Onset detection by thresholding**. *Interpolation in time domain using onset detection by thresholding. A database of 10° angular resolution in the horizontal plane was interpolated down to 5° resolution and compared to the original database of 5° resolution. The database of 40° and 20° angular resolution were made in a similar way. The presented HRTF is for azimuth angle $\phi = 5°$ and elevation angle $\psi = 0°$. (a) Time response. (b) Magnitude of frequency response (c) Phase response .*



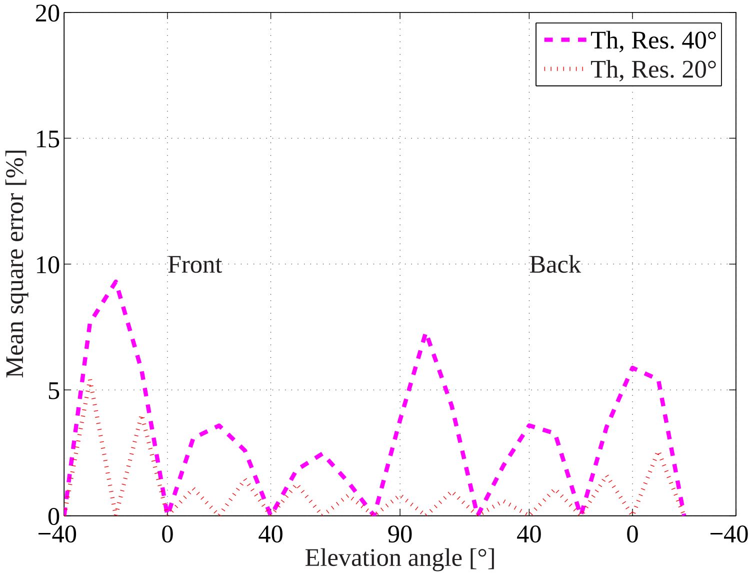
(a)

FIGURE 4.5: **Onset detection by thresholding**. *MSE for interpolation in time domain using onset detection by thresholding. A database of 10° angular resolution in the horizontal plane was interpolated down to 5° resolution and compared to the original database of 5° resolution. The database of 40° and 20° angular resolution were made in a similar way. The presented HRTF is for azimuth angle* $\phi = 5°$ *and elevation angle* $\psi = 0°$. .



(a)

Figure 4.6: **Interpolation in the horisontal plane using onset detection by thresholding**. *Mean Square Error (MSE) for a time domain interpolation technique in the horizontal plane. The interpolation is performed in the horizontal plane from* $\phi = 0°$ *to* $\phi = 355°$.



(a)

Figure 4.7: **Interpolation in the median plane using onset detection by thresholding**. *MSE for two time domain interpolation techniques in the median plane. The interpolation is performed in the median plane .*

Chapter 5

Filter design and the effectiveness of cross-talk cancellation

This chapter presents filter design techniques and an objective evaluation of cross-talk cancellation effectiveness for three virtual sound imaging systems, namely the SD, 2-way OSD and the 3-way OSD. It is shown how the geometry affects the filter design process and how to design the cross-talk cancellation networks using a minimum number of inverse filters. The filter design for OSD systems can be simplified by using superposition. This gives the advantage of reducing the number of inverse filters for the OSD to the same number as for the SD. The objective evaluation is carried out by simulating cross-talk cancellation effectiveness for different filter lengths and for asymmetric and symmetric listener positions of static and adaptive systems. The evaluation of cross talk cancellation effectiveness for different filter lengths aims to give an idea of which filter length to choose for a certain performance requirement. The benefit of using an adaptive virtual sound imaging system compared to a static virtual sound imaging system is demonstrated. The effectiveness of cross talk cancellation when using a static system is restricted to a relatively small area named the "sweet-spot". The improvements in area size where the listener can move, using an adaptive approach and steering the associated sweet-spot, is compared to the static case.

The evaluation is performed by simulations using measurement data from the anechoic chamber at the ISVR. The transfer functions used in the simulations both for the adaptive and the static systems include the HRTF and the loudspeaker response. This aims to evaluate the cross talk cancellation effectiveness under anechoic conditions. Cross-overs are included in the filter length evaluation but not included in the simulations of adaptive and the static systems. The simulations of adaptive and the static systems are performed for discrete sources and then the frequency band of each source are plotted together to show the full frequency band. The effect of fixed and adaptive cross-overs for adaptive systems are not included in the objective evaluation presented here.

The theory of cross-talk cancellation was introduced by Bauer [10] in an analogue implementation and followed up by Atal and Schroeder [4]. More recently, cross-talk cancellation techniques has been investigated by a number of researchers such as, Moller [68], Cooper and Bauck [20], Gardner [30], Kirkeby and Nelson [50], Ward and Elko [100] and Takeuchi [95]. Ward and Elko [100] showed that the transfer function matrix to be inverted became ill-conditioned when the path-length difference between one of the loudspeakers and two of the ears of the listener became equal to one half of the acoustic wavelength. This led to the development of the SD (Kirkeby and Nelson [49]), which ensured a well-conditioned inversion problem over a particularly useful frequency range. It was pointed out by Ward and Elko [100] that for broadband signals, ideally the loudspeaker spacing should vary with frequency. This theory was extended further by Takeuchi [95], [97] who presented the OSD loudspeaker and conducted a set of subjective experiments that showed the improvements in virtual sound source localisation that can be achieved with this loudspeaker type. Previous studies did not include the theory for how to design cross-talk cancellation filters for an OSD system using a minimum number of inverse filters, and this is presented here.

Different methods of loudspeaker response equalization using digital filters are compared by Karjalainen [45]. The advantages of Warped Finite Impulse Response (WFIR) and Warped Infinite Impulse Response (WIIR) filters are shown in comparison with traditional filter structures and are found to give a reduced filter order and increased numerical robustness. The disadvantage is the higher computational complexity of the warped structures. The required filter length for efficient cross-talk cancellation has previously been investigated by Farina [29]. Farina has implemented WFIRs for cross-talk cancellation and demonstrated the low frequency benefits and filter order reduction of such an approach. This chapter compares the cross-talk cancellation performance for the SD, 2-way OSD and the 3-way OSD. The filter design technique used here is the one presented by Kirkeby [50] and can be implemented using Finite Impulse Response (FIR) filters.

A subjective investigation of inverse filtering is presented by Norcross [79]. The idea with inverse filtering is to undo the filtering caused by a system such as a loudspeaker or a room. The inversion of an impulse response might cause undesired audible artefacts and the severity of these artefacts is affected by the impulse response of the system and the method used to compute the inverse filter. The artefacts from the inverse filtering can some times degrade the overall signal quality instead of improving it. This study has presented formal subjective tests to investigate potential limitations associated with inverse filtering techniques. The investigated inverse filtering techniques are the time domain least square (LS)-technique and frequency domain methods including the fast-deconvolution algorithm, both by Kirkeby and Nelson [50], [48]. Their subjective results show that the investigated inverse filtering techniques can sometimes improve the subjective audio quality and in other cases degrade it. The subjective audio quality depends on the impulse response, the chosen filtering technique and on the inverse filter length.

The robustness to head misalignment for cross-talk cancellation using the converted (60° loudspeaker span) stereo configuration was investigated by Gardner [30] and the SD and Stereo configuration was investigated by Takeuchi [96]. Misalignment of the head position and orientation give rise to inaccurate synthesis of the binaural signals at the ears of the listener. This is because the change in transfer functions between the loudspeakers and listener's ears and the result is deterioration in system performance. To overcome the limitations in "sweet-spot" size Rose [86] investigated an adaptive SD system, where the change in transfer function is compensated for by adapting the cross-talk cancellation scheme to the position and orientation of the listener. This study includes objective and subjective evaluations of cross-talk cancellation effectiveness in the lateral plane for the SD. In this chapter an objective evaluation of cross-talk cancellation effectiveness is presented for the SD, 2-way OSD and the 3-way OSD for the lateral plane, fore and aft plane and listener rotation together with filter lengths dependency and filter design techniques.

5.1 Filter design for virtual sound imaging systems

The solution to the problem of using a minimum number of inverse filters when designing filters for a FDL is to use superposition. This results in the same number of filters for the FDL as for a two-channel loudspeaker system like the SD. For example, without using superposition in the filter design process one would need a set of filters per transducer span, hence the 3-way OSD would need three times more filters than the SD. The filter design process for binaural loudspeakers systems using the minimum number of inverse filters is described for the following three examples, namely the SD, the 2-way OSD and the 3-way OSD.

The impulse responses for the SD are found by taking the impulse responses from azimuth angles 355° and 5° for the left and the right ear respectively. The transfer functions ($C_{11}(k)$, $C_{12}(k)$, $C_{21}(k)$ and $C_{22}(k)$) and the source angle α_1 for the SD are presented in Figure 5.5. The transfer function matrix $\mathbf{C}(k)$ is given by

$$\mathbf{C}(k) = \begin{bmatrix} C_{11}(k) & C_{12}(k) \\ C_{21}(k) & C_{22}(k) \end{bmatrix} \tag{5.1}$$

and the inverse of $\mathbf{C}(k)$ is

$$\mathbf{H}(k) \approx \mathbf{C}^{-1}(k)e^{-j\omega\Delta} \tag{5.2}$$

The combined transfer functions ($D_{11}(k)$, $D_{12}(k)$, $D_{21}(k)$ and $D_{22}(k)$), cross-overs ($G_{11}(k)$, $G_{12}(k)$, $G_{21}(k)$ and $G_{22}(k)$) and the source angles (σ_1 and σ_2) for the 2-way OSD are presented in Figure 5.5. The transfer function paths including cross-overs can be added together using superposition (Pierce [83]). To find the combined transfer functions, the following equations are used

$$D_{11}(k) = D_{11,1}(k)G_{11}(k) + D_{11,2}(k)G_{12}(k) \tag{5.3}$$

$$D_{12}(k) = D_{12,1}(k)G_{21}(k) + D_{12,2}(k)G_{22}(k) \tag{5.4}$$

$$D_{21}(k) = D_{21,1}(k)G_{11}(k) + D_{21,2}(k)G_{12}(k) \tag{5.5}$$

$$D_{22}(k) = D_{22,1}(k)G_{21}(k) + D_{22,2}(k)G_{22}(k) \tag{5.6}$$

The combined transfer function matrix $\mathbf{D}(k)$ is given by

$$\mathbf{D}(\mathbf{k}) = \begin{bmatrix} D_{11}(k) & D_{12}(k) \\ D_{21}(k) & D_{22}(k) \end{bmatrix} \tag{5.7}$$

and the inverse of $\mathbf{D}(k)$ is

$$\mathbf{J}(k) \approx \mathbf{D}^{-1}(k)e^{-j\omega\Delta} \tag{5.8}$$

Using the combined transfer functions, results in less effort to compute the inverse transfer functions, since otherwise one would need to compute inverse filters for all the components in Equations 5.3 - 5.6. Without using superposition one would also need twice the number of filters in the convolution engine.

The combined transfer functions ($E_{11}(k)$, $E_{12}(k)$, $E_{21}(k)$ and $E_{22}(k)$), cross-overs ($F_{11}(k)$, $F_{12}(k)$, $F_{13}(k)$, $F_{21}(k)$, $F_{22}(k)$, $F_{23}(k)$) and the source angles (φ_1, φ_2 and φ_3) for the 3-way OSD are presented in Figure 5.3. The transfer function paths can be added together using superposition, using the same method as for the 2-way OSD. To find the combined transfer functions, the following equations are used

$$E_{11}(k) = E_{11,1}(k)F_{11}(k) + E_{11,2}(k)F_{12}(k) + E_{11,3}(k)F_{13}(k) \tag{5.9}$$

$$E_{12}(k) = E_{12,1}(k)F_{21}(k) + E_{12,2}(k)F_{22}(k) + E_{12,3}(k)F_{23}(k) \tag{5.10}$$

$$E_{21}(k) = E_{21,1}(k)F_{11}(k) + E_{21,2}(k)F_{12}(k) + E_{21,3}(k)F_{13}(k) \tag{5.11}$$

$$E_{22}(k) = E_{22,1}(k)F_{21}(k) + E_{22,2}(k)F_{22}(k) + E_{22,3}(k)F_{23}(k) \tag{5.12}$$

The combined transfer function matrix $\mathbf{E}(k)$ is given by

$$\mathbf{E}(k) = \begin{bmatrix} E_{11}(k) & E_{12}(k) \\ E_{21}(k) & E_{22}(k) \end{bmatrix} \tag{5.13}$$

and the inverse of $\mathbf{E}(k)$ is

$$\mathbf{K}(k) \approx \mathbf{E}^{-1}(k)e^{-j\omega\Delta} \tag{5.14}$$

Again, by using the combined transfer functions in Equations 5.9 - 5.12, results in less effort to compute the inverse transfer functions. In this case, without using superposition one would need triple the number of filters in the convolution engine.

5.2 Filter length dependency

The effectiveness of cross talk cancellation for different filter lengths is an important issue. Cross-talk cancellation effectiveness for the SD, 2-way OSD and the 3-way OSD using different filter lengths is evaluated under anechoic conditions. The data used in this objective evaluation is based on the raw impulse responses that were obtained during ISVR-Samsung HRTF database measurement. The raw impulse response contains 4096 coefficients and includes the loudspeaker response and the KEMAR dummy head HRTF from the "large pinna". The plant model that was used contains 3000 filter coefficients of the raw impulse responses. The cross-overs were designed using the Matlab [66] function **fir1**() that implements the classical method of windowed linear-phase FIR filters [40] (Hann-window based). The cross-over frequency for the 2-way OSD is at 900 Hz, which is illustrated in Figure 2.8. The cross-over frequencies for the 3-way OSD are at 4000 Hz and 600 Hz, which is illustrated in Figure 2.6. The inverse of the transfer functions were found by using the fast-deconvolution algorithm. The regularisation factor β in the fast-deconvolution algorithm was set to 10^{-4} according to suggestions made by Kirkeby [50]. The same β value was used for all the three virtual sound imaging systems and only the number of coefficients in the filter was varied. Inverse filters with different number of coefficients N (32, 64, 128, 256, 512, 1048, 2048, 4096, 8192, 16384) were designed. The performance of the inverse filters with different numbers of coefficients was evaluated by simulating cross-talk cancellation effectiveness for the three virtual sound imaging systems.

The transfer functions and inverse transfer function are defined as in Section 5.1. The SD is associated with $\mathbf{C}(k)$, the 2-way OSD is associated $\mathbf{D}(k)$ and 3-way OSD is associated $\mathbf{E}(k)$. In the same manner, the inverse transfer functions for the SD, 2-way OSD and the 3-way OSD are defined as $\mathbf{H}(k)$, $\mathbf{J}(k)$ and $\mathbf{K}(k)$ respectively. For the SD, the first N_i coefficients of the matrix of transfer functions $\mathbf{C}(k)$ is named $\mathbf{C}_{N_i}(k)$ and the inverse matrix of $\mathbf{C}_{N_i}(k)$ is named $\mathbf{H}_N(k)$. The inverse transfer functions $\mathbf{H}_N(k)$ are found by zero padding $\mathbf{C}_{N_i}(k)$ up to $N = 4N_i$ coefficients and then using the fast-deconvolution algorithm. The number of coefficients of the impulse response N_i was varied (8, 16, 32, 64, 128, 256, 512, 1024, 2048). The inverse filter with $N = 16384$ was calculated using $N_i = 2048$ to make sure that N_i was not longer than the plant model of 3000 coefficients. The reason to choose $N = 4N_i$ was to minimize “wrap around” effects in the FFTs when carrying out fast-deconvolution. The plant model $\mathbf{C}_{N_p}(k)$ that was used contains $N_p = 3000$ coefficients of the transfer functions with a Hann-window applied, in order to suppress noise in the late response. The plant model aims to give a realistic objective evaluation of the system. The matrix of transfer functions, inverse matrix and plant model for the 2-way OSD and 3-way OSD are found in the same way as described for the SD.

The results from simulated cross-talk cancellation performance using measured transfer functions are presented for inverse filters with different number of coefficients. The simulated cross-talk performance is plotted in Figure 5.4 - Figure 5.9. Figure 5.4, Figure 5.5 and Figure 5.6 illustrates the control performance in the frequency domain $P_{11}(k)$ and $P_{21}(k)$ at the ears, as specified in Equation 2.30. The control performance in the frequency domain is also summarised in Table 5.1. Figure 5.7, Figure 5.8 and Figure 5.9 illustrates the control performance in the time domain $p_{11}(n)$ and $p_{21}(n)$ at the ears, as specified in Equation 2.30 and Equation 2.31. The impulse responses $p_{11}(n)$ and $p_{21}(n)$ are plotted using the decibel scale.

The channel separation parameter is used as a measure of cross-talk cancellation effectiveness. The channel separation in dB is defined as the ratio of the magnitude of the control performance at the ears.

$$S(k) = 20\log_{10}\left(\frac{|P_{11}(k)|}{|P_{21}(k)|}\right) \tag{5.15}$$

The bandwidth dependence of the number of filter coefficients is presented in Table 5.1. The bandwidth is here found by defining a minimum required channel separation of $S(k) > 20$ dB. It can be seen that the upper limit of the 2-way OSD is higher compared to the 3-way OSD, this can be because the HRTFs for the 6° source span contain only a small amount of energy in the highest frequencies. If one attempts to invert such a transfer function, the solution will boost frequencies just below the Nyquist frequency, which can reduce the cross-talk cancellation effectiveness.

Norcross [79] concluded that artefacts above -50 dB in the noise floor produced delay-type effects that where clearly audible to the subjects. Though the referenced study was made on a 1-channel inversion problem and the problem addressed here is 2-channel inversion, one would expect that the level of artefacts in the noise floor of the 1-channel system will still be audible in the 2-channel inversion case. The results in Table 5.2 indicates that the minimum filter length for not producing any audible artefacts using

Number of coefficients	Bandwidth SD $S(k) > 20$ dB [Hz]	Bandwidth 2-way OSD $S(k) > 20$ dB [Hz]	Bandwidth 3-way OSD $S(k) > 20$ dB [Hz]
32	-	-	-
64	-	-	-
128	-	-	-
256	-	-	-
512	2000-6000	2000-5000	4000-16000
1024	900-6000	2000-5000	4000-16000
2048	900-8000	1000-8000	700-16000
4096	800-8000	100-20000	50-16000
8192	800-19000	100-20000	30-16000
16384	700-19000	100-20000	30-16000

Table 5.1: **Control performance for different filter lengths and systems**.

the -50 dB limit is 4096 coefficients for all three systems.

Number of coefficients	Noise floor SD [dB]	Noise floor 2-way OSD [dB]	Noise floor 3-way OSD [dB]
32	-10	-10	-10
64	-10	-20	-15
128	-20	-20	-20
256	-25	-30	-30
512	-35	-40	-40
1024	-40	-40	-40
2048	-40	-40	-40
4096	-50	-55	-55
8192	-55	-60	-55
16384	-55	-60	-55

Table 5.2: **Noise floor for different filter lengths and systems**.

5.3 Cross-talk cancellation effectiveness

The effectiveness of cross-talk cancellation depends partly on the position of the listener and whether the system is static or adaptive. The evaluation is performed in the lateral plane (asymmetric), the fore and aft plane (symmetric) and for rotation (asymmetric). The listening position was moved from $-1 \leq X \leq 1$ m in the lateral plane ($Z = 0$) and from $-1 \leq Z \leq 1$ m ($X = 0$ m) in the fore and aft plane both with a step size of 0.01 m. The evaluation of listener rotation was performed by rotating the listener position from 0° to 90° in 1° steps. The evaluation is carried out for the static case and the adaptive case. In the static system, the cross talk cancellation filters are not updated when the listener is moving away from the optimal position. On the contrary, in the adaptive case the cross talk cancellation filters are updated when the listener is moving away from its optimal position. Finally, the effect of non-undividualised pinnae is investigated in the lateral plane for an adaptive SD system.

The transfer functions and inverse transfer function for the three virtual sound imaging systems under investigation are labelled as defined in Section 5.1, i.e. the SD is associated with $\mathbf{C}(k)$, the 2-way OSD is associated $\mathbf{D}(k)$ and 3-way OSD is associated $\mathbf{E}(k)$. Likewise, the inverse transfer functions are labelled as $\mathbf{H}(k)$, $\mathbf{J}(k)$ and $\mathbf{K}(k)$. The channel separation parameter in Equation 5.15 can be used to define the "sweet-spot" size, by setting a minimum required channel separation. The bandwidth estimates in this section are found by defining a minimum required channel separation of $S(k) > 20$ dB.

5.3.1 Static systems

The effectiveness of static cross-talk cancellation is simulated at asymmetric listener positions in the lateral plane, at symmetric listener positions in the fore and aft plane and for listener rotation. The cross-talk cancellation filter was computed for each listening position using the fast-deconvolution algorithm with a FFT size of 4096 coefficients and 1024 coefficients of the impulse responses. The impulse response was limited to 1024 coefficients to include the most important information of the HRTF and the loudspeaker response. The FFT size is again more than four times the number of coefficients, to minimize wrap around effects. The first $N_i = 1024$ coefficients of the transfer function $\mathbf{C}(k)$ is named $\mathbf{C}_{N_i}(k)$ and the inverse is named $\mathbf{H}_N(k)$. The inverse transfer functions $\mathbf{H}_N(k)$ were found by zero padding $\mathbf{C}_{N_i}(k)$ up to $N = 4096$ coefficients and then using the fast-deconvolution algorithm with the regularisation factor of $\beta = 10^{-4}$.

The plant model $\mathbf{C}_{N_p}(k)$ that was used contains $N_p = 3000$ coefficients of the transfer functions. A Hann-window was applied to the plant model $\mathbf{C}_{N_p}(k)$ in order to suppress noise in the late response. The plant model was chosen to reflect the system performance under realistic conditions. The matrix of transfer functions, inverse matrix and plant model for the 2-way OSD and 3-way OSD are found in the same way as for the SD.

The results for static cross-talk cancellation effectiveness at asymmetric and symmetric listener positions is presented here. Figure 5.10 - Figure 5.21 illustrates simulated control performance in the lateral, in the fore and aft plane and for listener rotation as a function of frequency and listener position for the SD, 2-way OSD and 3-way OSD. The elements $P_{11}(k)$, $P_{12}(k)$, $P_{21}(k)$ and $P_{22}(k)$ of the control performance matrix shows the effectiveness of the cross-talk cancellation scheme. The results show that the 2-way OSD and 3-way OSD outperforms the SD in terms of operative bandwidth as expected.

The "sweet-spot" size in the lateral plane is determined here for a channel separation of 20 dB. The sweet-spot size in the lateral plane for SD is limited to a region of ±0.01 m from 200 Hz up to 8 kHz. The sweet-spot size in the lateral plane for 2-way OSD is limited to a region of ±0.01 m from 60 Hz up to 20 kHz. Likewise, for the 2-way OSD the sweet-spot size for 3-way OSD is limited to a region of ±0.01 m from 60 Hz up to 20 kHz.

The low-pass cross-over frequency for the 32° source span of the 3-way OSD system can be seen as too high in Figure 5.15, since the ringing frequency appears to be within its frequency band. This is due to discrepancies in the analytical free field model used for determining the cross-over frequencies, where the condition number matrix of the HRTFs is partly different from that of the adjusted free field model presented in Section 2.4.1. The adjusted free field model with a variable receiver distance is here a linear function of frequency and source span but could be modified to fit the HRTF data at hand better by taking into account non-linearities. However, this can also be taken account for by lowering the low-pass cross-over frequency for the 32° source span and by lowering the high-pass cross-over frequency for the 6° source span.

The "sweet-spot" size in the fore and aft direction is much greater than in the lateral plane, since the ITD parameter is constant. The sweet-spot size in the fore and aft for SD is limited to a region ±0.15 m from 200 Hz up to 8 kHz for a channel separation of 20 dB. The limit for the 2-way OSD is ±0.15 m from 60 Hz up to 20 kHz. Similarly, the sweet-spot size for 3-way OSD is limited to a region of ±0.15 m from 60 Hz up to 20 kHz (if the cross-over frequency is lowered from 4000 Hz to about 3000 Hz). The "sweet-spot" size with respect to rotation and 20 dB channel separation is, for the SD ±2° within 500 Hz - 8 kHz, for the 2-way OSD ±2° within 120 Hz - 14 kHz and for the 3-way OSD ±2° within 120 Hz - 14 kHz.

The results for static virtual sound imaging systems are summarised in Table 5.3. The 2-way OSD and 3-way OSD perform better at both high and low frequencies compared to the SD. The difference in performance between the 2-way OSD and the 3-way OSD is small, even though it does not show in Table 5.3, the 3-way OSD performs slightly better than the 2-way OSD.

System	Bandwidth/ azimuth angle $\phi(\pm 2°)$ $S(k) > 20$ dB	Bandwidth/ lateral $X(\pm 0.01)$ m $S(k) > 20$ dB	Bandwidth/ fore and aft $Z(\pm 0.15)$ m $S(k) > 20$ dB	Bandwidth/ optimal position $(X = 0,\ Z = 0)$ $S(k) > 20$ dB
SD	500-8000	200-8000	200-8000	200-20000
2-way OSD	120-14000	60-20000	60-20000	60-20000
3-way OSD	120-14000	60-20000	60-20000[a]	60-20000

Table 5.3: **Control performance of the static system for asymmetric listener positions**. *The control performance is shown for azimuth angle, lateral displacement and fore and aft displacement. (a) There are some spikes present around the cross-over frequency at 3.5 - 4 kHz that can possibly be removed by lowering the cross-over frequency.*

To overcome these limitations with the "sweet-spot" size, one solution is to update the cross-talk cancellation filters according to the listeners location and orientation. This is demonstrated in the following section.

5.3.2 Adaptive systems

The cross-talk cancellation filter was computed for each listening position using the fast-deconvolution algorithm with a FFT size of 4096 coefficients and 1024 coefficients of the impulse responses as in the static case. For the SD, the first $N_i = 1024$ coefficients of the matrix transfer functions $\mathbf{C}(k)$ is named $\mathbf{C}_{N_i}(k)$ and the inverse matrix of $\mathbf{C}_{N_i}(k)$ is named $\mathbf{H}_N(k)$. The inverse transfer functions $\mathbf{H}_N(k)$ were found by zero padding $\mathbf{C}_{N_i}(k)$ up to $N = 4096$ coefficients and then using the fast-deconvolution algorithm as for the static system evaluation. Again, the regularisation parameter β in the fast-deconvolution algorithm was 10^{-4}. The “sweet-spot” size is also here determined for a channel separation of 20 dB in the lateral plane, fore and aft plane and rotation.

This paragraph presents the results for cross-talk cancellation effectiveness in the lateral plane. The cross-talk cancellation filters were adapted to the listener position. The listening position was moved from $-1 \leq X \leq 1$ m in the lateral plane for $Z = 0$ m. The cross-talk cancellation effectiveness for the SD, 2-way OSD and 3-way OSD in the lateral plane is illustrated in Figure 5.22 - Figure 5.24. The results suggests that it is possible for the listener to move at least from $-1 \leq X \leq 1$ m in the lateral plane with a channel separation better than 20 dB within a certain frequency range. The frequency range for the listener to move within $-1 \leq X \leq 1$ m is, for the SD 500 Hz - 8 kHz, for the 2-way OSD 120 Hz - 16 kHz and for the 3-way OSD 110 Hz - 16 kHz.

This paragraph presents the results for cross-talk cancellation effectiveness in the fore and aft plane. The listening position was moved from $-1 \leq Z \leq 1$ m in the fore and aft plane for $X = 0$ m. The cross-talk cancellation effectiveness in the fore and aft is illustrated in Figure 5.25 - Figure 5.27. The results shows that it is possible for the listener to move at least the from $-1 \leq Z \leq 1$ m in the fore and aft plane with a channel separation better than 20 dB within a certain frequency range. The frequency range when the listener moves within $-1 \leq Z \leq 1$ m is, for the SD 800 Hz - 8 kHz, for the 2-way OSD 120 Hz - 16 kHz (some spikes at 15 dB channel separation are present around 10 kHz when $-1 \leq Z \leq -0.8$) and for the 3-way OSD 120 Hz - 16 kHz. Some spikes are present at 15 dB channel separation around the cross-over frequency at 3.5 - 4 kHz (can possibly be removed by lowering the cross-over frequency) and at 10 kHz when $-1 \leq Z \leq -0.8$.

This paragraph presents the results for cross-talk cancellation effectiveness dependency on azimuth angle. The listener position was rotated from 0° to 90° in 1° steps. The cross-talk cancellation effectiveness for the SD, 2-way OSD and 3-way OSD for listener rotation is illustrated in Figure 5.28 - Figure 5.33. The results suggests that it is possible for the listener to rotate at least from ±71° in the fore and aft plane with a channel

System	Bandwidth/ azimuth angle $\phi(0°$ to $71°)$ $S(k) > 20$ dB	Bandwidth/ lateral $-1 \leq X \leq 1$ m $S(k) > 20$ dB	Bandwidth/ fore and aft $-1 \leq Z \leq 1$ m $S(k) > 20$ dB	Bandwidth/ optimal position $(X = 0, Z = 0)$ $S(k) > 20$ dB
SD	1000-8000	500-8000	800-8000	500-10000
2-way OSD	$150 - 16000^{\mathrm{a}}$	120-16000	$120 - 16000^{\mathrm{c}}$	120-16000
3-way OSD	$150 - 16000^{\mathrm{b}}$	110-16000	$120 - 16000^{\mathrm{d}}$	110-16000

Table 5.4: **Control performance of the adaptive system for asymmetric listener positions**. *The control performance is shown for azimuth angle, lateral displacement and fore and aft displacement. (a) There are some spikes present at 15 dB channel separation are present around the cross-over frequency at 900 Hz - 1 kHz that can possibly be removed by increasing the cross-over frequency. (b) There are some spikes present at 15 dB channel separation are present around the cross-over frequency at 3.5 - 4 kHz that can possibly be removed by lowering the cross-over frequency. (c) There are some spikes present at 15 dB channel separation are present around 10 kHz when $-1 \leq Z \leq -0.8$. (d) There are some spikes present at 15 dB channel separation are present around the cross-over frequency at 3.5 - 4 kHz (can possibly be removed by lowering the cross-over frequency) and at 10 kHz when $-1 \leq Z \leq -0.8$.*

separation better than 20 dB within a certain frequency range. The frequency range when the listener rotates within ±71° is, for the SD in the range of 1000 Hz - 8 kHz, for the 2-way OSD in the range of 150 Hz - 16 kHz (some spikes at 15 dB channel separation are present around the cross-over frequency at 900 Hz - 1 kHz that can possibly be removed by increasing the cross-over frequency) and for the 3-way OSD in the range of 150 Hz - 16 kHz (some spikes at 15 dB channel separation are present around the cross-over frequency at 3.5 - 4).

The results from the study of asymmetric and symmetric listener positions for adaptive virtual sound imaging systems are summarised in Table 5.4. As expected, in general the 2-way OSD and 3-way OSD perform better at both high frequencies and low frequencies compared to the SD. The difference in performance between the 2-way OSD and the 3-way OSD is relatively small and the 3-way OSD only performs slightly better than the 2-way OSD.

5.3.3 The effect of non-individualised head related transfer functions

Non-individualised HRTFs can degrade the cross-talk cancellation effectiveness of binaural sound reproduction systems. The effect of pinna mis-match is briefly investigated in this short section by conducting simulations of cross-talk cancellation effectiveness using different sets of HRTFs. The cross-talk cancellation performance for the SD is evaluated by using different HRTFs in the plant model compared to the inverse filters. The "small pinna" was used for the plant model and the "large pinna" was used for the inverse filters. The investigation was carried in the lateral plane for an adaptive SD system. The simulation was performed using the same procedure and configurations as in Section 5.3.2.

The results for cross-talk cancellation effectiveness in the lateral plane when the HRTF of the listener is different from the HRTFs that were used to compute the inverse filters are presented here. The listening position was moved from $-1 \leq X \leq 1$ m in the lateral plane for $Z = 0$ m. The cross-talk cancellation effectiveness for the SD in the lateral plane is illustrated in Figure 5.34. The results suggests that it is possible for the listener to move at least from $-1 \leq X \leq 1$ m in the lateral plane with a channel separation better than 20 dB within a frequency range of 900 Hz - 2 kHz. The cross-talk cancellation effectiveness is here clearly limited compared to when using individualised HRTFs, especially at higher frequencies above $\approx$ 7 kHz. The pinna mis-match also produces unwanted peaks and dips above $\approx$ 7 kHz that can degrade the sound quality.

5.4 Discussion

The inverse filter computed using the fast-deconvolution algorithm can possibly be shortened by applying a window function as suggested by Papadopoulos [82]. An inverse filter of length N with initial coefficients of $N_i = N/4$ can be shortened by applying a window function to the inverse filters and preliminary results show that it is possible to make the inverse filter shorter by a factor of two and achieve better results in terms of cross-talk cancellation effectiveness than with an inverse filter of the same length that has not been windowed.

The number of inverse filter coefficients can be reduced further, instead of using FIR, one can use other filter techniques such as WFIR (Karjalainen [45], Farina [29]), IIR (Jot [44]) and WIIR (Karjalainen [45]) filters. According to Farina [29], a WFIR filter can be implemented with a number of coefficients ten times lower than those of a FIR filter, but still featuring the same low-frequency equalization. However, for real-time implementations WFIR requires more computational power. In a comparison between of WFIR and FIR by Farina [29], an 45 coefficient WFIR yields better subjective results than an 225 coefficient FIR. The compared WFIR and FIR filters take approximately equal computational cost. Thus it might be possible to reduce the size of inverse filter by a factor of 225/45 by using WFIR instead of FIR with the same computational cost.

The length of the inverse filter can also be reduced by using the OSD principle, where each frequency band is reproduced from an optimal angle. If the sources are chosen so that the magnitude their frequency response is the inverse of the transfer function matrix, then the inverse filter only needs to be implementing the delay. This will result in a minimum filter length and equalisation effort. However, in an adaptive system the transfer functions will inevitably have to change as the listener moves away from the optimal listener positions. Hence, this principle is more useful for static systems.

The analytical evaluation presented in Chapter 2 is compared with the objective evaluation presented here. The cross-over frequencies found in the analytical evaluation (fixed cross-over at optimal position) were used in the objective evaluation. The result shows that the ringing frequency appears within the frequency band of the 32° for the 3-way OSD. This can be taken into account for by lowering the low-pass cross-over frequency for the 32° source span and by lowering the high-pass cross-over frequency for the 6° source span. Hence, it is desirable to adjust the model of receiver distance presented in Section 2.4.1, which is a linear function of frequency and source span. The model can be modified to fit the HRTF data at hand better by using a non-linear approach.

5.5 Conclusion

An objective evaluation of inverse filter design and cross-talk cancellation effectiveness for three different audio systems has been presented. The design process of the cross-talk cancellation networks using a minimum number of inverse filters has been shown. The cross-talk cancellation effectiveness has been evaluated by varying the filter lengths using the fast-deconvolution algorithm and by changing the listener position in the lateral plane, the fore and aft plane and by azimuth rotation. The limitations of the "sweet-spot" of static virtual sound imaging systems have been illustrated by simulations of the control performance under anechoic conditions. The benefit of using an adaptive virtual sound imaging where the cross-talk cancellation filters are updated to the listeners location and orientation have been demonstrated by Matlab [66] simulations using data from the database measurement presented in Chapter 3. The effect of non-individualised pinnae has been illustrated in the lateral plane for an adaptive SD system and the results suggests that the flatness of the frequency response can be degraded considerably above 7 kHz.

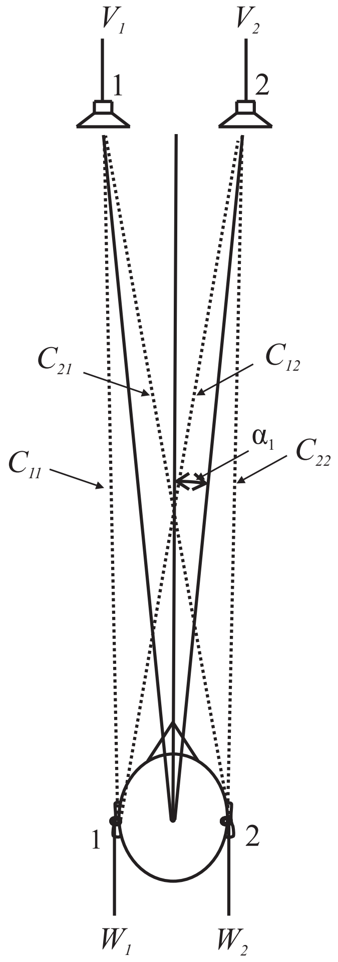


Figure 5.1: **Transfer function paths and source angle for the SD**.

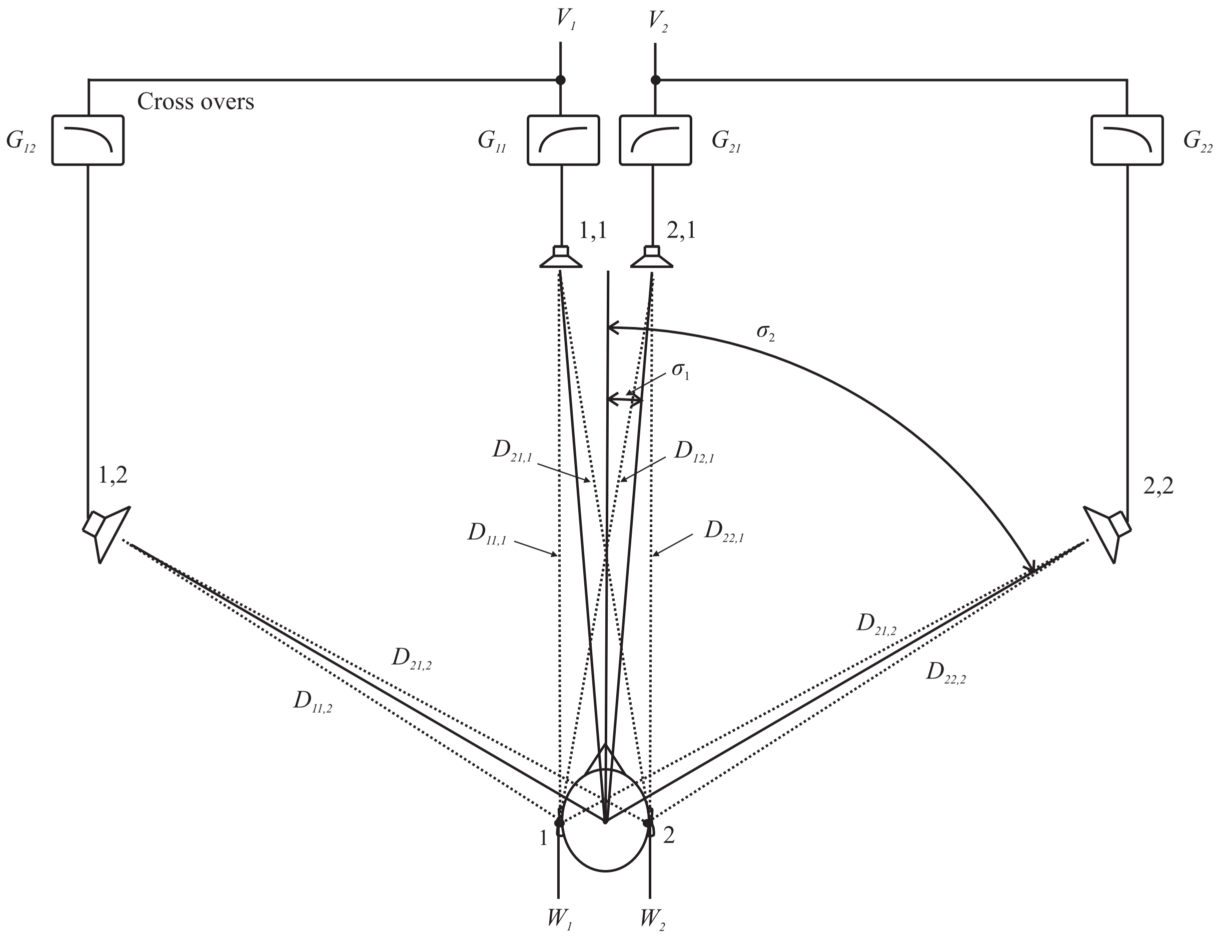


Figure 5.2: **Transfer function paths, source angles and cross-overs for the 2-way OSD**.

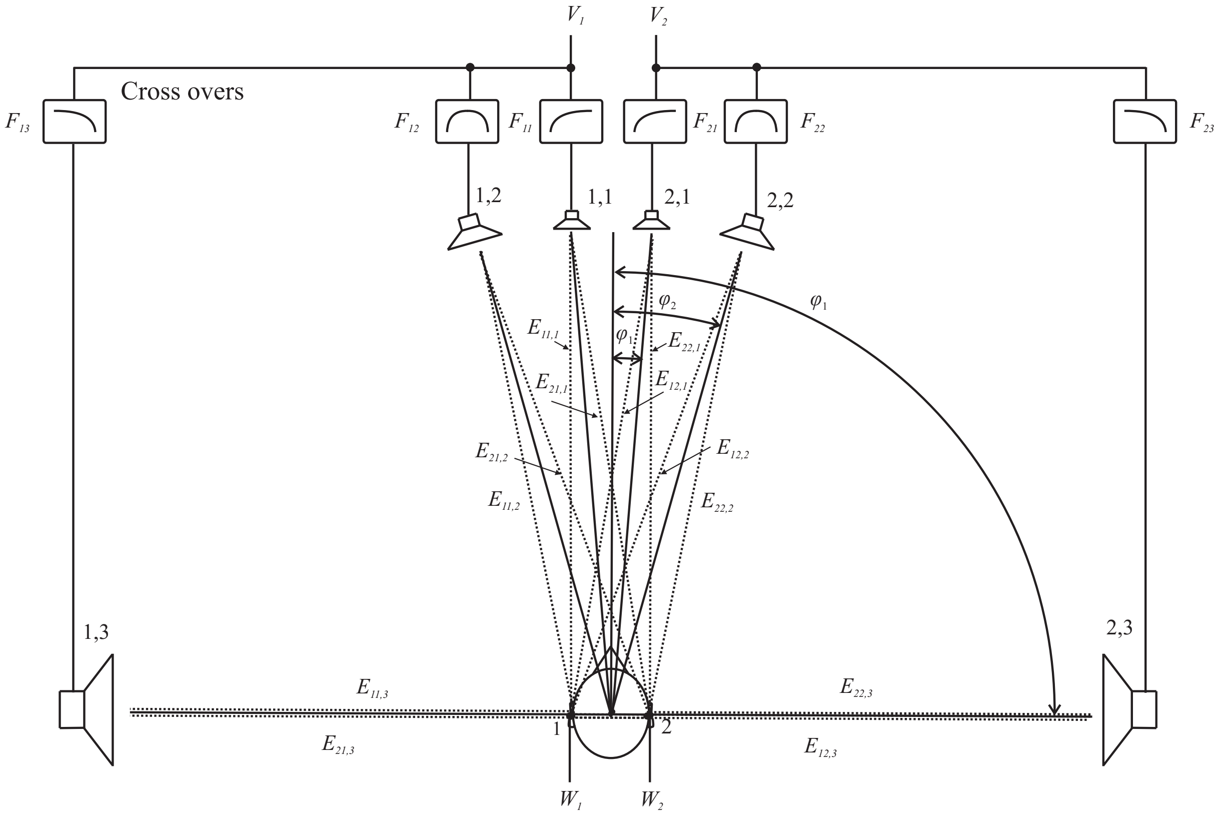
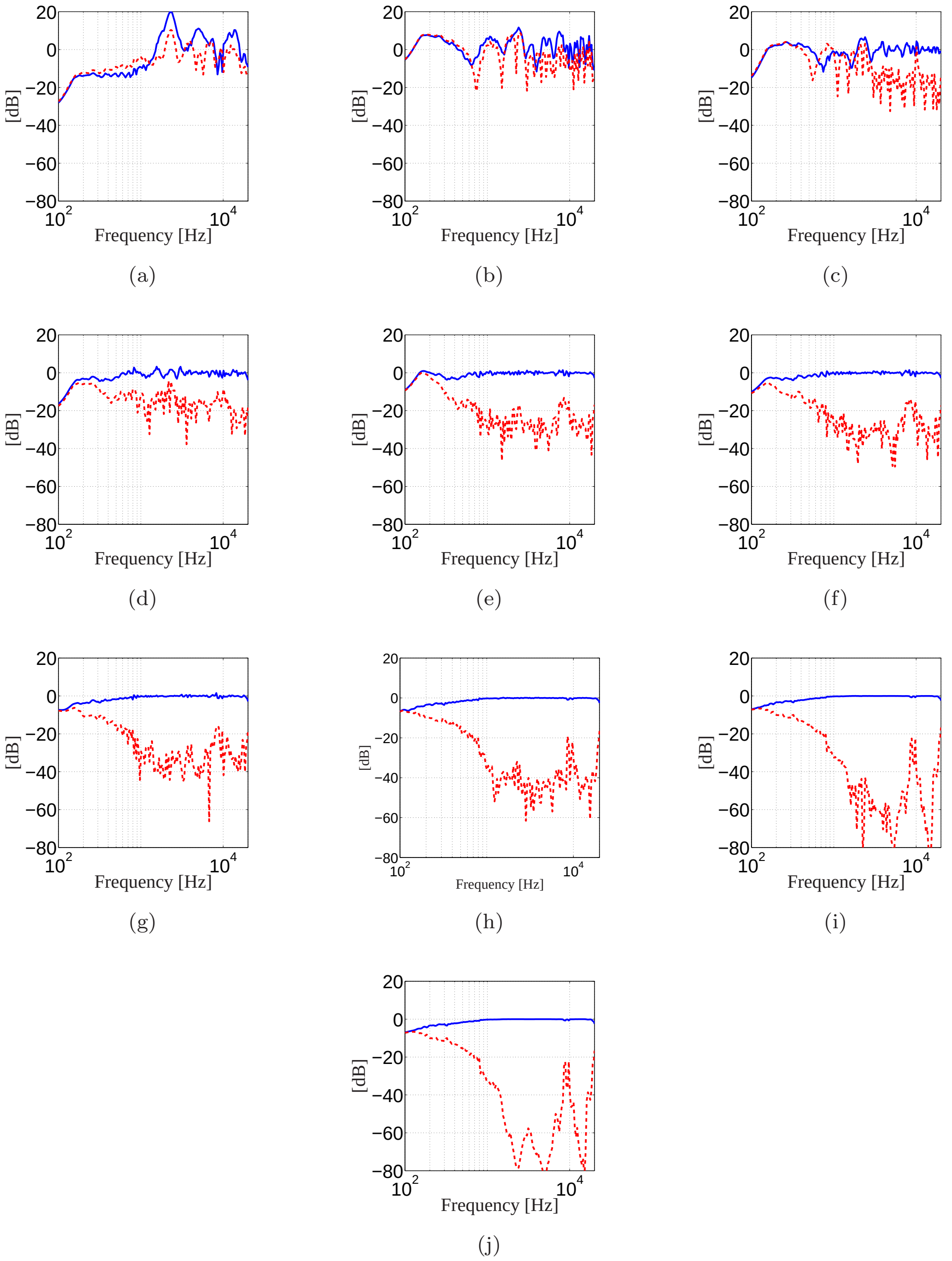
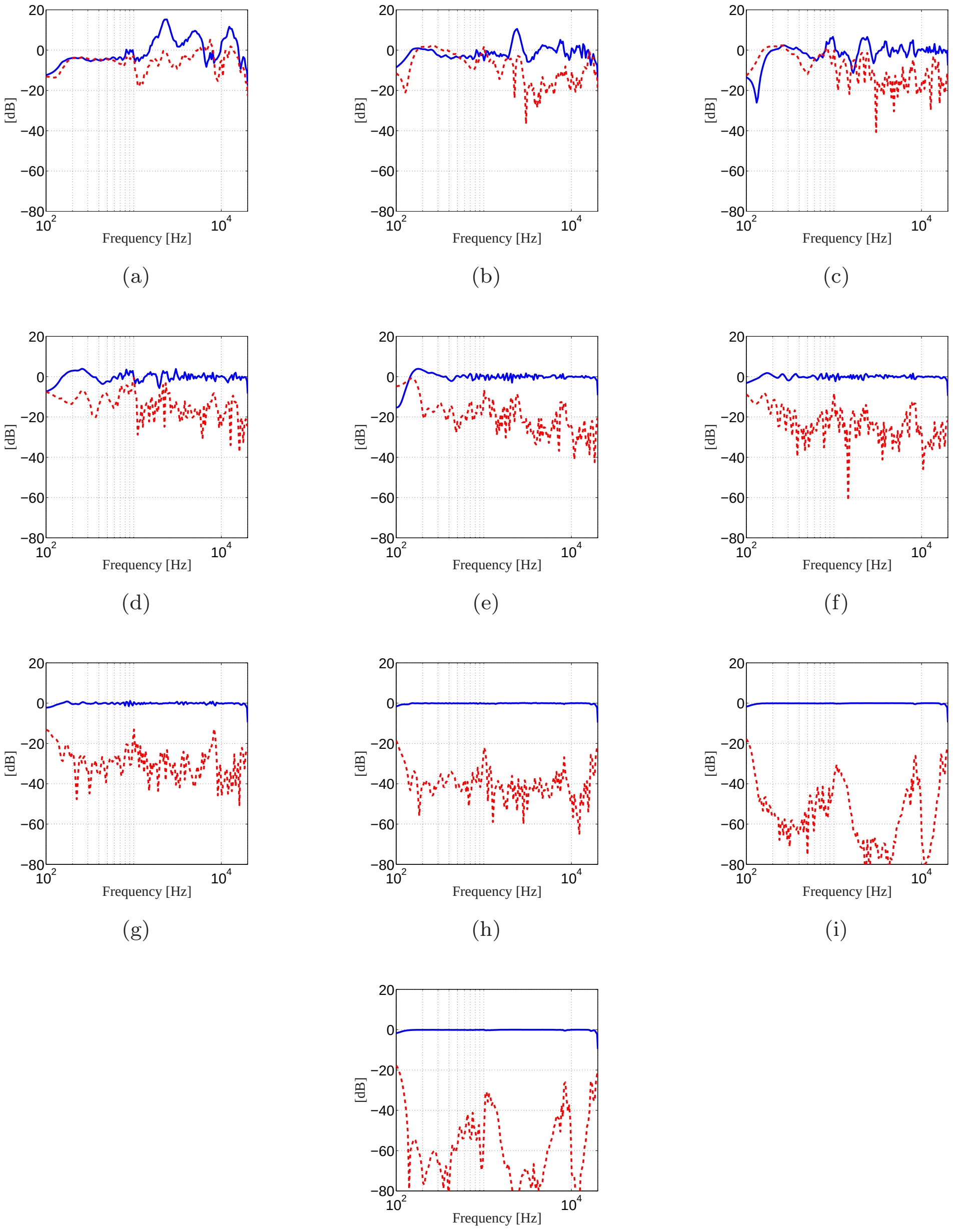


FIGURE 5.3: **Transfer function paths, source angles and cross-overs for the 3-way OSD.**



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Figure 5.4: **Simulated cross-talk cancellation performance for the SD**. *The solid line is $P_{11}(k)$ dashed line is $P_{21}(k)$ and the frequency axis is plotted from 100 to 20000 Hz on a log scale. The 3000 coefficient plant model is used. Results are illustrated for a number of inverse filter coefficients given by (a) 32. (b) 64. (c) 128. (d) 256. (e) 512. (f) 1024. (g) 2048. (h) 4096. (i) 8192. (j) 16384.*



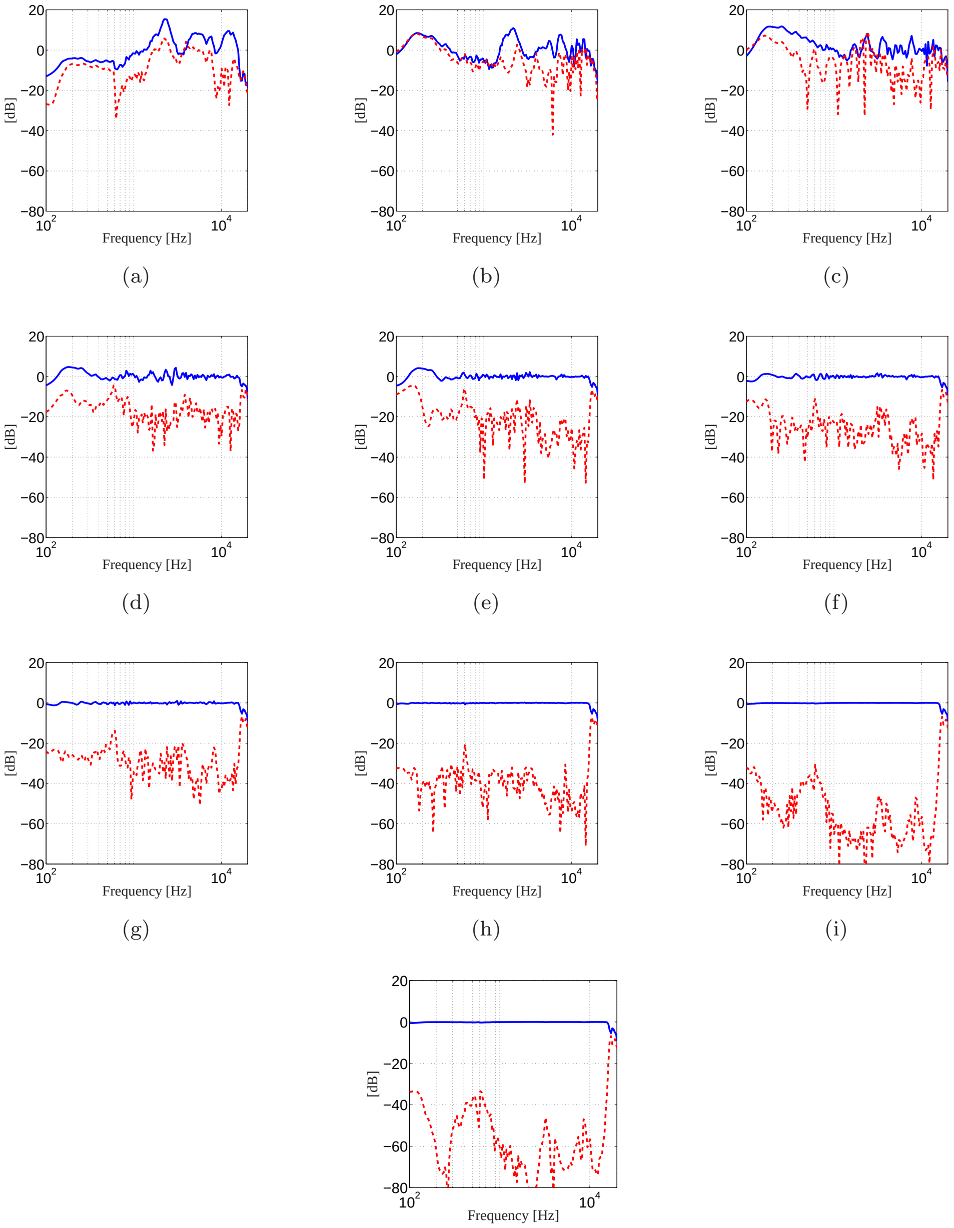
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

FIGURE 5.5: **Simulated cross-talk cancellation performance for the 2-way OSD**. *The solid line is $P_{11}(k)$ dashed line is $P_{21}(k)$ (the frequency axis is plotted from 100 to 20000 Hz on a log scale). The 3000 coefficient plant model is used. Results are shown for a number of inverse filter coefficients given by (a) 32. (b) 64. (c) 128. (d) 256. (e) 512. (f) 1024. (g) 2048. (h) 4096. (i) 8192. (j) 16384.*



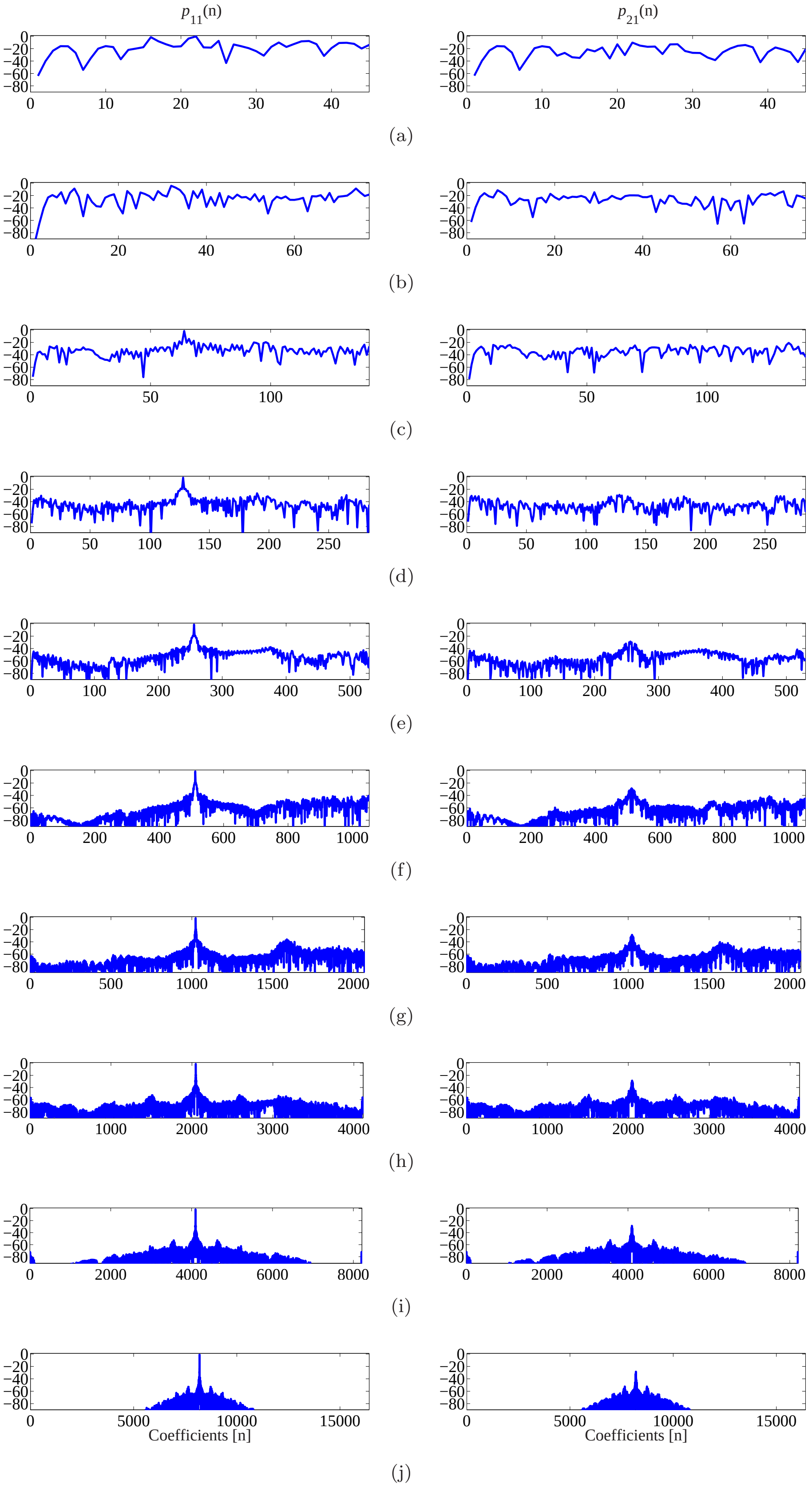
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

FIGURE 5.6: **Simulated cross-talk cancellation performance for the 3-way OSD**. *The solid line is $P_{11}(k)$ dashed line is $P_{21}(k)$ (the frequency axis is plotted from 100 to 20000 Hz on a log scale). The 3000 coefficient plant model is used. Results are shown for a number of inverse filter coefficients given by (a) 32. (b) 64. (c) 128. (d) 256. (e) 512. (f) 1024. (g) 2048. (h) 4096. (i) 8192. (j) 16384.*



(a)

(b)

(c)

(d)

(e)

(f)

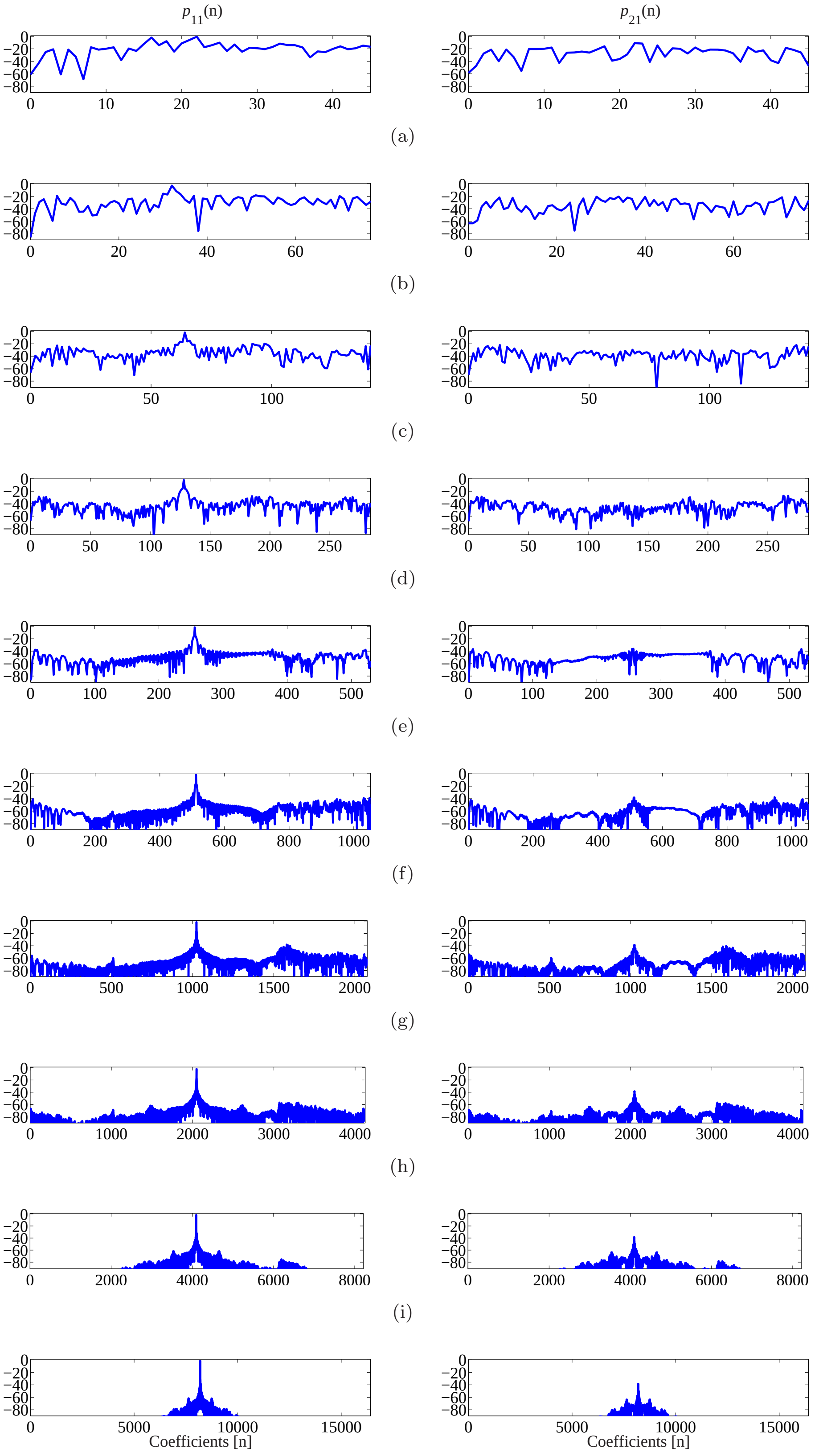
(g)

(h)

(i)

(j)

FIGURE 5.7: **Simulated cross-talk cancellation performance for the SD in the time domain**. *The amplitudes of $p_{11}(n)$ and $p_{21}(n)$ are plotted in dB. The 3000 coefficient plant model is used. Results are illustrated for a number of inverse filter coefficients given by (a) 32. (b) 64. (c) 128. (d) 256. (e) 512. (f) 1024. (g) 2048. (h) 4096. (i) 8192. (j) 16384.*



(a)

(b)

(c)

(d)

(e)

(f)

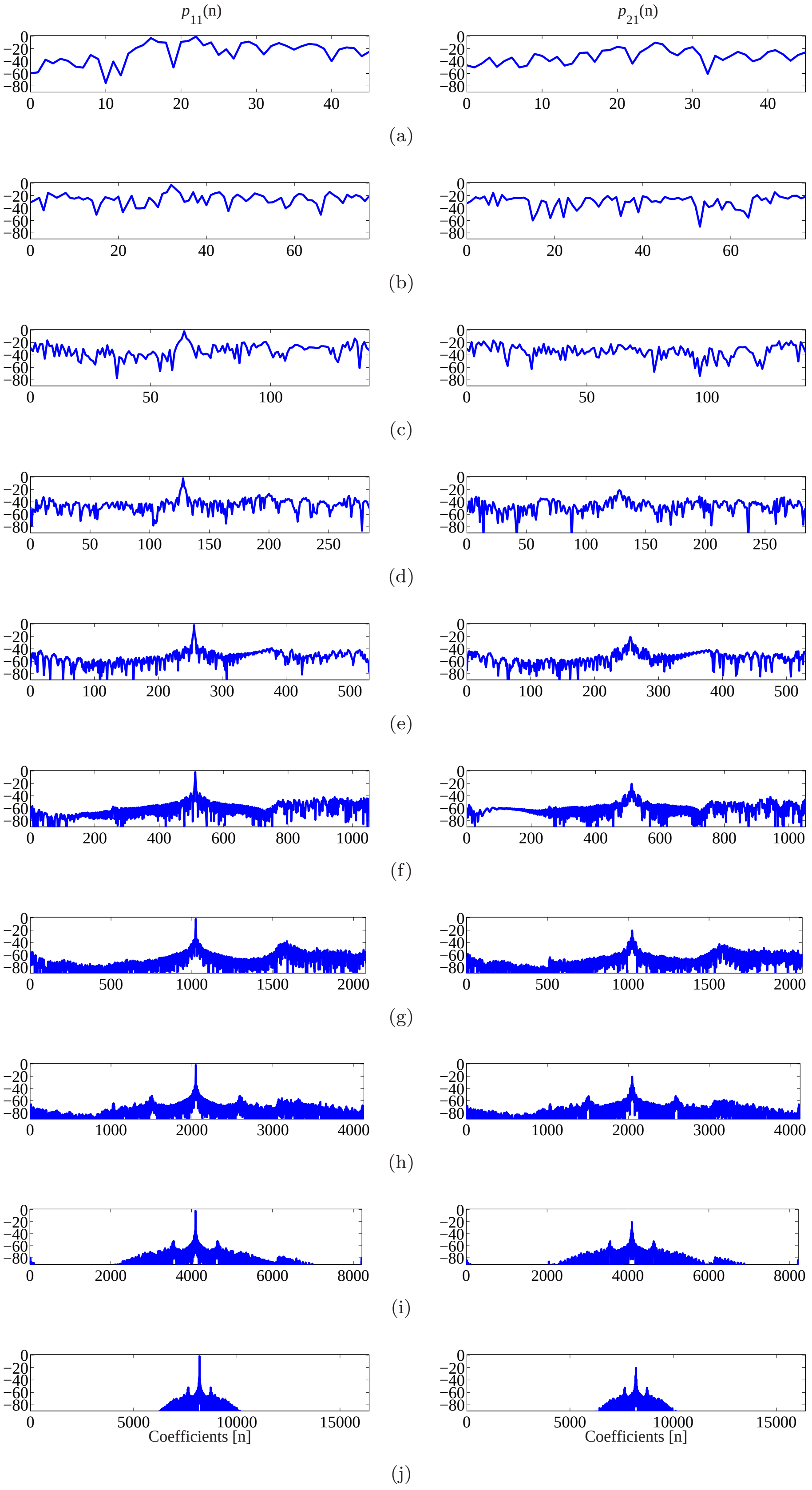
(g)

(h)

(i)

(j)

FIGURE 5.8: **Simulated cross-talk cancellation performance for the 2-way OSD in the time domain**. *The amplitudes of $p_{11}(n)$ and $p_{21}(n)$ are plotted in dB. The 3000 coefficient plant model is used. Results are illustrated for a number of inverse filter coefficients given by (a) 32. (b) 64. (c) 128. (d) 256. (e) 512. (f) 1024. (g) 2048. (h) 4096. (i) 8192. (j) 16384.*



(a)

(b)

(c)

(d)

(e)

(f)

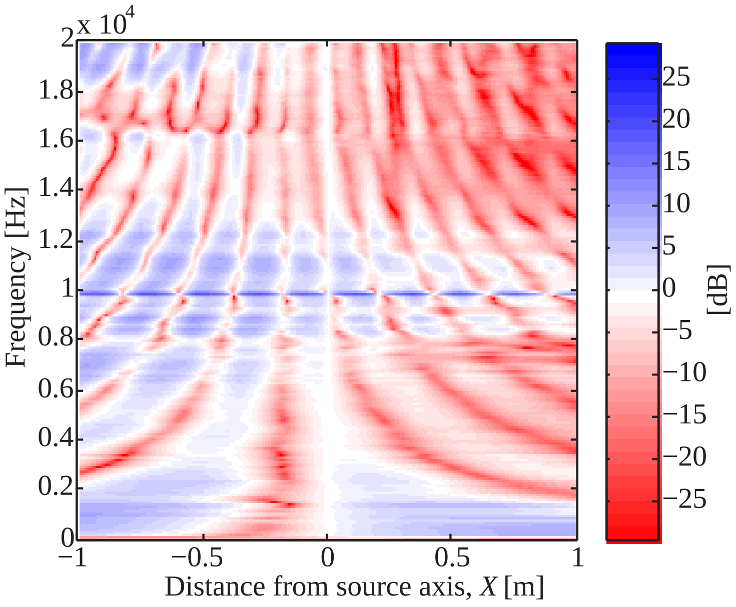
(g)

(h)

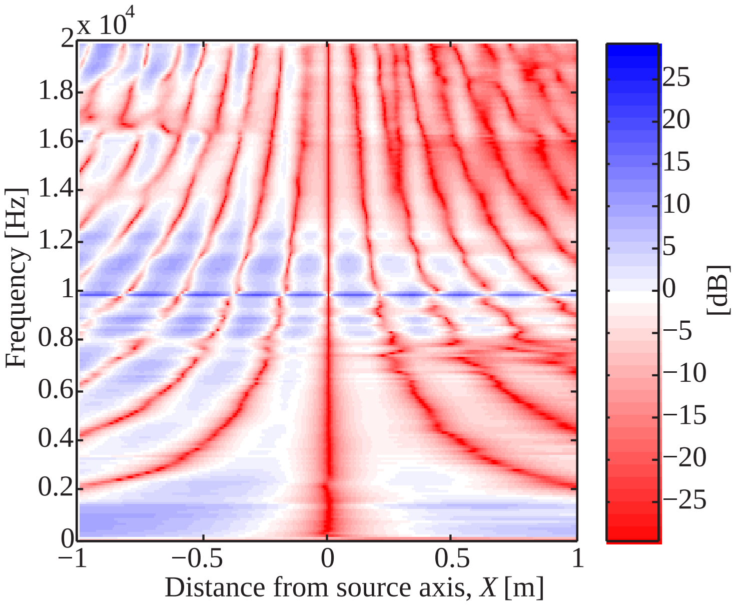
(i)

(j)

FIGURE 5.9: **Simulated cross-talk cancellation performance for the 3-way OSD in the time domain**. *The amplitude of $p_{11}(n)$ and $p_{21}(n)$ plotted in dB. The 3000 coefficient plant model is used. Results are illustrated for a number of inverse filter coefficients given by (a) 32. (b) 64. (c) 128. (d) 256. (e) 512. (f) 1024. (g) 2048. (h) 4096. (i) 8192. (j) 16384.*

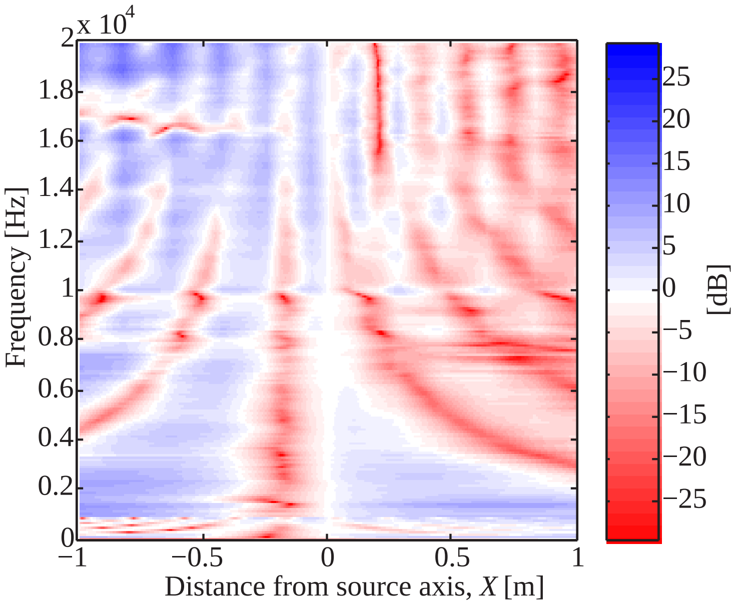


(a)

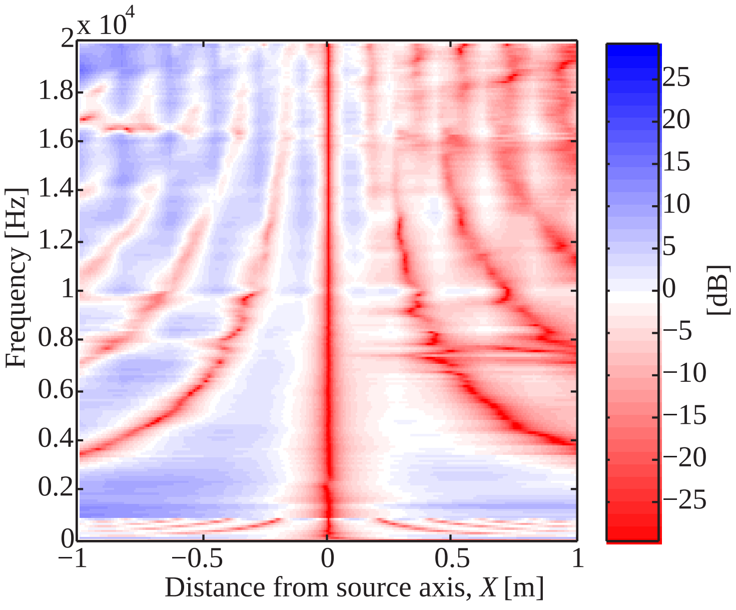


(b)

FIGURE 5.10: **Simulated static cross-talk cancellation as a function of lateral position in the horizontal plane for the SD**. *The ISVR HRTF database is used and the regularisation parameter is set to $\beta = 10^{-4}$. The inverse filter length is 4096. The control performance $P_{11}(k)$ and $P_{21}(k)$ is presented ($P_{22}(k)$ and $P_{12}(k)$ are mirrored versions of $P_{11}(k)$ and $P_{21}(k)$). (a) Left ear ($P_{11}(k)$). (b) Right ear ($P_{21}(k)$).*

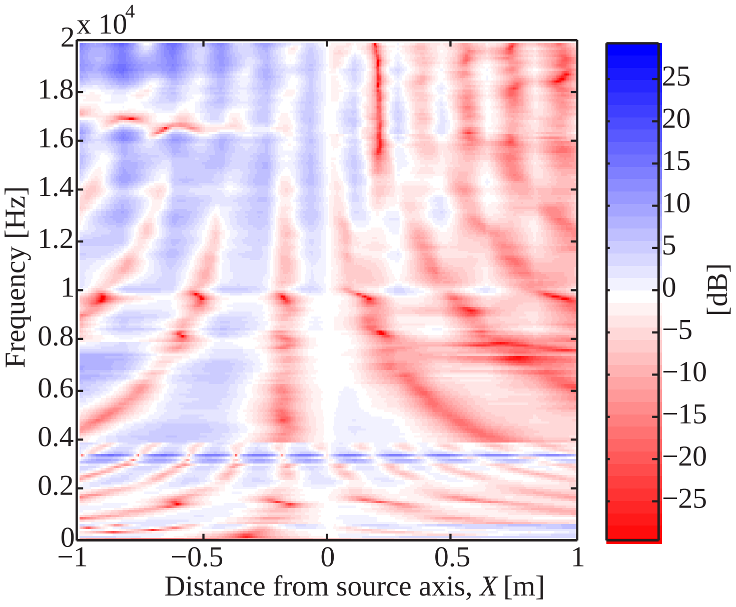


(a)

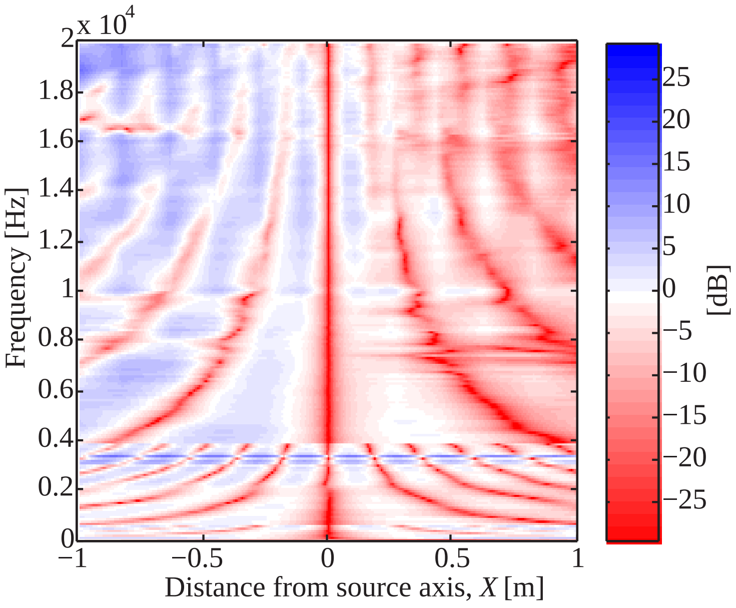


(b)

FIGURE 5.11: **Simulated static cross-talk cancellation as a function of lateral position for the 2-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^{-4}$. *The inverse filter length is 4096 and the cross-over frequency is at 900 Hz. The control performance* $P_{11}(k)$ *and* $P_{21}(k)$ *is presented (*$P_{22}(k)$ *and* $P_{12}(k)$ *are mirrored versions of* $P_{11}(k)$ *and* $P_{21}(k)$*). (a) Left ear (*$P_{11}(k)$*). (b) Right ear (*$P_{21}(k)$*).*

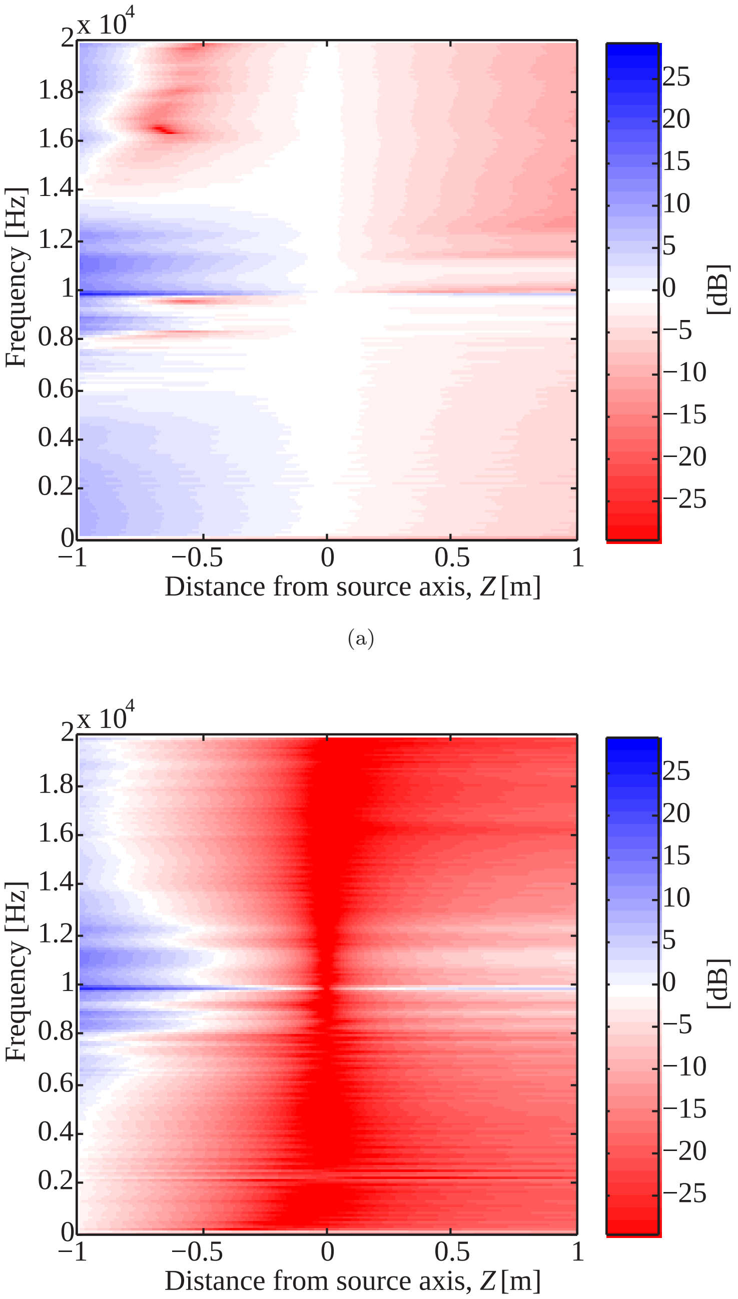


(a)



(b)

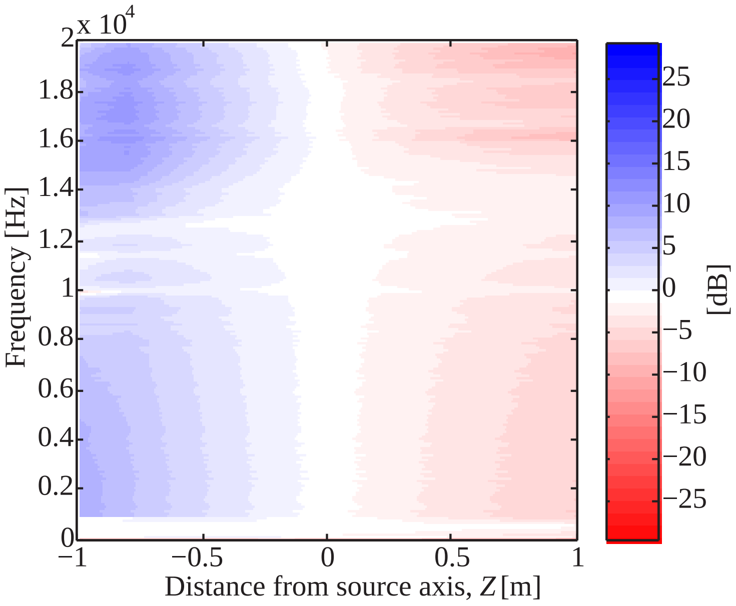
FIGURE 5.12: **Simulated static cross-talk cancellation as a function of lateral position for the 3-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^{-4}$. *The inverse filter length is 4096 and the cross-over frequencies are at 600 Hz and 4000 Hz. The control performance* $P_{11}(k)$ *and* $P_{21}(k)$ *is presented (*$P_{22}(k)$ *and* $P_{12}(k)$ *are mirrored versions of* $P_{11}(k)$ *and* $P_{21}(k)$*). (a) Left ear (*$P_{11}(k)$*). (b) Right ear (*$P_{21}(k)$*).*



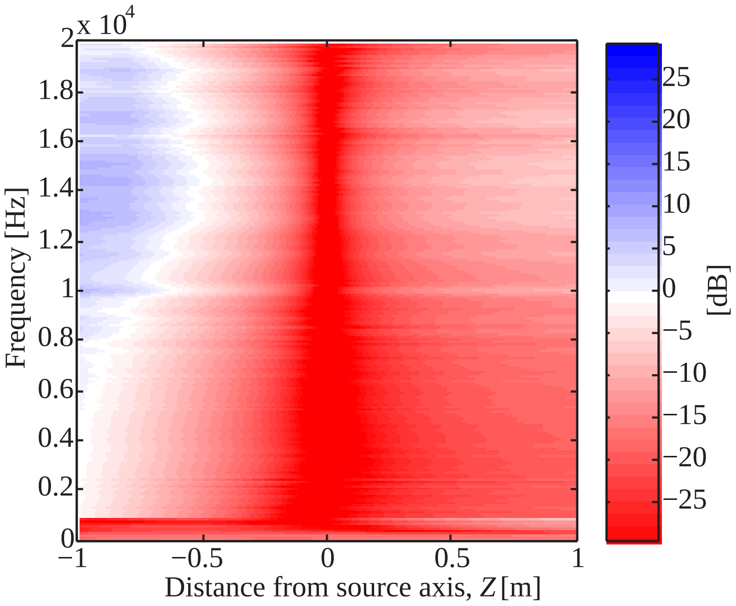
(a)

(b)

Figure 5.13: **Simulated static cross-talk cancellation as a function of fore and aft position for the SD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^{-4}$. *The inverse filter length is 4096. The control performance* $P_{11}(k)$ *and* $P_{21}(k)$ *is presented (*$P_{22}(k)$ *and* $P_{12}(k)$ *are mirrored versions of* $P_{11}(k)$ *and* $P_{21}(k)$*). (a) Left ear (*$P_{11}(k)$*). (b) Right ear (*$P_{21}(k)$*).*

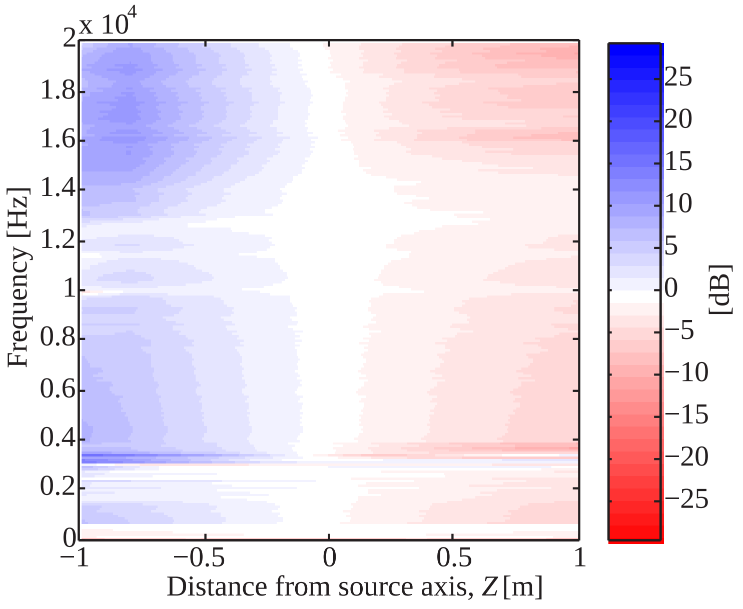


(a)

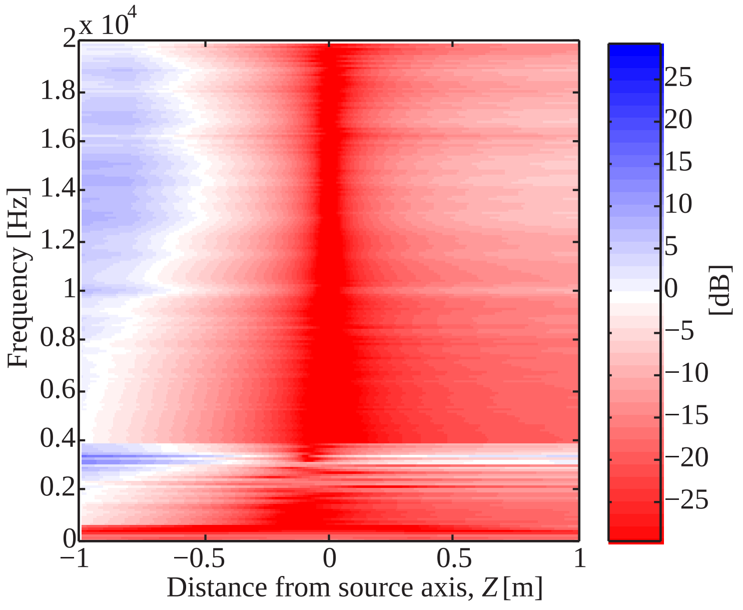


(b)

FIGURE 5.14: **Simulated static cross-talk cancellation as a function of fore and aft position for the 2-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^{-4}$. *The inverse filter length is 4096 and the cross-over frequency is at 900 Hz. The control performance* $P_{11}(k)$ *and* $P_{21}(k)$ *is presented (*$P_{22}(k)$ *and* $P_{12}(k)$ *are mirrored versions of* $P_{11}(k)$ *and* $P_{21}(k)$*). (a) Left ear (*$P_{11}(k)$*). (b) Right ear (*$P_{21}(k)$*).*

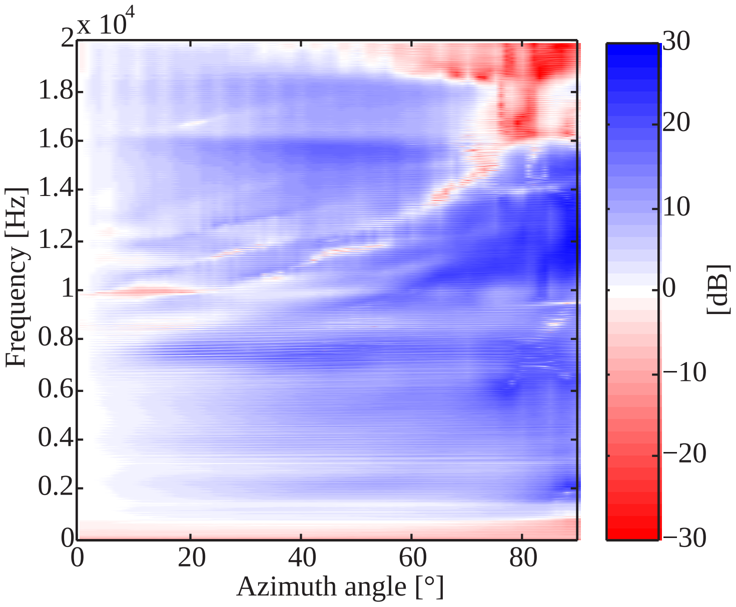


(a)

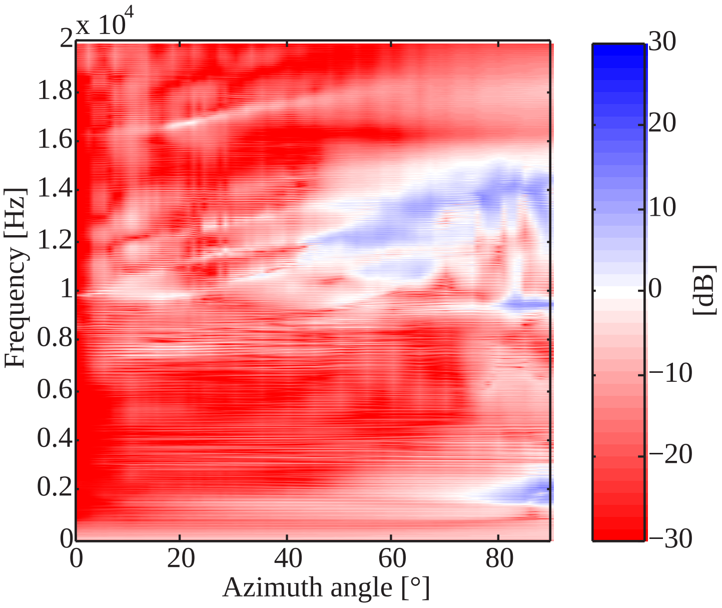


(b)

FIGURE 5.15: **Simulated static cross-talk cancellation as a function of fore and aft position for the 3-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to $\beta = 10^{-4}$. The inverse filter length is 4096 and the cross-over frequencies are at 600 Hz and 4000 Hz. The control performance $P_{11}(k)$ and $P_{21}(k)$ is presented ($P_{22}(k)$ and $P_{12}(k)$ are mirrored versions of $P_{11}(k)$ and $P_{21}(k)$). (a) Left ear ($P_{11}(k)$). (b) Right ear ($P_{21}(k)$).*

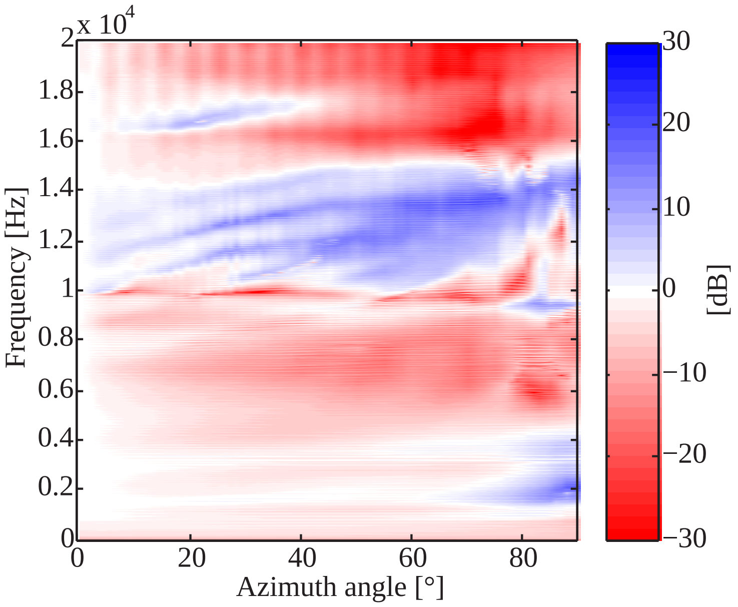


(a)

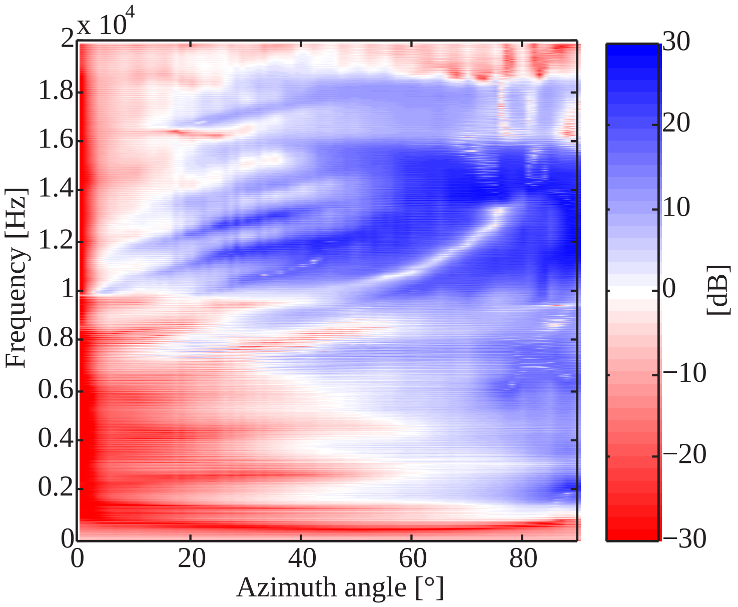


(b)

FIGURE 5.16: **Simulated static cross-talk cancellation as a function of azimuth angle for the SD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^{-4}$. *The inverse filter length is 4096. The control performance* $P_{11}(k)$ *and* $P_{21}(k)$. *(a) Left ear (*$P_{11}(k)$*). (b) Right ear (*$P_{21}(k)$*).*

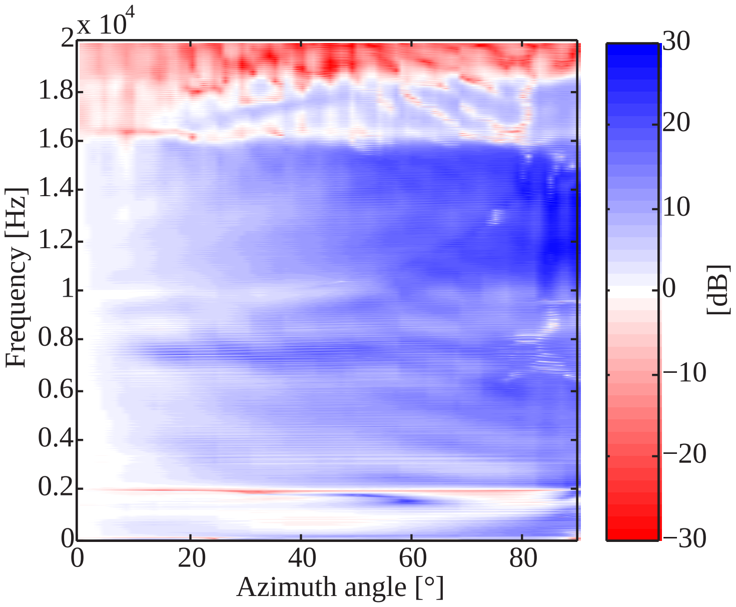


(a)

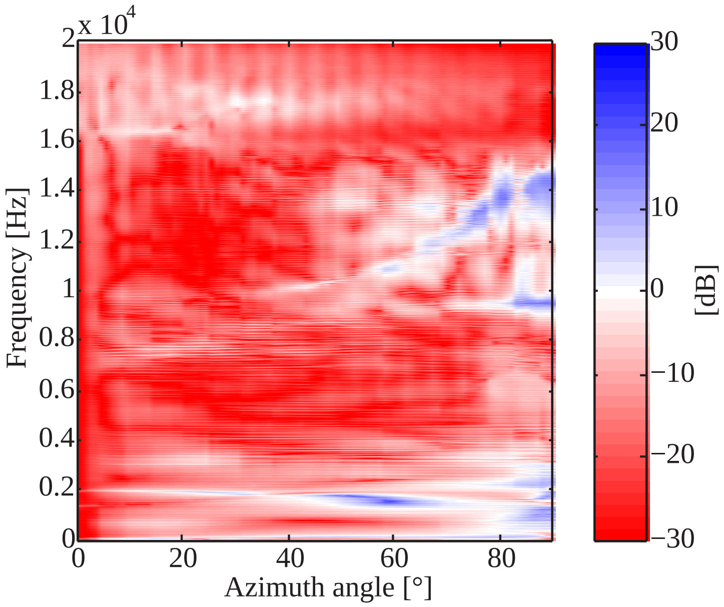


(b)

FIGURE 5.17: **Simulated static cross-talk cancellation as a function of azimuth angle for the SD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^{-4}$. *The inverse filter length is 4096. The control performance* $P_{12}(k)$ *and* $P_{22}(k)$ *is presented. (a) Left ear (*$P_{22}(k)$*). (b) Right ear (*$P_{12}(k)$*).*

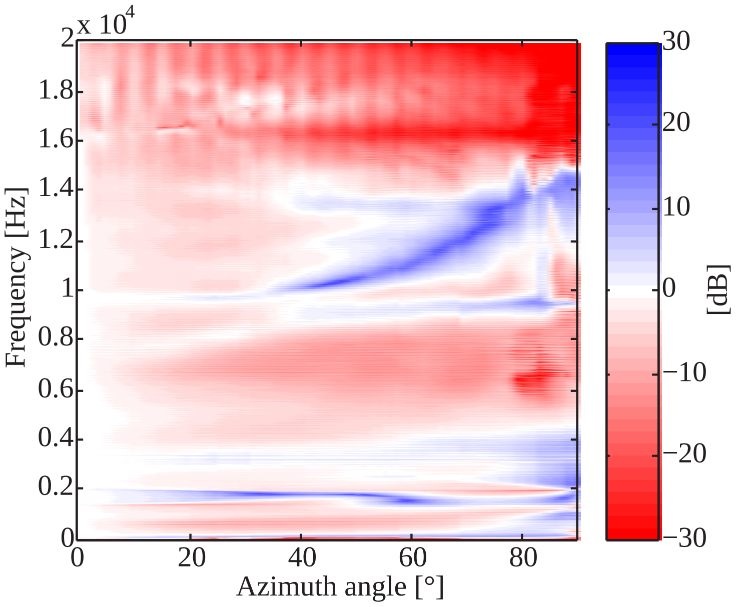


(a)

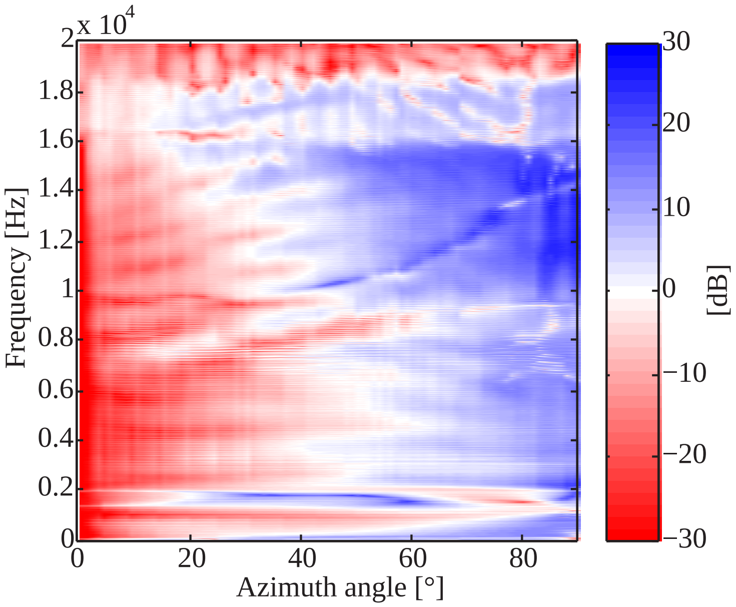


(b)

FIGURE 5.18: **Simulated static cross-talk cancellation as a function of azimuth angle for the 2-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to $\beta = 10^{-4}$. The inverse filter length is 4096 and the cross-over frequencies is at 900 Hz. The control performance $P_{11}(k)$ and $P_{21}(k)$ is presented. (a) Left ear ($P_{11}(k)$). (b) Right ear ($P_{21}(k)$).*

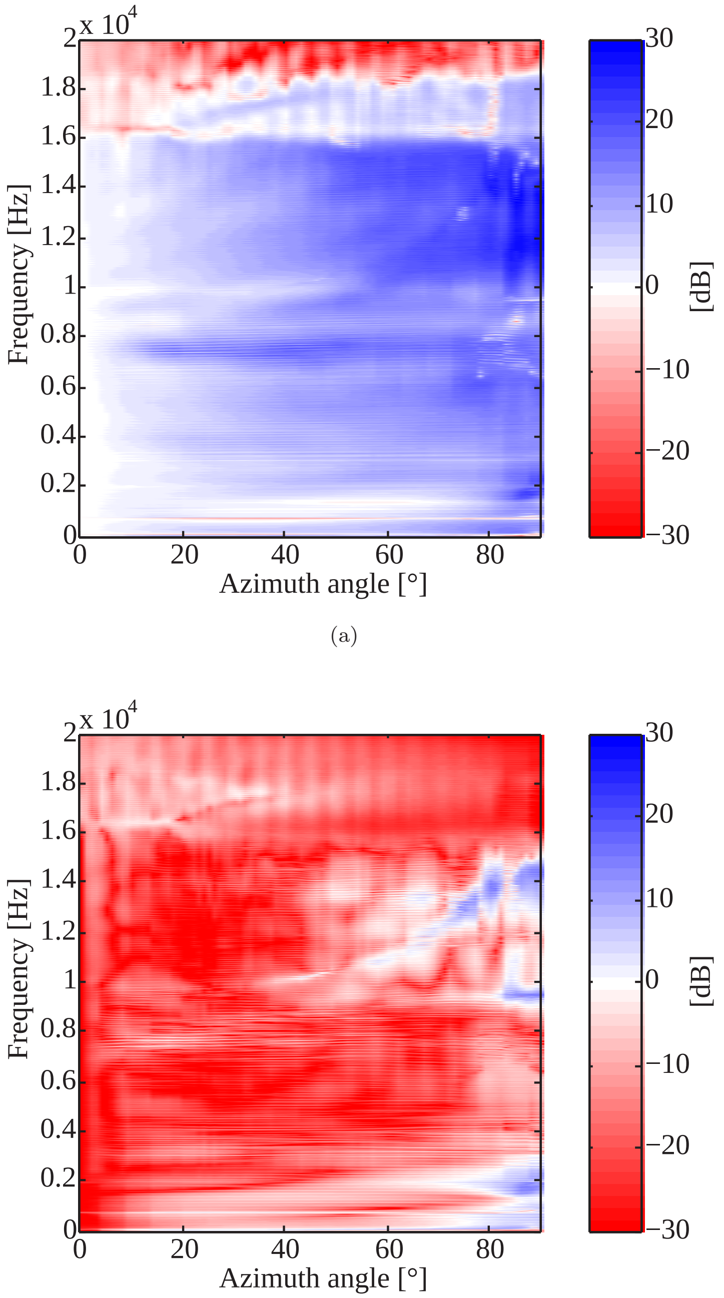


(a)



(b)

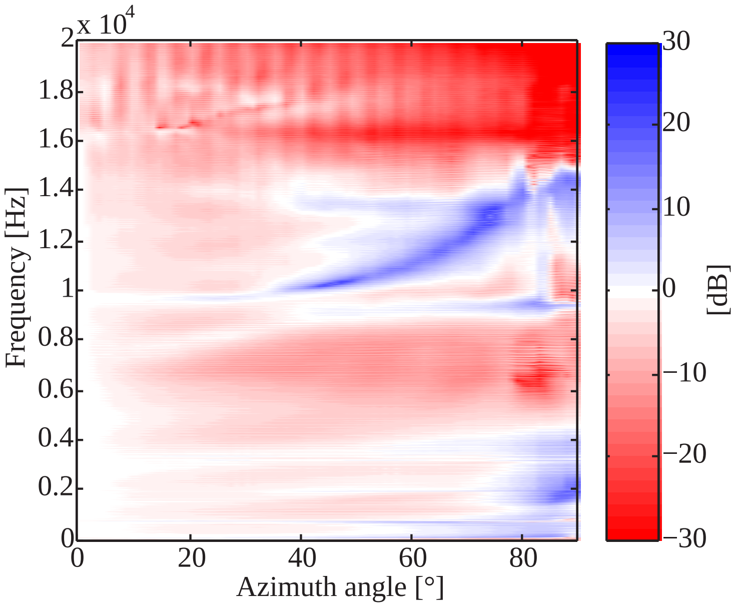
FIGURE 5.19: **Simulated static cross-talk cancellation as a function of azimuth angle for the 2-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to $\beta = 10^{-4}$. The inverse filter length is 4096 and the cross-over frequencies is at 900 Hz. The control performance $P_{12}(k)$ and $P_{22}(k)$ is presented. (c) Left ear ($P_{22}(k)$). (d) Right ear ($P_{12}(k)$).*



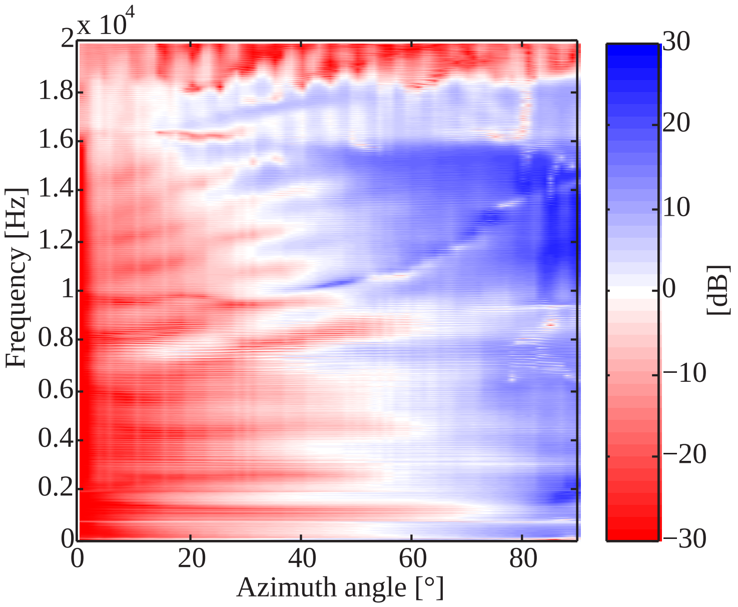
(a)

(b)

FIGURE 5.20: **Simulated static cross-talk cancellation as a function of azimuth angle for the 3-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^{-4}$*. The inverse filter length is 4096 and the cross-over frequencies are at 600 Hz and 4000 Hz. The control performance* $P_{11}(k)$ *and* $P_{21}(k)$ *is presented. (a) Left ear (*$P_{11}(k)$*). (b) Right ear (*$P_{21}(k)$*).*

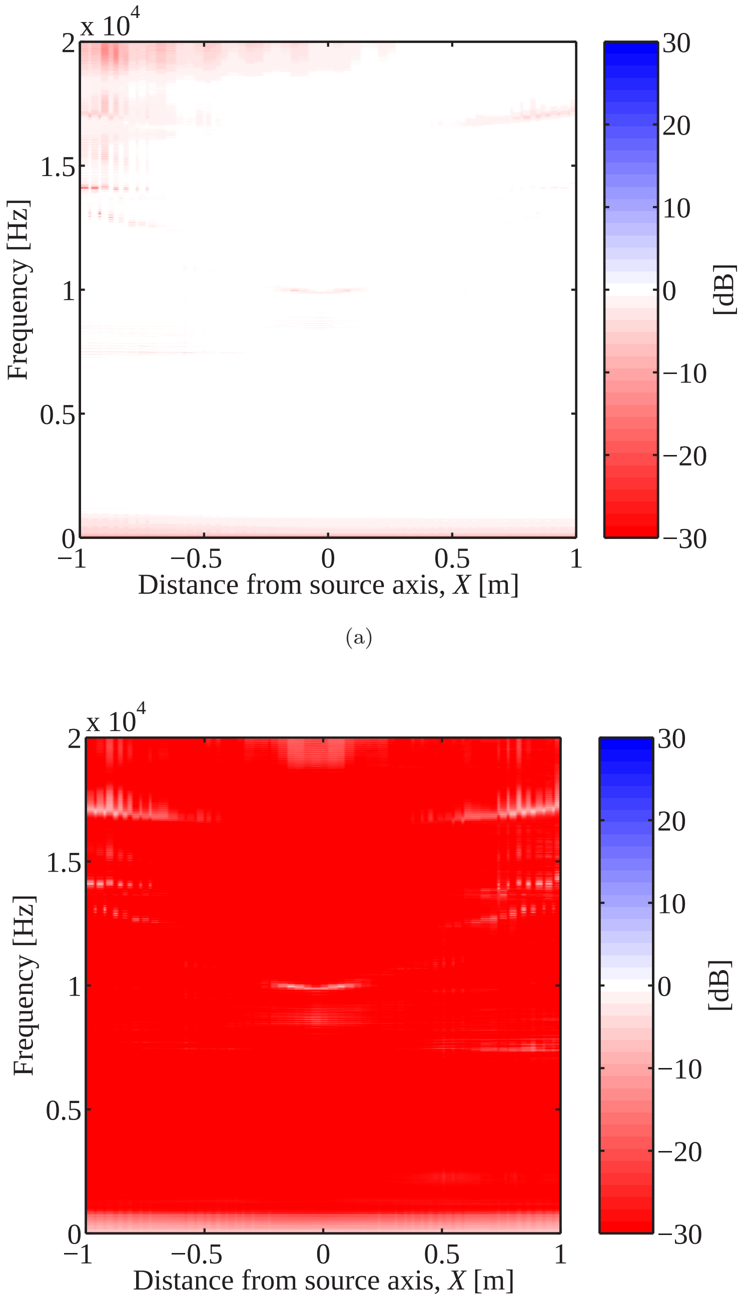


(a)



(b)

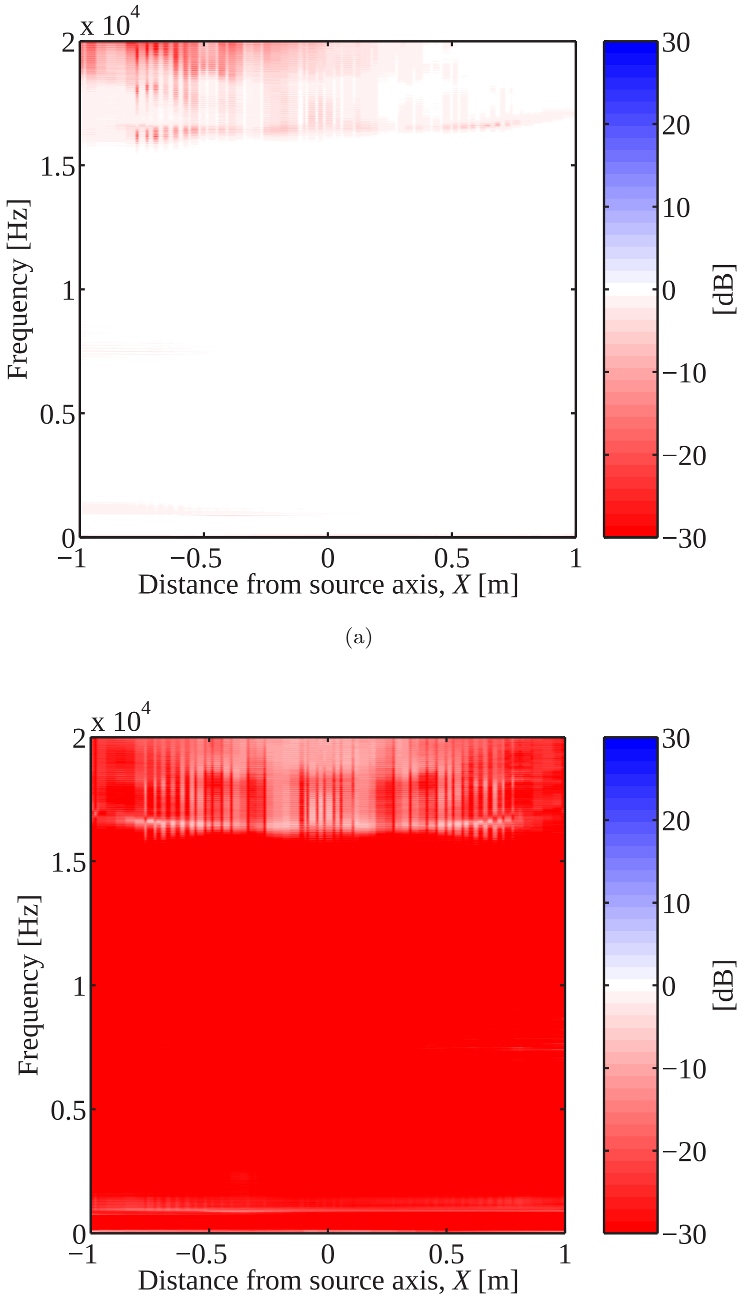
FIGURE 5.21: **Simulated static cross-talk cancellation as a function of azimuth angle for the 3-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to $\beta = 10^{-4}$. The inverse filter length is 4096 and the cross-over frequencies are at 600 Hz and 4000 Hz. The control performance $P_{12}(k)$ and $P_{22}(k)$ is presented. (c) Left ear ($P_{22}(k)$). (d) Right ear ($P_{12}(k)$).*



(a)

(b)

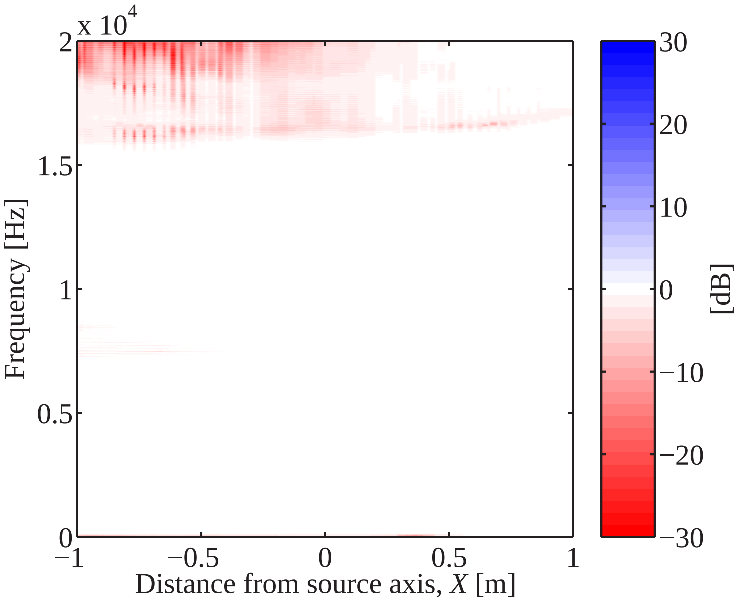
FIGURE 5.22: **Simulated adaptive cross-talk cancellation as a functions of lateral position in the horizontal plane for the SD**. *The ISVR HRTF database is used and the regularisation parameter is set to $\beta = 10^{-4}$. The inverse filter length is 4096. The control performance $P_{11}(k)$ and $P_{21}(k)$ is presented ($P_{22}(k)$ and $P_{12}(k)$ are mirrored versions of $P_{11}(k)$ and $P_{21}(k)$). (a) Left ear ($P_{11}(k)$). (b) Right ear ($P_{21}(k)$).*



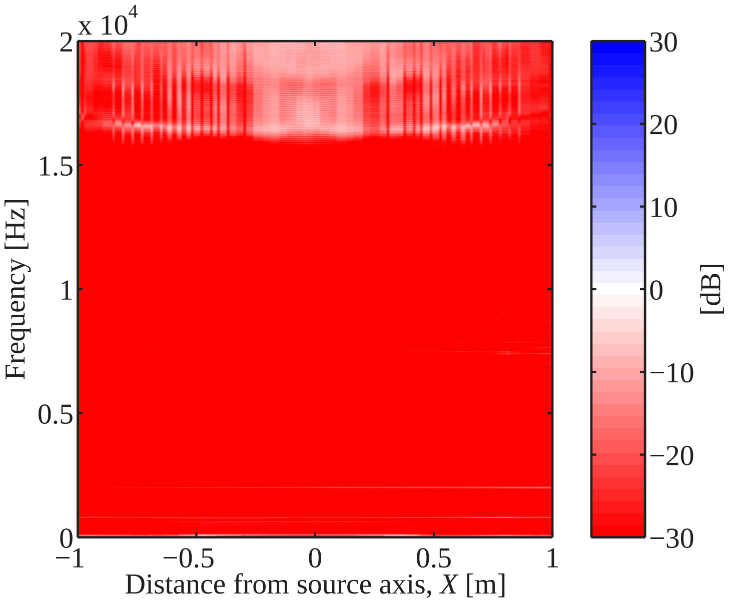
(a)

(b)

Figure 5.23: **Simulated adaptive cross-talk cancellation as a functions of lateral position for the 2-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^{-4}$. *The inverse filter length is 4096 and the cross-over frequency is at 900 Hz. The control performance* $P_{11}(k)$ *and* $P_{21}(k)$ *is presented (*$P_{22}(k)$ *and* $P_{12}(k)$ *are mirrored versions of* $P_{11}(k)$ *and* $P_{21}(k)$*). (a) Left ear (*$P_{11}(k)$*). (b) Right ear (*$P_{21}(k)$*).*

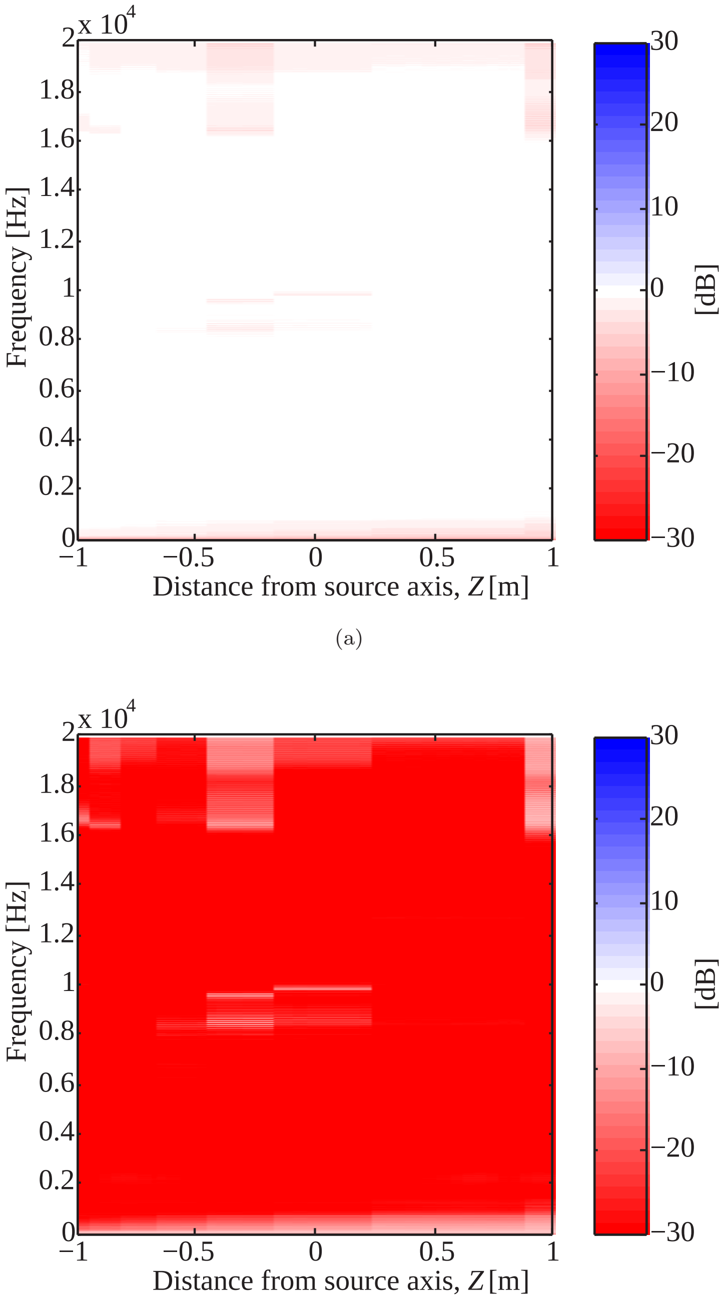


(a)



(b)

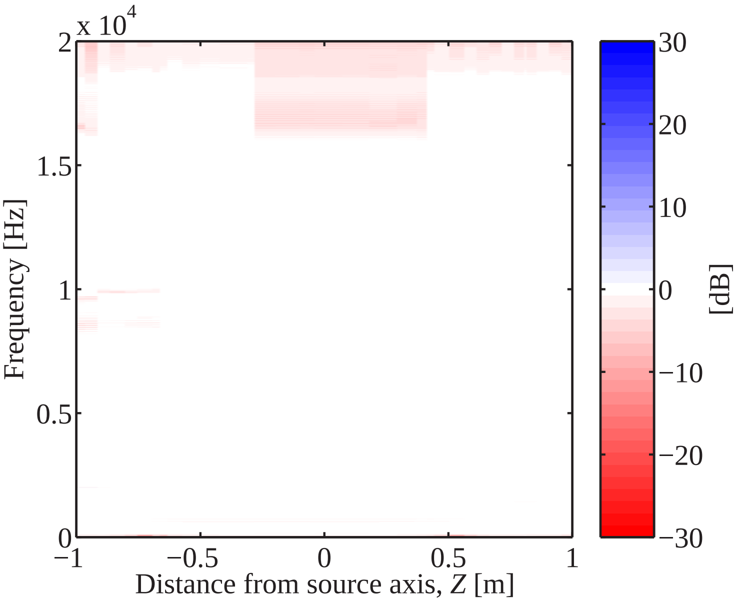
FIGURE 5.24: **Simulated adaptive cross-talk cancellation as a functions of lateral position for the 3-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to $\beta = 10^{-4}$. The inverse filter length is 4096 and the cross-over frequencies are at 600 Hz and 4000 Hz. IThe control performance $P_{11}(k)$ and $P_{21}(k)$ is presented ($P_{22}(k)$ and $P_{12}(k)$ are mirrored versions of $P_{11}(k)$ and $P_{21}(k)$). (a) Left ear ($P_{11}(k)$). (b) Right ear ($P_{21}(k)$).*



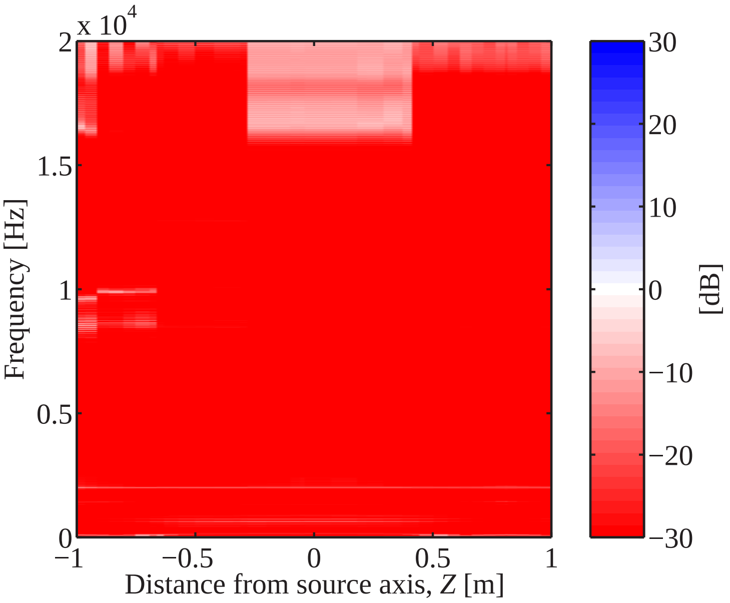
(a)

(b)

FIGURE 5.25: **Simulated adaptive cross-talk cancellation as a functions of fore and aft position in the horizontal plane for the SD**. *The ISVR HRTF database is used and the regularisation parameter is set to $\beta = 10^{-4}$. The inverse filter length is 4096. The control performance $P_{11}(k)$ and $P_{21}(k)$ is presented ($P_{22}(k)$ and $P_{12}(k)$ are mirrored versions of $P_{11}(k)$ and $P_{21}(k)$). (a) Left ear ($P_{11}(k)$). (b) Right ear ($P_{21}(k)$).*

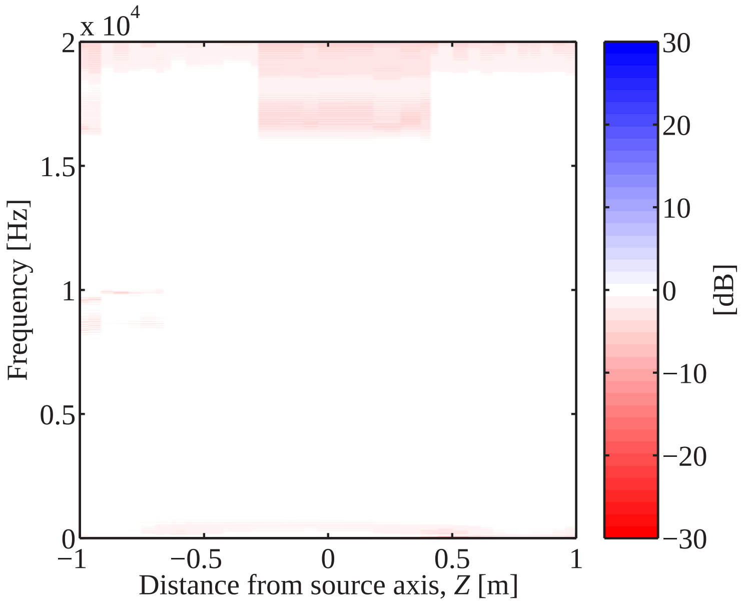


(a)

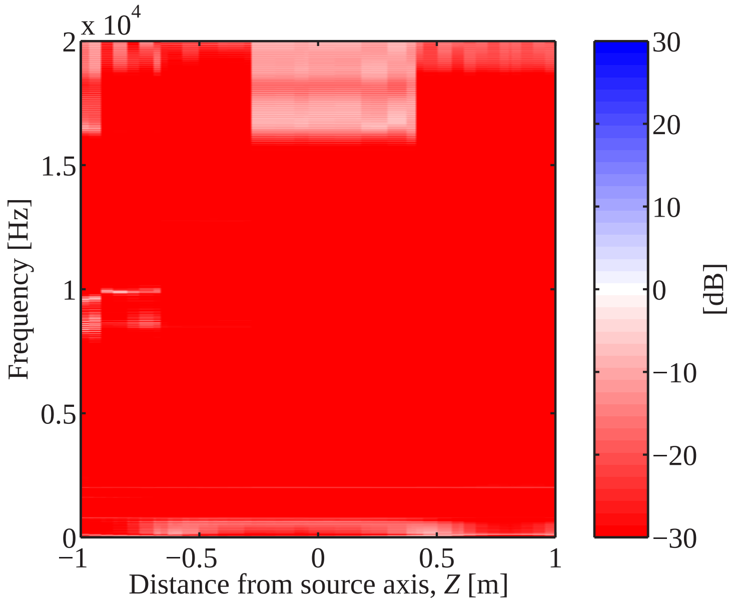


(b)

FIGURE 5.26: **Simulated adaptive cross-talk cancellation as a functions of fore and aft position for the 2-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to $\beta = 10^{-4}$. The inverse filter length is 4096 and the cross-over frequency is at 900 Hz. The control performance $P_{11}(k)$ and $P_{21}(k)$ is presented ($P_{22}(k)$ and $P_{12}(k)$ are mirrored versions of $P_{11}(k)$ and $P_{21}(k)$). (a) Left ear ($P_{11}(k)$). (b) Right ear ($P_{21}(k)$).*

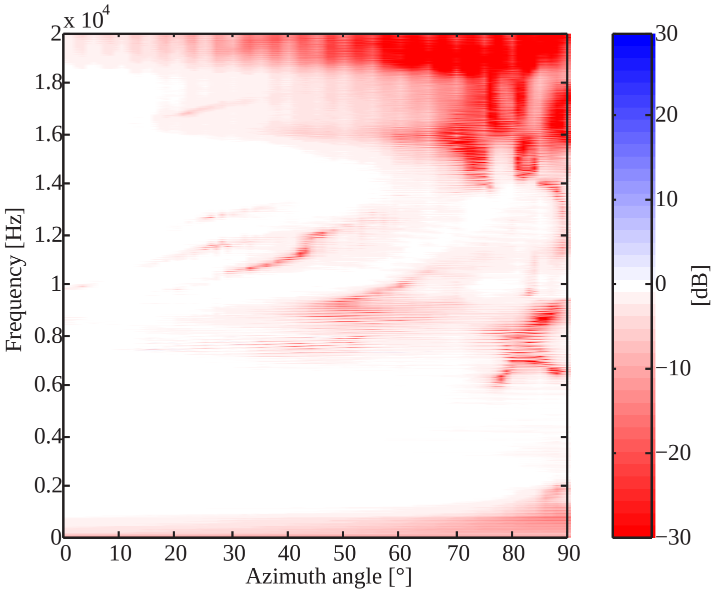


(a)

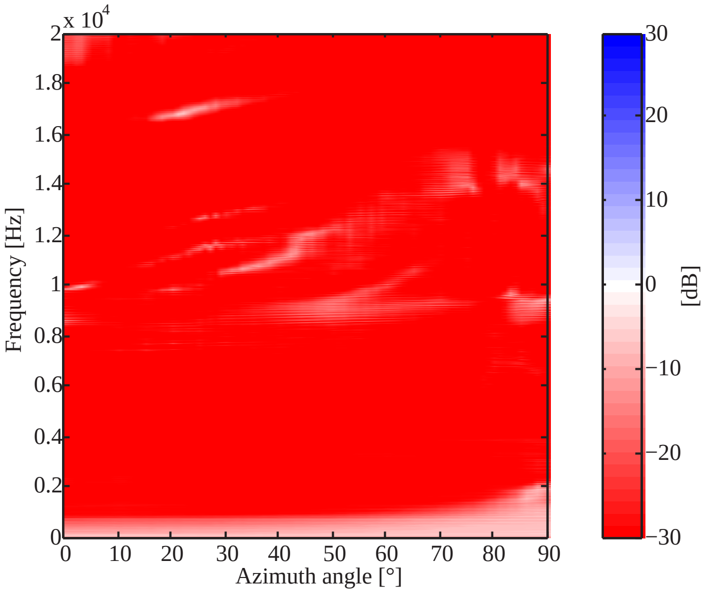


(b)

Figure 5.27: **Simulated adaptive cross-talk cancellation as a functions of fore and aft position for the 3-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^{-4}$*. The inverse filter length is 4096 and the cross-over frequencies are at 600 Hz and 4000 Hz. The control performance* $P_{11}(k)$ *and* $P_{21}(k)$ *is presented (*$P_{22}(k)$ *and* $P_{12}(k)$ *are mirrored versions of* $P_{11}(k)$ *and* $P_{21}(k)$*). (a) Left ear (*$P_{11}(k)$*). (b) Right ear (*$P_{21}(k)$*).*

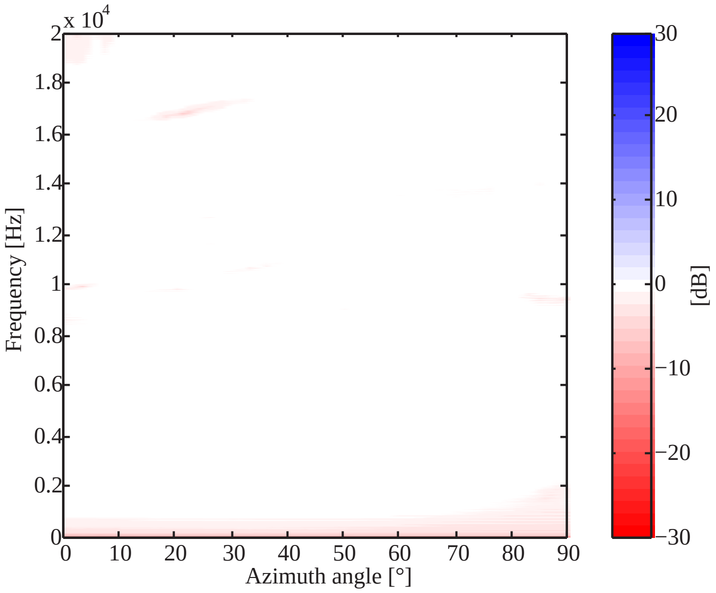


(a)

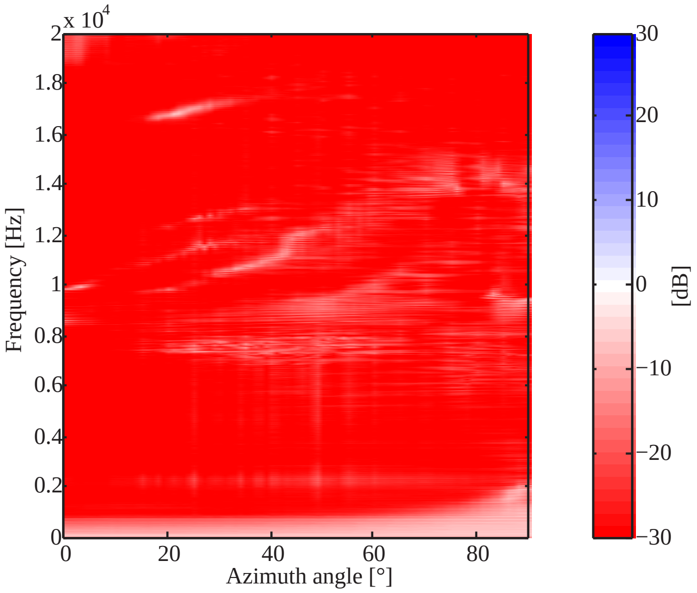


(b)

FIGURE 5.28: **Simulated adaptive cross-talk cancellation as a functions of azimuth angle for the SD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^{-}4$. *The inverse filter length is 4096. The control performance* $P_{11}(k)$ *and* $P_{21}(k)$ *is presented. (a) Left ear (*$P_{11}(k)$*). (b) Right ear (*$P_{21}(k)$*).*

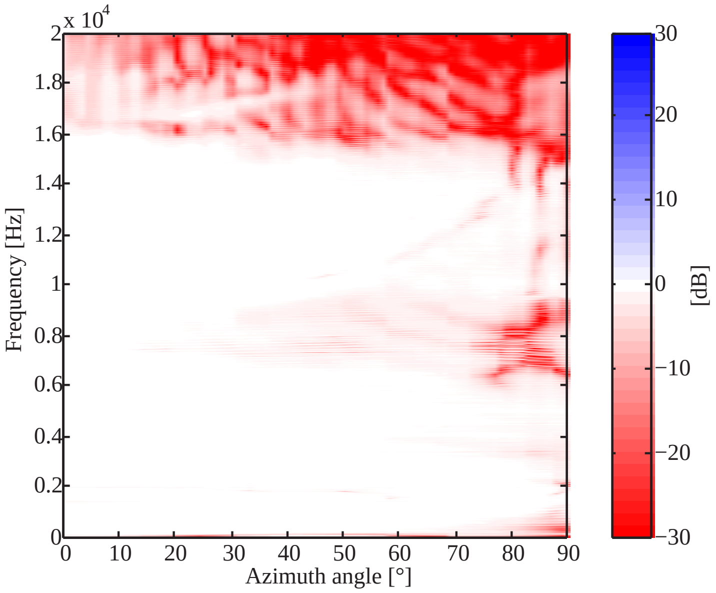


(a)

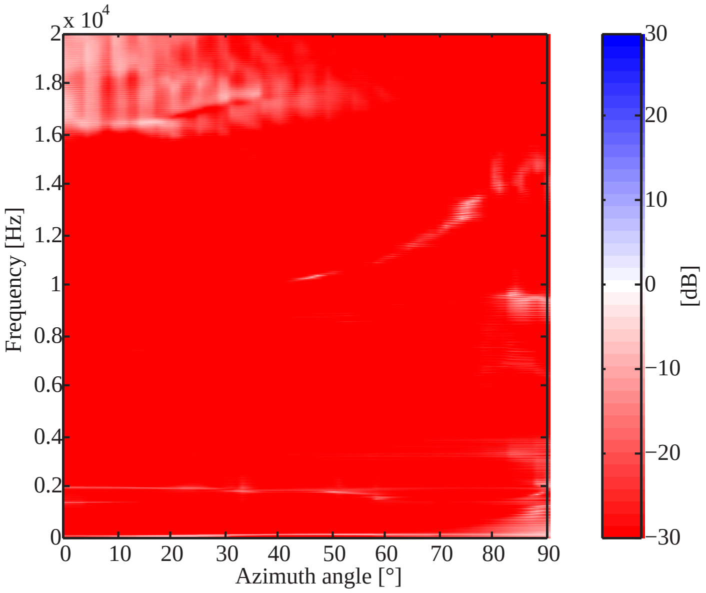


(b)

FIGURE 5.29: **Simulated adaptive cross-talk cancellation as a functions of azimuth angle for the SD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^-4$. *The inverse filter length is 4096. The control performance* $P_{12}(k)$ *and* $P_{22}(k)$ *is presented. (a) Left ear* ($P_{22}(k)$)*. (b) Right ear* ($P_{12}(k)$)*.*

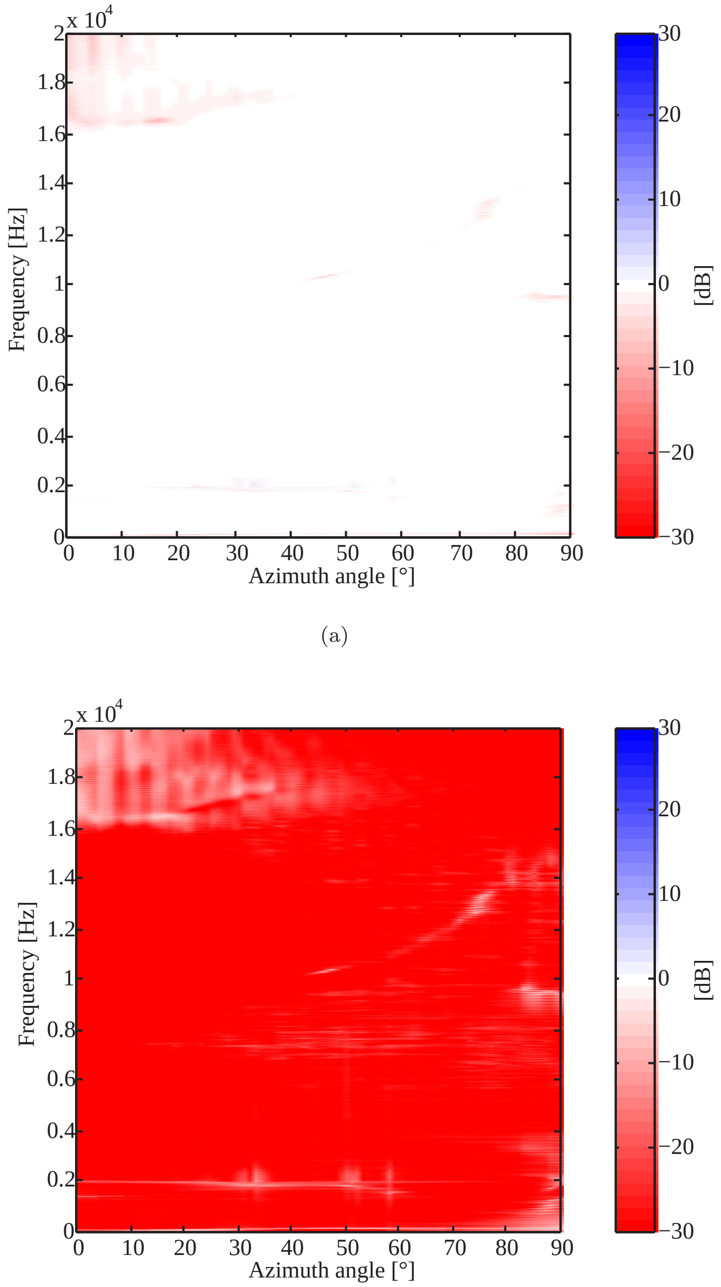


(a)



(b)

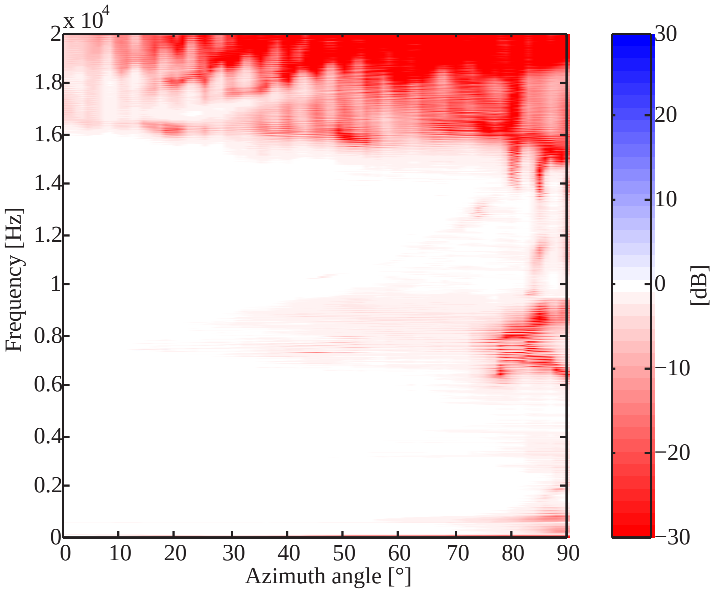
FIGURE 5.30: **Simulated adaptive cross-talk cancellation as a functions of azimuth angle for the 2-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^{-}4$. *The inverse filter length is 4096 and the cross-over frequency is at 900 Hz. The control performance* $P_{11}(k)$ *and* $P_{21}(k)$ *is presented. (a) Left ear* ($P_{11}(k)$)*. (b) Right ear* ($P_{21}(k)$)*.*



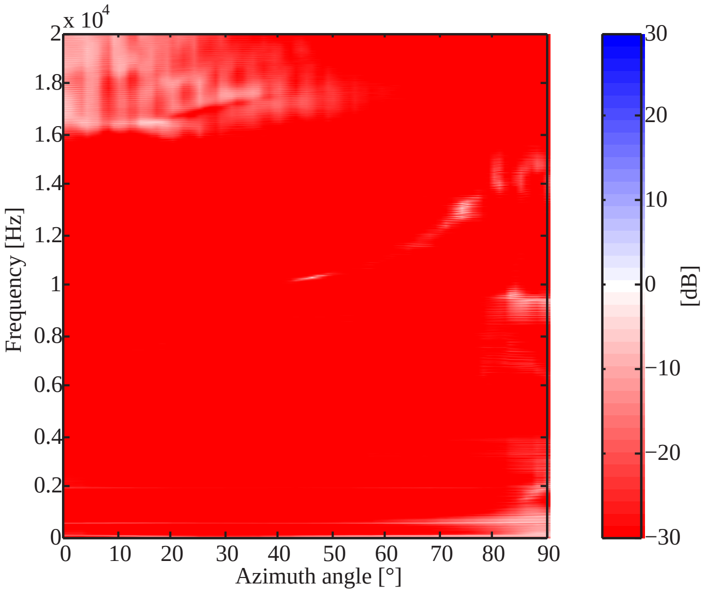
(a)

(b)

FIGURE 5.31: **Simulated adaptive cross-talk cancellation as a functions of azimuth angle for the 2-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^{-}4$*. The inverse filter length is 4096 and the cross-over frequency is at 900 Hz. The control performance* $P_{12}(k)$ *and* $P_{22}(k)$ *is presented. (a) Left ear (*$P_{22}(k)$*). (b) Right ear (*$P_{12}(k)$*).*

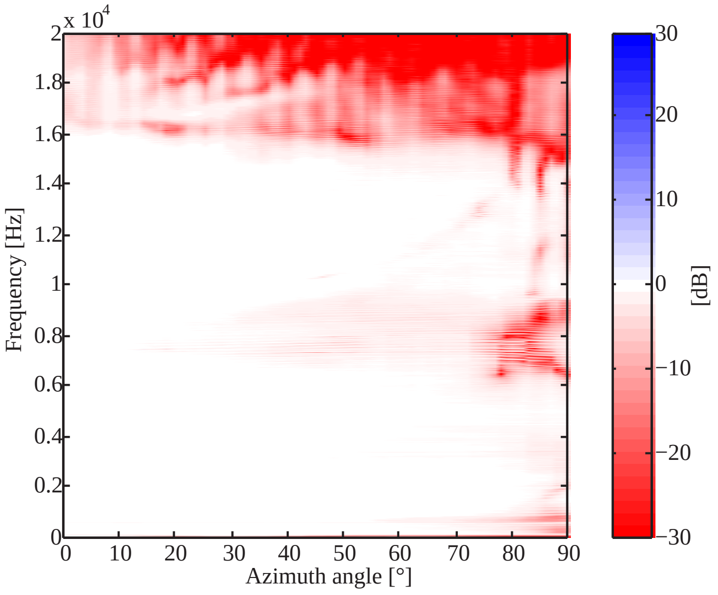


(a)

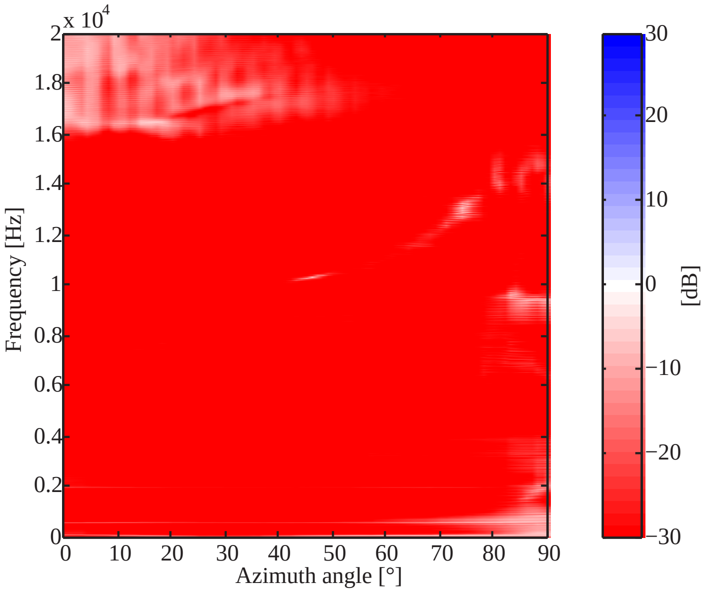


(b)

FIGURE 5.32: **Simulated adaptive cross-talk cancellation as a functions of azimuth angle for the 3-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^-4$. *The inverse filter length is 4096 and the cross-over frequencies are at 600 Hz and 4000 Hz. The control performance* $P_{11}(k)$ *and* $P_{21}(k)$ *is presented. (a) Left ear (*$P_{11}(k)$*). (b) Right ear (*$P_{21}(k)$*).*

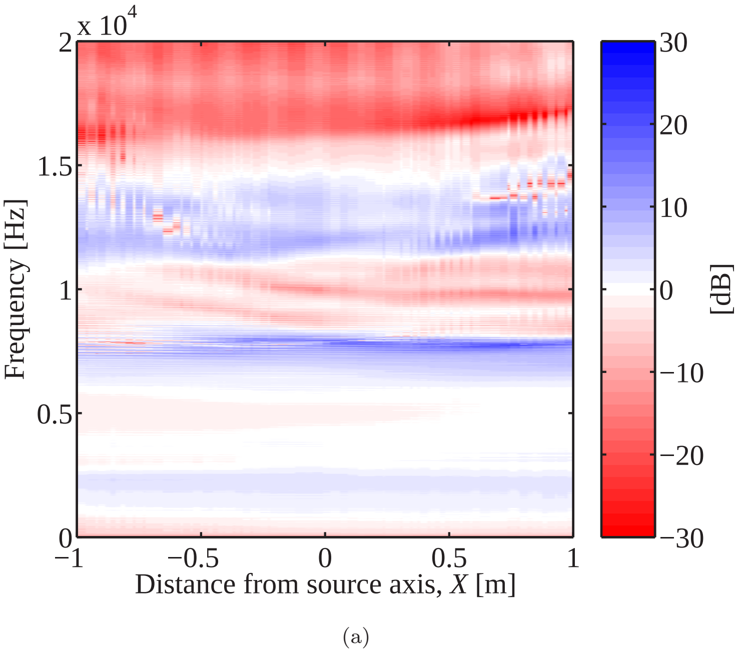


(a)

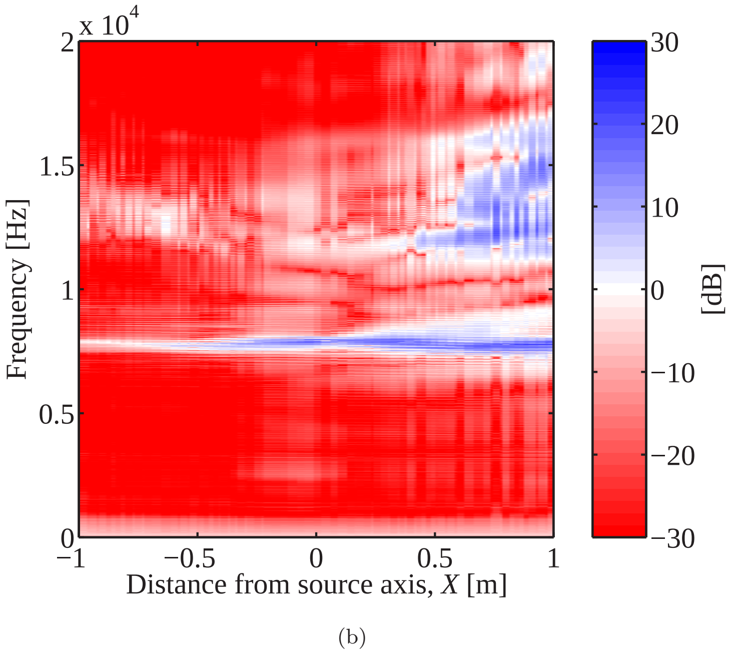


(b)

Figure 5.33: **Simulated adaptive cross-talk cancellation as a functions of azimuth angle for the 3-way OSD**. *The ISVR HRTF database is used and the regularisation parameter is set to $\beta = 10^-4$. The inverse filter length is 4096 and the cross-over frequencies are at 600 Hz and 4000 Hz. The control performance $P_{11}(k)$ and $P_{21}(k)$ is presented. (a) Left ear ($P_{11}(k)$). (b) Right ear ($P_{21}(k)$).*



(a)



(b)

FIGURE 5.34: **Simulated adaptive cross-talk cancellation as a functions of lateral position for the SD with pinna miss-match**. *The "small pinna" is used in the plant model and the "large pinna" is used in the inverse filter. The ISVR HRTF database is used and the regularisation parameter is set to* $\beta = 10^{-4}$. *The inverse filter length is 4096. The control performance* $P_{11}(k)$ *and* $P_{21}(k)$ *is presented. (a) Left ear* ($P_{11}(k)$)*. (b) Right ear* ($P_{21}(k)$)*.*

Chapter 6

Filter update techniques and subjective experiments

Filter updates are necessary when the listener is moving out of the "sweet-spot" since the transfer functions between the listener and the loudspeakers change with position. Two filter update techniques are presented and evaluated, namely direct filter updates using pre-computed inverse filters and filter updates using a reduced database of transfer functions. The filter update rate is an important parameter for virtual sound imaging systems, and determines how often the filter coefficients need to be updated so that the updates can be seamless for the listener. The filter update rate is determined here by defining two criteria: the Just Noticeable Difference (JND) criterion and the Just Noticeable Change (JNC) criterion. The JNC criterion determines the maximum distance between filter updates that does not create any *audible signal changes* for the listener. For example if the listener is changing position and the filter is updated on a too coarse grid then a noticeable "click" sound will appear. Hence, the JNC criterion determines the step size of the filter updates in the commutation process and aims to remove audible artifacts. The JND criterion determines the maximum distance between filter updates that does not create any *changes in the virtual sound image* perceived by the listener. For example if the listener is changing position and the filter is updated on a too coarse grid then it is perceived as if the virtual sound source is changing position while it is not intended to do so. The JND and JNC criteria for filter updates are determined through a set of subjective experiments.

The interpolation process is typically associated with the synthesis of an intermediate transfer function from a database of predefined filters. If a reduced database of transfer functions is used, the interpolation process serves to fill in the associated gaps. Commutation is the process of updating the filter coefficients in real-time while the filter is running (Jot [44]). The filter update can cause audible artifacts and the commutation process can remove those.

Filter updates can be dealt with by using different techniques. A common technique is "cross-fading" that uses two filters in parallel, and the output is computed as a weighted sum of the outputs from the filters using linear interpolation (Huopaniemi [38]). The filter update transition time is the time it takes to completely change the output from the current filter to the new filter. This technique doubles the computational cost of the filter implementation. The filters can be implemented by using techniques such as FIR, WFIR, IIR and WIIR among others (Karjalainen [45], Farina [29], Jot [44]). Another filter update technique that limits transients has been presented by Valimaki [99] and is based on state variables for updating the IIR filter coefficients. This technique can reduce the computational cost of commutation by 50% compared to cross-fading two filters in parallel.

Rose [86] conducted subjective experiments in order to determine a filter update criterion for an adaptive SD system. The results show that the filter update criterion is about 0.03 m to achieve the most stable virtual sound images. It was also found that virtual sound image locations far from the loudspeakers require a smaller "filter update movement increment" criterion than virtual sound images close to the loudspeakers. The "filter update movement increment" is similar to the JND criterion determined here.

Localisation blur is defined as the threshold for detecting changes in virtual source angles. A survey for localisation blur in the horizontal plane is presented by Blauert [13] where the localisation blur varies between 1.4° and 3.3° for broadband/narrowband noise signals. The localisation blur studies could be used to form a JND criterion using an objective evaluation of the filter updates. Moore [70] points out that the smallest detectable intensity change for wide-band noise and band-pass filtered noise is approximately a constant fraction of the intensity of the stimulus. This fraction is called Weber's law, which states that the smallest detectable change in a stimulus is proportional to the magnitude of the stimulus. The smallest detectable change for broadband noise is about 0.5 - 1 dB. This could be used as an alternative method to form a JNC criterion for the filter updates.

6.1 A review of alternative filter update techniques

The problem of updating inverse filters in real-time can be resolved by using the following two alternative techniques, direct filter updates using pre-computed inverse filters and filter updates using a reduced database of transfer functions. The presented filter update techniques uses direct filter coefficient updates for a total of four FIR filters to implement adaptive cross-talk cancellation for a single listener. These techniques require only half the computational cost for the filter updates compared to cross-fading that uses two filters in parallel. The filter update techniques can be used with both SD and FDL type of systems by using superposition of the different frequency bands as described in Section

5.1 in the FDL case. A block diagram of the adaptive virtual sound imaging system is presented in Figure 6.1. The signals and transfer functions in Figure 6.1 are defined as follows. The transducers produce source strengths from the input voltages defined by the elements of the complex vector $\mathbf{v}(k) = [V_1(k)\ V_2(k)]^T$. The resulting acoustic pressure signals are defined by the vector $\mathbf{w}(k) = [W_1(k)\ W_2(k)]^T$. The transfer functions for a generelised binaural sound reproduction system as a function of time t are named $\mathbf{\Psi}_t(k)$ and the corresponding inverse transfer functions are named $\mathbf{\Pi}_t(k)$. The acoustic pressure signals are found by

$$\mathbf{w}(k) = \mathbf{\Psi}_t(k)\mathbf{v}(k) \tag{6.1}$$

The two signals to be reproduced at the listener's ears are defined by the complex vector $\mathbf{d}(k) = [D_1(k)\ D_2(k)]^T$. The desired signals $\mathbf{d}(k)$ to be reproduced is a delayed version of the recorded or synthesised signals $\mathbf{u}(k)$, hence $\mathbf{d}(k) = \mathbf{u}(k)e^{-j\omega\Delta}$. The signals $\mathbf{u}(k)$ can be obtained from a dummy head recording or from filtering signals $\mathbf{x}(k)$ by a matrix of binaural filters $\mathbf{A}(k) = [A_1(k)\ A_2(k)]^T$. The source strengths can now be found by

$$\mathbf{v}(k) = \mathbf{\Pi}_t(k)\mathbf{d}(k) \tag{6.2}$$

where $\mathbf{\Pi}_t(k)$ contains inverse filters that also change with time t as the listener enters a new position such that

$$\mathbf{\Pi}_t(k) = \begin{bmatrix} \Pi_{11t}(k) & \Pi_{12t}(k) \\ \Pi_{21t}(k) & \Pi_{22t}(k) \end{bmatrix} \tag{6.3}$$

and thus

$$\mathbf{w}(k) = \mathbf{\Pi}_t(k)\mathbf{\Psi}_t(k)\mathbf{d}(k) \tag{6.4}$$

where

$$\mathbf{\Psi}_t(k) = \begin{bmatrix} \Psi_{11t}(k) & \Psi_{12t}(k) \\ \Psi_{21t}(k) & \Psi_{22t}(k) \end{bmatrix} \tag{6.5}$$

6.1.1 Direct filter updates

The results from the subjective filter update experiment for JNC in Section 6.2 can be used to establish a technique based on direct updates of inverse filters. The direct update of inverse filter technique works by using a large database of pre-computed inverse filters (look-up table) that are updated at steps of the JNC criterion for lateral movement. Hence, as the listener moves the inverse filters that correspond to the geometry of the loudspeakers and the listener are updated at the JNC criteria for lateral movement. The technique is based on filter coefficient updates. The algorithm essentially picks up the inverse filters that correspond to the position the listener that is located with a precision corresponding to the JNC criterion and then updates the filter coefficients accordingly. The advantage of this technique is that it requires only a small amount of computing power for updating the filters. The disadvantage is that it requires a large database of inverse filters, which increases the memory requirements.

6.1.1.1 Summary of algorithm

The database contains inverse filters for pre-defined listener positions that are spaced according to the JNC criterion.

(1) Acquire the listener position from the head tracker.

(2) If the threshold for filter update is breached, then update the inverse filters $\mathbf{\Pi}_t(k)$ directly at an update rate, which ideally is equal to or greater than the filter decay time (the number of FIR filter coefficients multiplied by the sampling period). The threshold is based on the JNC criterion.

(3) Return to (1).

Note that if the precision of the tracking system can not match the JNC criterion then it is possible establish an appropriate filter update threshold. The filter update threshold should ideally never exceed the JND criterion. The filter update is then carried out by performing commutation in between the new and the old listener position using pre-computed filters.

6.1.2 Filter updates using a reduced database of transfer functions

A filter update algorithm is presented that can reduce the size of the database to a minimum in terms of filter design positions. The technique uses commutation and real-time inversion. The reduced database holds only transfer functions for three listener positions in the lateral plane. The three sets of HRIRs correspond to the far end of each direction of the lateral plane where the listener is allowed to move and the centre position. The transfer functions in the database are denoted $\mathbf{\Psi}_l(k), \mathbf{\Psi}_c(k), \mathbf{\Psi}_r(k)$ where l represents the far left position, c represents the centre position and r corresponds to the far right position. The listener can in this example move ±0.7 m in the lateral plane relative to on-axis at a distance of 2 m in the fore and aft plane. The amount of listener movement is established by allowing interpolation for azimuth angles up to 20°. This will ensure that the average mean square interpolation error is less than about 1.74%, which is valid for frontal directions as stated in Table 4.1.

The HRIRs are stored in an ITD-equalised state such that interpolation can be applied directly to the coefficients. The interpolation is performed by using linear interpolation as described in Section 4.1. The appropriate amount of delay is inserted after the interpolation stage using a delay line. The delay line should be able to handle fractional delays (Laakso [57]) if the sampling rate is lower than approximately 100 kHz, since the ITD threshold is approximately 10 μs (Klumpp and Eady [52]). The algorithm computes the inverse filters in between two pre-measured filters in the database to achieve a spacing of no less than the JNC criterion during the filter update. The inverse filters are computed using the fast-deconvolution algorithm. A filter update is initiated every time the JND criterion is breached by the tracker. When a filter update is initiated, then the inverse filters are computed with a spacing of the JNC criterion in between the old and the new filter positions that has a spacing of the JND criterion. The number of inverse filters during update is given by $N_u = 1/((\text{JNC mod JND})/\text{JND})$. When the new inverse filters are available then the filters are updated with a step size of the JNC criterion and at an update rate that ideally is equal to or greater than the filter decay time. The advantage of this method is the reduced size of the database of transfer functions. The disadvantage with commutation between filters using a reduced database is the computational overhead that the interpolation and the fast-deconvolution algorithm are imposing.

6.1.2.1 Summary of algorithm

The spacing of the database is 20° in azimuth angle as in the simulation in Figure 4.4.

(1) Get the new listener position from the head tracker.

(2) If the threshold for filter update is breached, then compute N_u new sets of inverse filters $\mathbf{\Pi}_1(k)$,...,$\mathbf{\Pi}_{N_u}(k)$ using the fast-deconvolution algorithm and the database with transfer functions $\boldsymbol{\Psi}_l(k)$, $\boldsymbol{\Psi}_c(k)$, $\boldsymbol{\Psi}_r(k)$. The new set of inverse filter positions are spaced with the JNC criterion. The two transfer functions that are the nearest neighbours to the new position are used for computing the new set of inverse filters.

(3) Update the inverse filters $\mathbf{\Pi}_t(k)$ at an update rate, which ideally is equal to or greater than the filter decay time.

(4) Return to (1).

6.2 Subjective evaluation of filter update techniques

The filter update criteria are determined through a series of subjective experiments carried out both under anechoic conditions and in a listening room. The intended filter design position was moved relative to the listener at carefully controlled rates. The intended filter design position is defined as the position where the listener is ideally located for the cross-talk cancellation process to perform optimally. When the intended filter design position is moved away from the centre of the listener, then the performance of the cross-talk cancellation process starts to degrade, which in turn leads to a degradation in the sound source localisation ability of the listener. The listeners were wearing a mask in front of their eyes so that no visual cues were available. The subjective experiments that have been carried out here applies directly to the filter update techniques presented in Section 6.1.

A filter update criterion for an adaptive SD system was determined in a set of experiments by Rose [87]. The virtual imaging system was moved in the lateral plane in front of several subjects. The first experiment was performed for determining the "sweet-spot" size in a static case, where the subjects had a fixed position and the SD was moved into different positions in the lateral plane. The sound source localization ability was determined by asking the subject for the perceived angle of a virtual source. The results from this experiment show that the maximum tolerable distance between filter updates is about 0.03 m. The objective of the second experiment was to find the threshold parameters necessary for two real moving sound sources to produce a stationary virtual sound image. The subjects in the experiment were asked to respond to both the location and the stability of the virtual sound image as the SD was moving and the filters were updated. The results show that filter update increments corresponding to lateral movements less than 0.03 m achieve the most stable virtual sound image of the examined increments. Above this filter update rate, there was a steady deterioration of the stability of the virtual sound image.

A set of subjective experiments has been carried out to determine JND and JNC for binaural sound reproduction using loudspeakers under both anechoic and more realistic conditions (listening room). Firstly, the electro-acoustic transfer functions of the loudspeaker were measured at asymmetric listener positions. The loudspeaker system that was used is a 4-way FDL with the following source spans 2θ: 5.8°, 15.4°, 41.2° and 57.6°. The transfer functions were measured with a spacing of 0.025 m both in the lateral plane ($-0.175 \leq X \leq 0.175$) and in the fore and aft plane ($-0.175 \leq Z \leq 0.175$) using the KEMAR dummy head. The transfer functions were interpolated to achieve a spacing of 0.0125 m by using the interpolation algorithm described in Section 4.1. Then inverse filters for the listener positions were designed. Finally, the cross-talk cancellation performance was verified and filter update rates were determined by the subjective experiments presented in Section 6.2.1 and Section 6.2.2.

6.2.1 Filter update rates under anechoic conditions

A subjective experiment was carried out in order to determine the JND and JNC criteria. The binaural stimuli was created by convolving band limited white noise (300 Hz - 3000 Hz) with a set HRTFs corresponding to a virtual source at 40° to the right of the listener. The upper limit of the band-pass filter was chosen to 3000 Hz and the lower limit of the band-pass filter was chosen to cut off frequencies below 300 Hz. The "pinna notch" was excluded in order minimize uncertainties regarding differences between the subject's pinnae. It is noted that the high-pass cross-over frequency for the high frequency unit of the 4-way FDL is 4000 Hz, hence the high frequency unit did not contribute to the results in this experiment. The binaural stimuli for the JND criterion were convolved with the cross-talk cancellation filters for the intended filter design positions and recorded to a CD that was played back to the subjects. The stimuli for the JNC criterion was created by updating the FIR filters with the Huron work station and recording using a Digital Audio Tape (DAT) player and a KEMAR dummy head (Burkard [16]). The recorded binaural stimuli was then convolved with cross-talk cancellation filters and recorded to a CD, which was played back to the subjects. The JNC criterion that determines the maximum distance between filter updates was evaluated for an update rate of 0.05 m/s for changing the intended filter design position. The experiments were carried out by changing only the intended filter design position while the listener and the loudspeakers stayed in fixed positions.

The JND criterion was investigated by presenting a virtual noise source to the listener and changing the intended filter design position. The intended filter design position was changed in the lateral plane from $-0.175 \leq X \leq 0.175$ m with a spacing of 0.0125 m. Hence, the position was changed more than half a head width. The experiment was carried out by changing the intended filter design position in four directions as illustrated in Figure 6.2. The stimuli were convolved with the filter for each intended design position such that a virtual source at $\phi = 40°$ to the right would appear for the subject. The stimuli were presented for one second at each intended filter design position and after the stimuli a pause of five seconds was imposed so that the subject could point the perceived angle of the virtual noise source.

The JNC criterion was investigated by presenting a virtual noise source and changing the intended filter design position dynamically. The stimuli were convolved with the filter for each intended design position such that a virtual noise source at $\phi = 40°$ to the right appeared for the subject. The intended filter design position was switched in the lateral plane between two positions, where one of the positions was the centre of the head and the other an off-axis position. The spacing between the two positions was as follows: 0.0125 m, 0.025 m, 0.0375 m, 0.05 m and 0.065 m. The stimuli were presented for five seconds while the intended filter design positions were switched back and forth. After the stimuli a pause of five seconds was imposed such that the subject could report if the filter changes were audible or not. The intended filter design position was changed in the lateral directions (3) and (4) as illustrated in Figure 6.2.

A preliminary subjective experiment was carried out in the fore and aft plane, in order to determine the JND criterion in this direction. The experiment was carried out using the same process as in the lateral plane. The JND criterion was investigated by presenting a virtual noise source at $\phi = 40°$ to the right of the listener as in the previous case. The experiment was carried out by changing the intended filter design position in four directions as illustrated in Figure 6.3.

The results presented are from fifteen subjects out of twenty, where data from five of the subjects was disregarded by performing an ANOVA (Analysis of Variance) test. The Matlab [65] function **anova1**() was used for carrying out the ANOVA test with a p-value ("statistical significance" measure) of 0.05. The subjective responses in the lateral plane of the virtual source localisation test in the JND experiment is presented in Figure 6.4. The mean values and standard deviation from fourteen measurement points in the lateral plane with a spacing of 0.0125 m is presented for four directions of movement. Now the individual off-axis positions were investigated starting with $X = 0.025$ in order to determine the JND criterion with a certain amount of probability. Figure 6.5 illustrates the data for an off-axis listener position of 0.025 m. The mean virtual source localisation change for this off-axis position is 2.3° and the standard deviation is 4.0°. The combined data from all of the four types of movements were used here. The data was un-biased such that the directionality of the virtual source angle change was the same for all investigated directions of movement. Here, the localisation blur is defined as the amount of displacement of the position of the virtual sound source that is recognised by 50% of the subjects as a change in the the position of the auditory event. This definition is common practice in psychoacoustics (Blauert [13]). Under the assumption of a normal distribution, for 50% probability of a stable virtual sound image, then the localisation blur must be no less than ±3.2°. Then the off-axis position of $X = 0.0375$ was investigated and it was found that the mean virtual source localisation change is 5.6° and the standard deviation is 5.8°, which is greater than in the localisation blur studies summarised by Blauert [13]. Hence, the result from the JND experiment in the lateral plane suggests that the inverse filters should be updated as often as approximately 0.025 m for the sound image to be stable with respect to virtual source angle.

The subjective responses for the JNC experiment is illustrated in Figure 6.6. The histogram shows the probability density of the data that was collected in the anechoic chamber. The assumption was made that the threshold for detecting filter change is in between the two readings where the subjects could detect the filter change and where they could not detect it. For example, a subject that could detect filter change with a filter update rate of 0.025 m but not detect a filter update rate of 0.0125 m, then the threshold is given by (0.025-0.0125/2)=0.01875 m. Here, the mean value of the data is 0.016 m. The probability density of the collected data was then represented by the lognormal probability density function as illustrated in Figure 6.6 (b). The mean and standard deviation associated with the lognormal function was estimated using the Maximum Likelihood Estimator (MLE) function **mle** in Matlab [65]. Under the assumption that the collected data can be represented by a lognormal distribution, for 50% probability of not detecting any filter change, then the JNC threshold is 0.012 m in the anechoic case. The result from the JNC experiment in the lateral plane suggests that the inverse filters using the direct filter update technique should be updated more often than 0.012 m.

The result from the JND experiment in the fore and aft plane suggests that the inverse filters can be updated less often than 0.15 m for the image to be stable with respect to virtual source angle. This can be expected, since the ITD does not change significantly in the fore and aft direction. The subjective response for JND in the fore and aft plane is presented in Figure 6.7.

6.2.2 Filter update rates in a listening room

This section determines filter update rates through a series of subjective experiments carried out under realistic listening conditions with intended filter design positions moved relative to the listener at carefully controlled rates. The experiment was carried out in the listening room at the ISVR, which was designed to be equivalent to realistic conditions for homes according to the IEC 268-13 standard. The procedures of the experiment were the same as those described in Section 6.2 and was performed on fourteen subjects.

The results from the JND experiment in the listening room presented here are from thirteen subjects out of fourteen. One of the subject's data were disregarded by performing an ANOVA test as in Section 6.2.1. The results show good agreement between the listening room and the anechoic chamber. The subjective response represented by mean values and standard deviation of virtual source localisation in the lateral plane is presented in Figure 6.8. Again, the mean values and standard deviation from fourteen measurement points in the lateral plane with a spacing of 0.0125 m is presented for four directions of movement. Figure 6.9 illustrates the data for an off-axis listener position of 0.025 m where the mean virtual source localisation change for position is 3.3° and likewise the standard deviation is 3.3°. Under the assumption of a normal distribution, for 50% probability of a stable virtual image, a localisation blur of ±3.4° must be allowed, which is close to the result for the anechoic case. The result from the subjective experiment in the listening room shows that the inverse filters should be updated as often as approximately 0.025 m for the image to be stable with respect to virtual source angle, as in the anechoic case.

Figure 6.10 illustrates the JNC data from the experiment that was carried out in the listening room. The mean value of this data is 0.017 m. The data was processed presented in the same way as for the anechoic case where the mean and standard deviation associated with the lognormal function was estimated using the MLE. Under the assumption that the collected data can be represented by a lognormal distribution, for 50% probability of not detecting any filter change, then the JNC threshold is 0.012 m in the listening room. As in the anechoic case, the result from the JNC experiment shows that the inverse filters using the direct filter update technique should be updated more often than 0.012 m.

The subjective experiment in the listening room shows no significant deviations from the anechoic chamber. The main difference that can be seen is the higher variance when the intended filter design position was far away from the on-axis position. This is likely to be a result of room reflections that makes the sound more diffuse. The results indicates that sound spatialisation using loudspeakers works well both under anechoic and non-anechoic listening conditions. The threshold for detecting filter changes was also very close to the anechoic case. This is likely to be due to the "precedence effect" that has a strong influence on sound source localisation ability. The precedence effect describes a group of phenomena that are used in resolving competition for perception and localisation between a reflection and direct sound (Litovskya [61]). In a reverberant environment such as the listening room the sound reach the ears by taking several different paths and although these reflections would be audible in isolation, the first arriving wave front dominates the performance of sound source localisation.

6.3 Conclusion

Two different approaches to the solution of the problem of updating inverse filters in real-time have been proposed. The first and simplest technique is to update the inverse filters directly by using an inverse filter database with a fine spacing that complies with the JNC criterion. This technique should be used when it is required that the computational cost is low and when the memory can be relatively large. The second technique reduces the database to a minimum by using real-time interpolation and real-time inversion for updating the inverse filters. This technique should be used when it is required to have a small memory but the computational cost can be allowed to be relatively high. The reduced database uses only filters for three listener positions in the lateral plane.

Filter update rates have been determined by defining the JND criterion and the JNC criterion. The subjective evaluations in the anechoic chamber show that the JND criterion is approximately 0.025 m for movements in the lateral plane. The JNC criterion is approximately 0.012 m for movements in the lateral plane. The JND criterion for movements in the fore and aft plane is greater than the distance of ±0.15 m that was explored. A complementary subjective experiment was carried out in the listening room using the same procedure as in the anechoic chamber. The results from the listening room showed no significant differences compared to the anechoic chamber. These results suggest that the filter update techniques and sound source localisation with loudspeakers works well both under anechoic and non-anechoic listening conditions.

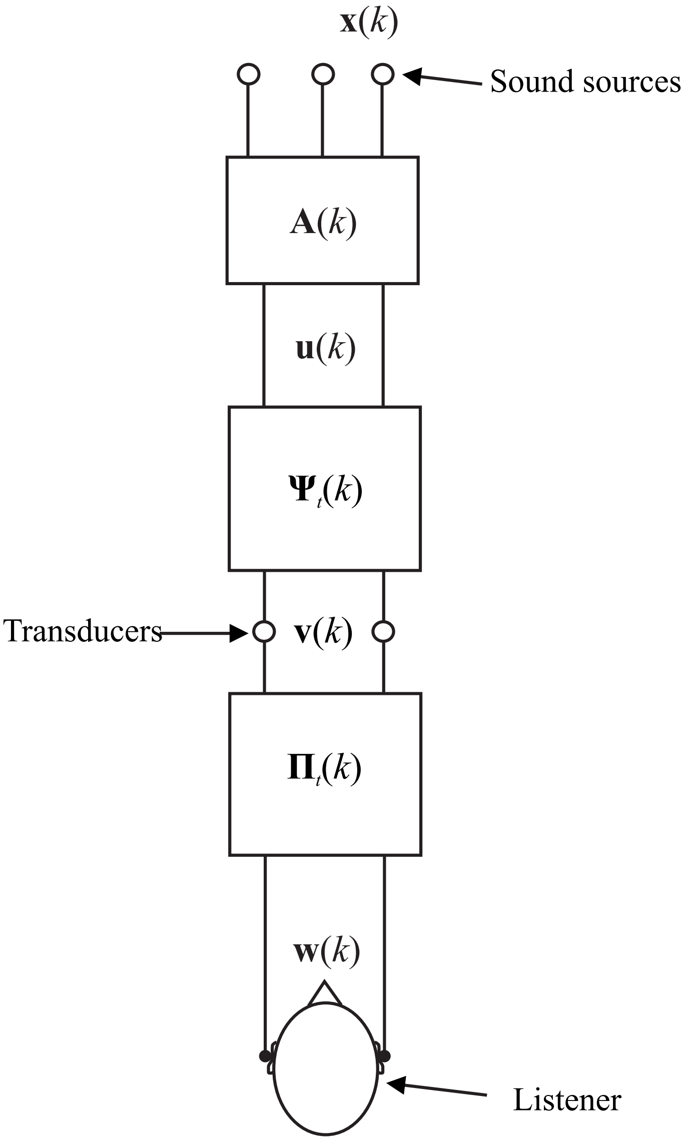


FIGURE 6.1: **Block diagram of the adaptive virtual sound imaging system**.

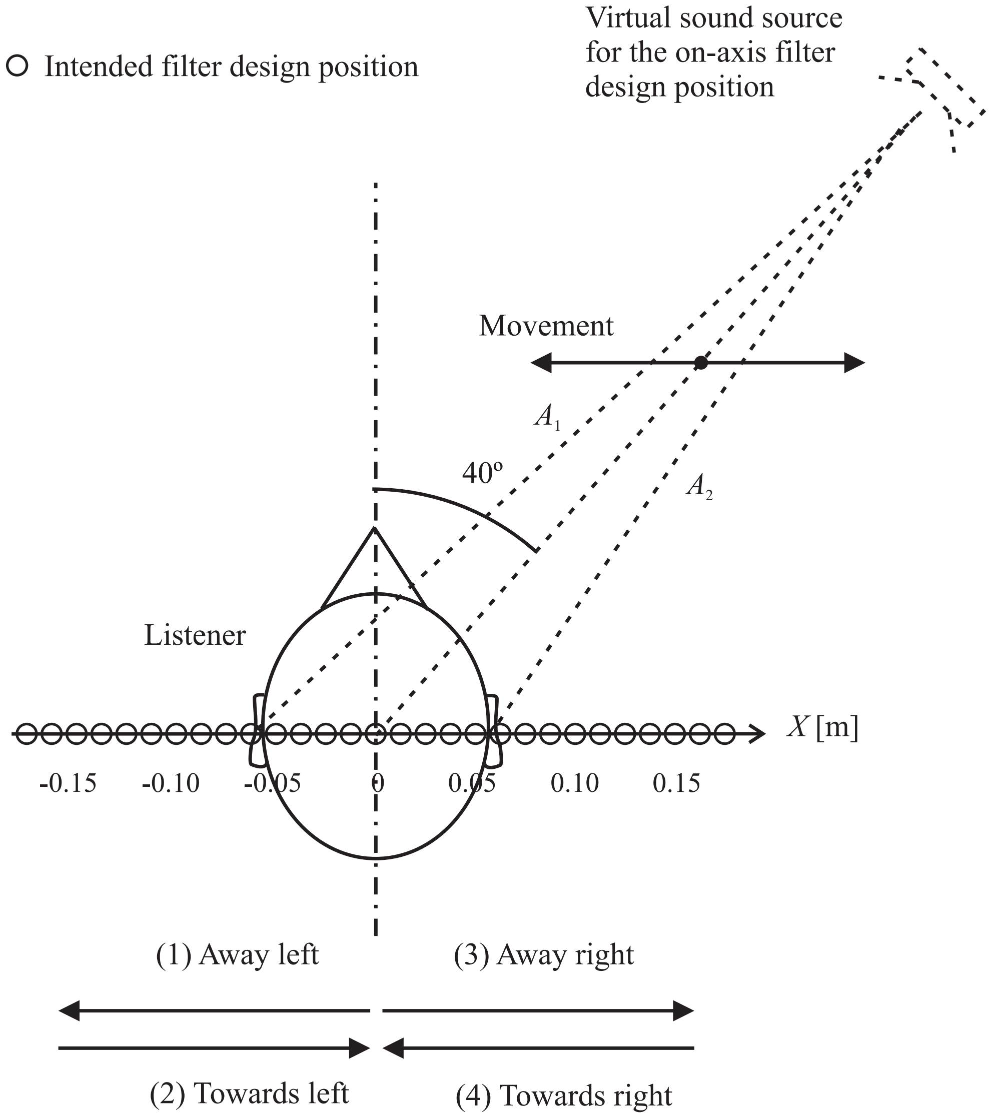


Figure 6.2: **Intended filter design positions and directions of filter change for the lateral plane**. *The signals to be reproduced are obtained from filtering the source signal by the binaural filters indicated as $A_1(k)$ and $A_2(k)$. The intended filter design position was moved in the lateral plane for four directions as indicated. This configuration was used for determining how far off the intended filter design position the listener is allowed to be while still perceiving a stable virtual sound image. The loudspeakers and the listener were in fixed positions while only the filters were updated.*

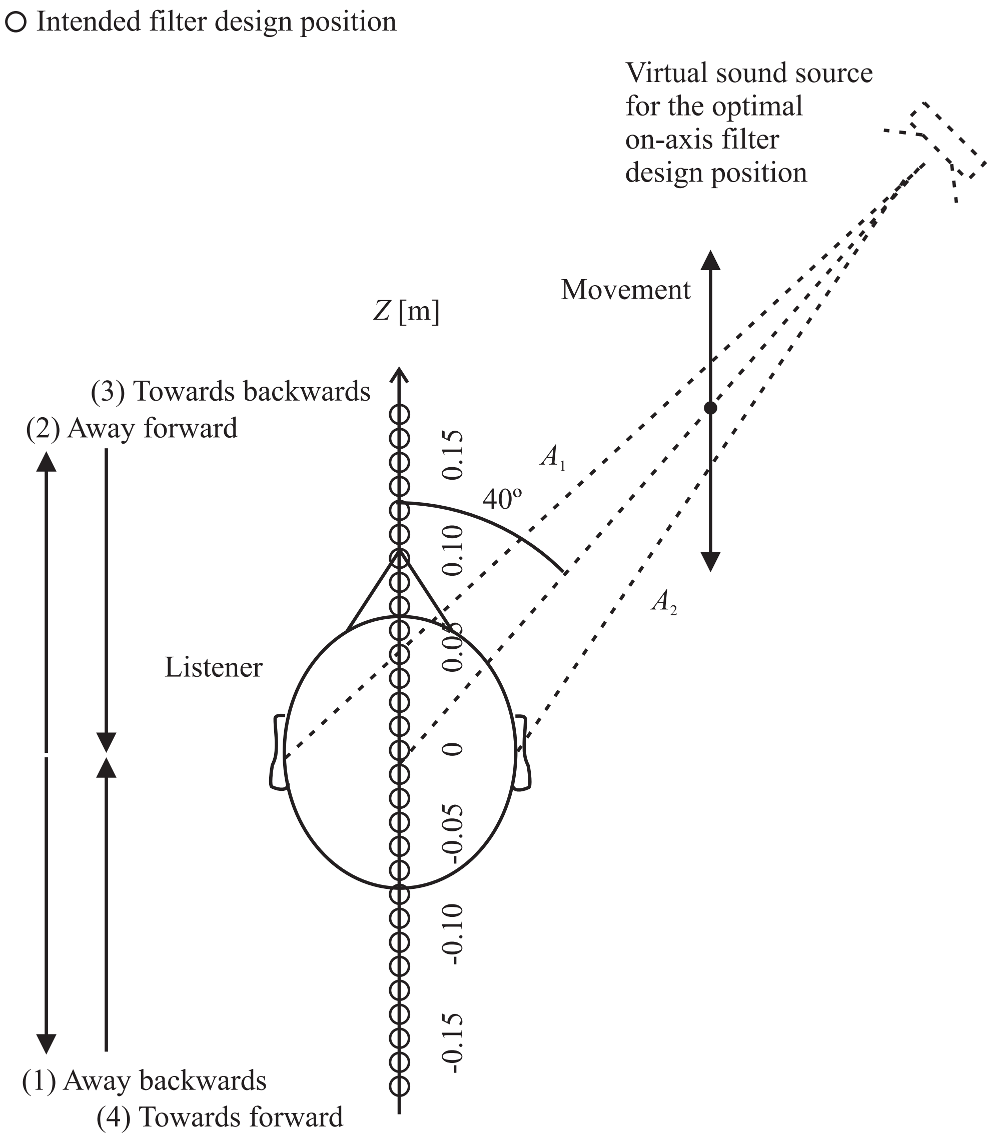
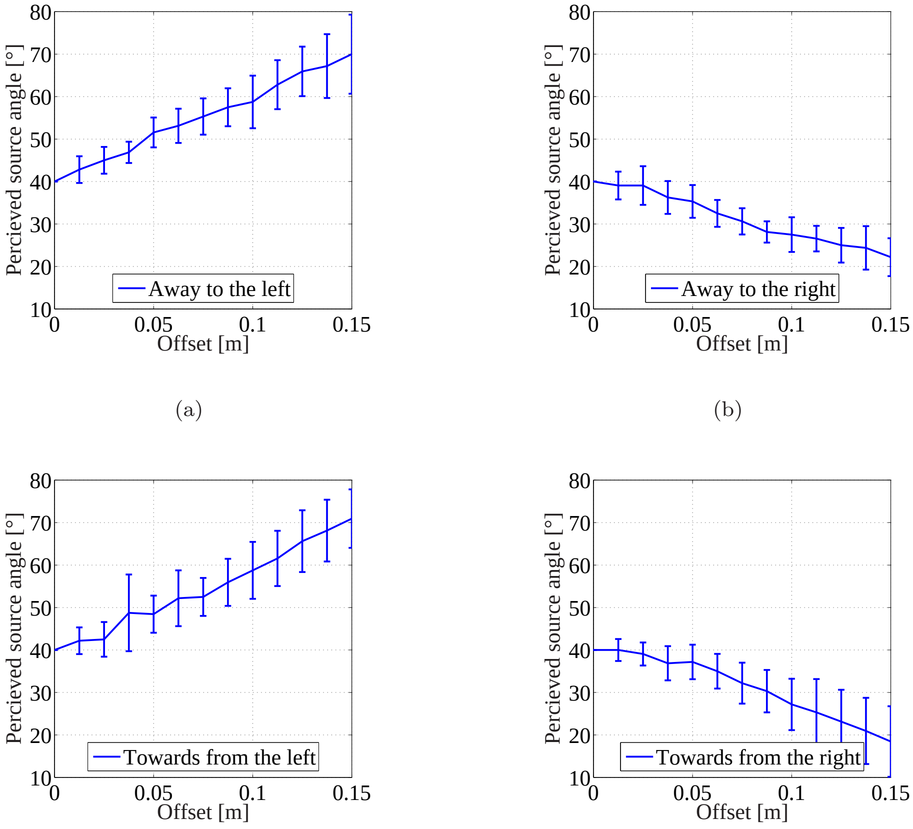


Figure 6.3: **Intended filter design positions and directions of filter change for the fore and aft plane**. *The signals to be reproduced are obtained from filtering the source signal by the binaural filters indicated as $A_1(k)$ and $A_2(k)$. The intended filter design position was moved in the fore and aft plane for four directions as indicated. This configuration was used for determining how far off the intended filter design position the listener is allowed to be while still perceiving a stable virtual sound image. The loudspeakers and the listener were in fixed positions while only the filters were updated.*



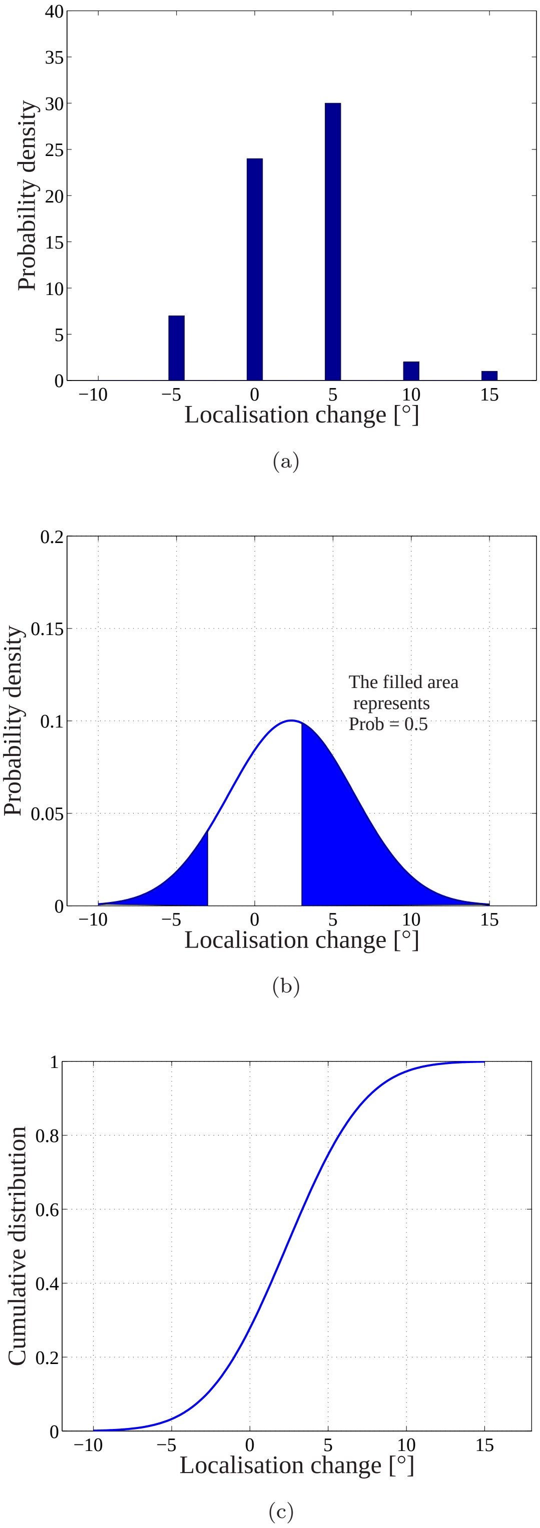
(a)

(b)

(c)

(d)

Figure 6.4: **Perceived virtual source angles in the anechoic chamber**. *The perceived virtual source angle ϕ was investigated in the lateral plane X for off-axis listener positions in the anechoic chamber. (a) Intended filter design position moving away from on-axis to the left. (b) Intended filter design position moving away from on-axis to the right. (c) Intended filter design position moving towards on-axis from the left. (d) Intended filter design position moving towards on-axis from the right.*

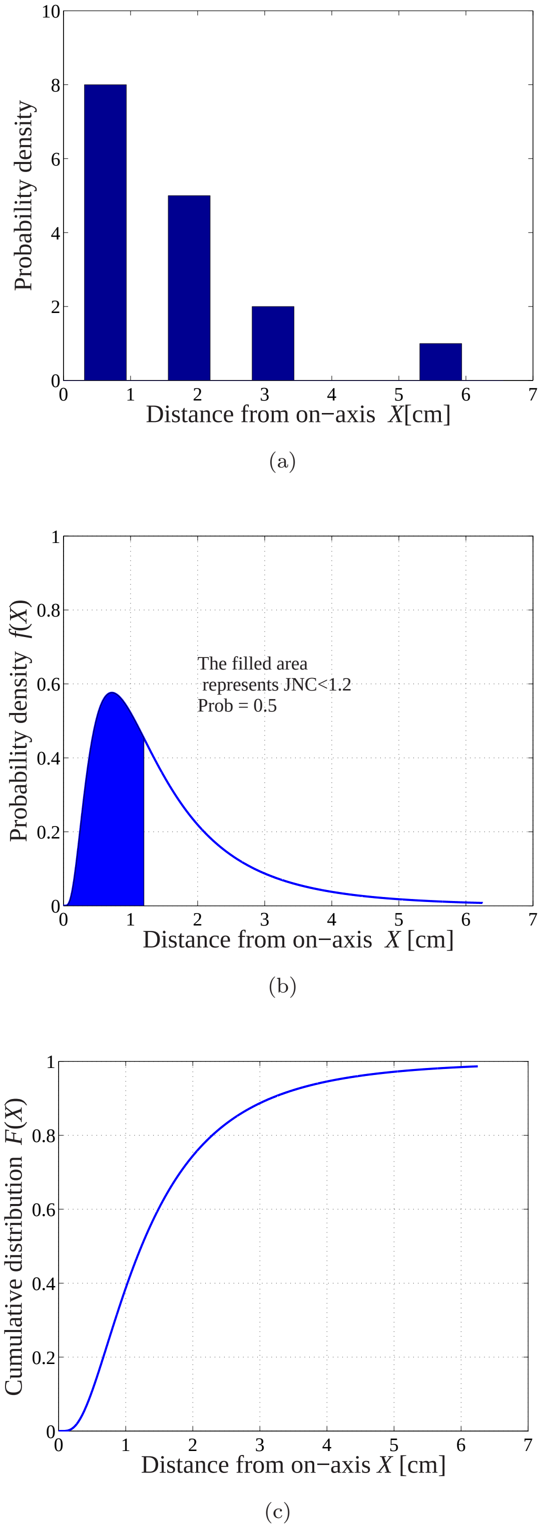


(a)

(b)

(c)

FIGURE 6.5: **Virtual sound source localisation in the lateral plane at an off-axis listener position of 2.5 cm in the anechoic chamber**. *The JND criterion determines the threshold for detecting changes in virtual sound source angles. (a) Histogram of the data. (b) The normal distribution of the data. The JND criterion is* < 2.5 *cm for* $100 \times (1 - 0.5) = 50\%$ *probability of not detecting a filter change assuming a localisation blur of* $\pm 3.2°$. *(c) The normal cumulative distribution function that was used to determine the confidence interval in (b).*

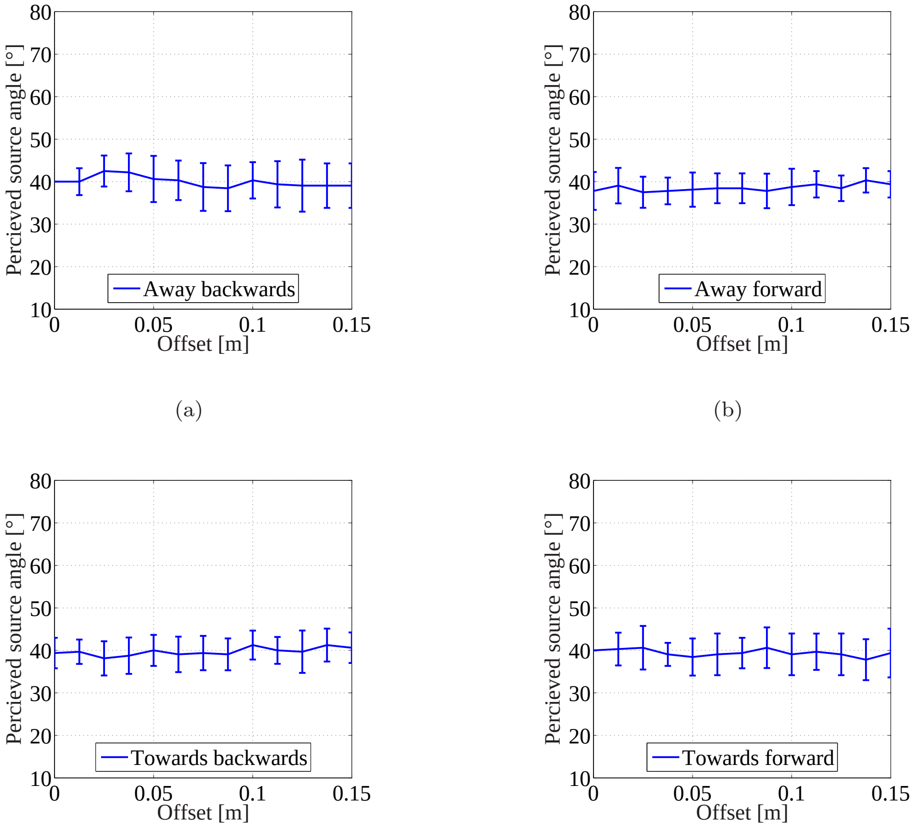


(a)

(b)

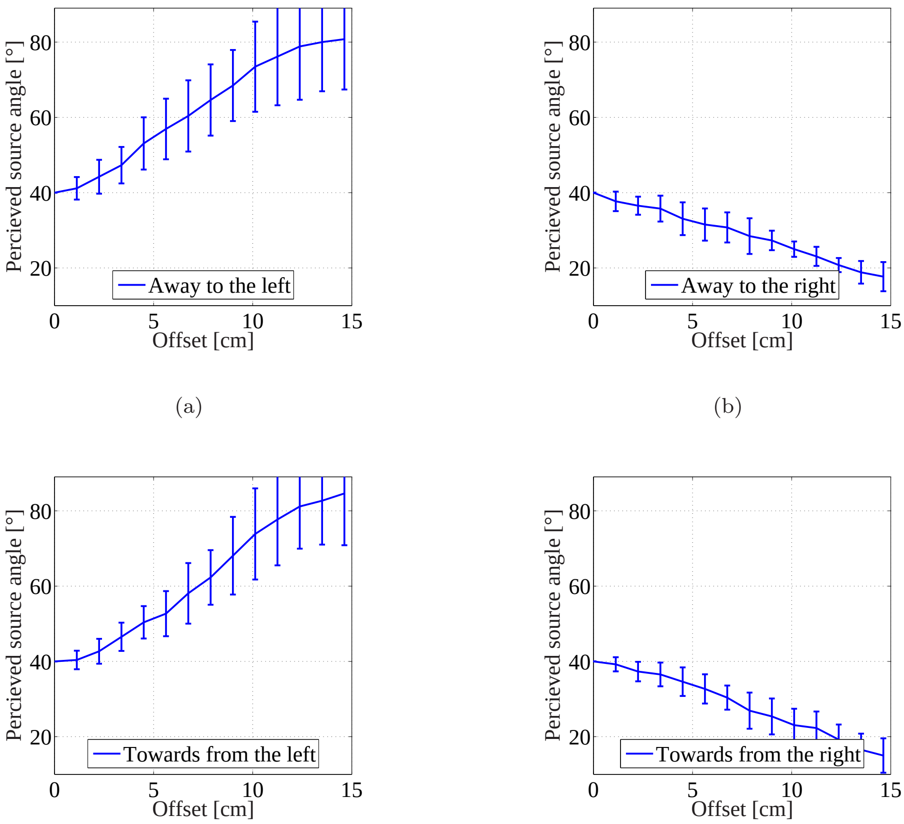
(c)

FIGURE 6.6: **JNC detection threshold in the anechoic chamber for lateral movements**. *The JNC criterion determines how many of the subjects that could not detect the filter change, for example at 0 cm on-axis none of the subjects could detect any filter change and at 6.25 cm off-axis all the subjects could detect filter change. (a) Histogram of the data. (b) The lognormal distribution fitted by MLE parameters computed from the data. The JNC criterion is* < 1.2 *cm for* $100 \times (1 - 0.5) = 50\%$ *probability of not detecting a filter change. (c) The lognormal cumulative distribution function that was used to determine the confidence interval in (b).*



(a) (b) (c) (d)

FIGURE 6.7: **Perceived virtual source angles for the fore and aft plane in the anechoic chamber**. *The perceived virtual source angle ϕ was investigated in the fore and aft plane Z for off-axis listener positions in the anechoic chamber. (a) Intended filter design position moving away from the starting point (2 m away from the loudspeakers on-axis) backwards. (b) Intended filter design position moving away from the starting point forward. (c) Intended filter design position moving towards the starting point backwards. (d) Intended filter design position moving towards the starting point forward.*



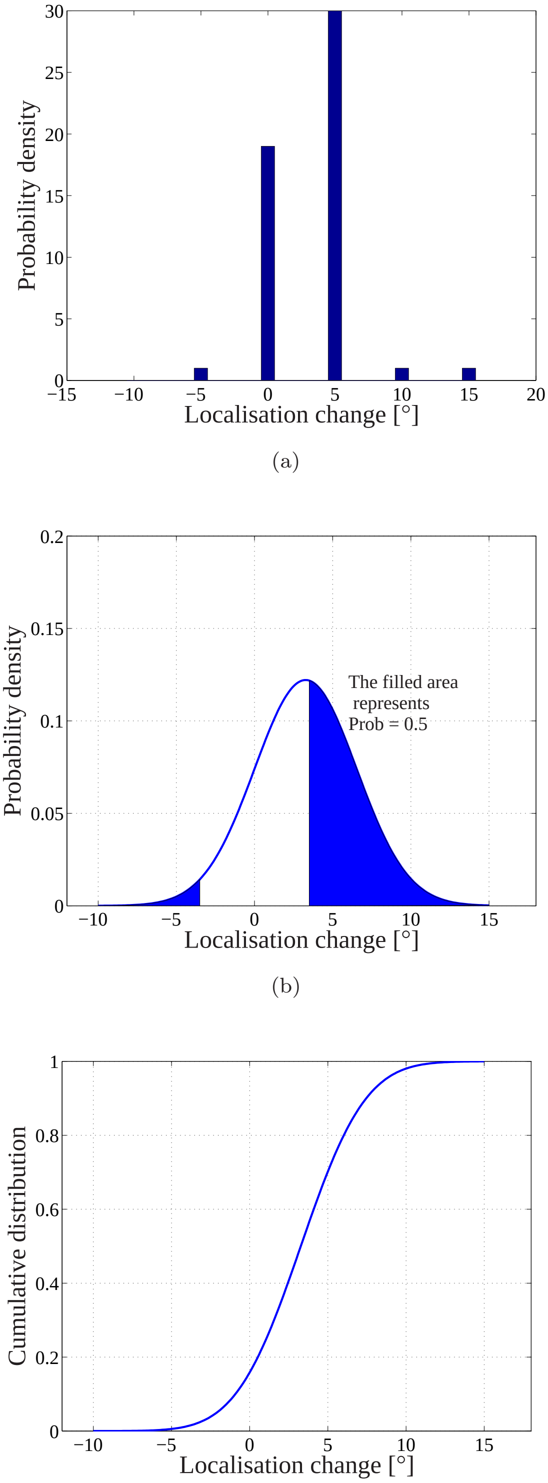
(a)

(b)

(c)

(d)

Figure 6.8: **Perceived virtual source angles in the listening room**. *The perceived virtual source angle ϕ was investigated in the lateral plane X for off-axis listener positions in the listening room. (a) Intended filter design position moving away from on-axis to the left. (b) Intended filter design position moving away from on-axis to the right. (c) Intended filter design position moving towards on-axis from the left. (d) Intended filter design position moving towards on-axis from the right.*

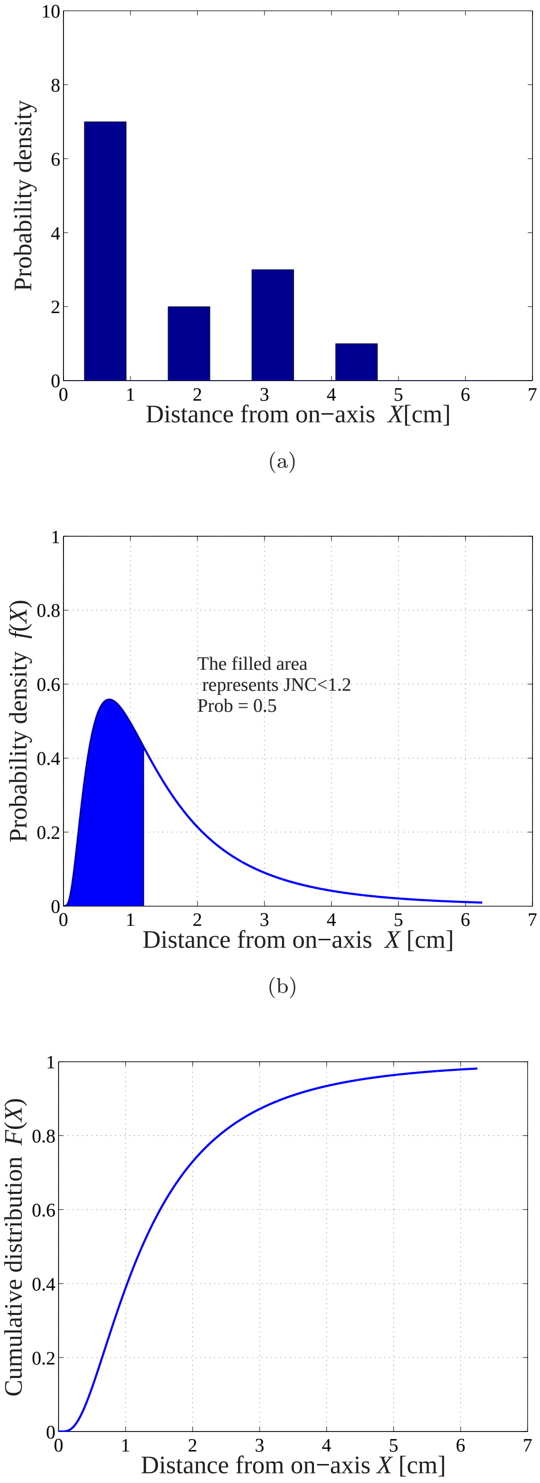


(a)

(b)

(c)

FIGURE 6.9: **Virtual sound source localisation in the lateral plane at an off-axis listener position of 2.5 cm**. *The JND criterion determines the threshold for detecting changes in virtual sound source angles. The experiment was carried out in the listening room. (a) Histogram of the data. (b) The normal distribution of the data. The JND criterion is* < 2.5 *cm for* $100 \times (1 - 0.5) = 50\%$ *probability of not detecting a filter change assuming a localisation blur of* $\pm 3.4°$. *(c) The normal cumulative distribution function that was used to determine the confidence interval in (b).*



(a)

(b)

(c)

FIGURE 6.10: **JNC detection threshold in the listening room for lateral movements X.** *The JNC criterion determines how many of the subjects that could not detect the filter change, for example at 0 cm off-axis none of the subjects could detect any filter change and at 6.25 cm off-axis all the subjects could detect filter change. (a) Histogram of the data. (b) The lognormal distribution fitted by MLE parameters computed from the data. The JNC criterion is < 1.2 cm for $100 \times (1 - 0.5) = 50\%$ probability of not detecting a filter change. (c) The lognormal cumulative distribution function that was used to determine the confidence interval in (b).*

Chapter 7

Image processing algorithms for listener head tracking

There is a demand for a head-tracking algorithm, because of the relatively small "sweet-spot" size of loudspeaker based binaural sound reproduction systems. Adding access to a video camera for the audio system gives the possibility to track head movements and update the inverse filters accordingly. There are several alternative methods for head tracking, such as, magnetic, infrared and laser tracking, though they typically require sensors to be worn by the user or are expensive. Visual tracking has emerged as one of the principal areas of research within the computer vision community. The increasing interest in visual tracking is due in part to the falling cost of computing power, video cameras and memory. A sequence of images grabbed at or near video rate typically does not change radically from frame to frame, and this redundancy of information over multiple images can be extremely helpful for analysing the input to track individual objects. The ideal algorithm for this application should be robust to background noise, track the position of the head (and ears) and be computationally efficient.

Preliminary work on this topic has been carried out at ISVR and has been based on very simple template matching algorithms that capture an image of the listener's head and then ensure a best estimate of the listener head position by finding the template position that best matches the current image. It is important that the image processing algorithms used are efficient in order to make the processing required as small as possible. A literature survey has been carried out in order to provide an initial evaluation of the candidate algorithms currently available.

There are many approaches possible for head tracking. Jones [42] concludes that it is possible to build powerful generic models for low-level image attributes like colour using simple histogram learning techniques. They describe the construction of colour models for skin and non-skin classes from a huge dataset of nearly 1 billion labelled pixels. These classes are found to exhibit a surprising degree of separability which can be exploited by building a skin pixel detector achieving a detection rate of 80% with 8.5% false positives. Bowden [15] demonstrates how to use red green blue (RGB) colour spaces to probabilistically label and segment regions of skin from image sequences for the location and tracking of the human face. He has also demonstrated that human skin clusters in a small region of colour space, since human skin colours differ more in intensity than actual colour, and under certain lighting conditions, a skin colour distribution can be characterised by a multivariate Gaussian distribution in a normalised colour space. The use of colour labelling can provide a rough estimate of the location of a head within the image frame to initialise a listener head tracking system and to track the translation of the listener in the following frames.

The distance cue for the listener can be found by using stereo area correlation tracking. Stereo analysis is the process of measuring the distance to an object based on a comparison of the object projection of two images. The stereo analysis technique developed by Konolige [53] uses area correlation to compare small patches among images. Stereo area correlation is a reliable technique for finding distance information in images. Stereo area correlation can be used to track people in natural environments by applying background subtraction and segmentation of the image with distance information. The major disadvantage with this tracking scheme is the background subtraction step, which is not desirable to carry out. Since, the background is changing frequently in most listening environments it would become a problem to calibrate the system.

Contours in images can be used for detecting the position and shape of objects. The original “snake” model was introduced by Kass [46] and this concept has been further developed by authors such as Gunn [34], [35]. The technique presented by Gunn can be used find the position and shape of the head. However, it is relatively sensitive to the quality of edge detection and the factors affecting this, namely illumination and boundary contrast. Blake [12] and Isard [41] established a stochastic framework for tracking curves in visual clutter using a Bayesian random sampling algorithm. The task was to track outlines and features of foreground objects, modelled as curves in a cluttered background. The curve models are named deformable contours, also known as “active contours” or “snakes”. A particle filtering technique named the “condensation algorithm” was used for tracking objects described by contours in highly cluttered environments. The main disadvantage of using the condensation algorithm is its computational inefficiency.

7.1 Overview of object localisation techniques

Much has been written about object localisation, and this section attempts to summarise the approaches used by previous researchers in computer vision. Table 7.1 gives an overview of approaches for detecting the position and shape of rigid objects. The table can used as a rough guide in choosing an appropriate approach for solving an object localisation problem. The algorithms for tracking objects in images can be divided into three types namely "global matching", "pose space search" and "correspondence space search" (Grimson [33]). The global matching approach to object localisation involves finding a transformation from a model to an image without determining the correspondence linking individual features of the model and the data. Here global refers to features or parameters that depend on the whole object, for example, area or volume. The usual approach is to compute a group of global parameters of a model and of the sensory data represented as a vector of parameters. An example of global parameter matching is to produce a binary image where each pixel with a value of 1 represents the object and the background pixels are set to 0. This can be achieved, for example, by thresholding an intensity image **F**.

$$\mathbf{F}(x,y) = \begin{cases} 1 & \text{object} \quad \text{pixel} \\ 0 & \text{otherwise} \end{cases} \tag{7.1}$$

$$\tag{7.2}$$

where the index x denotes the horizontal axis in the image and index y denotes the vertical axis of the image. Then the moments of each model is computed and compared to the moments of the binary image. For example, the zeroth order moment is given by

$$A = \sum_x \sum_y \mathbf{F}(x,y) \tag{7.3}$$

that computes the area of the object. Higher order moments are defined as

$$M_{p,q} = \sum_x \sum_y x^p y^q \mathbf{F}(x,y) \tag{7.4}$$

where the first moments gives the centre of mass (position) of the object. The skin colour tracking approach in Section 7.2 is another example of global matching.

The searching of correspondence space involves focusing on the matching process. Consider an object model represented by a group of features, each with an individual label that is associated with a set of attributes. Likewise, the sensory data results in a group of features with individual labels that are associated with a set of attributes. Then the search problem can be solved by considering the space of all possible matches of data features to model features. Given that there are d sensory data features and m model features, then we can imagine a d-dimensional space, along each axis of which there are $m + 1$ (+1 is due to the event of false detection) discrete possible values. Each point in correspondence space represents an interpretation of all of the data points. There are $(m+1)^d$ nodes in this space, which is highly computational demanding in most applications, therefore various techniques for imposing constraints on the correspondence space have been investigated by previous researchers (Grimson [33]).

The pose space search involves the evaluation of possible transformations of the model into the data. Unlike the correspondence space search that finds the association of data features and model features, while the pose of the object is of secondary importance. The pose space search focuses on finding poses of an object model that is imposed on to the sensory data. The pose space search can be represented by, for example a three-dimensional space having two axes for translational components and one for rotation. Then each point in the configuration of pose space corresponds to a different pose of the model and the idea is to determine which poses that best fit the data. This approach is further described in Section 7.4.

Many techniques for object localisation have concentrated on searching a correspondence space. The number of detected features can become large, which means that they have to be limited in some way. For example, probabilistic reasoning can be used to limit the number of features. However, the disadvantage of this approach is that the target need have easily detectable features, such as corners or colour, and detailed models of the target need to be known. In some cases, such as here where the target is a face, there are relatively easily detectable features. The features that are feasible to detect in face include skin colour, eyes and nose. The main advantage of this approach is the computational efficiency that can be achieved when the target features are easily detectable.

Correlation based techniques are fundamentally based on correlating a known template with various regions in an image and can be run effectively in hardware. However, in general the pose space needs to be of a very small dimension. Contour tracking using a particle filter is also a pose space search, which also requires the dimension of this space to be small. Experiments show that if the prior is diffuse over more than three or four dimensions, then the algorithm tends to be to slow for real time tracking (MacCormick [62]). The advantages are that it can be applied for a wide range of objects with minimal modelling requirements and it is very insensitive to cluttered environments.

Other techniques for feature detection include the Hough transform that is a well-known technique for localising corners and lines in images. The Hough transform can for example be used to detect the corners of the eyes in a face. Optical flow is another method for target localisation. However, it requires the target to move and the camera and the background to be fixed. The background subtraction step is surprisingly sensitive to camera movements and lighting changes.

A series of computer simulation studies have been undertaken in order to evaluate promising candidate algorithms, especially with regard to computational efficiency and

	Techniques	Advantages	Disadvantages	Examples
Global matching	Image segmentation (blobs etc.)	* Faster than pose space and correspondence space	* Objects need to be separated from background * Light sensitive	Grimson [33]; this chapter
Correspondence space search	Feature detection	* Faster than pose space search (of high dimension)	* Objects need to have clearly detectable features	Grimson [33]
Pose space search	Correlation	* Technology ready available	* Light sensitive * Can not handle irregular shapes	Ballard [8]; Cootes [22]
	Contour tracking using particle filters	* Simplified modelling of targets * Clutter resistance	* Not suitable for complex 3D structures * Occasional false detection	Blake [12]; this chapter
Other	Hough transform	* Effective when applicable	* Limited for other shapes than corners and lines	Duda [26]
	Optical flow	* Simple geometric modelling	* Requires moving target and fixed background	Subbarao [94]

Table 7.1: **Approaches for detecting the position and shape of objects**.

clutter resistance. The algorithms that have been evaluated for listener head tracking are colour tracking (of skin), stereo area correlation and contour tracking. The colour tracking algorithm evaluated here is based on statistical colour distribution in Red Green Blue (RGB) space and is presented in Section 7.2. The colour tracking algorithm finds the skin region in an image that represents the face of the listener and tracks listener translation. The main advantage of this relatively simple algorithm is that it can be run efficiently on available hardware. The presented stereo area correlation algorithm requires two cameras and is described in Section 7.3. The stereo algorithm calculates the distance of the objects in the camera field and hence the distance to the listener. The stereo area correlation technique is a well known algorithm that can find distance information in image sequences in real time. The contour tracking algorithm makes it possible to track the listener in an affine "shape space" of six DOF and is presented in Section 7.4. The contour tracking approach was investigated to improve the robustness of tracking through serious clutter and changes in illumination. The visual tracking system is implemented by using a stereo camera package (Small Vision System (SVS)).

7.2 Colour tracking

The colour content of an image is an important attribute. This section will discuss how colour can be used to provide a reliable feature for locating and tracking moving objects. The availability of colour generally provides a more distinguishable difference between foreground and background objects within an image. Cambridge research laboratory in Massachusetts (Jones [42]) has used large image data sets of photos on the World Wide Web to build powerful generic models from colour using histogram learning techniques and multivariate Gaussian classifiers. They have demonstrated that human skin clusters in a relatively small region of colour space. A skin colour distribution can be characterised by a multivariate Gaussian distribution in a colour space as previously investigated by Bowden [15], Jones [42], Azarbayejani [6], [7] and Kuchi [55]. This colour labelling can be used to provide an estimate of the location of a head within the image frame to initialise a head tracking system and also to use the colour feature for tracking.

By performing image processing upon a greyscale representation, calculated from the colour channels (typically the average intensity of the three colour channels) a considerable amount of information about object boundaries is lost. An RGB image that consists of three colour regions where each region has the same intensity in its colour channel for example: the red area has $R = 255$, $G = 0$, and $B = 0$; the green area has $R = 0$, $G = 255$ and $B = 0$; etc. By taking the average of the three colour channels at each pixel, the resulting image would have a constant intensity of 85 and no distinction would be possible between the various areas. However, in the colour image, it is visually apparent that such a distinction does exist and very clear boundaries are defined. It is clear that reducing colour information to one channel "throws information away", which may be invaluable to the application at hand. If an object of interest is sufficiently prominent within one of the colour channels, then the intensity of that channel can be used instead of the mean intensity.

7.2.1 Multivariate Gaussian probability distribution

The colour representation of a face in a video sequence is influenced by many factors such as ambient light and the object moving relative to a light source. Also human skin colours deviate in RGB space from person to person, although they still cluster in a relatively small space (Bowden [15]). When the skin colour of the person is known a priori then colour can be used as a robust feature for tracking skin colour in a video sequence with consistent ambient light. The skin colour of a subject head can be extracted by selecting a "training set" of skin pixels in the image. The training set of skin pixels was here chosen on a frame from the image sequence used in the tracking experiment in Section 7.2.2.

The set of pixels that constitute the foreground with skin pixels is denoted $\mathbf{X} \in \mathcal{F}$

where $\mathcal{F}$ represents foreground. The matrix $\mathbf{X}$ contains three image components (RGB) that are written as column vectors $[\mathbf{x}_R, \mathbf{x}_G, \mathbf{x}_B]$ of length N_X. Similarly, the matrix with background pixels is denoted $\mathbf{X} \in \mathcal{B}$, where $\mathcal{B}$ represents background. It is noted that the image of interest in its original form is represented by three $N \times M$ matrices that corresponds to the three colour channels (RGB). Then the image components are transformed into vectors of pixel indices in order to create the matrix $\mathbf{X}$.

$$\mathbf{X} = [\mathbf{x}_R, \mathbf{x}_G, \mathbf{x}_B] = \begin{bmatrix} \mathbf{x}_R(1), & \mathbf{x}_G(1), & \mathbf{x}_B(1) \\ \vdots & \vdots & \vdots \\ \mathbf{x}_R(N_X), & \mathbf{x}_G(N_X), & \mathbf{x}_B(N_X) \end{bmatrix} \tag{7.5}$$

The vector of mean values from the skin pixels is labeled $\mathbf{u} \in \mathcal{F}$ and contains the mean values of $\mathbf{X}$. Likewise, the vector of mean values from the background pixels is labeled $\mathbf{u} \in \mathcal{B}$.

$$\mathbf{u}^T = [u_R, u_G, u_B]^T = [E[\mathbf{x}_R], E[\mathbf{x}_G], E[\mathbf{x}_B]]^T \tag{7.6}$$

The mean values are put into a matrix $\mathbf{U}$ of the same size as the matrix $\mathbf{X}$ by

$$\mathbf{U} = (\mathbf{u}\mathbf{z}^T)^T \tag{7.7}$$

where $\mathbf{z}$ is a unit vector of length N_X. The covariance matrix of the mean values is now given by

$$\mathbf{\Sigma} = E\left[(\mathbf{X} - \mathbf{U})^T (\mathbf{X} - \mathbf{U})\right] \tag{7.8}$$

The pixel indices of the "training set" are used to estimate the parameters of a multivariate Gaussian probability distribution. The likelihood of the pixel data $\mathbf{x}$ given the skin parameters, can be estimated by the following Equation

$$p(\mathbf{x}|\mathcal{F}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} e^{-(\mathbf{x}-\mathbf{u})^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mathbf{u})/2} \tag{7.9}$$

where $\mathbf{u} \in \mathcal{F}$ is the 1×3 skin mean values, $\mathbf{\Sigma} \in \mathcal{F}$ is the 3×3 skin covariance matrix, $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$ and $d = 3$, which is the dimension (three dimensions: R, G and B). The data $\mathbf{x}$ is the RGB value of pixel i and is given by $\mathbf{x} = [\mathbf{x}_R(i), \mathbf{x}_G(i), \mathbf{x}_B(i)]$ The output of Equation 7.9 is the conditional probability value for a pixel of RGB data.

By using Bayes' theorem we can find the probability $p(\mathcal{F}|\mathbf{x})$ of skin given the data.

$$p(\mathcal{F}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{F})p(\mathcal{F})}{p(\mathbf{x})} \tag{7.10}$$

where $p(\mathcal{F})$ is prior probability and $p(\mathbf{x})$ is the evidence that is given by

$$p(\mathbf{x}) = p(\mathbf{x}|\mathcal{F})p(\mathcal{F}) + p(\mathbf{x}|\mathcal{B})p(\mathcal{B}) \tag{7.11}$$

The prior $p(\mathcal{F}) = 0.09$ is found from the training database (Jones [42]) used by Sigal [90]. The training database holds 80,306,243 foreground (skin) pixels and 861,142,189 background pixels. The prior probability for background is $p(\mathcal{B}) = 1 - p(\mathcal{F}) = 1 - 0.09 = 0.91$.

The probability of the data given the foreground for each pixel of the image is named $\mathbf{L}_{\mathcal{F}}$ and is computed using Equation 7.9. The same procedure can be carried out for a "background matrix" $\mathbf{L}_{\mathcal{B}}$ in order to get the probability of the data given the background. The un-normalized $\mathbf{L}_{\mathcal{F}}$ and $\mathbf{L}_{\mathcal{B}}$ values can be interpreted as the "conditional probability density" or alternatively the "likelihood". The final output of the algorithm should be normalised "posterior probabilities" which can be found by normalising the values of each matrix such that they add up to unity for each pixel location. The probability of foreground given the data is found by

$$\mathbf{L}_{\mathcal{F}}^{'} = \mathbf{L}_{\mathcal{F}}p(\mathcal{F}) \div (\mathbf{L}_{\mathcal{F}}p(\mathcal{F}) + \mathbf{L}_{\mathcal{B}}p(\mathcal{B})) \tag{7.12}$$

where $\div$ denotes matrix element-by-element division. The probability of background given the data is found by

$$\mathbf{L}_{\mathcal{B}}^{'} = \mathbf{L}_{\mathcal{B}}p(\mathcal{B}) \div (\mathbf{L}_{\mathcal{F}}p(\mathcal{F}) + \mathbf{L}_{\mathcal{B}}p(\mathcal{B})) \tag{7.13}$$

The sum of the posterior probability matrices sums up to one, i.e. all the elements of the matrix are unity.

$$\mathbf{L}_{\mathcal{F}}^{'} + \mathbf{L}_{\mathcal{B}}^{'} = \mathbf{1} \tag{7.14}$$

The probability of skin in a given image can now be calculated by using equations 7.10, 7.12 and the result is illustrated in Figure 7.1.

7.2.2 Colour segmentation

Here, multivariate Gaussian colour segmentation has been performed on an image with an un-cluttered background. Figure 7.1 illustrates the skin matrix, the background matrix obtained with a multivariate Gaussian classifier and the original colour image. The skin parts of the image are clearly segmented from the background. The mean and covariance values for skin and background have here been extracted from the original RGB image in Figure 7.1 (a). Further image segmentation can be used to enhance the skin colour tracker, which is illustrated in Figure 7.2. Firstly, greylevel thresholding is performed in order to create a binary image of skin pixels and background pixels. The threshold can be set manually or by adopting for example the Otsu method that chooses the threshold to minimize the intraclass variance of the thresholded black and white pixels (Otsu [81]). Secondly, binary image dilation is performed with a line-structuring element. Thirdly, the holes in the image are filled to create a solid object. Finally, morphological opening is used to remove small objects from the image while preserving the shape and size of larger objects in the image. Appendix B (MathWorks [64]) describes morphological image processing operations in more detail. If there are several objects detected in the image, the size of the "blobs" can be estimated and the head can be assumed to be the largest blob.

The image in Figure 7.3 illustrates a subject with his hand present in a highly cluttered background, and the same segmentation as in Figure 7.2 has been applied. Two "blobs" are present in the image, the large blob is the head of the subject and the small blob is the hand of the subject. Some of the background objects appear to be skin, though the image segmentation process manages to filter out the undesired features and finds the head. The mean and covariance values for skin and background have been extracted from the original RGB image in Figure 7.3 (a).

7.2.3 Summary of the algorithm

To track coloured objects in video frame sequences, the colour image data can be represented as a probability distribution. The centre of the colour object in the initial frame is found by applying colour segmentation on the entire image. The process for continuous tracking needs to be simplified in order to save processing time. One alternative is to limit the search for skin using a region of interest window (ROI). Only pixels that are within the ROI need to be converted and classified, which significantly speeds up the procedure. Also background clutter outside the ROI cannot be misclassified. This procedure produces much cleaner and more efficient segmentation without the need for morphological image segmentation as previously described.

In order to limit processing to the ROI, a procedure for moving the window must be devised. If the assumption is made that the binary segmented object has a central white mass surrounded by black background, then the centre of the "blob" should be at the centre of the window. A simple translation can then be calculated to position the centre of the window at the centre of the segmented feature for the next iteration of the algorithm. The size of the window can be configured to be 10% larger than the "blob" that represents the head.

Figure 7.4 illustrates the probability of skin in the first frame of an image sequence. The ROI has been determined by performing morphological image segmentation and it fits closely to the head. A threshold is configured for the skin pixel probabilities in the ROI in order to get a binary image with skin and not skin pixels. The binary image is calculated with a simple algorithm that searches every pixel in the ROI and sets the pixel to zero if the pixel value is less then the threshold and sets the pixel to one if the pixel value is greater then the threshold. The threshold level can be found by using Receiver Operating Characteristics (ROC) (Sigal [90]) or by experiment.

The following provides a summary of the algorithm:

(1) Initialisation

(1.1) Construct a multivariate probability function for skin and background with Equation 7.9. Calculate the probability of the data $\mathbf{x}$ given skin for each pixel in the input image. Compute the probability of skin given $\mathbf{x}$ by using Bayes' theorem Equation 7.12.

(1.2) Perform morphological image segmentation as illustrated in Figure 7.2. The output image obtained is a binary image of skin pixels and non-skin pixels.

(1.3) Search the output image for the largest "blob".

(1.4) Calculate the centre of the largest "blob" and initialise the window to this position.

(2) Tracking

(2.1) Calculate the size of the largest "blob" and make the window approximately 10% larger.

(2.2) Capture the new input image.

(2.3) Segment the image in the window using the probability density function as in point (1.1). The priors in the ROI are estimated to be approximately equal $p(F) = p(B) = 0.5$. Then threshold the probabilities map in order to obtain the output image, which is a binary image of skin pixels and not skin pixels (see Figure 7.4).

(2.4) Calculate the mean white pixel in order to find the centre.

(2.5) Move the window to the centre.

(2.6) Enlarge the window by a suitable number of pixels; here 5 pixels have been proven to work well. Then repeat steps (2.3), (2.4) and (2.5) again and then fit a new window with the original size to the centre of the skin pixels. Perform this a number of times (3 times has proved to work well without adding to much processing time) in order to make sure that the window is fitted to the centre of the skin pixels.

(2.7) Return to (2.2).

7.2.4 Results

The performance of the colour tracking algorithm is compared to a magnetic tracking system from Polhemus. The Polhemus tracking system is here used as a reference system. The magnetic tracking system can track with six DOF with an update rate of 120 Hz. The tracking precision of the Polhemus is 0.8 mm within a distance of 0.8 m between the transmitter and the receiver. The colour tracking algorithm was executed on a Pentium 4 computer and the stereo camera was used to capture the images. The listener was moving between approximately $-0.5 \leq X \leq 0.5$ m, and from $-0.5 \leq Z \leq 0.5$ m and from about $-0.1 \leq Y \leq 0.1$ m.

The result from a tracking sequence of a listener moving mainly in the X-direction with a relatively uncluttered background is illustrated in Figure 7.6 and Figure 7.7. The standard deviation in the horizontal plane X is about 0.02 m and in the vertical plane Y about 0.05 m. The deviation for tracking in the Y-direction is here greater than in the X-direction. The large deviation in Y-direction is mainly due to pitch movements of the listener that are not captured with the colour tracking algorithm. Without the pitch movements the tracking performance in the Y-direction should be close to the tracking performance in the X-direction. The pitch movements can be seen since the Polhemus receiver was mounted on the back head of the listener. The deviations that have been seen can be acceptable in the application of adaptive virtual sound imaging systems.

7.3 Stereo area correlation

This section presents a stereo vision algorithm that can be used to find the distance to the listener. Stereo vision algorithms compute distance information to objects by using triangulation (Ayache [5]). Two cameras are used to capture two images from different viewpoints and an estimation of the relative position of a feature as it appears in the two images makes it possible to calculate the distance of the feature (Catleman [17]). The stereo vision system that is used here, has a built in stereo area correlation function. This built in function returns a disparity image, which is used to estimate the distance information of the listener. The full details of the stereo area correlation process and the stereo vision system named SVS are presented by Konolige [53], [54].

7.3.1 Stereo analysis

Stereo analysis is the process of measuring distance to an object based on a comparison of the object projection on two or more images. The problem to be solved is to find corresponding elements between the images and once a match is made, then the distance to the object can be computed using the image geometry as described in Appendix B. The common method for stereo analysis is to use area correlation, which compares small patches or windows among images using correlation. The window size will always be a compromise, since small windows are more likely to be similar in images with different viewpoints, and large windows increase the signal to noise ratio.

In Figure 7.8 area correlation and the pixel-shift calculation is illustrated. The stationary window is chosen to be in the left image and hence the moving window in the right window. For each pixel, the area correlation is calculated by moving the window in the right image. The area correlation is calculated by the Laplacian of Gaussian (LOG) transform and L1 norm (absolute difference). The LOG transform measures directed edge intensities over the area smoothed by the Gaussian. The LOG transform and L1 norm have proven to give good results (Konolige [54]).

7.3.2 Summary of the algorithm

A brief summary of the stereo area correlation algorithm is presented here. The stereo area correlation algorithm can be divided into five steps as follows

(1) Geometry correction: distortions in the input images are corrected.

(2) Image transform: a local operator transforms each pixel in the greyscale image into a more appropriate form, e.g., normalizes it based on average local intensity.

(3) Area correlation: each small area is compared with other areas in its search window. The area correlation is calculated by the Laplacian of Gaussian (LOG) transform and L1 norm (absolute difference).

(4) Extrema extraction: the extreme value of the correlation at each pixel is determined, which results in a disparity image. The disparity measure is further described in Appendix B. In the disparity image, each pixel value is the disparity between left and right image patches at the best match.

(5) Post-filtering: Interest operator and left/right check filters are applied in order to clean up noise in the disparity image.

The stereo area correlation algorithm performs the five steps above for each frame and the result is a disparity image of each frame (Konolige [54]).

The stereo area correlation algorithm is described in Figure 7.9. The disparity, which is a function of distance to the objects in the image, is calculated by using stereo area correlation. When the disparity is calculated for each pixel in the colour tracker ROI, the result is a number of different disparities because the listeners head is not flat and also the stereo area correlation process does not give 100 % reliable results. To minimize the effect of pixels that are not from the head, only the disparity pixels from a smaller search window (50 % of the colour tracker ROI) are used. A histogram is applied to the pixel values of the smaller ROI and the most common disparity value is chosen in order to find the most common distance to the tracked object. The distance is calculated from this disparity measure. The lowest disparity values are disregarded, which will result in a histogram with only disparity values from the tracked object. Figure 7.10 illustrates the original RGB image and the disparity image with the ROI indicated by a rectangle.

7.3.3 Results

The results from the stereo area correlation algorithm are presented here. The distance to listeners head is calculated from the disparity value as described in Section 7.3.2. The stereo tracking algorithm was running on a Pentium 4 computer using the SRI stereo camera. The tracking sequence is illustrated in Figure 7.6 and Figure 7.11. The results show that the standard deviation is 0.05 m for fore and aft movements Z.

The relatively large deviation in Z-direction is mainly due to limited accuracy in the disparity measure from the stereo area correlation algorithm. The standard deviation could possibly be reduced by applying a suitable smoothing filter. A moving average filter was applied to smooth the distance measure in the implementation that is described in Chapter 8.

7.4 Contour tracking

In applications where the listener is allowed to move in a seriously cluttered environment and through changing illumination, then it is necessary to use a more advanced algorithm than the previously presented colour tracking algorithm. This section describes “active contours” for visual head tracking using affine transformations. The shape of objects is here represented by B-spline curves in an image sequence. The active contour framework was developed by Blake [12], Kass [46] and a number of collaborators such as Isard [41] and MaCormick [62]. It will be shown how contours can provide a reliable algorithm for locating and tracking moving objects. This section will briefly cover the theory of splines, “shape spaces”, dynamical models using auto-regressive processes and particle filters.

7.4.1 Spline curves

Visual curves can be represented in terms of parametric spline curves, which is common in computer graphics. The parametric spline curves are used because they can efficiently represent sets of boundary curves in an image. This section forms the basis for contour tracking. The curves are denoted, $\mathbf{r}(s) = (x(s), y(s))$ where s is a parameter that increases as the curve is traversed, and x and y are functions of s known as splines. The parameterised curve can be written as:

$$\mathbf{r}(s) = (\mathbf{b}(s)\mathbf{q}^x, \mathbf{b}(s)\mathbf{q}^y) \quad for \quad 0 \leq s \leq L \tag{7.15}$$

where $\mathbf{b}(s)$ is a vector $(b_0(s), ..., b_{N_B-1}(s))^T$ of B-spline basis functions, $\mathbf{q}^x$ and $\mathbf{q}^y$ are vectors of B-spline control point coordinates and L is the number of spans (concatenated polynomial segments as described by Blake [12]). For example, the elements of $\mathbf{q}^x$ define the x-coordinates of the control points. To simplify the notation, the B-spline control points can be combined in to a spline matrix.

$$\mathbf{Q} = \begin{pmatrix} \mathbf{q}^x \\ \mathbf{q}^y \end{pmatrix} \tag{7.16}$$

Thus equation 7.15 can be written in matrix form as

$$\mathbf{r}(s) = \mathbf{U}(s)\mathbf{Q} \quad for \quad 0 \leq s \leq L \tag{7.17}$$

Where the following definition holds

$$\mathbf{U}(s) = \mathbf{I}_2 \otimes \mathbf{b}(s)^T = \begin{pmatrix} \mathbf{b}(s)^T & 0 \\ 0 & \mathbf{b}(s)^T \end{pmatrix} \tag{7.18}$$

and $\mathbf{U}(s)$ a matrix of size $2 \times 2N_Q$. In this expression $\otimes$ denotes "Kronecker product" of two matrices and $\mathbf{I}_m$ denotes an $m \times m$ identity matrix. An example of a quadratic ($d = 3$), parametric spline curve $\mathbf{r}(s)$ is illustrated in Figure 7.12.

7.4.2 Shape space models

The "shape space" is introduced to reduce the dimension. Arbitrary manipulation of the spline vectors $\mathbf{Q}$ is often too general in practice, therefore a restricted class of transformations is introduced that is parameterised by a lower dimensional configuration vector $\mathbf{y}$ termed a shape space vector. Typical objects require 5 - 20 control points to produce a contour that for a human observer, appears to match the object closely (Blake [12]). The vector space of such contours has 10 - 40 dimensions, which is often undesirably large. Therefore we work in a vector subspace of $\Re^{2n}$ termed the shape space and denoted $\mathcal{S}$. An element $\mathbf{y} \in \mathcal{S}$ is related to the control point coordinates by

$$\mathbf{Q} = \mathbf{W}\mathbf{y} + \mathbf{Q}_0 \tag{7.19}$$

where $\mathbf{W}$ is a $N_Q \times N_Y$ "shape matrix". $\mathbf{Q}_0$ is called a constant offset, which is a template curve against which shape variations are measured. The shape matrix corresponds to, for example, translation, Euclidian similarities or affine transformation.

The planar affine "shape space" is adopted in this thesis, since it can represent six DOF, to a good approximation, which sufficient for the virtual sound imaging system under investigation. The figure below illustrates how a template can be translated, scaled and rotated in the planar affine shape space representation.

The "shape matrix" used here is given by

$$\mathbf{W} = \begin{pmatrix} 1 & 0 & \mathbf{q}_0^x & 0 & 0 & \mathbf{q}_0^y \\ 0 & 1 & 0 & \mathbf{q}_0^y & \mathbf{q}_0^x & 0 \end{pmatrix} \tag{7.20}$$

The first two columns of $\mathbf{W}$ represent horizontal and vertical translation. The remaining four affine motions can be expressed as simple linear combinations of the four last columns of $\mathbf{W}$.

Some examples of transformations are given below.

1. $\mathbf{y} = (0,0,0,0,0,0)^T$ represents the original template $\mathbf{Q}_0$.
2. $\mathbf{y} = (10,0,0,0,0,0)^T$ represents the original template translated 10 units to the right.
3. $\mathbf{y} = (0,0,1,1,0,0)^T$ represents the original template doubled in size.
4. $\mathbf{y} = (0,0,\cos(\theta)-1,\cos(\theta)-1,-\sin(\theta),\sin(\theta))^T$ represents the original template rotated with an angle θ.

7.4.3 Feature detection

This paragraph describes the process of feature detection and introduces the concept of a limited search region. Feature detection processes are effective up to a point but cannot retrieve entire geometric structures. "Snakes" were therefore introduced by Kass [46] to deal with these limitations of low-level processing. The idea is to take a feature map $\mathbf{F}$ (an image that is processed in order to enhance certain features) such as that in Figure 7.13 and to treat $\mathbf{F}$ as a landscape on which a deformable curve (snake) $\mathbf{r}(s)$

can slide. For efficiency, the deformable templates described in this section are driven towards a distinguished feature curve $\mathbf{r}_f(s)$ instead of over the entire image landscape $\mathbf{F}$ that is used in the original snake model.

To reduce the computational cost of applying filters across the entire image, one can define a search region, in which the corresponding image feature is likely to lie. The search region is defined from an estimate of the position of a tracked image contour. Image processing can now effectively be restricted to this search region as illustrated in Figure 7.14.

The normals to $\mathbf{r}(s)$ are constructed at points $s = s_i, i = 1,, N$ along the curve $\mathbf{r}(s)$, which will result in a sequence of sampled points $\mathbf{r}_f(s_i)$ $i = 1,, N$ along the feature curve $\mathbf{r}_f(s)$. The sample points can be spaced arbitrarily along the curve, for example it may be desirable to concentrate the sample points in regions where measurements are expected to be particularly informative. In reality several features might be detected on each normal, particularly when tracking in cluttered images. In Section 7.4.5 it is described how to choose the desired features by using the condensation algorithm.

In Figure 7.15, a contour is initialised to the centre of a head (found by using colour tracking) and the result from the feature detection process (using edge detection) is illustrated. At fixed points along the B-spline so called measurement lines, which are normal to the contour, are introduced onto the image. The length of the measurement lines has been fixed in advance and it is symmetric about the measurement point. The next step is to apply a one dimensional feature detector. The edge detector is applied to the image intensity along each measurement line. The edges are located by applying an operator mask $\mathbf{c}(n), \quad -N_c \leq n \leq N_c$, by discrete convolution, to the samples of the measurement line $\mathbf{f}(n), \quad 1 \leq n \leq N_f$, in order to find the strength of the edges

$$\mathbf{e}(n) = \sum_{m=-N_c}^{N_c} \mathbf{c}(m)\mathbf{f}(n+m) \tag{7.21}$$

The result for each line is a list of numbers describing the distance from each feature to the measurement point. The two highest values of $\mathbf{e}(n)$ are located and the point that is closest to the previously detected edge (in the previous time step) is chosen. The edge detector is operating on the Gaussian probability distribution $p(\mathcal{F}|\mathbf{x})$ given by Equation 7.10. Hence, the probability of skin given the data of each measurement line is computed before the edge detector is applied.

7.4.4 Dynamical models

The tracking algorithm requires a model of how the system is expected to evolve over time. The approach here is to use an auto regressive process (ARP), which is commonly used in speech and vision applications. The second order ARP is chosen because it can capture a rich variety of motions of interest and it is straightforward to learn from training data, also it is easy and efficient to implement in real time algorithms. A second order ARP, represented in discrete time is expressed as

$$\mathbf{y}_t = \mathbf{A}_2\mathbf{y}_{t-2} + \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{D}_0 + \mathbf{B}\mathbf{w}_t \tag{7.22}$$

where $\mathbf{w}_t$ are independent vectors of independent standard normal variables, $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{B}$ are fixed matrices representing the deterministic and stochastic components of the dynamical model and $\mathbf{D}_0$ is a fixed offset. Blake [12] shows how to set $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{B}$ for various tracking problems. To simplify the notation an augmented state vector can be used

$$\mathbf{Y}_t = \begin{pmatrix} \mathbf{y}_{t-1} \\ \mathbf{y}_t \end{pmatrix} \tag{7.23}$$

such that

$$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_{t-1} + \mathbf{D} + \mathbf{B}\mathbf{w}_t \tag{7.24}$$

where

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{A}_2 & \mathbf{A}_1 \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} \mathbf{0} \\ \mathbf{D}_0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{0} \\ \mathbf{B}_0 \end{pmatrix} \tag{7.25}$$

This also shows that the Markovian properties of the dynamical model, hence $\mathbf{Y}_t$ depends only on the previous state $\mathbf{Y}_{t-1}$.

The system can be described as a set of damped oscillators, whose modes, damping constants and natural frequencies are determined by the parameters $\mathbf{A}_1$ and $\mathbf{A}_2$. The oscillators are driven by random accelerations coupled into the dynamics via $\mathbf{B}$ from the noise term $\mathbf{B}\mathbf{w}_t$. The values for $\mathbf{A}_1$, $\mathbf{A}_2$ and $\mathbf{B}$ can be configured by hand by specifying oscillator parameters. The oscillator parameters consists of a damping constant β, a natural frequency f_0 and a root mean square average displacement ρ (Blake [12]). A more appropriate approach is to specify the parameters of the dynamical model by learning from expected tracking sequences. The learning approach was taken here using a least squares algorithm. The theory for this is presented by Neumaier [78] and Schneider [89].

7.4.5 Particle filters

Particle filters have been used in many fields of applied science. The use of particle filters are due to their generality and ease of implementation, which make them suitable for many signal processing problems. They have previously been used in for simulations in fluid mechanics, statistical mechanics, econometrics, signal processing and computer vision, in order to mention a few (MacCormick [62]). Several closely related algorithms are known under the names sequential Monte Carlo methods, bootstrap filters, condensation, Monte Carlo filters, interacting particle approximations and survival of the fittest [25]. The particle filter that has been applied in this thesis is the condensation algorithm (Blake [12], Isard [41], MacCormick [62]). It has been chosen due to its suitability in computer vision applications and its reputation for being able to track objects in considerable clutter.

The condensation algorithm can be used to track known objects in a video sequence where the background in the video sequence is not assumed to be rigid or stationary and its three dimensional structure is unknown. It is particularly useful in cases where there is a significant amount of clutter in the background. Clutter is defined as elements in the background that may be similar to parts of foreground features. The probability density for $\mathcal{Y}_t = \{\mathbf{Y}_1, ...\mathbf{Y}_t\}$ is multi-modal, hence not even approximately Gaussian, therefore the widely used Kalman filter (based on Gaussian distributions) is not suitable for this task.

The condensation algorithm is suitable in this more general situation and also it is in many cases a considerably simpler algorithm than the Kalman filter. The use of random sampling is often considered as computationally inefficient, despite the fact that the condensation algorithm runs in near real-time. This can be achieved by using relatively tight distributions for shape, which is achieved by using accurate learned models of shape and motion. The theory of the condensation algorithm is briefly explained in Appendix C.

7.4.6 Results

In this section, the contour tracker using the condensation algorithm is tested on cluttered and uncluttered image sequences. The contour tracking algorithm is made to track a single target using a single camera. The “shape space” for tracking is built from a hand drawn template of the head, which is allowed to deform via planar affine transformations (and six DOF). The initial template $\mathbf{Q}_0$ can be extracted from the first frame in the video sequence. The control points are positioned around the edge of the head and spline interpolation is performed between the control points. The initialisation of the contour tracker is a crucial matter because it will affect the performance of the whole tracking sequence. The initialisation of the position of the contour in the image is performed with the image segmentation technique described in Section 7.2. The ARP motion model parameters are set by learning dynamics from expected tracking sequences.

7.4.6.1 Contour tracking with a plain background

The ability to track a moving persons head is tested using a video sequence of 300 frames. Figure 7.16 and Figure 7.17 illustrate the tracking sequence. Representing the state density with $N = 70$ samples at each time step proves sufficient for successful tracking. The contour tracker correctly estimates the head position throughout the video sequence. The algorithm is implemented in Matlab [66] and runs at a couple of Hz on a Pentium 4 computer. The initialisation of the spline template is performed by using the same initialisation that the skin colour tracking algorithm uses in Section 7.2.

Figure 7.16 illustrates a number of frames from a video sequence with a plain background. The subject is mainly moving in the lateral plane. As the subject is moving around the contour tracker follows the centre of the head closely and the shape of the head is also retained relatively well. Figure 7.17 illustrates a number of frames from a video sequence with a plain background and occlusion. The subject is holding its hand in front of the head for a couple of seconds and then moving his hand in away from the head in order to distract the contour tracker. The hand movement distracts the tracker but the tracker recovers and the contour is fitted to head again when the hand is removed.

7.4.6.2 Contour tracking with a cluttered background

Figure 7.18 illustrates frames from a video sequence with a cluttered background. Again, the subject is mainly moving in the lateral plane and as the subject is moving around the contour tracker follows the centre of the head closely. The shape of the head is also here retained relatively well. Figure 7.19 illustrates frames from a video sequence with a plain background and occlusion. The subject is moving his hand in front of the head (as for the plain background case) in order to distract the contour tracker. The hand movement distracts the tracker but the tracker does not lose its target and the shape of the contour fits the head closely after the hand is removed.

7.5 Discussion

The motivation for using a camera system for listener head tracking is the relatively low cost of implementation and that the listener does not need to wear any sensors. The presented colour tracking is relatively robust to background noise, tracks the position of the head accurately enough in terms of translation and is computationally efficient.

One problem with the algorithm is light sensitivity, and to improve this parameter it is recommended that Nearly Infra Red (NIR) and Infra Red (IR) tracking is investigated. The combination of RGB bands with three narrow NIR bands is suggested by Storring [93] to robustly detect skin under changing illumination and to distinguish it from objects similar to skin colour. The disadvantage with this approach is the need for a dedicated sensor instead of a standard RGB camera and the added computational complexity of using six dimensions instead of three dimensions. Another shortcoming is that for some applications it is desirable to include listener rotation, which the colour tracking algorithm can not handle. The colour tracking algorithm can be suitable for applications where the listener is facing in a certain known direction, such as when a screen is present. The colour tracking algorithm can be extended to robust tracking of multiple listeners.

The stereo area correlation process for finding the distance to the listener can possibly be replaced by using "blob" size as a measure for distance. Since, the size of the colour blob that represents the head is proportional to the distance between the camera and the head. However, the accuracy of the distance measure is likely to be significantly less than for the stereo area correlation algorithm. A drawback with using stereo tracking is obviously the need for two cameras.

The contour algorithm is recommended if the listener is allowed to move more freely and when there is plenty of clutter in the environment. The contour algorithm is more robust with respect to background noise compared to the colour tracking algorithm. The contour tracking algorithm can also be extended to robust tracking of multiple listeners (MacCormick [62]) as well as to track listener rotation. The particle filter that was incorporated in contour tracking framework adds predictive capabilities that can possibly reduce filter update times. The main disadvantage of the contour tracking algorithm is the computational effort it requires.

An interesting algorithm that has not been evaluated in this thesis is eye tracking. The possibility of using eyes as a feature to track is interesting especially for systems with a screen, such as a TV or desktop computer. Eye tracking can also be used to verify that the listener is still present in the view field of the camera and can be applied for example in the ROI of the colour tracking algorithm.

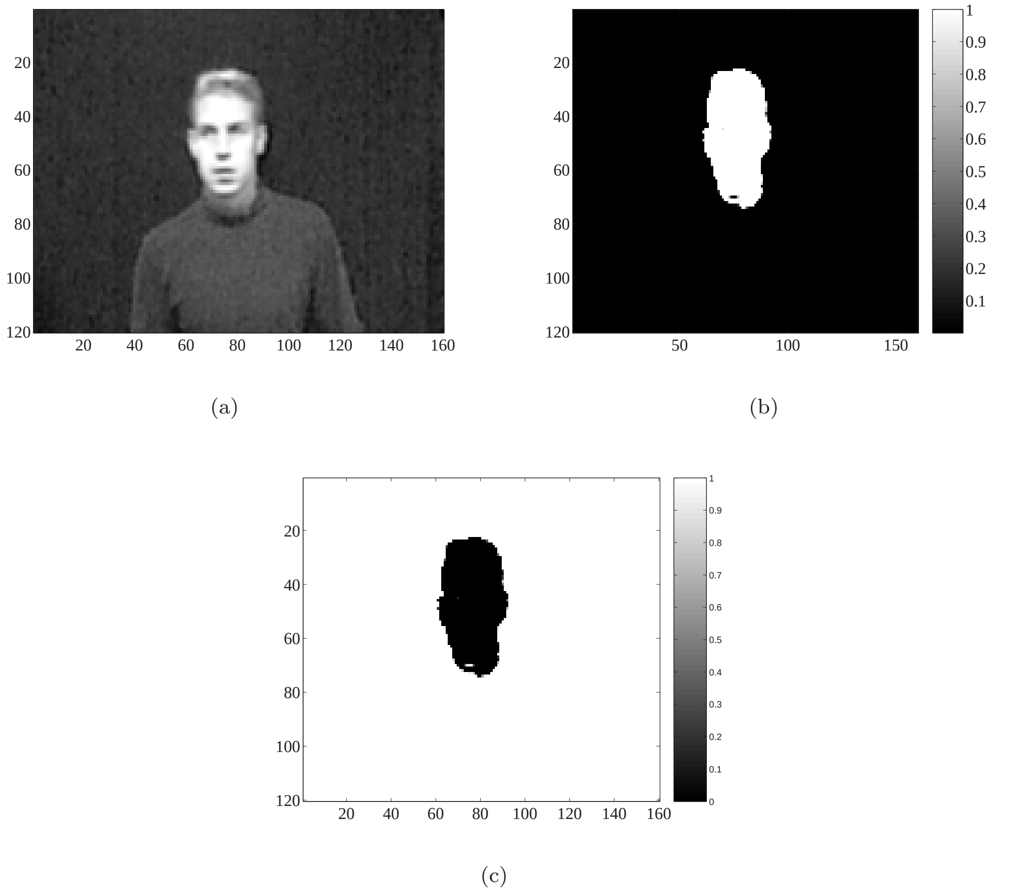
Another promising application for image processing in virtual sound imaging is pinna shape detection. Image processing can be used to detect the shape of the listeners pinna, which then can be mapped to an approximately individual HRTF [56].

7.6 Conclusion

The development of image processing algorithms for virtual sound imaging systems has been presented. A fundamental understanding has been reached of the main factors affecting system performance with respect to listener head tracking using a camera system. A skin colour distribution can be characterised by a multivariate Gaussian distribution. This can provide a rough estimate of the location of a head within the image frame to initialise a listener head tracking system and to track the translation of the listener in the following frames. The measured standard deviation of the colour tracking algorithm in the X-direction is 0.02 m. Stereo area correlation tracking has been used to find the distance of the listener. The standard deviation of the stereo area correlation algorithm in the Z-direction is 0.05 m. The tracker can be suitable for virtual sound imaging systems, since the JND criterion established in Chapter 6 is 0.025 m in the X-direction and > 0.15 m in the Z-direction.

A contour tracking algorithm has been briefly investigated and its capabilities have been demonstrated. The main reasons for investigating contour tracking was to capture the motions that the colour tracking algorithm can not track and to improve the robustness with respect to clutter. The contour algorithm tracks with six DOF compared to the two DOF for the presented colour tracking algorithm. However, the contour algorithm did not run in real-time in this Matlab [66] implementation. Therefore the combined colour tracking and stereo area correlation algorithm was applied in the implementation presented in Chapter 8 using frame rate of 30 Hz.

To conclude, image processing algorithms for listener head tracking can improve the overall performance of virtual sound imaging systems significantly. Listener head tracking in virtual sound can increase the effective listening area by steering the “sweet-spot” and thereby improve the quality of the reproduced sound. The main advantage of using visual tracking compared to other tracking techniques is that the listener does not need to wear any sensors.



(a)

(b)

(c)

Figure 7.1: **Multivariate Gaussian skin classifier**. *(a) Original RGB image. (b) Image of skin probabilities $\mathbf{L}^{'}_{\mathcal{F}}$ obtained with a multivariate Gaussian classifier. (c) Background image $\mathbf{L}^{'}_{\mathcal{B}}$ multivariate Gaussian classifier.*

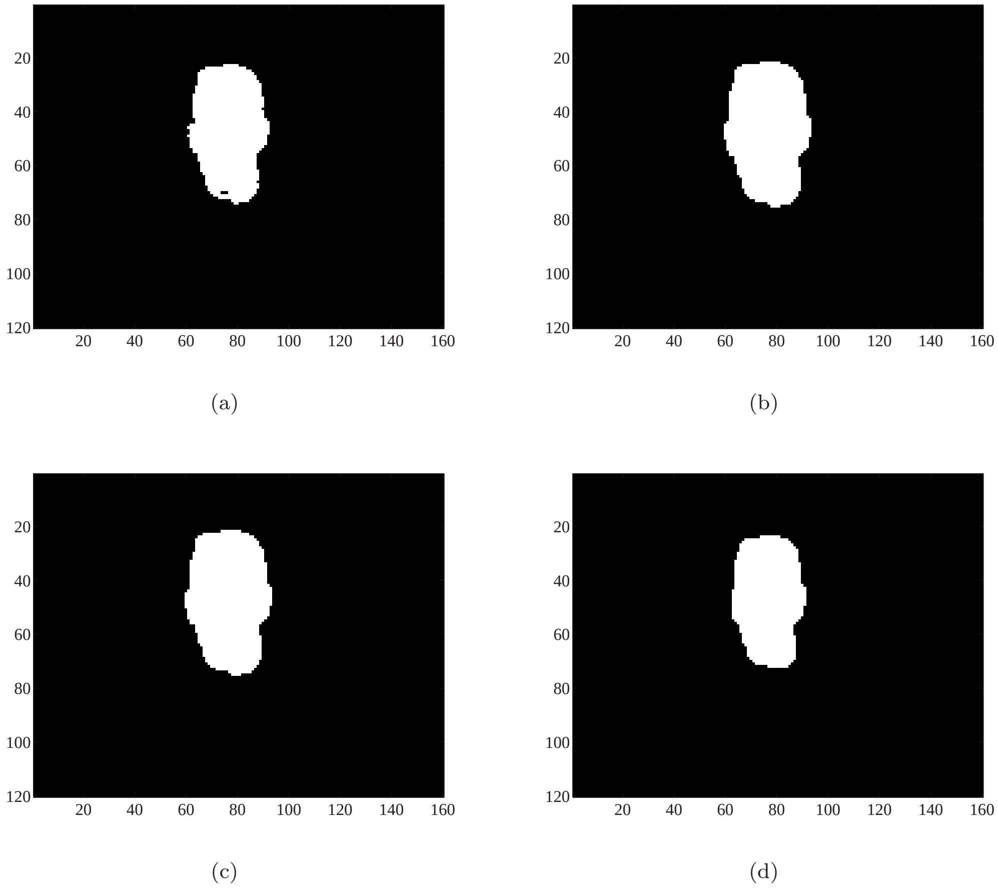
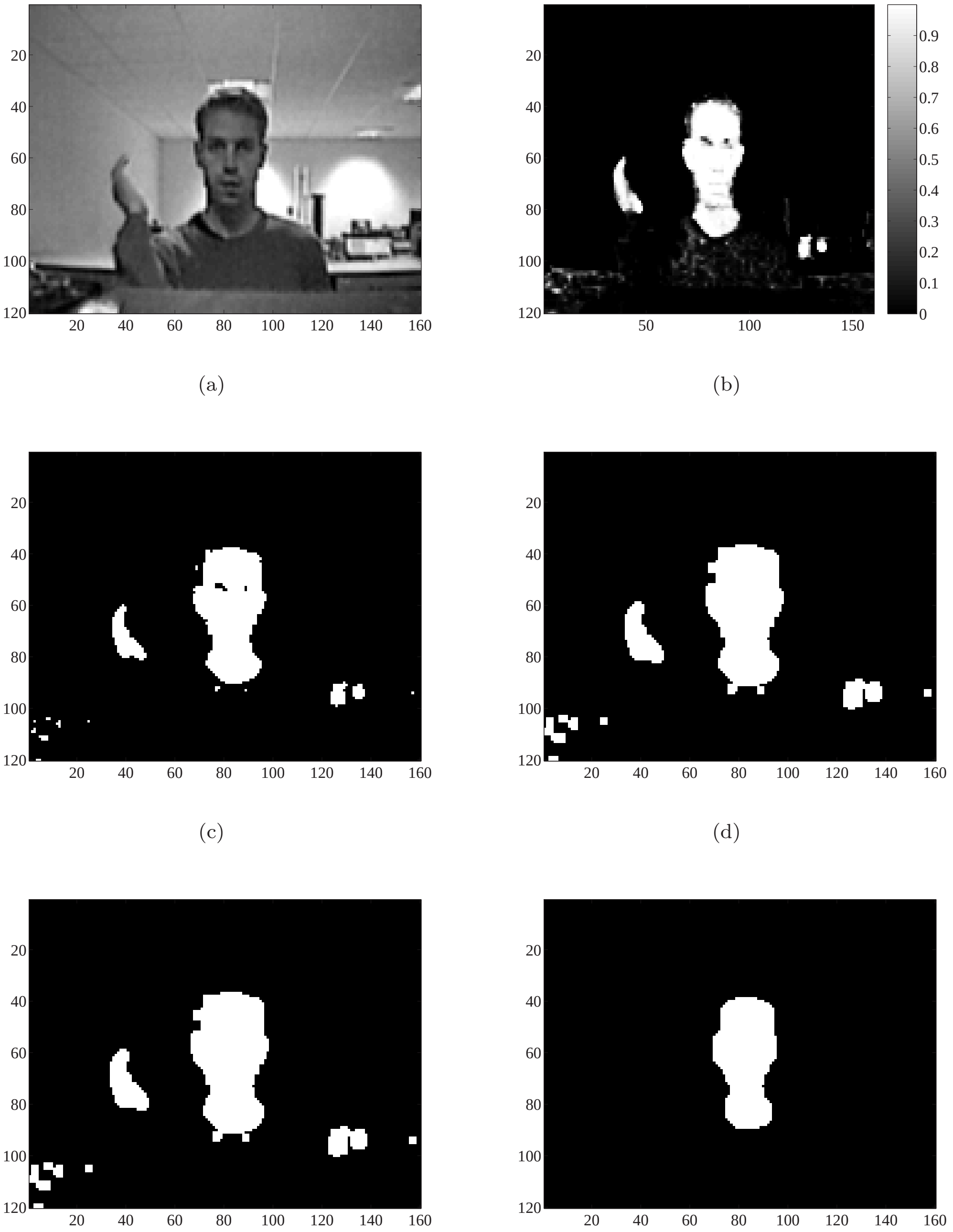


FIGURE 7.2: **Image segmentation with a plain background**. *Segmentation of the skin image* $\mathbf{L}'_{\mathcal{F}}$ *in Figure 7.1 (b). (a) Greylevel thresholding. (b) Apply dilated gradient mask. (c) Fill the holes in the image. (d) Perform morphological opening.*



(a)

(b)

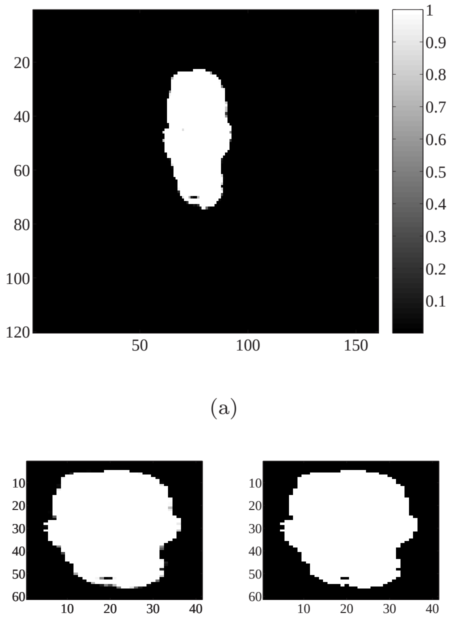
(c)

(d)

(e)

(f)

Figure 7.3: **Image segmentation with clutter**. *Segmentation of skin image* $\mathbf{L}_{\mathcal{F}}^{'}$ *where the objects hand is present. (a) Original RGB image. (b) Image of skin probabilities* $\mathbf{L}_{\mathcal{F}}^{'}$ *obtained with a multivariate normal classifier. (c) Greylevel thresholding. (d) Apply dilated gradient mask. (e) Fill the holes in the image. (f) Perform morphological opening.*



(a)

(b)

(c)

Figure 7.4: **Initialisation of the skin colour tracker and the region of interest**. *(a) Image of skin probabilities* $\mathbf{L}^{'}_{\mathcal{F}}$ *obtained with a multivariate normal classifier. (b). ROI cropped from (a). (c) Thresholded ROI.*

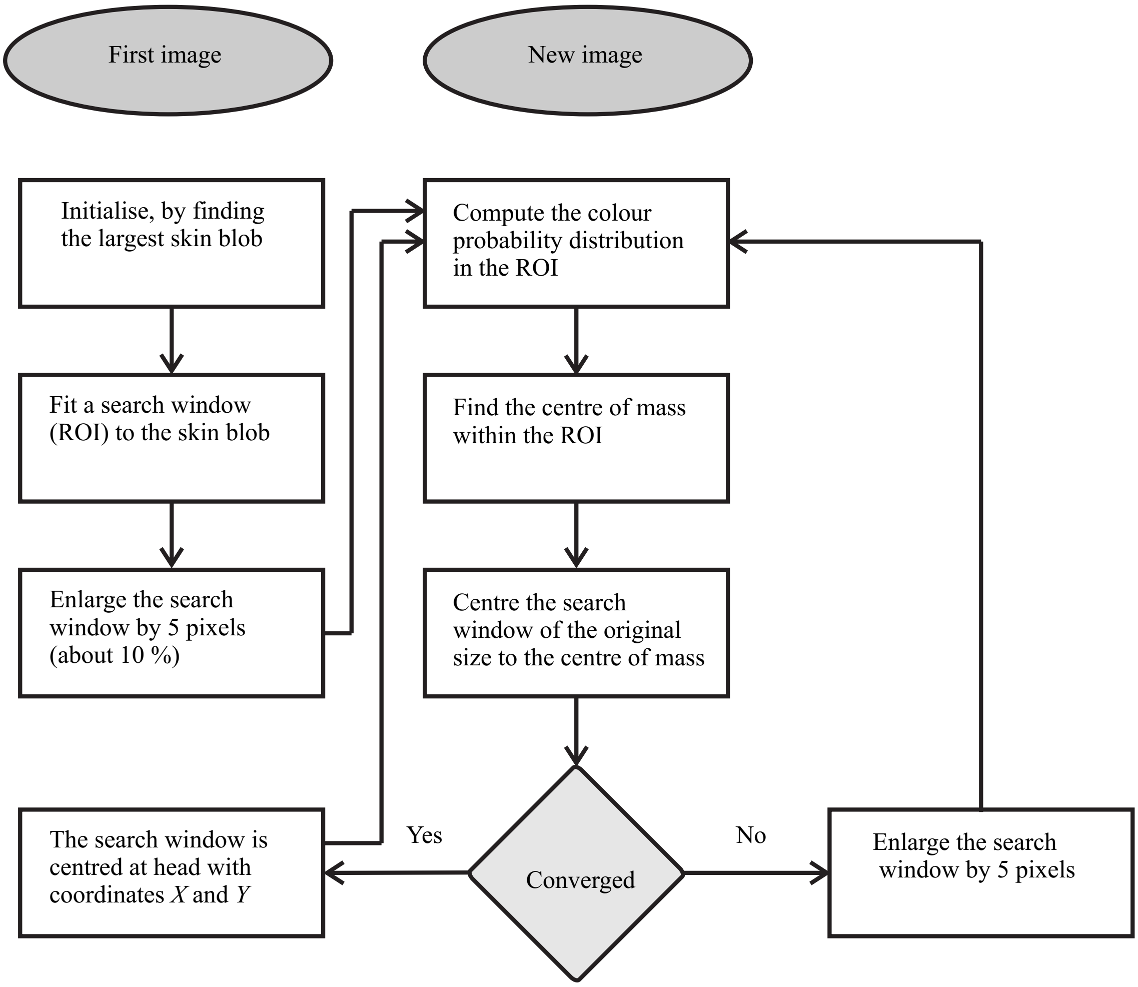
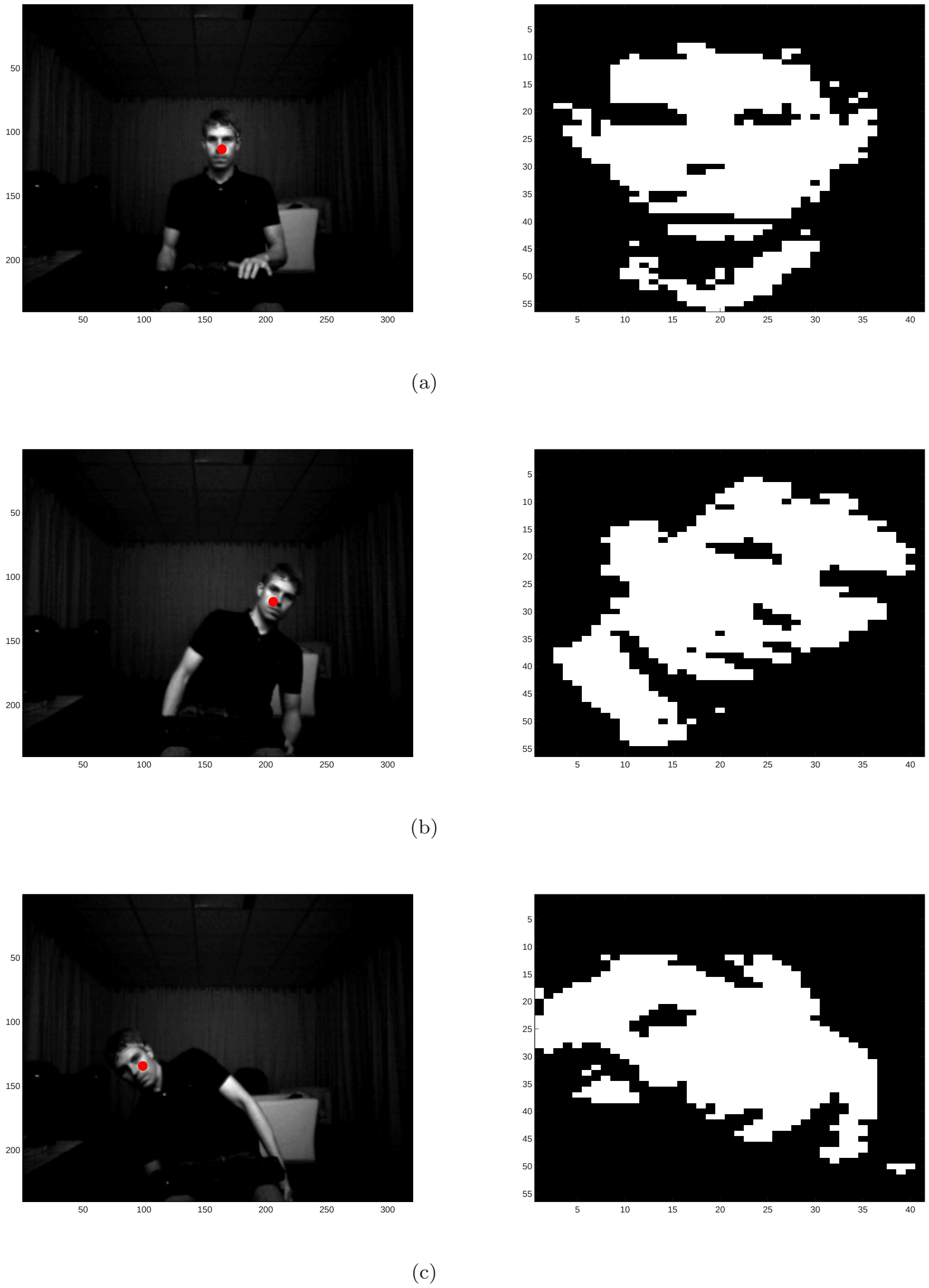

First image
New image
Initialise, by finding the largest skin blob
Fit a search window (ROI) to the skin blob
Enlarge the search window by 5 pixels (about 10 %)
Compute the colour probability distribution in the ROI
Find the centre of mass within the ROI
Centre the search window of the original size to the centre of mass
Converged
Yes
No
The search window is centred at head with coordinates X and Y
Enlarge the search window by 5 pixels

Figure 7.5: **Flowchart of the colour tracking algorithm**.

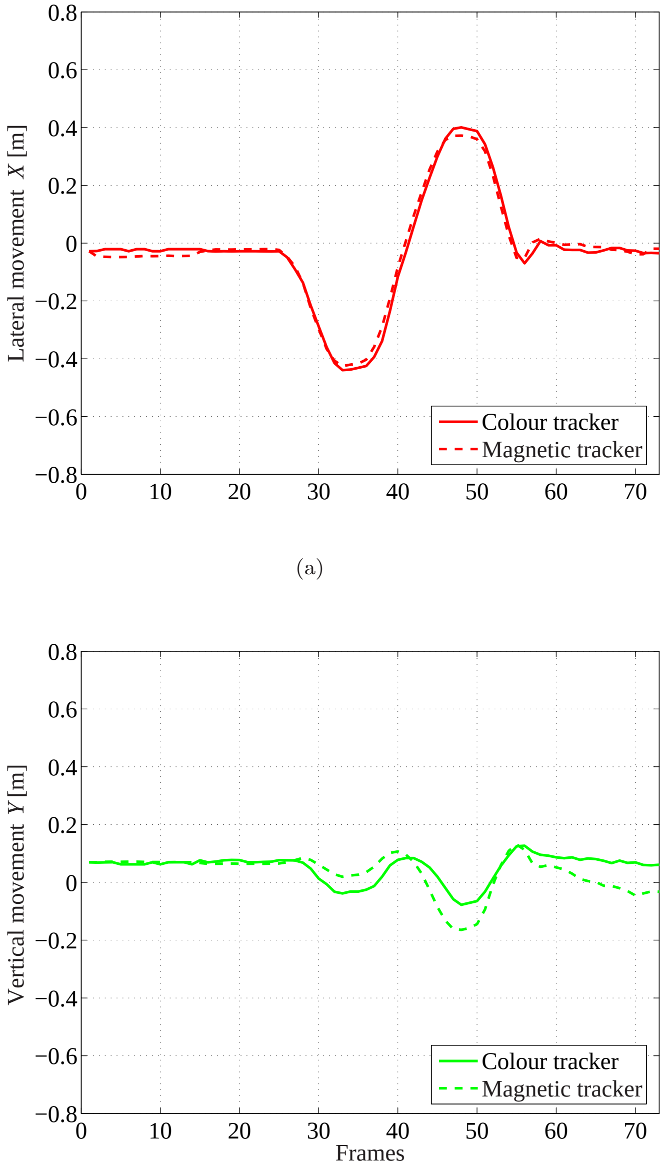


(a)

(b)

(c)

Figure 7.6: **Visual tracking using skin colour detection**. *Tracking sequence for the evaluation of colour tracking and magnetic tracking (Polhemus). The listener was moving mainly in the lateral plane and in the fore and aft plane. (a) Frame 1. (b) Frame 30. (c) Frame 48.*



(a)

(b)

FIGURE 7.7: **The performance of visual tracking using skin colour detection compared to magnetic tracking**. *Comparison of the colour tracking algorithm to a magnetic tracking system (Polhemus). The data was collected from the tracking sequence in Figure 7.6. (a) Lateral movement X. (b) Vertical movement Y.*

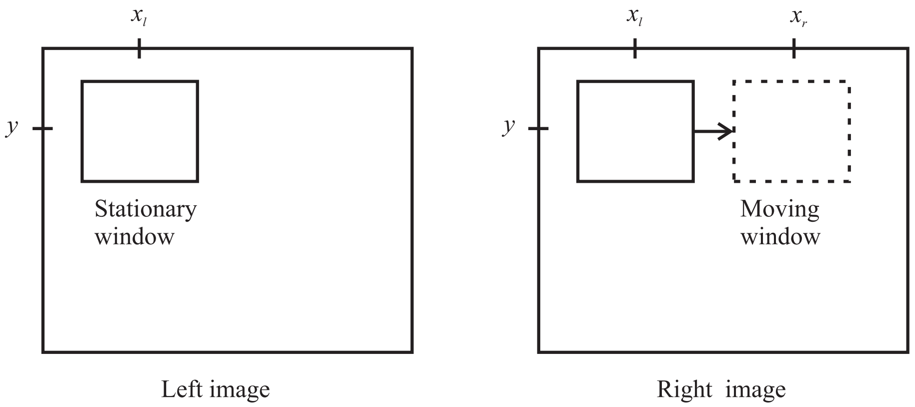


Figure 7.8: **Stereo area correlation using the pixel shift method**. *The stationary window is in the left image and hence the moving window in the right window. For each pixel, the area correlation is calculated by moving the window in the right image. Then the area correlation is calculated by the Laplacian of Gaussian (LOG) transform and L1 norm (absolute difference).*

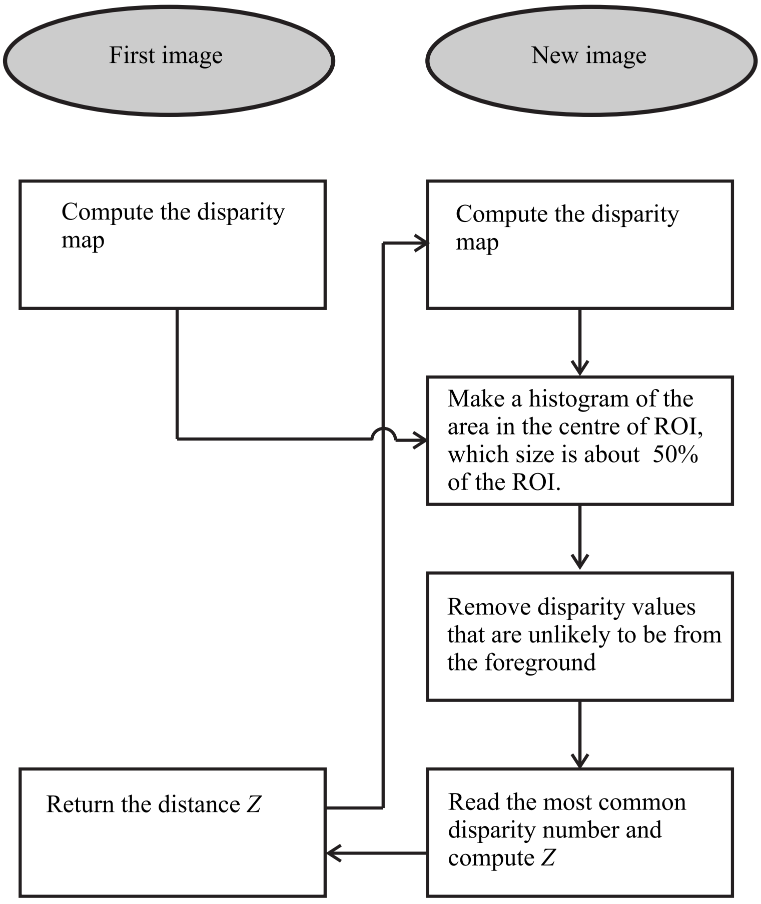

First image
New image
Compute the disparity map
Compute the disparity map
Make a histogram of the area in the centre of ROI, which size is about 50% of the ROI.
Remove disparity values that are unlikely to be from the foreground
Return the distance Z
Read the most common disparity number and compute Z

Figure 7.9: **Flowchart of the stereo area correlation algorithm**.

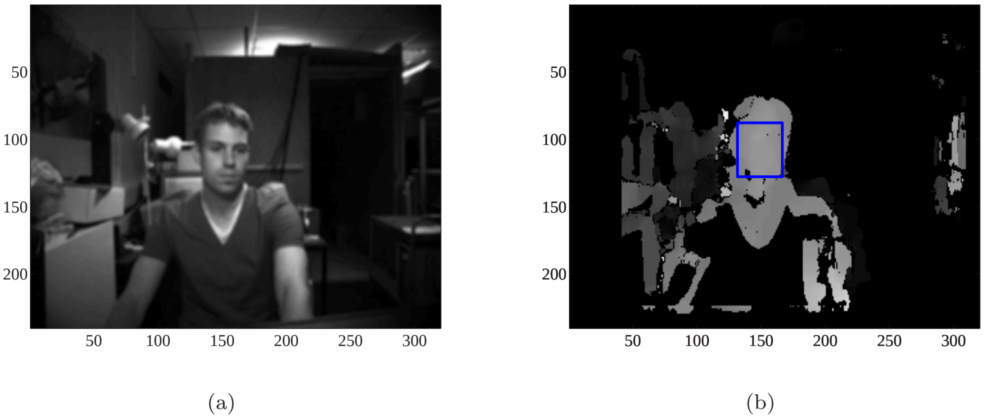


FIGURE 7.10: **The disparity image from the stereo area correlation algorithm**. *(a) Original input image. (b) The disparity image with the ROI indicated by the rectangle.*

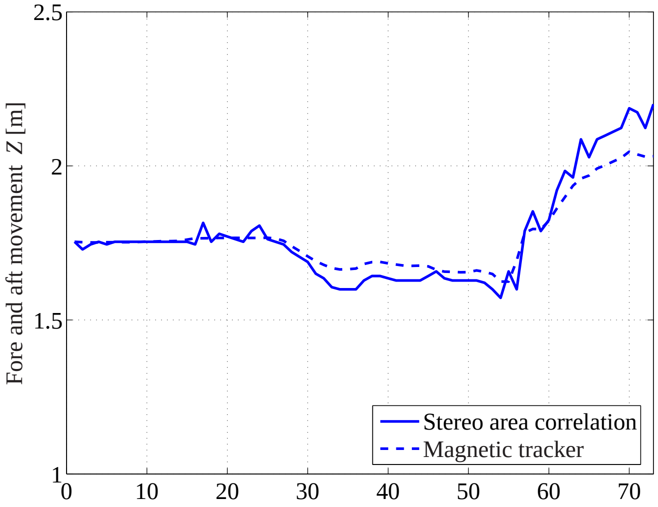


FIGURE 7.11: **The performance of visual tracking using stereo tracking compared to magnetic tracking**. *Comparison of the stereo area correlation processing compared to a magnetic tracking system (Polhemus). The data was collected from the tracking sequence in Figure 7.6.*

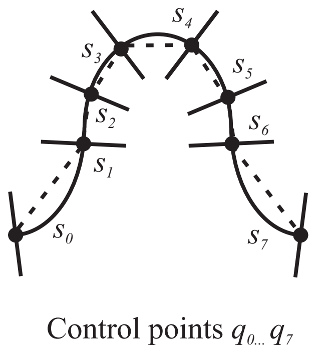


FIGURE 7.12: **Contour with eight control points**. *The illustrated contour is a quadratic parametric spline curve that is open. The dotted line is the control polygon and the curve is an approximation. The control polygon is formed from a sequence of control points* $q_0, ..., q_7$.

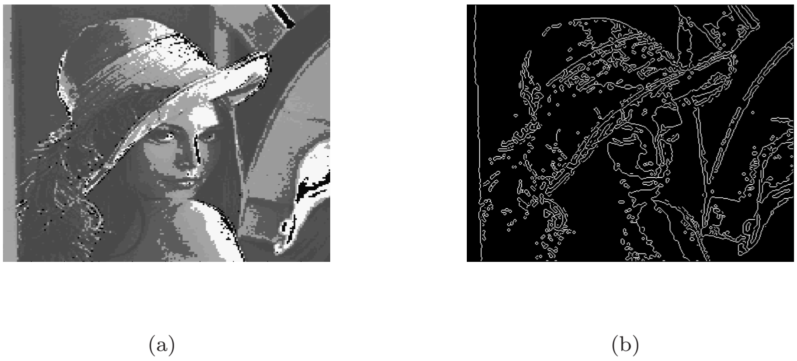


FIGURE 7.13: **Feature map**. *A feature map is created from applying the Sobel edge detector (Appendix A) on the Lena image. (a) Original image. (b) Feature map* **F**.

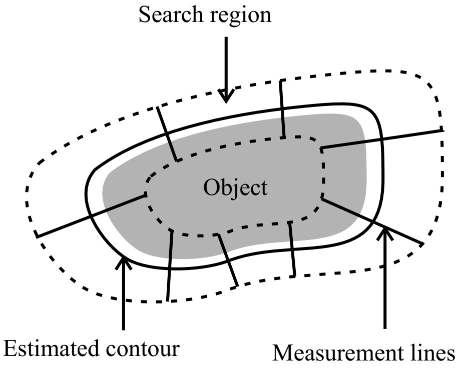


FIGURE 7.14: **Contour fitting**. *The principle of fitting a contour to an object. Contour with a search region, which corresponds to where the image features are likely to lie.*

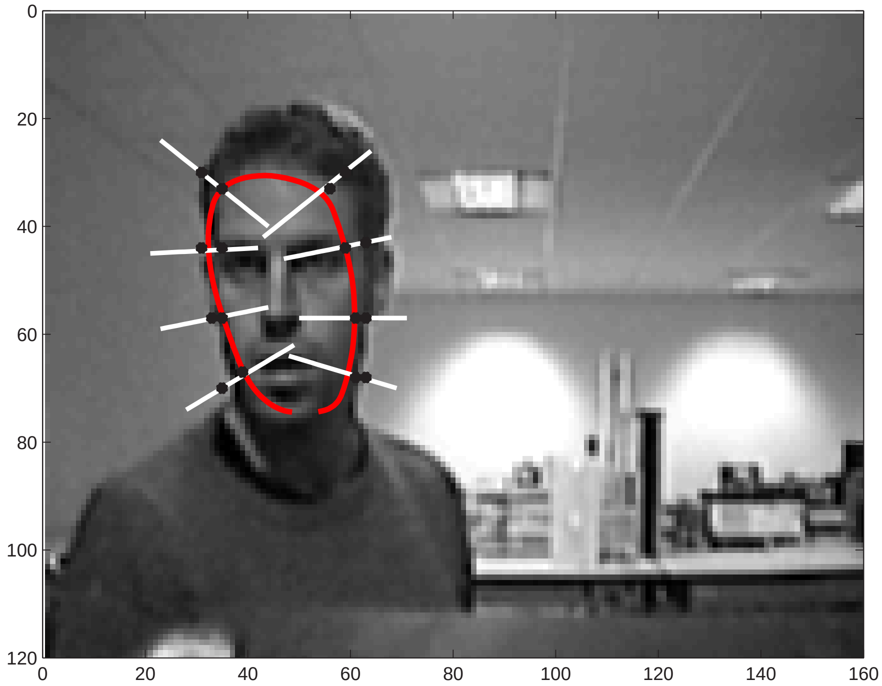
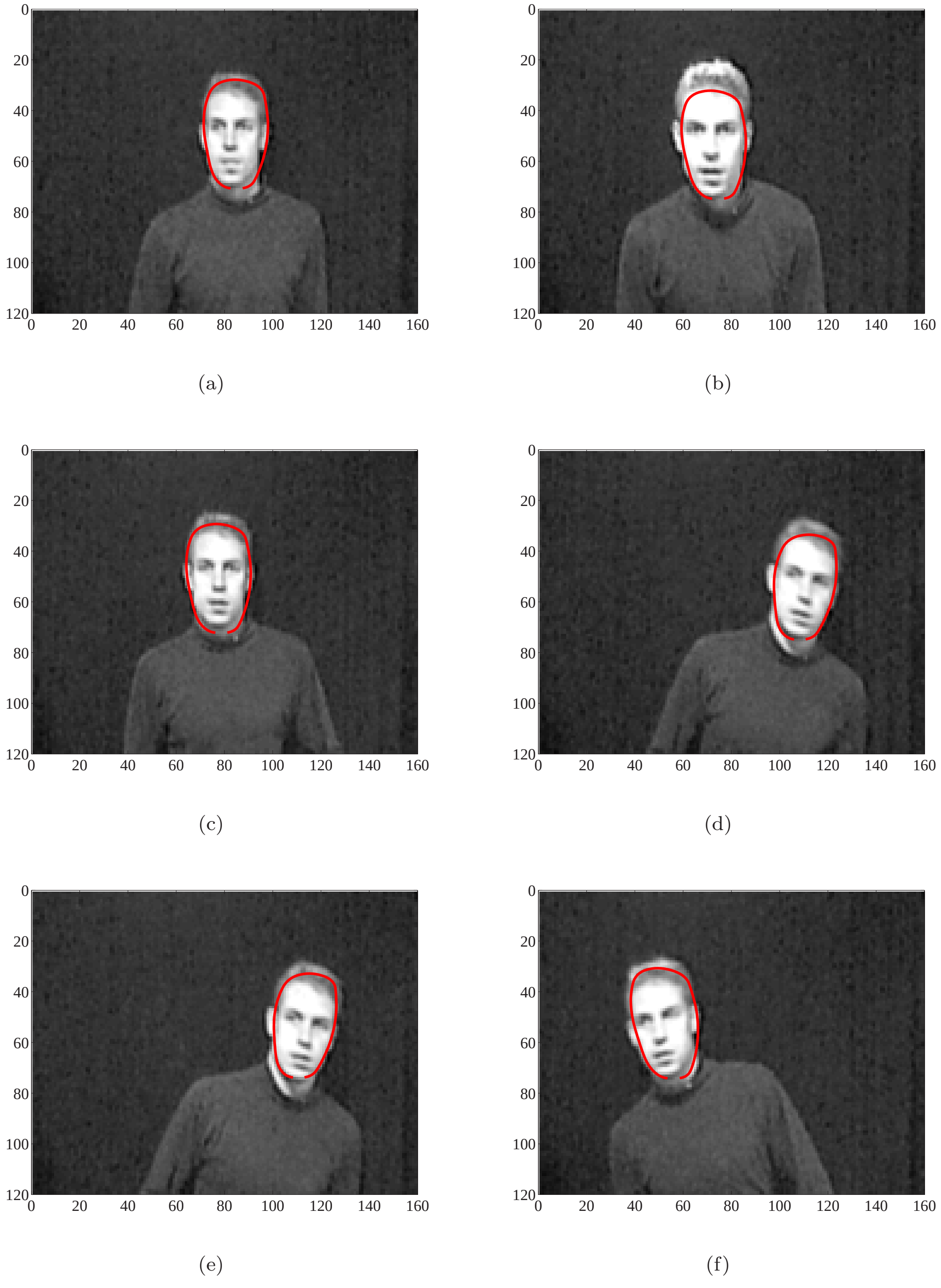


Figure 7.15: **Contour with measurement lines that is fitted to the head of a listener**. *Edge detection is carried out on each measurement line in order to to fit the contour to the object. The edge detection scheme is enhanced by using skin colour detection on each measurement line prior to applying the edge detector.*



(a)

(b)

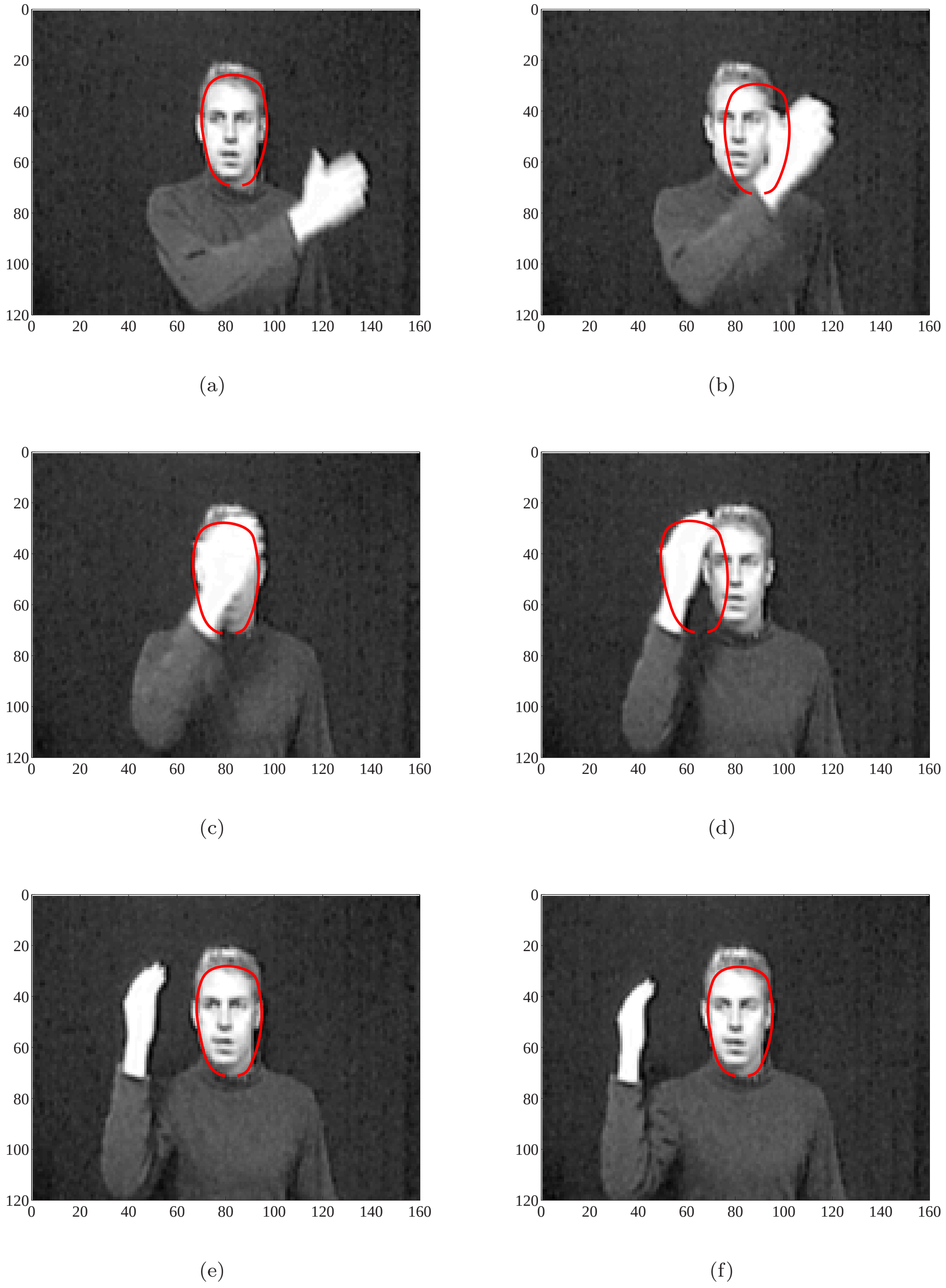
(c)

(d)

(e)

(f)

Figure 7.16: **Contour tracking with a plain background**. *The contour tracker is operating on an image sequence with a plain background. (a) Frame 10. (b) Frame 20. (c) Frame 30. (d) Frame 40. (e) Frame 50. (f) Frame 60.*



(a)

(b)

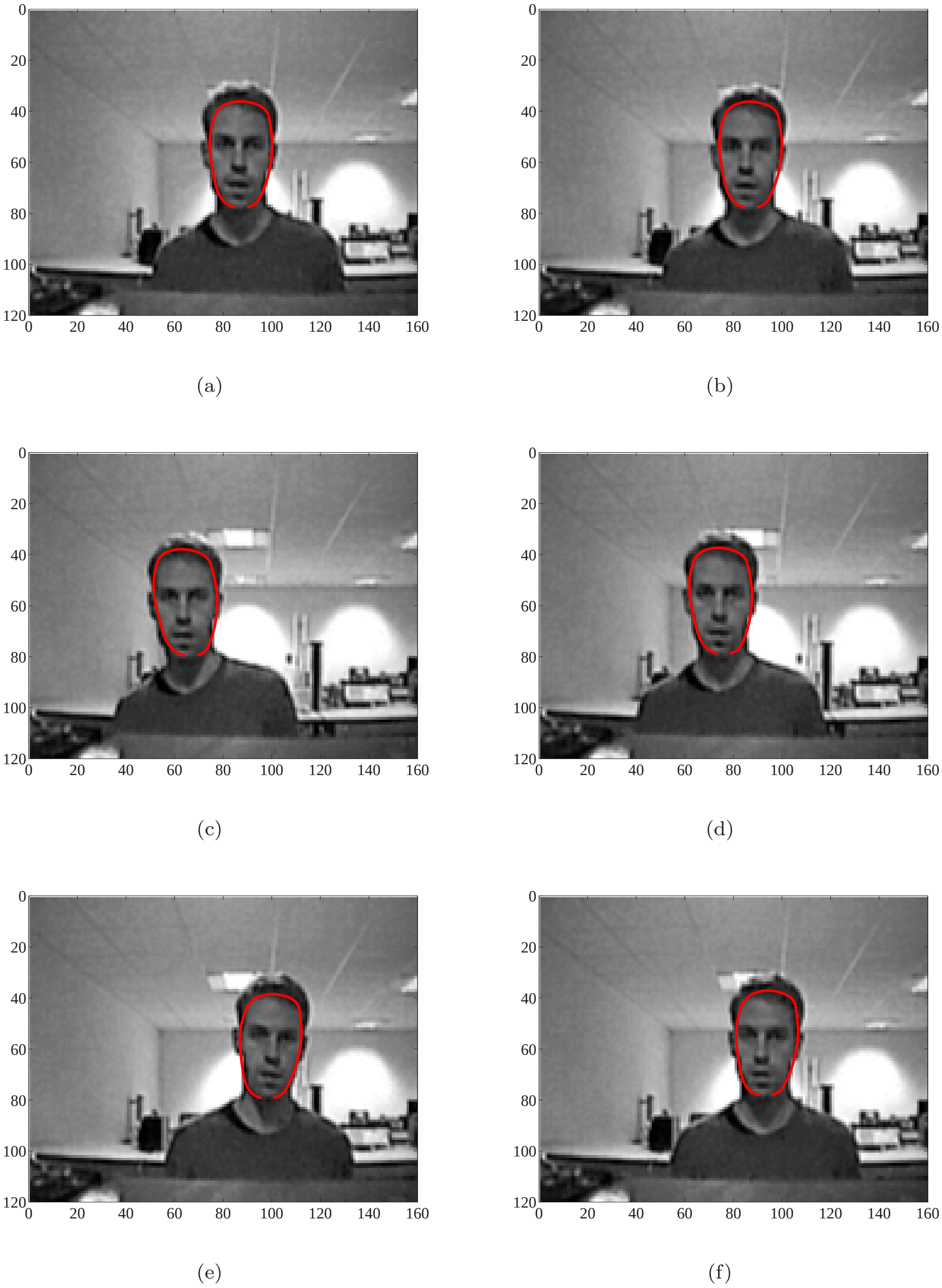
(c)

(d)

(e)

(f)

FIGURE 7.17: **Contour tracking with a plain background and occlusion**. *The contour tracker is operating on an image sequence with a plain background including a moment of occlusion. (a) Frame 118. (b) Frame 120. (c) Frame 122. (d) Frame 124. (e) Frame 126. (f) Frame 128.*



(a)

(b)

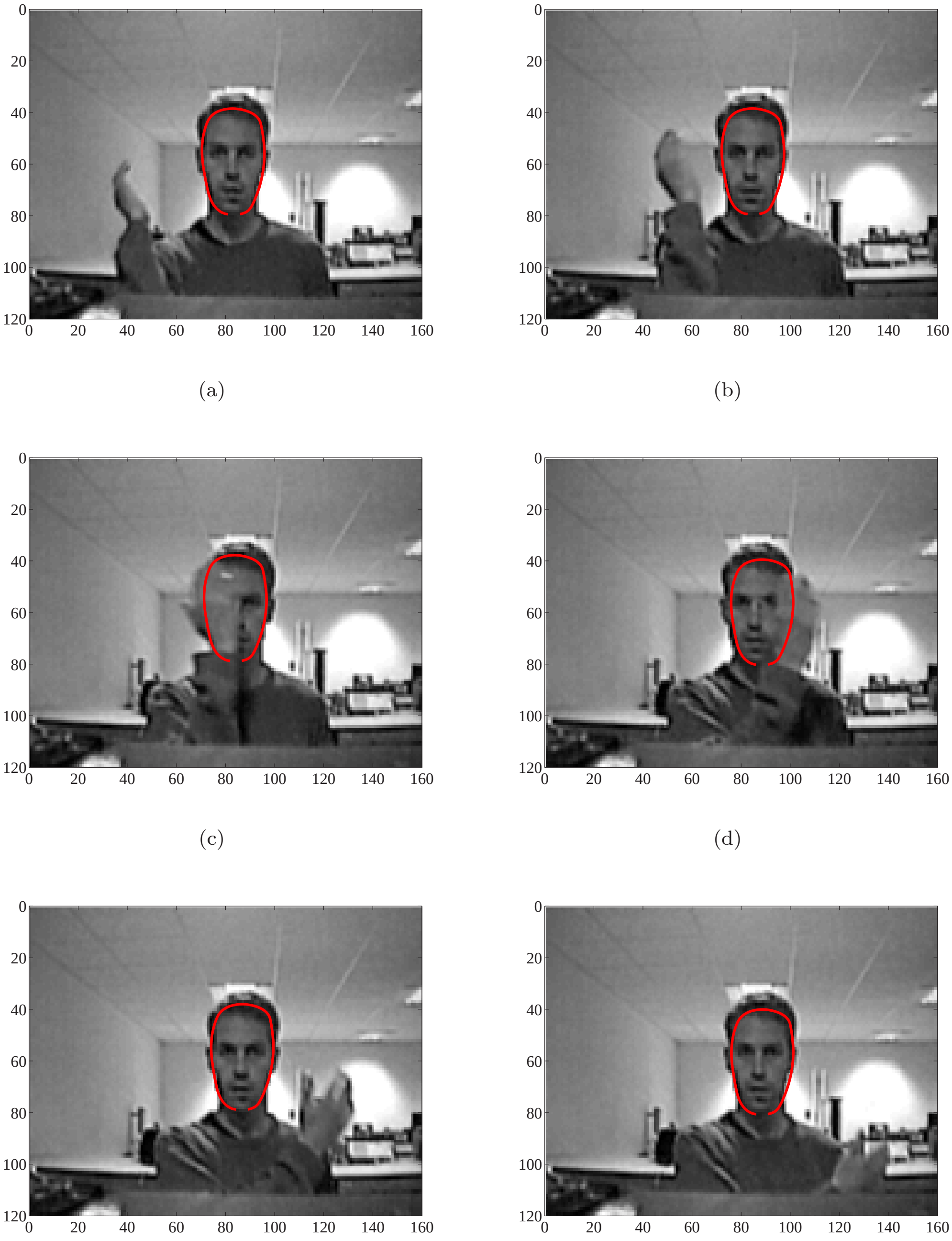
(c)

(d)

(e)

(f)

FIGURE 7.18: **Contour tracking with cluttered background**. *The contour tracker is operating on an image sequence with a cluttered background. (a) Frame 60. (b) Frame 70. (c) Frame 80. (d) Frame 90. (e) Frame 100. (f) Frame 110.*



(a) (b)

(c) (d)

(e) (f)

Figure 7.19: **Contour tracking with cluttered background and occlusion**. *The contour tracker is operating on an image sequence with a cluttered background and a moment of occlusion. (a) Frame 288. (b) Frame 290. (c) Frame 292. (d) Frame 294. (e) Frame 296. (f) Frame 298.*

Chapter 8

Experiments on an integrated system

The implementation of a visually adaptive SD system is introduced that combines visual head tracking and binaural sound reproduction. The system is capable of tracking the position of the listener in the lateral plane and in the fore and aft plane while dynamically updating the cross-talk cancellation filters. The algorithms are implemented using a software approach on a Pentium 4 computer. The realized filter update algorithm is described in detail. The performance of cross-talk cancellation filters and the filter update algorithm is objectively evaluated through measurements in an anechoic chamber. The results from the objective evaluation are compared to previously presented subjective experiments.

3D sound for virtual reality and multimedia applications is presented by Begault [11] and to quote him "*it is impossible to predict the position of the listener or of the speakers in any given situation*", although this is true for any given situation, under certain constraints it is possible to track the position of the listener relative to the loudspeakers. Here it has been shown that it is possible to track the position of a listener that is not wearing any sensors using image processing techniques. The constraints on the implemented algorithm require that the background has to have a different colour distribution compared to the face of listener and the light during tracking has to be consistent with the light when the listener skin sample was chosen.

A realisation of an adaptive cross-talk cancellation system for a moving listener is presented by Lentz [59]. The system is based on static cross-talk cancellation that is updated depending on the listener's position. Listening tests show that the dynamic cross-talk cancellation produces impressive results and that the listener can move in an area of about 1 m^2. This system uses a magnetic head tracking system with a sensor attached to the listener. More recently Lentz [60], [58] has developed the adaptive cross-talk cancellation system further to include four loudspeakers, which can support 360° listener rotation. The system includes room-specific simulation, dynamic cross-talk cancellation and multi-track binaural synthesis. The head tracking is performed with a commercial infrared tracking system that requires the user to wear markers that the infrared cameras can detect. The system is configured in a CAVE like environment, which is a five sided projection system of a rectangular shape installed at RWTH Aachen University. An other example of virtual sound imaging system is the DIVA project [88], which was designed in order to produce a real-time environment for immersive audiovisual processing.

8.1 System

The visually adaptive system that was developed at the ISVR, University of Southampton is described here. The system is implemented in a first version and it is open to any extended virtual sound modelling applications in real-time. The system is named VSI (Virtual Sound Imaging) and developed using object oriented programming (C++). The aim was to write a program that could be used on a standard computer. The current version is optimised for a Pentium 4 computer and runs on a single processor. In order to run both the image processing and the audio signal processing on a single processor requires highly efficient algorithms. Therefore, the image processing algorithm that was implemented needed to be relatively simple in order to keep the computational cost to a minimum. The currently implemented colour tracking fulfils the efficiency requirements. The most demanding parts of the audio signal processing comprise of the fast-deconvolution algorithm and the partitioned convolution process.

There are other solutions already available for the audio signal processing like the Lake Huron system. Though, the Huron is a highly specialised system that is also limited in terms of the real time filter switching capabilities that was required for the VSI system. The VSI system is different from previous approaches from several viewpoints. The system adapts the sweet spot position dynamically for a moving listener without the need for the user to wear any sensors. The system is not limited to any constrained hardware requirements, such as specific audio DSP technology. The audio signal processing part of the system is realised using a software approach and can be used with readily available computer hardware.

The system layout is illustrated by the diagram in Figure 8.1. The head tracking algorithm finds the listener position and passes the value to the thresholding part of the system. The threshold for lateral movement corresponds to a 1° change in azimuth angle and the threshold for fore and aft movement is configured to 0.03 m. The fore and aft coordinates are smoothed by a moving average filter using 10 averages. This introduces some un-wanted lag. However, the precision of the estimate of the fore and aft coordinate does not need to be as precise as for the lateral coordinate, and the introduced lag can be accepted if the listener mainly moves in the lateral plane. New inverse filters are computed when the threshold for listener movement is breached. The filters are computed by choosing the closest HRTF pairs from an ITD-equalised database with a resolution of 1°. The delay information in the measured HRTF database is removed in advance by the interpolation algorithm described in Section 4.1. The database that was used here holds 512 coefficients including the on-axis loudspeaker response. The delay and amplitude scaling models for the filter calculation are based on the far field spherical radiation approximation. A Hanning window of the same length as the inverse filter is applied in order to avoid ringing effects. The inverse filters are then updated as described by the partitioned convolution and commutation process described in Section 8.2. The sound output device is configured with the Audio Stream Input Output (ASIO) protocol, which was designed for professional audio recording. The ASIO protocol provides an interface between an application and the sound card.

8.2 Partitioned convolution and commutation

Binaural sound reproduction using loudspeakers requires multichannel convolution and for adaptive systems it is also desirable to have a low latency between the filter updates. Partitioned convolution provides low input/output delay in the convolution process. For an adaptive loudspeaker system it is necessary to update the inverse filters according to where the listener is positioned. The filter update process results in audible artifacts if the inverse filters are updated on too coarse a grid. Therefore commutation can be used in order to remove the audible artifacts. This section shows the theory for partitioned convolution and how a commutation scheme can be implemented directly in the partitioned convolution algorithm in order to make it suitable for adaptive cross-talk cancellation.

The cross-talk cancellation scheme is performed by convolving the two input signals with four inverse filters. These inverse filters are computed by chosing appropriate HRTFs and using the fast-deconvolution algorithm. The convolution task can be simplified by performing FFTs and IFFTs, since time domain convolution becomes multiplication in the frequency domain. The FFT assumes inherently that the analysed segment is periodic, which leads to unsatisfactory results if a straightforward block by block implementation in the frequency domain is used. Therefore, the periodicity of the FFTs needs to be removed from the output sequence. This can be achieved by using the overlap and save algorithm, which is illustrated in Figure 8.2. The input window of M samples is shifted to the right over the input sequence of $M - N$ samples, and then convolution is performed for the following segment. The disadvantage is that FFTs of lengths $M > N$ are necessary. A common configuration for the overlap-and-save algorithm is to have $M = 2 \times N$, which also gives the best efficiency (Torger [98]).

The partitioned overlap-and-save algorithm is the one that has been used in this project. It was first proposed by Stockham [92] and further developed for real-time implementation by Soo [91]. Torger [98] has previously used partitioned convolution for binaural sound reproduction. Firstly, the impulse response of the inverse filter $s_t(n)$ is partitioned into a suitable number of P blocks of equal size, which is illustrated in Figure 8.3. Then each partition is convolved by a standard overlap-and-save process with FFTs of lengths L (here $L = 2 \times K$). In the process, zero padding is used for making each block L samples in length, and the transformation into the frequency domain is carried out with the FFT, which results in a number of frequency domain filters $\Pi_{i,t}(k)$. These frequency domain filters $\Pi_{i,t}(k)$ are multiplied by the FFTs of the input blocks and then IFFTs are used to revert to the time domain and the result is summed as described in Figure 8.4. This process gives the same result as the un-partitioned convolution, but with the main difference that the latency of the partitioned convolution process is only L samples compared to M samples. Hence, the input/output delay can be controlled by choosing a suitable value for L.

The commutation process is implemented directly in the partitioned convolution algorithm. The relationship between old, new and commutated coefficients are as follows

$$\Pi_{i,c}(k) = \left(\frac{N_c - i}{N_c}\right)\Pi_{i,a}(k) + \frac{i}{N_c}\Pi_{i,b}(k) \tag{8.1}$$

where the commutated inverse filter block is labeled $\Pi_{i,c}(k)$, the previous inverse filter $\Pi_{i,a}(k)$ and the new inverse filter $\Pi_{i,b}(k)$. The commutated inverse filter $\Pi_{i,c}(k)$ is then updated for each new block $\Pi_{1,c}(k), \Pi_{2,c}(k) \ldots, \Pi_{N_c,c}(k)$. For example, if the filter is segmented into 8 blocks and there are 16 commutation steps, then each of the inverse filter coefficients are switched in 2 steps, as illustrated in Figure 8.5.

The commutation process is carried out in real-time by dynamically updating the coefficients. The commutation technique described here uses FIR filtering and an ITD-equalised non-minimum phase HRTF database. The non-minimum phase characteristic can result in comb filtering effects in the frequency domain due to phase differences between the interpolated HRTFs. The non-minimum phase FIR approximation has been found to produce severe comb filtering effects if the phase delays vary considerably. This may occur if the resolution of the HRTF database is $> 10^\circ - 15^\circ$ (Huopaniemi [38]). However, this can be minimized by using a fine resolution in the HRTF database, like the 1° resolution used here.

8.3 Filter update rate

The time it takes between the update of a new listener position and for the output signal to be generated with the newly calculated filters is the latency of the system.

The output buffer of the sound card and the block length of the partitioned convolution algorithm is 256 coefficients. This results in a latency of 5.8 ms at a sampling rate of 44.1 kHz for the output of a single block. The fast-deconvolution algorithm takes 6 ms to compute a new set of inverse filters (2048 coefficients) in the current configuration (Pentium 4 computer). A filter length of 2048 coefficients at a sampling frequency of 44.1 kHz correspond to 45.2 ms in the time domain. The image processing algorithm takes 10 ms to compute a new position for the current implementation of the colour tracking and stereo area correlation algorithms. However, the image processing algorithm updates the images at a frame rate of 30 Hz, which is the upper limit of the current hardware configuration, which leads to a delay of 33.3 ms for finding the new listener position. If the commutation and partitioned convolution algorithm is configured to 8 partitions and 8 steps of commutation, the resulting time it takes for a complete new inverse filter to be present on the output is $45.2 + 5.8 + 33.3 + 6 = 90.3$ ms. If the commutation part is left out and the filters are updated directly, then the latency for the new filter to become active would be $5.8 + 33.3 + 6 = 45.1$ ms. This could be further reduced if the camera system would run at higher frame rate of for example 60 Hz, which would result in a latency of $5.8 + 16.7 + 6 = 28.5$ ms.

The update rate of the real-time system is an important aspect here. The filter update rate has been studied intensively by subjective experiments. Two criteria for filter updates namely the JND and JNC criteria were established in Chapter 6. In order to apply these criteria it is important to know the velocity of the movements of the listener. A field study by Lentz [60] shows that the mean translational velocity is at 0.15 m/s for their CAVE configuration. The CAVE allows the listener to walk around in an area of a couple of m^2. The listener position in the VSI system is updated for lateral movement when the angle of an HRTF has changed by 1°. This results in a filter update every 0.035 m at a distance of 2 m away from the loudspeakers. If the listener moves at 0.15 m/s in the lateral plane then a filter update occurs every 233 ms. If listener rotation is allowed, then less delay would be required for the filter update, according to field studies by Lentz [60], a recalculation of the inverse filters needs to be carried out every 10-20 ms with a filter update rate of 35 ms.

The filter update algorithm that was implemented is close to matching the JND criteria at 2 m distance for the listener. The filter updates in the lateral plane every 1° of change in HRTF angle. This corresponds to 0.035 m at 2 m distance on-axis. The JND criterion in the lateral plane is slightly higher and was found to be approximately 0.025 m in Chapter 6. However, the filter update criteria was established using bandlimited white noise as a stimulus and preliminary listening tests suggests that when using music and speech signals the update rate can be coarser to some extent. The system configuration with 8 partitions and 24 commutation steps also fulfills the JNC criterion. The filter coefficients are on average stepped 0.012 m in this configuration (8 partitions, 24 commutation steps), which is close to the JNC criterion that was determined to be approximately 0.012 m. Hence, the filter update algorithm can be configured with 8 partitions and 24 commutation steps as a default in order to comply with data from subjective experiments. The time it takes for a complete filter update using this configuration is $(3 \times 45.2) + 5.8 + 33.3 + 6 = 180.7$ ms.

The filter update rate can be reduced by using shorter inverse filters. The inverse filter length of 2048 coefficients used here can possibly be reduced by cropping out the 1024 most important coefficients and still keeping the delay intact (Lentz [60]). It could be reduced further by using a database of HRTFs that contains less coefficients. An alternative filter update technique could possibly be used. The outputs from two adjacent inverse filters could be "cross-faded" instead of the commutation process described in Section 8.2. The result from this has not been investigated here. However, it would take twice as many inverse filters in the convolution process, i.e. 8 filters instead of 4 filters. This leads to a more computationally demanding algorithm but with the advantage of possibly reducing the filter update rate. The cross-fading technique can also be used with IIR filters, which can reduce the number of filter coefficients and the computational effort. Another method for reducing the time lag in the system is to use a listener tracking algorithm that predicts the future listener position, such as the "condensation" algorithm (Blake [12]).

8.4 Evaluation of filter updates

An objective evaluation of the implemented filter update algorithm was undertaken. The aim was to determine how well the filter update algorithm would perform for different configurations with respect to the number of partitions and commutation steps. Three different configurations using different numbers of partitions and commutation steps were evaluated.

The system was configured in an anechoic chamber at the ISVR as is illustrated in Figure 8.6. The SD loudspeaker configuration was used and the KEMAR dummy head was used for measuring the filter updates. The distance between the loudspeakers and the KEMAR was 2 m and the source signal was a 1 kHz sine wave. The loudspeakers were moved in front of the KEMAR and the visually adaptive system was configured to track the dummy head. The loudspeakers were moving at a rate of 0.01 m/s, which was the maximum allowed rate of change for the loudspeaker moving rig that was available. The loudspeakers were moved from on-axis to 0.2 m to the right of the dummy head in the lateral plane. The source signal used in this experiment was a 1 kHz sine wave that was played from the VSI system and then recorded with the KEMAR dummy head connected to the Huron measurement system. The database in the software was the small pinna with open ear canal and with the on-axis loudspeaker response included. The same loudspeakers that were used for the HRTF database measurement were used here so that the loudspeaker response could be included. The KEMAR had the same configuration as when the HRTF database was measured, i.e. configured with the small pinna and the open ear canal.

In order to evaluate the filter update scheme, a different number of partitions and commutation steps were investigated for a filter length of 2048 samples. The first configuration holds 2 partitions and 2 steps of commutation, the aim was to simulate a direct filter update (the software could not be configured for un-partitioned convolution without rewriting the code). The second configuration uses 8 partitions and 8 steps in the commutation scheme. This aims to show how the software can perform in a configuration with a block size of 256 samples using a minimum amount of commutation. The third configuration uses 8 partitions and 48 steps of commutation. The number of commutation steps are increased to show the smoothness that can be achieved in a filter transition.

The results are presented in the time domain in Figure 8.7 for a 1 kHz source signal (reference). Figure 8.7 (a) illustrates the case when the loudspeaker is fixed at on-axis and 1 kHz source signal is played during 20 s. It can be seen that the configuration with 2 partitions and 2 commutation steps results in unwanted spikes during the filter update. The configuration with 8 partitions and 8 commutation steps illustrates that the spikes are removed but the filter updates seems to be quite sharp still. The last configuration with 8 partitions and 48 commutation steps illustrates that the spikes are removed and that a much more smooth transition is achieved.

The same measurement data as in Figure 8.7 are here presented in a time-frequency plot in Figure 8.8. The spectra is here represented by the magnitude of the short-time Fourier transform as described in Oppenheim [80]. The time-frequency plots were created using the following parameters: a block size of 2048 samples, a frequency discretisation of 4096 samples and an overlap of 1536 samples. These values where chosen to enhance the clarity of the filter updates. The extra components in the reference spectra shown in Figure 8.8 (a) are due to harmonics produced by the loudspeakers and the amplifier. In Figure 8.8 (b)-(d) there are additional components created by the step motor that was unfortunately quite noisy. The vertical lines in the spectra are due to the filter updates. The results in Figure 8.8 (b) clearly indicates high frequency noise that is audible by human hearing. Figure 8.8 (c) illustrates that the filter update is not as pronounced as in the previous case when increasing the number of commutation steps. Figure 8.8 (d) illustrates that the an increasing number of commutation step results in a significantly smoother filter transition.

8.5 Evaluation of the effectiveness of cross-talk cancellation

The cross-talk cancellation performance was measured in an anechoic chamber using the same system configuration as in the filter update evaluation above. The measurement was carried out in the lateral plane from -35 cm to 35 cm relative to on-axis in 5 cm steps. The measurement was carried out using the Huron workstation. The source signal was bandlimited white noise (300-20000 Hz) of 32768 samples long and 10 times averaging was applied in the evaluation of spectra. The source signal was loaded into the software and played from there. The KEMAR was connected to the Huron in order to record the resulting signals. The measurement data was further averaged 20 times during post-processing.

The measured system cross-talk cancellation performance is illustrated for the static system in Figure 8.9 and for the adaptive system in Figure 8.10. Figure 8.9 (a) illustrates the signal $P_{11}(k)$ (as defined in Equation 2.29). It can be observed that there are both peaks and dips present in the frequency response as the listener moves off-axis. Figure 8.9 (b) illustrates the signal $P_{21}(k)$. The cross-talk cancellation performance is in the region of 15 dB on-axis and degrades to around 0 dB with additional peaks and dips depending on listener position. The sweet-spot appears to be in the region of a couple of cm depending on frequency and channel separation.

Figure 8.10 (a) illustrates the signal $P_{11}(k)$ and it can be observed that the frequency response is similar for all lateral positions with some difference mainly at frequencies above 7000 Hz. Figure 8.10 (b) illustrates the signal $P_{21}(k)$ and it can be seen that the cross-talk cancellation performance is in the region of 15 dB for all measured lateral positions up to approximately 7000 Hz, where the pinna resonance occurs.

8.6 Conclusion

An evaluation of a visually adaptive virtual sound imaging system has been presented. The system is capable of reproducing binaural signals at the ears of a moving listener that is not wearing any sensors. The listener is allowed to move in the lateral plane and in the fore and aft plane. The system is implemented using a software approach and readily available computer equipment. The software that has been designed can be extended to include any desired virtual sound application. An innovative filter update algorithm has been introduced that can be implemented directly in the partitioned convolution process. The performance of the filter update algorithm has been evaluated by measurements in an anechoic chamber. The results shows that the filter update algorithm manages to produce smooth filter updates in its current application. The results from the filter update evaluation have been compared to previous subjective experiments. It has been illustrated how the filter update algorithm can be configured to comply with the subjective filter update criteria, JND and JNC that were established in Chapter 6. The cross-talk cancellation performance for static and adaptive systems has been evaluated under anechoic conditions. The features of the measurement data shows similarities to those in the simulations presented in Section 5.3.

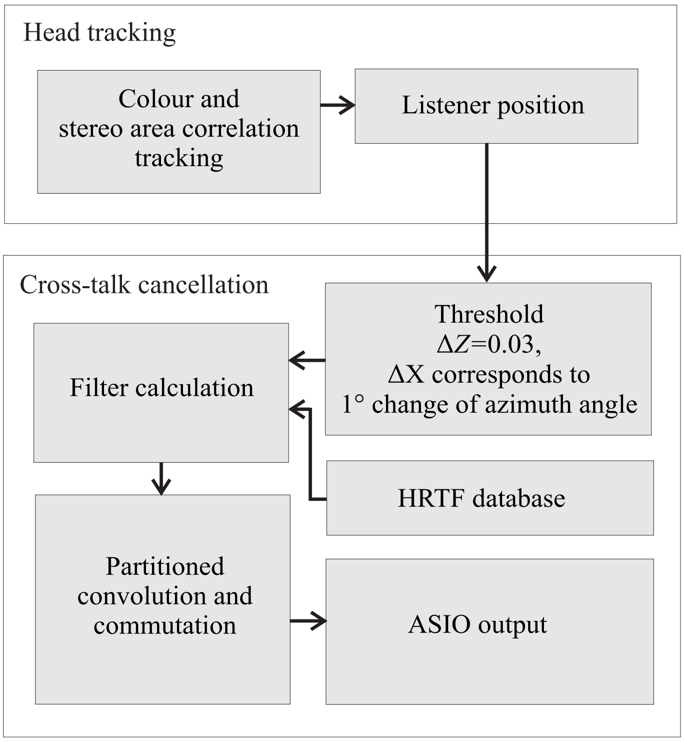

Head tracking
Colour and stereo area correlation tracking
Listener position
Cross-talk cancellation
Threshold ΔZ=0.03, ΔX corresponds to 1° change of azimuth angle
Filter calculation
HRTF database
Partitioned convolution and commutation
ASIO output

Figure 8.1: **Software implementation**. *The system is named VSI (Virtual Sound Imaging) and developed using object oriented programming (C++).*

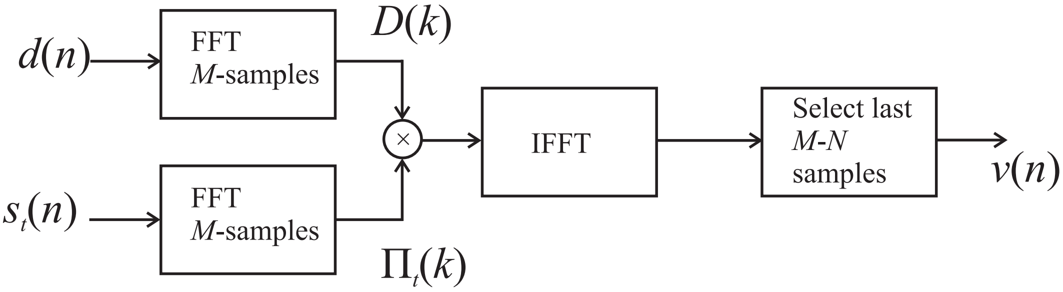


FIGURE 8.2: **The overlap-and-save algorithm**. *The impulse response $s_t(n)$ and the input signal $d(n)$ are convolved using an overlap-and save process. The process outputs $M - N$ convolved samples as the input window of size M is shifted to the right over the input sequence of $M - N$ samples. Then convolution is performed for the following segment.*

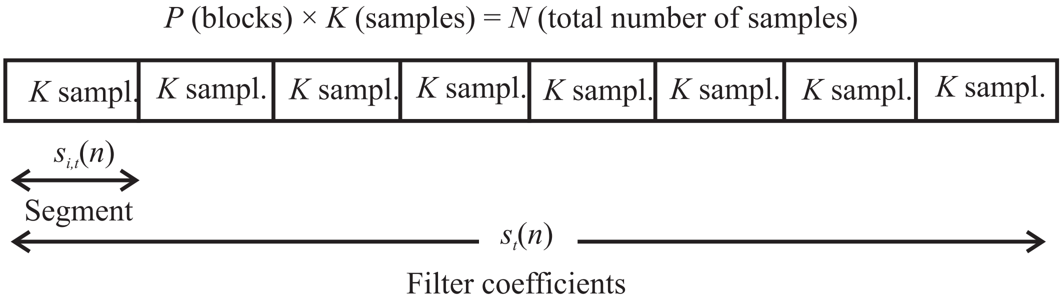


FIGURE 8.3: **Partitioning of the filter**. *The impulse response of the filter $s_t(n)$ is partitioned into a suitable number of P blocks of equal size.*

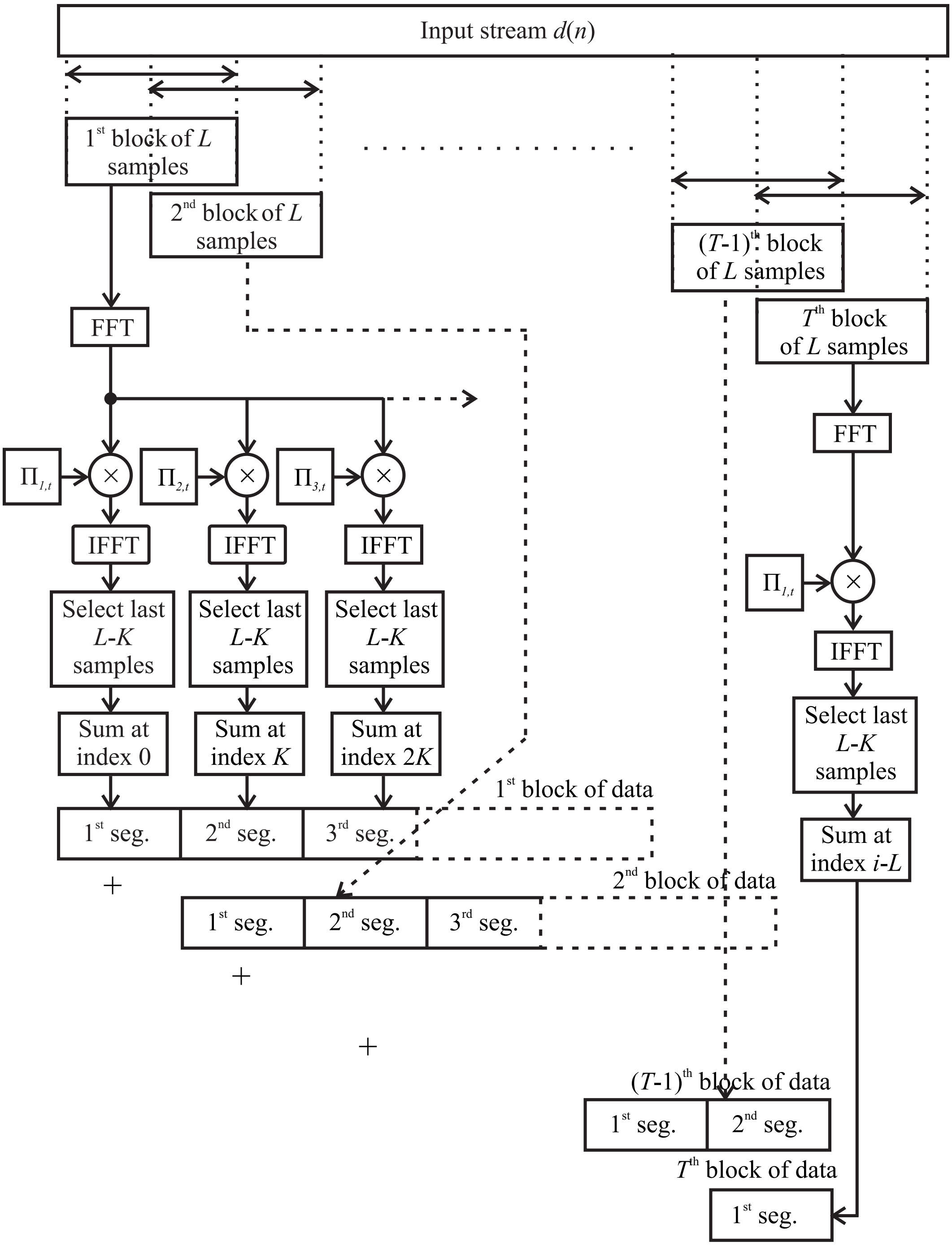


FIGURE 8.4: **The partitioned convolution algorithm**. *The latency of the partitioned convolution process can be controlled by choosing the number of samples L compared to M samples in overlap-and-save process.*

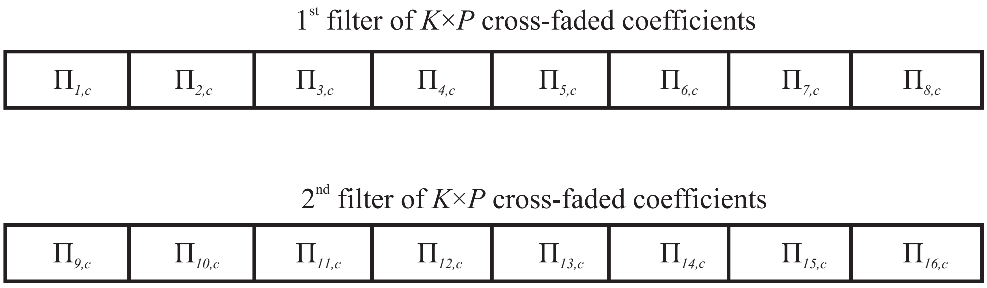

1st filter of K×P cross-faded coefficients
$\Pi_{1,c}$ $\Pi_{2,c}$ $\Pi_{3,c}$ $\Pi_{4,c}$ $\Pi_{5,c}$ $\Pi_{6,c}$ $\Pi_{7,c}$ $\Pi_{8,c}$
2nd filter of K×P cross-faded coefficients
$\Pi_{9,c}$ $\Pi_{10,c}$ $\Pi_{11,c}$ $\Pi_{12,c}$ $\Pi_{13,c}$ $\Pi_{14,c}$ $\Pi_{15,c}$ $\Pi_{16,c}$

Figure 8.5: **The commutation step in the partitioned convolution algorithm**. *In this example, the filter is partitioned into 8 blocks and switched in 16 steps.*

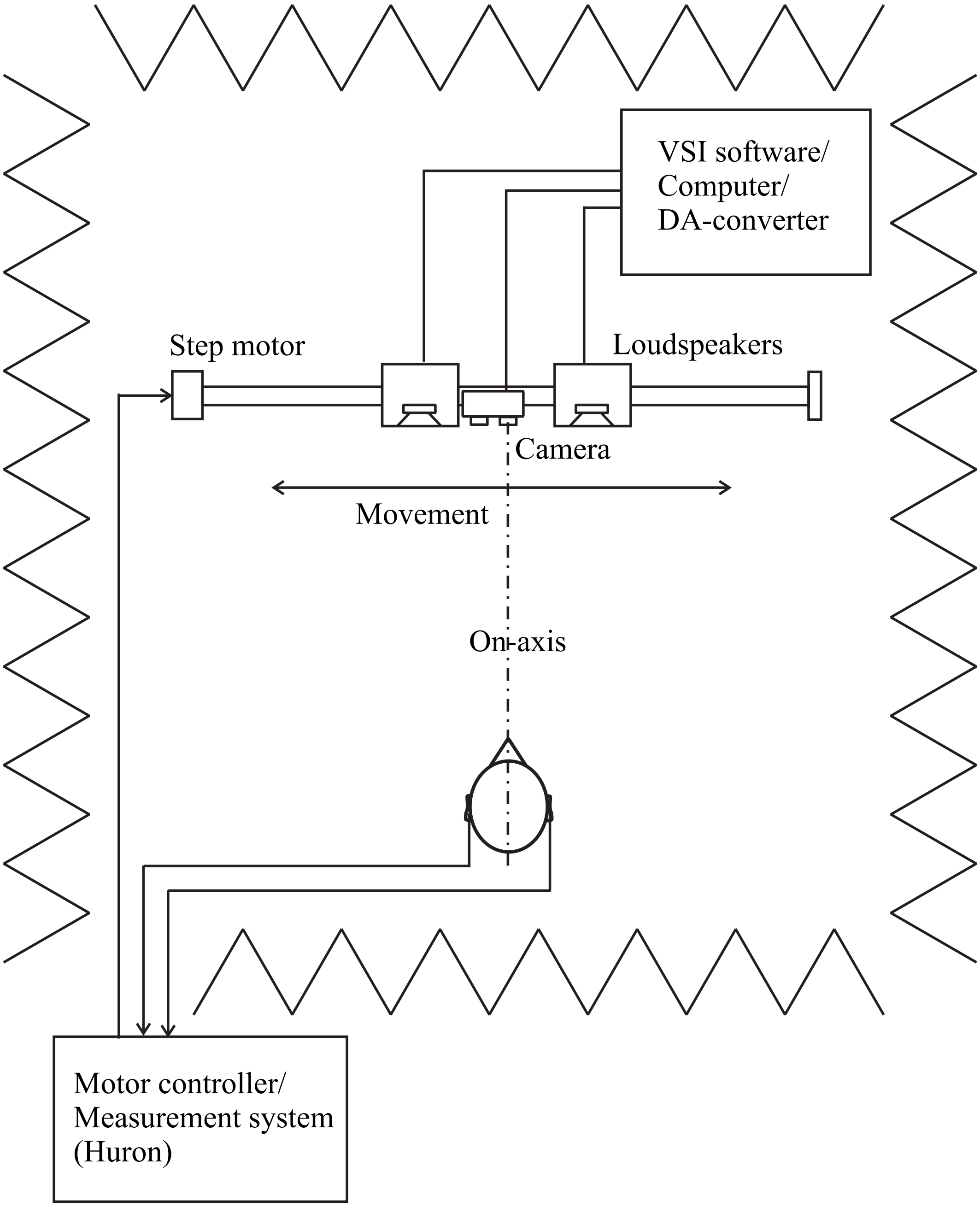
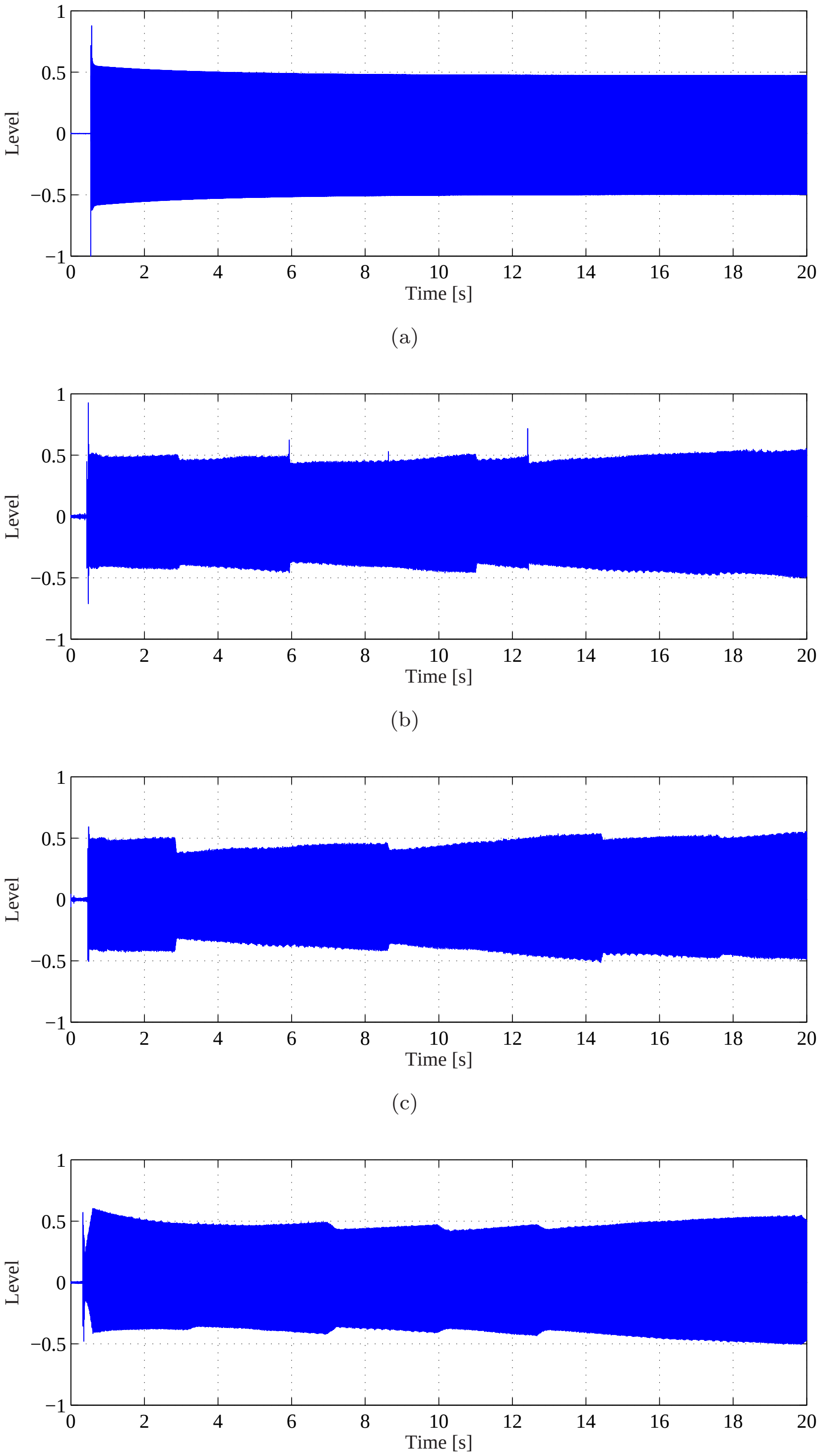


Figure 8.6: **The measurement set-up for evaluating the performance of the VSI system in an anechoic chamber**. *The SD loudspeaker configuration was used and the KEMAR dummy head was used for measuring filter updates and the effectiveness of cross-talk cancellation.*



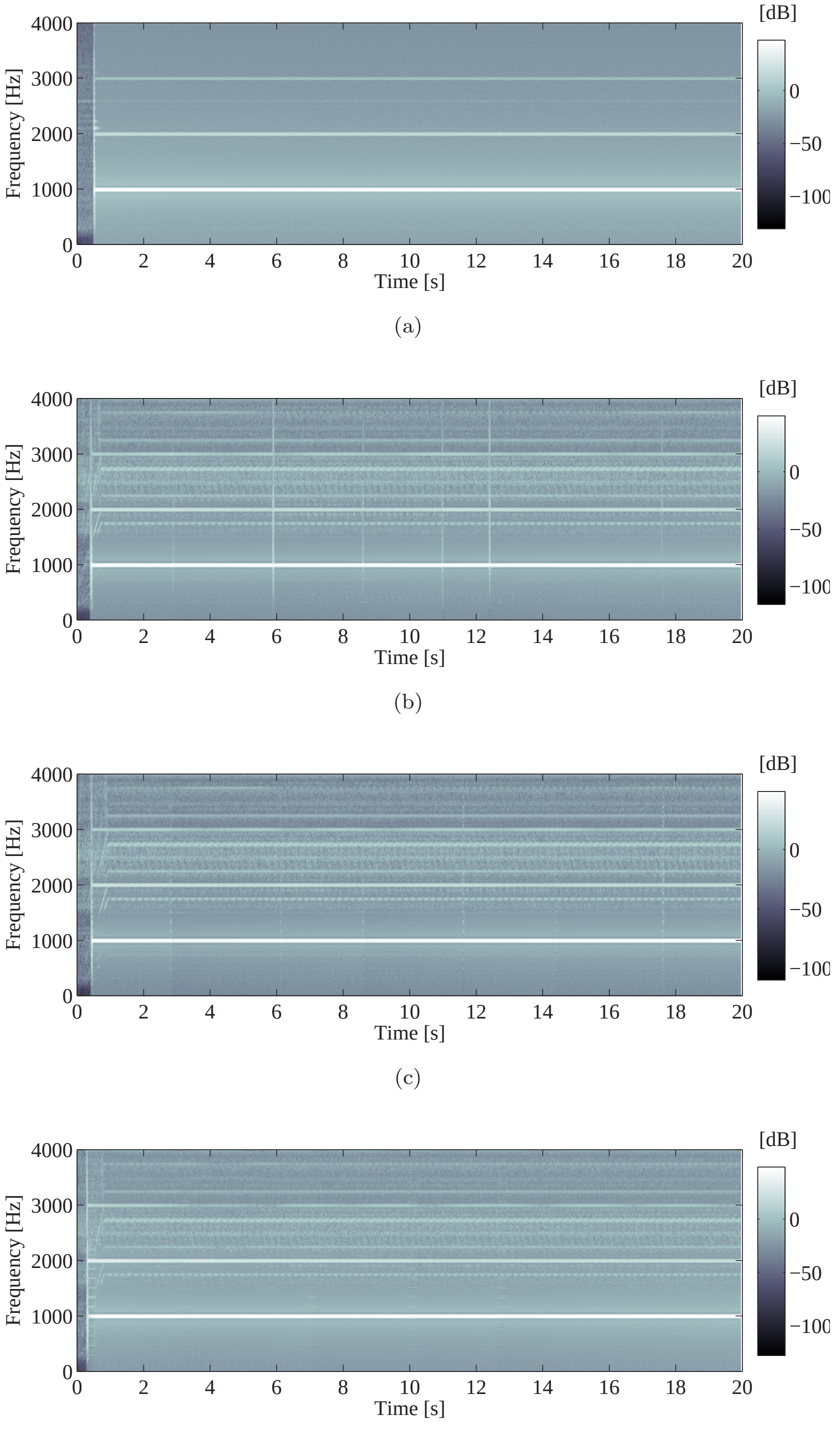
(a)

(b)

(c)

(d)

Figure 8.7: **Measured filter updates with a 1 kHz sine wave as the source signal**. *The loudspeaker was moved in the lateral plane in front of the KEMAR from on-axis to 0.2 m off-axis to the right at 1 cm/s. (a) Reference signal (1 kHz sine) played through the VSI system and measured with the KEMAR dummy head. (b) 2 partitions and 2 steps of commutation. (c) 8 partitions and 8 steps of commutation. (d) 8 partitions and 48 steps of commutation.*



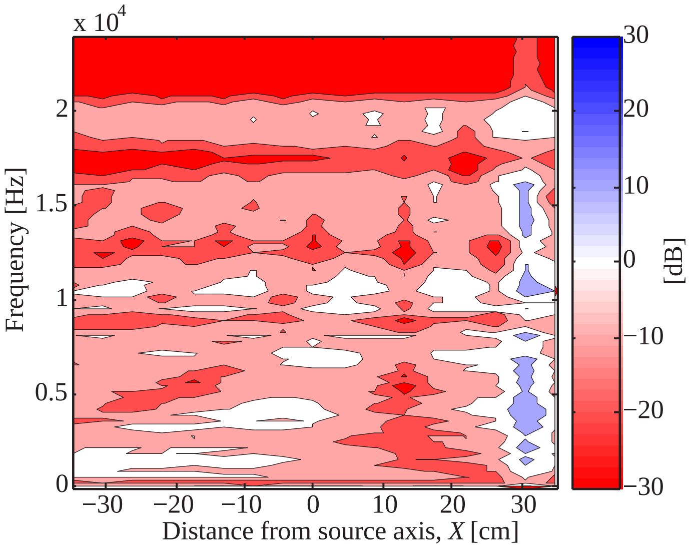
(a)

(b)

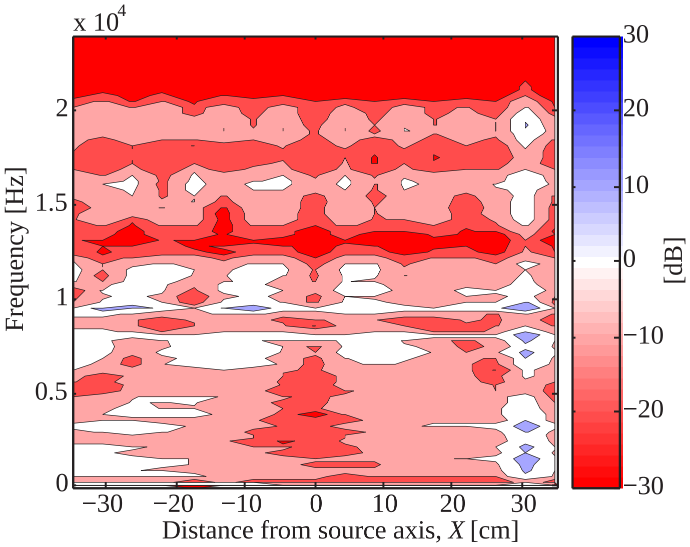
(c)

(d)

FIGURE 8.8: **Time-frequency plot for the measured filter updates using a 1 kHz sine wave as the source signal**. *The loudspeaker was moved in the lateral plane in front of the KEMAR from on-axis to 0.2 m off-axis to the right at 1 cm/s. (a) Reference signal (1 kHz sine) played through the VSI system and measured with the KEMAR dummy head. (b) 2 partitions and 2 steps of commutation. (c) 8 partitions and 8 steps of commutation. (d) 8 partitions and 48 steps of commutation.*

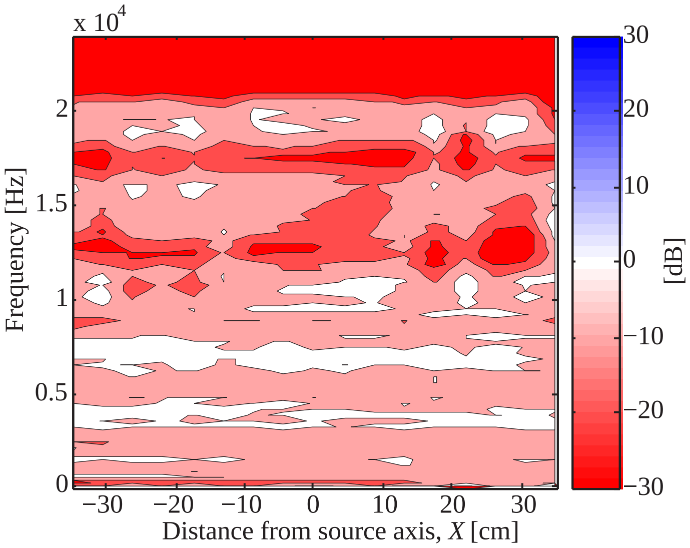


(a)

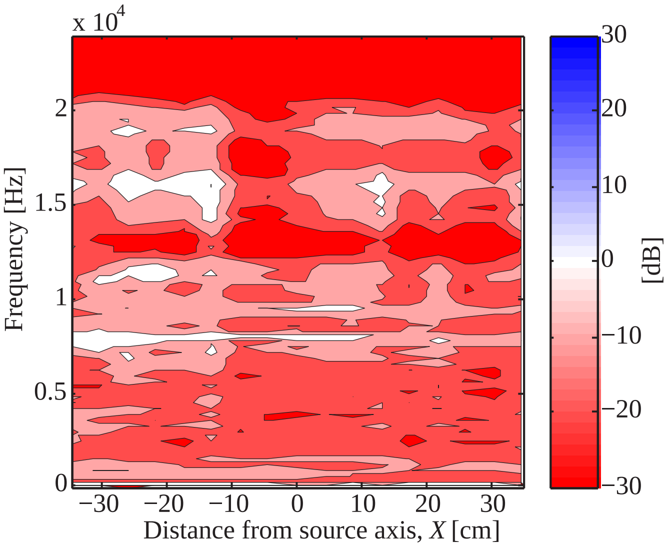


(b)

FIGURE 8.9: **The effectiveness of cross-talk cancellation of the static virtual sound imaging system**. *The SD loudspeaker configuration was used and the KEMAR dummy head was used for measuring the effectiveness of cross-talk cancellation. (a) $P_{11}(k)$. (b) $P_{21}(k)$.*



(a)



(b)

FIGURE 8.10: **The effectiveness of cross-talk cancellation of the adaptive VSI system**. *The SD loudspeaker configuration was used and the KEMAR dummy head was used for measuring the effectiveness of cross-talk cancellation. (a) $P_{11}(k)$. (b) $P_{21}(k)$.*

Chapter 9

Conclusions

It has been demonstrated that visually adaptive virtual sound imaging systems can generate convincing virtual sound images under both anechoic and realistic listening conditions. The problem addressed has been to update the inverse filters without creating any audible changes for a moving listener that does not wear any sensors. The adaptation to listener position has been carried out by using a video camera to track head movements and update the inverse filters accordingly. Image processing algorithms that successfully tracks the listener have been presented together with filter update algorithms and the extent of the "operational area". The filter update algorithms have been evaluated by subjective experiments in both an anechoic environment and in a listening room. The geometry of the system has been evaluated by an analytical evaluation and an objective evaluation. The results show the extent of the operational area under certain performance criteria for different configurations of loudspeaker systems. Cross-over design techniques using a fixed and adaptive approach have been introduced since the optimal cross-over frequencies change with the geometry of the system.

There is a strong relationship between cross-talk cancellation performance for virtual sound imaging systems and the condition number of the matrix of transfer functions that needs to be inverted. The dependency of condition number on frequency has been demonstrated for asymmetric and symmetric listener positions. The "operational area" where the "sweet-spot" can be moved within can be estimated using an analytical model. The size of the operational area using a performance criterion has been shown for the Stereo Dipole (SD) and the 3-way Optimal Source Distribution (OSD). The problem of designing cross-overs has been addressed where the respective advantages between using fixed or adaptive cross-overs have been illustrated and recommendations have been made.

The measurement of a database of Head Related Transfer Functions (HRTFs) based on the KEMAR dummy head has been carried out. The most important features of the HRTFs were illustrated. The database was measured for six different combinations of pinna and ear canal in order to provide extensive future modelling opportunities. The combination of all the six databases will enable the clear identification of the effect of the pinna on the HRTF. Also it is possible that "no pinna" models may be useful in the design of virtual sound imaging systems. However, the choice of database may vary between loudspeaker and headphone based systems.

Interpolation between Head Related Impulse Responses (HRIRs) have been investigated and it has been found that a time domain technique named onset detection by thresholding performs surprisingly well. The average percentage mean square error is less than 0.52% when head shadowing does not occur, in the presence of head shadowing the error goes up to approximately 7.52%. The advantage of this technique is its robustness and inherent simplicity. This interpolation algorithm has proven to be a useful tool for inverse filter design and it can also be applied in filter update algorithms.

Inverse filter design and cross-talk cancellation effectiveness for three different audio systems has been investigated. The filter design process of cross-talk cancellation networks using a minimum number of inverse filters has been demonstrated. The presented filter design technique can reduce the number of filters for Frequency Distributed Loudspeakers (FDL) to the same number as for 2-channel systems. Cross-talk cancellation effectiveness using the fast-deconvolution algorithm has been evaluated for various filter lengths and by varying the listener position/orientation. The "sweet-spot" of static loudspeaker based binaural sound systems is very limited, which have been illustrated by simulations of the control performance under anechoic conditions. The benefit of using an adaptive approach where the cross-talk cancellation filters are updated to the listeners location and orientation have been demonstrated and compared to the static case. The results show that the area within which the listener can move is significantly larger in the adaptive case compared to the static case. The size of the area depends on the degradation of performance that can be accepted compared to the optimal listening position and has been determined by introducing a performance criterion.

The problem of updating inverse filters in real-time have been investigated and two different approaches to the solution have been presented. The filter update rates have been determined by defining two criteria namely, the Just Noticeable Difference (JND) criterion and the Just Noticeable (JNC) criterion. Subjective evaluations from both an anechoic chamber and a listening room show that the JND criterion is approximately 0.025 m for movements in the lateral plane. The JND criterion for movements in the fore and aft plane is greater than the distance of ±0.15 m that was explored. The JNC criterion is approximately 0.012 m for movements in the lateral plane in both the anechoic chamber and in the listening room. These results suggest that the filter update techniques and sound source localisation using binaural loudspeaker based systems works effectively under anechoic as well as under non-anechoic listening conditions.

It has been demonstrated that image processing algorithms can be used to find the position of the listener and initiate filter updates in real-time. Three image processing algorithms for listener tracking have been presented. The first algorithm uses human skin colour characterised by a multivariate Gaussian distribution in a normalised colour space. This can provide a rough estimate of the location of a head within the image frame to initialise a listener head tracking system and to track the translation of the listener in the following frames. The measured standard deviation between the colour tracking algorithm and a reference tracker in the lateral plane is 0.02 m. The second investigated algorithm is stereo area correlation, which can be used to track the distance of the listener. The process of measuring the distance to an object is based on a comparison of the object projection on two images. The standard deviation between the stereo area correlation algorithm and the reference tracker in the fore and aft plane is 0.05 m. However, the standard deviation can be reduced further by applying a suitable smoothing filter. The results suggests that the investigated algorithms can be used together with the presented filter update algorithms, since the JND criterion established is 0.025 m in the lateral plane and > 0.15 m in the fore and aft plane. The third investigated algorithm is a contour tracking algorithm that uses a particle filter named the condensation algorithm. The contour tracking algorithm is robust to background clutter, occlusion, and it tracks with six degrees of freedom. The contour tracking algorithm did not run in real-time the investigated Matlab [66] implementation, though it can be configured to run in real-time on a single processor computer when coded in for example C++ (Isard [41]).

The design and implementation of a visually adaptive binaural sound reproduction system named Virtual Sound Imaging (VSI) has been carried out. The system was designed using a software approach in order make it flexible for future development. The outcome is a prototype that is capable of reproducing binaural signals at the ears of a moving listener that is not wearing any sensors. The system can track the listener in the horizontal plane. The implemented real-time filter update technique produces smooth filter updates that can be seamless for the listener. The filter update technique is implemented directly in the partitioned convolution algorithm that is used for convolving the binaural signals with the inverse filters. The system performance has been objectively evaluated in an anechoic chamber with respect to real-time filter updates and cross-talk cancellation effectiveness. The objective evaluation has been compared to the subjective filter update criterion JNC that was determined in Chapter 6. The results from the objective evaluation shows that in order to avoid transients one need to update the filters at a rate close to the subjective JNC criterion.

Further works include, subjective experiments for proving the benefits of a visually head tracked virtual sound imaging system in different applications such as, mobile phones, video games, virtual reality, home cinema and domestic hifi. It would be interesting to compare the performance to existing virtual imaging system such as the SD and the OSD without head tracking and also to more established systems such as Stereo and multi-channel surround (Dolby digital 5.1 for example).

The concept of using individual pinna responses is likely to significantly improve the system performance. The use of non-individual pinna responses can be a major drawback for binaural sound reproduction systems and results presented here suggests that the frequency response above 7 kHz can become considerably degraded. Promising projects for this are numerical modelling of pinnae and image processing algorithms that can convert an image of a pinna to HRTFs. With the increasing interest in biometric research it can become reality that each person carries a biometric identity where the pinnae shapes are included. The pinnae shapes could then be converted to HRTFs and used to configure the sound system with individualised pinnae responses

The listener tracking algorithm can be approved further by using a fusion architecture, where for example colour tracking is combined with eye tracking and nose tracking. The presented contour tracking algorithm can be extended to track rotation and also multiple listeners for a system that can perform adaptive multi-channel inversion. More specialised sensors than the RGB cameras that were used here could be considered, for example, IR and NIR cameras.

The VSI software can be extended to include the FDL principle together with adaptive cross-over frequencies such that each frequency band is reproduced from an optimal angle relative to the listener. The system can be extended to be able to handle a number of sound sources that can be placed anywhere in the sound space in order to fully demonstrate the capabilities of this technology. An improved head tracking scheme that allows for listener rotation is also desirable to incorporate in the future.

The adaptive system can maintain the sound source localisation capabilities in a larger operational area than a static binaural sound reproduction system. The sweet-spot size of static binaural loudspeaker systems is limited to few centimeters in the lateral plane. This makes it difficult to find promising applications where it is beneficial to apply such technology. When the system is made adaptive to listener position the full benefit of binaural technology can be achieved and there are now plenty of applications that can benefit from it. The most promising applications in the near future are TV, home cinema and video gaming. When there are more binaural recordings available on the market then it will also be beneficial within domestic hifi. In a more far future there might be enough biometric information readily available about the listeners so that one can make use of individualised HRTFs and all the advantages these can expect to give in terms of sound quality and virtual sound imaging performance. The main competing technologies to binaural sound reproduction are wave field synthesis (suitable for the multiple listener problem) and intensity panning techniques. Both of these technologies requires many loudspeakers to reproduce the sound in three dimensions. Where as the binaural approach only requires two loudspeakers in the most basic configuration. This thesis have laid out the theory for adaptive binaural sound reproduction using visual tracking and loudspeakers.

Appendix A

Image processing

This paragraph describes binary edge detection and gives a brief description of the more advanced edge detection method named the Sobel operator. Binary gradient edge detection is a simple form of edge detection and is performed in the following manner. Firstly the intensities along a line x in a grey scale image are illustrated in the histogram in Figure A.1(a) along with the continuous approximation to the histogram. The first derivative this data is illustrated in Figure A.1(b). The peak of this first derivative provides a position along the line for the best fit edge.

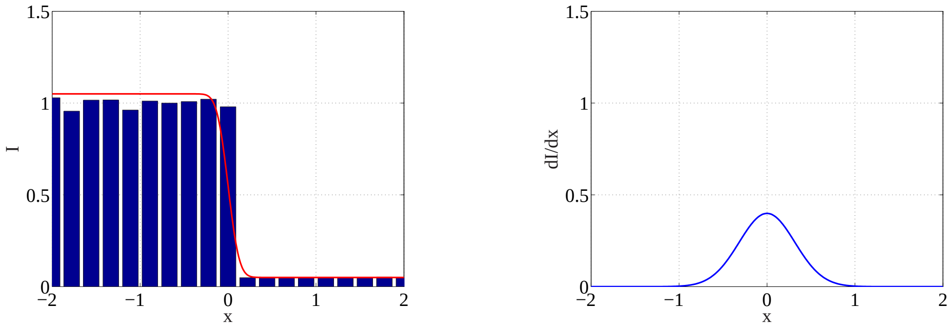
The Sobel operator performs a 2D spatial gradient measurement on an image in order to emphasize regions of high spatial gradient that correspond to edges. It is used in here to find the approximate absolute gradient magnitude at each point in an input greyscale image. The Sobel operator consists of a pair of 3 × 3 convolution masks, which are illustrated in the Figure A.2. These masks are designed to respond maximally to edges running vertically and horizontally relative to the pixel grid, one mask for each of the two perpendicular orientations. The masks can be applied separately to the input image, to produce separate measurements of the gradient component in each orientation ($\mathbf{G}_x^2$ and $\mathbf{G}_y^2$). These can then be combined to find the absolute magnitude of the gradient at each point, and the gradient magnitude is given by Equation A.1.

$$|\mathbf{G}| = \sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2} \tag{A.1}$$

Morphological image processing In morphological image processing dilation and erosion are often used in combination to implement image processing operations. For example, the definition of a "morphological opening" of an image is erosion followed by dilation, using the same structuring element for both operations. The related operation, "morphological closing" of an image is the reverse: it consists of dilation followed by erosion with the same structuring element. One can use morphological opening to remove small objects from an image while preserving the shape and size of larger objects in the image. For example, one can use the Matlab [64] function **imopen** to remove all spurioses from the original image, creating an output image that contains only the larger shapes, such as heads.

An essential part of the dilation and erosion operations is the structuring element used to probe the input image. Two-dimensional structuring elements consist of a matrix of zeros and ones, typically much smaller than the image being processed. The centre pixel of the structuring element identifies the pixel of interest, i.e. the pixel being processed. The pixels in the structuring element containing ones define the neighbourhood of the structuring element. These pixels are also considered in the dilation or erosion processsing. One can typically choose a structuring element the same size and shape as the desired objects one wants to process in the input image. The structuring element should be large enough to remove the lines when one erodes the image, but not large enough to remove the desired shapes. It should consist of all ones, so it removes everything but large continuous patches of foreground pixels. In the morphological dilation and erosion operations, the state of any given pixel in the output image is determined by applying a rule to the corresponding pixel and its neighbours in the input image. The rule used to process the pixels defines the operation as dilation or erosion.

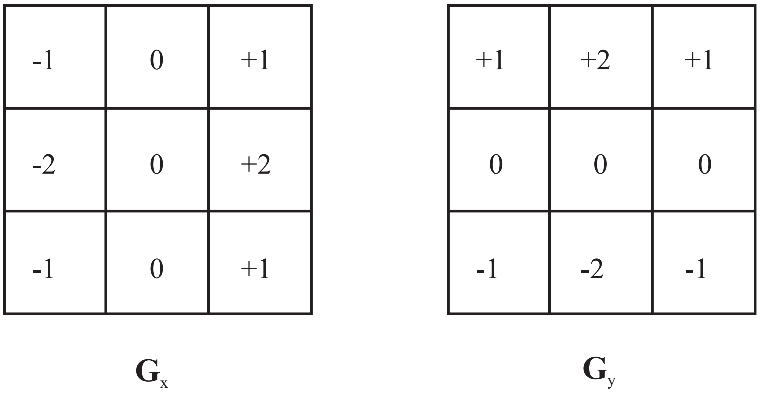
Dilation can be described as: the value of the output pixel is the maximum value of all the pixels in the input pixels neighbourhood. In a binary image, if any of the pixels is set to the value one, the output pixel is set to one. Erosion can be described as: the value of the output pixel is the minimum value of all the pixels in the input pixels neighbourhood. In a binary image, if any of the pixels is set to 0, the output pixel is set to 0. The erosion removes the spurioses in the image, but also shrinks the desired shapes. To restore the desired shapes to their original size, the eroded image can be dilated using the same structuring element.



(a)

(b)

FIGURE A.1: **The principle of gradient edge detection**. *(a) Intensity histogram. (b) Gradient edge detection.*



-1	0	+1
-2	0	+2
-1	0	+1

$\mathbf{G}_x$

+1	+2	+1
0	0	0
-1	-2	-1

$\mathbf{G}_y$

FIGURE A.2: **Edge detection using Sobel convolution masks**.

Appendix B

Stereo vision

Disparity The disparity measure can be obtained by acquiring two images of the same scene from different viewing positions. An estimation of the relative position of a feature as it appears in the two images makes it possible two calculate the distance of the feature, where the distance is between the baseline of the imaging devices and the feature.

Figure B.1 illustrates the geometry of the stereo vision system in the horizontal plane. The principle is two cameras acquire two images of the same object from different viewpoints. The two viewpoints are noted as A and B and are the positions of the lenses of the two cameras. The distance between the viewpoints is the camera separation, and is noted as d. The point $P(X, Y, Z)$ in the real world where the coordinate system has an origin O, at the left camera (at point A). The X-direction is along the line between the two lenses and the Y-direction is at right angles to X and parallel with the image planes. The focal length of the lenses is denoted by f. The distance from the image centre to the object are noted as x_l and x_r and the image centres are denoted by O_l and O_r. The positions of P in the left and the right images are denoted by x_l and x_r with respect to the local coordinate systems for each image (O_l and O_r).

The disparity measure is calculated as follows. Firstly, using similar triangles in the X, Z-plane and in the X, Y-plane shows that a line from P through the centre of the left camera lens intersects the $Z = -f$ image plane. Similarly for the right camera, it follows that

$$X_l = -X_o \frac{f}{Z_o} \qquad Y_l = -Y_o \frac{f}{Z_o} \tag{B.1}$$

$$X_r = -(X_o + d) \frac{f}{Z_o} \qquad Y_r = -Y_o \frac{f}{Z_o} = -Y_o \frac{f}{Z_o} \tag{B.2}$$

Secondly, a two dimensional coordinate system is now defined in each image plane and rotated 180° from the main coordinate system, in order to counteract the rotation that is intrinsic to the imaging process. Thus

$$x_l = -X_l \qquad y_l = -Y_l \qquad x_r = -X_r \qquad y_r = -Y_r \tag{B.3}$$

Now the coordinates of the point images are given by the equations below.

$$x_l = X_o \frac{f}{Z_o} \qquad y_l = Y_o \frac{f}{Z_o} \tag{B.4}$$

$$x_r = (X_o + d) \frac{f}{Z_o} \qquad y_r = Y_o \frac{f}{Z_o} \tag{B.5}$$

Rearranging gives the following equation. It is also worth noticing that the y-coordinate of the point is the same in both images.

$$X_o = x_l \frac{f}{Z_o} = x_r \frac{f}{Z_o} - d \tag{B.6}$$

Solving this for the distance Z_o produces the following equation.

$$Z_o = \frac{fd}{x_r - x_l} \tag{B.7}$$

This relates the Z_o component to the amount of pixel shift between the two images. Finally the disparity measure is given by Equation B.8.

$$\Delta x = x_r - x_l = \frac{fd}{Z_0} \tag{B.8}$$

The disparity measure is inversely proportional to the distance of the point P and it is directly proportional to the focal length f and the camera separation d.

The distance calculation assumes that the cameras are perfectly aligned, with parallel image planes although in practice this is often not the case. The disparity measure returned by the stereo area correlation algorithm will be offset from the ideal disparity by some amount X_{off}. The offset can be adjusted so that it is possible to capture the active areas of interest in the image.

The disparity measure can be used to find pixels that correspond in the two images. The disparity measure is here specified in units of 1/16 pixel. Then the X coordinates are related by

$$x_r^1 = x_l^1 + 16\Delta x \tag{B.9}$$

The equation above assumes that there is no X_{off} between the images and that the calibration was specified as having zero disparity at infinity. If there is an offset X_{off} then the equation becomes

$$x_r^1 = x_l^1 + 16\Delta x - X_{off} \tag{B.10}$$

Horopter For stereo algorithms it is normal to search only a window of disparities, for example 16 or 32 disparities. Hence the distance of objects that can be successfully determined is restricted to a limited interval. The "horopter" is introduced for describing this search window, and can be described as the 3D volume that is covered by the search range of the stereo algorithm. The horopter is illustrated in Figure B.2. The "horopter" depends on the following parameters: camera parameters, stereo camera separation, the disparity search range, and X_{off}.

The stereo algorithm searches a pixel range of disparities in order to find a match. An object that has a valid match must lie in the region between the two planes illustrated in Figure B.2 and the nearest plane has the highest disparity (15) and the furthest plane has the lowest disparity (0).

The placement of the "horopter" can be varied by changing the X_{off} offset between the two images. This will change the disparity search window for a stereo match, which is depicted in Figure B.3. The cameras are slightly converged which results in that the zero disparity plane occurs at a finite distance in front of the cameras. The top arrow represents a search with 5 disparities without any offset X, hence the search is from disparity 0 to disparity 4. The lower arrow represents the offsetting X of one image by n pixels, the "horopter" can be changed to go from -n to 5-n disparities. The X_{off} offset can be used to compensate for camera convergence or divergence so that the furthest "horopter" plane is at infinity, which is generally desirable. The reason for setting the furthest "horopter" plane at infinity is because it is usually possible to control how close objects get to the camera, but not how far away. The offset that put the far "horopter" plane at infinity is called X_{inf}. The disparity with offset X_{inf} is 0 and indicates an infinitely far object.

The location and size of the "horopter" depends on the application. The objects that fall outside the "horopter" will result in a random distribution of disparities. In the left image the disparity search window is correctly positioned so that all the objects in the image fall inside the "horopter". The right image, the disparity search window has moved back so that objects have higher disparities. The closer objects are now outside the "horopter" and their disparity image becomes a random pattern. It is thus important for the “horopter” to be large enough to encompass the distance of objects in the application. In the general case, the upper end of the “horopter” is positioned at infinity and makes the search window large enough to see the closest objects. The “horopter” can be adjusted by the following parameters:

(1) Search window. (2) Offset X. (3) Baseline. (4) Focal length. (5) Pixel width.

The “horopter” can be enlarged by increasing the search window. The offset X_{off} can be adjusted so that the “horopter” can capture the active areas in the image. When the cameras are moved closer together their viewpoints also become closer, and image differences and disparity are lessened. This results in a larger “horopter”. By decreasing the focal length of the cameras the image geometry changes, which results in smaller perceived sizes. This also results in a larger “horopter”. Another way to make the “horopter” larger is to make the pixels wider. The three last methods (change of camera separation, focal length and pixel width) all changes the camera geometry, and thus have a corresponding effect on the distance resolution, which decreases. The only way to increase the “horopter” size and maintain distance resolution is to increase the size of the disparity search window. This will lead to an increase in computation. Multi-resolution methods can be used to minimize the computation (Konolige [54]).

The distance resolution is important to know, in order to determine the minimum change that the stereo engine can detect. The distance resolution is a function of the distance, and is given by the following equation

$$\Delta Z = \frac{Z_o^2}{df}(\Delta x) \tag{B.11}$$

At close distances, the resolution is better than at further distances. The distance resolution is the smallest change in distance that can be detected given a change in disparity. It is obvious from the equation above that the distance resolution goes up (gets worse) with the square of the distance. The camera separation and the focal length have an inverse influence on the resolution, hence larger baselines and focal lengths (telephoto for example) increases distance resolution (makes it better). The pixel size also has an influence, and smaller pixel size makes the distance resolution better.

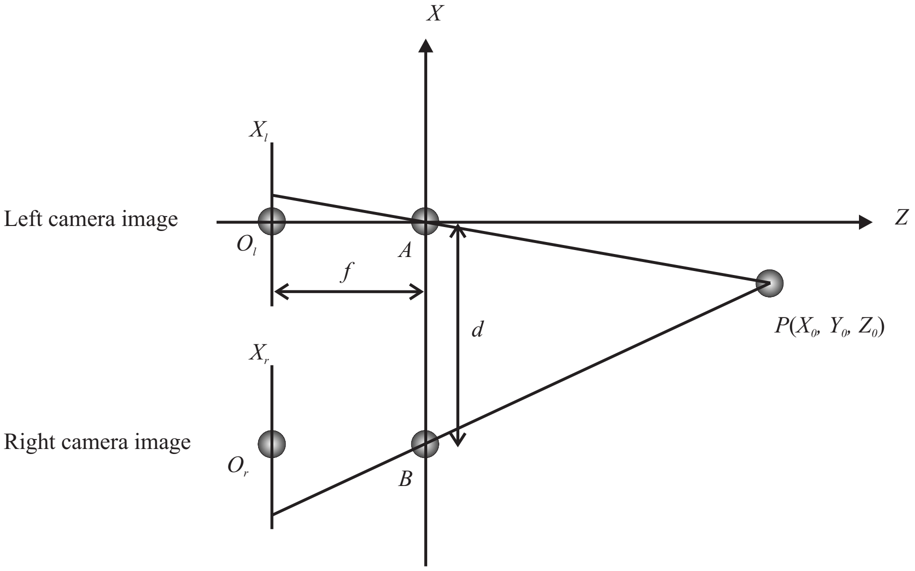


FIGURE B.1: **The geometry of the stereo vision system viewed from above**.

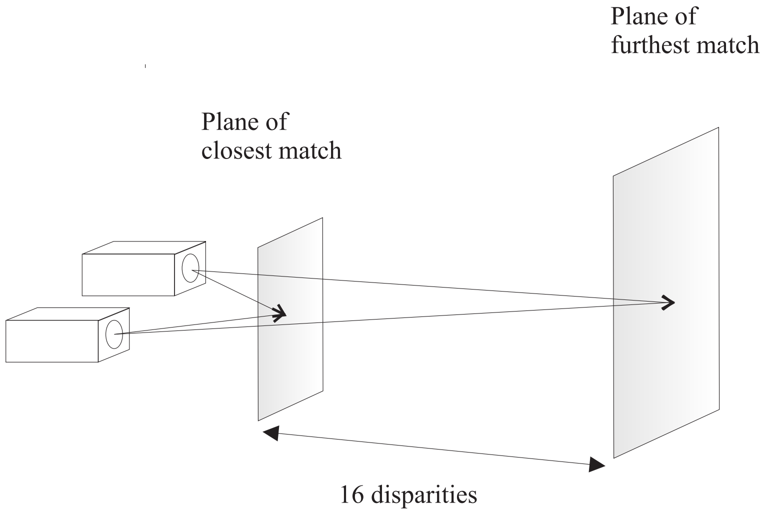


FIGURE B.2: **Horopter planes for disparity search**.

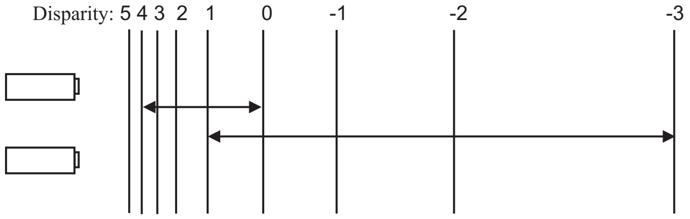


FIGURE B.3: **Planes of constant disparity for a converged stereo camera**. *A search range of 5 pixels is shown to cover two different horopters, that depend on the search offset X.*

Appendix C

Condensation

The theory of the condensation algorithm is briefly described here. It is assumed that there are T frames of data to be processed and that at time t only data from times $1, t-1$ are available. The measurements acquired from frame t are denoted $\mathbf{Z}_t$ and will consist of a list of edges observed in the image, as discussed in Section 7.4.3. The state of the modelled object at time t is $\mathbf{Y}_t$. The measurements acquired up to frame t are denoted $\mathcal{Z}_t = \{\mathbf{Z}_1, ...\mathbf{Z}_t\}$ and the state of the modelled object up to frame t is denoted $\mathcal{Y}_t = \{\mathbf{Y}_1, ...\mathbf{Y}_t\}$.

Dynamics The assumption that the object dynamics form a temporal Markov chain is made for the probabilistic framework, so that

$$p(\mathbf{Y}_t|\mathcal{Y}_{t-1}) = p(\mathbf{Y}_t|\mathbf{Y}_{t-1}) \tag{C.1}$$

The new state is dependent only on the immediately preceding state and is independent of earlier states. The dynamics are now entirely determined by the form of the conditional density $p(\mathbf{Y}_t|\mathbf{Y}_{t-1})$.

Observations The observations $\mathbf{Z}_t$ are assumed to be mutually independent and also independent with respect to the dynamical process. The observation process is therefore defined by specifying the observation density $p(\mathbf{Z}_t|\mathbf{Y}_t)$ at each time t.

Propagation The conditional state-density at time t is defined by $p(\mathbf{Y}_t|\mathcal{Z}_t)$ (assuming a continuous valued Markov-chain with independent observations). This represents all information about the state at time t that can be deduced from the entire data stream up to time t. The equation for propagation of state density over time is given by

$$p(\mathbf{Y}_t|\mathcal{Z}_t) = k_t p(\mathbf{Z}_t|\mathbf{Y}_t)p(\mathbf{Y}_t|\mathcal{Z}_{t-1}) \tag{C.2}$$

where

$$p(\mathbf{Y}_t|\mathcal{Z}_{t-1}) = \int_{\mathbf{Y}_{t-1}} k_t p(\mathbf{Y}_t|\mathbf{Y}_{t-1})p(\mathbf{Y}_{t-1}|\mathcal{Z}_{t-1})d\mathbf{Y}_{t-1} \tag{C.3}$$

and k_t is a normalisation constant that does not depend on $\mathbf{Y}_t$ (this is proved by Isard [41]). The propagation of state density can be interpreted as Bayes' theorem for inferring posterior state density from data, for the time varying case. The prior $p(\mathbf{Y}_t|\mathcal{Z}_{t-1})$ is a prediction taken from the posterior $p(\mathbf{Y}_{t-1}|\mathcal{Z}_{t-1})$ from the previous time step by superimposing one time-step from the dynamical model. The multiplication by the observation density $p(\mathbf{Z}_t|\mathbf{Y}_t)$ then applies the reactive effect expected from observations. Equation C.2 is the filter that the condensation algorithm must approximate.

Factored sampling This paragraph describes the factored sampling algorithm that deals with non-Gaussian observations in single images, which will be extended in the following paragraph to deal with temporal image sequences.

Given a static image $\mathbf{Z}$ where the problem is to find an object parameterised as $\mathbf{Y}$ with prior $p(\mathbf{Y})$. The posterior density $p(\mathbf{Y}|\mathbf{Z})$ represents all the knowledge about $\mathbf{Y}$ that is deducible from the data. This can be evaluated by applying Bayes' theorem:

$$p(\mathbf{Y}|\mathbf{Z}) = kp(\mathbf{Z}|\mathbf{Y})p(\mathbf{Y}) \tag{C.4}$$

Where k is a normalisation constant that is independent of $\mathbf{Y}$. When $p(\mathbf{Z}|\mathbf{Y})$ is sufficiently complex that $p(\mathbf{Y}|\mathbf{Z})$ cannot be evaluated in closed form, iterative sampling techniques can be used (MacCormick [62]). The factored sampling algorithm generates a random variate $\mathbf{Y}'$ from the approximation of the posterior distribution denoted $\tilde{p}(\mathbf{Y})$. First a sample set $\{\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(N)}\}$ is generated from the prior density $p(\mathbf{Y})$ and then each index $i \in \{1, \ldots, N\}$ is assigned with probability π_i , where

$$\pi_i = \frac{p_z(\mathbf{s}^{(i)})}{\sum_{j=1}^N p_z(\mathbf{s}^{(j)})} \tag{C.5}$$

$$p_z(\mathbf{Y}) = p(\mathbf{Z}|\mathbf{Y}) \tag{C.6}$$

The index i determines the value $\mathbf{Y}' = \mathbf{Y}_i$ and by choosing $\mathbf{Y}'$ in this way gives it a distribution that approximates the posterior $p(\mathbf{Y}|\mathbf{Z})$ with increasing accuracy as N increases. Moments of $\mathbf{Y}$ can be estimated using Equation C.7.

$$E[g(\mathbf{Y})|\mathbf{Z}] = \frac{\sum_{n=1}^{N} g(\mathbf{s}^{(n)}) p_z(\mathbf{s}^{(n)})}{\sum_{n=1}^{N} p_z(\mathbf{s}^{(n)})} \qquad \text{(C.7)}$$

The mean can be estimated by setting $g(\mathbf{Y}) = \mathbf{Y}$.

Summary of the algorithm The condensation algorithm uses factored sampling extended to apply iteratively to image sequences. The process at each time step is an iteration of factored sampling; and the output of an iteration will be a weighted, time-stamped sample set, denoted $\{\mathbf{s}_t^{(n)} \;,\; n = 1, \ldots, N\}$ with weights $\pi_t^{(n)}$, representing approximately the conditional state density $p(\mathbf{Y}_t|\mathcal{Z}_t)$ at time t. The process of obtaining the sample set is described as follows: the process begins with an effective prior density for time step t denoted $p(\mathbf{Y}_t|\mathcal{Z}_{t-1})$ (which is multi-modal in general and no functional representation is available), this prior is derived from the sample set representation $\{\mathbf{s}_{t-1}^{(n)}, \pi_{t-1}^{(n)} \;,\; n = 1, \ldots, N\}$ of $p(\mathbf{Y}_{t-1}|\mathcal{Z}_{t-1})$ (the output from the previous time step), to which the prediction must be applied.

The iterative process applied to sample sets is described in Figure C.1. At the top of the diagram, the output from time-step $t-1$ is the weighted sample-set $\{\mathbf{s}_{t-1}^{(n)}, \pi_{t-1}^{(n)} \;,\; n = 1, \ldots, N\}$. The algorithm must have sample sets of fixed size N, so that it can be guaranteed to run within a given computational resource. Therefore the first operation is to sample with replacement N times from the sample set $\mathbf{s}_{t-1}^{(n)}$, choosing a given element with probability $\pi_{t-1}^{(n)}$. The elements with high weights may be chosen several times, leading to identical copies of elements in the n^{th} set. The elements with relatively low weights may not be chosen at all. Each element chosen from the new set is now subjected to the predictive step. The motion model causes identical elements to split and diffuse through state-space. Now the sample set $\mathbf{s}_t^{(n)}$ for the new time step has been generated but without its weights. It is approximately a random sample from the effective prior density $p(\mathbf{Y}_t|\mathcal{Z}_{t-1})$ for time step t. Finally the observation step from factored sampling is performed, generating weights from the observation density $p(\mathbf{Z}_t|\mathbf{Y}_t)$, in order to obtain the sample set representation $\{\mathbf{s}_t^{(n)}, \pi_t^{(n)}\}$ of the state density for time t.

The condensation algorithm makes use of cumulative probabilities $c_t^{(n)}$ calculated from $\pi_t^{(n)}$:

$$\begin{aligned} c_t^{(n)} &= 0 \\ c_t^{(n)} &= c_t^{(n-1)} + \pi_t^{(n)} \qquad (n = 1, \ldots, N) \end{aligned} \qquad \text{(C.8)}$$

In step (1), a base sample $\mathbf{s}'^{(n)}_t = \mathbf{s}_{t-1}^{(j)}$ is chosen from the sample set $\{\mathbf{s}_{t-1}^{(n)}, \pi_{t-1}^{(n)}, c_{t-1}^{(n)}\}$ with probability $\pi_{t-1}^{(j)}$. This can be done by first generating a random number $r \in [0, 1]$, uniformly distributed, then find by binary subdivision the smallest j for which $c_{t-1}^{(j)} \geq r$ and finally set $\mathbf{s}'^{(n)}_t = \mathbf{s}_{t-1}^{(j)}$.

After any time-step in the condensation algorithm, it is possible to report the current state, for example by evaluating moments of the state density.

The following provides a summary of the algorithm:

From the old sample set $\{\mathbf{s}_{t-1}^{(n)}, \pi_{t-1}^{(n)}, c_{t-1}^{(n)} \quad , \quad n = 1, \ldots, N\}$ at time step $t-1$, construct a new sample set $\{\mathbf{s}_{t}^{(n)}, \pi_{t}^{(n)}, c_{t}^{(n)} \quad , \quad n = 1, \ldots, N\}$ and keep on iterating. Construct the n^{th} of new samples as follows:

(1) Select a sample $\mathbf{s'}_t^{(n)} = \mathbf{s}_{t-1}^{(j)}$ with probability $\pi_{t-1}^{(j)}$ as discussed above.

(2) Predict by sampling from $p(\mathbf{Y}_t|\mathbf{Y}_{t-1} = \mathbf{s'}_t^{(n)})$ to choose each $\mathbf{s}_t^{(n)}$. The new sample value is generated by applying the dynamics in equation 7.22.

(3) Measure and weight the new position in terms of the measured features $\mathbf{Z}_t$ by using $\pi_t^{(n)} = p(\mathbf{Z}_t|\mathbf{Y}_t = \mathbf{s}_t^{(n)})$, then normalise so that $\sum_n \pi_t^{(n)} = 1$ and store together with cumulative probability as $\{\mathbf{s}_t^{(n)}, \pi_t^{(n)}, c_t^{(n)}\}$

When the N samples have been constructed, one can then estimate moments of the tracked position at time-step t using the Minimum Mean Square Estimator (MMSE) as

$$E[f(\mathbf{Y}_t)] = \sum\nolimits_{n=1}^{N} \pi_t^{(n)} f(\mathbf{s}_t^{(n)}) \tag{C.9}$$

Where the mean position is obtained using $f(\mathbf{Y}) = \mathbf{Y}$.

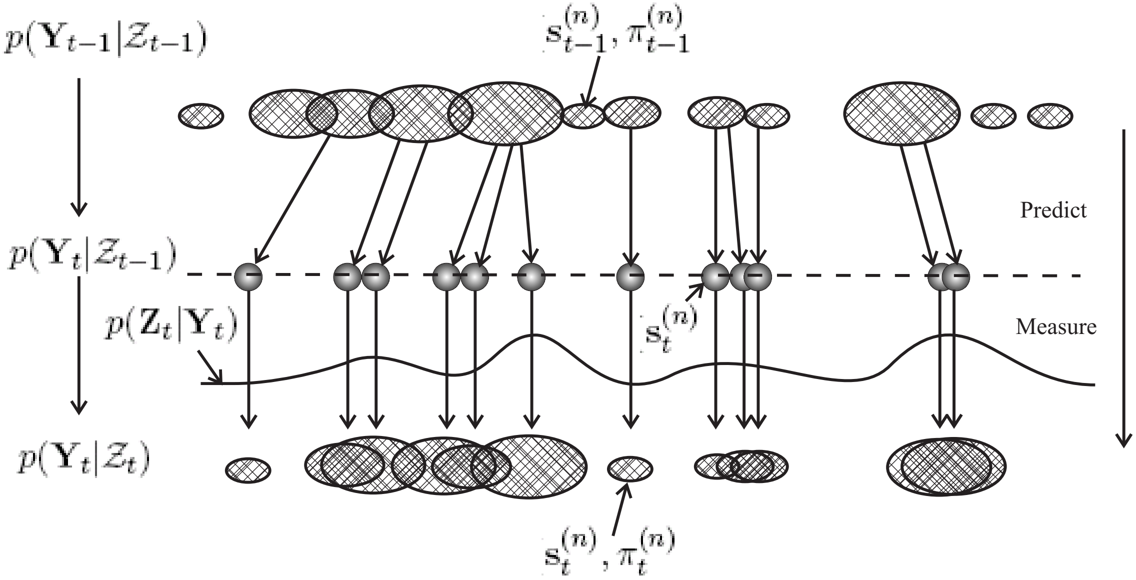

$p(\mathbf{Y}_{t-1}|\mathcal{Z}_{t-1})$
$\mathbf{s}_{t-1}^{(n)}, \pi_{t-1}^{(n)}$
Predict
$p(\mathbf{Y}_t|\mathcal{Z}_{t-1})$
$p(\mathbf{Z}_t|\mathbf{Y}_t)$
$\mathbf{s}_t^{(n)}$
Measure
$p(\mathbf{Y}_t|\mathcal{Z}_t)$
$\mathbf{s}_t^{(n)}, \pi_t^{(n)}$

FIGURE C.1: **One time step in the "condensation algorithm" (Blake [12])**. *The three steps, drift, diffuse and measure describes the probabilistic propagation process of the condensation algorithm. The drift and diffusion steps are illustrated as one combined step, i.e. the prediction step.*

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* Dover, 1965.

[2] V. Algazi, R. Duda, and D. Thompson. The cipic hrtf database. Technical report, CIPIC, 2001.

[3] V. Algazi, A. Ralph, C. Avendano, and R. Duda. Estimation of a spherical-head model from anthropometry. *Journal of the Audio Engineering Society*, 49:472–479, 2001.

[4] B. S. Atal and M. R. Schroeder. Apparent sound source translater. *US Patent*, 3,236,949, 1966.

[5] N. Ayache and P. T. Sander. *Artificial Vision for Mobile Robots: Stereo Vision and Multisensory Perception.* MIT Press, 1991.

[6] A. Azarbayejani, T. Darrell, and A. Pentland. Real-time tracking of the human body. *Institute of Electrical and Electronics Engineers, Transactions on pattern analysis and machine intelligence*, 19:780–785, 1997.

[7] A. Azarbayejani and A. Pentland. Real time self-calibrating stereo person tracking using 3-d shape estimation from blob features. *ICPR'96 Vienna*, 1996.

[8] D. H. Ballard and C. M. Brown. *Computer Vision.* Prentice Hall, 1982.

[9] J. L. Bauck. A simple loudspeaker array and associated cross-talk canceller for improved 3d audio. *Journal of the Audio Engineering Society*, 49:3–13, 2001.

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

[11] D. R. Begault. *3-D Sound for Virtual Reality and Multimedia.* Morgan Kaufmann Pub, 1994.

[12] A. Blake and M. Isard. *Active Contours.* Springer, 1998.

[13] J. Blauert. *Spatial Hearing.* The MIT Press, 2001.

[14] M. A. Blommer and G. H. Wakefield. Pole-zero approximations for head-related transfer functions using a logarithmic error criterion. *Institute of Electrical and Electronics Engineers Transactions Speech Audio Processing*, 5:278–287, 1997.

[15] R. Bowden. *Learning Non-linear Models of Shape and Motion.* Phd, Brunel University, 1999.

[16] M. D. Burkard. Manikin measurements. Technical report, Knowles, 1978.

[17] K. R. Catleman. *Digital Image Processing.* Prentice Hall, 1996.

[18] J. Chen, B Van Veen, and K. Hecox. A spatial feature extraction and regularization model for the head-related transfer function extraction. *Journal of the Acoustical Society of America*, 97, 1995.

[19] D. H. Cooper and J. L. Bauck. Prospects for transaural recording. *Journal of the Audio Engineering Society*, 37:3–19, 1989.

[20] D. H. Cooper and J. L. Bauck. Generalised transaural stereo and applications. *Journal of the Audio Engineering Society*, 44:683–705, 1992.

[21] D. H. Cooper and J. L. Bauck. Head diffraction compensated stereo system with loudspeaker array. *US Patent*, 5,333,200, 1994.

[22] T. Cootes, G. J. Page, C. B. Jackson, and C. Taylor. Statistical grey-level models for object location and identification. *Proceedings of British Machine Vision Conference*, pages 533–542, 1995.

[23] R. Crochiere. *Multirate Digital Processing.* Prentice Hall, 1983.

[24] P. Damaske and V. Mellert. Sound reproduction of the upper semi-space with directional fidelity using two loudspeakers. *Acoustica*, 22:153–162, 1969.

[25] A. Doucet, N. Freitas, and N Gordon. *Sequential Monte Carlo Methods in Practice.* Springer, 2001.

[26] R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, 1973.

[27] E. A. Durant and G. H. Wakefield. Efficient model fitting using a genetic algorithm: Pole-zero approximations of hrtfs. *Institute of Electrical and Electronics Engineers Transactions Speech Audio Processing*, 1:18–27, 2002.

[28] M. Evans, J. Angus, and A. Tew. Analyzing head-related transfer function measurements using surface spherical harmonics. *Journal of the Acoustical Society of America*, 104, 1998.

[29] A. Farina, G. Glasgal, E. Armelloni, and A. Torgers. Ambiophonic principles for the recording and reproduction of surround sound for music. *Audio Engineering Society 19th International Conference*, 2001.

[30] W. G. Gardner. *3-D audio using loudspeakers.* Phd thesis, MIT Media Laboratory, Cambridge, MA, 1997.

[31] W. G. Gardner and K. Martin. Hrtf measurement of a kemar dummy-head microphone. Technical report, MIT Media Lab, 1994.

[32] G. H. Golub and C. F. Van Loan. *Matrix computations.* The John Hoopkins University Press, Baltimore, MD, 1996.

[33] W. Grimson. *Object recognition by computer.* MIT press, 1990.

[34] S. R. Gunn. *Dual Active Contour Models for Image Feature Extraction.* Phd, University of Southampton, 1996.

[35] S. R. Gunn and M. S. Nixon. Snake head boundary extraction using local and global energy minimisation. *Institute of Electrical and Electronics Engineers, Proceedings of the International Conference on Pattern Recognition ICPR*, B:581–585, 1996.

[36] E. R. Hafter and C Trahiotis. Functions of the binaural system. *Encyclopaedia of Acoustic*, 3:1461–1479, 1997.

[37] H. Hamada, N. Ikeshoji, Y. Ogura, and T. Miura. Relation between physical characteristics of orthostereophonc system and horizontal plane localization. *Journal of the Acoustical Society of Japan*, 6:143–154, 1996.

[38] J. Huopaniemi. *VIRTUAL ACOUSTIC AND 3-D SOUND IN MULTIMEDIA SIGNAL PROCESSING.* PhD thesis, Helsinki University of Technology, 1999.

[39] J. Huopaniemi and M. Karjalainen. Comparison of digital filter design methods for 3-d sound. *Institute of Electrical and Electronics Engineers, Nordic Signal Processing Symposium (NORSIG96)*, Sep, 1996.

[40] IEEE. *Programs for Digital Signal Processing.* IEEE Press New York, 1979.

[41] M. Isard. *Visual Motion Analysis by Probablistic Propagation of Conditional Density.* Phd, Oxford University, 1998.

[42] M. Jones and J. Rehg. Statistical color models with application to skin detection. Technical report, Cambridge Research Laboratory Compaq Computer Corporation One Cambridge Center Massachusetts, 2002.

[43] J-M. Jot. Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces. *Multimedia Systems*, 7, 1999.

[44] J-M. Jot, V. Larcher, and O. Warusfel. Digital signal processing issues in the context of binaural and transaural stereophony. *Audio Engineering Society, Convention*, 98:3980, 1995.

[45] M. Karjalainen, Piirila E., and Jarvinen A. Loudspeaker response equalisation using warped digital filters. *Proceedings of NorSig-96*, pages 367–370, 1996.

[46] M. Kass, A. Witkin, and Terzopoulos P. Snakes: Active contour models. *In Proceedings 1st International Conference On Computer Vision*, pages 259–268, 1987.

[47] O. Kirkeby and P. A. Nelson. Local sound field reproduction using two closely spaced loudspeakers. *Journal of the Acoustical Society America*, 104(3):1973–1981, 1998.

[48] O. Kirkeby and P. A. Nelson. Digital filter design for inversion problems in sound reproduction. *Journal of the Audio Engineering Society*, 47, 1999.

[49] O. Kirkeby, P. A. Nelson, and H. Hamada. Stereo dipole - a virtual source imaging system using two closely spaced loudspeakers. *Journal of the Audio Engineering Society*, 46(5):387–395, 1998.

[50] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante. Fast deconvolution of multi-channel systems using regularisation. *Proceedings - International Conference on Noise Control Engineering*, (6):2829, 1996.

[51] O. Kirkeby, P. A. Nelson, F. Orduna-Bustamante, and H. Hamada. Local sound field reproduction using digital signal processing. *Journal of the Acoustical Society of America*, 100(3):1584, 1996.

[52] R. G. Klumpp and H. R. Eady. Some measurements of interaural time difference thresholds. *Journal of Acoustic Society of America*, 28:859–860, 1956.

[53] K. Konolige. Small vision systems: hardware and implementation. *In Eighth International Symposium on Robotics Research*, pages 111–116, 1997.

[54] K. Konolige and D. Beymer. Sri small vision system, users manual. Technical report, SRI, 2004.

[55] P. Kuchi, P. Gabbur, P. Subbana, and S. David. Human face detection and tracking using skin color modeling and connected component operators. *Journal of the Institution of Electronics and Telecommunication Engineers*, 48:289–293, 2002.

[56] C. Kyriakakis, T. Holman, J Lim, H. Hong, and H. Neven. Signal processing, acoustics, and psychoacoustics for high quality desktop audio. *Journal of Visual Communication and Image Representation*, 9:51–61, 1998.

[57] T. I. Laakso, V. Valimaki, M. Karjalainen, and Laine U. K. Splitting the unit delay. tools for fractional delay filter design. *Institute of Electrical and Electronics Engineers Signal Processing Magazine*, 13:30–60, 1996.

[58] T. Lentz and G. Behler. Dynamic cross-talk cancellation for binaural synthesis in virtual reality environments. *Audio Engineering Society 117th Convention*, 2004.

[59] T. Lentz and O. Schmitz. Realisation of an adaptive cross-calk cancellation system for a moving listener. *Audio Engineering Society 21st Conference*, 2002.

[60] T. Lentz, D. Schroder, M. Vorlander, and Assenmacher I. Virtual reality system with integrated sound field simulation and reproduction. *EURASIP Journal on Advances in Signal Processing*, 2007.

[61] R. Y. Litovskya, S. H. Colburn, W. A. Yost, and S. J. Guzman. The precedence effect. *Journal of the Acoustical Society of America*, 104(4), 1999.

[62] J. MacCormick. *Stochastic Algorithms for Visual Tracking.* Springer, 2002.

[63] MathWorks. Signal processing toolbox user's guide version 5, 2000.

[64] MathWorks. Image processing toolbox user's guide version 5.1, 2005.

[65] MathWorks. Statistics toolbox user's guide version 5.1, 2005.

[66] MathWorks. Matlab, the language of technical computing, r2007a, 2007.

[67] J. C. Middlebrooks. Narrow-band sound localization related to external ear acoustics. *Journal of Acoustical Society of America*, (92):2607–2624, 1992.

[68] H. Moller. Fundamentals of binaural technology. *Applied Acoustics*, (0003), 1992.

[69] H. Moller, M. Sorensson, D. Hammershoi, and C. Jensen. Head related transfer functions of human subjects. *Audio Engineering Society*, 43(5), 1995.

[70] B. C. Moore. *An introduction to the Psychology of Hearing.* Academic Press, 2003.

[71] P. A. Nelson. Active control for virtual acoustics. *Proceeding of Active 2002: The 2002 International Symposium on Active Control of Sound and Vibration*, 1:67–89, 2002.

[72] P. A. Nelson and S. J. Elliott. *Active Control of Sound.* Academic Press, London, 1992.

[73] P. A. Nelson, H. Hamada, and S. J. Elliott. Adaptive inverse filters for stereophonic sound reproduction. *Institute of Electrical and Electronics Engineers Transactions on Acoustics, Speech and Signal Processing*, 40:1621–1632, 1992.

[74] P. A. Nelson, O. Kirkeby, and H. Hamada. Local sound field reproduction using two closely spaced loudspeakers. *Journal of the Acoustical Society of America.*, 104:1973–1981, 1998.

[75] P. A. Nelson, F. Orduna-Bustamante, and D. Engler. Multi-channel signal processing techniques in the reproduction of sound. *Journal of the Audio Engineering Society*, 44:973–989, 1996.

[76] P. A. Nelson and J. F. W. Rose. The time domain response of some systems for sound reproduction. *Journal of Sound and Vibration*, 296:461–493, 2006.

[77] G. Neu, E. Mommertz, and A. Shmitz. Investigateions on true directional sound reproduction by playing head-referred recording over two loudspeakers. *Acoustica 76*, 76:183–192 (In German), 1992.

[78] A. Neumaier and T. Schneider. Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM transactions on Mathematical Software*, 27:25–57, 2001.

[79] S. G. Norcross, G. A. Souldre, and Lavoie M. C. Subjective investigations of inverse filtering. *Journal of the Audio Engineering Society*, 52, 2004.

[80] A. V. Oppenheim and W. Schafer, R. *Discrete Time Signal Processing.* Prentice Hall, NJ, 1989.

[81] N. Otsu. A threshold selection method from gray-level histograms. *Institute of Electrical and Electronics Engineers: Transactions Systems, Man and Cybernetics*, 9:62–66, 1979.

[82] T. Papadopoulos and P. A. Nelson. The effectiveness of cross-talk cancellation and equalisation of electroacoustic plants in virtual acoustic imaging systems. *Proceedings of the Institute of Acoustics*, page 26(2), 2004.

[83] A. D. Pierce. *Acoustics, An Introduction to Its Physical Principles and Applications.* Acoustical Society of America - American Institute of Acoustics, 1994.

[84] J. G. Proakis and D. G. Manolakis. *Digital signal processing principles, algorithms, and applications.* New York, Macmillan, 1992.

[85] Lord. Rayleigh. On our perception of sound direction. *Philosophical magazine*, 13:214–232, 1907.

[86] J. Rose. *A Visually Adaptive Virtual Sound Imaging System.* Phd, University of Southampton, 2003.

[87] J. Rose, P. A. Nelson, B. Rafaely, and T. Takeuchi. Sweet spot size of virtual acoustic imaging systems at asymmetric listener locations. *Journal of the Acoustical Society of America*, 112(5 I):1992–2002, 2002.

[88] L. Savioja, J. Huopaniemi, T. Lokki, and R. Vaananen. Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society*, 47:675–705, 1999.

[89] T. Schneider and A. Neumaier. A matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM transactions on Mathematical Software*, 27:58–65, 2001.

[90] L. Sigal, S. Sclaroff, and V Athitsos. Skin color-based video segmentation under time-varying illumination. Technical report, Boston University Computer Science, 2003.

[91] J. S. Soo and K. K. Pang. Multidelay block frequency adaptive filter. *IEEE Transactions Acoutics, Speech, and Signal Processing*, 38, 1990.

[92] T. G. Stockham Jr. High speed convolution and correlation. *AFIPS Procedings 1966 Spring Joint Computer Conference*, 28:229–233, 1966.

[93] M. Storring, T. Kocka, H. Andersen, and E. Granum. Tracking regions of human skin through illumination changes. Technical report, Computer Vision and Media Technology Laboratory, Aalborg University., 2003.

[94] M. Subbarao. *Interpretation of Visual Motion: A Computational Study.* Pitman Publishing, London, 1988.

[95] T. Takeuchi. *Systems for virtual acoustic imaging using the binaural principle.* PhD thesis, University of Southampton, 2001.

[96] T. Takeuchi and P. A. Nelson. Robustness to head misalignment of virtual acoustic imaging systems. *Journal of the Acoustical Society of America*, 109:958–971, 2001.

[97] T. Takeuchi and P. A. Nelson. Optimal source distribution for binaural synthesis over loudspeakers. *Journal of the Acoustical Society of America*, 112:2786–2797, 2002.

[98] A. Torger and A. Farina. Real-time partitioned convolution for ambiophonics and surround. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

[99] V. Valimaki and T. Laksoo. Suppression of transients in variable recursive digital filters with a novel and efficeint cancellation method. *IEEE Transactions on Signal Processing*, 46:3408–3414, 1998.

[100] D. Ward and G. Elko. Optimum loudspeaker spacing for robust cross-talk cancellation. *Institute of Electrical and Electronics Engineers, Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1:3514–3544, 1998.