UNIVERSITY OF SOUTHAMPTON

Systems for virtual acoustic imaging using the binaural principle

by

Takashi Takeuchi

Doctor of Philosophy

FACULTY OF ENGINEERING AND APPLIED SCIENCE

INSTITUTE OF SOUND AND VIBRATION RESEARCH

September 2001

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND APPLIED SCIENCE

INSTITUTE OF SOUND AND VIBRATION RESEARCH

Doctor of Philosophy

SYSTEMS FOR VIRTUAL ACOUSTIC IMAGING USING THE BINAURAL PRINCIPLE

by Takashi Takeuchi

This work concerns binaural control through the use of loudspeakers in systems used for virtual acoustic imaging. Various aspects involved in the process are investigated. An interpolation and extrapolation scheme is described that provides an accurate method for estimating the head related transfer function at arbitrary locations. A number of practical issues are described, discussed and clarified. These include microphone location, the treatment of ear canal responses and transducer responses, and the design of inverse filters.

Features of the "Stereo Dipole", a system with closely spaced control transducers, are discussed. The principles behind such a system and the practical implementation of a system with finite source separation are described and evaluated. Factors contained in the plant that could lessen the performance such as individual differences in head related transfer functions, misalignment of head and control transducers, and reflections in the reproduction space are examined closely.

The system inversion (cross-talk cancellation) is difficult over certain frequency ranges. Furthermore, the process involved gives rise to a number of problems such as a loss of dynamic range. Following a detailed investigation, a number of methods are described that aim to overcome or mitigate the problems associated with virtual acoustic imaging. The "Optimal Source Distribution" introduces the concept of a variable transducer span that enables natural and effective spatial sound reproduction with the minimum manipulation of sound signals. A number of practical solutions such as discretization are suggested for the realization of such optimally distributed transducers. Characteristics of the plant at various elevation positions of the control transducers are also investigated. As a consequence, a position in the frontal plane above the listener's head is found to be an attractive alternative to the conventional horizontal position.

Table of Contents

Acknowledgement

First and most importantly, I would like to thank Prof. Nelson for his tremendous help not only with this research but also with my entire career. Without it, my life would not be as it is now. Secondly, I would like to thank all the staff and students at the ISVR for their vital help with discussions, advice, construction of devices, experiments etc. I would also like to thank all the sponsors, principally Kajima Corporation and its people, all of whom supported this research. Last but not least, I am grateful to my family, especially my wife Yumiko for her devoted support.

1 Introduction

1.1 A brief review of spatial sound reproduction systems

Having two ears, human beings can to some extent discriminate sound spatially as well as resolve its frequency and temporal components. Many trials have been undertaken to reproduce the spatial impression given to a listener through the use of electroacoustic systems. One of the earliest attempts to reproduce spatial aspects of sound was probably made by Ader with a system installed in the Paris Opera House and the Electrical Exhibition Hall in 1881. The system seems to have tried to create spatial sound by transmitting the different signals from a number of microphones on the stage to each of a number of pairs of telephone receivers[1][2].

1.1.1 Stereophony

Since the use of loudspeakers became popular for sound reproduction systems in the 1920's, there were many trials to provide directional information with two or more loudspeakers rather than headphones. It had been known that a phantom image could be produced by intensity differences between two loudspeakers located in front of a listener. Blumlein was one of the earliest workers in this field to exploit this knowledge, which is the basis of stereophonic sound reproduction systems. In 1933 [3], he proposed a system which converted the phase difference between closely placed microphones in the original sound field into the intensity difference of the loudspeakers in the reproducing space. At the same time, work in this area was also being undertaken at Bell Laboratories in 1934 which made use of an additional loudspeaker in the centre [4].

Frontal localisation with Stereophony is achieved by having the loudspeakers in front of the head. Lateralisation with Stereophony mainly depends on interaural level difference which is induced by placing one loudspeaker to the right and the other to the left, though it can not be controlled fully due to cross-talk. Cross-talk is the process by which sound from the right speaker is received by the left ear and vice versa. Thanks to head diffraction, the cross-talk in the mid-high frequency range diminishes and interaural level difference is reasonably well reproduced. Thus, localisation in the horizontal plane between two speakers in front can be realised fairly well.

A number of derivatives of Stereophony appeared later, which includes quadraphony, Dolby stereo, Dolby surround, ITU-R(CCIR) and so on. The main difficulty with these reproduction systems is that they rely strongly on the so-called phantom source effect. The cues which humans use for sound localisation are different for each arrival direction of sound. Consequently, the theory of phantom sources is no longer effective for back - front or up – down localisation. Other methods for reproducing spatial information have been proposed such as "Ambisonics"[5], "Delta Stereophony" [6], "Omnimax" [7], and "Wave field synthesis" [8].

1.1.2 Binaural reproduction

Two channel dummy head recordings and headphone reproductions have also been made by many workers. The most popularly known earliest trial is the dummy "Oscar" at the Century of Progress Exposition in Chicago in 1933 [1]. Since then, the art of dummy head transmission has undergone vast improvement by Boer and Vermeulen in 1939 [9], Damaske in 1968 [10] and subsequently by others [11]-[13].

These systems are based on the idea that human beings primarily make use of the information from the two ears to listen to sound, i.e. the skin of the face, vibrations propagating through the skull, and other factors have little effect on sound perception. Moreover, to obtain acoustical information including directional information through the use of the ears, it is necessary in principle only to reproduce the sound pressures at the listener's two ears that are exactly the same as those in the original sound field.

1.1.3 Binaural reproduction over loudspeakers

To ensure independent sound reproduction at both ears with loudspeakers, the cross-talk compensation method was first proposed by Schroeder and Atal in 1963 [14][15]. This method cancels cross-talk between loudspeakers and ears by feeding each corresponding signal through a matrix of cross-talk cancellation filters and then to the loudspeakers. The cross-talk cancellation filter matrix can in principle be found by performing the inverse Fourier transform of the inverse matrix of transfer functions between two loudspeakers and both ears. Damaske verified in 1971 [16] by subjective experiment that cross-talk compensation enabled binaural reproduction using loudspeakers. He obtained results that showed good localisation over most directions on the horizontal plane.

1.2 Objectives

Our primary objective is to create all the information contained in a sound environment in order to ensure the accurate synthesis of the sound environment. Therefore, binaural technology is adopted here as a fundamental principle. The principle of this technology is to control the sound field at the listener’s ears so that the reproduced sound field coincides with what would be produced when he is in the desired real sound field. Binaural technology is often used to present an accurate virtual acoustic environment to a listener. The superiority of this binaural technique lies in its capability of providing very accurate spatial impression to a listener. A system based on the binaural technique can produce the accurate illusion of a virtual acoustic space, which includes direction, distance and movement of sound sources, and the acoustic response of the surrounding space such as rooms. Producing the correct ear sound pressures should lead to almost the same sensation as the listener would experience in the real sound field for most realistic sound signals. Then the listener would experience an extremely realistic three dimensional sound environment. Appropriate control of directional information associated with direct and reflected sounds, as well as information regarding reflecting surfaces, together with the distance that the sound has travelled and information from the sound source itself, is essential to creating a convincing virtual auditory space. In exchange for the ability of such precise control, the area of control is limited to only around the ears of the listener. A binaural based sound system is essentially a personal system although it is possible in principle to extend it to use by multiple listeners.

Binaural technology requires the control of the sound at each of two ears independently. It is necessary to feed each corresponding sound signal into each ear in order to achieve independent sound reproduction at both ears. One way of achieving this is to use a pair of headphones or similar types of transducers. Headphones ensure that the sound signal for each ear is only heard by that ear. However, presenting the correct sound signals is not as easy. Errors induced by the whole process often result in perceptual errors, such as inside or on the head localisation and bias localisation errors in the rear hemisphere.

An alternative to this is to use a pair of loudspeakers in a listening space with the help of signal processing to ensure that appropriate binaural signals are obtained at the listener's ears [14]-[20]. This approach has many advantages. One of them is that the listener does not need to attach any objects such as headphones. Most of the advantages the loudspeaker as a transducer has over headphones apply to this case as well. Not only the ears but the listener's head, even the whole body, is exposed to the sound so the listener obtains a much more realistic sensation of the virtual sound environment. Although the system controls the sound signals at both ears only, it often results in reasonably correct reproduction of sound signals in the region of other parts of the body at low frequencies (which such body parts could sense), apart from the path through the opening of the nose and mouth. However, the method has a number of drawbacks as well. Many of them are associated with the inversion of the plant that constitutes the transfer functions between the transducers and the ears. The typical error induced with this approach is the bias localisation error which tends to produce the perception of virtual source positions towards the transducer positions. Since the presented acoustic environment, especially the spatial aspects of it, is different from what the listener would expect from the position of the electro-acoustic transducers and the property of the listening space (the real acoustic environment), the method of binaural reproduction over loudspeakers is also often referred to as virtual acoustic imaging.

The objective of this work is to investigate the performance of such virtual acoustic imaging systems. Various aspects of this kind of system are explored. Factors contained in the plant that could lessen the performance of such systems received particularly intense attention during this project. Through the investigation presented here, a number of methods are described that aim to overcome or mitigate the problems associated with virtual acoustic imaging. The performance of the sound reproduction system includes the quality of sounds as well as the ability to produce the spatial aspects of the intended sound field. However, we shall largely concentrate on the performance related to the spatial aspects. In many cases, the ability to localise the direction of a single sound wave is investigated as the most fundamental element comprising a spatial sound environment.

1.3 Organisation of this thesis

This thesis consisting of 9 chapters is a compilation of the projects undertaken by the author at the Institute of Sound and Vibration Research, University of Southampton, in the period between October 1995 and September 1997, and after an interruption, between February 1999 and January 2001. The first chapter serves as an introduction with a brief review of spatial sound reproduction systems and human auditory function of spatial hearing. The second chapter explains the principles of virtual acoustic imaging using binaural control with loudspeakers that receives the main attention in this study. A model for the design and evaluation of binaural synthesis over loudspeakers was defined. An interpolation and extrapolation scheme described here provides a very accurate method of estimating the head related transfer function at arbitrary locations. A number of practical issues are described, discussed or clarified such as microphone location, treatment of the ear canal responses and transducer responses, and inverse filter design. In Chapter 3, features of the "Stereo Dipole", a system with closely spaced control transducers, are discussed. The principles of the Stereo Dipole and a practical system with finite source separation are described and evaluated. Factors such as individual differences in head related transfer functions, misalignment of head and control transducers, and reflections in the reproduction space that are contained in the plant could lessen the performance of such systems and studies of these are presented from Chapter 4 to Chapter 6. In Chapter 7, the system inversion involved in such systems that gives rise to a number of problems such as a loss of dynamic range and other problems discussed in previous chapters is investigated in depth. A method of overcoming these fundamental problems is proposed. The "Optimal Source Distribution" introduces the concept of variable transducer span that enables natural and effective spatial sound reproduction with the minimum manipulation of sound signals. A number of practical solutions such as discretization are suggested for the realization of such optimally distributed transducers. Chapter 8 presents a study of the characteristics of various elevation positions of the control transducers, particularly for the position in the frontal plane above the listener's head that is found to be an attractive alternative to the conventional horizontal position. The last chapter concludes with a summary.

1.4 The psychological and physiological function of spatial hearing

Humans have the ability to sense the arrival direction and travel distance of a sound wave. This ability plays an important role in the spatial perception of sound. The listener can extract information about the environment by analysing the structure of the impinging sound waves, as well as the information within the sound source signals themselves. When designing or evaluating spatial sound reproduction systems, it is important to understand the way that human beings perceive the spatial attributes of sound. In order to manipulate a listener's spatial auditory perception, the system needs to modify accordingly the information that the auditory system utilises. On the other hand, knowing the information that the sound reproduction system provides to the listener enables to some extent the estimation of the auditory perception, and therefore, the estimation of the performance of such a system. The following section presents a summary of the current understanding of the psychological and physiological function of spatial hearing with regard to the localisation of a single sound source.

However, knowledge of the human auditory system is still limited despite the extensive research undertaken which has revealed much about the human perception of sound. Since the sound signals provided by a virtual acoustic imaging system to a listener can often be in a manner that can never happen in a real environment, experiments with human subjects are also necessary to supplement the analysis based on available psycho-acoustical knowledge.

1.4.1 Directional Cues

It is widely agreed that humans extract the difference of arrival time at both ears (the interaural time difference) and use this information to locate the direction of sound along the interaural axis (lateral localisation). Anatomical and physiological studies strongly suggest that the interaural time difference (ITD) information is extracted together with the interaural cross-correlation of the auditory-nerve responses in the superior olivary complex and are then further processed at a higher level of the auditory pathway [21][22]. The ITD associated with high frequency sound can also be extracted in the form of signal envelope delay by the cross-correlation system as well as enabling the extraction of the phase delay of low frequency signals [23][24]. This is thought to be the primary cue for lateral localisation. The threshold for ITD discrimination is considered to be approximately 10μs [25] which corresponds to about 1° in azimuth localisation of a far field source.

The head (and the pinae and the body of the listener) modifies the spectrum of sound in a distincitive manner that depends strongly upon the direction of arrival. The frequency analysis system in the cochlea can be used to extract this information as well as perceiving the frequency of sound signal contents. Separating (monaurally) the change of spectral shape due to the transmission of sound from the spectra of the sound source signal itself requires additional information. Usually, through his experience, the listener has a memory of both the spectra of sound source signal and spectral change due to transmission so that separation of these two provides little problem. Since the ITD cue supplies very strong information for lateral localisation, the change of the spectral shape of the sound source signal (the spectral cue) is primarily used to localise in vertical directions and in fore-and-aft directions. The monaural spectral shape is regarded as an important cue to identify one direction of arrival out of a number of directions of arrival which have the same interaural differences. However, the monaural spectral cues also have a supplemental role in localisation along the interaural axis (lateral localisation) [26]. The interaural difference of spectra could have two roles. The major role is to localise along the interaural axis with interaural level difference (ILD). This is particularly useful when the ITD cue has ambiguity, such as in the case of high frequency sound with little envelope change. It could also provide another cue to resolve confusion among directions with no interaural time difference by utilising the pattern of frequency dependent interaural spectral differences [27]. The advantage of this cue over the monaural spectral shape cue in practice would be that it does not depend on the spectrum of the sound source signal.

Dynamic change of the cues also supplements the perception of direction. For example, when the listener does not have a memory of the spectrum of the sound source signal or the spectral change due to transmission, movement of the listener's head helps the separation of spectra. The spectrum of the source signal is likely to stay the same while the way the head modifies the spectrum changes according to the relative direction of the sound source. Another example is the dynamic change of ITD. Rotation of the head produces change in ITD that is dependent on the source location. For example, in the case of a sound wave arriving from the right side, a right turn of the head will decrease ITD caused by the sound when the sound wave is arriving from front and vice versa when it arrives from the rear. Such information can give additional cues for localisation.

Dynamic cues are particularly useful when developing spatial hearing as a child, when localising unknown sound signals (as often the case with laboratory experiments), and in localising sound when other important directional cues are poorly presented by a sound reproduction system.

1.4.2 Distance cues

Although directional perception (which is a two dimensional problem) usually receives most attention in spatial hearing, distance perception adds another dimension to spatial perception. When a sound source is located far away relative to the size of the head (in the far field), the loudness of the sound is the primary cue used to estimate distance. In order to estimate the distance with the loudness, the listener needs additional information, such as the sound energy emitted by the source, in the same way as the monaural spectral cue requires information regarding the source signal spectra. Again, this provides little problem in normal circumstances where the auditory system has memory of such information gained by experience.

When a sound source is located relatively close to the head (in the near field), the information discussed in the previous section (such as spectral shape) cues has a much stronger distance dependence. There are relatively few studies reported which have investigated the distance cues in the near field region. However, there are well known phenomena that headphone presentation of artificial signals to ears without these cues results in a perception of distance that is very close to the head.

The ratio of direct and reverberant sound also gives additional information regarding distance perception as well as the amount of dynamic change in direction relative to the listener's head movement.

1.5 Coordinate system

The spherical co-ordinate system used to define direction throughout this thesis is shown in Fig. 1.1. This interaural polar coordinate system is used since it coincides well with the way the auditory system perceives the location of a sound source. The origin is at the intersection of the interaural axis and the median plane. The polar axis coincides with the interaural axis. The azimuth angle ranges from -90° to 90° as the direction changes from the pole at the left to the other pole at the right. A cone of constant azimuth is approximately the same as the cone of confusion where there are constant ITDs. The elevation angle ranges from -180° on the horizontal plane behind the head to -90° below, 0° on the horizontal plane in front, 90° above the head to 180° again on the horizontal plane behind.

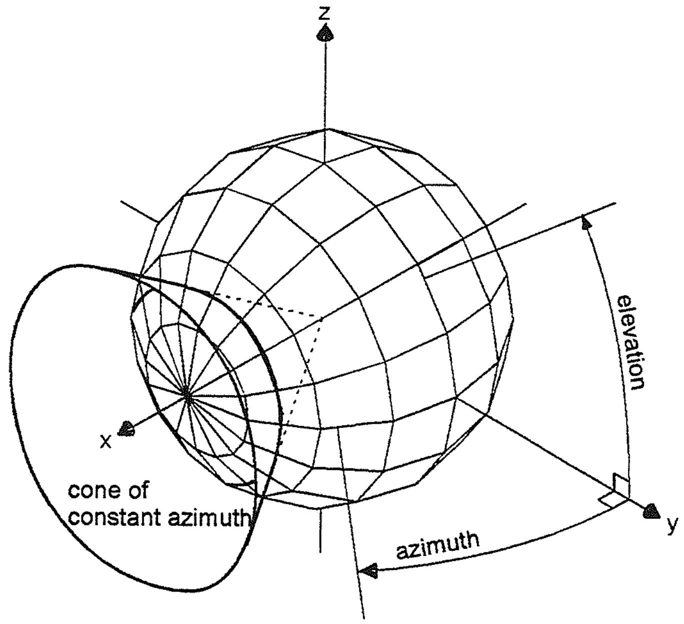


Fig. 1.1 The spherical co-ordinate system used to define the direction of sound sources relative to the listener's head position and orientation. An example of "cone of constant azimuth" is illustrated.

2 Principles of binaural synthesis over loudspeakers

2.1 Elements of the sound environment

When there are sound sources in a space, each source emits acoustic signals into the surrounding air. The fluctuation of the air pressure propagates as sound waves. Some of the waves arrive at the listener directly from the source. Others are reflected, diffracted, scattered, or transmitted by objects in the space (including the boundary of the space) or refracted in the air, before arriving at the listener. The arrival direction and travel distance of each sound wave is governed by the location of the sound sources and the listener, and the shape and size of the space defined by the boundaries. The time-frequency structure of each sound wave is also modified depending upon the characteristics of the objects and the air. Therefore, the sound signals arriving at the listener contain spatial information regarding the environment as a superposition of these sound waves as well as information associated with the sound source signals. Furthermore, the spatial information changes in time when the sound sources, the listener, objects, or air move in the space, or when the temperature or humidity change in time. The movements also introduce the non-linear phenomenon referred to as the Doppler effect. Usually, the term “spatial sound” also includes this dynamic change of the spatial information.

2.2 A model for binaural synthesis

Binaural signals contain most of the information regarding the sound environment to be synthesised (Section 1.1.2). A sound environment can be modelled as follows. This model is also useful in the design or evaluation of a spatial sound reproduction system. An infinite number of possible situations could comprise different sound environments. Therefore, verifying against all of them is impossible. The model used here enables the investigation of the very basic elements that are likely to lead to the maximum understanding of the system behaviour.

Binaural signals are modelled as convolution of the sound source signals with pairs of binaural filters. A pair of binaural filters can be defined as the transfer functions between a sound source and the ears of a listener in order to represent the static spatial information conveyed by each sound source. This can be obtained by measuring the transfer functions between a sound source and two ears in an existing space (Fig. 2.1), or in an acoustic scale model (Fig. 2.2). The dynamic change of the spatial information may be modelled with time-varying binaural filters. The Doppler effect is usually modelled by modifying the sound source signal to be convolved rather than by modifying the transfer functions. Binaural recordings contain this dynamic spatial information as well as sound source signals. They can be obtained by recording sound signals with a recording head (e.g. a dummy head) in an existing sound environment.

In order to obtain a pair of binaural filters, the response of the space can also be calculated. Since solving the wave equation over the entire audible frequency range over a reasonable size of space is not yet feasible, the most popular approach is based on geometrical acoustic models in order to obtain the arrival direction and course of travel of each sound wave. Then the static spatial information is represented by a superposition of each direct, reflected, diffracted, scattered, transmitted, or refracted sound wave.

Now the arrival direction and travel distance of a single sound wave, direct or indirect, is considered as the basic element of spatial sound. Most of the information is contained in a pair of head related transfer functions (HRTFs) and can be obtained by measuring or calculating the transfer functions between a single sound source and the two ears in an anechoic environment. Since the HRTF pairs of arbitrarily located sound sources are required in order to construct arbitrary static spatial information, the measurements are performed at a number of sampled locations in space. In the far field, a spherical surface is sampled directionally and interpolation of direction and extrapolation of distance gives the HRTFs for an arbitrary location. In the near field, it is necessary to sample a sphere both in direction and distance, and interpolate accordingly. The density of sampling in space must be fine enough to avoid spatial aliasing. The detail of the interpolation and extrapolation scheme used in this study is described in Appendix 1.

2.3 Binaural control with loudspeakers

The block diagram defining each of the signals and transfer functions involved in binaural synthesis over loudspeakers is illustrated in Fig. 2.3. The following is described with a frequency domain representation of the acoustic paths (transfer functions) and acoustic signals.

A pair of binaural signals **d** are expressed with a vector of sound source signals **s** and pairs of binaural transfer functions which form the matrix **A**. Each pair of binaural filters corresponds to the transfer functions between each sound source and both of the listener's ears. Thus

$$\mathbf{d} = \mathbf{A}\mathbf{s} \tag{2.1}$$

where

$$\mathbf{d} = \begin{bmatrix} \mathrm{D}_1(\mathrm{j}\omega) \\ \mathrm{D}_2(\mathrm{j}\omega) \end{bmatrix}, \quad \mathbf{s} = \begin{bmatrix} \mathrm{S}_1(\mathrm{j}\omega) \\ \vdots \\ \mathrm{S}_n(\mathrm{j}\omega) \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathrm{A}_{11}(\mathrm{j}\omega) & \cdots & \mathrm{A}_{1n}(\mathrm{j}\omega) \\ \mathrm{A}_{21}(\mathrm{j}\omega) & \cdots & \mathrm{A}_{2n}(\mathrm{j}\omega) \end{bmatrix}$$

(2.2a, b, c)

When there is only one sound source, the vector **s** becomes a scalar $\mathrm{S}(\mathrm{j}\omega)$ and the matrix **A** becomes a vector **a** given by

$$\mathbf{a} = \begin{bmatrix} \mathrm{A}_1(\mathrm{j}\omega) \\ \mathrm{A}_2(\mathrm{j}\omega) \end{bmatrix}$$

(2.3)

The number of rows is 2 because only two points in the sound field (at listener's ears) are to be controlled. Even though the principle here is general to these multi-channel or multi-listener cases, a case with two ears, i.e. one listener, is taken as an example. When there are more than two positions at which to control sound, **A** becomes a matrix with a number of rows equal to the number of positions.

System inversion is used for multi-channel sound control including binaural reproduction over loudspeakers. Independent control of two signals (such as binaural sound signals) at two receivers (such as the ears of a listener) can be achieved with two electro-acoustic transducers (such as loudspeakers), by filtering the input signals to the transducers with the inverse of the transfer function matrix of the plant. When the signals at both ears of the listener are controlled by two transducers in the listening space, the 2x2 matrix **C** of transfer functions can be defined between the transducers and the ears. Two monopole transducers produce source strengths defined by the elements of the complex vector **v**. The resulting acoustic pressure signals are given by the elements of the vector **w**. Thus

$$\mathbf{w} = \mathbf{C}\mathbf{v} \tag{2.4}$$

where

$$\mathbf{C} = \begin{bmatrix} C_{11}(j\omega) & C_{12}(j\omega) \\ C_{21}(j\omega) & C_{22}(j\omega) \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} V_1(j\omega) \\ V_2(j\omega) \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} W_1(j\omega) \\ W_2(j\omega) \end{bmatrix} \tag{2.5a, b, c}$$

The two signals to be synthesised at the receivers are the elements of the complex vector **d**. In order to present the binaural signals at each ear, the signals **d** are filtered through a 2x2 matrix **H** of control filters which contains the pseudo-inverse of the transfer function matrix **C**. Hence

$$\mathbf{v} = \mathbf{H}\mathbf{d} \tag{2.6}$$

where

$$\mathbf{H} = \begin{bmatrix} H_{11}(j\omega) & H_{12}(j\omega) \\ H_{21}(j\omega) & H_{22}(j\omega) \end{bmatrix} \tag{2.7}$$

Then the synthesised binaural signals **w** and the sound source signals **s** are related by

$$\mathbf{w} = \mathbf{CHd} = \mathbf{CHAs} \tag{2.8}$$

For convenience in later analysis, the control performance matrix **X** and vector of synthesised binaural transfer functions $\mathbf{a}_\mathrm{S}$ are defined as follows

$$\mathbf{X} = \mathbf{CH} \tag{2.9}$$

$$\mathbf{a}_\mathrm{S} = \mathbf{CHa} = \mathbf{Xa} \tag{2.10}$$

A number of filter design methods have been presented and a few of these are described in the next Section. In short, with the use of a modelling delay Δ and a regularisation parameter for causal stable inversion, **H** is designed so that the vector **w** is a good approximation to the vector **d** with a certain delay [28]-[31]. When

$$\mathbf{X} = \mathbf{CH} \approx z^{-\Delta}\mathbf{I} \tag{2.11}$$

is satisfied where **I** is the identity matrix, this ensures the synthesised signals **w** are a good approximation to the original binaural signals **d**. Thus, from Eq. (2.10) and (2.11),

$$\mathbf{w} \approx z^{-\Delta}\mathbf{d} \tag{2.12}$$

2.4 Inverse filter design

The realisation of the inverse filter matrix relies on the invertibility of the transfer function matrix. There is a risk of producing an unrealisable cross-talk cancellation matrix if the components of the transfer function matrix are non-minimum phase. A filter design procedure which is capable of dealing with the presence of non-minimum phase components is proposed by Nelson et al [28]-[31]. A least squares approach was used to design the inverse filter matrix and this gives great flexibility for filter design.

The difference between the desired binaural signals **d** and the synthesised signals **w** are defined as the error signal vector **e**. Then

$$\mathbf{e} = \mathbf{d} - \mathbf{w} \tag{2.13}$$

where

$$\mathbf{e} = \begin{bmatrix} \mathrm{E}_1(\mathrm{j}\omega) \\ \mathrm{E}_2(\mathrm{j}\omega) \end{bmatrix} \tag{2.14}$$

A cost function J is defined as the sum of two terms: a "performance error" term $\mathbf{e}^{\mathrm{H}}\mathbf{e}$, which is a measure of how well the desired signals are reproduced at the receivers, and "effort penalty" term $\beta\mathbf{v}^{\mathrm{H}}\mathbf{v}$, which is a measure proportional to the total source power. The superscript H denotes the Hermitian transpose of a matrix; that is the complex conjugate of the transpose [32]. Thus

$$J = \mathbf{e}^{\mathrm{H}}\mathbf{e} + \beta\mathbf{v}^{\mathrm{H}}\mathbf{v}$$

(2.15)

The positive real number β is a regularisation parameter that determines how much weight to assign to the effort term. By varying β from zero to infinity, the solution to minimise the cost function J changes gradually from minimising only the performance error to minimising only the effort cost. When $\beta \triangleright 0$, J is minimised in the least squares sense by a vector **v** of source strengths that is given by

$$\mathbf{v} = \left[\mathbf{C}^{\mathrm{H}}\mathbf{C} + \beta\mathbf{I}\right]^{-1}\mathbf{C}^{\mathrm{H}}\mathbf{d}$$

(2.16)

If **v** is to take its optimal value for any choice of **d**, then according to Eq.(2.6), the matrix of inverse filters **H** must be given by

$$\mathbf{H} = \left[\mathbf{C}^{\mathrm{H}}\mathbf{C} + \beta\mathbf{I}\right]^{-1}\mathbf{C}^{\mathrm{H}}$$

(2.17)

2.5 Practical implementation

In order to implement a virtual acoustic imaging system in practice, it is necessary to obtain sound signals **s** and transfer functions **A** and **C** in Eq. (2.8). Their definition and relationship are illustrated in detail in Fig. 2.4. A case with only one sound source is explained here as an example.

2.5.1 Recording information in the sound environment

When a listener obtains information in a real sound environment, a scalar original source signal $S(j\omega)$ is filtered by a vector **p** of transfer functions between the original source and the listener's ear canal entrances. When there are number of sound sources, **p** becomes a matrix **P** with a number of columns equal to the number of sources. The vector **p** contains the HRTFs of the listener. The resulting sound signals are multiplied by a diagonal matrix **Y** of ear canal transfer functions to produce a vector **d** of sound signals at both ear drums (Fig. 2.4a). The relationship between signals at each point is given by:

$$\mathbf{d} = \mathbf{a}S(j\omega) = \mathbf{Y}\mathbf{p}S(j\omega) \tag{2.18}$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1(j\omega) & 0 \\ 0 & Y_2(j\omega) \end{bmatrix} \tag{2.19}$$

The transfer functions **a** between sound source and ear drums are separated into two (ref. Eq. (2.1)), because in the audible frequency range, sound propagates as plane waves in the ear canals so that **Y** can be regarded as being independent of source direction. It is **p** that contains directional and spatial information.

The virtual acoustic imaging system attempts to synthesise these signals **d** as closely as possible by using a limited number of transducers. The easiest way to obtain the vector of signals **d** is to record the sound source signal $S(j\omega)$ with a recording head which is usually an artificial head (Fig. 2.4b). In contrast to Eq. (2.1), the recorded signals $\mathbf{f}_d$ obtained also inevitably contain a diagonal matrix **M** of microphone responses as well as the desired binaural signals $\mathbf{d}_d$. Thus the recorded signals are expressed by:

$$\mathbf{f}_d = \mathbf{M}\mathbf{d}_d = \mathbf{M}\mathbf{a}_d \, S(j\omega) = \mathbf{M}\mathbf{Y}_d\mathbf{p}_d \, S(j\omega) \tag{2.20}$$

where

$$\mathbf{M} = \begin{bmatrix} M_1(j\omega) & 0 \\ 0 & M_2(j\omega) \end{bmatrix} \tag{2.21}$$

The subscript d is necessary since the recording head may be different from the listener's own head. Dynamic spatial information as well as the sound source signals are obtained in this way.

2.5.2 Measurement of spatial information

Alternatively, it is possible to measure the vector of transfer functions **a** that contains the spatial information associated with a sound environment. Transfer function measurement is normally possible only when the sound environment is static. Static binaural filters can be measured as an acoustic response of a space (e.g. room response) including HRTFs. Alternatively, an HRTF pair associated with an arbitrary location of a sound source may be measured and the static binaural filters are constructed from them as described in Section 2.2. In both cases, transfer functions between a sound source and listener's ears are measured with an electro-acoustic transducer located whose response is expressed as a scalar $L_a(j\omega)$ at the position of a sound source and microphones whose responses are $\mathbf{M}$ at one of the listener's ears (Fig. 2.4c). The measured matrix of transfer functions $\mathbf{B}_a$ contains transfer functions $\mathbf{p}_d$ and $\mathbf{Y}_d$ of the measurement head (which is again usually an artificial head) and can be expressed as follows.

$$\mathbf{B}_a = \mathbf{M}\mathbf{Y}_d\mathbf{p}_d L_a(j\omega) \tag{2.22}$$

In order to separate transducer responses from the acoustic propagation characteristics, it is necessary to obtain $L_a(j\omega)$ and $\mathbf{M}$ by a free field measurement with microphones and a loudspeaker. Then $L_a(j\omega)$ and $M_1(j\omega)$ or $M_2(j\omega)$ are deconvolved with a inverse filter designed from the measurement to obtain

$$\begin{aligned}\mathbf{a}_d &= [\mathbf{M}L_a(j\omega)]^{-1}\mathbf{B}_a \\ &= L_a(j\omega)^{-1}\mathbf{M}^{-1}\mathbf{M}\mathbf{Y}_d\mathbf{p}_d L_a(j\omega) \\ &= \mathbf{Y}_d\mathbf{p}_d\end{aligned} \tag{2.23}$$

2.5.3 Measurement of the plant matrix

The transfer functions between the control transducers and the listener's ear drums can be separated again (ref. Eq. (2.8)) into a diagonal matrix $\mathbf{L}$ which contains the responses of the control transducers, a matrix $\mathbf{Q}$ of transfer functions, including HRTFs, between the output of the transducers and the entrance of the listener's ear canals, and the diagonal matrix $\mathbf{Y}$ of ear canal transfer functions. Hence

$$\mathbf{C} = \mathbf{YQL} \tag{2.24}$$

where

$$\mathbf{L} = \begin{bmatrix} \mathrm{L}_1(\mathrm{j}\omega) & 0 \\ 0 & \mathrm{L}_2(\mathrm{j}\omega) \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} \mathrm{Q}_{11}(\mathrm{j}\omega)\mathrm{Q}_{12}(\mathrm{j}\omega) \\ \mathrm{Q}_{21}(\mathrm{j}\omega)\mathrm{Q}_{22}(\mathrm{j}\omega) \end{bmatrix}$$

(2.25a, b)

It is **Q** that contains directional and spatial information relevant to the listening space. To design the system inversion filters **H**, the matrix **C** can be measured with the same microphones used in obtaining **p** together with the same transducers that are used for the binaural control (Fig. 2.4d). The measured matrix of transfer functions $\mathbf{B}_\mathrm{C}$ contains transfer functions $\mathbf{Q}_\mathrm{d}$ and $\mathbf{Y}_\mathrm{d}$ with the measurement head that is again usually an artificial head. When the same head is used as that which is used to obtain **p**, the measured plant transfer function matrix $\mathbf{B}_\mathrm{C}$ is expressed as

$$\mathbf{B}_\mathrm{C} = \mathbf{M}\mathbf{Y}_\mathrm{d}\mathbf{Q}_\mathrm{d}\mathbf{L}$$

(2.26)

where

$$\mathbf{B}_\mathrm{C} = \begin{bmatrix} \mathrm{B}_{11}(\mathrm{j}\omega)\mathrm{B}_{12}(\mathrm{j}\omega) \\ \mathrm{B}_{21}(\mathrm{j}\omega)\mathrm{B}_{22}(\mathrm{j}\omega) \end{bmatrix}$$

(2.27)

In order to separate the transducer responses from the acoustic propagation characteristics, it is necessary to obtain **L** and **M** by free field measurements with the microphones and the control transducers. Then **L** and **M** are deconvolved with inverse filters designed from the measurement to obtain

$$\mathbf{C}_\mathrm{d} = \mathbf{Y}_\mathrm{d}\mathbf{Q}_\mathrm{d} \tag{2.28}$$

2.5.4 Design of the inverse filters

The inverse filter matrix **H** can be designed by inverting $\mathbf{B}_\mathrm{C}$. When the matrix inversion is perfect,

$$\mathbf{H} = \mathbf{B}_\mathrm{C}^{-1} = [\mathbf{M}\mathbf{Y}_\mathrm{d}\mathbf{Q}_\mathrm{d}\mathbf{L}]^{-1} = \mathbf{L}^{-1}\mathbf{Q}_\mathrm{d}^{-1}\mathbf{Y}_\mathrm{d}^{-1}\mathbf{M}^{-1} \tag{2.29}$$

The inverse filter matrix could also be designed from the acoustic propagation characteristics in the plant after the transducer responses are separated. From Eq. (2.28),

$$\mathbf{H} = \mathbf{Q}_\mathrm{d}^{-1}\mathbf{Y}_\mathrm{d}^{-1} \tag{2.30}$$

2.5.5 Synthesised binaural signals

When a binaural recording is used for reproduction, the recorded signals **f** are used in place of **d** in Eq. (2.8). Combination of the signals and transfer functions in Eq. (2.21), (2.24), and (2.29) gives

$$\begin{aligned}\mathbf{w} &= \mathbf{C}\mathbf{H}\mathbf{f}_{\mathrm{d}} \\ &= \mathbf{Y}\mathbf{Q}\mathbf{L}\mathbf{L}^{-1}\mathbf{Q}_{\mathrm{d}}^{-1}\mathbf{Y}_{\mathrm{d}}^{-1}\mathbf{M}^{-1}\mathbf{M}\mathbf{Y}_{\mathrm{d}}\mathbf{p}_{\mathrm{d}}\,S(j\omega) \\ &= \mathbf{Y}\mathbf{Q}\mathbf{Q}_{\mathrm{d}}^{-1}\mathbf{Y}_{\mathrm{d}}^{-1}\mathbf{Y}_{\mathrm{d}}\mathbf{p}_{\mathrm{d}}\,S(j\omega) \\ &= \mathbf{Y}\mathbf{Q}\mathbf{Q}_{\mathrm{d}}^{-1}\mathbf{p}_{\mathrm{d}}\,S(j\omega)\end{aligned} \tag{2.31}$$

Note that using this procedure ensures that the microphone and loudspeaker responses are deconvolved at the same time. If we assume that $\mathbf{p}_{\mathrm{d}} \approx \mathbf{p}$ and $\mathbf{Q}_{\mathrm{d}} \approx \mathbf{Q}$, Eq. (2.31) becomes

$$\mathbf{w} \approx \mathbf{Y}\mathbf{p}\,S(j\omega) = \mathbf{d} \tag{2.32}$$

Therefore in this case, perfect binaural reproduction is achieved. It is therefore not necessary to include ear canal responses in the recordings, but it is also possible to include this response. In that case, ear canal responses $\mathbf{Y}_{\mathrm{d}}$ of the recording head included in **f** are deconvolved by **H** anyway. The ear canal of a dummy head with microphones at the ear drum position is used in a similar way to using a probe tube microphone to obtain sound pressure at the ear canal entrance. The advantage of this method is that there is no additional object such as a probe tube in the measurement space which could be a source of error.

Even when a probe is used for the measurement, it does not matter, in principle, at which position along the ear canal the signal is measured, as long as it is obtained at the same position and with the same ear canal for recording **f** and measuring **C**. However, the existence of ear canals may be important since they may be a vital part to interweave correct information with the transfer functions themselves. In practice, transfer functions along the canal may at certain frequencies affect the signal to noise ratio of the recorded/measured signals or transfer functions, since it is known that the ear canal has a resonance within the audible frequency range. It is of course desirable from the point of view of safety to measure and record as far away as possible from the ear drum when human subjects are used for recording or measurement.

It is also possible to synthesise a virtual acoustic environment with measured transfer functions $\mathbf{B}_a$. In such a case, combination of obtained signals and transfer functions in Eq. (2.22), (2.24), and (2.29) together with a recorded sound source signal S(jω) (e.g. free field recording by a microphone with a response $M_S(j\omega)$) gives

$$\begin{aligned}\mathbf{w} &= \mathbf{CHB}_a M_S(j\omega)S(j\omega)\\ &= \mathbf{YQLL}^{-1}\mathbf{Q}_d^{-1}\mathbf{Y}_d^{-1}\mathbf{M}^{-1}\mathbf{MY}_d\mathbf{p}_d L_a(j\omega)M_S(j\omega)S(j\omega)\end{aligned}$$

(2.33)

Since $\mathbf{LL}^{-1} = \mathbf{I}$ and $\mathbf{MM}^{-1} = \mathbf{I}$,

$$\begin{aligned}\mathbf{w} &= \mathbf{YQQ}_d^{-1}\mathbf{Y}_d^{-1}\mathbf{Y}_d\mathbf{p}_d L_a(j\omega)M_S(j\omega)S(j\omega)\\ &= \mathbf{YQQ}_d^{-1}\mathbf{p}_d L_a(j\omega)M_S(j\omega)S(j\omega)\end{aligned}$$

(2.34)

Again, we assume that $\mathbf{p}_d \approx \mathbf{p}$ and $\mathbf{Q}_d \approx \mathbf{Q}$, and therefore, Eq. (2.34) becomes

$$\begin{aligned}\mathbf{w} &\approx \mathbf{Yp} L_a(j\omega)M_S(j\omega)S(j\omega)\\ &= \mathbf{d} L_a(j\omega)M_S(j\omega)\end{aligned}$$

(2.35)

If a virtual acoustic environment is synthesised with measured transfer functions from which the transducer responses are deconvolved (Eq.(2.23), Eq, (2.30)), together with Eq. (2.24), the synthesised signals become

$$
\begin{aligned}
\mathbf{w} &= \mathbf{CHa}_d \, M_S(j\omega)S(j\omega) \\
&= \mathbf{YQLQ}_d^{-1}\mathbf{Y}_d^{-1}\mathbf{Y}_d\mathbf{p}_d \, M_S(j\omega)S(j\omega) \\
&= \mathbf{YQLQ}_d^{-1}\mathbf{p}_d \, M_S(j\omega)S(j\omega)
\end{aligned}
\tag{2.36}
$$

In Eq.(2.35), the responses of the loudspeakers and microphones affect monaural cues but they do not affect the binaural cues. Here, the difference in responses between two control transducers induce binaural error in the HRTF synthesis. However, if $L_1(j\omega) = L_2(j\omega) = L_a(j\omega)$, i.e. their responses are identical, Eq.(2.36) becomes

$$
\begin{aligned}
\mathbf{w} &= \mathbf{YQQ}_d^{-1}\mathbf{p}_d L_a \, M_S(j\omega)S(j\omega) \\
&\approx \mathbf{Yp}L_a \, M_S(j\omega)S(j\omega) \\
&= \mathbf{d}L_a \, M_S(j\omega)
\end{aligned}
\tag{2.37}
$$

which is identical to Eq. (2.35). The effect of the loudspeakers becomes independent of binaural part of synthesised HRTFs and is equivalent to degrading the sound source signal. The responses of the loudspeakers of course still affect monaural cues, but they do not affect the binaural cues. Therefore, for binaural synthesis, it is important to use a well-matched pair of loudspeakers.

In Eq.(2.35) or Eq.(2.37), the synthesised signal includes a surplus microphone response and a loudspeaker response. In order to remove the transducer response, it is necessary to obtain $L_a(j\omega)M_S(j\omega)$ by a free field measurement and deconvolve them with another inverse filter designed from the measurement.

If the sound source signal itself is also synthesised (e.g. with a computer generated signal), Eq.(2.35) becomes

$$\begin{aligned} \mathbf{w} &\approx \mathbf{Yp}\,L_a(j\omega)S(j\omega) \\ &= \mathbf{d}\,L_a(j\omega) \end{aligned} \qquad (2.38)$$

The synthesised binaural signals contain a surplus loudspeaker response $L_a(j\omega)$ only, therefore, this is equivalent to the case that all the virtual sound sources are loudspeakers each of them playing a corresponding sound source signal. However, it is not possible to obtain $L_a(j\omega)$ alone in principle. In practice, it is easier to obtain a microphone than a loudspeaker that has reasonably good response. Therefore, $L_a(j\omega)$ may be measured with a good microphone in a free field and an inverse filter can be designed in order to deconvolve the loudspeaker response. The inverted microphone response may be ignored, provided that it has a flat response over the audible frequency range.

2.6 Conclusions

This chapter described the principles of binaural synthesis over loudspeakers. A model of a sound environment for the design and evaluation of the method of binaural synthesis has been defined which forms the basis of the analysis and the experiments described in the following chapters. The spherical interpolation and extrapolation scheme described in (Appendix 1) provides a very accurate method for estimating the head related transfer function at arbitrary locations from a sampled database. This enabled the detailed analysis of a realistic HRTF based model in following Chapters. A number of practical issues such as microphone location, treatment of the ear canal responses and transducer responses, and inverse filter design have been described, discussed and clarified.

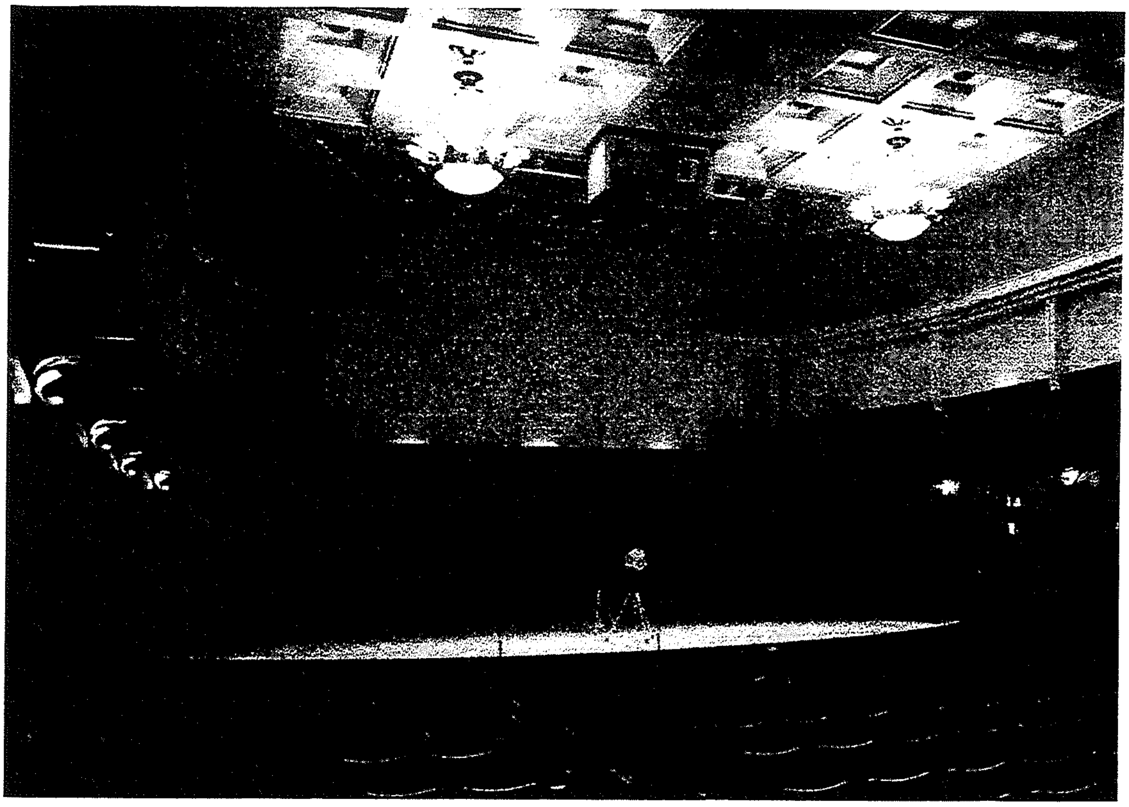


Fig. 2.1 Measurement of the transfer functions between a sound source and two ears in an existing space.

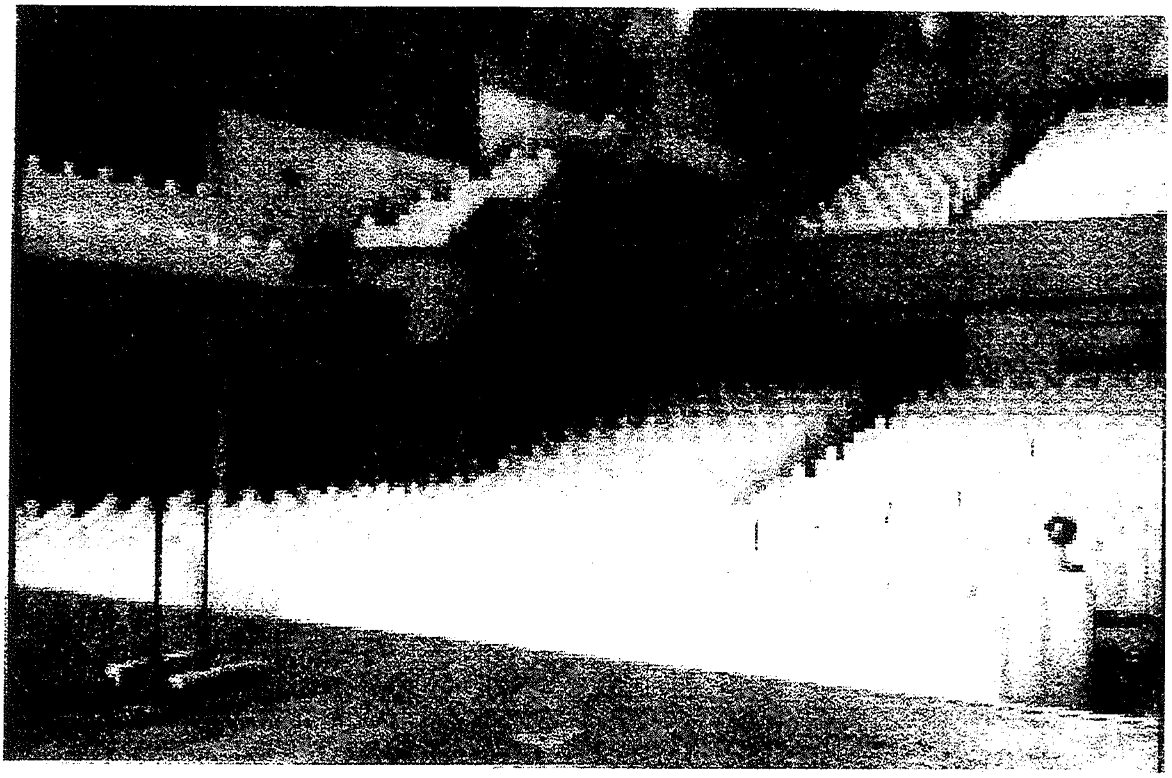


Fig. 2.2 Measurement of the transfer functions between a sound source and two ears in an acoustic model.

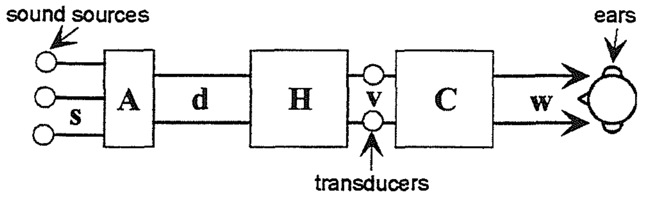


Fig. 2.3. Block diagram illustrating binaural synthesis over loudspeakers.

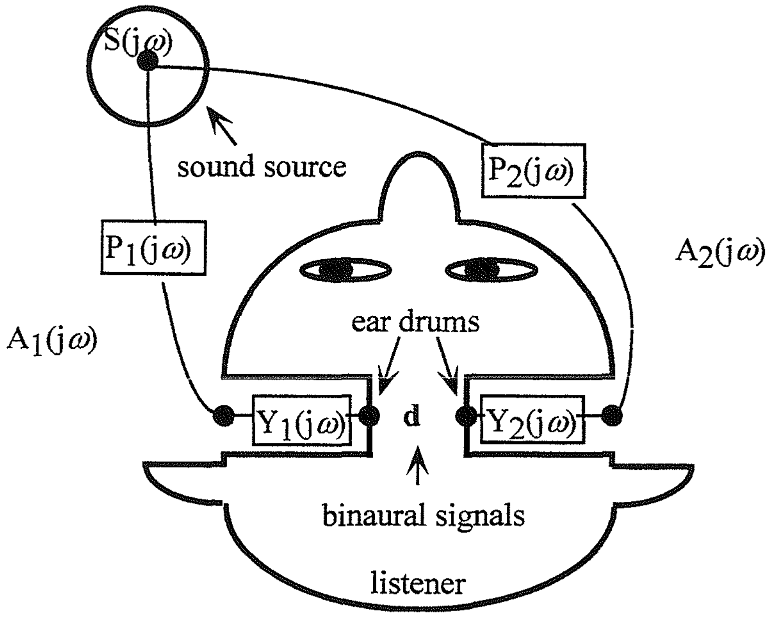


Fig. 2.4a A model of the acoustic system between a real sound source and the listener's ear drums.

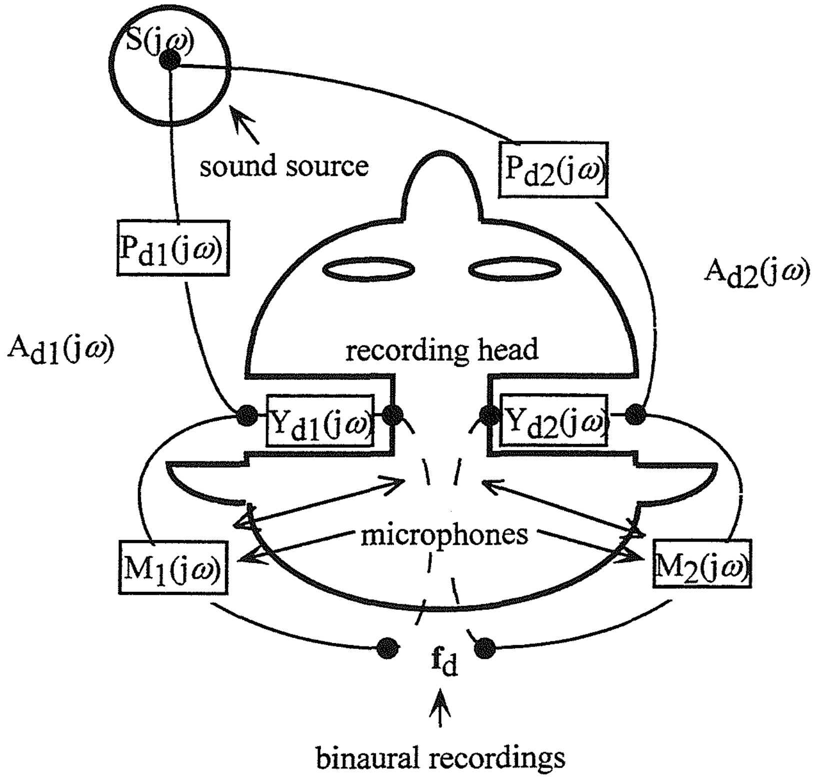


Fig. 2.4b A model of the acoustic system between a real sound source and the microphone output signals of the recording head.

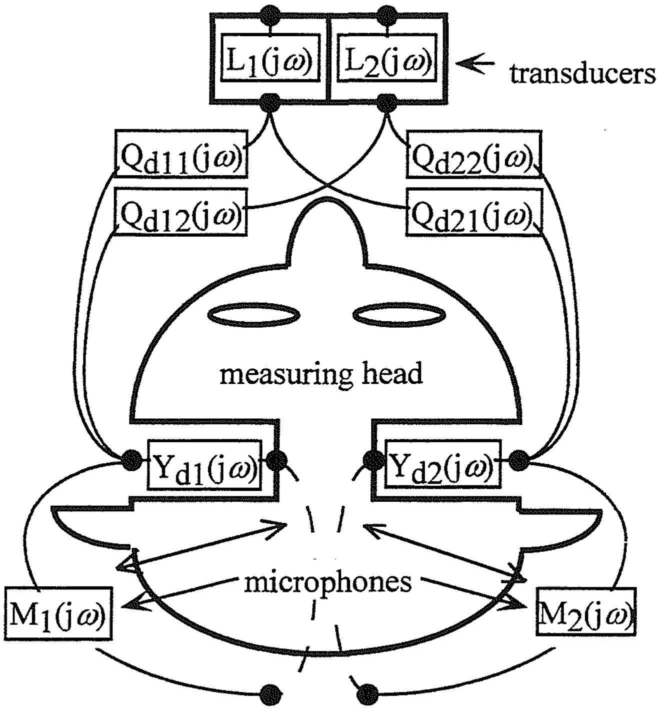


Fig. 2.4c A model of the acoustic system between an electro-acoustic transducer at the position of a sound source and the microphone output signals of the measurement head.

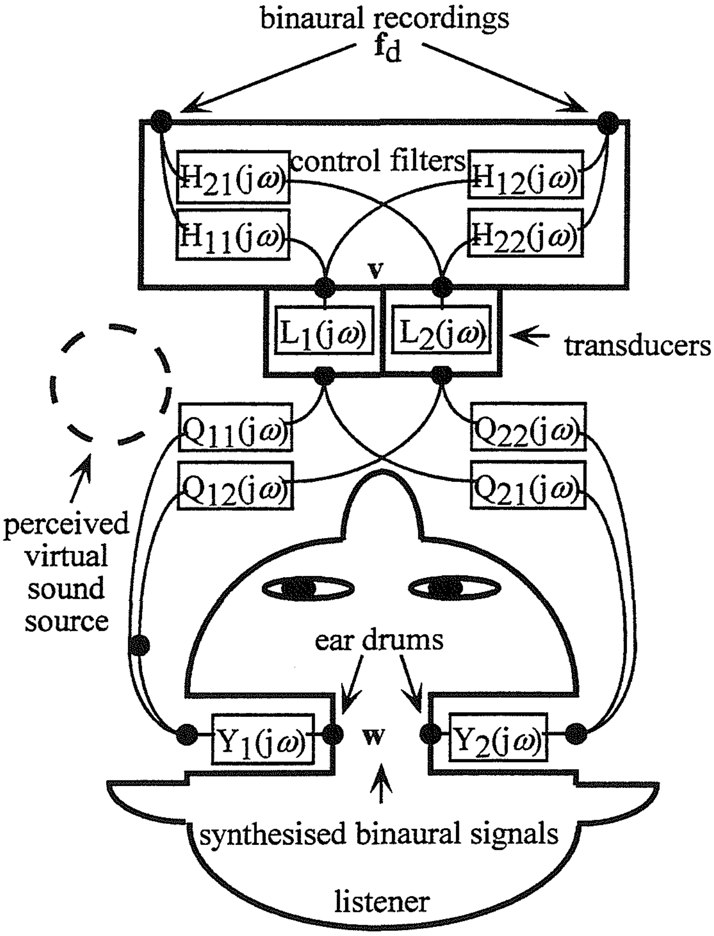


Fig. 2.4d A model of the acoustic system between the control transducers and the microphone output signals of the measurement head.

3 The Stereo Dipole system

3.1 Introduction

The location of transducers for binaural reproduction over loudspeakers has received little attention so far. Two transducers are usually placed symmetrically in front of a listener subtending an angle of about 60°. This conventional arrangement is probably adopted from the arrangement for the conventional stereophony that is still widely used to this day. The purpose of this Chapter is to point out that the binaural synthesis over loudspeakers can also be made to operate remarkably effectively, and arguably more effectively, by using a pair of loudspeakers that are placed very close together. Such a system is referred to as a "Stereo Dipole" [33][34]. Such a loudspeaker arrangement appears to have received little attention in the past, although it has been referred to in a recent paper [35]. There are also some very early experiments by Lauridsen [36], who used a combination of a conventional boxed loudspeaker together with an open-backed loudspeaker in an alternative scheme for stereophonic sound reproduction. It is shown that it is possible to achieve independent control of the sound signal at two ears with a monopole transducer and a dipole transducer at the same position. When two closely spaced monopole transducers are used, the sound field produced is a good approximation to that produced by a point monopole and a point dipole transducer up to a given frequency. Such an arrangement has the advantage that the transducer array is compact. In this chapter, the basic behaviour of the sound fields generated by virtual sound imaging systems are analysed. It is demonstrated that the use of two closely spaced loudspeakers also approximates such a source combination. Subjective experiments are also performed to establish the basic understanding of virtual spatial sound reproduction performance of such a system compared to a real sound field.

3.2 Principle of the "Stereo Dipole"

In this analysis, for simplicity, a simple case involving the control of two monopole receivers with two monopole transducers (sources) under free field conditions is considered here in order to understand the physics underlying binaural synthesis over loudspeakers. The effect of Head Related Transfer Functions (HRTFs) is also analysed in the later chapters as a more realistic plant. In such a case, the acoustic response of the human body (pinnae, head, torso and so on) also comes to influence the problem. The analysis in the following section can be found in reference [A 1].

3.2.1 The sound field radiated by two monopole transducers

A symmetric case with the inter-source axis parallel to the inter-receiver axis is considered for an examination of the basic properties of the system. The geometry is illustrated in Fig. 3.1. In the free field case, the plant transfer function matrix can be modelled as

$$\mathbf{C} = \frac{\rho_0}{4\pi}\begin{bmatrix} \mathrm{e}^{-\mathrm{j}kl_1}/l_1 & \mathrm{e}^{-\mathrm{j}kl_2}/l_2 \\ \mathrm{e}^{-\mathrm{j}kl_2}/l_2 & \mathrm{e}^{-\mathrm{j}kl_1}/l_1 \end{bmatrix} \tag{3.1}$$

where $\mathrm{e}^{\mathrm{j}\omega t}$ is assumed as time dependency with $k = \omega/c_0$. ρ_0 and c_0 are the density and sound speed.

The inverse of this matrix can be obtained analytically. The solution is that a desired value of signal is obtained at one of the ears, while the other ear receives no signal. Now consider the case

$$\mathbf{d} = \frac{\rho_0 e^{-jkl_1}}{4\pi l_1}\begin{bmatrix} D(j\omega) \\ 0 \end{bmatrix}$$

(3.2)

i.e., the desired signals for the ear 1 is the acoustic pressure signals which would have been produced by the closer control source alone whose values is D(jω) without disturbance due to the other source (cross-talk). A zero signal is desired at ear 2. This normalization also ensures a causal solution. When the ratio of and the difference between the path lengths connecting one source and two receivers are defined as $g = l_1/l_2$ and $\Delta l = l_2 - l_1$, the elements of **H** can be obtained from the exact inverse of **C** and can be written as

$$\mathbf{H} = \mathbf{C}^{-1} = \frac{1}{1 - g^2 e^{-2jk\Delta l}}\begin{bmatrix} 1 & -ge^{-jk\Delta l} \\ -ge^{-jk\Delta l} & 1 \end{bmatrix}$$

(3.3)

Therefore, the source outputs given by Eq. (2.7) can be written as

$$\mathbf{v} = \frac{D(j\omega)}{1 - g^2 e^{-2j\omega\tau}}\begin{bmatrix} 1 \\ -ge^{-j\omega\tau} \end{bmatrix}$$

(3.4)

where $\tau = \Delta l/c_0$. The denominator of this expression can be written in series form as follows

$$(1-x)^{-1}=\sum_{n=0}^{\infty}x^n \text{ (for } |x|<1)$$

and therefore, Eq (3.4) becomes as

$$\mathrm{v} = \mathrm{D}(\mathrm{j}\omega)\sum_{n=0}^{\infty}g^{2n}\mathrm{e}^{-2n\mathrm{j}\omega\tau}\begin{bmatrix}1\\ -\mathrm{g}\mathrm{e}^{-\mathrm{j}\omega\tau}\end{bmatrix} \tag{3.5}$$

As described in [14], the solution for the required source strengths has a recursive structure within. In the time domain, the solution can be written in the form

$$\begin{aligned} v_1(t) &= d(t)*[1+g^2\delta(t-2\tau)+g^4\delta(t-4\tau)+\ldots] \\ v_2(t) &= -gd(t-\tau)*[1+g^2\delta(t-2\tau)+g^4\delta(t-4\tau)+\ldots] \end{aligned} \tag{3.6}$$

where the asterisk denotes convolution. If, for example, $d(t)$ is a pulse the duration of which is short compared to the delay τ, control source 1 first emits a pulse $d(t)$ that travels to ear 1 to give the desired signal $d_1(t)$. This pulse then arrives at ear 2 but is cancelled by the pulse $-gd(t - \tau)$ that has been emitted from control source 2. This pulse however, causes an unwanted pulse at ear 1. This in turn is cancelled by the pulse $g^2d(t - 2\tau)$ emitted from control source 1, and so on. This process is illustrated in Fig. 3.2 which shows a sequence of "snapshots" of the instantaneous pressure field produced by the two control sources when the system is trying to synthesise a hanning pulse given by

$$d_1(t)=\begin{Bmatrix} 0 & t\prec 0, t\succ 2\pi/\omega_0 \\ (1-\cos\omega_0 t)/2 & 0\le t\le 2\pi/\omega_0 \end{Bmatrix} \tag{3.7}$$

as desired signals where ω_0 is 6.400π (which implies that the first zero in the spectrum is at 6.4kHz). The intervals of the illustrations are $0.1/c_0$. Each illustration is calculated over an area of 1m × 1m. The control sources are spanned to give the source span Θ of 60° and Δr = 0.18m.

3.2.2 The sound field produced by a monopole transducer and a dipole transducer

When the pair of control sources are placed close together, the spatial aspect of sound field generated in the region of the listener's head becomes considerably simple while still achieving the same objective. In this case, the two control sources can be used to synthesise a close approximation to the sound field that would be produced in the region of the listener's head by the superposition of a point monopole type control source and a point dipole type control source. These two equivalent control sources are placed at the mid-point of the two monopole type control sources. The strengths of these two equivalent sources are those necessary to produce the desired ear signals. This can be found by evaluating the monopole and dipole moments associated with the two sources in the limit $\omega\tau \rightarrow 0$ and $g \rightarrow 1$.

The monopole moment is given by

$$v(\mathrm{j}\omega) = v_1(\mathrm{j}\omega) + v_2(\mathrm{j}\omega) = D(\mathrm{j}\omega)\left(1 - g\mathrm{e}^{-\mathrm{j}\omega\tau}\right)/\left(1 - g^2\mathrm{e}^{-2\mathrm{j}\omega\tau}\right) \tag{3.8}$$

Since the denominator of this expression can be written as $\left(1 - g\mathrm{e}^{-\mathrm{j}\omega\tau}\right)\left(1 + g\mathrm{e}^{-\mathrm{j}\omega\tau}\right)$, this expression becomes

$$v(\mathrm{j}\omega)=D(\mathrm{j}\omega)/\left(1+g\mathrm{e}^{-\mathrm{j}\omega\tau}\right) \tag{3.9}$$

Using the series expansion $\mathrm{e}^x=1+x+x^2/2+\ldots$, in the limit $\omega\tau \to 0$ and g → 1, Eq. (3.9) becomes

$$v(\mathrm{j}\omega)\approx D(\mathrm{j}\omega)/2 \tag{3.10}$$

Similarly, the dipole moment can be written as

$$f(\mathrm{j}\omega)=\frac{\rho_0\Delta s}{2}\left(v_1(\mathrm{j}\omega)-v_2(\mathrm{j}\omega)\right)=\frac{\rho_0\Delta s}{2}D(\mathrm{j}\omega)\left(1+g\mathrm{e}^{-\mathrm{j}\omega\tau}\right)/\left(1-g^2\mathrm{e}^{-2\mathrm{j}\omega\tau}\right) \tag{3.11}$$

Again using a series expansion of the denominator, together with the approximation $|\Delta l|=|\Delta s|\sin\alpha$, where the angle α is defined in Fig. 3.1, in the limit $\omega\tau \to 0$ and g → 1, Eq. (3.11) becomes

$$f(\mathrm{j}\omega)\approx\frac{\rho_0 D(\mathrm{j}\omega)}{2\sin\alpha}\left(\frac{1}{l}+\frac{\mathrm{j}\omega}{c_0}\right)^{-1} \tag{3.12}$$

It is assumed that the acoustic wavelength is much larger than the path length difference Δl between one of the control sources and the ears.

Identical results can be derived by assuming that the two control sources used to produce the same desired ear signals are a superposition of point monopole and dipole sources. The sound field produced when the combination of point monopole and dipole sources is used to produce the same desired ear signals is shown in Fig. 3.3. The sound field produced has a far less complex spatial behaviour than that produced by the widely spaced monopole type control sources. It is possible to achieve the desired objective with a single pulse emitted by this type of control source combination.

3.3 Aspects of the system with a finite source separation

3.3.1 High frequency limit for approximation

Since practical transducers have a certain size, there is a physical limit to placing them close together. Therefore, the combination of two monopole transducers with a finite source separation is capable of producing this form of sound field over only a limited frequency range. Two monopole transducers can approximate one monopole transducer or one dipole transducer when Δs in Fig. 3.1 is small compared to the wavelength. This is expressed as following.

$$f << \frac{c_0}{\Delta s} \tag{3.13}$$

In a typical example of the closely placed case, in which the sources span 10° with the transducers being 1.4m away from the listener's head, $c_0/\Delta s$ is only about 1.5kHz. Therefore, this approximation holds only up to a few hundred Hertz (Fig. 3.4). However, if the sound field control is limited to the relatively small region perpendicular to the dipole axis, which with this arrangement is around the direction of the listener's ears, two monopole transducers can approximate one monopole transducer or one dipole transducer when $\Delta s \sin\alpha$ is small compared to the wavelength. This is expressed as

$$f << \frac{c_0}{\Delta s \sin\alpha} \tag{3.14}$$

For the same example as before, the frequency limit goes up to about 5kHz (Fig. 3.5).

3.3.2 Time domain structure

When the monopole and dipole combination is approximated by the two monopole transducers with finite separation distance, the time difference τ always remains finite and the recursive behaviour of the sources described by Eq. (3.6) is always present, except that the time duration between successive source output pulses changes as the path length difference Δl changes. This time scale τ is given by the time at which the sound takes to travel the distance Δl. As shown in Fig. 3.6, τ becomes smaller as the source span becomes smaller. On the other hand, each successive pulse reduces its amplitude from the previous pulse by the factor of g. The reduction of amplitude of the successive pulses also becomes smaller as the source span becomes smaller as shown in Fig. 3.7. Therefore, as a result, the time for these successive pulses to converge (the amplitude becomes small enough to be truncated without perceptual consequences), i.e., the necessary length of the inverse filters, stays roughly the same regardless the source span. The convergence time to –60dB of the initial pulse, which is directly related to the necessary length of the inverse filters is plotted in Fig. 3.8. It is seen to be about 30ms regardless the control source span, which is about 1200 ~ 1400 samples at a sampling frequency of 44.1kHz. However, once these pulses are emitted into the sound field, the frequency components within the high frequency limit which has large wavelength compared to the separation of the successive pulses are cancelled by the pulses from the other transducer. As a result, a single pulse similar to the monopole-dipole combination is formed by the two monopole sources within a certain frequency range. The simple sound field spatially as seen in Fig. 3.3 can be obtained when the power spectrum of the Hanning pulse is concentrated below the high frequency limit.

3.4 Subjective evaluation

The ability to convey spatial information through the use of the "Stereo Dipole" system was investigated by using subjective localisation experiments. As the very basic components of a virtual sound environment, generation of a single incident sound wave is taken as an example here. Therefore, the experiments were carried out in an anechoic chamber. Experiments with real sound sources were also performed to establish the accuracy of the experimental procedure itself. As a comparison, another arrangement of transducers spanning 60° is investigated. Source directions on the horizontal plane were chosen to be examined since this covers the whole range of azimuth directions and two alternative elevation directions, i.e. 0° (front) and 180° (rear), in each cone of constant azimuth.

3.4.1 Procedure

An initial experiment with pink noise as a test signal showed that the high frequency components resulted in the dominant components perceptually. Listeners heard very loud high frequency noise at the control transducer position while lower amplitude mid-frequency sound were localised in space virtually. The consequence of large high frequency discrepancy between the HRTFs of the subjects and the KEMAR [37] HRTFs used in the filter design procedure (Section 2.4) were suspected to be the cause. In order to minimise this, a weighted noise signal (EAIJ RC-7603) was used as source signal. The signal has a flat spectrum between 200Hz and 2kHz and gradually rolls off towards lower and higher frequencies (Fig. 3.9). The relative level is about -2dB at 5kHz, -5dB at 10kHz, -13dB at 20Hz and 20kHz with respect to the level between 200Hz and 2kHz. Each stimulus consisted of a reference signal and a test signal. A reference signal was presented at 0° azimuth and 0° elevation, i.e., directly in front of the listener before each test signal. Both signals had the same sound source signal with duration of 3 seconds for the reference signal and 5 seconds for the test signal with a gap of 3 seconds in between. In order to avoid the effect of presentation order, the order of presentation from different directions was randomised. The reference stimulus not only cancelled the order effect, but also gave subjects prior knowledge of the sound source signal spectrum that is important for the monaural spectral cue. The stimuli, consisting of a set of reference and test signals, were repeated when subjects had difficulty in making a judgement.

The spherical co-ordinate system used to define the direction of the perceived sound sources and of the transducers is shown in Fig. 3.10. Two different transducer arrangements are investigated for comparison. In both cases, two transducers are placed in front of the listener on the horizontal plane (0° elevation) and aligned symmetrically with respect to the median plane. The transducers positioned spanning 60° as seen by the listener (±30° azimuth) are representative of the popular arrangement. The span of 10° (±5° azimuth) represents close spacing, the "Stereo Dipole". The loudspeakers as control transducers and as real sound sources were placed 1.4m from the origin of the co-ordinate system (Fig. 3.11). The precision of the arrangement of the loudspeakers and listener's head was of the order of ±10mm. The loudspeakers used had a fairly flat response between about 250Hz and 5kHz which gradually rolls off towards lower and higher frequencies (Fig. 3.12). The relative level at 20kHz is about 10dB smaller with respect to the frequency that gives maximum response.

Subjects were required to choose the closest marker to the perceived direction of sound. The markers were placed all around the head in the horizontal plane 1m from the origin of the co-ordinate with 10° intervals (Fig. 3.11). It was found in preliminary experiments that the subjects can produce a large error when they report a direction without seeing the reference marker. The magnitude of the error in reporting the direction is as large as 40° especially when the direction is in the rear. The visible reference marker reduces this error down to about 5° at the expense of increasing visual related error mainly in the front hemisphere where localisation accuracy is much finer than 5°. The subjects were allowed to choose more than one marker when they perceived two or more separate directions of sound. In order to avoid introducing dynamic cues that relate to head movement, the subject was instructed not to move the head nor body while the stimuli were presented. However, the subject was allowed to turn his head to see markers after each test stimulus had stopped. The subject's head was not physically fixed but supported by a small headrest. The subject was surrounded by a thin black curtain placed between markers and loudspeakers in order to minimise the effect of visual information (Fig. 3.11). Subjects were all European males with normal hearing function.

The results from the subjective experiments are presented in the following format. The area of each circle in the figures is proportional to the number of subjects who perceived the source to be in the given direction. In cases where the subjects perceived sound sources in more than two directions, the area of the circle is distributed into those positions in accordance with the number of the responses. The dash-dot line shows the position of the circles when the perceived direction is the same as the presented direction. The dotted line is in a symmetric position to the dash-dot line with respect to the interaural axis. Therefore, the subjective responses due to front-back confusion fall around these lines.

3.4.2 Real sound sources

Nine loudspeakers as real sound sources were placed at 10° increments at different azimuthal angles except ±20° and ±90°, and two sources were placed at azimuth 0° with different elevations of 0° and 180° (front and rear). Five of them were positioned in front (elevation 0°) and four of them were positioned in the rear (elevation 180°). Four of them were positioned to the left (negative azimuth) and five of them were positioned to the right (positive azimuth). The sound source signal was recorded on a DAT recorder and played back during the experiment via an amplifier to each loudspeaker.

The performance with real sound sources (Fig. 3.13) shows the localisation performance of the subjects and the accuracy of the experimental procedure itself. This therefore implies the maximum precision achievable with the following experiments with synthesised virtual sound sources. More than 60% of the responses resulted in the correct marker being chosen and more than 90% of the responses resulted within the smallest (±10°) measurable error with the method. The judgements are more accurate for smaller azimuth directions than for larger azimuth directions. The repeatability of the response is exceptionally good in that the responses associated with a particular direction for a particular subject almost always (more than 95%) resulted at the same marker (even for the wrong marker). The accuracy can be observed best at small azimuth directions (closer to the median plane) and deteriorates towards large azimuth directions (the side of the listener). There are no obvious signs of confusion along the cone of constant azimuth, i.e., front-and-back confusion. The subjects reported after the experiments that the task was very easy and did not have any ambiguity in deciding which marker to choose.

3.4.3 Virtual sound sources

Localisation experiments with binaural synthesis over loudspeakers were carried out. For the virtual source localisation experiment, the physical acoustic paths **a** and **C** are modelled with free field (absence of any effects other than head) head related impulse responses (HRIRs: the time domain representation of HRTFs). A database comprising directionally discrete HRIRs on a virtual spherical surface 1.4m from a KEMAR dummy head was obtained from MIT Media Lab [38]. Each HRTF is the result of a measurement in an anechoic chamber at a sampling frequency of 44.1 kHz. The "compact" data set was used. (The researchers at MIT deconvolved the loudspeaker response from the data. The process induced a phase distortion. In addition, different microphones were used for HRTF measurements and the transducer response measurements. Therefore, discrepancy of microphone responses remains within the data.) The control filter matrix **H** is determined by the frequency domain deconvolution method described in Section 2.4. The frequency response of the inverse filters **H** and the effectiveness of the independent control at each ear **X** is shown in Fig. 3.14 and Fig. 3.15. A reasonably good control performance is achieved between around 1kHz and 8kHz. Computationally, where the largest source of error is rounding error and more than 180dB of dynamic range is available, it is possible to design an inverse filter matrix whose control effect looks much better. However, in practice, such filters prove to be more harmful than effective, and result in numerous problems such as severe colouration, distortion, low signal to noise ratio, etc. This matter is analysed and treated in more detail in Chapter 7. For the time being, regularisation is used to find the fine balance in this trade-off. The inverse filters were implemented by digital filters using an MTT Lory Accel digital signal processing system. Each filter had 1600 coefficients at a sampling frequency of 44.1 kHz. The output of the digital filters were recorded on a DAT recorder and played back during the experiment via a 2-channel amplifier to two pairs of loudspeakers.

The two control transducers are placed in front of the listener on the horizontal plane (0° elevation) and aligned symmetrically with respect to the median plane. The control transducers positioned spanning 60° as seen by the listener (±30° azimuth) are representative of a standard arrangement as in Stereophony. The span of 10° (±5° azimuth) represents the “Stereo Dipole”. Fig. 3.16 shows the difference of response of the two pairs of loudspeakers. The characteristics of the pair of control transducers were well-matched (0.5dB difference in amplitude and a few degrees difference in phase response). Therefore, it is equivalent to synthesising a virtual loudspeaker as in Eq.(2.38) despite the fact that the loudspeaker responses were not deconvolved. Subjects localise the position of a virtual loudspeaker, but not an ideal monopole sound source. Nevertheless, the loudspeaker pairs for different transducer arrangements were swapped for half of the subjects with the aim of minimising bias errors which are induced by different responses between the loudspeakers. Sixteen virtual sound sources were placed at 0°, ±20°, ±40°, ±60° and ±80° azimuth with 0° elevation (front) and 0°, ±20°, ±40° and ±70° azimuth with 180° elevation (rear).

Fig. 3.17 and Fig. 3.18 show the localisation performance for 11 subjects. The localisation performance is much poorer than that for the real sound sources. The localisation in azimuth is again more accurate for smaller azimuth than larger azimuth.

However, azimuth localisation error in general is much larger than the localisation of real sound sources.

It was also revealed that there was a population of subjects for whom the synthesis of virtual sound sources works reasonably well ("good" subjects) whereas it does not work so effectively for the rest of the subjects ("poor" subjects). Only a few front-back confusions can be observed with the 7 "good" subjects. However, the 4 "poor" subjects did not localise the virtual sound sources in the rear half of the horizontal plane correctly, and instead, localised them around symmetric positions in the front. Moreover, virtual sound sources at large azimuth directions (around ±90° azimuth) were perceived at the much offset position systematically towards the centre (smaller azimuth angle).

The reason why this happens is not certain. However, there are a couple of possible hypotheses. One is that the HRTFs used to design the control filters **H** are very different from these subjects' own HRTFs. As a result, sound signals at the ears are not synthesised well enough for their auditory system to localise sound source well. Another is that these two groups of subjects make use of or give an importance to different information to distinguish the sound source at the front and the back. Then, the information which is used by the "good" group is synthesised well but that used by the "poor" group is not. However, it is clear that the grouping of subjects has no relation to the different transducer span. It also has no relation to the ability of the subjects to localise real sound sources (Subjects in Fig. 3.13 includes both from Fig. 3.17 and Fig. 3.18). This aspect is investigated further and described later in Chapter 4.

In principle, different control transducer arrangements should not produce much difference in performance when the listener's head is at the intended position. Nevertheless, the 10° control transducer span showed slightly better performance for "good" subjects around 0° azimuth where it showed no front-back confusion, contrary to the considerable amount of confusions with the 60° span. The increase in front-back confusion with the 60° control transducer span suggests that spectral cues may have been degraded. The 60° span transducer arrangement has a slight advantage in azimuth localisation, both by the "good" subjects and by the "poor" subjects. This observation may accord with the better control performance at the lower frequency region by the 60° control transducer arrangement which is important for the synthesis of time related cues. These aspects are investigated further in Chapter 5.

3.5 Conclusions

A combination of a monopole source and a dipole source is shown to work as a control transducer for binaural synthesis over loudspeakers. It is also demonstrated that the use of two closely spaced monopole sources can approximate such a source combination. Subjective experiments are performed to establish the basic understanding of virtual spatial sound reproduction performance of such a system compared to a pair of conventional monopole control transducers widely spaced. Both types of system performed equally well in terms of presentation of spatial information, with different advantages and disadvantages. The localisation performance is considerably worse with both systems compared to real sound sources. The most significant observation is a difference in performance among subjects rather than between different type of systems. The elevation localisation error (front-back confusion) was significant. The error was biased in that more rear images are perceived in front than vice-versa.

Related publications

[A 1] P. A. Nelson, O. Kirkeby, T. Takeuchi, and H. Hamada, "Sound fields for the production of virtual acoustic images," J. Sound. Vib. **204** (2), 386-396 (1997).

[A 2] O. KIRKEBY, P. A. NELSON, T. TAKEUCHI and H. HAMADA, "Sound Fields Generated By Virtual Source Imaging Systems", Proceedings of the Active 97, 941-954 (1997)

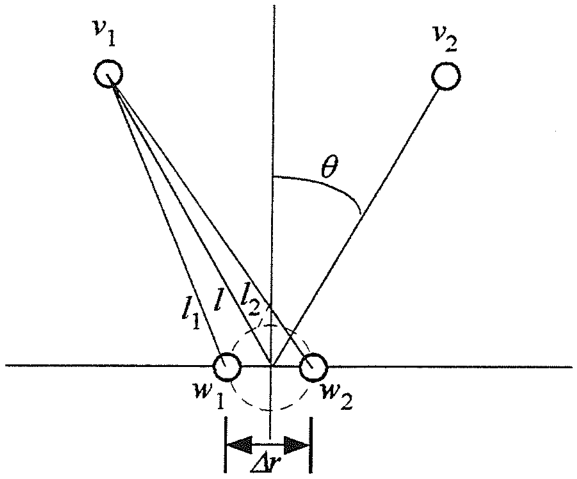


Fig. 3.1 Geometry of the virtual source imaging system .

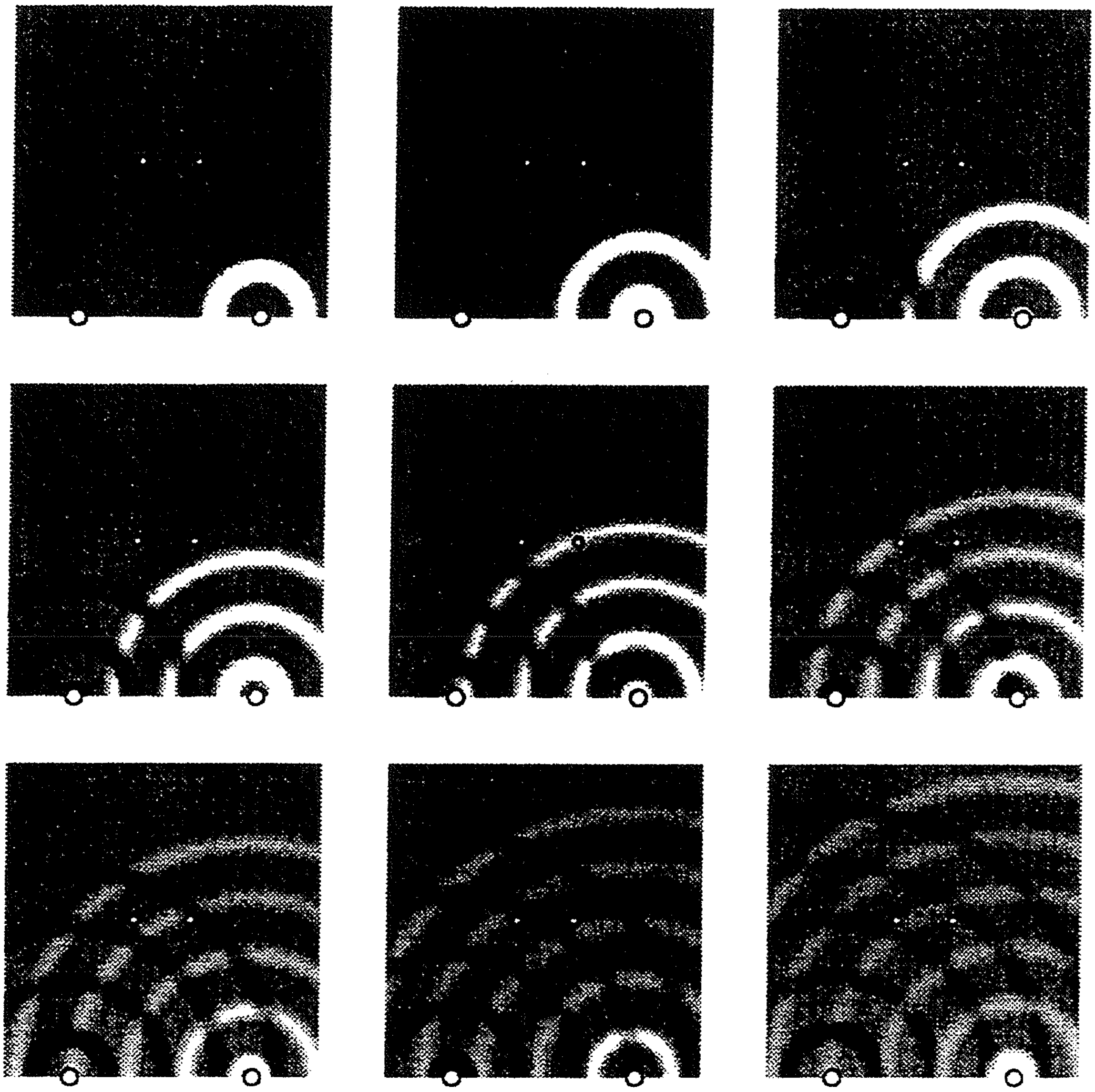


Fig. 3.2 A series of illustrations of the sound field produced when two point monopole sources are used to generate a desired pulse at ear 1 while producing zero pressure at ear 2. Each illustration depicts the magnitude of the acoustic pressure on a grey scale, with lighter shading denoting positive pressures and darker shading denoting negative pressures.

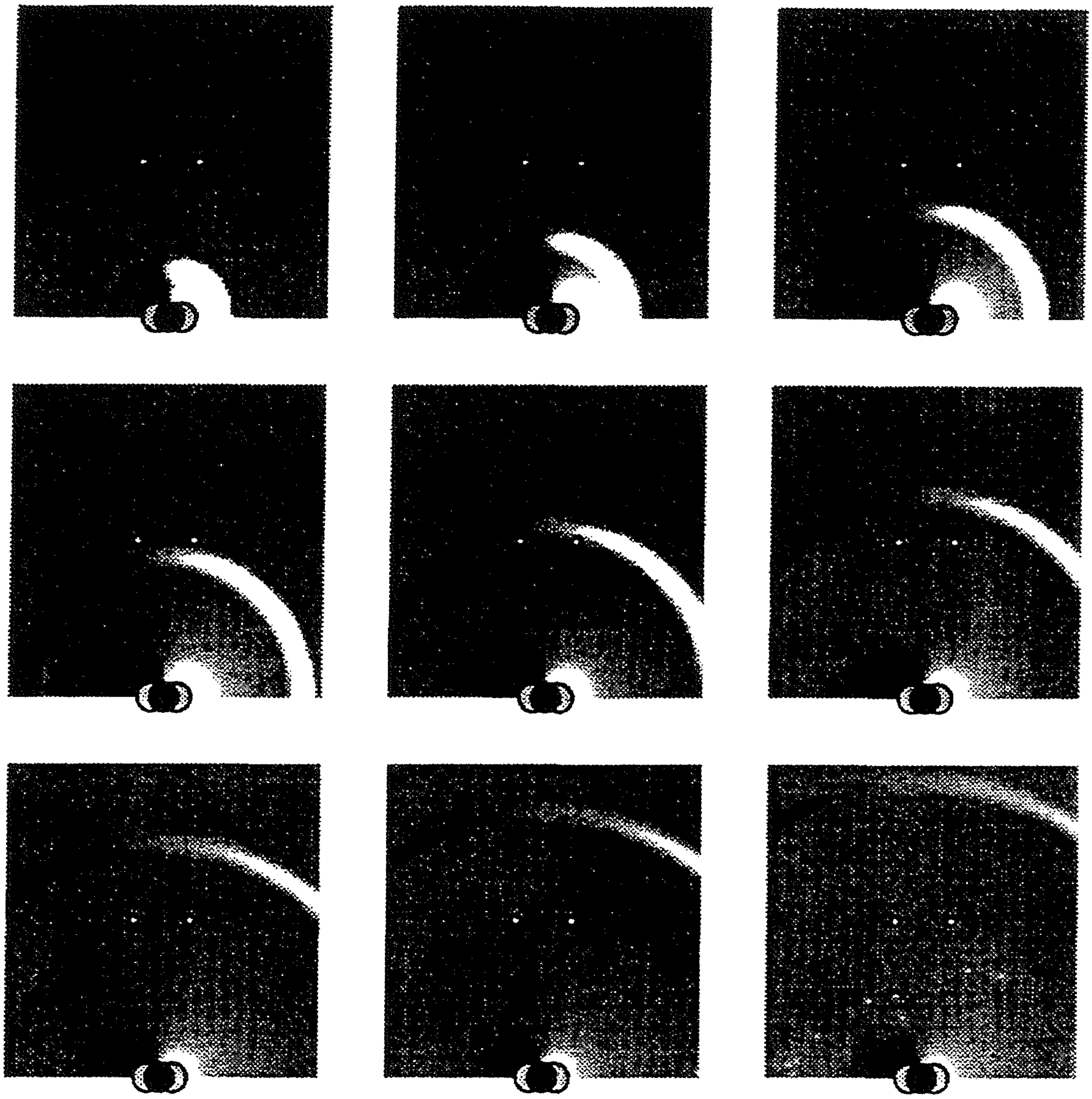


Fig. 3.3 A series of illustrations that are equivalent to those shown in Fig. 3.2, except that the desired field is produced by using the superposition of a point monopole and a point dipole.

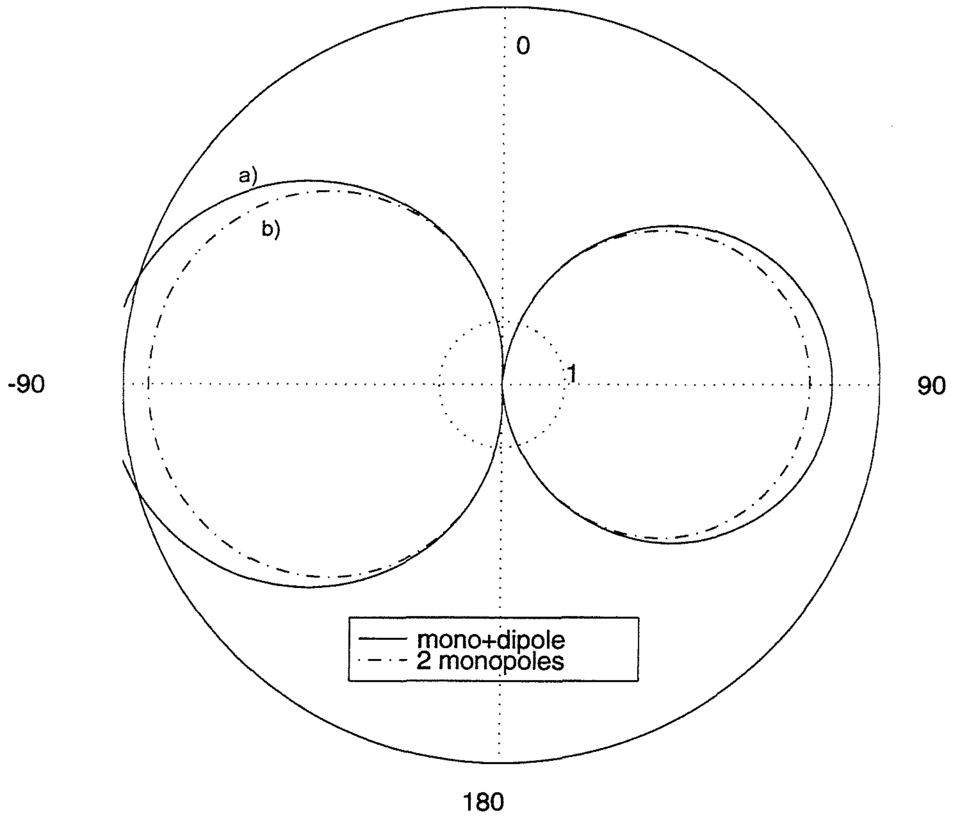


Fig. 3.4 Directivity pattern of transducer pairs. a) monopole and dipole combination. b) approximation by two monopole transducers ($f \ll c_0/\Delta s$).

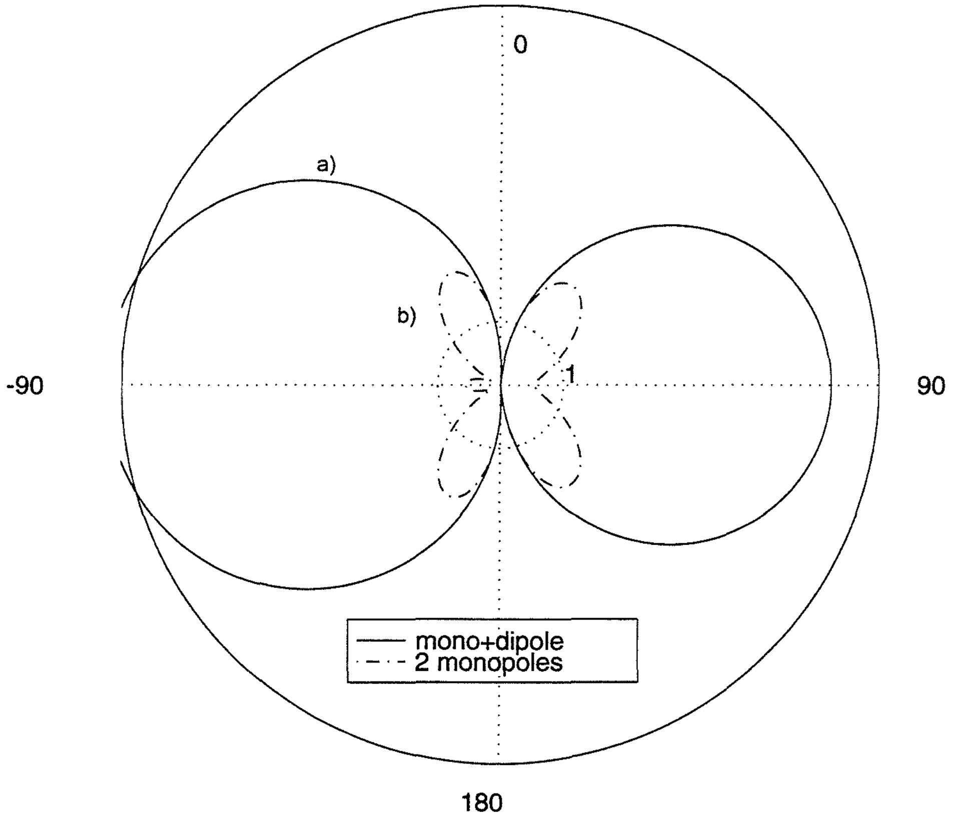


Fig. 3.5 Directivity pattern of transducer pairs. a) monopole and dipole combination. b) approximation by two monopole transducers ($f \ll c_0/\Delta s \sin\alpha$).

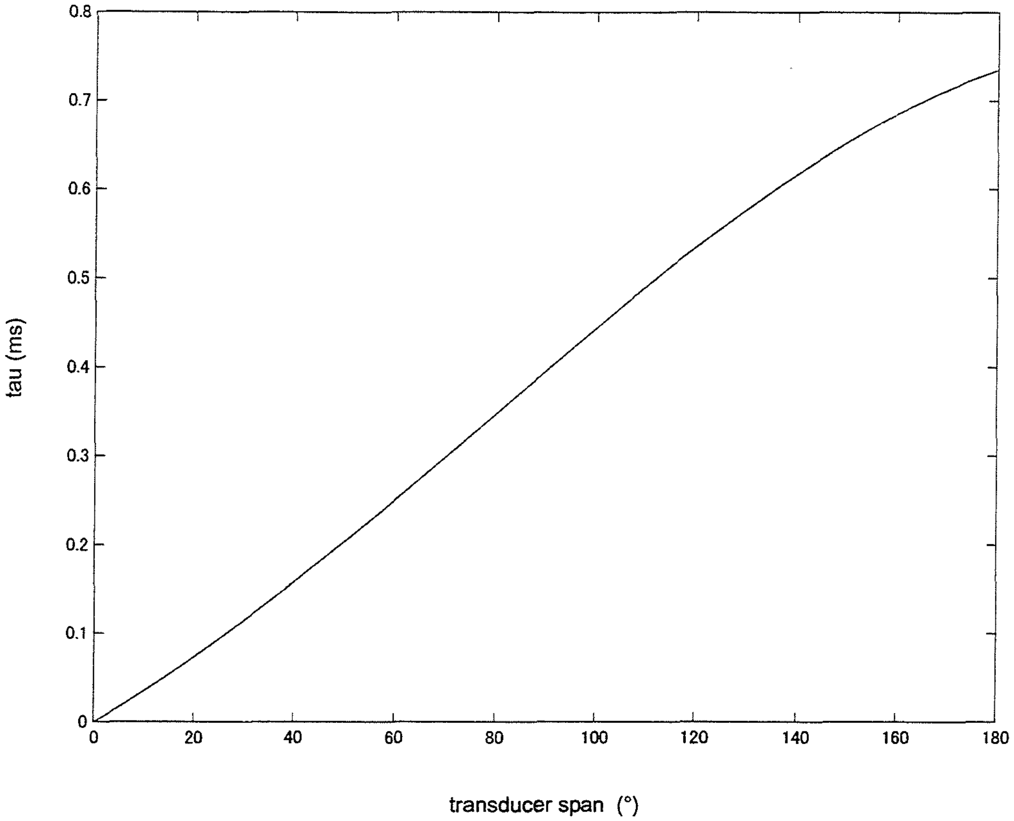


Fig. 3.6 Time difference *t* of the successive pulse as a function of the source span.

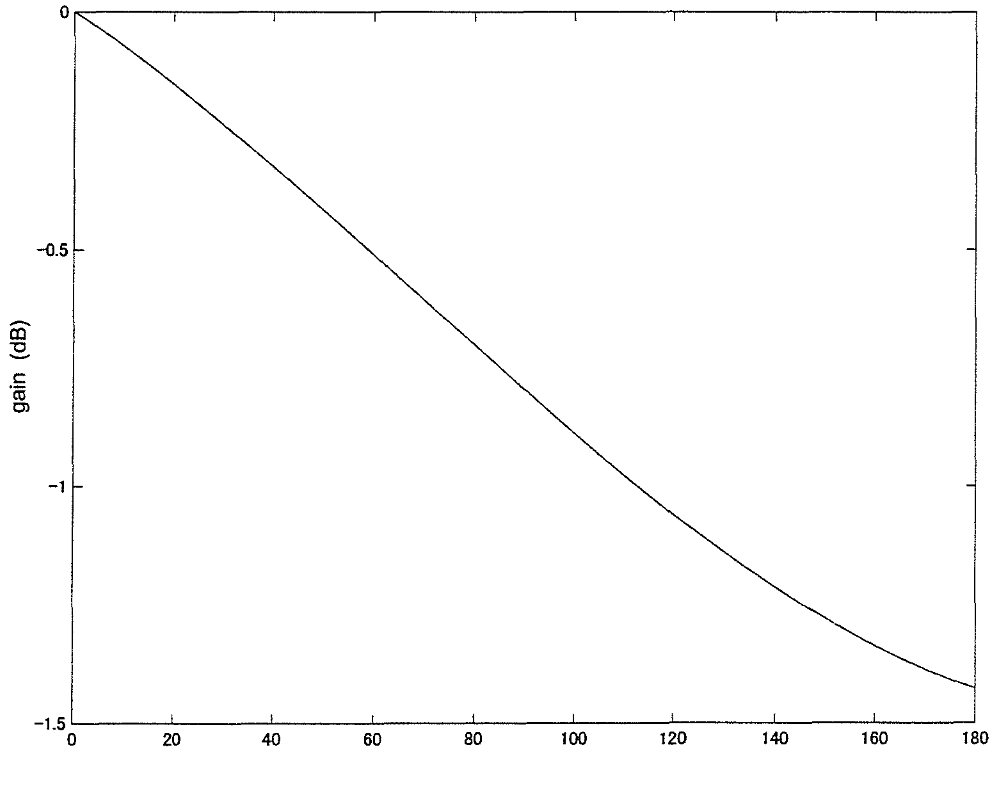


Fig. 3.7 Reduction of amplitude of the successive pulse as a function of the source span.

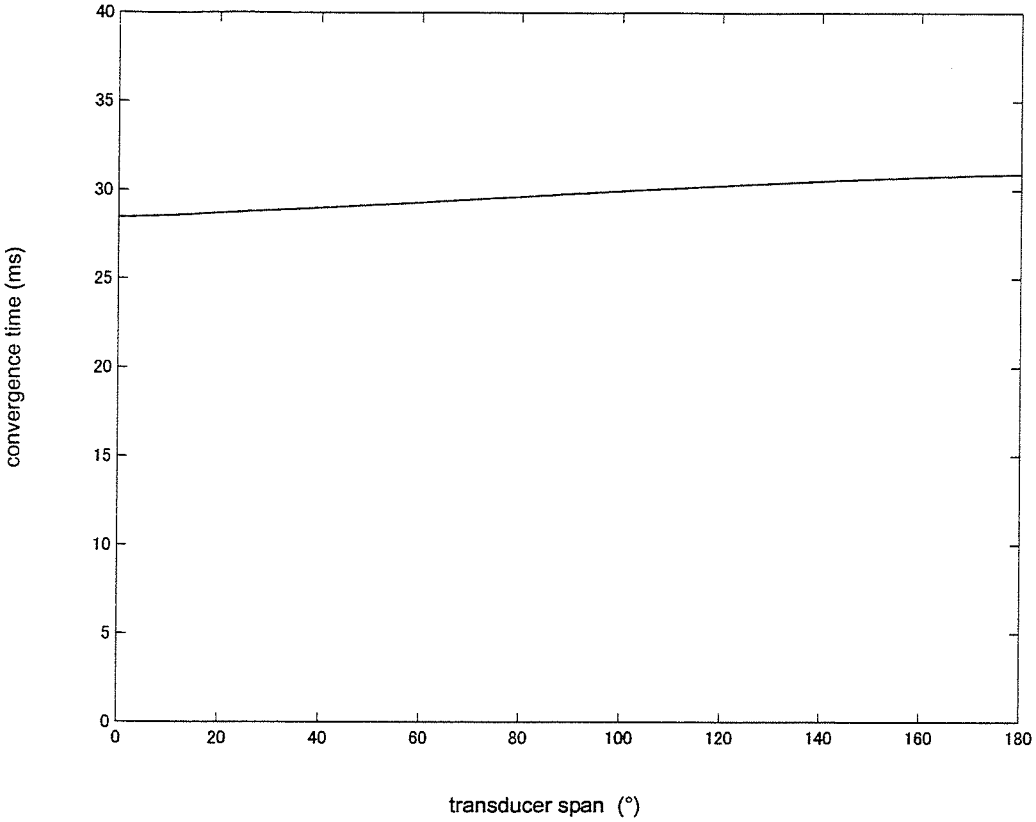


Fig. 3.8 Convergence time of inverse filters (-60dB) as a function of the source span.

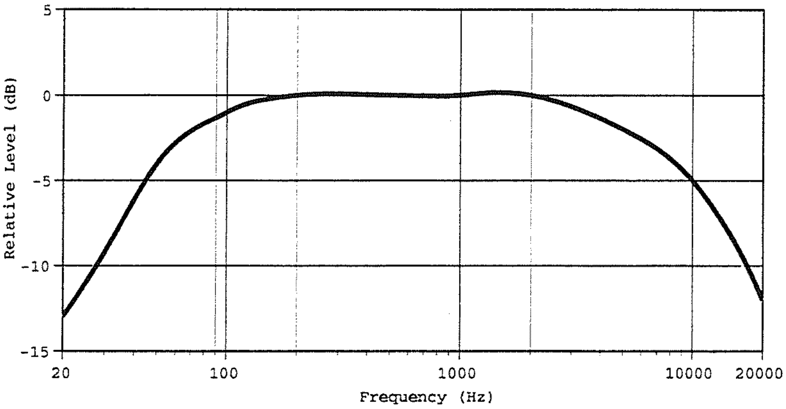


Fig. 3.9. The spectrum of weighted noise signal (EAIJ RC-7603) used as source signal.

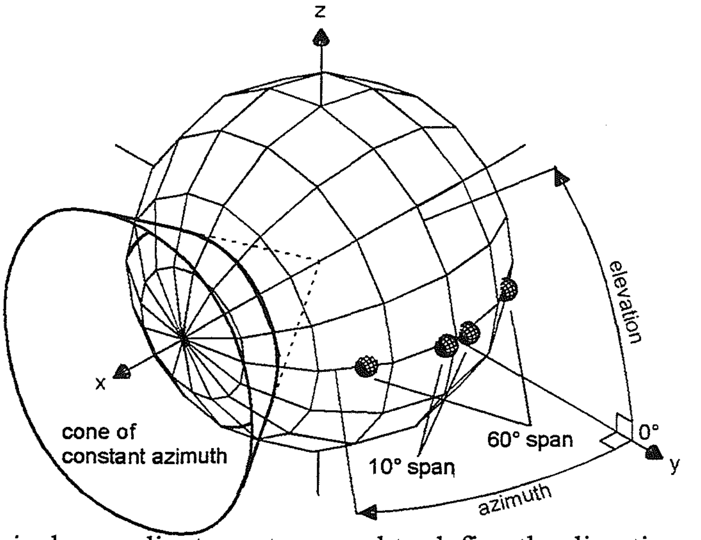


Fig. 3.10 The spherical co-ordinate system used to define the direction of sound sources relative to the listener's head position and orientation. An example of "cone of constant azimuth" is illustrated. The two different transducer arrangements investigated are also shown (relative to the optimal head position and orientation).

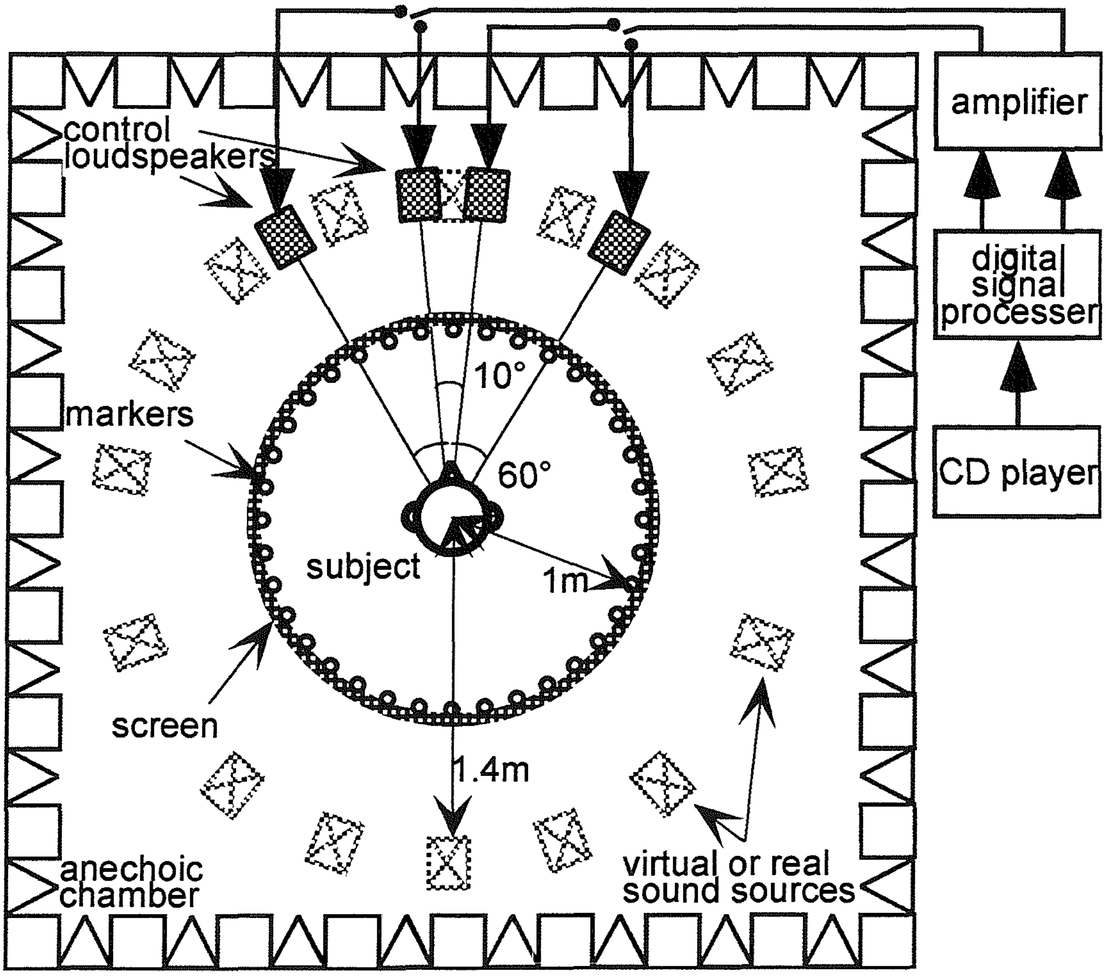


Fig. 3.11. The arrangement for the subjective experiment.

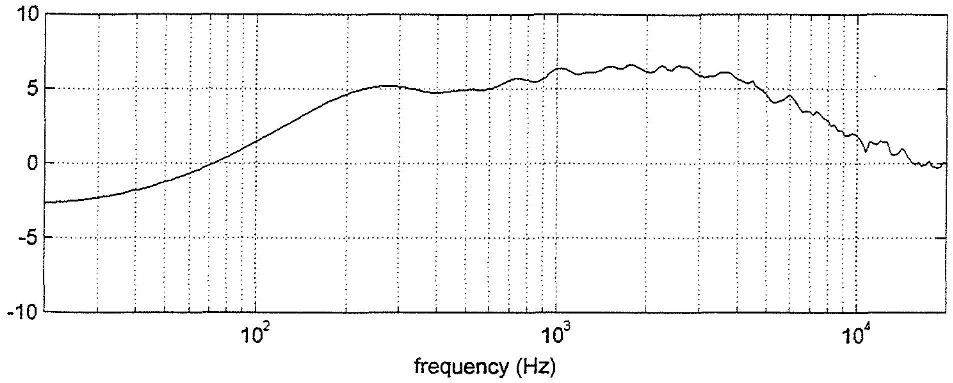


Fig. 3.12. The frequency response of one of the loudspeaker.

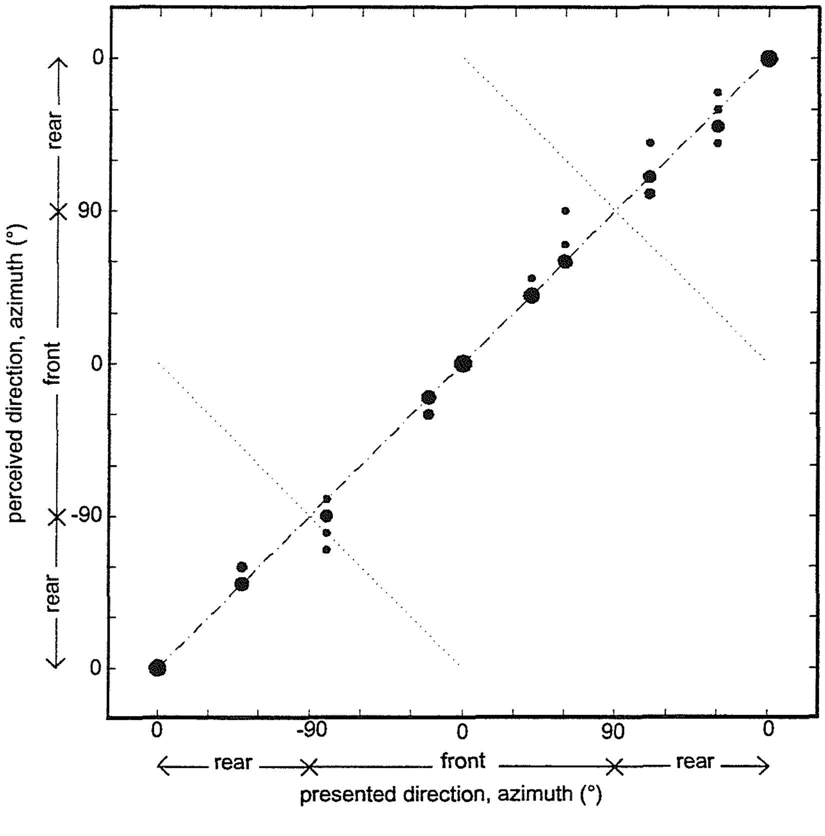


Fig. 3.13. Results of the subjective experiment for localising real sound sources. 6 subjects were tested.

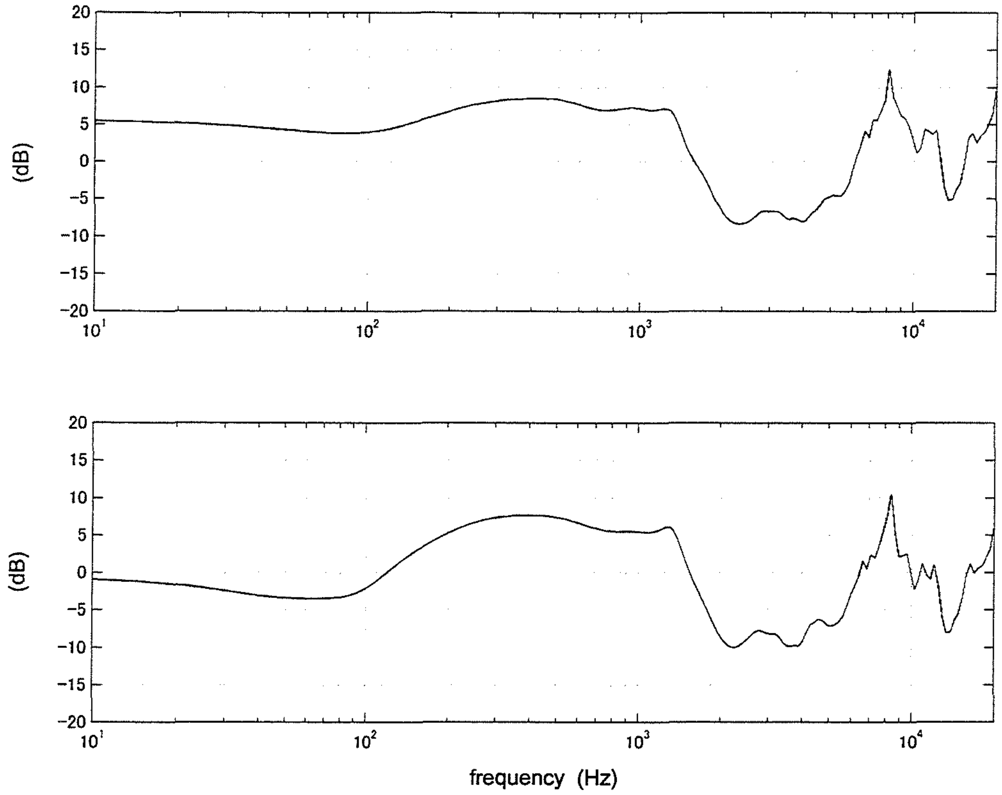


Fig. 3.14 Inverse filter matrix response. Top) diagonal term. Bottom) non-diagonal term.

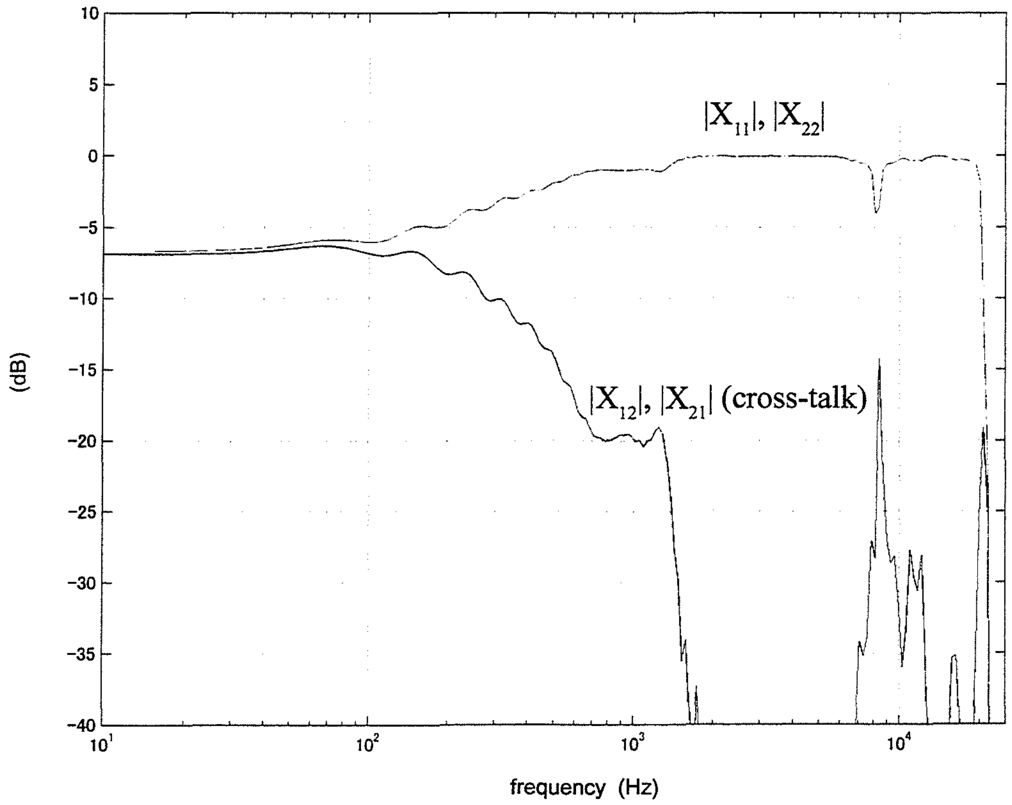
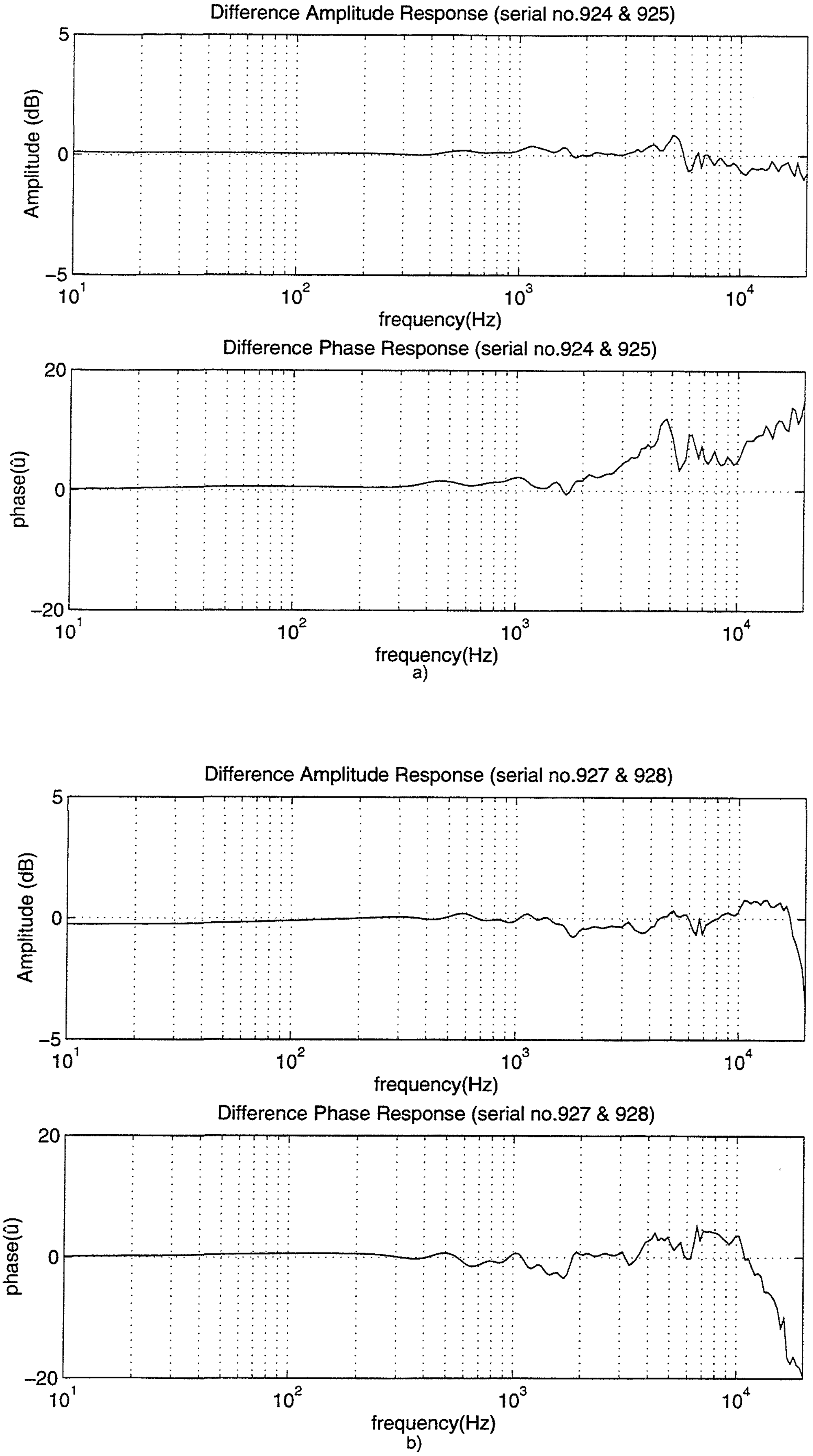


Fig. 3.15 Control performance.



a)

b)

Fig. 3.16 The difference of response of two pairs of loudspeakers. a) Pair 1. b) Pair 2.

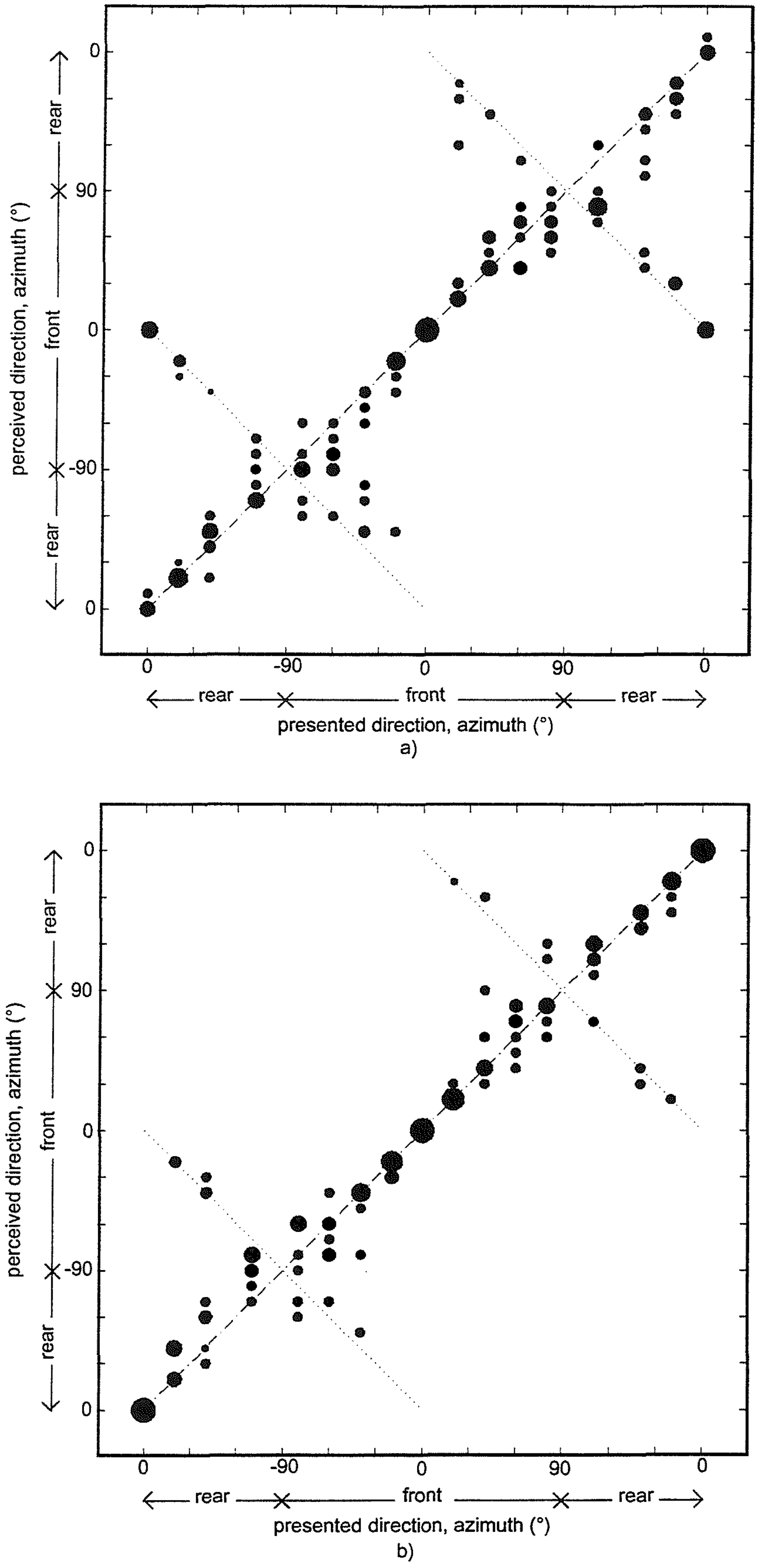


Fig. 3.17. Results of the localisation experiment with binaural synthesis over loudspeakers. The listener's head is at the optimal position and orientation. 11 subjects were tested. Responses by the 7 subjects for whom the systems work well. a) 60° transducer span. b) 10° transducer span.

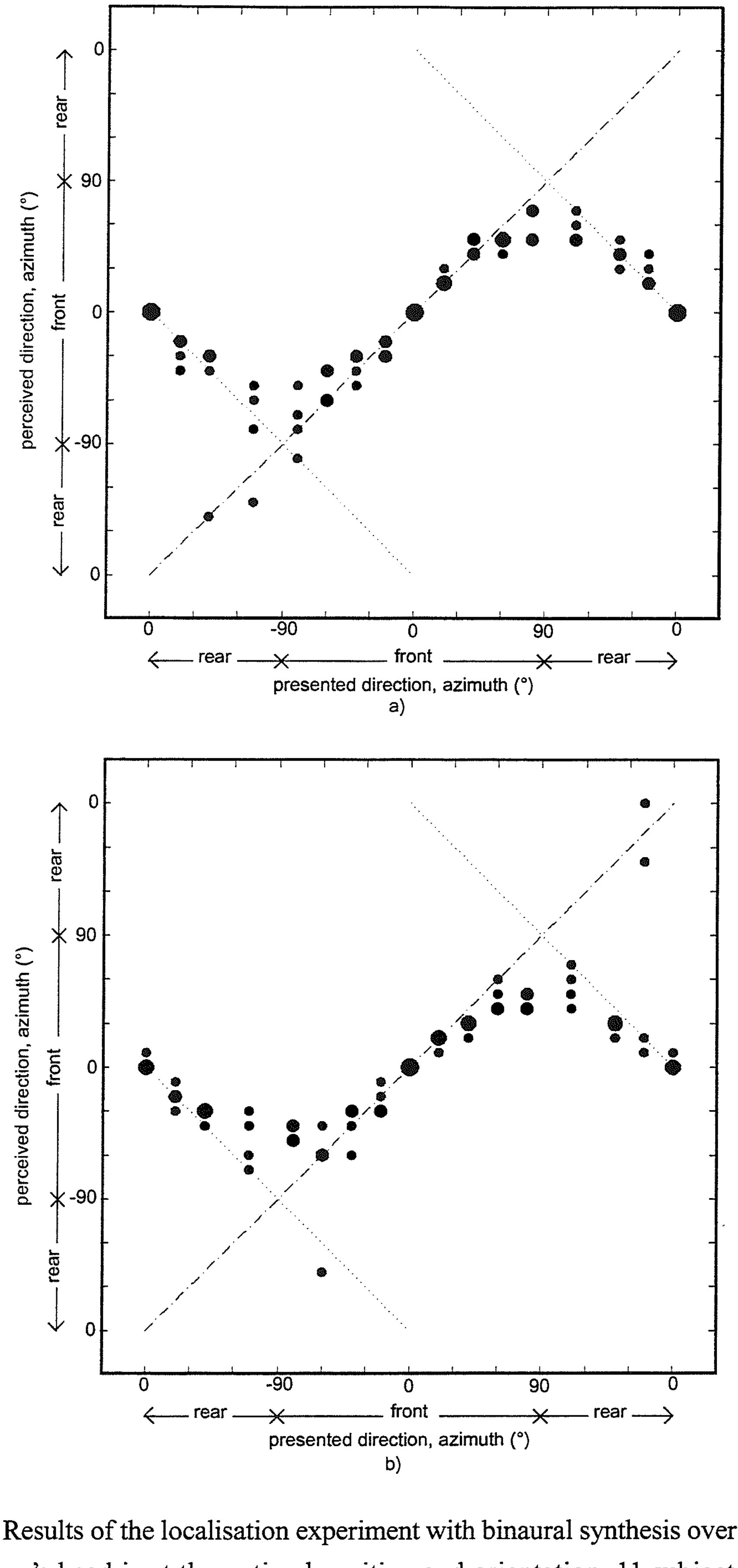


Fig. 3.18. Results of the localisation experiment with binaural synthesis over loudspeakers. The listener's head is at the optimal position and orientation. 11 subjects were tested. Responses by the 4 subjects for whom the systems do not work well. a) 60° transducer span. b) 10° transducer span.

4 Influence of individual differences in head related transfer functions

4.1 Introduction

Differences in performance among individual subjects were observed through the localisation tests in the previous Chapter. According to these results, it was possible to define two groups of subjects. For one group the virtual acoustic imaging system works well, but does not work so effectively for the other group. The latter subjects did not correctly localise the virtual sound source in the rear half of the horizontal plane and localised them at axisymmetrical positions in front with respect to the interaural axis. Moreover, the virtual sound sources at both sides are usually perceived at an offset position towards the front.

A further investigation was carried out in order to find out the cause of these differences. Among the types of error observed in the test, we put emphasis on front-back confusion, since it is widely reported that such binaural based virtual reality systems often have problems with discriminating source positions along the cone of confusion. It is widely agreed that there are two types of cues used to locate a sound source along the cone of confusion [11]. One type is the cues related to the spectra of the binaural signals and the other type consists of cues in conjunction with head movement. When the number of positions where the sound pressure is to be controlled is limited to two, it is simpler to control the sound field if the listener's head is fixed. As a result, information that relates to head movement cannot be obtained by the listener. Thus, subjects mainly made use of spectral cues to locate sound sources along the cone of confusion in the previous experiment where head movement was restricted. Therefore, it is suspected that the spectral cues are distorted in a certain manner together with the interaural time and level difference cues, and that this produces those systematically biased results.

The objective of this Chapter is to investigate the variation in the performance of the virtual acoustic imaging system among individuals. When referring the performance of the sound reproduction system, it contains the ability to reproduce the quality of sounds as well as the spatial impression of the original sound field. However, we shall concentrate on how the spatial impression is preserved, more strictly, how much the ability to localise a single reproduced sound source is preserved. The "Stereo Dipole" system whose transducers are placed close together was taken as an example. The influence of the individual Head Related Transfer Functions (HRTFs) are thought to be responsible for the variety.

It is known that headphone binaural presentation can be improved by individual recordings **[39]**-**[41]**. In such cases, there is only one set of HRTFs in the virtual source synthesis process which is related to virtual source locations. Furthermore, the errors caused by headphone presentation are biased that perception of front virtual sources results at the rear. In the case of virtual acoustic imaging, there are two sets of HRTFs, one for the virtual source position and the other for the control which is related to transducer locations. Contrary to headphone reproduction, the errors observed among the "poor" subjects are biased that perception of rear sources results in front.

To make detailed investigations possible, the HRTFs of each subject were measured. Another set of localisation tests, with control filters specifically designed for the subject with the individually measured HRTFs, were repeated in order to confirm the hypothesis. The synthesised HRTFs for each subject with the control filters which had been designed with dummy head HRTFs are compared with the subjects' own HRTFs. The reason why some subjects make mistakes in discriminating front and back, (in most cases localising a rear image in front), is explained by looking at the similarity of the synthesised HRTFs and either of the individual HRTFs corresponding to front or rear sound sources.

4.2 Measurements of the individual HRTFs

Among many possible causes of the differences in performance of the system among the individuals observed in the previous subjective experiment (Section 3.4.3), individually different HRTFs in **p** and **Q** defined in Section 2.5 are suspected. Because the other factors are common to all subjects and that they do not explain why the system works well for some subjects and does not for other subjects. To make detailed investigations possible, the HRTFs in **p** and **Q** of each subject are measured. These are obtained using a maximum length sequence (MLS) measurement technique with a sampling frequency of 44.1kHz in an anechoic chamber. Measurements are performed in an anechoic chamber to obtain HRTFs only and in order to exclude the acoustic response of the environment. The Knowles XL-9073 probe microphone assembly [42] is placed at the entrance of the ear canals. The high frequency limit of the microphone is about 8 kHz. The KEF C-35 loudspeaker is placed 1.4 m from the centre of the interaural axis at various directions on the horizontal plane. Measured transfer functions are deconvolved with loudspeaker-microphone response measured in a free field (Fig. 4.1). The binaural transfer functions **p** of all the directions which is used in the localisation test are measured for 3 "good" subjects and 5 "poor" subjects. The measurements are performed for all the "poor" subjects who took part in the experiments in Section **3.4.3** and Section **6.4**. The overlap in subjects between the two experiments made the total number of 5 "poor" subjects. The 3 "good" subjects were sampled from these experiments.

The measured HRTFs in **Q** of all subjects together with those of the KEMAR dummy head are shown in Fig. 4.2. Fig. 4.2a shows HRTFs in **Q** for "good" subjects and those of "poor" subjects are shown in Fig. 4.2b. For comparison, they are normalised at a frequency below 300Hz where small differences in detail of the head are not considered to affect its response. It can be clearly seen that deviation becomes larger as frequency becomes higher. However, the HRTFs of the "good" subjects seem to be close to each other and close to the KEMAR HRTFs below 2 kHz and around 3~4 kHz.

Similar tendencies to **Q** are observed in the HRTFs in **p** for all directions. Distinctively different patterns of the HRTFs were observed between the sound source in front and in the rear at the axisymmetric positions for all directions. An example of the measured HRTFs in **p** of all subjects are shown in Fig. 4.3 together with $\mathbf{p}_d$ of the KEMAR dummy head. Fig. 4.3a shows HRTFs in **p** for "good" subjects and those of "poor" subjects are shown in Fig. 4.3b. The direction of the sound source is at ±40° and ±140° and HRTFs for the ear closer to the sound source is plotted. These directions are at the axisymmetric positions with respect to the interaural axis and that mistakes in choosing alternatives between the two are often observed in localisation tests.

4.3 Localisation test with the filters designed using the individual HRTFs

The differences in performance among the individuals seem to be due to differences in the amount of mismatch in the HRTFs used in designing digital filters and those of the subjects. To confirm this, another localisation test with digital filters specifically designed for the subject with the individually measured HRTFs were performed. The same procedure was undertaken except that the individually measured **p** and **Q** in the previous section were used as $\mathbf{p}_d$ and $\mathbf{Q}_d$ to design digital filters specifically for each subject. The experimental procedure is fundamentally the same as that described in Section 3.4

Fig. 4.4 shows the result of 4 “poor” subjects from the previous localisation tests (Section 3.4.3) for whom the virtual acoustic imaging system did not work so effectively. They showed as good a performance as "good" subjects. They perceived virtual sound sources in the rear as well as the “good” subjects did. The localisation test showed no significant difference between these two groups of subjects. It is confirmed that what degraded the performance of the virtual acoustic imaging system for the “poor” subjects were mismatch of HRTFs in design and synthesis.

4.4 Comparison between the synthesised and individual HRTFs

Among the errors observed in the localisation test, misjudgement of the front-back discrimination at the axisymmetric position with respect to the interaural axis are paid particular attention here, since even when the threshold of misjudgement is delicate, the resulting absolute error angle is substantial.

It is assumed that human beings acquire memories of the transfer functions **p** as modifications of spectra in accordance with the different sound source position through every day life. When one perceives a familiar sound, one can estimate the transfer functions and make use of it for localisation. When it is not a familiar sound, one can estimate the change of transfer functions by moving one’s head then localise the sound source utilising memories of change of transfer functions with head movement. It is also known that a human can become able to use spectral cues of unknown sound with very limited experience. In the previous subjective experiments, the listener was assumed not to move their head when the digital filters had been designed. The subjects were told not to move their head during the subjective experiments so that the cues that relate to head movement did not result in the deterioration of the judgement. There is a small possibility that subjects might have unconsciously used the movement related cues since their heads were not physically fixed. As a result, there might have been some misjudgement due to this. However, it has been proven that the digital filters specifically designed for the subject with the individually measured HRTFs improved the performance. It has been shown that humans can discriminate broad band sound between sources from the front and those from the back, even when their head is fixed [12]. Therefore, it is suspected that most probable cause in this case is that the spectral cues were distorted by mismatch of the HRTFs.

4.4.1 Spectral shape of the HRTFs

If HRTFs of the listener in **p** and **Q** are very different from those that were used to design digital filters, the resulting synthesised HRTFs in $\mathbf{p}_s$ become different from the transfer functions **p** in one's memory. The vector $\mathbf{p}_s$ can be expressed as follows from Eq. (2.19) and Eq. (2.31):

$$\mathbf{p}_s = \mathbf{Q}\mathbf{Q}_d^{-1}\mathbf{p}_d \tag{4.1}$$

When a sound signal is presented with $\mathbf{p}_s$, the listener can nominate either of two directions at axisymmetric positions in the horizontal plane through the interpretation of the interaural time and level differences. Then, it is possible that the listener chooses the direction whose **p** is closer to the presented $\mathbf{p}_s$. To find out if this is the case, $\mathbf{p}_s$ synthesised with KEMAR HRTFs $\mathbf{p}_d$ for 0°, ±20°, ±40°, ±140°, ±160° and 180° are compared with the listener's own HRTFs $\mathbf{p}$ for the same direction and axisymmetric direction for the subjects. The HRTFs corresponding to the sound source in the front half are denoted as $\mathbf{p}_f$ and $\mathbf{p}_b$ is that of the rear out of the two candidates. Fig. 4.5 shows an example of $\mathbf{p}_s$ and two candidates $\mathbf{p}_f$ and $\mathbf{p}_b$ for the direction of ±160° for the ear closer to the sound source. They are again normalised at a frequency below 300Hz.

It was observed that all the subjects' judgements in the localisation test and the similarity of the HRTFs between $\mathbf{p}_s$ and $\mathbf{p}$ coincides very well. As observed in the subjective test, the synthesised HRTFs in $\mathbf{p}_s$ for the rear half of the source positions for the "poor" subjects were actually often closer to those in front $\mathbf{p}_f$. Probably, this is because the loudspeakers for reproduction are actually placed in front and the components in $\mathbf{Q}$ related to frontal image were not cancelled properly. This explains why the misjudgement of the "poor" subjects were biased to judge rear images in front but not to misjudge in front and rear randomly.

4.4.2 Discriminant analysis

To give this observation an analytical measure, discriminant analysis is performed [43]. Discriminant analysis is a kind of regression analysis that enables the distinctions between groups by a linear function of variables when several characteristics can be measured as the variables on each of the individuals. As a measure of similarity of the transfer functions between $\mathbf{p}_s$ and two candidates $\mathbf{p}_f$ and $\mathbf{p}_b$, mean values of difference of normalised $\mathbf{p}_s$ and either $\mathbf{p}_f$ or $\mathbf{p}_b$ within a certain frequency band are defined as ε_f and ε_b. The *n*th elements of ε_f and ε_b are defined by

$$\varepsilon_{\mathrm{f}n} = \frac{1}{\omega_h - \omega_l} \int_{\omega_l}^{\omega_h} \left(\left|20\log\left|\mathrm{A}_{\mathrm{f}n}(\mathrm{j}\omega)\right| - 20\log\left|\mathrm{A}_{\mathrm{s}n}(\mathrm{j}\omega)\right|\right|\right) d\omega$$

$$\varepsilon_{\mathrm{b}n} = \frac{1}{\omega_h - \omega_l} \int_{\omega_l}^{\omega_h} \left(\left|20\log\left|\mathrm{A}_{\mathrm{b}n}(\mathrm{j}\omega)\right| - 20\log\left|\mathrm{A}_{\mathrm{s}n}(\mathrm{j}\omega)\right|\right|\right) d\omega$$

(4.2)

where ω_l and ω_h are lower and upper limit of the angular frequency. The logarithm is adopted from Weber's Law. If ($\varepsilon_{\mathrm{bn}}$ - $\varepsilon_{\mathrm{fn}}$) is positive, $\mathbf{p}_{\mathrm{s}}$ is close to $\mathbf{p}_{\mathrm{f}}$ in that frequency range.

In this trial, the frequency range was divided into 1/3 octave bands whose centre frequencies are from 315 Hz to 6300 Hz. The independent (predictor) variables are ($\varepsilon_{\mathrm{bn}}$ - $\varepsilon_{\mathrm{fn}}$) in each frequency band. 8 subjects, 5 of them are "poor" subjects, and 10 directions presented are taken as the cases. The cases which subjects perceived the sound sources both in front and in the rear simultaneously were excluded from the analysis so there were 74 cases. The subjects' judgements of the source position are the dependant (criterion) variables. To take variance into account, the Mahalanobis' generalised distance is used **[43]**. The distance D_0 between the centre of a group $\left(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p\right)$ and a particular point $\left(x_{01}, x_{02}, \cdots, x_{0p}\right)$ can be given with variance $\mathrm{V}^{nn'}$ by

$$D_0^2 = \sum_{n=1}^{p} \sum_{n'=1}^{p} \left(x_{0n} - \bar{x}_n\right)\left(x_{0n'} - \bar{x}_{n'}\right) \mathrm{V}^{nn'}$$

(4.3)

where $x_{\mathrm{n}} = \varepsilon_{\mathrm{bn}} - \varepsilon_{\mathrm{fn}}$. By comparing the distances from the centre of two groups, it is determined whether each synthesised spectrum resulted in a perception in front or in the rear.

Fig. 4.6 shows the results of the discriminant analysis in quantifying the ability of subjects to discriminate between front and back. The predicted directions with this analysis are presented against the actual subjects' judgements. It turned out that this model explains the discrimination of front and back very well. The prediction rate was 96%. The coefficients of the discriminant function suggest that there are some frequencies which may be more important than the others. The variables corresponding to 315 Hz ~ 500 Hz and 1000 Hz ~ 2500 Hz bands seem to have more influence than the others. As we expect, coefficients for the ear closer to the virtual sound source are larger than the other ear.

4.5 Conclusions

The reason for the variety in the performance of the virtual acoustic imaging system among individuals was investigated. It was revealed to be due to the variety among the individual HRTFs by a localisation test which is performed with the digital filters designed with the individually measured HRTFs. The synthesised HRTFs for each subject with filters which are designed with a dummy head were compared with those of the subjects' own. It was found that the synthesised HRTFs for the rear half of the positions which were localised in front were actually closer to the subjects' own HRTFs corresponding to the sources in front at the axisymmetric positions. It is concluded that when mismatch of the HRTFs is large, it results in presenting wrong information to discriminate front and back.

Related publications

[A 3] T. Takeuchi, P.A. Nelson, O. Kirkeby and H. Hamada, "Influence of Individual Head Related Transfer Function on the Performance of Virtual Acoustic Imaging Systems", 104th AES Convention Preprint 4700 (P4-3), (1998).

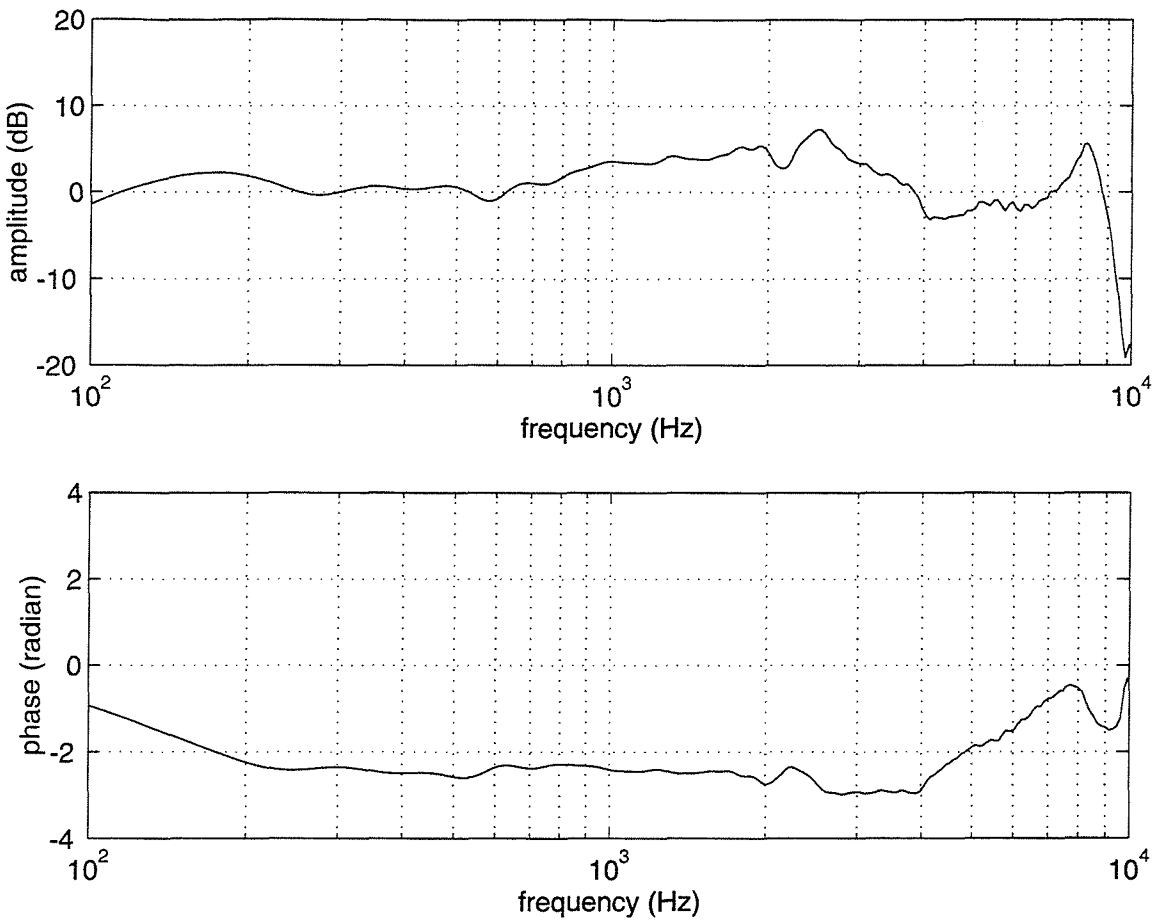


Fig. 4.1 The response of the microphones and the loudspeakers used to measure the individual HRTFs

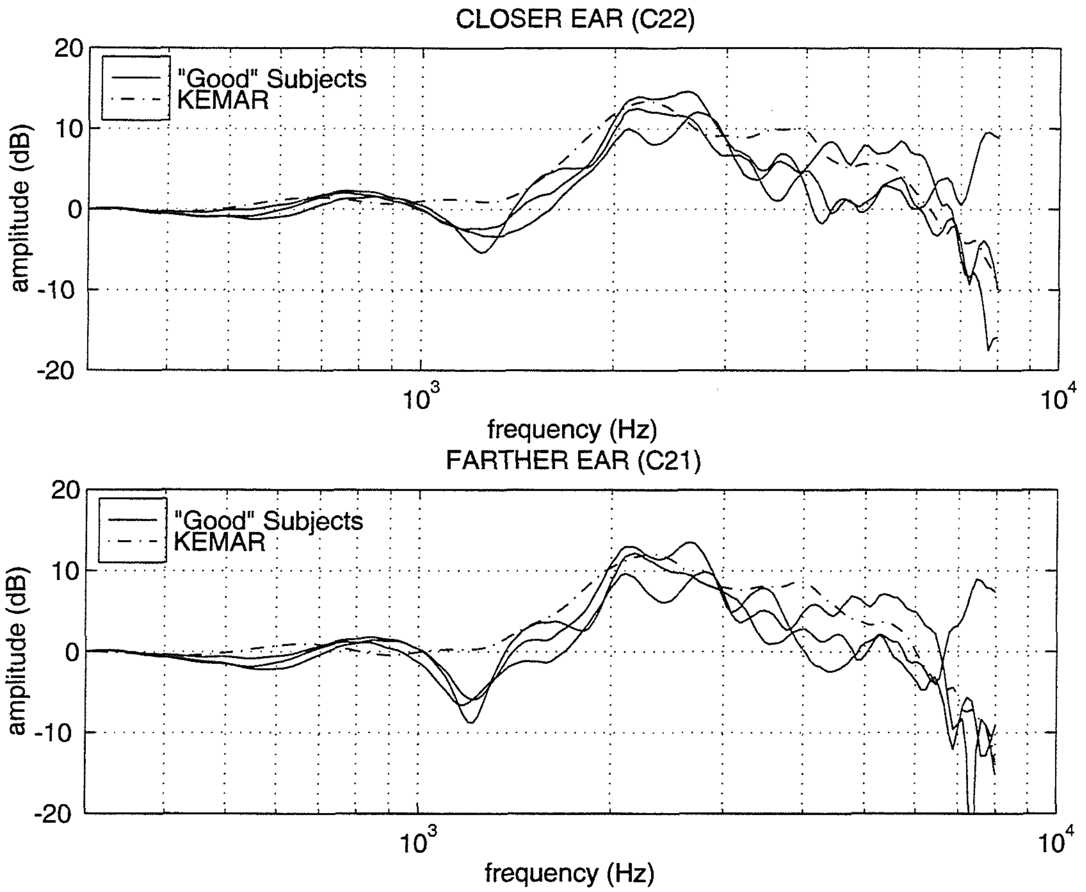


Fig. 4.2a Measured HRTFs in **Q** for the "good" subjects

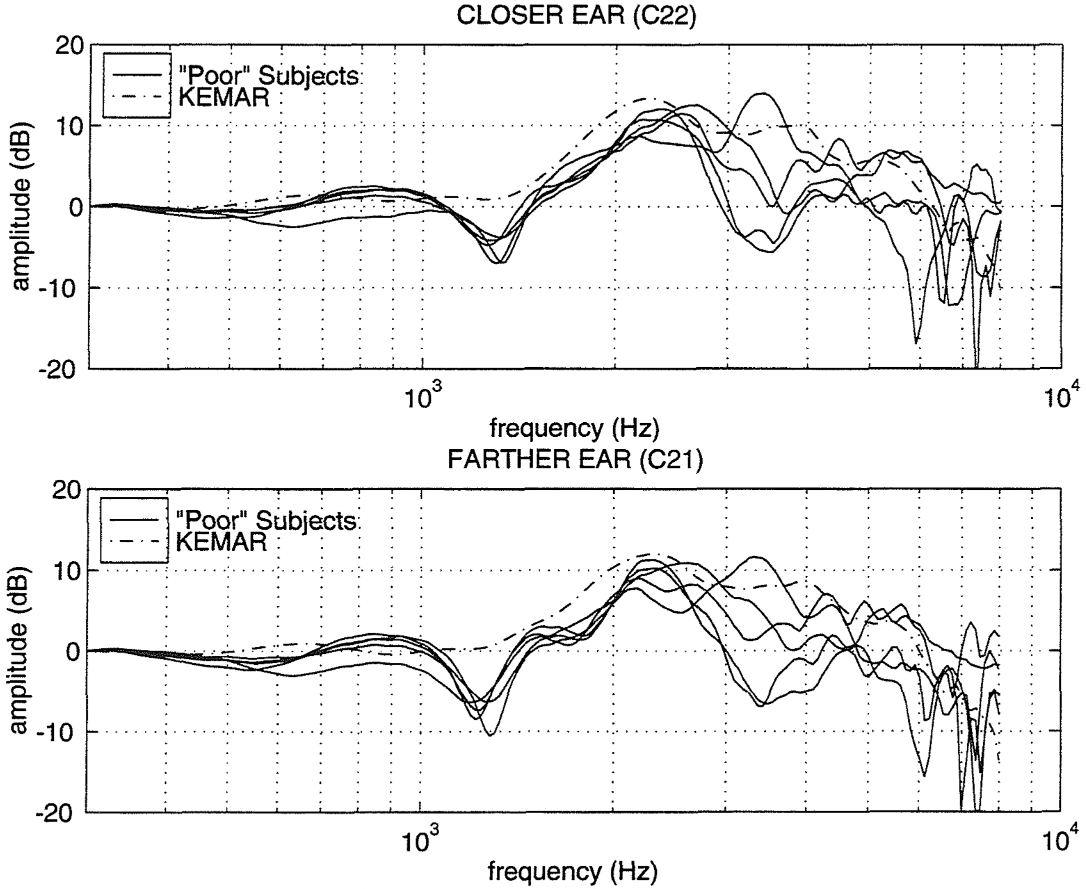


Fig. 4.2b Measured HRTFs in **Q** for the "poor" subjects

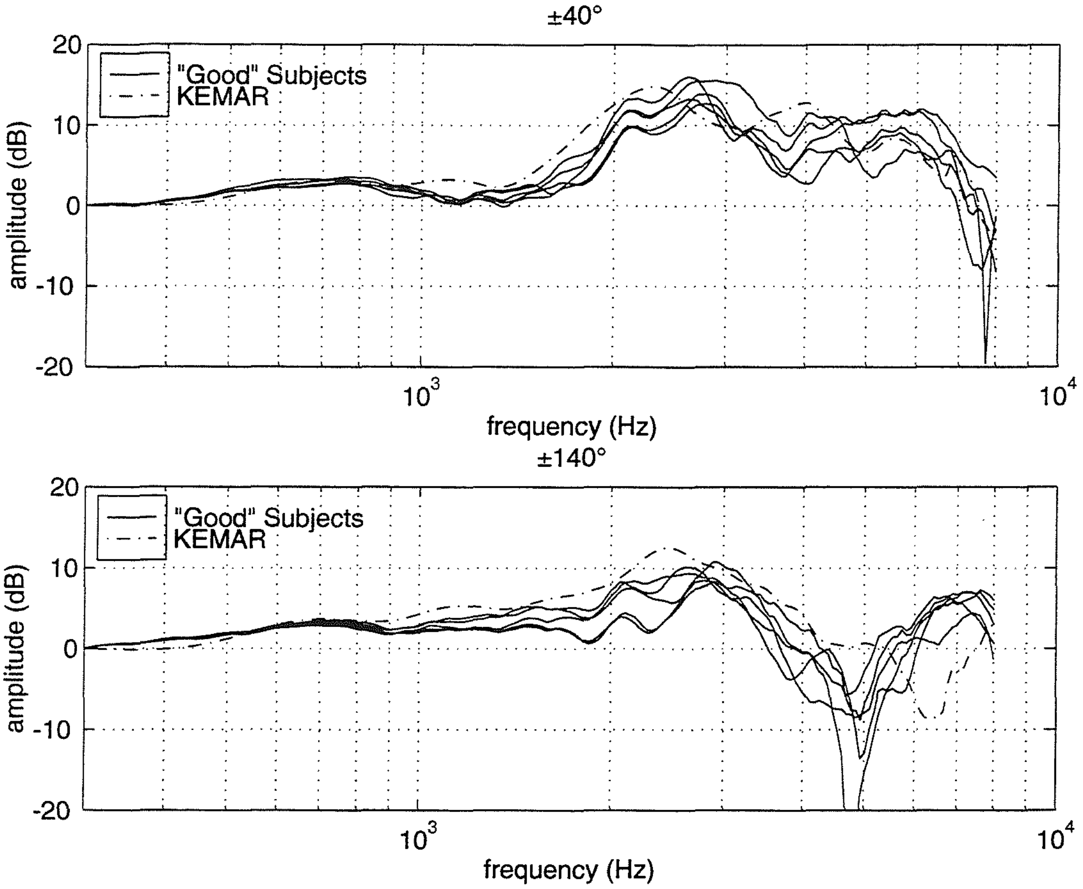


Fig. 4.3a Measured HRTFs in **p** for the closer ear of the "good" subjects, sound source at ±40° (front) and ±140° (rear)

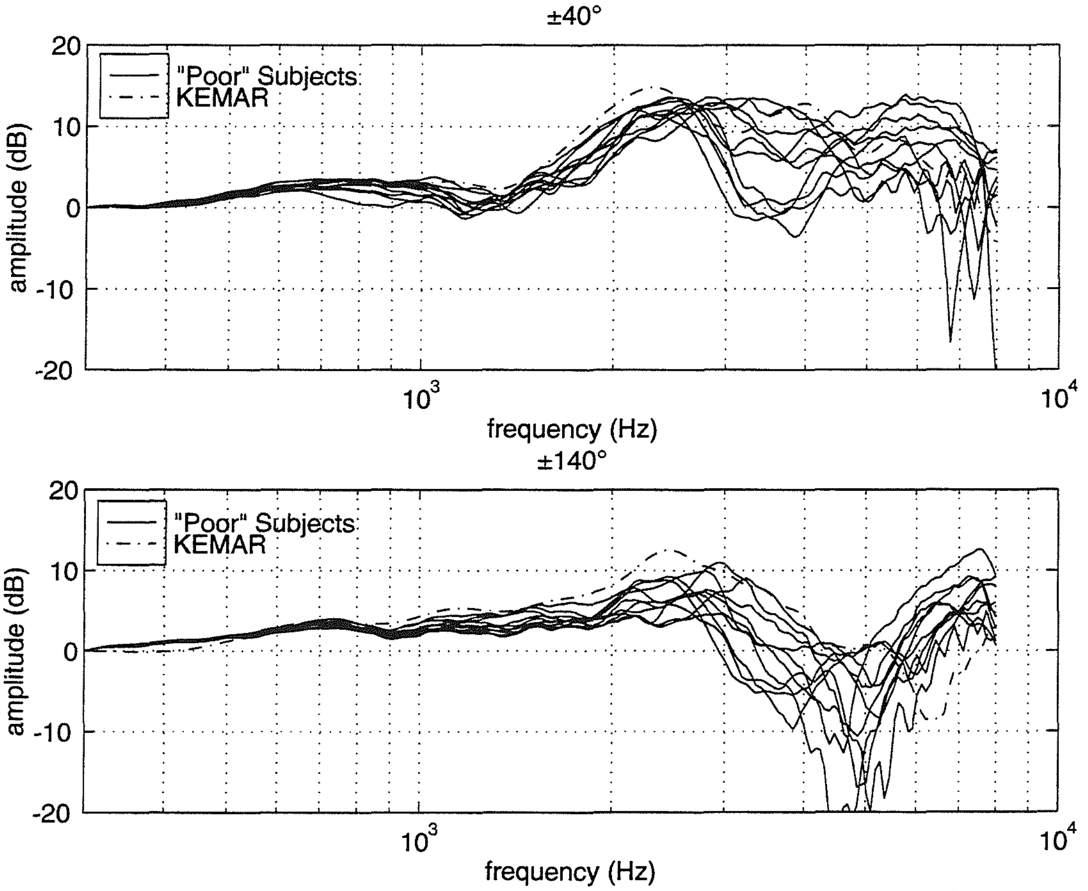


Fig. 4.3b Measured HRTFs in **p** for the closer ear of the "poor" subjects, sound source at ±40° (front) and ±140° (rear)

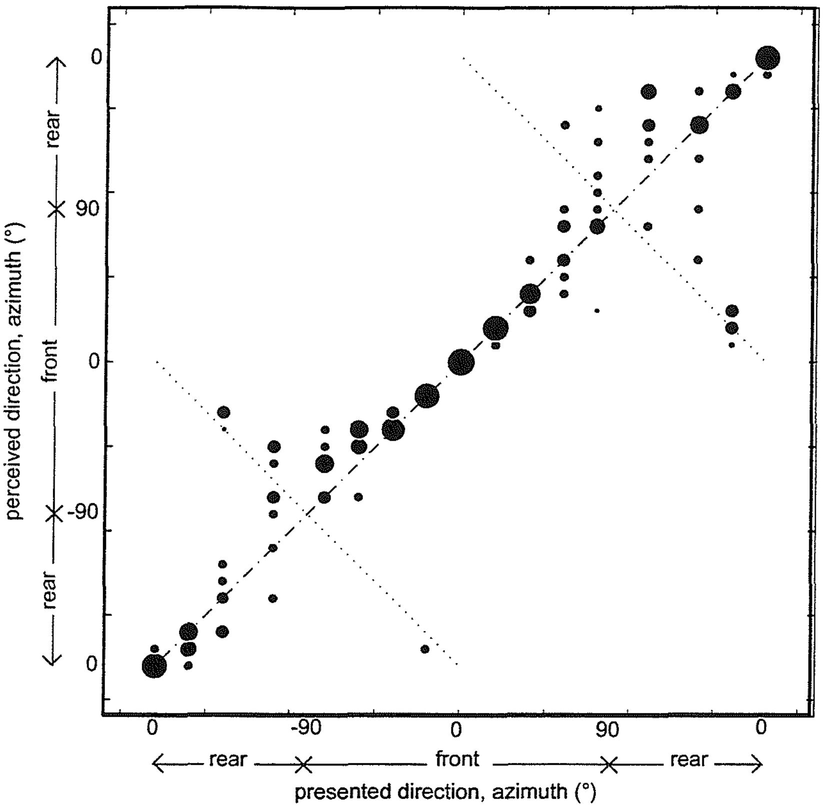


Fig. 4.4 The results of the localisation test of the virtual acoustic imaging system specifically designed for each subject ("Poor" subjects)

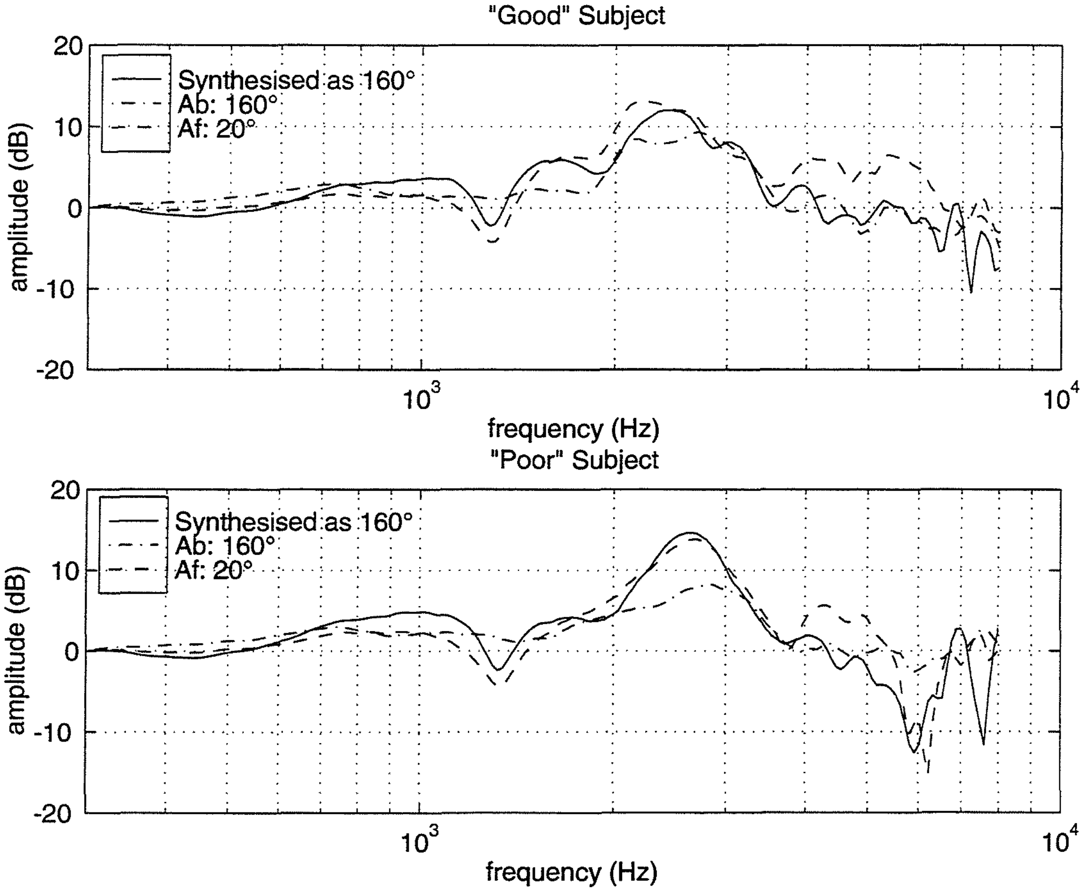
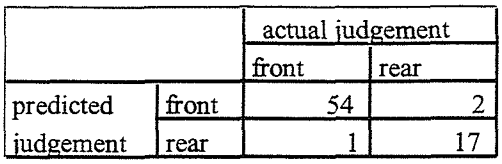


Fig. 4.5 Synthesised HRTFs in $\mathbf{p}_s$ and two candidate $\mathbf{p}_f$ and $\mathbf{p}_b$, for the closer ear, sound source at ±160°



		actual judgement	
		front	rear
predicted judgement	front	54	2
	rear	1	17

Fig. 4.6 The results of the discriminant analysis.

5 Robustness to head misalignment

5.1 Introduction

One disadvantage of binaural sound reproduction over loudspeakers is that the size of space where a reasonable control is obtained is not so large. The listener's ears must be in the relatively small region in a listening space at which the control is effective. This is because the independent control of the sound signals at each ear relies on interference of sound waves radiated from the loudspeakers. Misalignment of the head position and orientation results in the inaccurate synthesis of the binaural signals at the ears. This results from the change in the transfer functions between the transducers and the listener's ears. Consequently, the performance of the system deteriorates, i.e., directional information associated with the sound is smeared as is other information.

A certain amount of misalignment is inevitable in the practical use of such a system. Therefore, it is obviously advantageous if the system is more robust to misalignment. When, for example, the system is used to attempt to track head movement, it is essential to know the threshold at which the perception of the virtual acoustic environment collapses. The filter update would inevitably be discrete in time and hence the listener's movement would result in a certain amount of displacement from the intended position before the filters are updated for the new head position. In other words, the filters must be updated before the head moves a distance that is greater than the tolerances. Having larger tolerances will be advantageous in keeping track of head movement up to a higher velocity with a given update rate.

The objective of this Chapter is to investigate the robustness of the performance of such a system when the listener's head is misaligned. Comparison between two different transducer arrangements is made; two transducers placed close together as in the "Stereo

Dipole" arrangement and the conventional arrangement where the two transducers are spaced apart. As described in Chapter 3, the sound field generated by the "Stereo Dipole" has a distinct character in that its rate of change over space is much smaller than that generated by two monopole transducers spaced apart. As a consequence, it is expected to be more robust to misalignment of the position and orientation of the listener's head.

The consequences of three translational and three rotational displacements of the head are examined. Much emphasis is put upon the preservation of directional information which depends mostly upon the head related transfer functions (HRTFs). First, the effectiveness of control is investigated by synthesis of a unit impulse at both ears in both the time and the frequency domains. Presentation of an incident sound from various directions are then investigated as the very basic components of a virtual sound environment. The characteristics of the synthesised binaural signals are examined in several ways. In the temporal domain, the interaural time difference (ITD) of the synthesised binaural sound signals is investigated. The monaural spectral shape of the signals is also investigated since this will influence the spectral localisation cue. Further consideration is also given to the binaural spectral difference, i.e., the interaural level difference (ILD) that is used to localise along the interaural (azimuth) direction and also the interaural difference of spectral shape that is used to localise around the interaural axis (elevation). Cues related to the dynamics of head movement are outside the scope of this study. Subjective localisation experiments are performed for displacements for which notable differences in performance are expected from the previous analysis.

5.2 Analysis method

5.2.1 Model

As the listener's head is displaced away from the exact position for which control filters **H** are calculated, the transfer functions **C** change gradually. Thus the pseudo-identity matrix **X** and, as a consequence, the synthesised binaural HRTFs are degraded and may result in the wrong subjective perception.

As the very basic components of a virtual sound environment, generation of a single incident sound wave is taken as an example here. The physical acoustic paths **a** and **C** are modelled with free field (absence of any effects other than head) head related impulse responses (HRIRs: the time domain representation of HRTFs). A database comprising directionally discrete HRIRs on a virtual spherical surface 1.4m from a KEMAR dummy head is obtained from MIT Media Lab [38]. The "full" data set was used and the loudspeaker response is deconvolved from the data and thus each control transducer of the system is modelled as an ideal monopole source. Those between sampled directions are obtained by bilinear interpolation on the virtual spherical surface of magnitude and phase spectra in the frequency domain (Appendix 1). Those at a different distance from a head are obtained by extrapolation with an appropriately chosen delay and spherical attenuation (Appendix 1). The control filter matrix **H** is determined by the frequency domain deconvolution method [29].

The listener's head is displaced with respect to six orthogonal axes (three translational and three rotational) as in Fig. 5.1 and Table 5.1. Since the robustness to relatively small displacement of the head position and orientation is of interest here, the robustness of the virtual sound image is evaluated relative to the listener's head, not relative to the listening space. In other words, when the listener's head is displaced, he should ideally perceive the same virtual sound image as in the optimal position and orientation, unlike those applications where the listener may want to move around in a virtual sound environment.

The spherical co-ordinate system used to define direction of sound and of transducers is shown in Fig. 5.2. Two different transducer arrangements are investigated for comparison. In both cases, two transducers are placed in front of the listener on the horizontal plane (0° elevation) and aligned symmetrically with respect to the median plane. The transducers positioned spanning 60° as seen by the listener (±30° azimuth) are representative of a standard arrangement as in Stereophony. The span of 10° (±5° azimuth) represents close spacing, i.e. the "Stereo Dipole".

5.2.2 Indices for analysis

In the temporal domain, the ITD is analysed as the most important cue to locate the direction of sound along the interaural axis (azimuth discrimination). The interaural cross-correlation function $\Psi(\tau)$ of HRIRs $\mathbf{a}(t)$ corresponding to the real source direction are examined and the time lag which gives the peak values of $\Psi(\tau)$ is used as an estimate of ITD of the acoustic signals at the two ears. The interaural cross-correlation function $\Psi(\tau)$ is expressed as follows in terms of the elements of $\mathbf{a}(t)$.

$$\Psi(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \mathrm{a}_1(t)\mathrm{a}_2(t+\tau)dt \tag{5.1}$$

There are other possible methods for estimating ITD, for example, by detecting the leading-edge in the HRIRs, or by computing the phase spectrum or group delay of the binaural signals. However, the leading-edge method may misjudge ITD by detecting the less potent onset ITD rather than the ongoing ITD to which neurones are sensitive [44]-[46]. There is no indication that the nervous system could detect the high-frequency phase spectrum nor group delay. Anatomical and physiological studies strongly suggest that ITD information is extracted with the interaural cross-correlation of the auditory-nerve responses to the stimuli in the superior olivary complex then further processed at a higher level of auditory pathway [21][22]. The envelope delay of high frequency signals as an ITD cue [23][24] can be extracted by the cross-correlation method as well as the phase delay of low frequency signals. However, while this method extracts a single number ITD in binaural acoustic signals, it does not attempt to model the complex human auditory system which transduces acoustic signals at the ears into vibration of the auditory organs and then into nerve signals which are subsequently processed. Therefore, the absolute value of ITD may not be completely significant although it can extract tendencies and enables comparison between the two different conditions studied here.

In the spectral domain, a spectral analysis of synthesised HRTFs is performed over a logarithmic scale both in frequency and magnitude to account for the basic property of auditory filters. The monaural spectral shape is analysed as an important cue to identify a paricular direction out of a number of directions with no interaural differences. This cue utilises the change of the spectral shape of the sound source signal due to the HRTF for each ear. The monaural spectral cues also have supplemental role in localisation along the interaural direction [26]. Interaural difference of spectra is also analysed for two different objectives. One of them is as a cue to localise along the interaural direction (azimuth discrimination) with interaural level difference (ILD). The other is as another cue to resolve confusion among directions with no interaural time difference (elevation discrimination) by utilising the pattern of frequency dependent interaural spectral difference [27]. Again it should be noted that this does not attempt a complete model of the human auditory system.

5.3 Robustness of temporal cues

First, the effectiveness of control as a function of head displacement is evaluated by analysing the matrix of electro-acoustic paths **X**(t) which is independent of the direction of the virtual source. Following this, the synthesised HRIRs $\mathbf{a}_s(t)$ with head displacement are analysed in order to demonstrate what happens to temporal cues as a function of the relative direction of the virtual sound source.

5.3.1 Control performance (Temporal)

When the inputs to **X**(t) is a pair of simultaneous delta functions $\delta(t)$ rather than binaural signals, the interaural cross-correlation function, $\Psi_\xi(\tau)$, of the synthesised signals is expressed as

$$\Psi_\xi(\tau) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T}\xi_1(t)\xi_2(t+\tau)dt \tag{5.2}$$

where

$$\xi(t) = \begin{bmatrix}\xi_1(t)\\ \xi_2(t)\end{bmatrix} = \begin{bmatrix}\mathrm{x}_{11}(t)+\mathrm{x}_{12}(t)\\ \mathrm{x}_{21}(t)+\mathrm{x}_{22}(t)\end{bmatrix} \tag{5.3}$$

When the listener's head is at the optimal position and orientation, the synthesised signals $\xi(t)$ are approximately delta functions with an identical delay. Thus $\Psi_{\xi}(\tau)$ is a delta function with ITD = 0 (μs). In this way, the directional dependence in $\mathbf{a}(t)$ can be excluded from the analysis of the interaural cross-correlation functions. As the head is displaced away from the optimal position and orientation, the synthesised signals $\xi(t)$ are no longer delta functions. Thus $\Psi_{\xi}(\tau)$ is also no longer a delta function. A degraded $\Psi_{\xi}(\tau)$ indirectly suggests the degradation of the ITD cue of the synthesised HRIRs for all directions. A shift of the peak in $\Psi_{\xi}(\tau)$ suggests a shift in the ITD of the synthesised HRIRs and multiple peaks in $\Psi_{\xi}(\tau)$ may cause ambiguity or result in the wrong perception among multiple directions of sound.

Fig. 5.3 shows the degradation of $\Psi_{\xi}(\tau)$ (the interaural cross-correlation functions for the synthesised simultaneous unit impulses) with lateral displacement over the range of ±200mm. The maximum value of $\Psi_{\xi}(\tau)$ '1' at 0 lag can be observed at 0mm displacement (the optimal position) for both transducer arrangements. When the listener's head is displaced laterally, an ITD shift for the 60° transducer arrangement increases significantly as displacement increases, which is at the rate of approximately 2.7μs/mm. For example, 25mm displacement results in about 65μs ITD shift which corresponds to about 8° shift in azimuth direction. The threshold for ITD discrimination is considered to be approximately 10μs [25] and corresponds to about 4mm displacement with the 60° arrangement. On the other hand, the rate of shift is much less for the 10° transducer arrangement (0.2μs/mm) and so 50mm displacement would be just enough to produce the threshold value for ITD discrimination. When the listener's head is rolled, the ITD shift is again greater for the 60° arrangement though the difference between two arrangements is much smaller (about 1.2μs/° and 0.4μs/°) than the lateral displacement (Fig. 5.4). Yaw displacement showed the same ITD shift (about 8μs/°) which corresponds exactly to the yaw displacement angle for both of the two transducer arrangements (Fig. 5.5). However, better preservation (smaller amplitude of additional maxima) of the interaural cross correlation function can be observed for the 10° arrangement. $\Psi_\xi(\tau)$ for fore-and-aft displacement (Fig. 5.6) shows no shift of the original peak for both transducer arrangements, as expected from the symmetry, but slightly better preservation (smaller amplitude of additional maxima) of the interaural cross correlation function can be observed for the 10° arrangement. Vertical (Fig. 5.7) and pitch (Fig. 5.8) displacement did not show any ITD shift for both arrangements. The results for the six types of displacement are summarised in Table 5.2.

Comparisons can be made between the six types of displacement by normalising the results by the amount of displacement of the ears produced by each of the six types of head displacement. The synthesised ITD cue is the most sensitive to yaw displacement followed by lateral and roll displacements. It is very robust to fore-and-aft, pitch and vertical displacement. However, the difference in the robustness of the ITD cue between two different transducer arrangements is most significant for lateral displacement followed by roll displacements. There are no obvious differences other than additional maxima between two transducer arrangements for the other four displacements (yaw, fore-and-aft, vertical, pitch).

5.3.2 Accuracy of synthesis (Temporal)

By analogy with $\Psi(\tau)$, the interaural cross-correlation functions of synthesised HRIRs, $\Psi_s(\tau)$, is expressed as follows in terms of the elements of $\mathbf{a}_s(t)$.

$$\Psi_s(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \mathrm{a}_{s1}(t)\mathrm{a}_{s2}(t+\tau)dt \tag{5.4}$$

As the ITD cue is regarded as the most salient cue that is used to determine the azimuth direction [47], directions on the horizontal plane which contain two sets of all the azimuth directions are taken as examples to show the interaural cross-correlation functions of HRIRs (Fig. 5.9). That of the original HRIRs, $\Psi(\tau)$, is shown in Fig. 5.9a and that of synthesised HRIRs, $\Psi_s(\tau)$, when the listener's head is displaced 25 mm laterally are shown in Fig. 5.9b and Fig. 5.9c. In Fig. 5.9a, it can be observed that ITD is increasing almost linearly with respect to azimuth angle over most of the range. (Note that the variation is not sinusoidal which would be the case if there were no head in the sound field). $\Psi_s(\tau)$ is severely degraded with the 60° transducer arrangement (Fig. 5.9b); a few large additional local maxima (especially around ±250μs, corresponding to ±30° azimuth which are the control transducer directions) can be observed over wide range of virtual source directions as well as a shift (about 65μs, 8° azimuth) of the original peak. However, $\Psi_s(\tau)$ is better preserved with the 10° arrangement (Fig. 5.9c) except for very minor local maxima at virtual source directions around -90° azimuth, 0μs lag (the largest around -60μs which again corresponds to the control transducer directions).

ITD estimated from the synthesised HRIRs without displacement for both transducer arrangements are identical to the estimate from the original HRIRs for all the directions around the head. The estimated ITD from the synthesised HRIRs when the listener's head is displaced 25 mm laterally for most of the directions around the head is plotted in Fig. 5.10 and Fig. 5.11. There are no data points on the bottom part of the spherical plot. In general, it is observed that cones of constant ITD are shifted from the original value (Fig. 5.10a and Fig. 5.11a) for the 60° arrangement (Fig. 5.10b and Fig. 5.11b) but little shift is observed for the 10° arrangement (Fig. 5.10c and Fig. 5.11c), as observed in Fig. 5.3. The system with the 10° transducer arrangement preserved the synthesised ITD value for larger azimuth directions better than that of the 60° arrangement. A slightly worse performance is expected on the left side of the head than the other side (right) for the 10° arrangement. Whereas the right side shows worse performance than the left side for the 60° arrangement. The loss of a large ITD value around large azimuth directions (e.g. $|\text{azimuth}| > \pm 30°$ in Fig. 5.10b and Fig. 5.11b, around -90° azimuth in Fig. 5.10c) is primarily because the additional peaks in the interaural cross-correlation function became larger than the original peak. When the head is displaced, large additional peaks which give ITD values corresponding to the direction of the control transducers appear. In cases when these additional peaks are larger than the original peaks, if the largest peak is taken to estimate ITD, the virtual sound source would vanish and the listener would localise the sound source in the direction of the control transducers. However, with the existence of the other types of cue such as monaural spectral shape cues, the smaller magnitude of the original peak could be more plausible in estimating ITD. If it is taken to estimate ITD, it would result in much better preserved ITD value and thus better preserve the direction of virtual sound. This is down to psychological function at higher levels of the nervous system. It is likely, inferring from the results from subjective experiment presented in a later section, that a smaller but more plausible original peak would result in the estimated ITD for head displacements below a certain value.

5.4 Robustness of spectral cues

As in the analysis of temporal cues, the effectiveness of control as a function of head displacement is evaluated first by analysing the matrix of transfer functions **X** which is independent of the virtual source direction. Then, synthesised HRTFs $\mathbf{a}_s$ are analysed in order to demonstrate what happens to spectral cues depending on the direction of the virtual sound source.

5.4.1 Control performance (Spectral)

When the control system is required to synthesise particular spectra at two ears, head displacement results in leakage of some of the signal intended for one of the ears to the other ear. This is the so called "cross-talk" component of the signals, i.e. the component of the signal for right ear fed to the left ear and vice versa. This can be regarded as a noise component in the intended signal. From Eq.(2.10), the components of the synthesised HRTFs $\mathbf{a}_S$ are given by

$$\mathbf{a}_S = \mathbf{Xa} = \begin{bmatrix} X_{11}(j\omega)A_1(j\omega) + X_{12}(j\omega)A_2(j\omega) \\ X_{21}(j\omega)A_1(j\omega) + X_{22}(j\omega)A_2(j\omega) \end{bmatrix} \tag{5.5}$$

where $X_{11}(j\omega)$ and $X_{22}(j\omega)$ are the elements which contribute towards the correct synthesis of the HRTFs but $X_{12}(j\omega)$ and $X_{21}(j\omega)$ are noise elements which smear the synthesis. For the left ear, the signal (signal intended for the left ear) to noise (signal intended for the right ear) ratio of the control system is estimated from $|X_{11}(j\omega)|/|X_{12}(j\omega)|$. This is the case when the time histories of the inputs to **X** are a pair of identical delta functions. This again excludes the effect of **a**, i.e. the direction dependence. Fig. 5.12 shows the degradation of the S/N for the HRTF synthesis at the left ear with lateral displacement over the range of ±250mm. The signal to noise ratio (S/N) at the right ear, $|X_{22}(j\omega)|/|X_{21}(j\omega)|$, can be obtained by flipping over the left and right of the figure. Much larger displacements which maintain good S/N over wide frequency range (>500Hz) are allowed for the 10° transducer arrangement (roughly

±40mm for 20dB S/N) compared to the 60° transducer arrangement (roughly ±8mm for 20dB S/N). The dip in S/N around 9kHz and 13kHz even when the head is at the optimal position is due to low signal to noise ratio of the measurement of the HRTFs. Good S/N with larger displacement for the 10° arrangement can also be observed for fore-and-aft (roughly ±410mm compared to ±120mm for 20dB S/N) and yaw (roughly ±12° compared to ±6° for 20dB S/N) displacement as shown in Fig. 5.13 and Fig. 5.14. The 60° transducer arrangement has the advantage at frequencies below 500Hz, however. This is where ILD cues are less potent than ITD cues. There are not large differences between two arrangements for the other three displacements (roll, vertical, pitch). However, a slightly better S/N is preserved with the 60° arrangement for pitch (Fig. 5.15) and vertical (Fig. 5.16) displacement. With rotation about the interaural axis, transducers being at large azimuth angle means less change of transducer direction than transducers being around the median plane. There are not large differences between two arrangements for roll displacements (Fig. 5.17). The results for six types of displacements are summarised in Table 5.3.

When compared in the same way as used in the temporal cue analysis, synthesised spectral cues are most sensitive to lateral and roll displacement followed by yaw, pitch, vertical and fore-and-aft displacements. However, the difference in robustness of spectral cues between two different transducer arrangements is most significant for lateral displacement followed by fore-and-aft and yaw displacements. Note that 20dB S/N is roughly sufficient to synthesise the monaural spectra for the ipsi-lateral ear but much better S/N is required for the contra-lateral ear. This is because, if the level of two desired ear signals $\mathbf{d}(z)$ is compared, the level of the signal for the ipsi-lateral ear is smaller than that for the other ear over most of the frequency range and for most directions. As a result, at the contra-lateral ear, binaural synthesis is affected by a smaller signal input with a much larger noise input in addition to the response of the control performance of the system.

5.4.2 Accuracy of synthesis (Spectral)

As the role of the monaural spectral shape cue is primarily to determine the elevation direction of sources located on the cone of confusion, directions along the cone of 50° azimuth are taken as examples to illustrate the monaural spectral shape cue in the HRTFs. Fig. 5.18 shows examples of monaural spectral shape in HRTFs for the ipsi-lateral (right) ear at directions along the cone of constant azimuth (50°). Significant differences in spectrum pattern between sources below (at negative elevation) and above (positive elevation) the horizontal plane can be observed easily in Fig. 5.18a for real sound sources (estimated from $|A_2(j\omega)|$). There are less significant differences between sources in front (0~±90°) and in the rear (±90°~±180°) except on the horizontal plane where a significant dip in spectra around ±180° compared to those around ±0° can be seen in the mid frequency range. The synthesised monaural spectral shape (estimated from $|A_{s2}(j\omega)|$) when the listener's head is displaced 40 mm laterally are shown in Fig. 5.18b and Fig. 5.18c. The elevation dependency is less clear for that of the 60° arrangement (Fig. 5.18b). However, the synthesised monaural spectral shape for the 10° transducer arrangement (Fig. 5.18c) shows similar elevation dependent monaural spectra to the original spectra (Fig. 5.18a). The consequence of degraded monaural spectral shape would be an increased number of confusions among the directions on the constant azimuth cone. The degradation of this cue may also affect the azimuth localisation since the monaural spectral cue has a supplemental role for azimuth discrimination, especially when the interaural cross-correlation function $\Psi_s(\tau)$ is degraded to present ambiguity in estimating ITD due to a multiple choice of peaks.

Fig. 5.19 shows examples of monaural spectral shape in HRTFs for the contra-lateral (left) ear (estimated from $|A_1(j\omega)|$ and $|A_{s1}(j\omega)|$) at directions along the cone of constant azimuth (50°). As for the ipsi-lateral ear, differences in spectrum pattern for real sound sources is more significant between sources below and above than between front and rear (Fig. 5.19a). When the listener's head is displaced 40 mm laterally, the monaural spectral shape cue for the synthesised contra-lateral (left ear) HRTF is dominated by the noise, i.e., the cross-talk component, even for the 10° transducer arrangement due to the low S/N (Fig. 5.19b and Fig. 5.19c). The requirement for the preservation of monaural spectra for the contra-lateral ear is much more severe than that of the ipsi-lateral ear as pointed out in the previous section. For example, a lateral displacement of not more than 25 mm even for the 10° transducer arrangement and less than 5 mm for the 60° arrangement is required for the 50° azimuth directions. Obviously, the requirement varies as the direction of virtual sound source varies. The variation of azimuth direction (along the interaural axis) has more influence on it than the variation of the elevation direction (around the interaural axis).

Naturally, the same requirement as the contra-lateral monaural spectral shape cue, which is more severe than ipsi-lateral ear, applies for the both of the binaural spectral cues. In terms of analysis, these binaural spectral cues are essentially identical to the difference between the two monaural spectral shapes and estimated from $|A_{s2}(j\omega)|/|A_{s1}(j\omega)|$. Examples of the interaural spectral shape difference at directions along the cone of constant azimuth (50°) is shown in Fig. 5.20. As a matter of course, it has features of monaural spectral shape for both ears (Fig. 5.20a). For preservation of this cue (as well as monaural spectral shape for the contra-lateral ear), Fig. 5.20b and Fig. 5.20c also shows that 25mm lateral displacement is just within the limit with the 10° transducer arrangement for the cone of 50° azimuth directions but not with the 60° transducer arrangement. Above all, these monaural and binaural spectral shape cues are well preserved by the 10° transducer arrangement, so less confusion along the cone of confusion is expected with this arrangement.

Examples of another type of binaural spectral cue, the interaural level difference (ILD), are shown in Fig. 5.21 for sound source directions on the horizontal plane. As can be seen in Fig. 5.21a, which shows the ILD with real sound sources, it is not a simple task to allocate one ILD value to a particular azimuth angle. Since complex interference at higher frequencies yields multiple (often more than 4) azimuth angles for one ILD value at each frequency. In addition, the ILD value for a particular azimuth direction varies depending on frequency. The ILD with synthesised HRTFs when the listener's head is displaced 25mm laterally are shown in Fig. 5.21b and Fig. 5.21c. The ILD with the 60° transducer arrangement is degraded severely but those with the 10° span preserved well. Generally speaking, the ILD value for larger azimuth angles cannot be achieved without a very good preservation of monaural spectra for the contra-lateral ear. For example, with the 60° transducer arrangement with 50mm lateral head displacement, the ILD value (averaged over the mid-frequency range) for azimuth directions larger than ±30° cannot be achieved.

5.5 Subjective evaluation

The virtual directional information synthesised with two different arrangements of monopole transducers when the listener's head is misaligned were investigated by using subjective localisation experiments. Source directions on the horizontal plane were chosen to be examined since this covers the whole range of azimuth directions and two alternative elevation directions, i.e. 0° (front) and 180° (rear), in each cone of constant azimuth.

5.5.1 Procedure

The experimental procedure is fundamentally the same as that described in Section 3.4. The same data for the acoustic paths **a** and the control filter matrix **H** as those used in the analysis were implemented by digital filters using an MTT Lory Accel digital signal processing system. However, now the measurement head and the listener's head are different so that **a** becomes $\mathbf{a}_d$ and that **C** for the inverse filter matrix becomes $\mathbf{C}_d$. It is very important to bear in mind that there is considerable amount of variability of the HRTFs among individuals as described in Chapter 4. Inevitably, the matrix **C** containing each subject's HRTFs in this experiment is different from $\mathbf{C}_d$ that assumed when the matrix **X** is designed. This is the largest source of error when comparing the results with the analysis. The loudspeakers, rather than the listener's head, were displaced in both the lateral direction and in the fore-and-aft direction in order to achieve the displacement of the listeners head from the optimal position. The precision of the arrangement of the loudspeakers and listener's head was of the order of ±10mm. All the subjects are the same as those took part in the subjective experiment described in Section 3.4.3

5.5.2 Virtual sound sources

Now some of the observations made in the previous subjective experiment shown in Fig. 3.17 in Section 3.4.3 may be explained by the result of the above analysis. In principle, different transducer arrangements should not produce much difference in performance when the listener's head is at the optimal position and orientation. However, it was observed in Section 3.4.3 that the 10° transducer span showed slightly better performance, especially for "good" subjects around 0° azimuth where it showed no front-back confusion, contrary to the considerable number of confusions with the 60° span. Although the listener's head was supposed to be at the optimal position and orientation in this experiment, some misalignment of the head is inevitable in practice. This may have caused the increase in front-back confusion with the 60° transducer span. The slight unintended displacement of the head would have probably exceeded the severe limit for the good synthesis of spectral cues for the 60° transducer arrangement (see Table 5.3, Fig. 5.12), even though the same displacement may have been within the required limit for the 10° transducer arrangement.

In Section 3.4.3, the 60° span transducer arrangement showed a slight advantage in azimuth localisation, both by the "good" subjects and by the "poor" subjects. This observation accords with the better control performance at the lower frequency region by the 60° control transducer arrangement observed in Fig. 5.12 ~ Fig. 5.14 which is important for the synthesis of time related cues.

5.5.3 Head displacement

Further experiments with head displacement were carried out only with the 7 "good" subjects. The results when the listener's head is displaced 50mm to the right are shown in Fig. 5.22. The subjects reported after this experiment that the task was very difficult since sometimes they did not perceive a clear direction and sometimes they perceived the source to be at multiple directional locations. The multiple perception may be the consequence of multiple maxima in the interaural cross-correlation function. Discrepancy in different cues (e.g. ITD and ILD) could also be the cause. Virtual sound sources presented by the 60° transducer arrangement intended at 0° azimuth angle (both in front and rear) are often perceived at 10°~20° offset direction, whereas the virtual sources were mostly perceived in the intended direction by the 10° arrangement. These results agree with predicted direction by the ITD analysis where a 16° offset is expected from the 60° arrangement but a 0° offset is expected from the 10° arrangement. This systematic shift can not be clearly seen at higher azimuth directions where the random localisation error is much larger. Nevertheless considerable offset around ±40°~±60° azimuth is also noticeable for the 60° arrangement. The azimuth localisation error is now more apparent with the 60° arrangement contrary to the previous case when the head is at the optimal position. More front-back confusions for the 60° arrangement than the 10° arrangement can still be observed. Degradation of spectral shape cues does not seem to affect the performance very much since little increase of front-back confusion can be observed, although some effect may have already been in the results at the optimal head position as discussed earlier. Another possibility is that the head displacement may not have degraded the spectral shape very much more than the disparity between each individual HRTFs and the KEMAR HRTFs. A slightly better performance is observed on the side to which the head is displaced (right) for the 10° arrangement, whereas the other side (left) shows better performance for the 60° arrangement as predicted by ITD analysis. Contrary to the poor ILD values obtained, azimuth localisation seems surprisingly accurate. Considering that the additional local maxima of the cross-correlation function start to become larger than the original maximum around 25 mm displacement for the 10° arrangement and at a much smaller displacement for the 60° arrangement, the performance of azimuth estimation is more likely to be determined by a more plausible local maximum than by the absolute maximum of the interaural cross-correlation function, as discussed in the analysis of temporal cues.

When the listener's head is displaced 200mm and 400mm to the rear, the 10° span transducer arrangement showed slightly better performance than the 60° arrangement for both azimuth localisation and front-back discrimination (Fig. 5.23). However, in both cases, the difference in performance between the two transducer arrangements are much less significant compared to lateral displacement.

5.6 Conclusions

In binaural synthesis over two loudspeakers, yaw, lateral and roll displacement results in a shift of ITD as well as the generation of additional local maxima in the interaural cross-correlation function. Fore-and-aft, vertical and pitch displacement results only in the generation of additional local maxima. There is less degradation of temporal cues for lateral, roll, yaw and fore-and-aft displacements when two loudspeakers are placed close together.

Any displacement induces more "cross-talk" components in synthesised spectra. There is less degradation of the spectral cue for lateral, fore-and-aft and yaw displacement when two loudspeakers are placed close together.

The ITD cue is the most robust to head misalignment followed by the monaural spectral cue for the ipsi-lateral ear. The monaural spectral cue for the contra-lateral ear is the least robust together with binaural spectral cues (including ILD cues).

Subjective experiments confirmed that two closely spaced loudspeakers have an advantage in performance with regard to the misalignment of the listener's head. The localisation performance with subjective experiments were better than those predicted with any one individual localisation cue. This suggests the importance of the combination of different localisation cues.

Related publications

[A 4] T. Takeuchi, P. A. Nelson, O. Kirkeby, and H. Hamada, “Robustness of the Performance of the "Stereo Dipole" to Misalignment of Head Position,” 102nd AES Convention Preprint 4464(17), (1997).

[A 5] T. Takeuchi and P. A. Nelson, “Robustness of the Performance of the "Stereo Dipole" to Head Misalignment,” ISVR Technical Report No.285, University of Southampton (1999).

[A 6] T. Takeuchi, P. A. Nelson, and H. Hamada, "Robustness to Head Misalignment of Virtual Sound Imaging Systems," J. Acoust. Soc. Am. 109(3), 958-971 (2001)

Tables and Figures for Chapter 5

Description	Terminology
Translation along x-axis	lateral
Translation along y-axis	fore-and-aft
Translation along z-axis	vertical
Rotation about x-axis	pitch
Rotation about y-axis	roll
Rotation about z-axis	yaw

Table 5.1. Terminology used to describe head displacement.

type of displacement	rate of ITD shift		displacement at 10 μs ITD shift	
	60° span	10° span	60° span	10° span
lateral	2.7 μs/mm	0.2 μs/mm	4 mm	50 mm
fore-and-aft	0 μs/mm	0 μs/mm	-	-
vertical	0 μs/mm	0 μs/mm	-	-
pitch	0 μs/°	0 μs/°	-	-
roll	1.2 μs/°	0.4 μs/°	8°	24°
yaw	-8 μs/°	-8 μs/°	1.3°	1.3°

Table 5.2. Estimated rate of ITD shift and displacement which gives the threshold value of ITD discrimination (10μs) for six types of displacement and two different transducer arrangements.

type of displacement	displacement at 20dB S/N	
	60° span	10° span
lateral	±8 mm	±40 mm
fore-and-aft	±120 mm	±410 mm
vertical	±220 mm	±190 mm
pitch	±18°	±14°
roll	±9°	±9°
yaw	±6°	±12°

Table 5.3. Estimated displacement which gives 20dB signal to noise ratio of the control system for six types of displacement and two different transducer arrangements.

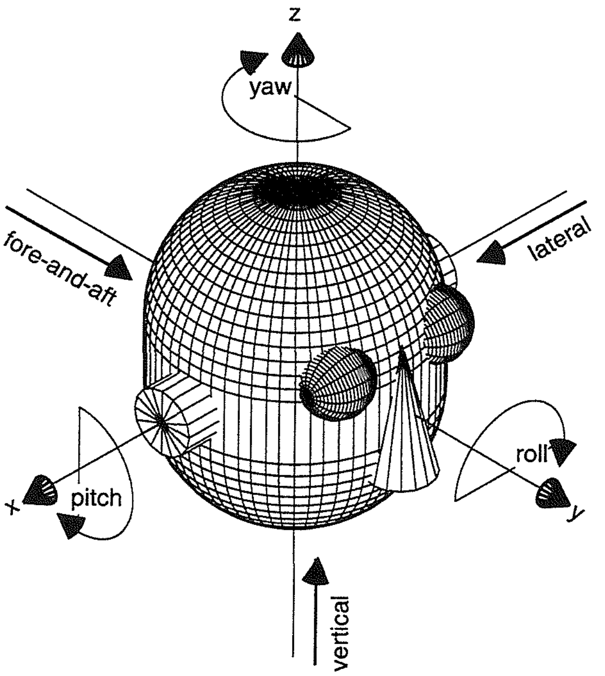


Fig. 5.1. The Cartesian co-ordinate system used to define head displacement relative to the optimal head position and orientation.

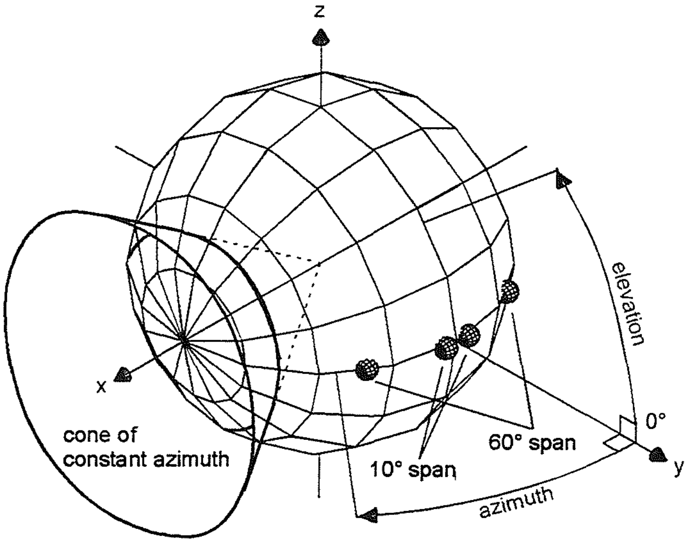


Fig. 5.2. The spherical co-ordinate system used to define the direction of sound sources relative to the listener's head position and orientation. An example of "cone of constant azimuth" is illustrated. The two different transducer arrangements investigated are also shown (relative to the optimal head position and orientation).

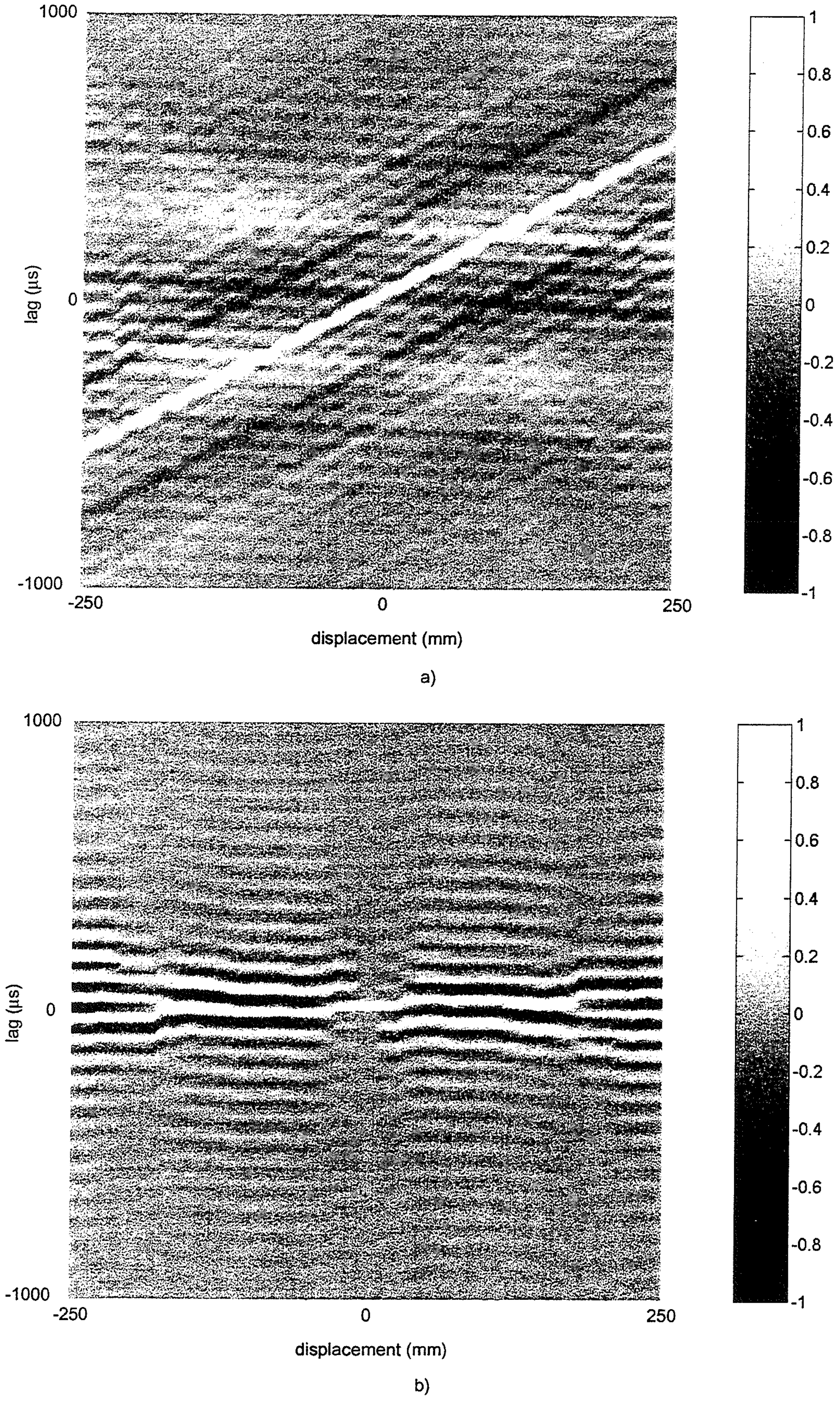
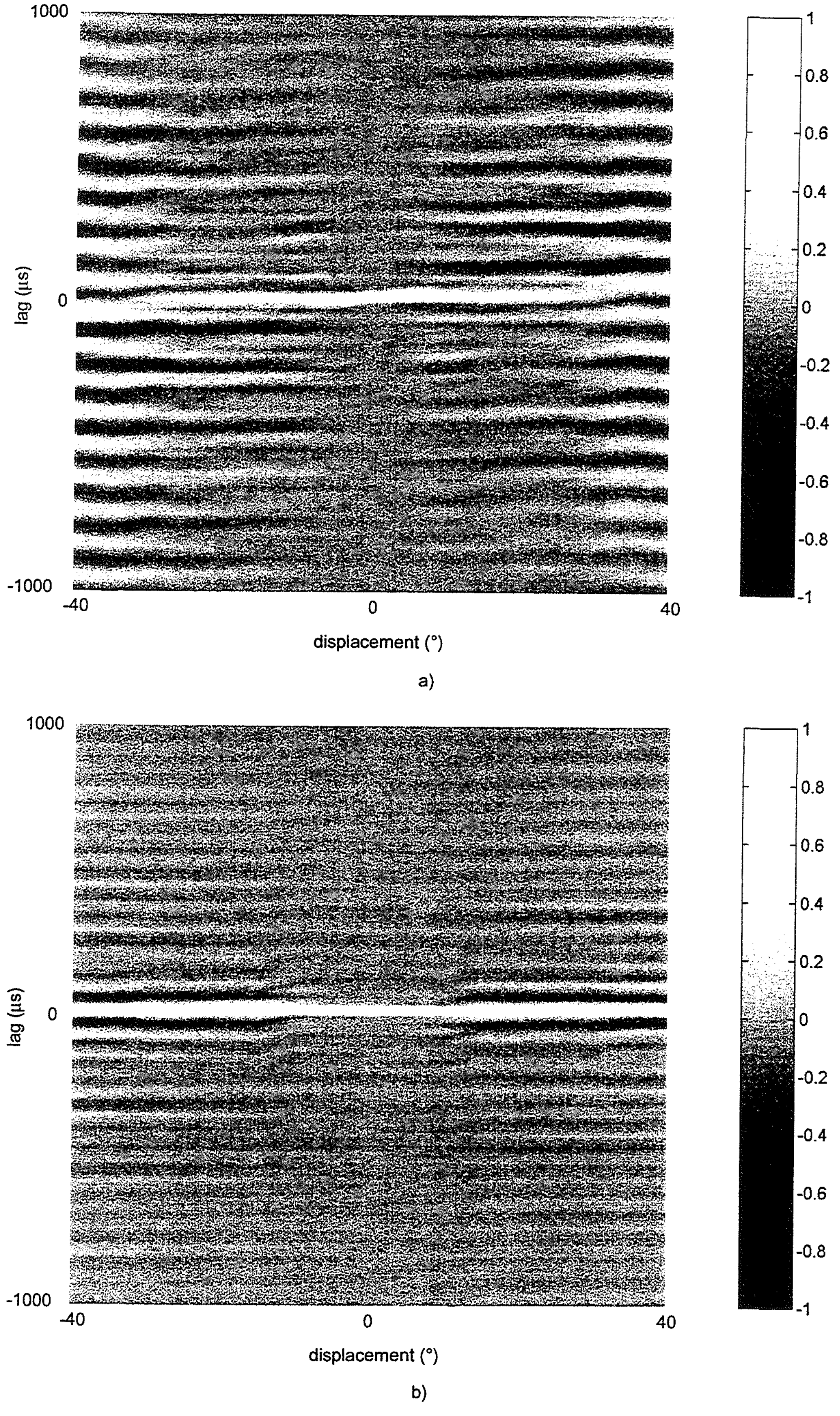


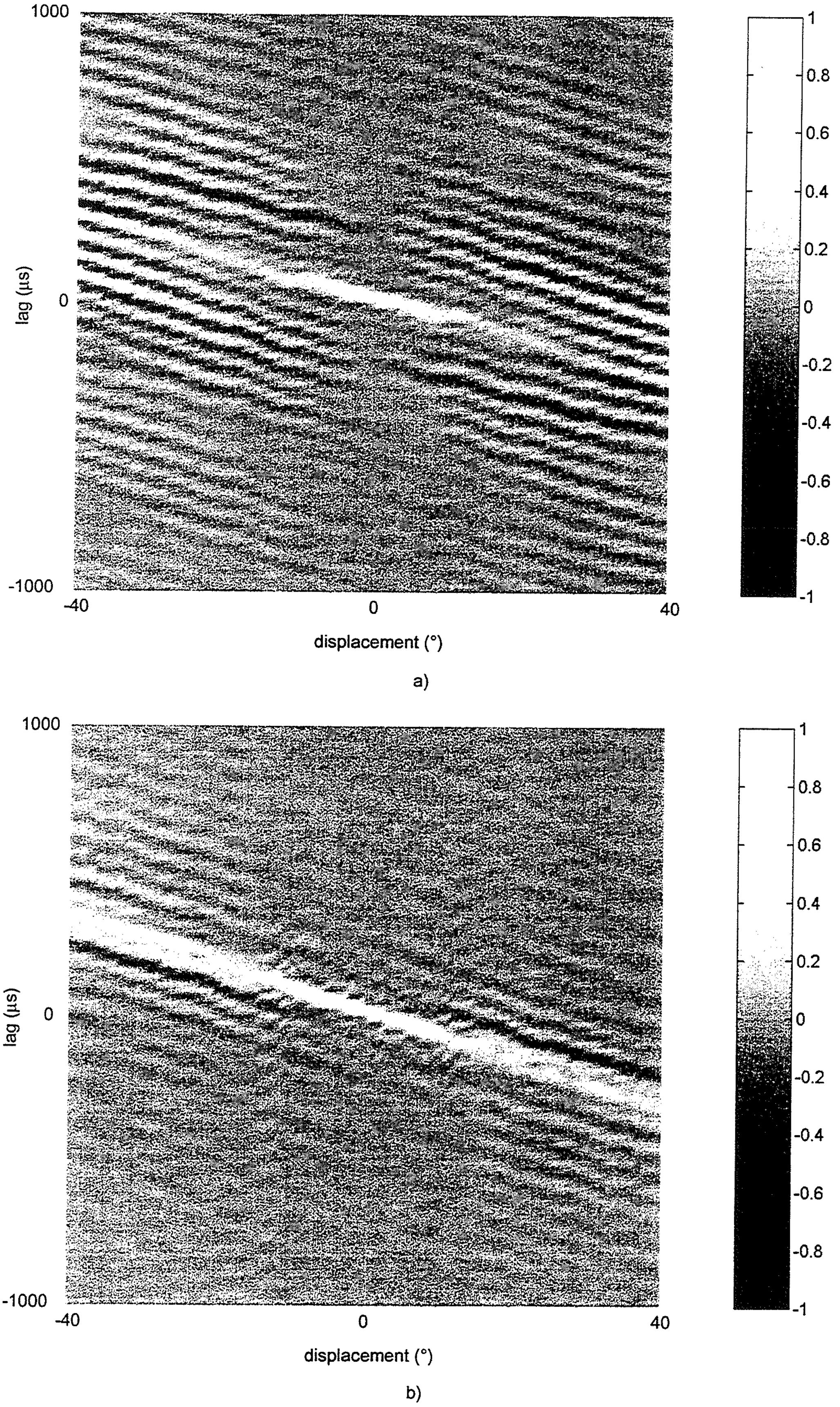
Fig. 5.3. The effect of lateral displacement on the interaural cross-correlation functions for the synthesis of simultaneous unit impulses. a) 60° transducer span. b) 10° transducer span.



a)

b)

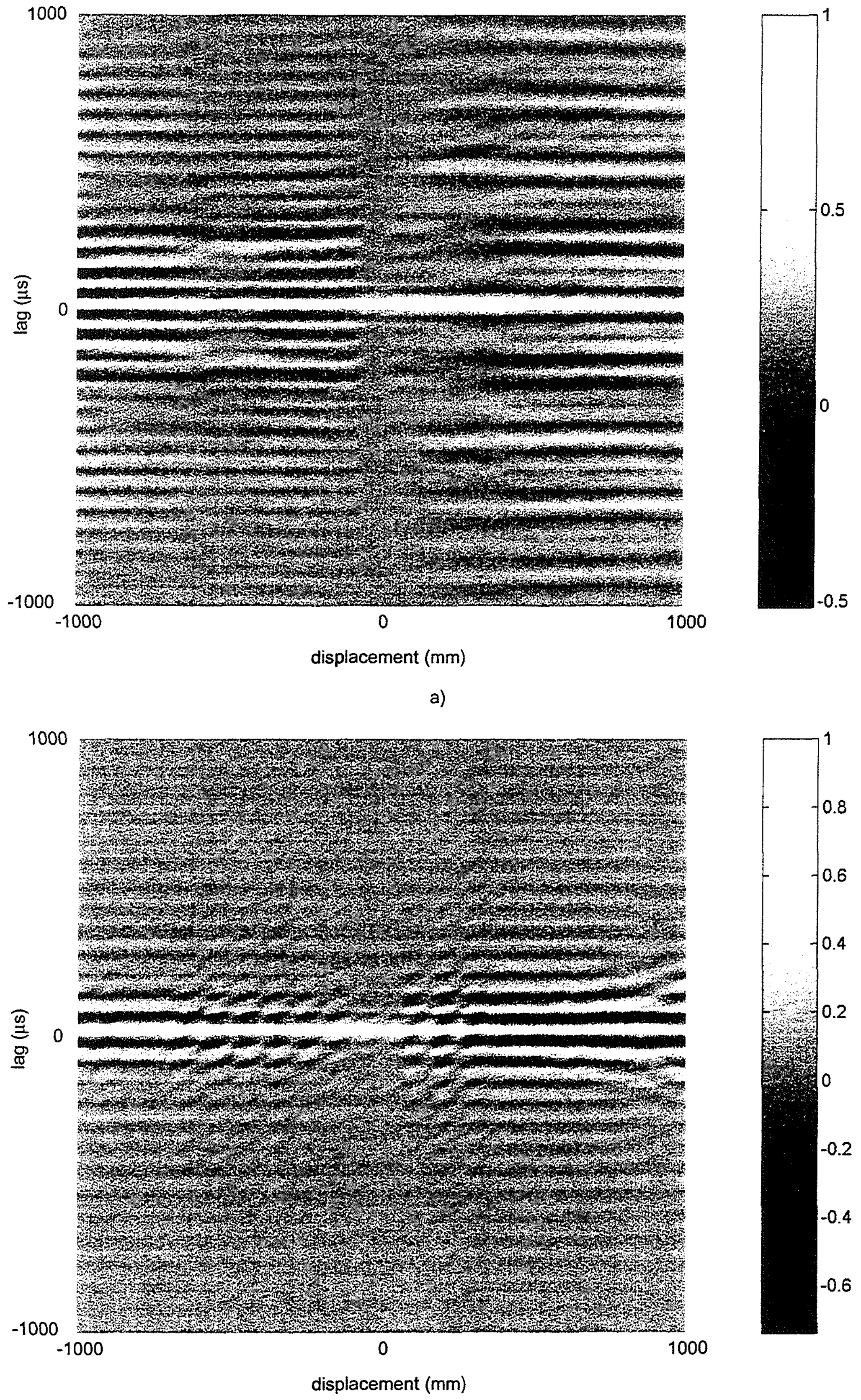
Fig. 5.4. The effect of roll displacement on the interaural cross-correlation functions for the synthesis of simultaneous unit impulses. a) 60° transducer span. b) 10° transducer span.



a)

b)

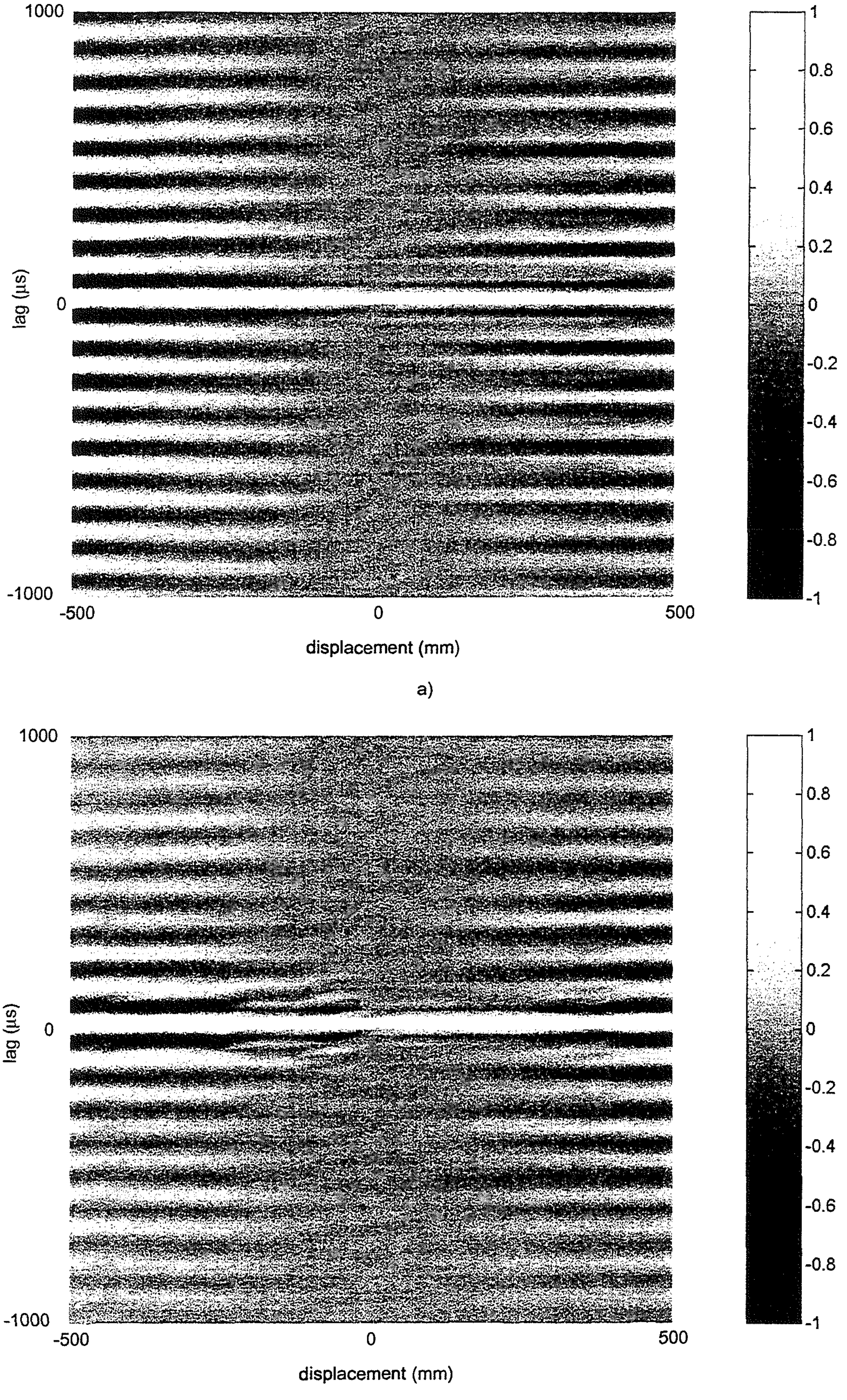
Fig. 5.5. The effect of yaw displacement on the interaural cross-correlation functions for the synthesis of simultaneous unit impulses. a) 60° transducer span. b) 10° transducer span.



a)

b)

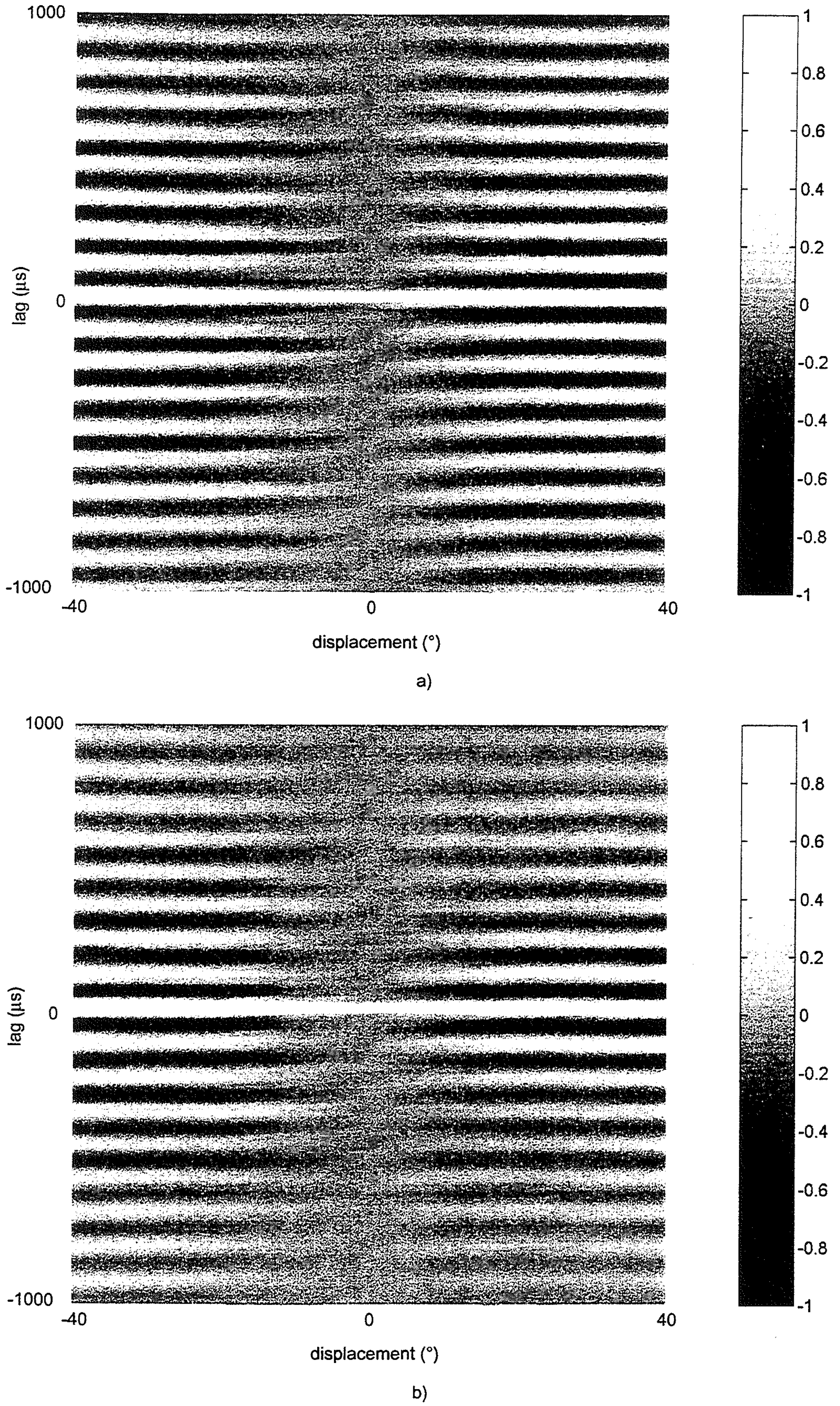
Fig. 5.6. The effect of fore-and-aft displacement on the interaural cross-correlation functions for the synthesis of simultaneous unit impulses. a) 60° transducer span. b) 10° transducer span.



a)

b)

Fig. 5.7. The effect of vertical displacement on the interaural cross-correlation functions for the synthesis of simultaneous unit impulses. a) 60° transducer span. b) 10° transducer span.



a)

b)

Fig. 5.8. The effect of pitch displacement on the interaural cross-correlation functions for the synthesis of simultaneous unit impulses. a) 60° transducer span. b) 10° transducer span.

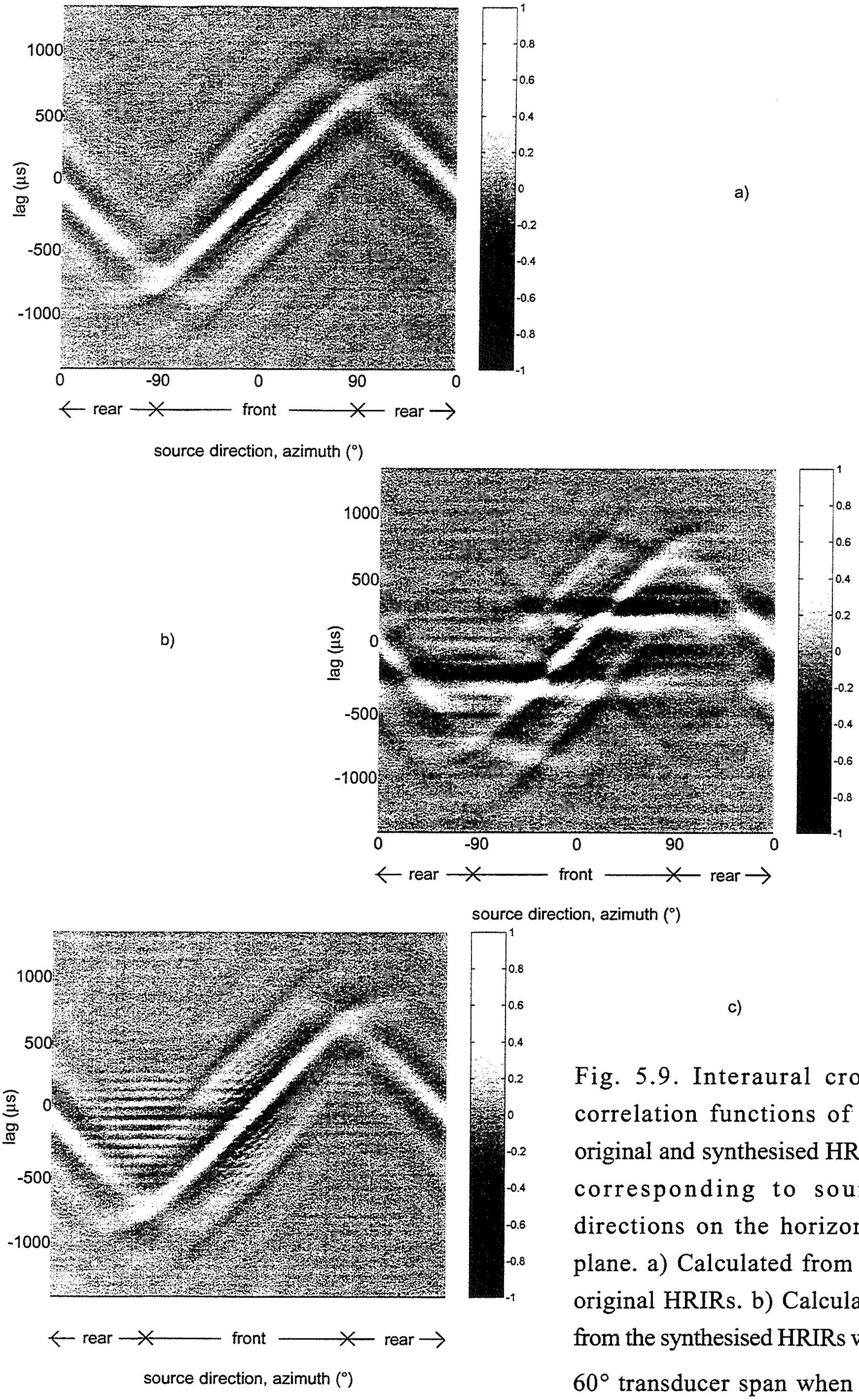


Fig. 5.9. Interaural cross-correlation functions of the original and synthesised HRIRs corresponding to source directions on the horizontal plane. a) Calculated from the original HRIRs. b) Calculated from the synthesised HRIRs with 60° transducer span when the listener's head is displaced 25 mm laterally. c) Calculated from the synthesised HRIRs with 10° transducer span when the listener's head is displaced 25 mm laterally.

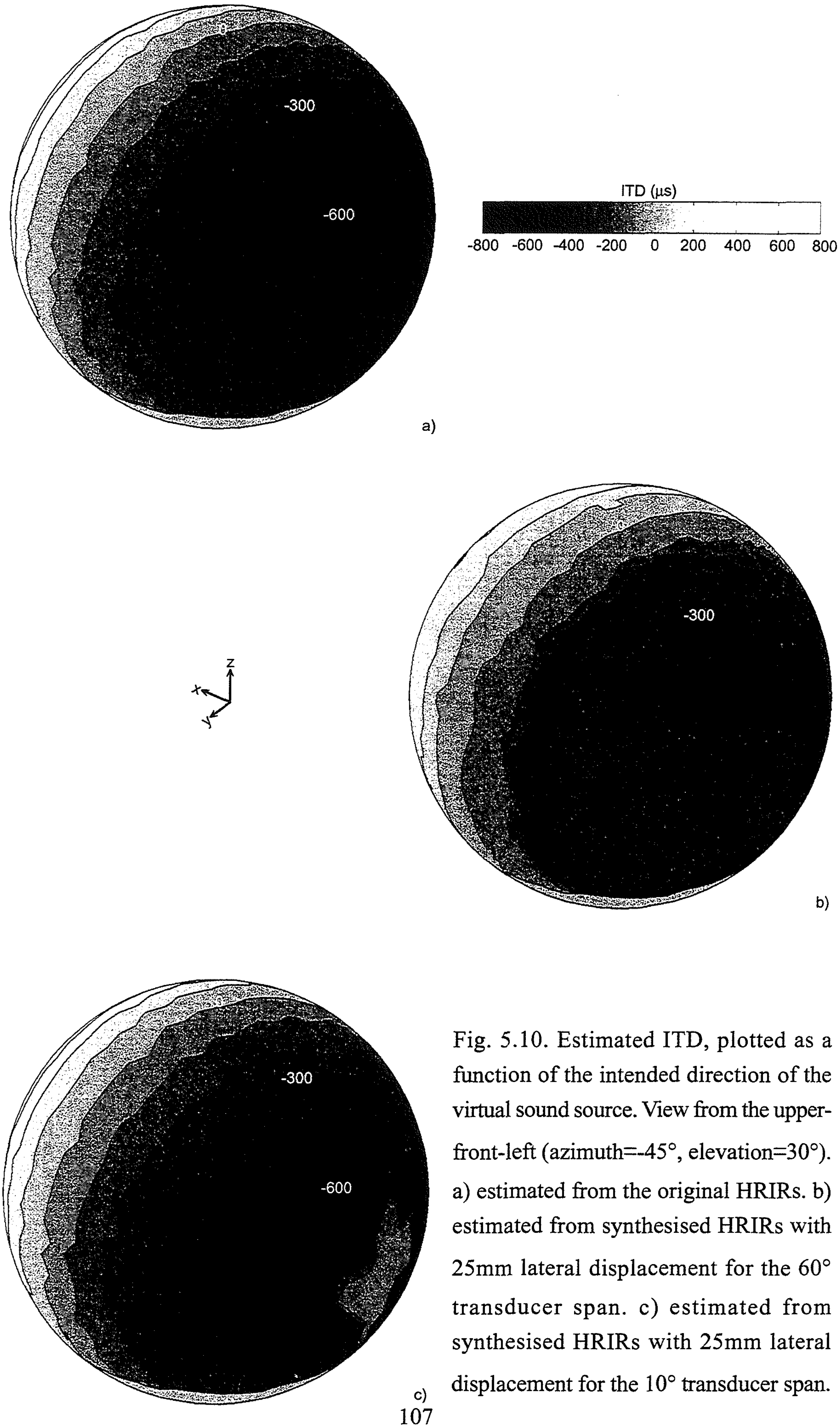


Fig. 5.10. Estimated ITD, plotted as a function of the intended direction of the virtual sound source. View from the upper-front-left (azimuth=-45°, elevation=30°). a) estimated from the original HRIRs. b) estimated from synthesised HRIRs with 25mm lateral displacement for the 60° transducer span. c) estimated from synthesised HRIRs with 25mm lateral displacement for the 10° transducer span.

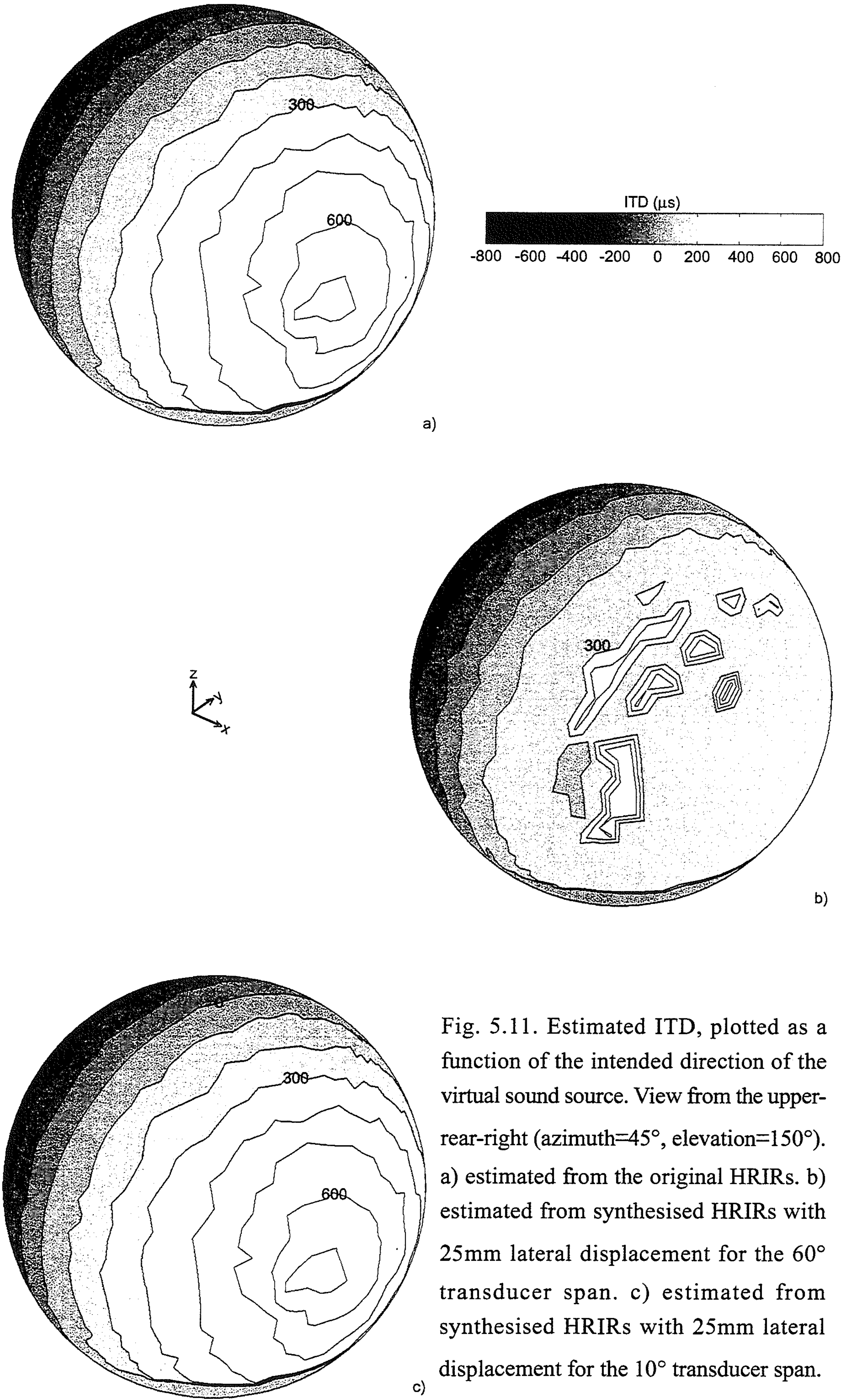
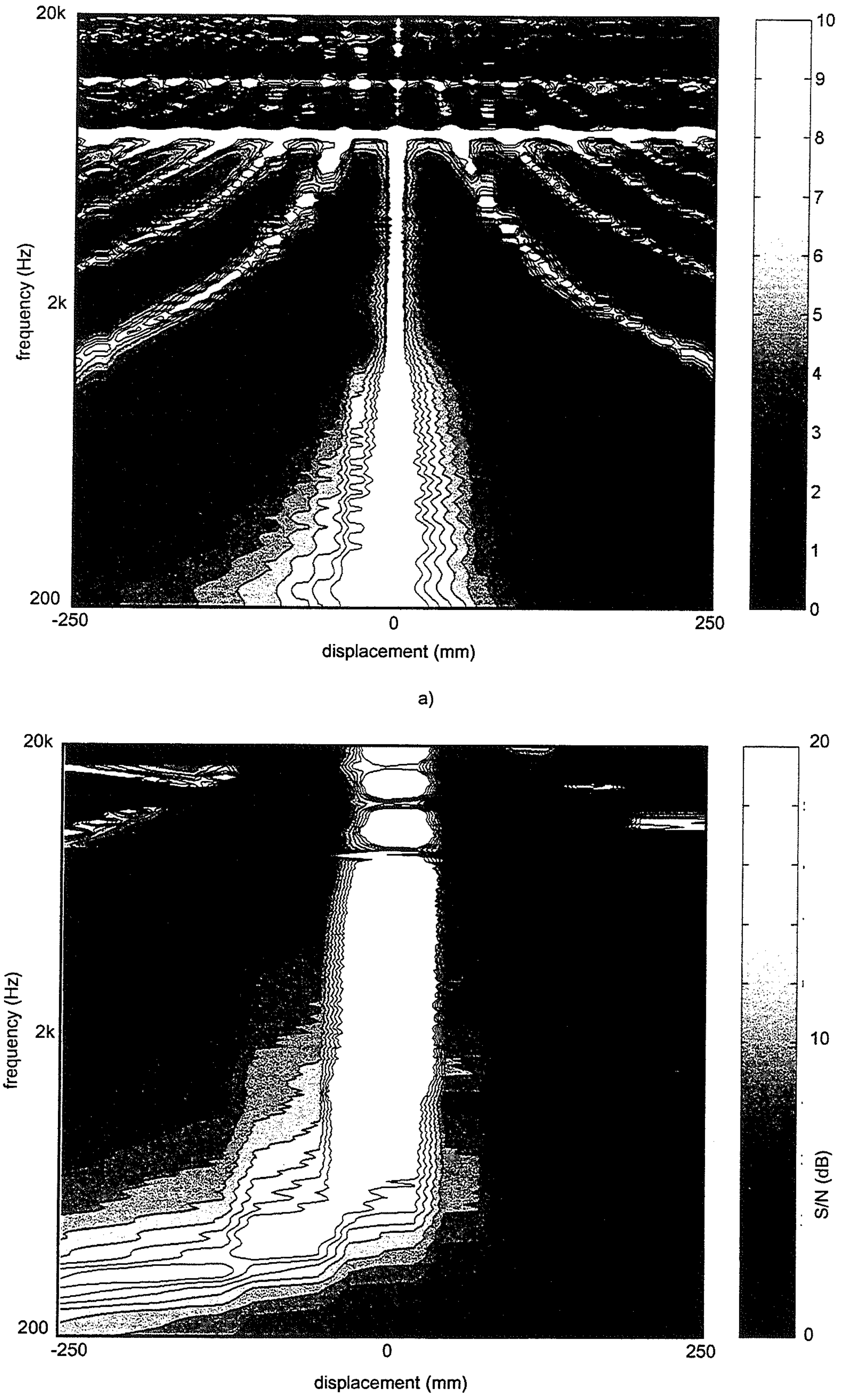


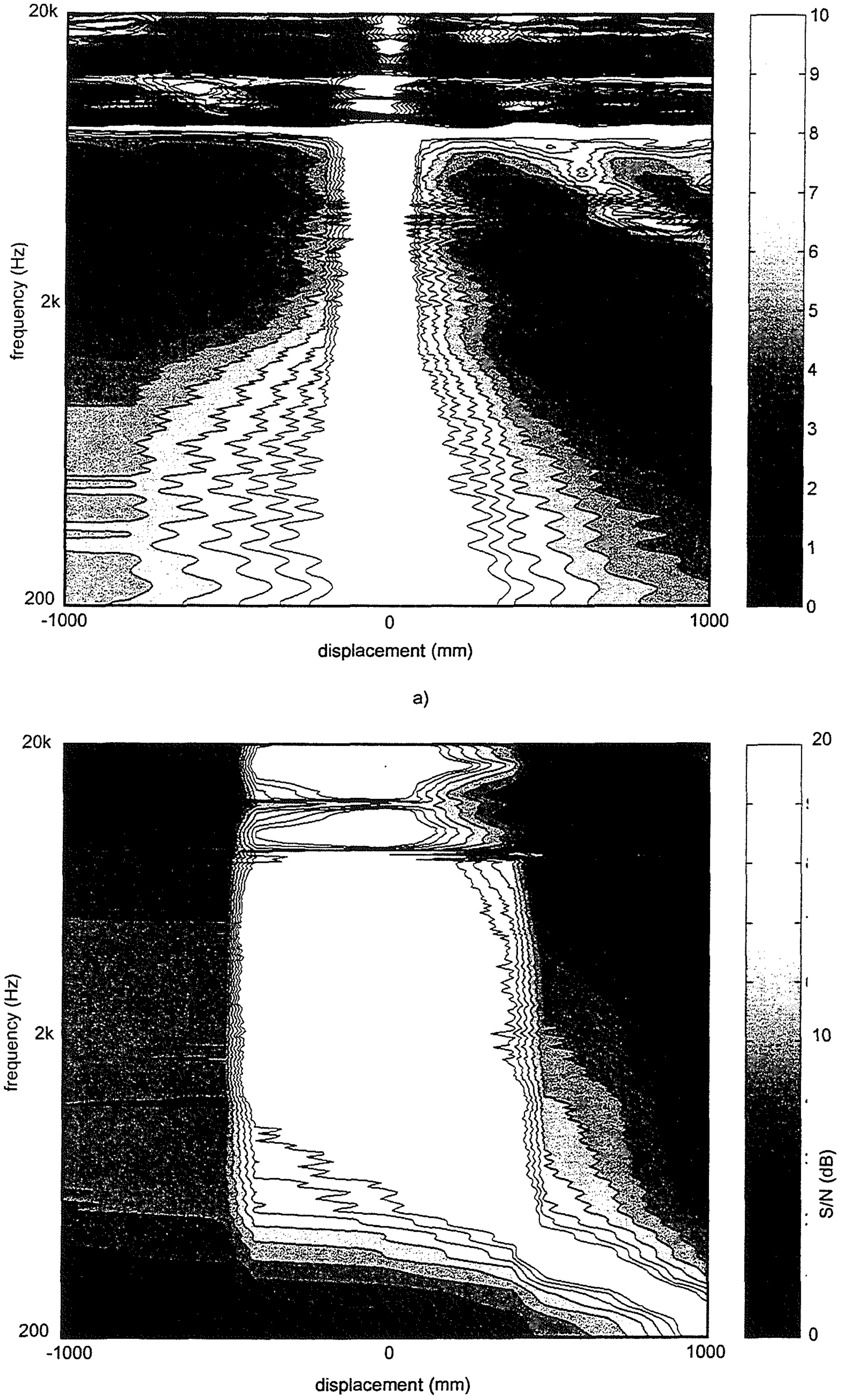
Fig. 5.11. Estimated ITD, plotted as a function of the intended direction of the virtual sound source. View from the upper-rear-right (azimuth=45°, elevation=150°). a) estimated from the original HRIRs. b) estimated from synthesised HRIRs with 25mm lateral displacement for the 60° transducer span. c) estimated from synthesised HRIRs with 25mm lateral displacement for the 10° transducer span.



a)

b)

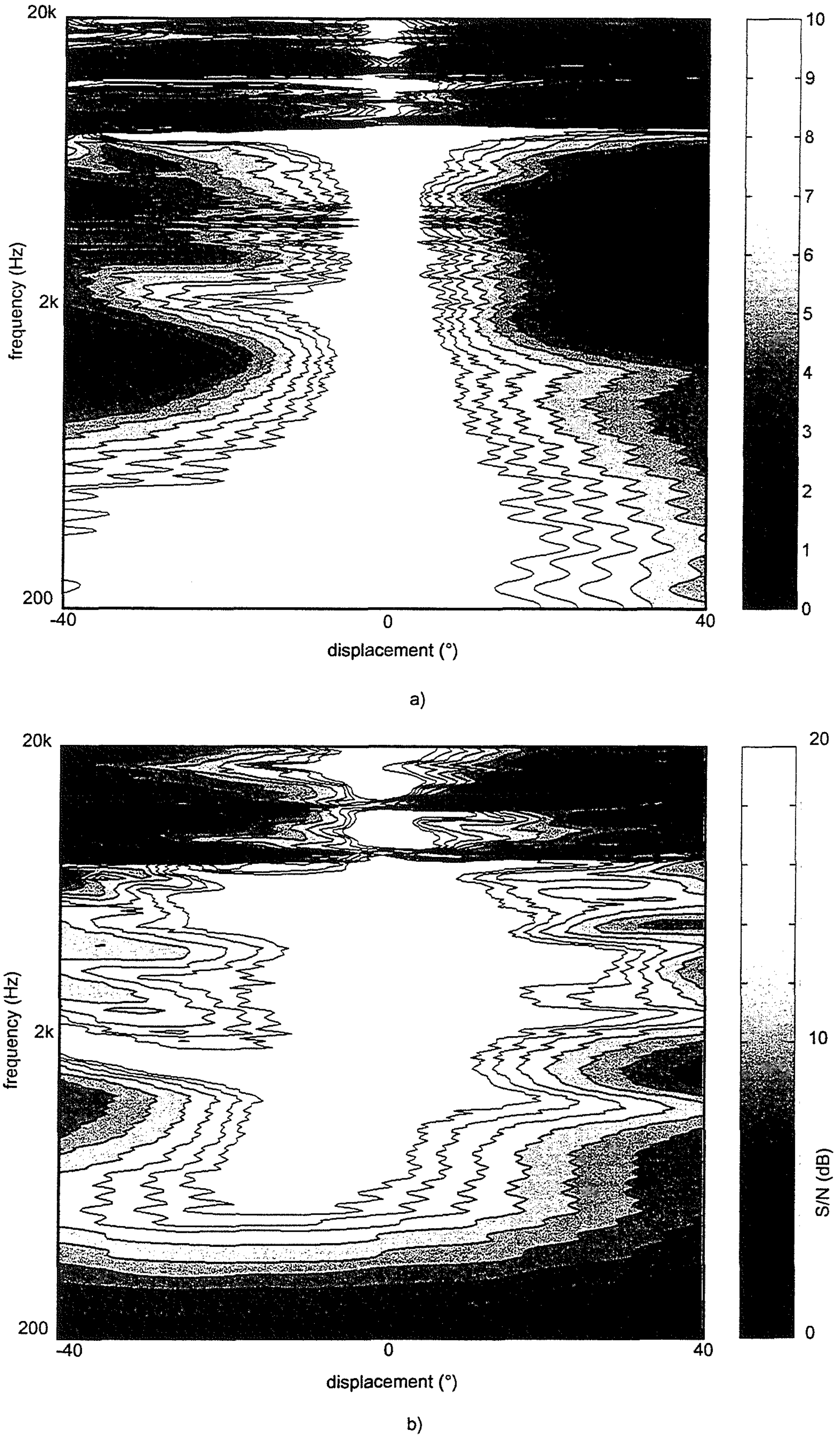
Fig. 5.12. Signal to noise ratio for the HRTF synthesis at the left ear as a function of lateral displacement. a) 60° transducer span. b) 10° transducer span.



a)

b)

Fig. 5.13. Signal to noise ratio for the HRTF synthesis at the left ear as a function of fore-and-aft displacement. a) 60° transducer span. b) 10° transducer span.



a)

b)

Fig. 5.14. Signal to noise ratio for the HRTF synthesis at the left ear as a function of yaw displacement. a) 60° transducer span. b) 10° transducer span.

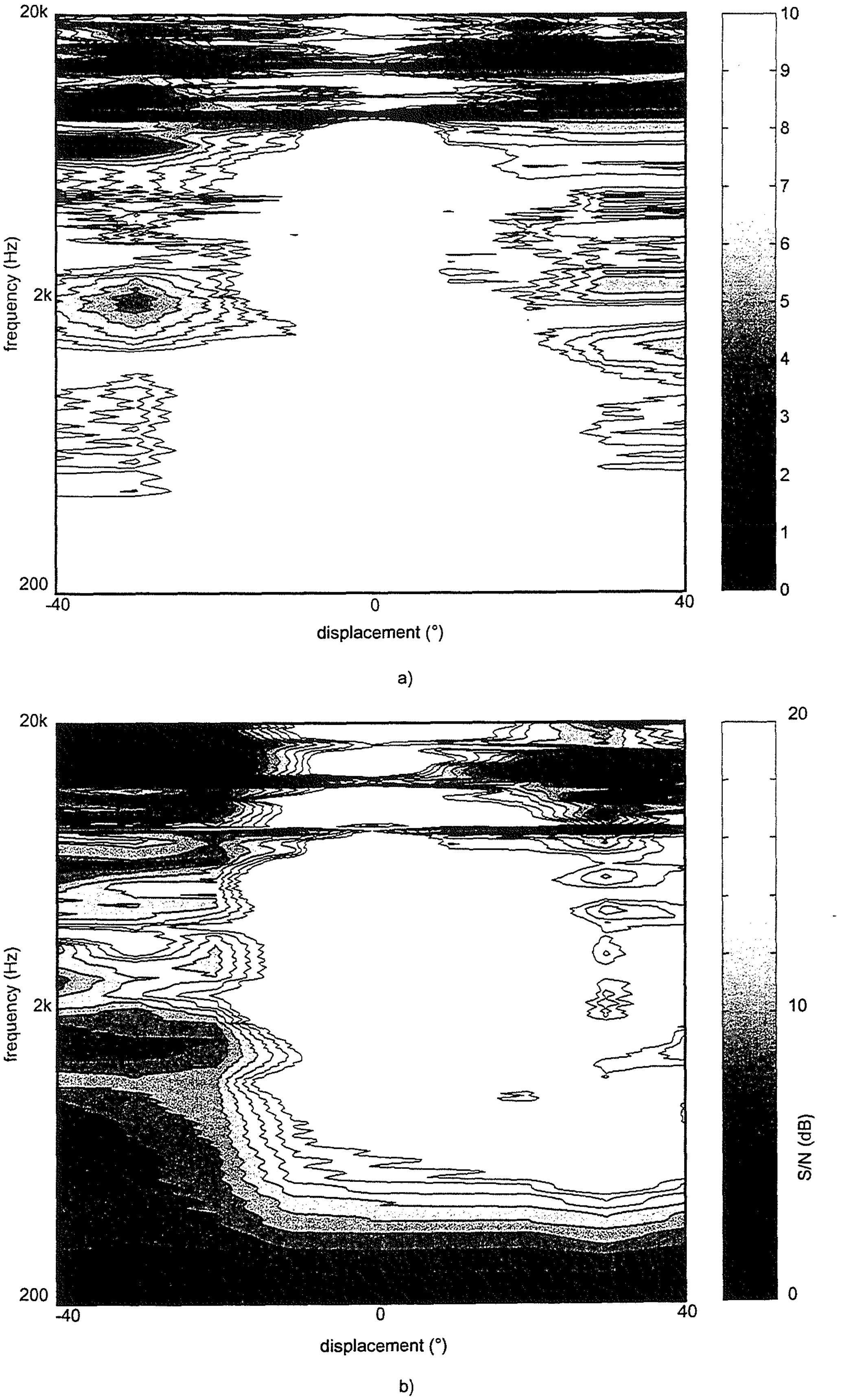


Fig. 5.15. Signal to noise ratio for the HRTF synthesis at the left ear as a function of pitch displacement. a) 60° transducer span. b) 10° transducer span.

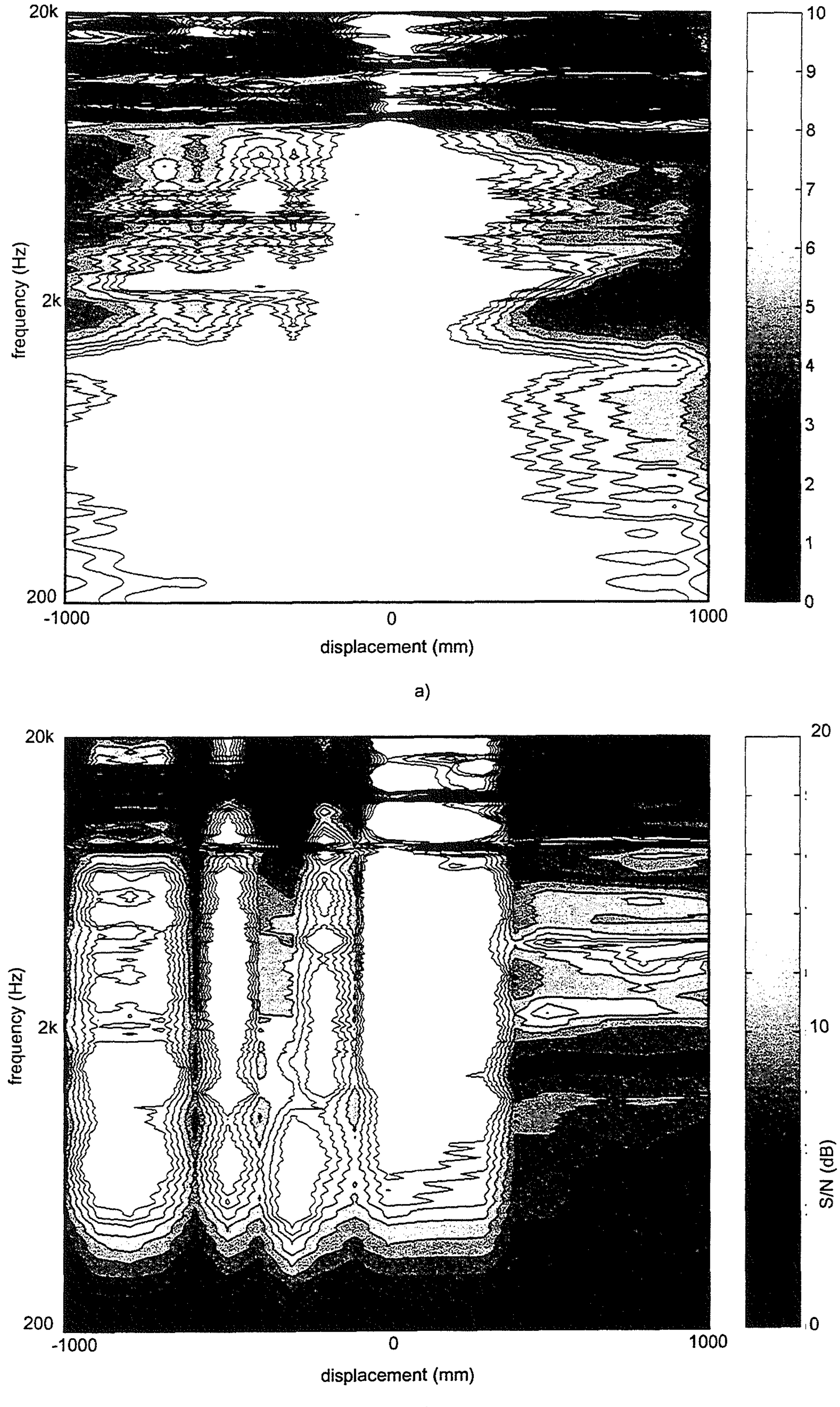
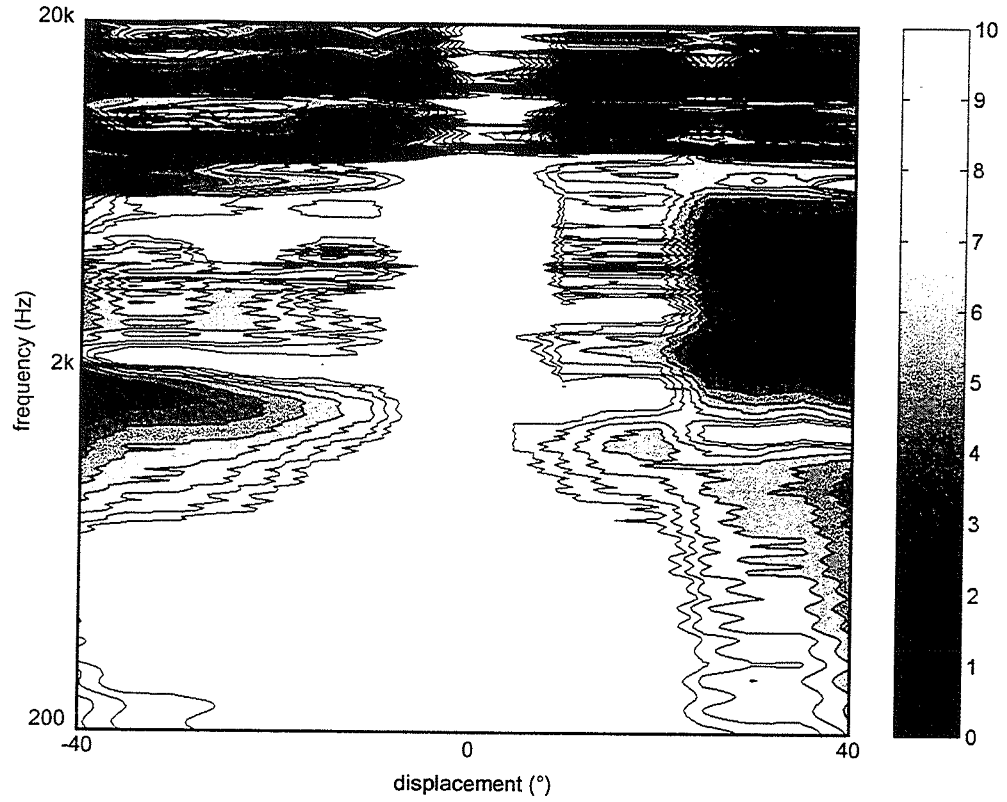
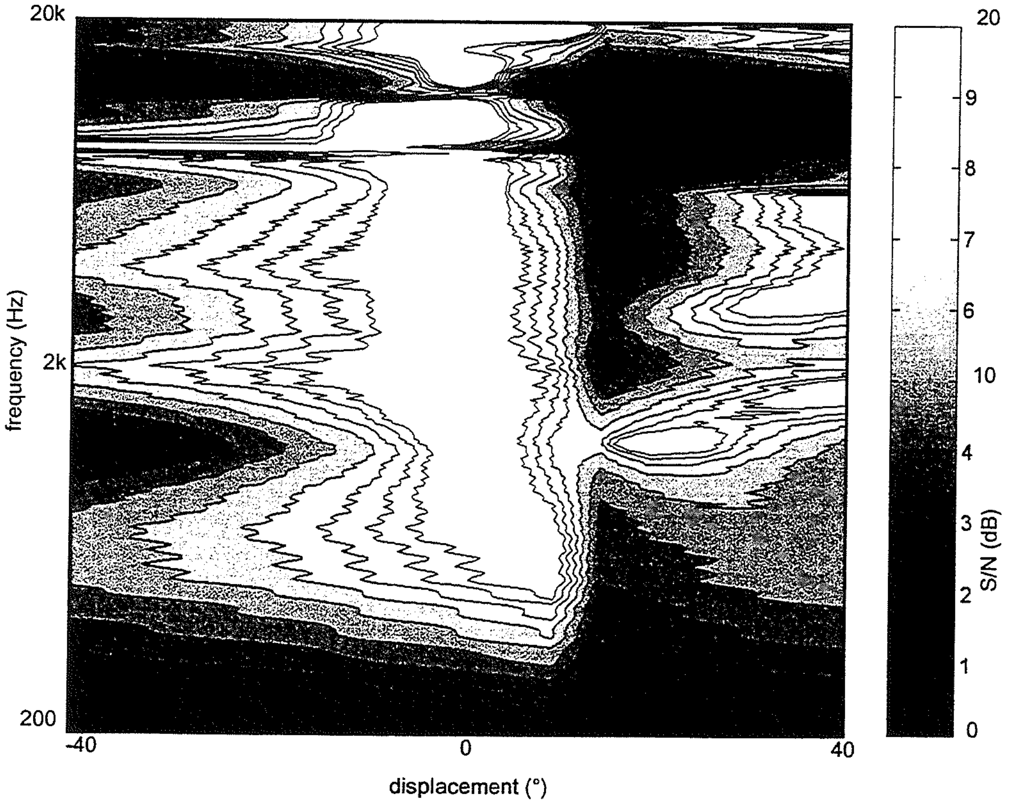


Fig. 5.16. Signal to noise ratio for the HRTF synthesis at the left ear as a function of vertical displacement. a) 60° transducer span. b) 10° transducer span.



a)



b)

Fig. 5.17. Signal to noise ratio for the HRTF synthesis at the left ear as a function of roll displacement. a) 60° transducer span. b) 10° transducer span.

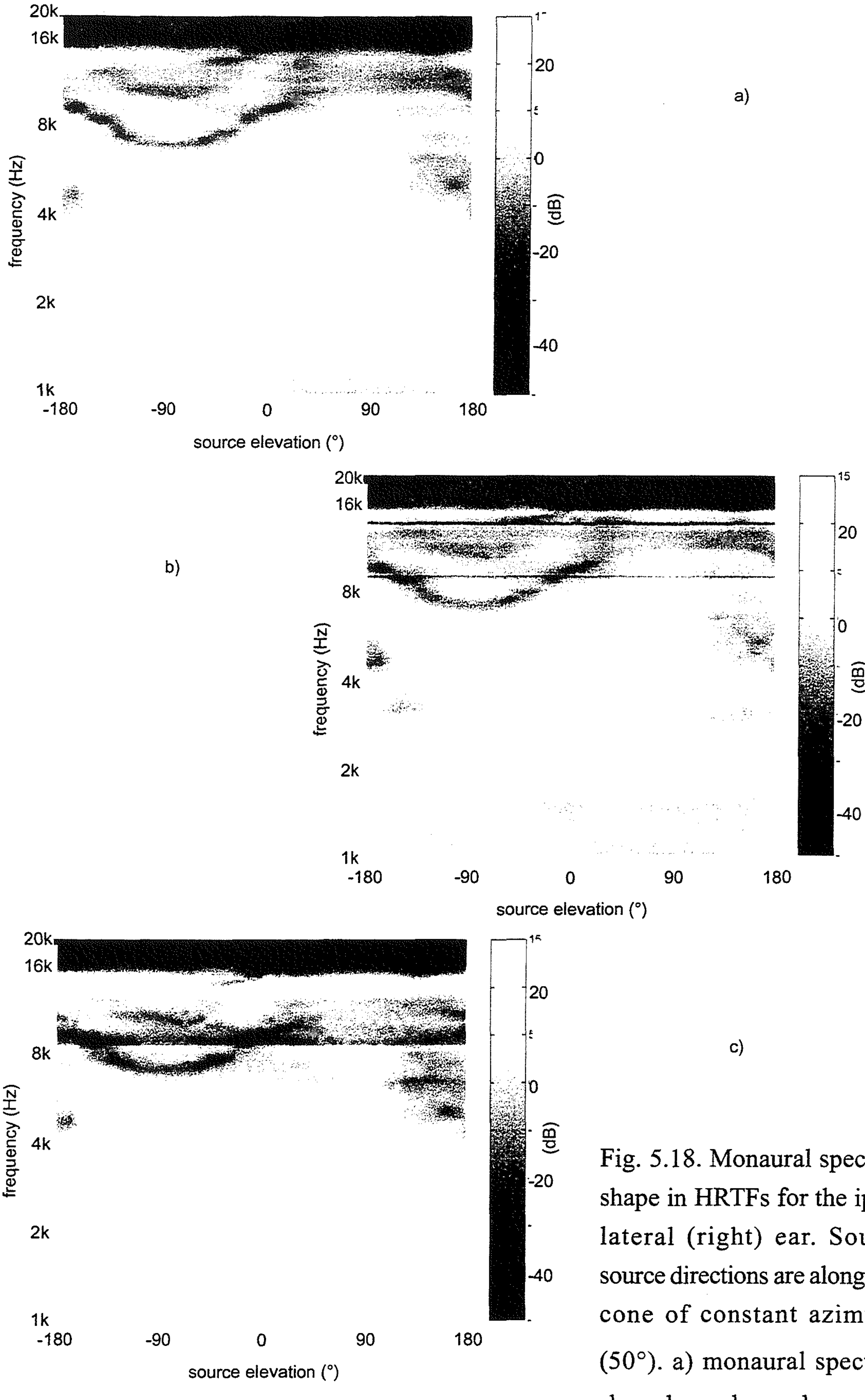


Fig. 5.18. Monaural spectral shape in HRTFs for the ipsilateral (right) ear. Sound source directions are along the cone of constant azimuth (50°). a) monaural spectral shape by real sound sources. b) monaural spectral shape synthesised by the 60° transducer arrangement. The listener's head is displaced 40mm laterally. c) monaural spectral shape synthesised by the 10° transducer arrangement. The listener's head is displaced 40mm laterally.

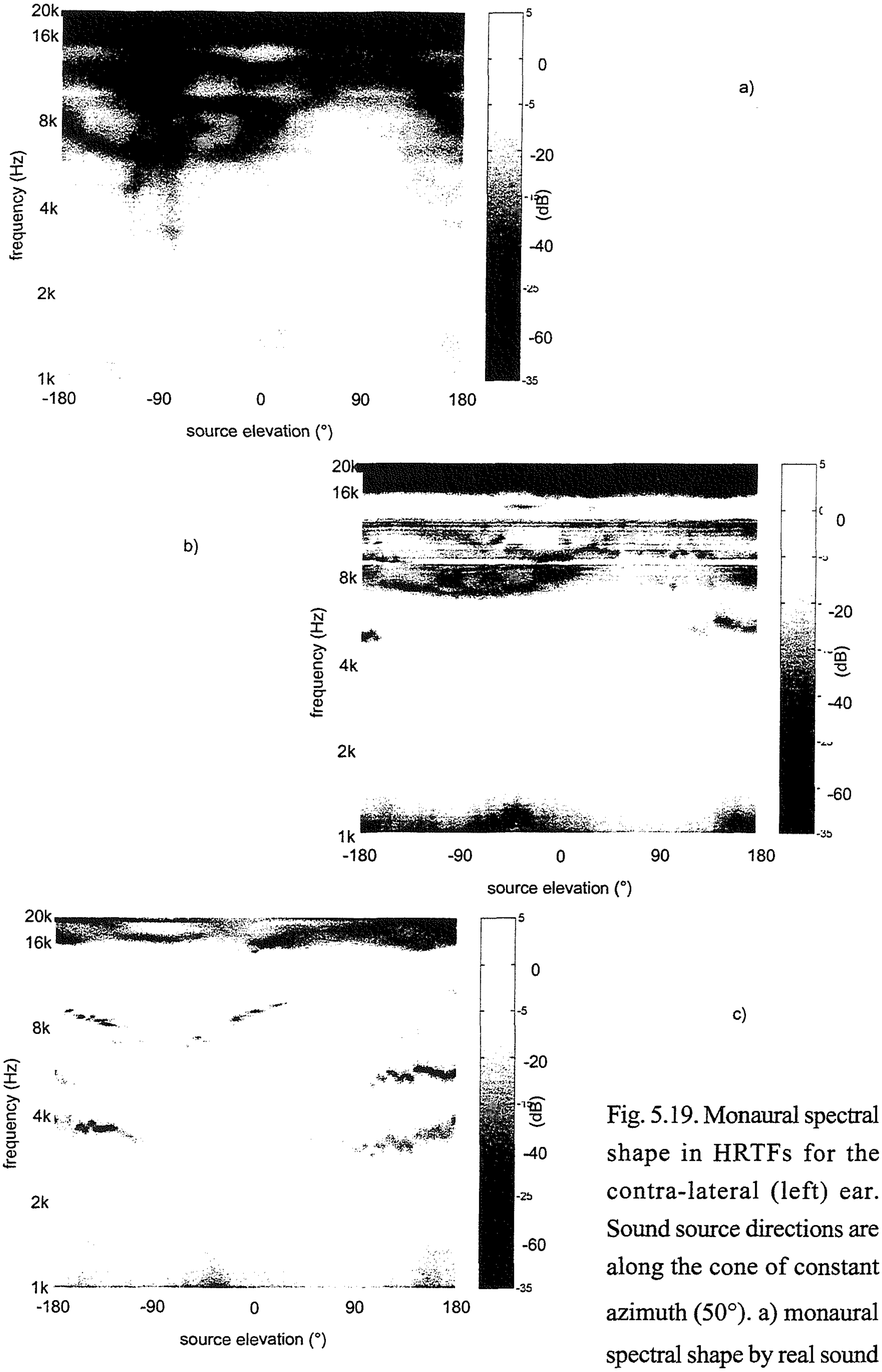


Fig. 5.19. Monaural spectral shape in HRTFs for the contra-lateral (left) ear. Sound source directions are along the cone of constant azimuth (50°). a) monaural spectral shape by real sound sources. b) monaural spectral shape synthesised by the 60° transducer arrangement. The listener's head is displaced 40mm laterally. c) monaural spectral shape synthesised by the 10° transducer arrangement. The listener's head is displaced 40mm laterally.

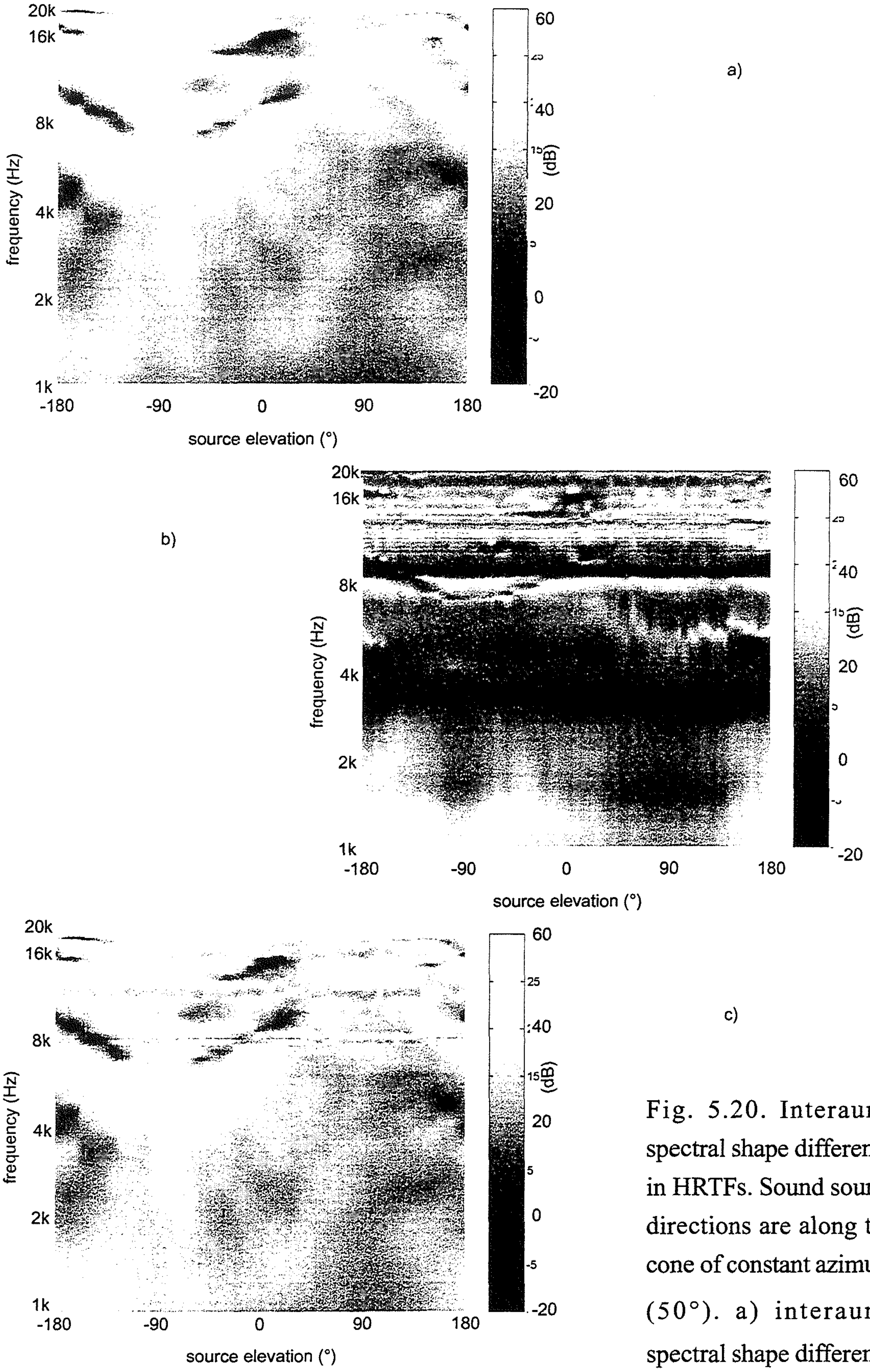


Fig. 5.20. Interaural spectral shape difference in HRTFs. Sound source directions are along the cone of constant azimuth (50°). a) interaural spectral shape difference by real sound sources. b) interaural spectral shape difference synthesised by the 60° transducer arrangement. The listener's head is displaced 25mm laterally. c) interaural spectral shape difference synthesised by the 10° transducer arrangement. The listener's head is displaced 25mm laterally.

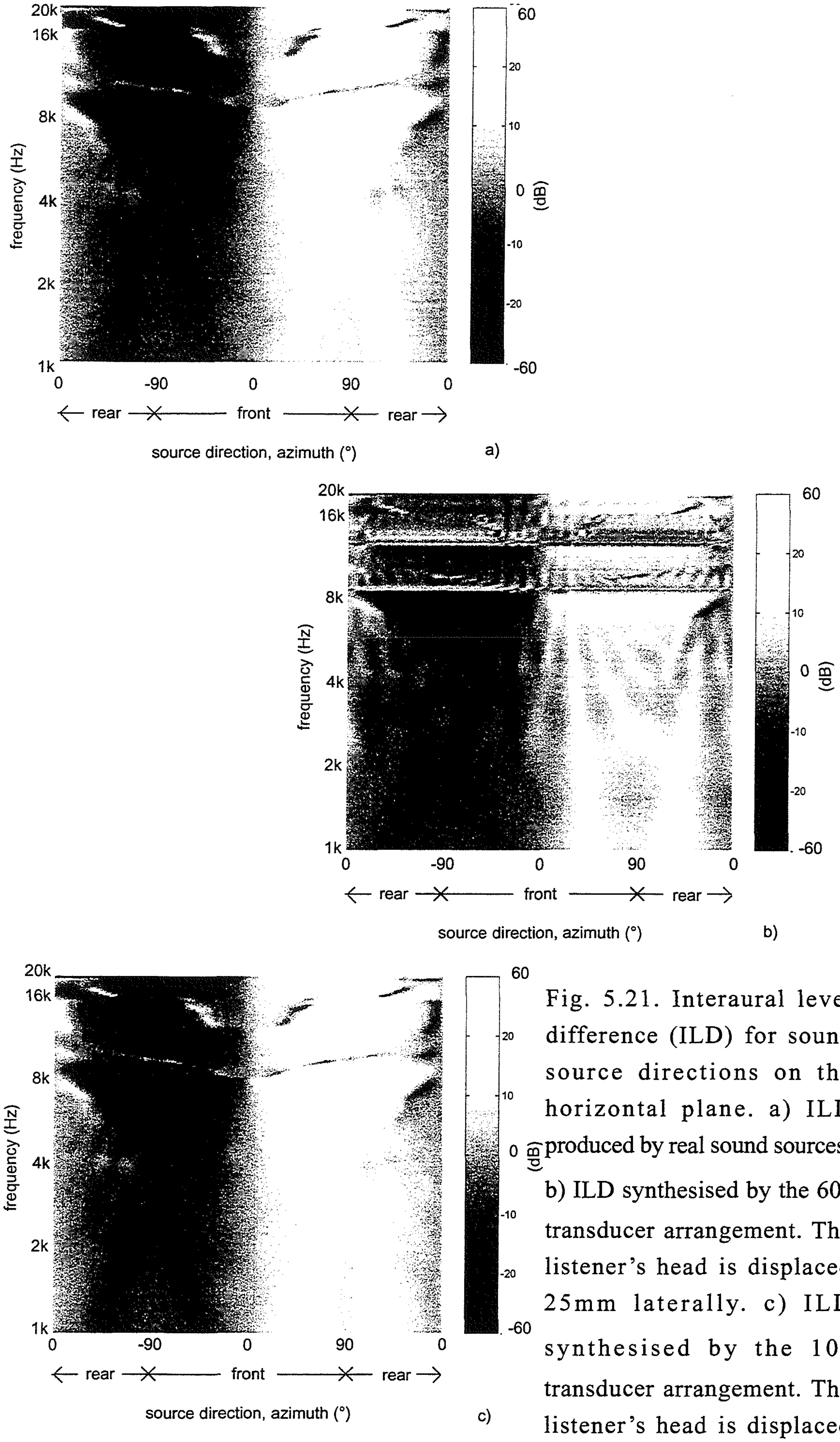
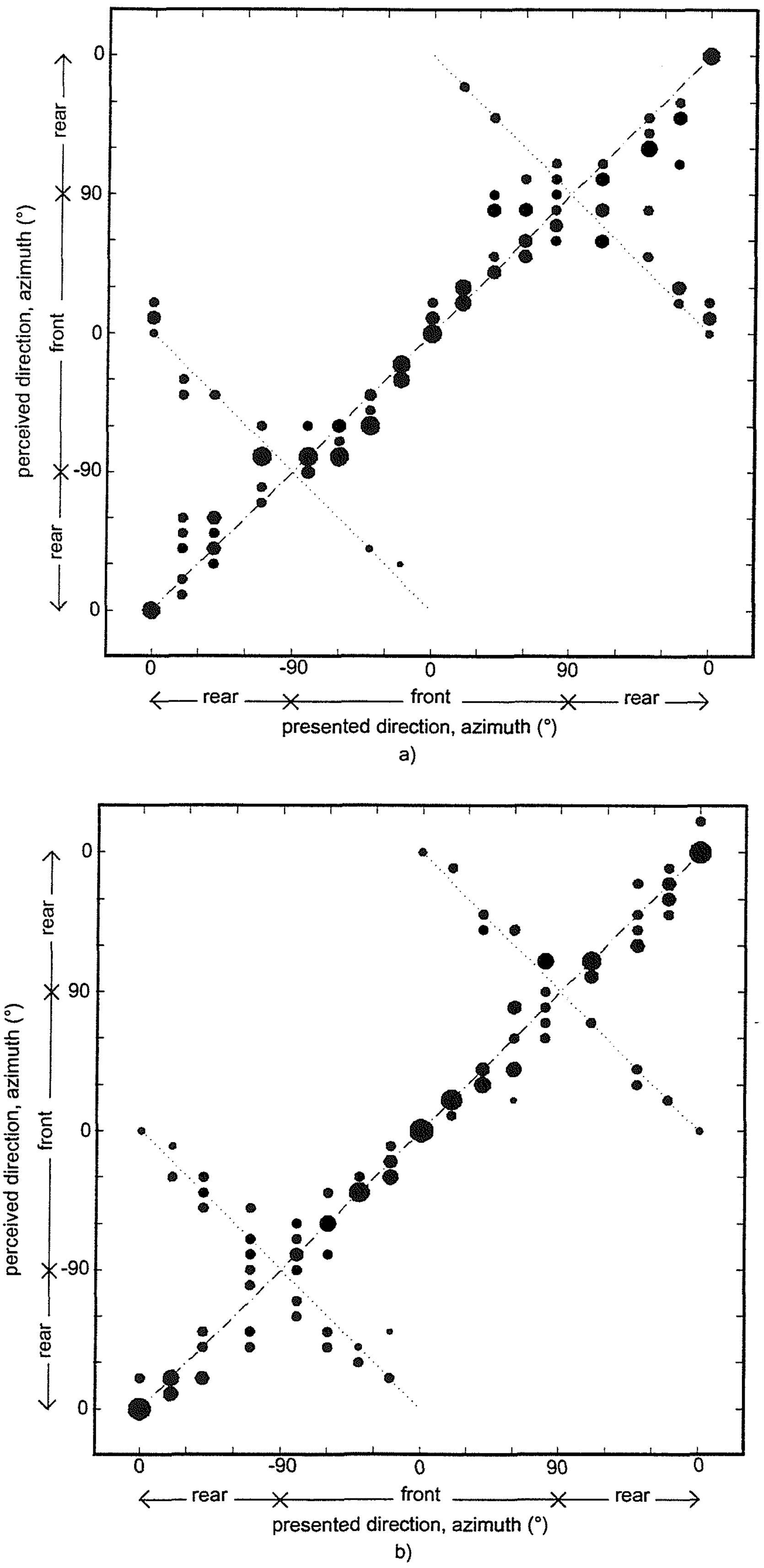


Fig. 5.21. Interaural level difference (ILD) for sound source directions on the horizontal plane. a) ILD produced by real sound sources. b) ILD synthesised by the 60° transducer arrangement. The listener's head is displaced 25mm laterally. c) ILD synthesised by the 10° transducer arrangement. The listener's head is displaced 25mm laterally.



a)

b)

Fig. 5.22 Results of the localisation experiment with binaural synthesis over loudspeakers when the listener's head is displaced 50mm laterally (to the right). 7 subjects were tested. a) 60° transducer span. b) 10° transducer span.

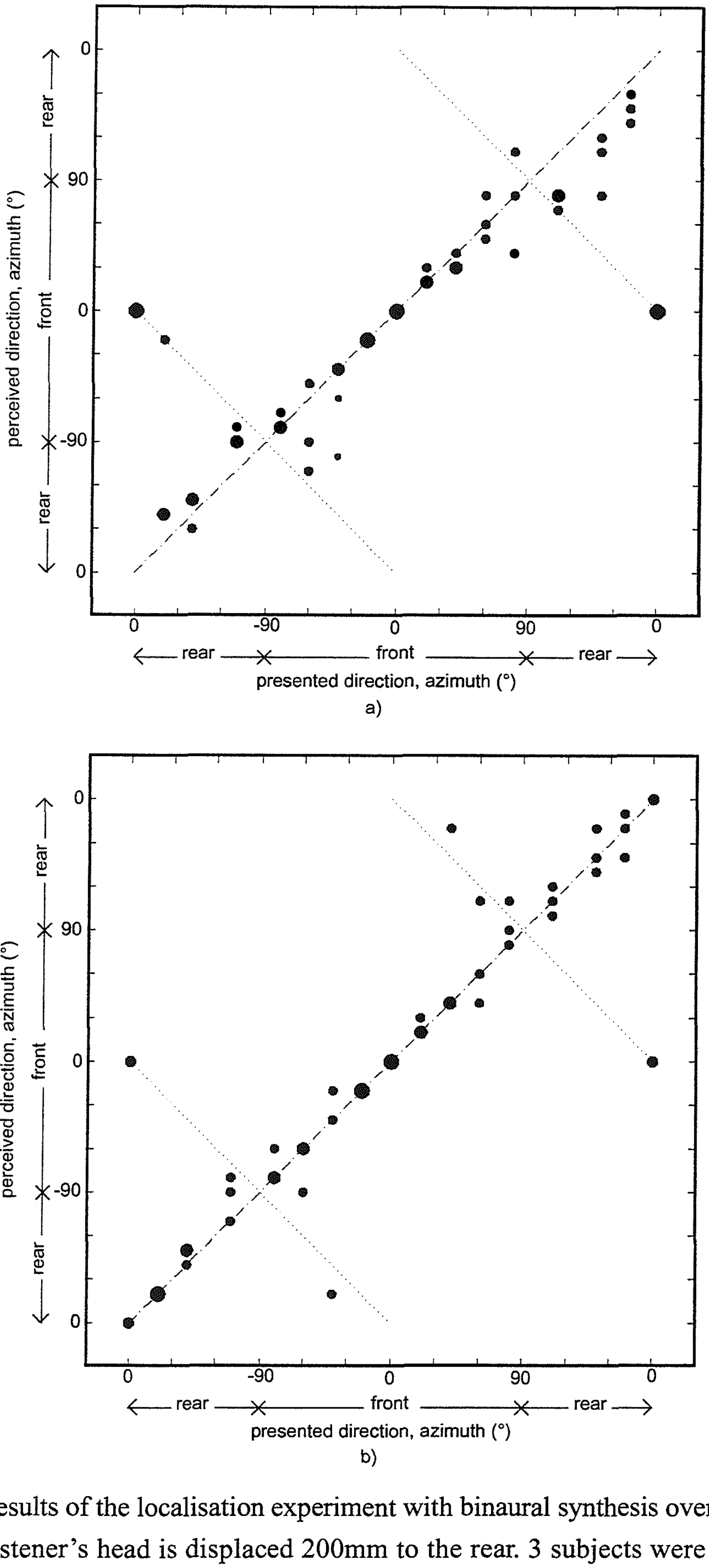


Fig. 5.23 Results of the localisation experiment with binaural synthesis over loudspeakers when the listener's head is displaced 200mm to the rear. 3 subjects were tested. a) 60° transducer span. b) 10° transducer span.

6 Effects of reflections

6.1 Introduction

Since the sounds generated by the loudspeakers are radiated into a listening space, the performance of the system is affected by the change of environment. The objective of this Chapter is to investigate the effect of reflections on the performance of systems for the binaural synthesis over loudspeakers. When we talk about the performance of the sound reproduction system, this includes the ability to reproduce the quality of sounds as well as the spatial impression of the original sound field. However, we shall again concentrate on how spatial impression is preserved, and more strictly, how much a system preserves the ability to localise a single reproduced sound source.

As long as the effect of reflections is small, sound localisation is expected to be well retained. As the effect of reflection gets stronger, frontal localisation will be degraded severely because the frequency response of the room modifies the spectrum of the sound which is the main cue for front-back discrimination. However, the ability of subjects to perform lateral localisation will be preserved to some extent by the precedence effect.

As a basic investigation, the effect of a single reflecting surface was examined. To begin with, simulations based on geometrical acoustics were carried out. Since the inter-aural time and level differences and the spectra of the sound are said to be the cues for sound localisation, the time history and the spectra of the sound were examined. The effects of an infinite uniform reflecting surface which is parallel to the front-back axis with respect to the listener are examined. The factors which have been changed here are the reflection coefficient, the distance between the listener and the reflecting surface, and the relative level between left and right ear. Then, subjective experiments were undertaken to examine the effect of different reflection coefficients.

6.2 Model for analysis

6.2.1 Model of room reflections

When the system containing a matrix of inverse filters **H** for the plant transfer functions of an anechoic environment is brought into a normal listening room, reflections from the room surfaces degrade the performance of the system. The system in a reverberant environment can be considered to be equivalent to the system which has a pair of control sources which aim to reproduce signals at the listener's ears plus infinite pairs of mirror image control sources. Reproduction sources and their corresponding mirror image sources are fed with same input signals **v** (Fig. 6.1). S_{m0} are the original reproduction sources and S_{mk} are the mirror image sources.

Then, from Eq. (2.5a, b, c), the vector of the reproduced signals can be written as

$$\mathbf{w} = [\mathbf{C}_0 + \sum_{k=1}^{\infty} \mathbf{\Omega}_k]\mathbf{v} \tag{6.1}$$

where $\mathbf{C}_0$ is the transfer function matrix associated with the original reproduction sources and $\mathbf{\Omega}_k$ is the transfer function matrix associated with the image sources (kth order) which includes the reflection coefficients of walls. The matrix $\mathbf{\Omega}_k$ has the following structure.

$$\mathbf{\Omega}_k = \begin{bmatrix} \mathrm{R}_{k11}(\mathrm{j}\omega)\mathrm{C}_{k11}(\mathrm{j}\omega) & \cdots & \mathrm{R}_{k1S}(\mathrm{j}\omega)\mathrm{C}_{k1S}(\mathrm{j}\omega) \\ \vdots & \ddots & \vdots \\ \mathrm{R}_{kR1}(\mathrm{j}\omega)\mathrm{C}_{kR1}(\mathrm{j}\omega) & \cdots & \mathrm{R}_{kRS}(\mathrm{j}\omega)\mathrm{C}_{kRS}(\mathrm{j}\omega) \end{bmatrix} \tag{6.2}$$

where $R_{krs}(j\omega)$ are products of the complex reflection coefficients of the walls by which the kth reflection is produced. Strictly speaking, the $R_{krs}(j\omega)$ are unique to each element of the matrix because they are angularly dependent. However, in this simulation, $R_{krs}(j\omega)$ shall be considered as independent of incident angle. This is the case when the sound path is relatively long compared to the distances between transducers or sources, or when reflection coefficients are angularly independent. As a result, this can be considered as a single complex number $R_k(j\omega)$. Then, Eq. (6.1) becomes

$$\mathbf{w} = [\mathbf{C}_0 + \sum_{k=1}^{\infty} \mathrm{R}_k(\mathrm{j}\omega)\mathbf{C}_k]\mathbf{v} \tag{6.3}$$

where $\mathbf{C}_k$ is the plant transfer function matrix associated with image sources of the kth reflection. This excludes the reflection coefficients of the walls and has the following structure.

$$\mathbf{C}_k = \begin{bmatrix} \mathrm{C}_{k11}(\mathrm{j}\omega) & \cdots & \mathrm{C}_{k1S}(\mathrm{j}\omega) \\ \vdots & \ddots & \vdots \\ \mathrm{C}_{kR1}(\mathrm{j}\omega) & \cdots & \mathrm{C}_{kRS}(\mathrm{j}\omega) \end{bmatrix} \tag{6.4}$$

Fig. 6.2 shows the block diagram which illustrates this.

6.2.2 Model for a single reflection

As the simplest case having reflections, the effect of only one reflection will be evaluated first. In this case, Eq. (6.3) can be written as follows.

$$\mathbf{w}=[\mathbf{C}_0+\mathrm{R}_1(\mathrm{j}\omega)\mathbf{C}_1]\mathbf{v} \tag{6.5}$$

From Eq.(2.7), Eq. (6.5) can also be written as

$$\mathbf{w}=[\mathbf{C}_0+\mathrm{R}_1(\mathrm{j}\omega)\mathbf{C}_1]\mathbf{Hd}=\mathbf{C}_0\mathbf{Hd}+\mathrm{R}_1(\mathrm{j}\omega)\mathbf{C}_1\mathbf{Hd} \tag{6.6}$$

Given that with Eq.(2.11), to a good approximation,

$$\mathbf{C}_0\mathbf{H}\approx z^{-\Delta}\mathbf{I} \tag{6.7}$$

where **I** is the identity matrix, substituting this into Eq. (6.6) shows that

$$\mathbf{w}\approx z^{-\Delta}\mathbf{d}+\mathrm{R}_1(\mathrm{j}\omega)\mathbf{C}_1\mathbf{Hd} \tag{6.8}$$

The reproduced signals are the sum of the delayed versions of the desired signals and the signals which are reflected by the wall which degrades the performance of the system.

Hence

$$\mathbf{X}\approx z^{-\Delta}\mathbf{I}+\mathrm{R}_1(\mathrm{j}\omega)\mathbf{C}_1\mathbf{H} \tag{6.9}$$

The signals which should be fed into image sources are written as follows;

$$\mathrm{v}_1 = \mathrm{R}_1(\mathrm{j}\omega)\mathbf{Hd} = \mathrm{R}_1(\mathrm{j}\omega)\mathbf{v} \tag{6.10}$$

6.3 Analysis

An infinite uniform reflecting surface in an anechoic environment which is parallel to the front-back axis with respect to the listener was simulated. The geometry for the simulation is illustrated in Fig. 6.3. The distance between each reproduction transducer and the middle of the listener's two ears is 1.4 m. The two transducers are located at symmetric positions with respect to the axis. The spanning angle was chosen to be 10° as an example. This arrangement is adapted from the "Stereo Dipole" system described in Chapter 3. The transducers used to simulate a reflection are placed at the mirror image position with respect to the wall to be simulated. For simplicity, the reflection coefficient $\mathrm{R}_1(\mathrm{j}\omega)$ is assumed to be equal for all the frequencies with a real number R in this simulation.

The HRTFs for the electroacoustic transfer functions matrix $\mathbf{C}_k$ were taken from the MIT Media Lab's database which has been made available for researchers over the Internet [38]. The "compact" version of the database was used. The corresponding HRTFs data is taken from the database for a certain direction of the control transducer. Transfer functions for the transducers at a direction between sampled directions are obtained by bilinear interpolation on the virtual spherical surface of magnitude and phase spectra in the frequency domain. Those for the transducers at different distance from a head are obtained by extrapolation with an appropriately chosen delay and spherical attenuation (Appendix 1).

The filters in the matrix **H** were obtained by the fast deconvolution method which is fully explained in reference [31]. The MIT Media Lab's database was also used at the control filter design stage. Each control filter has 1024 coefficients. The control performance matrix **X** is analysed for most of the cases unless otherwise stated in order to enable basic investigation. The factors which have been changed here are the reflection coefficient R and the distance between the listener and the reflecting surface.

6.3.1 The performance of the reproduction system in an anechoic environment

The frequency response and impulse response of the transfer function between the input of the system and the ears of the listener in an anechoic environment are shown in Fig. 6.4. The performance of cross-talk cancellation was about 20dB for most of the frequency range between 1kHz and 20kHz and gradually deteriorated below 1kHz.

Fig. 6.5 shows the reproduced time history of the signal when an impulse is fed to the input of the system. The logarithm of the squared sound pressure is plotted. The level of the desired signal was reproduced about 3 dB lower than the target signal due to the inversion. The signal to noise (error signal) ratio (S/N) is more than 25dB for the left ear and noise floor is about -40dB for the other ear. However, these error signals may not be perceived by the listener owing to forward and backward masking [48]. Fig. 6.6 shows the spectra of the reproduced signals. The almost flat spectrum was obtained for the left ear though the residual error signal for the right ear increases as the frequency decreases. Fig. 6.5 and Fig. 6.6 shows the maximum performance of the cross-talk cancellation of this reproduction system.

6.3.2 The general effects of a reflection

Fig. 6.7 shows the spectra and the time history of the reproduced signal when an infinite reflecting surface was introduced. The distance between the listener and the reflecting surface is 1.0m and the reflection coefficient of the surface is R = 1.0. In this arrangement the reflecting surface is placed at the opposite side of the ear to which the pure impulse is to be fed. The reflected signal can be seen following the direct sound signal in the time history of the signals. The error signal due to the reflection is bigger at the right ear than that of the left ear. At the left ear, the error signal due to the reflection is considerably smaller than the direct sound and it arrives about 8ms after the direct sound. Therefore, this error signal is likely to be masked by the direct sound. On the other hand, the error at the right ear is increased from about -40dB to more than -10dB. This can potentially be a problem.

The result when the reflecting surface is placed at the same side of the ear which the pure impulse is to be fed is shown in Fig. 6.8. Although the error signals are similar to the former result, the effect of reflection seems to be less since the larger error signal was hidden by the direct sound. In real life, this extreme situation, that is, the sound level at one ear is 40dB lower than at the other hardly happens. It can be concluded that a reflecting surface placed at the same side of the ear which is to receive a smaller signal causes more problems than in the opposite case.

In general, the spectra of the reproduced sound are degraded severely. In the time history, it can be seen that the reflected signal with long response follows the direct sound signal. The long response of the reflected signal is due to the response of filters in **H** which is not cancelled because the transfer functions of reflection path are different from that of direct path. Therefore, it is expected that even when the precedence effect helps a listener to localise sound sources correctly, the listener will get some reverberant feeling. However, it is also possible that this 'artificial reverberation' might make the listener have an impression of the reproduced sounds improved.

6.3.3 The effect of the reflection coefficient

The reproduced signals were simulated by changing its reflection coefficient. The distance between the listener and the reflecting surface was fixed to be 1m. It is obvious that when the reflection coefficient is small, the error due to the reflection will be small. The reflection coefficient affects the level of the reflected signal but does not affect the delay of the signal. It is said that the reflected sound should be 10dB - 15dB more intense than the direct sound in order to override the precedence effect [48]. The level difference between the two ears is mostly less than 30dB in the case of real sound sources. Therefore, the reflected sound can hardly be 10dB louder than the direct sound even in the case shown in Fig. 6.9. In this case, the target impulse for the right ear is about 20dB smaller than that for the left ear. Therefore, the reflection coefficient might not be important from the point of view of the time domain cues. However, in this case, the reflected signal has the response of the inverse filter which has not been cancelled by the proper **C** matrix. Therefore, the total energy of the reflected signal is larger than that of the reflection in the real situation. This might alter the threshold of the precedence effect. This point might be clarified through subjective experiment.

The result when the reflection coefficient was set to be R = 0.4 is shown in Fig. 6.10. The inter-aural level difference is almost 0dB for most of the frequency range. The sound which has most of the power in the middle frequency range may not be localised correctly in this kind of situation. However, this situation can happen quite often.

Fig. 6.11 shows the reproduced spectra with R = 0.1. The spectra of the right ear are dominated by the reflected signal which is to produce the signal for the left ear. The spectra for the right ear are about 10dB smaller than the spectra of the left ear at its maximum. In order to get good spectral information, the target signal for the right ears should be about 10dB larger than the error signal due to the reflection. That is, the signal should be about the same level as the signal for the other ear, even when the reflection coefficient is reasonably small. This is only the case for the sound located in median plane. This means that good spectral information cannot be obtained for most of the directions of the virtual source. In other words, it is difficult to avoid the degradation of spectra by simply changing the reflection coefficient.

6.3.4 The effect of the distance from the reflecting surface

When the distance between the listener and the reflecting surface is changed, the delay, level and the direction of the reflected sound will be changed. Obviously, when the reflecting surface is moved further away, the delay of the reflected sound increases. When the distance between the listener and the reflecting surface was changed from 0.5 m to 2.5 m, the resulting delay time of the reflected sound was varied from 5 ms to 15 ms. It is said that the upper limit of the precedence effect varies from 5 ms for a single click to 40 ms for sounds which have more complex character such as music and speech, depending on the types of the sound [48]. Therefore, the degradation of the ability to localise sound sources will be strongly dependent on the nature of sound.

6.4 Subjective evaluation

The analysis in the previous section suggests that the time domain cues would suffer from few problems due to the single reflection whereas the frequency domain cues would be severely degraded. Therefore, the effect of reflections on the virtual acoustic imaging system was investigated further by using subjective localisation experiments. Source directions on the horizontal plane were chosen to be examined since this covers the whole range of azimuth directions and two alternative elevation directions, i.e. 0° (front) and 180° (rear), in each cone of constant azimuth.

6.4.1 Experimental Procedure

The experimental procedure is fundamentally the same as that described in Section 3.4 except for a few points. The system used for the subjective experiments is illustrated in Fig. 6.12. Each control filter has 800 coefficients at sampling frequency of 44.1kHz. Geometrical arrangements of transducers used here are the same as those used in the analysis (Section 6.3). The wall to be simulated is located at left side of the subject. Therefore, the transducers used to simulate reflections are placed at the left side. The distance between the centre of the head and the reflecting surface was set to be 1.0m. Therefore, the distance between two pairs of loudspeakers was set to be 2.0m. The reflection coefficient of the imaginary wall R was set to be 0.0 (No reflection), 0.1, 0.3 and 1.0 in order to examine its effect. 13 European male subjects all with normal hearing were tested.

6.4.2 Experimental Results

Fig. 6.13 shows the results when there were no reflection (R = 0.0). It was found again that there are two groups of subjects. For one group the virtual acoustic imaging system works well, but does not work so effectively for the other group. Fig. 6.13a shows the results for those subjects for whom these systems work well. Much more front and back confusion can be observed compared to the previous experiments (Fig. 3.17, Section 3.4.3) but this is likely to be due to the shorter control filter length which resulted in inferior control performance. The results for those subjects for whom these systems do not work well are shown in Fig. 6.13b. They did not localise the virtual sound source at the rear half of the plane correctly and localised them at symmetric positions in the front. Moreover, the virtual sound sources at around θ=±90° are perceived at the offset position towards the centre (θ=0°).

Fig. 6.14a~c show the results with changing reflection coefficient. These experiments were carried out only with the 10 subjects for whom these systems work reasonably well. It can be seen that the ability of localisation becomes slightly less accurate as the reflection coefficient gets larger. However, as a whole, the ability to localise a sound source were preserved very well. It should be noted that there are no left and right confusion even when R = 0.3, which is the case that interaural level difference is almost 0 dB as simulated in Section 6.3.3. This suggests the superiority of time domain cues, especially the precedence effect, over the frequency domain cues. It is also noted that there are very slight left and right confusions when R = 1.0. However, this only happened for the angular positions of the virtual sources at opposite side of the reflecting surface as predicted in Section 6.3.2. The result of R = 1.0 also suggests that two independent systems working at the same time may work reasonably well.

6.5 Conclusions

The effect of reflections on the performance of the virtual acoustic imaging system was investigated by simulations and subjective experiments. An infinite uniform reflecting surface which is parallel to the front-back axis with respect to the listener was examined. The following were predicted from the known information on auditory function and the simulations.

(1) Whether the precedence effect helps the listener to localise sound source or not will be strongly dependent on the types of sound.

(2) The quality of sounds will be degraded by reflection even when the listener can localise the sound source correctly. However, the 'artificial reverberation', a by-product of the reflection might make a listener think that the sound is better.

(3) The inter-aural level difference can be almost 0dB or even negative for most of the frequency range when reflection exists. The sound for which inter-aural level difference is important for localisation may not be localised correctly.

(4) Having only a small level of reflected sound, good spectral information cannot be obtained for most of the angular positions of the virtual source.

The following predictions were confirmed by the subjective experiments.

(i) A reflecting surface placed at the same side of the ear which is to receive a smaller signal causes more problems than the opposite case.

(ii) A listener's ability to localise sound is preserved much more than is predicted from the spectral domain cues. The precedence effect seems to help preserve localisation ability.

As a whole, it was confirmed that the performance of the virtual acoustic imaging system is not degraded severely with existence of a single reflection.

Related publications

[A 7] T. Takeuchi, P.A. Nelson, O. Kirkeby and H. Hamada, "The Effects of Reflections on the Performance of Virtual Acoustic Imaging Systems", pp. 955-966, in Proceedings of the Active 97, The international symposium on active control of sound and vibration, Budapest, Hungary, August 21-23, (1997), OPAKFI

[A 8] Unpublished ISVR report

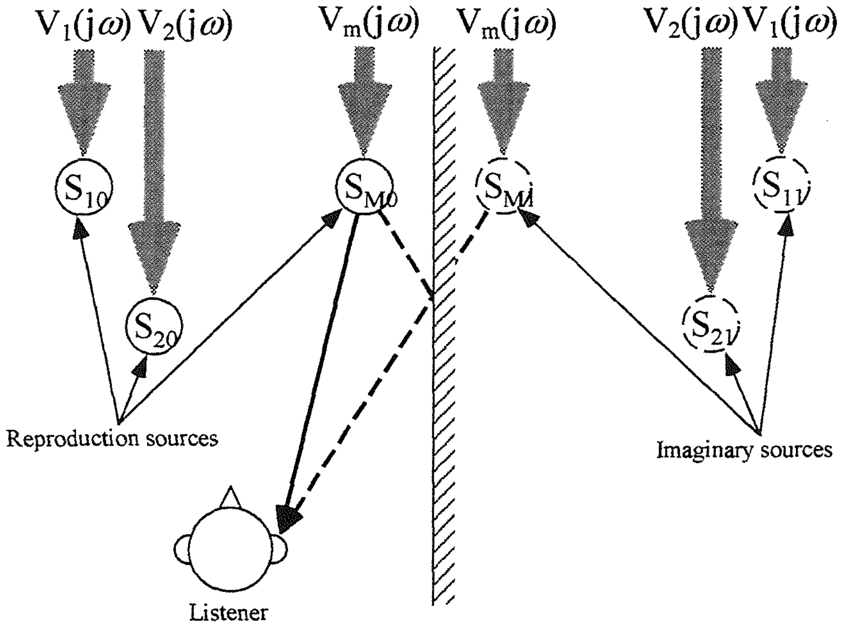


Fig. 6.1 Mirror image sources

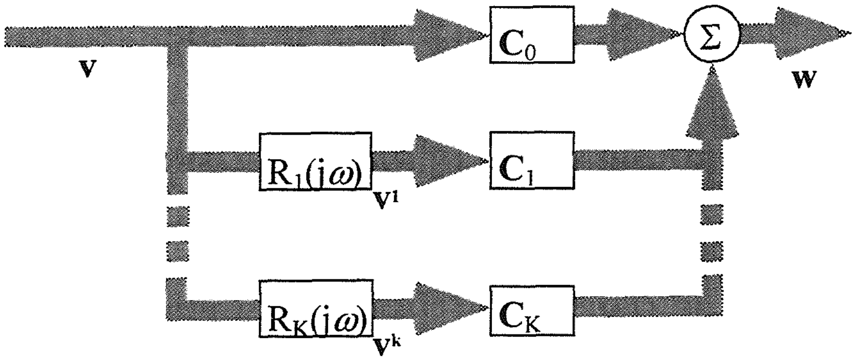


Fig. 6.2 Block diagram for Eq.(6.3)

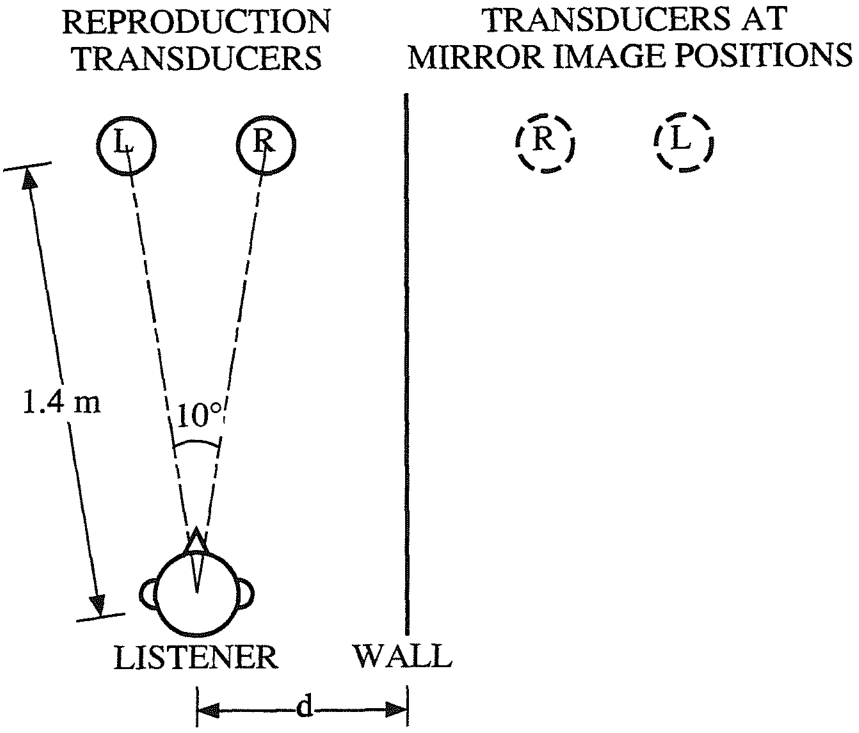


Fig. 6.3 Geometry for the simulations

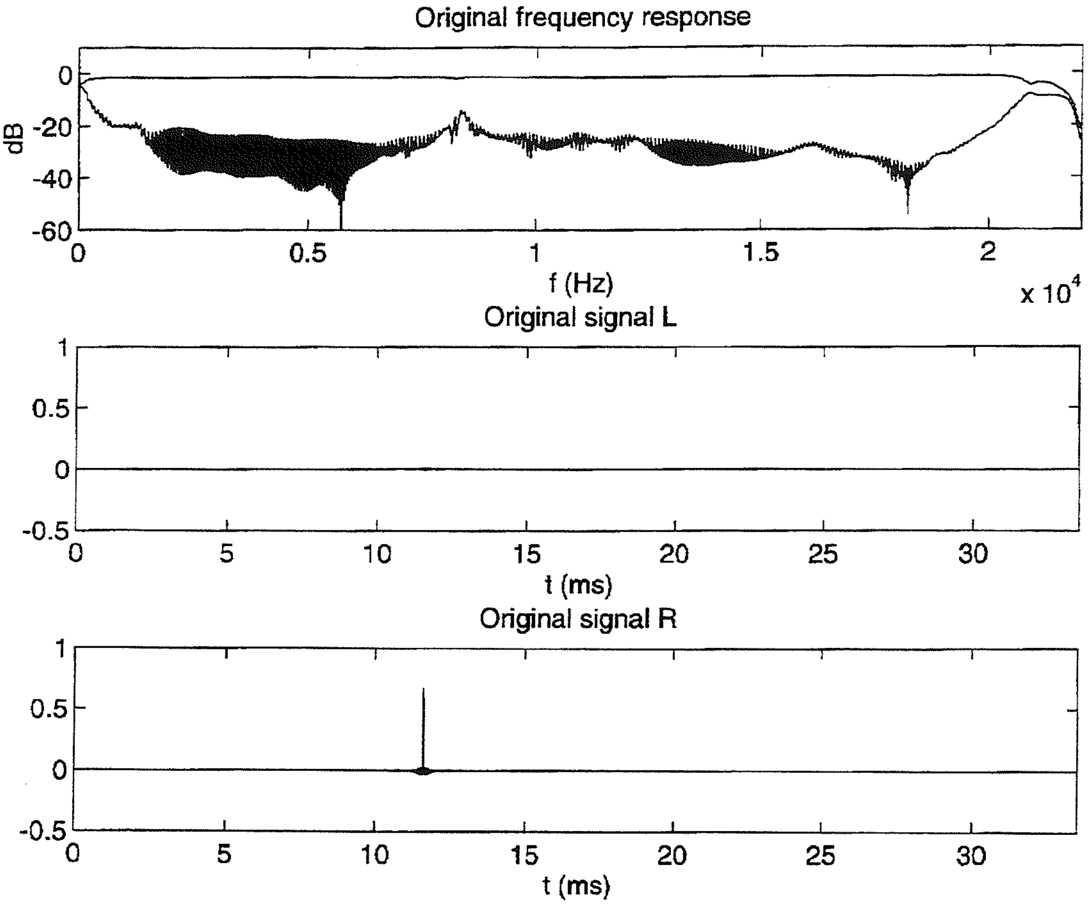


Fig. 6.4 Frequency response and impulse response of the transfer function between the input of the system and the ears of the listener without reflections

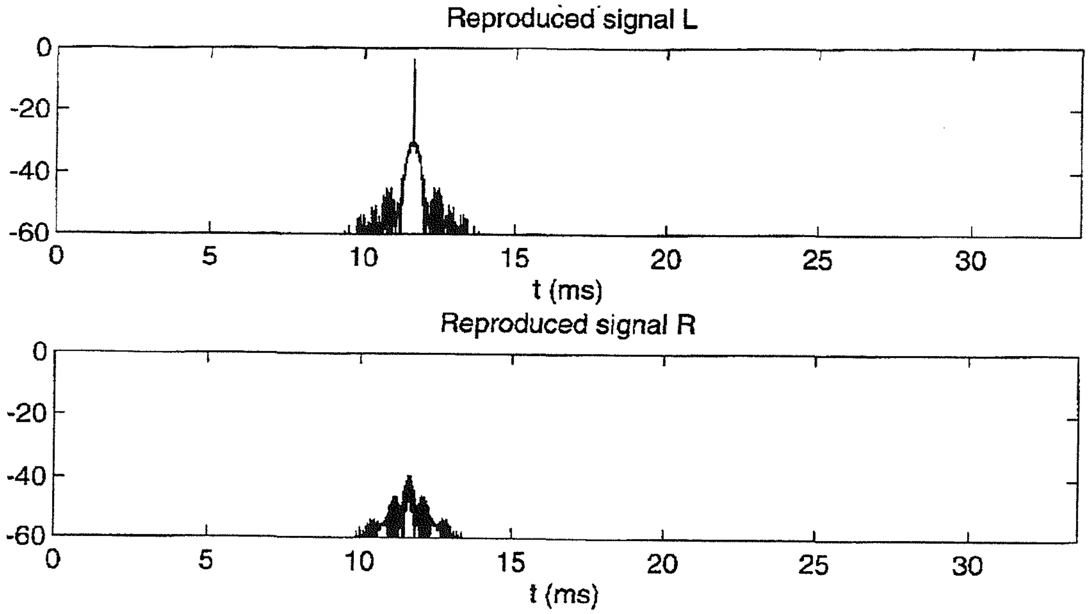


Fig. 6.5 Time history of the original reproduced signal (unit: dB)

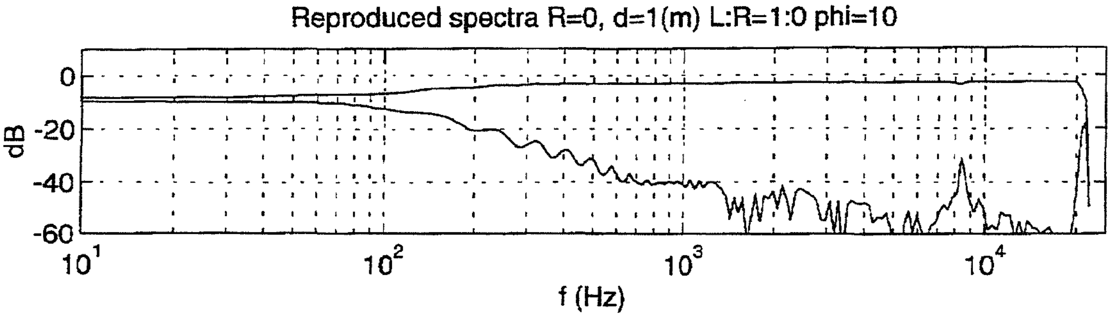


Fig. 6.6 Spectra of the original reproduced signal

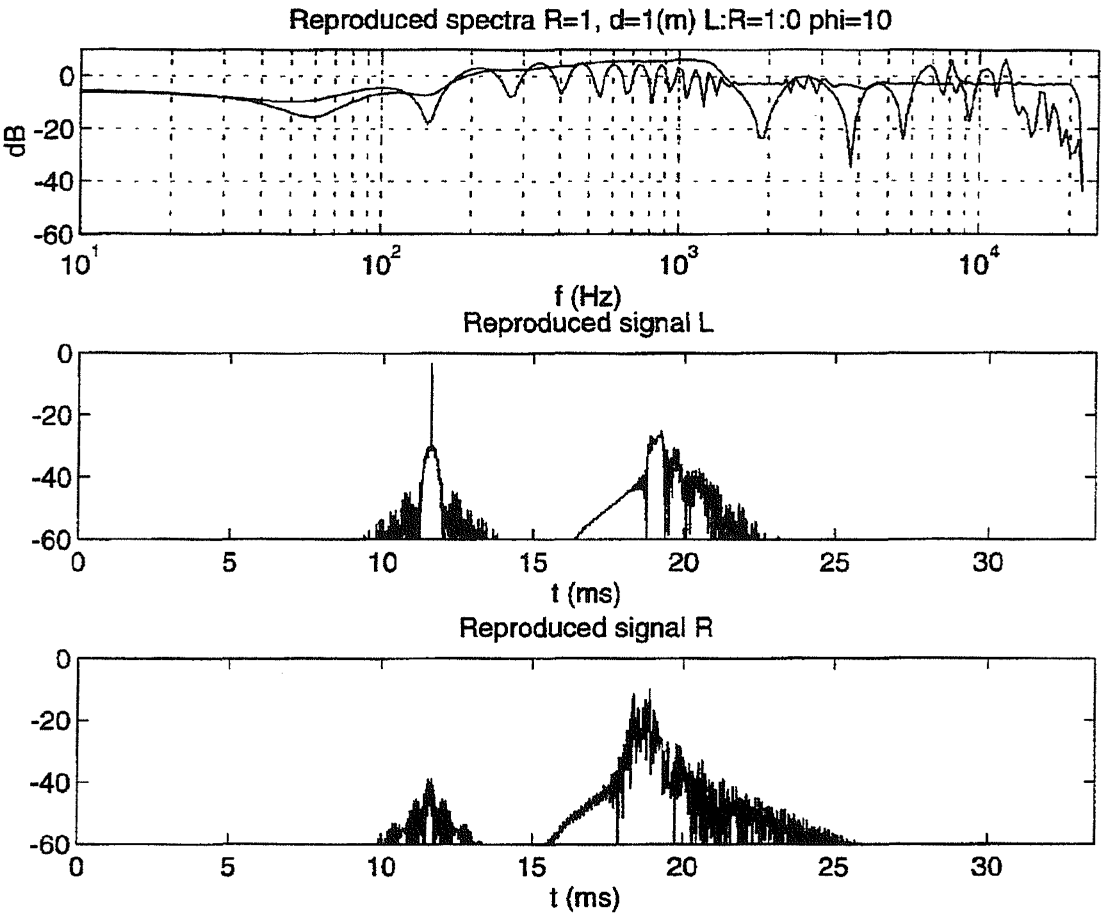


Fig. 6.7 Spectra and the time history of the reproduced signal with reflection (reflecting surface on the contra-lateral side, R = 1) (unit: dB)

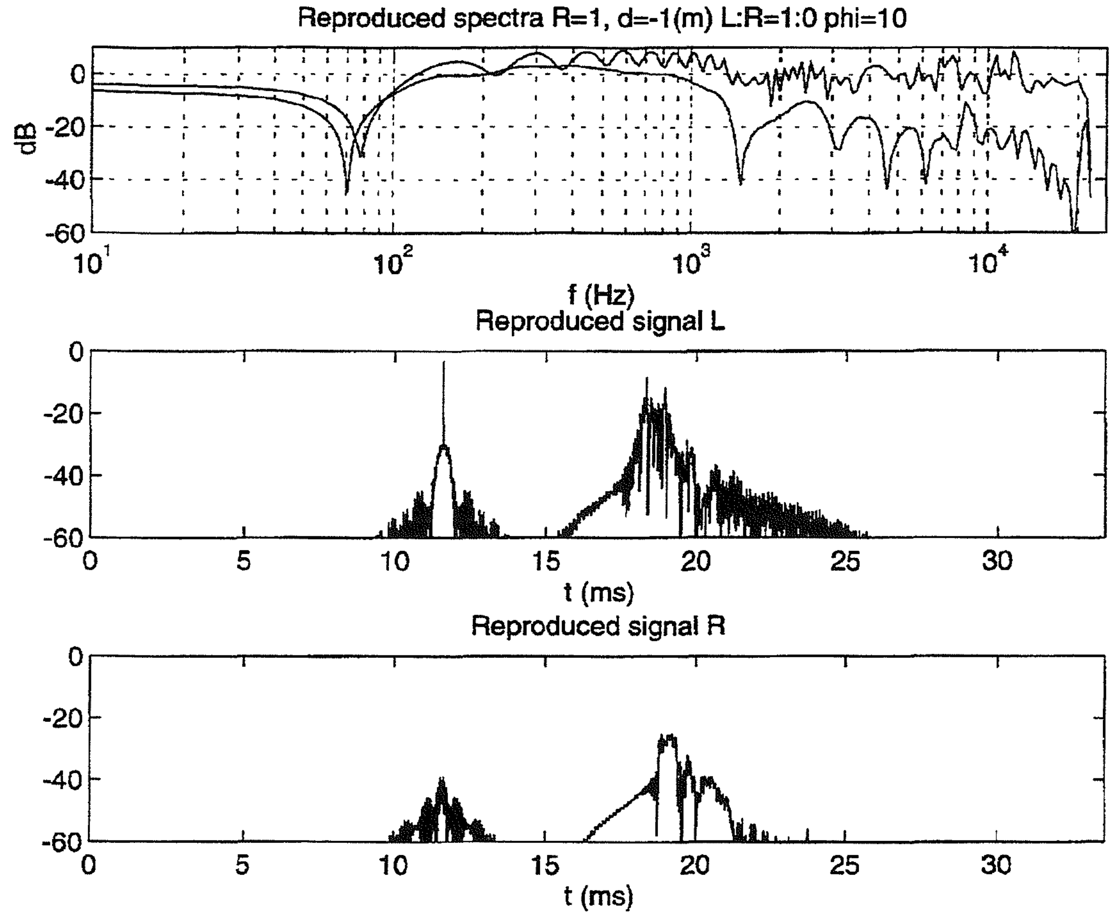


Fig. 6.8 Spectra and the time history of the reproduced signal with reflection (reflecting surface on the ipsi-lateral side, R = 1) (unit: dB)

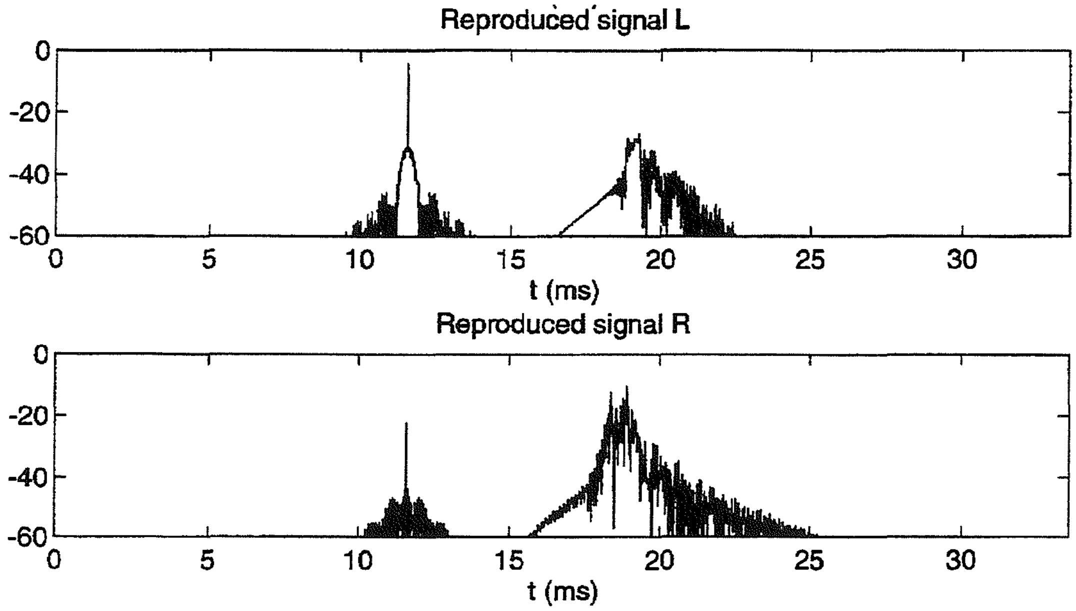


Fig. 6.9 Time history of the reproduced signal with reflection (reflecting surface on the contra-lateral side, level of the target impulse for the right ear is about 20dB smaller than that for the left ear)

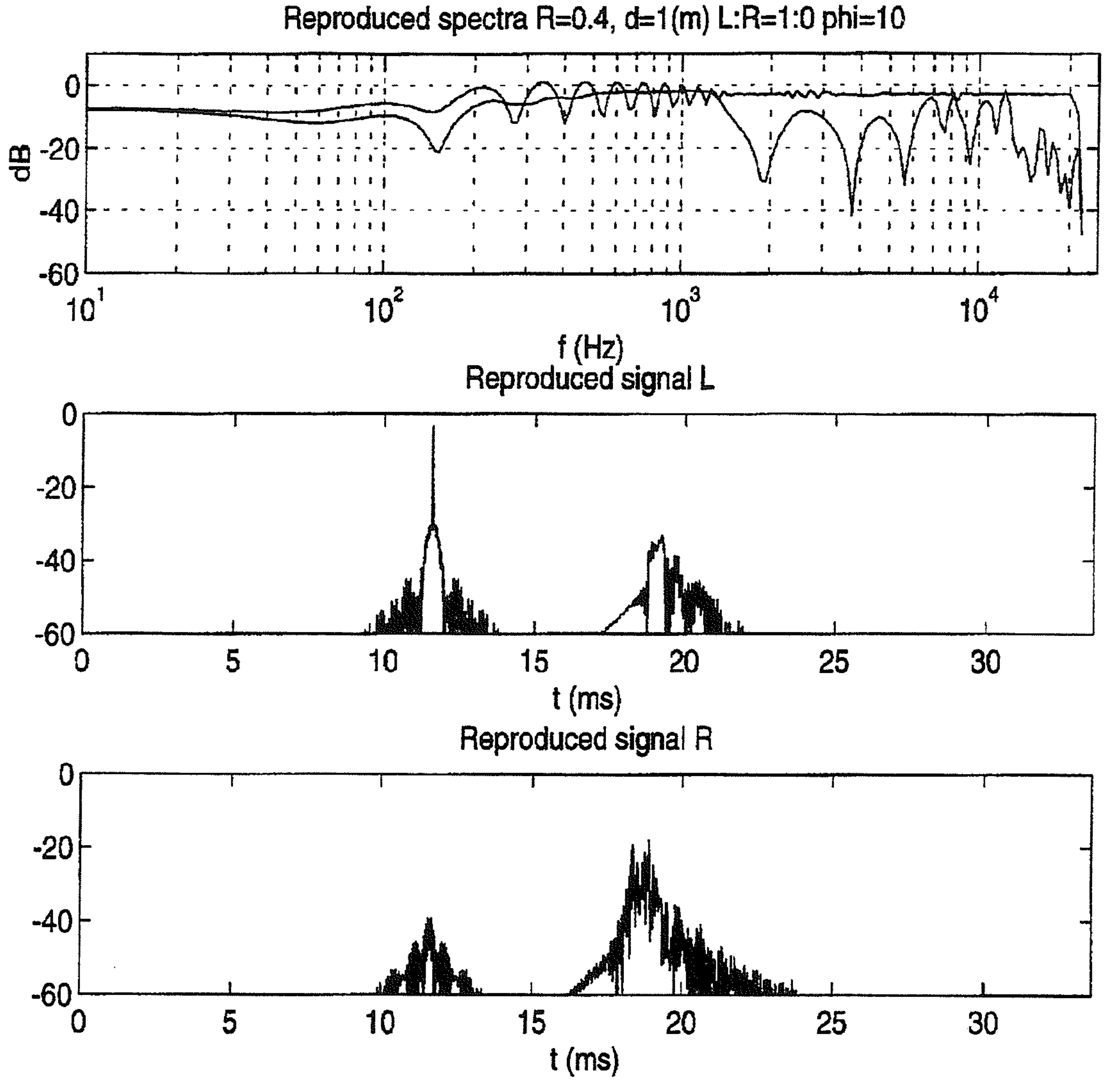


Fig. 6.10 Spectra and time history of the reproduced signal with reflection (reflecting surface on the contra-lateral side, R = 0.4) (unit: dB)

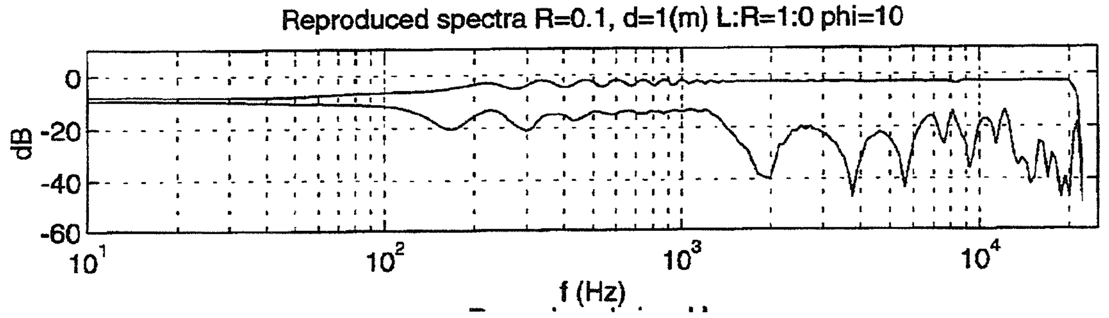


Fig. 6.11 Spectra of the reproduced signal with reflection (reflecting surface on the contra-lateral side, R = 0.1)

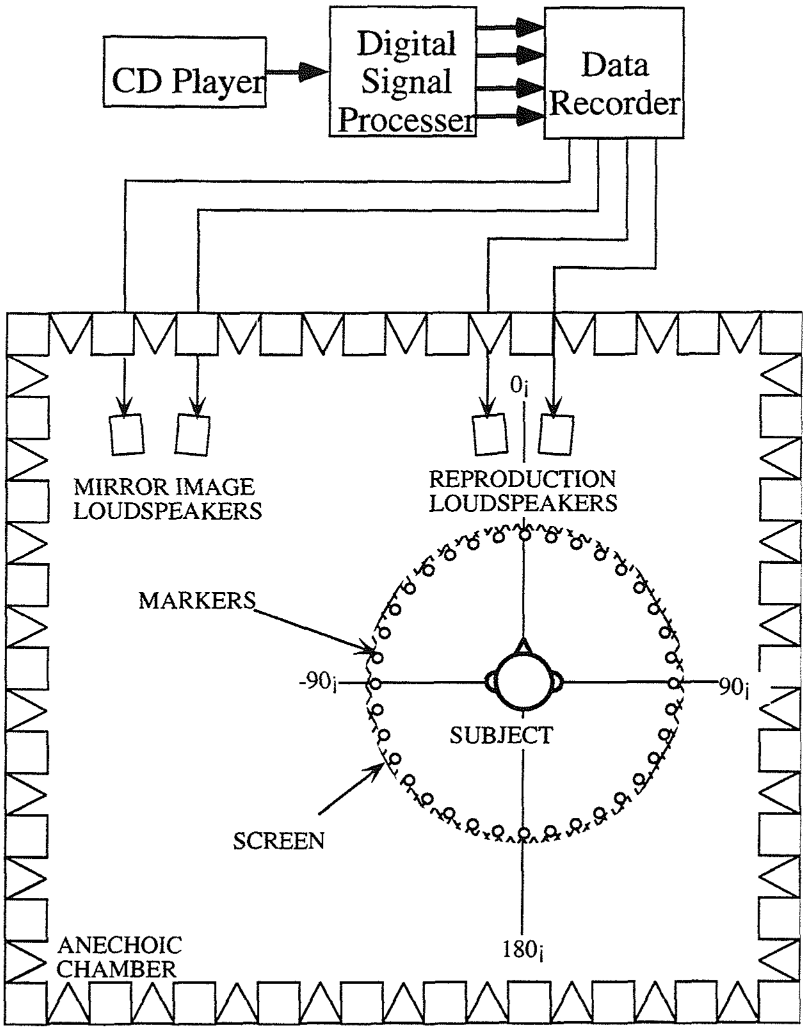


Fig. 6.12 Arrangements of the system for the subjective experiments

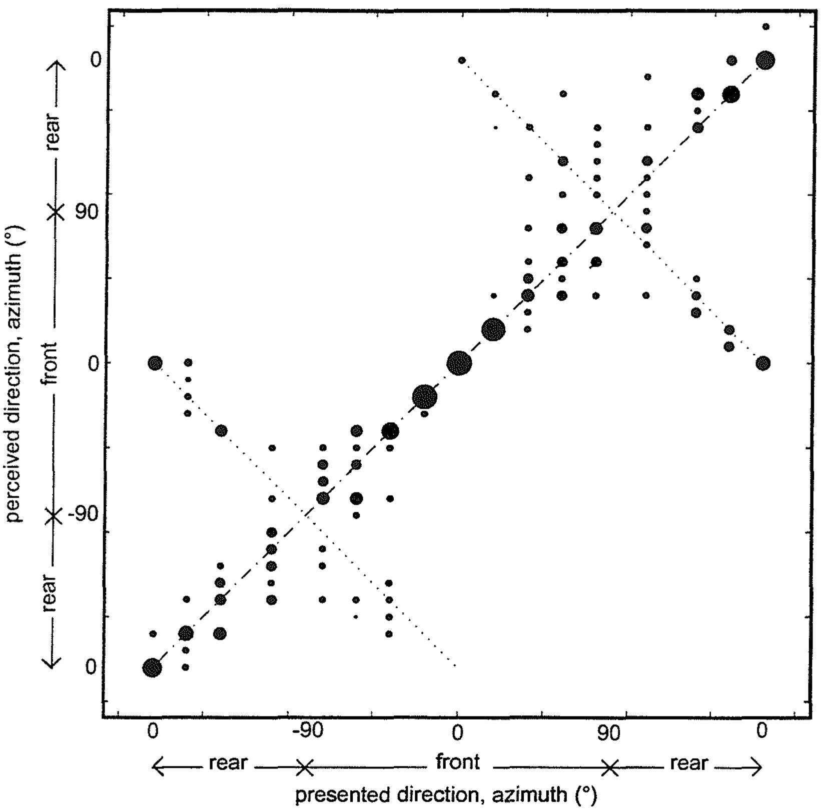


Fig. 6.13a Results of the subjects for whom the virtual acoustic imaging systems work well

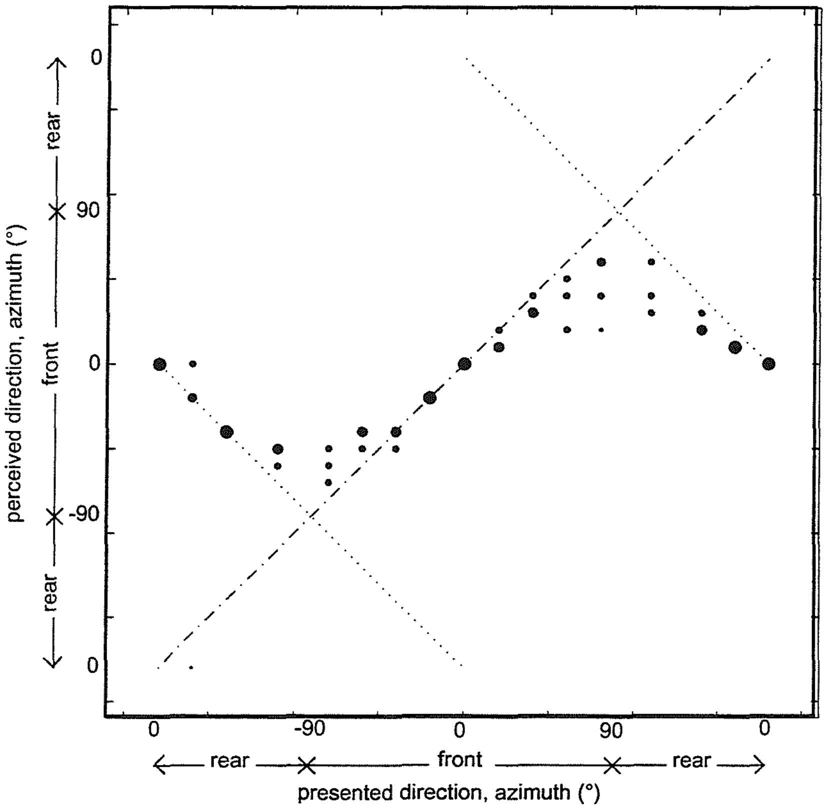


Fig. 6.13b Results of the subjects for whom virtual acoustic imaging systems do not work well.

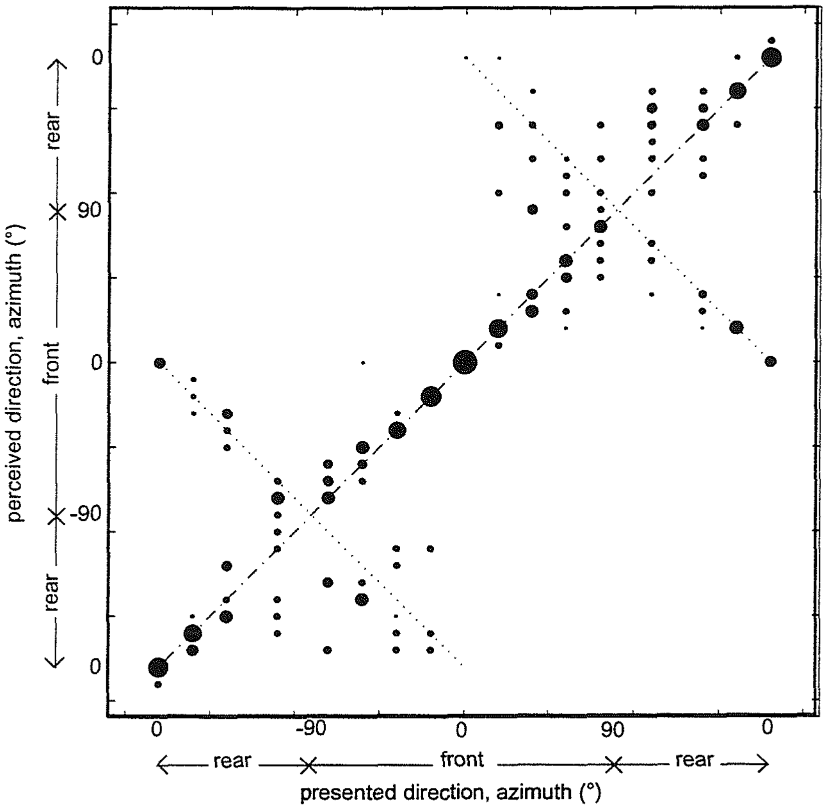


Fig. 6.14a Results when R = 0.1

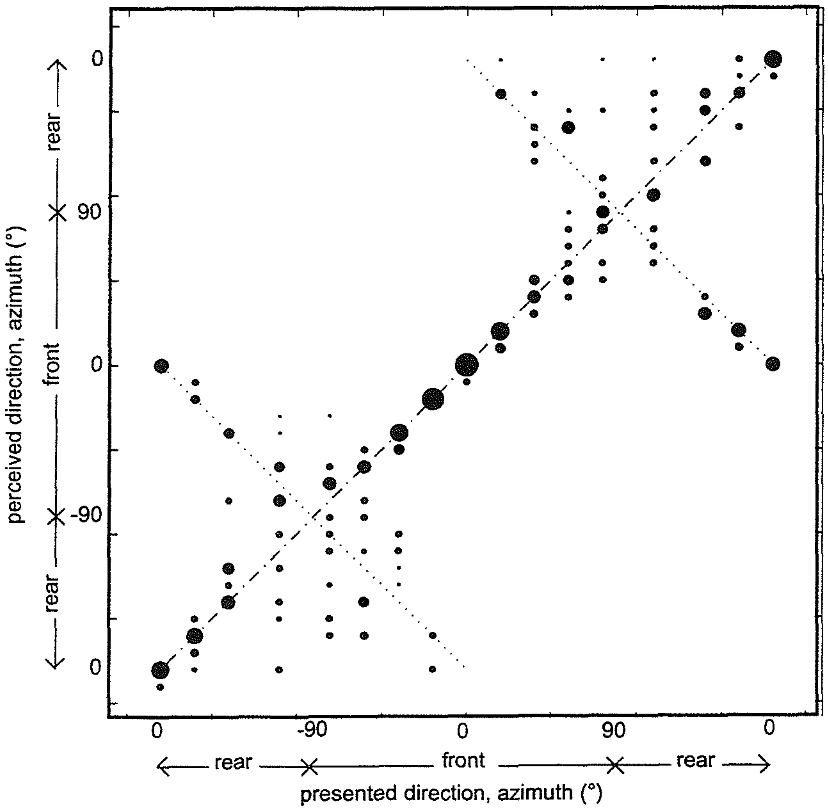


Fig. 6.14b Results when R = 0.3

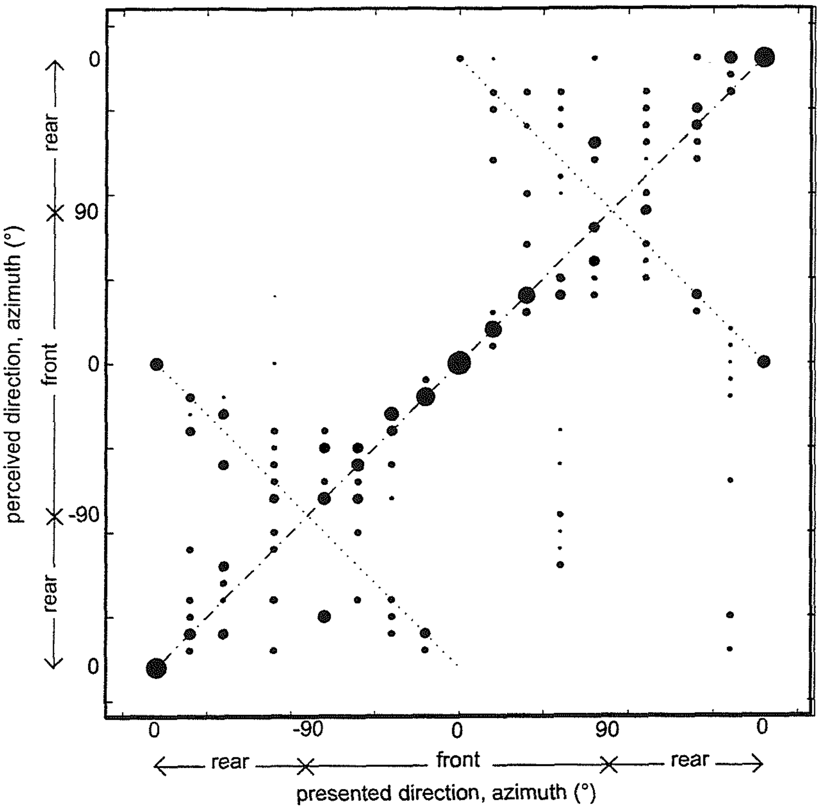


Fig. 6.14c Results when R = 1.0

7 Optimal source distribution

7.1 Introduction

One of the objectives of this chapter is to investigate a number of problems that arise from the multi-channel system inversion involved in binaural synthesis over loudspeakers. A basic analysis with a free field transfer function model illustrates the fundamental difficulties that such systems can have. The singular value decomposition helps to understand the role of the inverse filters more intuitively [A 10]-[A 15]. The amplification required by the system inversion results in loss of dynamic range. The inverse filters obtained are likely to contain large errors around ill-conditioned frequencies. Regularisation is often used to design practical filters but this also results in poor control performance around those frequencies. Sound radiation by transducers in directions other than that of the listener can be very large and this results in severe reflection which can degrade control performance. Further analysis with a more realistic plant matrix, where the sound signals are controlled at a listener's ears in the presence of the listener's body (pinnae, head...), demonstrates that this is still the case. Such problems are often noted as noise, distortion, fatigue of transducers, loss of directional and spatial perception, and colouration.

The investigation has resulted in the proposal of a system concept that we refer to as the Optimal Source Distribution (OSD) [A 10]-[A 15]. The OSD system overcomes these fundamental problems by means of a conceptual pair of monopole transducers whose span varies continuously as a function of frequency. This is where two singular values are balanced and requiring minimum amplification by the inverse filters. The underlying theoretical principle is described in detail. The significance is that all of the above problems that are associated with the multi-channel system inversion are solved by using this principle. The limitations with this principle are also made clear in terms of the operational frequency range. Several examples of practical solutions that can realize a variable transducer span are also described. One of them is discretisation and this enables the use of conventional transducer units and cross-over filter networks with only a little decrease in performance from the theoretical limit. Consequences of the discretisation are also investigated in detail. Practical ways to tackle the sub-low frequency region where the frequency-span relationship is remote from the optimal are also described. All passive, active and digital cross-over filters can be used for the discrete OSD system. The effect of differences in discretisation number of the discrete OSD system are investigated.

Among the number of examples investigated, two of them are found to be most practical and working systems with this principle are realised [A 16]. The practical working OSD system are realised by discretising a continuously variable span into 2 or 3 pairs of transducers. Their performance are investigated by a series of objective and subjective evaluations by comparing the OSD system with its predecessor "Stereo-Dipole" system (SD system). The practical system realized has a very good performance over a wide frequency range (e.g. over the whole audible frequency range). A systematic localisation test is carried out to investigate the effect of this principle on spatial perception. Since better control performance over wider frequency range is expected to give a better spatial perception.

During the course of this research, a research letter [55] was published independently of these researches [A 10]-[A 16]. It greatly enriched the discussion that will be presented in Section 7.3.2. It also suggests a similar system to the 2-way discrete OSD system that was described in reference [A 10]-[A 16] and in Section 7.5.2. The results of other independent research [56] were also published after the completion of this research and these depict a few drawings of how some examples of the discretised OSD systems might look in practical applications.

7.2 Analysis with a free field model and the singular value decomposition

Although the following analysis aims primarily at binaural synthesis over loudspeakers, the same discussion applies to many other cases of audio applications, where pair of desired signals are the signals that would produce a desired virtual auditory sensation when fed to the two ears independently. A simple case involving the control of two monopole receivers with two monopole transducers (sources) under free field conditions is first considered here in order to improve understanding of the physics underlying binaural synthesis over loudspeakers. The fundamental problems with regard to system inversion can be illustrated in this simple case where the effect of path length difference dominates the problem. A matrix of Head Related Transfer Functions (HRTFs) is also analysed in the later section as an example of a more realistic plant. In such a case, the acoustic response of the human body (pinnae, head, torso and so on) also comes to affect the problem. However, the fundamental difficulties inherent to such systems are still clearly evident.

7.2.1 Inverse filter matrix

A symmetric case with the inter-source axis parallel to the inter-receiver axis is considered for an examination of the basic properties of the system. The geometry is illustrated in Fig. 7.1. In the free field case, the plant transfer function matrix can be modelled as

$$\mathbf{C} = \frac{\rho_0}{4\pi}\begin{bmatrix} \mathrm{e}^{-\mathrm{j}kl_1}/l_1 & \mathrm{e}^{-\mathrm{j}kl_2}/l_2 \\ \mathrm{e}^{-\mathrm{j}kl_2}/l_2 & \mathrm{e}^{-\mathrm{j}kl_1}/l_1 \end{bmatrix} \tag{7.1}$$

where an $\mathrm{e}^{\mathrm{j}\omega t}$ time dependence is assumed with $k = \omega/c_0$, and where ρ_0 and c_0 are the density and sound speed. When the ratio of and the difference between the path lengths connecting one source and two receivers are defined as $g = l_1/l_2$ and $\Delta l = l_2 - l_1$,

$$\mathbf{C} = \frac{\rho_0\,\mathrm{e}^{-\mathrm{j}kl_1}}{4\pi l_1}\begin{bmatrix} 1 & g\,\mathrm{e}^{-\mathrm{j}k\Delta l} \\ g\,\mathrm{e}^{-\mathrm{j}k\Delta l} & 1 \end{bmatrix} \tag{7.2}$$

Now consider the case

$$\mathbf{d} = \frac{\rho_0 \mathrm{e}^{-\mathrm{j}kl_1}}{4\pi l_1}\begin{bmatrix} \mathrm{D}_1(\mathrm{j}\omega) \\ \mathrm{D}_2(\mathrm{j}\omega) \end{bmatrix} \tag{7.3}$$

i.e., the desired signals are the acoustic pressure signals which would have been produced by the closer sound source alone whose values are either $\mathrm{D}_1(\mathrm{j}\omega)$ or $\mathrm{D}_2(\mathrm{j}\omega)$ without disturbance due to the other source (cross-talk). This normalization enables a description of the effect of system inversion as well as ensuring a causal solution. The elements of **H** can be obtained from the exact inverse of **C** and can be written as

$$\mathbf{H}=\mathbf{C}^{-1}=\frac{1}{1-g^{2}\mathrm{e}^{-2\mathrm{j}k\Delta l}}\begin{bmatrix}1 & -g\mathrm{e}^{-\mathrm{j}k\Delta l}\\ -g\mathrm{e}^{-\mathrm{j}k\Delta l} & 1\end{bmatrix}$$

(7.4)

When $l >> \Delta r$, we have the approximation $\Delta l \approx \Delta r\sin\theta$ where $\Theta = 2\theta$ is the source span (hence $0 < \Theta \leq \pi$) and under these conditions,

$$\mathbf{H}=\frac{1}{1-g^{2}\mathrm{e}^{-2\mathrm{j}k\Delta r\sin\theta}}\begin{bmatrix}1 & -g\mathrm{e}^{-\mathrm{j}k\Delta r\sin\theta}\\ -g\mathrm{e}^{-\mathrm{j}k\Delta r\sin\theta} & 1\end{bmatrix}$$

(7.5)

7.2.2 Singular value decomposition

The singular value decomposition helps to understand the role of the inverse filter matrix **H** more intuitively. As described in the appendix, the inverse filter matrix **H** can be expressed as

$$\mathbf{H}=\mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{V}^{\mathrm{H}}=\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\end{bmatrix}\begin{bmatrix}\sigma_i & 0\\ 0 & \sigma_o\end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix}\sqrt{\frac{1+g\,\mathrm{e}^{\mathrm{j}k\Delta l}}{1+g\,\mathrm{e}^{-\mathrm{j}k\Delta l}}} & \sqrt{\frac{1+g\,\mathrm{e}^{\mathrm{j}k\Delta l}}{1+g\,\mathrm{e}^{-\mathrm{j}k\Delta l}}}\\ \sqrt{\frac{1-g\,\mathrm{e}^{\mathrm{j}k\Delta l}}{1-g\,\mathrm{e}^{-\mathrm{j}k\Delta l}}} & -\sqrt{\frac{1-g\,\mathrm{e}^{\mathrm{j}k\Delta l}}{1-g\,\mathrm{e}^{-\mathrm{j}k\Delta l}}}\end{bmatrix}$$

where $$\sigma_i=\frac{1}{\sqrt{\left(1+g\mathrm{e}^{-\mathrm{j}k\Delta \mathrm{r}\sin\theta}\right)\left(1+g\mathrm{e}^{\mathrm{j}k\Delta \mathrm{r}\sin\theta}\right)}}$$

and $$\sigma_o=\frac{1}{\sqrt{\left(1-g\mathrm{e}^{-\mathrm{j}k\Delta \mathrm{r}\sin\theta}\right)\left(1-g\mathrm{e}^{\mathrm{j}k\Delta \mathrm{r}\sin\theta}\right)}}$$

(7.6)

The unitary matrix $\mathbf{V}^{\mathrm{H}}$ extracts the in-phase and out-of phase components out of the binaural signals. It also introduces phase rotation according to the property of the plant but does not change their amplitude. The two singular values are denoted by σ_{i} and σ_{o}, and correspond to orthogonal components of the inverse filters. The singular value σ_{i} corresponds to the amplification factor of the in-phase component of the binaural signals and the other singular value σ_{o} corresponds to the amplification factor of the out-of-phase component of the binaural signals. The unitary matrix $\mathbf{U}$ distributes the suitably amplified in-phase and out-of phase components into the pair of transducers.

The net effect of the inverse filter matrix $\mathbf{H}$ depends largely on the content of the input signals $\mathbf{d}$, i.e. the characteristics of the sound source signal contents and the auditory virtual space being created. However, the maximum amplification of the source strengths required for the arbitrary binaural signal input at each frequency can be found from the 2-norm of $\mathbf{H}$ ($||\mathbf{H}||$). Since $||\mathbf{U}|| = ||\mathbf{V}|| = 1$, this is equal to the largest of the singular values. Thus

$$\|\mathbf{H}\| = \|\Sigma\| = \max(\sigma_{\mathrm{i}}, \sigma_{\mathrm{o}}) \tag{7.7}$$

Plots of σ_{i} , σ_{o}, and $||\mathbf{H}||$ with respect to $k\Delta r\sin\theta$ are illustrated in Fig. 7.2. The examples throughout this chapter use a typical value of the distance between the adult human ears for Δr (more detailed discussion can be found in Section 7.4.3). As seen in Eq. (7.6) and Fig. 7.2, the singular values σ_{i} and σ_{o} interchange their amplitude as a function of frequency and source span, periodically giving peaks of $||\mathbf{H}||$ where k and θ satisfy the following relationship with even values of the integer number n.

$$k\Delta r\sin\theta = \frac{n\pi}{2} \tag{7.8}$$

The singular value σ_i has peaks at n = 2, 6, 10, ... where the system is required to use large effort to reproduce the in-phase component of the desired signals. The singular value σ_o has peaks at n = 0, 4, 8, ... where the system is required to use large effort to reproduce the out-of-phase component. The low frequency boost as a consequence of the peak at $n = 0$ has often been addressed in several papers but the other features, especially the peak for the in-phase component, has drawn less attention.

7.3 Fundamental problems of binaural reproduction over loudspeakers

7.3.1 Loss of dynamic range

In practice, since the maximum source output is given by $||\mathbf{H}||_{\max}$, this must be within the range of the system in order to avoid clipping of the signals. The required amplification results directly in the loss of dynamic range illustrated in Fig. 7.3. The level of the output source signals $\mathbf{v}$ and the resulting level of the acoustic pressure at listener's ears $\mathbf{w}$ are plotted both with and without system inversion assuming that the maximum output levels and dynamic range of the systems are the same. Where $||\mathbf{H}||$ is large, the transducers are emitting very large sound output most of which is cancelled to leave small level of synthesised binaural signals at the listener's ears. The given dynamic range is distributed into the system inversion and the remaining dynamic range that is to be used by the binaural auditory space synthesis, and also most importantly, by the sound source signal itself. Thus the signal to noise ratio becomes low. Since the transducers are working much harder than normally to produce usual sound level at the ears, non-linear distortion becomes more significant and is often audible. For the same reason, fatigue of the transducers is more severe. Conventional driver units are not designed to be used in this manner and they can be easily destroyed by fatigue.

The dynamic range loss is defined by the difference between the signal level at the receiver with one monopole source and the signal level reproduced by two sources having the same maximum source strength when the system is inverted. The frequency of the peaks of ||**H**|| do not affect the amount of dynamic range loss but the magnitude of the peaks do. Since ||**H**|| here is normalised by the case without system inversion by Eq. (7.3), the dynamic range loss Γ is given by

$$\Gamma = \|\mathbf{H}\|_{\max} = \frac{1}{1-g} \tag{7.9}$$

The dynamic range loss given by Eq. (7.9) as a function of source span is shown in Fig. 7.4. Since $g \approx 1 - \Delta r \sin\theta / l$, then Γ can be approximated as

$$\Gamma \approx \frac{l}{\Delta r \sin\theta} \tag{7.10}$$

as a function of θ. Fig. 7.4 and Eq. (7.10) show that the larger the source span, the less is the dynamic range loss. It varies from more than 70dB when two transducers are very close together to about 15dB when they are on opposite sides of the ears. When there is a head between the ears, this is relaxed a little.

7.3.2 Robustness to error in the plant

Equation (2.5a, b, c) implies that the system inversion (which determines **v** and leads to the design of the filter matrix **H**) is very sensitive to small errors in the assumed plant **C** (which is often measured and thus small errors are inevitable) where the condition number of **C**, κ(**C**), is large [49]. Such errors include individual differences of HRTFs (Chapter 4) and misalignment of the head and loudspeakers (Chapter 5).

The condition number of **C** is given by

$$\begin{aligned}\kappa(\mathbf{C}) &= \|\mathbf{C}\|\|\mathbf{C}^{-1}\| = \|\mathbf{C}\|\|\mathbf{H}\| = \|\mathbf{H}^{-1}\|\|\mathbf{H}\| \\ &= \max\left(\sqrt{\frac{\left(1+ge^{-jk\Delta r\sin\theta}\right)\left(1+ge^{jk\Delta r\sin\theta}\right)}{\left(1-ge^{-jk\Delta r\sin\theta}\right)\left(1-ge^{jk\Delta r\sin\theta}\right)}}, \sqrt{\frac{\left(1-ge^{-jk\Delta r\sin\theta}\right)\left(1-ge^{jk\Delta r\sin\theta}\right)}{\left(1+ge^{-jk\Delta r\sin\theta}\right)\left(1+ge^{jk\Delta r\sin\theta}\right)}}\right)\end{aligned} \tag{7.11}$$

and is shown in Fig. 7.5. As seen in Eq. (7.11) and Fig. 7.5, κ(**C**) has peaks where Eq. (7.8) is satisfied with an even value of the integer number *n*. The frequencies which give peaks of κ(**C**) are consistent with those which give the peaks of ||**H**||.

Around the frequencies where κ(**C**) is large, the system is very sensitive to small errors in **C** [50] [55]. The calculated inverse filter matrix **H** is likely to contain large errors due to small errors in **C** and results in large errors in the reproduced signal **w** at the receiver. This is because such errors are magnified by the inverse filters but remain uncancelled in the plant. On the contrary, κ(**C**) is small around the frequencies where *n* is an odd integer number in Eq. (7.8). For the same value of *n*, the robust frequency range becomes lower as the source span becomes larger. With a logarithmic frequency scale, which is related to the perceptual attributes of the human auditory system, the frequency range of robust inversion is more or less constant for different source spans for the same value of n, even though it looks wider for smaller source spans on a linear frequency scale.

7.3.3 Robustness to error in the inverse filters

In addition, since

$$\mathbf{v} = \mathbf{C}^{-1}\mathbf{w} \tag{7.12}$$

and $\kappa(\mathbf{C}^{-1}) = \kappa(\mathbf{C})$, a practical and close to ideal inverse filter matrix **H** is easily obtained where $\kappa(\mathbf{C})$ is small. However, the reproduced signals **w** are less robust to small changes in the inverse of the plant matrix $\mathbf{C}^{-1}$, hence **H**, where $\kappa(\mathbf{C})$ is large. Even if **C** does not contain any errors, the reproduction of the signals at the receiver is too sensitive to the small errors within the inverse filter matrix **H** to be useful.

One common example of such an error is that due to regularisation, where a small error is deliberately introduced to improve the condition of matrix to design practical filters. It is also possible to reduce the excess amplification and hence the dynamic range loss by means of regularisation, where the pseudo inverse filter matrix **H** is given by Eq. (2.17) where β is a regularisation parameter. The regularisation parameter penalises large values of **H** and hence limits the dynamic range loss of the system. Since $\|\mathbf{H}\|$ is normalised by the case without system inversion by Eq. (7.3), the regularisation parameter limits the dynamic range loss to less than about

$$\Gamma \approx -10\log_{10}\beta - 6 \quad \text{(dB)} \tag{7.13}$$

However, the regularisation parameter intentionally, hence inevitably, introduces a small error in the inversion process. This gives rise to a problem for filter design at frequencies where κ(**C**) is large. An example of this is illustrated in Fig. 7.6. The dynamic range loss is reduced by regularisation from about 27dB (without regularisation) as in Fig. 7.6a to 14dB as shown in Fig. 7.6b ($\beta = 10^{-2}$). However, it can be clearly seen that the control performance of the system deteriorates around the frequencies where n is an even integer number in Eq. (7.8). The contribution of the correct desired signals (R_{11} and R_{22}) is reduced only slightly but the contribution of the wrong desired signals (R_{12} and R_{21}, the cross-talk component) is increased significantly. In other words, the system has little control (cross-talk cancellation) around these frequencies. This problem is significant at lower frequencies (n<1 in Eq. (7.8)) in the sense that the region without cross-talk suppression is large, and at higher frequencies (n>1 in Eq. (7.8)), in the sense that there are many frequencies at which the plant is ill-conditioned. With an equivalent dynamic range loss, making the source span larger leads to a better control performance at lower frequencies but a poorer performance at higher frequencies (Fig. 7.7a). On the contrary, making the source span smaller leads to better control performance at higher frequencies but poorer performance at lower frequencies (Fig. 7.7b).

7.3.4 Robustness to reflections

The amplification by the inverse filter also results in severe reflection ([51], Chapter 6). Fig. 7.8 shows an example ($n \approx 2$) of far field sound radiation by the control transducers with reference to the receiver directions. The horizontal axis is the inter-source axis and the receivers (ears) are at the directions of the vertical axis. At frequencies where Eq. (7.8) is not satisfied with an odd value of the integer number n, as in this example, the sound radiation in directions other than receiver directions can be significantly larger (typically +30dB ~ 40dB) than those at the receiver directions (0dB and -∞dB). The maximum amount of this excessive radiation is the same as the amount of dynamic range loss as in Eq. (7.9) and Fig. 7.4. When the environment is not anechoic, as is normally the case, this obviously results in severe reflections and the control performance of the system deteriorates. In addition, the sound radiated in directions other than that of the receiver has a peaky frequency response due to the response of inverse filter matrix **H** and normally result in severe colouration.

7.4 A system to overcome the problems

As discussed above, there is a trade-off between dynamic range, robustness and control performance. However, a system that aims to overcome these fundamental problems is proposed in what follows.

7.4.1 Principle of the Optimal Source Distribution

Equation (7.8) can be rewritten in terms of the source span Θ as

$$\Theta = 2\theta = 2\arcsin\left(\frac{n\pi}{2k\Delta r}\right) \tag{7.14}$$

As seen from the analysis above, systems with the source span where n is an odd integer number in Eq. (7.14) give the best control performance as well as robustness. This implies that the optimal source span must vary as a function of frequency.

We now consider a pair of conceptual monopole transducers whose span varies continuously as a function of frequency in order to satisfy the requirement for n to be an odd integer number in Eq. (7.14). This is where σ_i and σ_o are balanced and this is illustrated in Fig. 7.9 and Fig. 7.10. The source span becomes smaller as frequency becomes higher. With this concept, Eq. (7.5) becomes very simple as

$$\mathbf{H} = \frac{1}{1+g^2}\begin{bmatrix} 1 & -jg \\ -jg & 1 \end{bmatrix} \tag{7.15}$$

Note that $\|\mathbf{H}\| = 1/\sqrt{2}$ for all frequencies. Therefore, there is no dynamic range loss compared to the case without system inversion. In fact, there is a dynamic range gain of 3dB since the two orthogonal components of the desired signals are $\pi/2$ out of phase. This means that the system has good signal to noise ratio and is advantageous with respect to distortion or fatigue of transducers. The inverse filters have a flat frequency response so there is no colouration at any location in the listening room, even outside the sweet area. When the listener is far away from the sweet spot, the spatial information perceived may not be ideal. However, the spectrum of the sound signals are not changed by the inverse filters. Therefore, the listener can still enjoy the natural production of sound together with some remaining spatial aspects. The sound radiation by the transducer pair in all directions is always smaller than those at the receiver directions, which is also smaller than the sound radiation by a single monopole transducer producing the same sound level at the ears. An example when $n = 1$ is shown in Fig. 7.11. Therefore, the system is also robust to reflections in a reverberant environment, and these small reflections do not have any coloration other than those caused by the reflecting materials. In practice, the directivity of loudspeakers helps to reduce the effect of reflection further. Note also that $\kappa(\mathbf{C}) = 1$ which is the smallest value possible for all frequencies. The error in calculating the inverse filter is small and the system has very good control over the reproduced signals. The system is also very robust to the changes in plant matrix.

Also note that when $l \gg \Delta r$, $g \approx 1$ therefore,

$$\mathbf{H} \approx \frac{1}{2}\begin{bmatrix} 1 & -\mathrm{j} \\ -\mathrm{j} & 1 \end{bmatrix} \tag{7.16}$$

This implies that independent control of the two signals is nearly achieved just by addition of the desired signals with a $\pi/2$ relative phase shift between them.

7.4.2 Aspects of the proposed system

From Eq. (7.14), the range of variable source span Θ is given by the frequency range of interest as can be seen from Fig. 7.10. A smaller value of n gives a smaller source span for the same frequency. Therefore, the smallest source span Θ_h for the same high frequency limit is given by $n = 1$ and this is about 4° to give control of the sound field at two positions separated by the distance between two ears (about 0.13m for KEMAR dummy head) up to a frequency of 20kHz.

Equation (7.8) can also be rewritten in terms of frequency as

$$f = \frac{nc_0}{4\Delta r \sin\theta} \tag{7.17}$$

The smallest value of n gives the lowest frequency limit for a given source span. Since $\sin\theta \leq 1$,

$$f \geq \frac{nc_0}{4\Delta r} \tag{7.18}$$

i.e., the physically maximum source span of $\Theta = 2\theta = 180°$ gives the lowest frequency limit, f_l, associated with this principle. A smaller value of n gives a lower low frequency limit so the system given by $n = 1$ is normally the most useful among those with an odd integer number n. The low frequency limit given by $n = 1$ of a system designed to control the sound field at two positions separated by the distance between two ears is about f_l = 300 ~ 400 Hz.

7.4.3 Consideration of the head related transfer function model.

The condition number $\kappa(\mathbf{C})$ of the plant matrix plotted as a function of frequency and source span is shown in Fig. 7.12 for the audible frequency range (20Hz ~ 20kHz). Fig. 7.13 shows the condition number of the more realistic plant matrix with HRTFs. The HRTFs were measured with the KEMAR dummy head at MIT Media Lab [38] and the loudspeaker response was deconvolved later. Those between sampled directions are obtained by bilinear interpolation on the virtual spherical surface of magnitude and phase spectra in the frequency domain (Appendix 1). A similar trend can clearly be seen as in the free field case. However, additional "ill-conditioned frequencies" can be observed around 9kHz and 13kHz where the HRTFs have minima. It is possible that the signal to noise ratio of the measured data around these frequencies is poor.

It should also be noted that where the incidence angle θ is small, the peak frequencies obtained with the HRTF plant matrix are similar to those of the free field plant with the receiver distance $\Delta r \approx 0.13$. This corresponds to the shortest distance between the entrances of the ear canals of the KEMAR dummy head. However, where the incidence angle θ is large, the peak frequencies obtained with the HRTF plant matrix are similar to that of the free field plant with the receiver distance $\Delta r \approx 0.25$. This is a much larger distance than the shortest distance between the entrances of the ear canals of the KEMAR dummy head and is a result of diffraction around the head. A correction to the receiver distance Δr can be made in order to match the frequency-span characteristics of the free field model. The following is an example of a linear approximation which seems to be fairly accurate. Thus

$$\Delta r = \Delta r_0 (1 + \Theta/\pi) \tag{7.19}$$

where Δr_0 is the geometrical distance between the ears.

7.4.4 Transducers for the Optimal Source Distribution

This principle requires a pair of monopole type transducers whose position from which sound is radiated varies continuously as frequency varies. This might, for example, be realized by exciting a plate at each position individually (Fig. 7.14a). The requirement of such a transducer is that a certain frequency of vibration is excited most at a particular position such that sound of that frequency is radiated mostly from that position. Such characteristics may be achieved by exciting a triangular shaped plate at one end whose width and stiffness varies along its length in a controlled manner (Fig. 7.14b). The narrow and stiff excited end radiates most high frequency sound whereas the wide and "floppy" end of the plate radiates the lower frequency sound. Alternatively, a similar effect might be obtained by changing the width of a slot along an acoustic waveguide (Fig. 7.15). In both cases, the vibration characteristics of the plate or air particles would differ along the length, and so as the radiation impedance. Then, transducers that effectively distribute each of the frequency components to a desired position may be designed. A relatively large damping would be necessary in order to suppress peaks at resonance.

7.4.5 A discrete system

In practice, a monopole transducer whose position varies continuously as a function of frequency is not easily available. However, it is possible to realise a practical system based on this principle by discretising the transducer span as illustrated in Fig. 7.16. With a given span, the frequency region where the amplification is relatively small and plant matrix **C** is well conditioned is relatively wide around the optimal frequency. In other words, the valleys in Fig. 7.12 and Fig. 7.13 are U-shaped. Therefore, by allowing n to have some width, say $\pm\nu$ $(0 < \nu < 1)$, which results in a small amount of dynamic range loss and slightly reduced robustness, a certain transducer span can nevertheless be allocated to cover a certain range of frequencies where control performance and robustness of the system is still reasonably good (Fig. 7.17). Consequently, it is possible to discretise the continuously varying transducer span into a finite number of discrete transducer spans. A system with a smaller value of n gives a wider region with the same performance on a logarithmic scale as can be seen in Fig. 7.12 and Fig. 7.13.

It is important to design the system to ensure that $\|\mathbf{H}\|$ and $\kappa(\mathbf{C})$ are as small as possible over a frequency range that is as wide as possible. Therefore, the transducer spans for each pair of transducers in each frequency range can be decided to ensure that the smallest possible values of n are used over the whole frequency range of interest above f_1.

It is possible to discretise, i.e., decide the transducer spans and frequency ranges to be covered by each pair of driver units (i.e. range of n), in terms of a tolerable dynamic range loss. Fig. 7.18 shows the frequency/span region in terms of dynamic range loss. The required dynamic range loss of the entire system is now given by the maximum value among those values given by each discretised transducer span. Once a tolerable dynamic range loss is decided, the frequency/span region to be used for a discrete OSD system can be found from Fig. 7.18. The maximum amount of excess sound radiation in directions other than receiver directions is also given by the same figure.

It is also possible to design the system in terms of the control performance (cross-talk cancellation performance) as defined in Section 5.4.1. As an example, Fig. 7.19 illustrates the cross-talk cancellation performance as a function of frequency and source span when there exists 5% of error in the plant or in the inverse filters. This is approximately the same value as the error that is induced by the regularisation to limit dynamic range loss to be less than 20dB (Eq. (7.13)). The frequency/span region to be used can be decided from the required cross-talk cancellation performance.

7.4.6 Consequence of the discretisation of variable source span

The discretisation is extremely useful and practical because a single transducer which can cover the whole audible frequency range is not practically available either. Therefore, this principle also gives the ideal background for multi-way systems for binaural reproduction over loudspeakers which maximise the frequency range to be produced and controlled. Conventional driver units and cross-over filters can easily be accommodated to be used for this system. It should be noted that this is still a simple "2 channel" control system where only two independent control signals are necessary to control any form of virtual auditory space. This in principle can synthesise an infinite number of virtual source locations with different source signals with any type of acoustic response of the space. The difference for this discrete system from the conventional 2-channel system is that the two control signals are divided into multiple frequency bands and fed into the different pairs of driver units with different spans. Ironically, substantial effort has been invested in conventional multi-way loudspeakers for stereophony in order to approximate a point source by multiple driver units. The discrete OSD system requires just the opposite; different driver units are required to be at different locations. A "poor" performance unit in the sense of stereophony which has relatively narrow operational frequency range may perform very well with this principle.

It should be noted that the low frequency limit f_l given by odd integer numbers n in Eq. (7.18) is extended towards a lower frequency by discretisation because now the region for frequency and transducer span where n is not an integer number is also used. For example, a practical system discretised from the ideal system with $n = 1$ can now make use of the region $1-\nu < n < 1+\nu$ so that the low frequency limit is given by $n = 1-\nu$.

As can be seen from Fig. 7.10 and Fig. 7.17, in the higher frequency range where the source span is very small, the frequency range to be covered is very sensitive to small differences in transducer span. On the contrary, it is very insensitive to the source span at lower frequencies. Consequently, the range of practical span for the low frequency units is very large, which can practically be anywhere from about 60° to 180° with only a very slight increase of low frequency limit.

7.4.7 Considerations for the sub-low frequency region

At the frequencies below f_l ($n < 1\text{-}\nu$) where $\|\mathbf{H}\|$ and $\kappa(\mathbf{C})$ is larger than other frequencies, the requirement for dynamic range loss and robustness of the system are more severe than at other frequencies. Fig. 7.20 illustrates the 2-norm of $\mathbf{H}$ and the two singular values (σ_i and σ_o) with the "OSD" principle. As described in section 7.4.1, $\|\mathbf{H}\|$ shows the flat amplitude response of the inverse filters above f_l. However, below f_l, it still increases moderately as frequency becomes lower. In this region, although the system has difficulty in reproducing the out-of-phase component of the desired signal, it still can produce the in-phase component as well as before.

When f_l is reasonably low, where interaural difference may not be crucial for binaural reproduction, one can avoid system inversion and simply add a single sub-woofer unit for this frequency region to avoid the extra dynamic range loss required by this region. As seen in Eq. (7.6), adding two channels of signals results in complete cancellation of the out-of-phase component of the binaural signals and producing the in-phase component only. Then, there is no independent control of binaural signals in this region.

It is possible to cover this sub-low frequency region with the lowest frequency pair of units without sacrificing performance for other frequencies. A large value of regularization parameter can disable the large σ_o, the out-of-phase component, in this region. Even though little cross-talk suppression is available, the low frequency pair can still work as a sub-woofer mostly producing the in-phase component, while it is working perfectly within the OSD frequency range. In the sub-low region, the control performance deteriorates severely due to heavy regularization. However, $||\mathbf{X}||$ and hence the norm of the reproduced signal, is the same as that without regularization. This may be acceptable in binaural reproduction since the difference between the two desired signals is normally not so large and sometimes negligible in the very low frequency range.

When slight dynamic range loss is acceptable, the regularization can be used to limit the amplification, and hence avoid too much dynamic range loss, without sacrificing robustness for other frequencies. The cross-talk performance with regularization in the frequency range below f_l is not as good as at the other frequencies. However, there can still be reasonable cross-talk suppression available. If more dynamic range loss is allowed, a smaller regularization parameter can be used to suppress the out-of-phase component in the sub-low region. The cross-talk cancellation performance in this region is very sensitive to the allocated dynamic range loss. Therefore, it is possible to design the system by selecting the required low frequency cross-talk cancellation performance. The amount of the dynamic range loss required by the discretisation often gives relatively good control performance also in the sub-low frequency region, especially when the discretisation is coarse.

One might choose to allow all the dynamic range loss necessary for the full control of the sub-low frequency region. The overall dynamic range loss is determined by the lowest frequency pair, which has the largest span. As discussed in section 7.3.1, the dynamic range loss by the largest span is the smallest value among all other pairs.

7.5 Examples of a discrete “OSD” system

The design of a practical OSD system is relatively flexible and the system can be adopted in accordance with any particular application.

7.5.1 “3-way” systems and more

An example of 3-way systems with $0 < n < 2$ is illustrated in Fig. 7.21 ~ Fig. 7.23. This example aims to ensure a condition number that is as small as possible over a frequency range that is as wide as possible. Therefore, the transducer spans (Θ) for the high frequency units and the low frequency units were chosen at two extreme positions which gives $\nu = 0.7$. A pair of high frequency units spanning 6.2° is chosen to cover the frequency range up to 20kHz ($n = 1.7$) while a pair of low frequency units spanning 180° is chosen to cover as low a frequency as possible. The span for the mid frequency units is 32°. The driver units may, for example, be housed into 3 cabinets (with the mid and high frequency units in one cabinet). The dynamic range loss of about 7dB can be achieved with 3 pairs of units. This arrangement gives $f_l \approx$ 110Hz ($n = 0.3$ with low frequency pair) and a sub-woofer may be added to deal with the range below this frequency. The cross-over frequencies are given by $n = 0.3$ and $n = 1.7$ ($\nu = 0.7$) for each pair of units and at around 600Hz and 4kHz. The far field sound pressure level produced by the transducer pairs becomes a maximum at the cross-over frequencies and is shown in Fig. 7.24. Note that those around the middle of the frequency range for each transducer pair are as shown in Fig. 7.11 .

By limiting the amplification of the low frequency pair for frequencies below f_l to 7 dB with regularisation, the low frequency units can also cover frequencies down to about 100Hz with reasonable cross-talk cancellation of more than 20dB and cover below 100Hz with reduced interaural difference (Fig. 7.25, Fig. 7.26).

When more dynamic range loss is allowed, it is possible to use smaller regularisation parameters hence low frequency cross-talk performance improves (Fig. 7.27). By allowing dynamic range loss of 13dB, the low frequency units spanning 180° can cover frequencies down to 20 Hz with more than 20dB cross-talk suppression.

Alternatively, it is possible to use a smaller ν, i.e., transducer spans to improve the robustness of the system in the higher frequency range at the expense of the low frequency cross-talk performance, there being plenty to spare in the previous example. An example of this strategy is described in the following section for "2-way" systems.

As the variable transducer span is discretised more finely, e.g., by using 4-way or 5-way systems and so on, the smaller the width of n ($\pm\nu$) becomes. Hence, the system becomes more robust at frequencies above f_l. However, the performance gain becomes smaller and smaller as the number of driver units is increased. Obviously, the finer the discretisation, the closer the design is to the principle of the continuously variable transducer span. However, the number of driver pairs increases and hence the trade-off between performance gain and cost becomes more significant.

7.5.2 "2-way" systems

An example of a 2-way system with $0 < n < 2$ is illustrated in Fig. 7.28 and Fig. 7.30. This example is again designed to ensure small condition numbers over a wide frequency range so the transducer spans were chosen at 6.9° and 120° which gives $\nu \approx 0.9$. A dynamic range loss of about 18dB can be achieved with only 2 pairs of units without regularisation. A pair of mid-high frequency units spanning 6.9° is used to cover the frequency range up to 20kHz ($n = 1.9$) while a pair of mid-low frequency units spanning

120° gives a value of f_l of about 20Hz ($n = 0.1$ with low frequency pair). The cross-over frequency is given by $n = 0.1$ and $n = 1.9$ for each pair of units and is at around 900Hz. The maximum far field sound pressure level produced by the transducer pairs at the cross-over frequencies are shown in Fig. 7.31. Note again the sound pressures produced around the middle of the frequency range for each transducer pair are as shown in Fig. 7.11.

As discretisation becomes coarser, the more frequency regions become severely ill-conditioned. It is possible to reduce transducer spans to improve robustness at higher frequencies at the expense of the low frequency cross-talk performance. Fig. 7.32 ~ Fig. 7.34 shows another example of a 2-way system which is obtained by omitting the pair of woofer units from the 3-way system ($\nu \approx 0.7$) described in the previous section. The dynamic range in this example is maintained to be the same as that in the previous example of the 2-way system (as in Fig. 7.28 ~ Fig. 7.30) by means of regularisation. The span for the high frequency units is 6.2°. The span for mid-low frequency units is 32° which also covers the frequency range below $f_l \approx$ 600Hz with a cross-talk cancellation performance of more than 20dB. The mid-low frequency pair can also cover the range below 200Hz where the cross-talk cancellation performance becomes less than 20dB. All the driver units may be housed in just one cabinet. The cross-over frequency is now at around 4kHz. The conditioning above $f_l \approx$ 600Hz is as good as the 3-way system and it can be seen that the condition number becomes very small compared to the previous example illustrated in Fig. 7.30.

There is a difference in the transducer spanning to that described in reference [55]. This is probably due to the different approaches taken between the two research projects. The spanning of 6.9° and 120° are drawn as a result of discretisation from the continuous variable span [A 10], while in reference [55] it was suggested that a 40° span low frequency pair was added to the 10° span of the "Stereo-Dipole" pair.

7.5.3 "1-way" systems

The coarsest discretisation is given by an example of a 1-way virtual acoustic imaging system with $0 < n < 2$ as illustrated in Fig. 7.35 ~ Fig. 7.37. The transducer span is 7.2°. The benefit available is very limited for a 1-way system with this principle. Since the frequency range to be covered with a single pair of transducers is the whole audible frequency range (20Hz ~ 20kHz), the width of n is nearly ±1 ($\nu = 0.998$). The dynamic range loss is more than 40dB and very large condition numbers are notable in the wide range of low frequencies and at the high frequency end. When regularisation is used to limit the dynamic range loss to 18dB, the cross-talk cancellation performance below 1kHz is less than 20dB (Fig. 7.38).

This is not practical anyway since a practical single transducer which can be used over this frequency range is not available. It is possible to come to a compromise design to reduce the width of n ($\pm\nu$) by sacrificing the high and low frequency ranges which a practical full-range unit can not cover. The "Stereo Dipole" system which has a pair of transducers spanning 10° is one such system, limiting the independent control of binaural signals between about 1kHz and 10kHz in effect.

7.5.4 Comments on multi-region systems

It is also possible to compromise further to utilise two or more regions of n. Then there is no distinction from conventional systems. However, it is still possible to optimise their performance by utilising a similar discussion to that presented above but extending it into multiple regions of n. This approach is beneficial when one attempts to cover a wider frequency range with a smaller number of transducer pairs. The simplest example with a single pair of transducers utilising the regions of $0 < n < 2$ and $2 < n < 4$ is illustrated in Fig. 7.39 ~ Fig. 7.41. The frequency range of 20Hz ~ 20kHz is covered with a single pair of transducers spanning 14°. The required amplification is about 40dB so the example illustrated is regularised to 18dB dynamic range loss. It can be seen in Fig. 7.41 that the cross-talk cancellation performance in the low frequency range is improved from the 1-way system in Fig. 7.38. This example shows more than 20dB cross-talk cancellation performance down to about 400Hz (which was 1kHz in Fig. 7.38). However, there is an unusable region around 10kHz ($1+\nu < n < 3-\nu$) where the system has little control and is not robust.

It may be an idea to match this unusable region to the frequencies where HRTFs have minima ($\|\mathbf{C}\|$ is small) since inversion of minima requires further amplification in **H** and dynamic range loss. In addition, the position of minima in the higher frequency range can vary considerably between individuals. Therefore, it may not be practical to provide inversion around these frequencies where the HRTFs used for filter design have minima.

7.6 Considerations for cross-over filters and inverse filters

Cross-over filters (low pass, high pass or band pass filters) are used to distribute signals of the appropriate frequency range to the appropriate pair of driver units of the discrete (multi-way) "OSD" system. Since an ideal filter which gives a rectangular window in the frequency domain can not be realised practically, there are frequency regions around the cross-over frequency where multiple pairs of driver units are contributing significantly to the synthesis of the reproduced signals **w**. Therefore, it is important to ensure this "cross-over region" is also within the region of this principle.

7.6.1 “2 by 2” plant matrix

If the plant matrix **C** is obtained when including a cross-over network as illustrated in Fig. 7.42, it consists of a single 2 by 2 matrix of electro-acoustic transfer functions between two outputs of the filter matrix **H** and two receivers that contain the responses of the cross-over networks and the interaction between different pairs of driver units around the cross-over frequency. The plant matrix **C** for inverse filter design can also contain the transducer responses and the acoustic response of the human body and the surrounding environment. The obtained 2 by 2 inverse filter matrix **H** designed from this plant matrix **C** automatically compensates for all those responses contained in order to synthesise the correct desired signals at the listener’s ears.

7.6.2 Multiple “2 by 2” plant matrix

Alternatively, one can design inverse filter matrices $\mathbf{H}_1$, $\mathbf{H}_2$, ... for plants $\mathbf{C}_1$, $\mathbf{C}_2$, ... of each pair of driver units (Fig. 7.43). The cross-over filters for each pair of driver units ensure that the signals contain the corresponding frequency range of the signals for the particular pair of units. In this case, around cross-over frequencies, a virtual acoustic environment is synthesised with two different inverse filter matrices. Since both reproduced signals at the ears synthesised with both pairs of driver units are correct, the correct desired signals are reproduced at the ears as a simple sum of those two (identical but different in level) desired signals, provided that the cross-over filters behave well. Since the system inversion is now independent of the cross-over filters, the cross-over filters can also be applied to signals prior to the input to the inverse filters which can be after(Fig. 7.43b) or even before the binaural synthesis.

7.6.3 "2 by (2 × multiple)" plant matrix

It is also possible to obtain the plant matrix **C** as a 2 by (2 × *m*) matrix where *m* is a number of driver pairs (Fig. 7.44). The system is underdetermined and a (2 × *m*) by 2 matrix of the pseudo inverse filter matrix **H** is given by

$$\mathbf{H} = \mathbf{C}^{H}\left[\mathbf{C}\mathbf{C}^{H} + \beta\mathbf{I}\right]^{-1} \qquad (7.20)$$

where β is a regularisation parameter. This solution ensures that the "least effort" (smallest output) of the transducers is used in providing the desired signals at the listener's ears. The net result is similar to the case with a single 2 by 2 plant matrix inversion described in section 7.6.1.

7.6.4 Type of filters

In any case, the cross-over filters can be passive, active or digital filters. Obviously, when the cross-over filters are applied prior to the inverse filters, they can also be applied prior to the binaural synthesis filters **A** in Fig. 2.3. If they are digital filters, they can also be included in the same filters which implement the system inversion in the exactly the same way as the filters for binaural synthesis. As Eq. (7.15) suggests, the inverse filter matrix **H** can also be realised as analogue (active or passive) filters when the "OSD" principle is approximated reasonably well by means of fine discretisation or an ideal variable transducer such as that depicted in Fig. 7.14 and Fig. 7.15.

7.7 Comments on multi-channel systems

When the cross-over filters are not used, then the problem becomes a conventional multi-channel system, contrary to the "OSD" system which is a multi-way 2-channel system. In this case where m is a number of driver pairs, the plant matrix is again a 2 by $(2 \times m)$ matrix of electro-acoustic transfer functions between $(2 \times m)$ outputs of the filter matrix **H** and 2 receivers where $(2 \times m)$ is the number of channels. The pseudo inverse filter matrix **H** is given by Eq. (7.20). The obtained inverse filter matrix **H** is a $(2 \times m)$ by 2 matrix which distributes signals automatically to different drivers so that least effort is required. As an example, the magnitude of the elements of **H** ($|H_{mn}(j\omega)|$) which has 6 channels of transducers at the same position as the drivers used for the examples of the 3-way "OSD" systems with $\nu = 0.7$ (Fig. 7.21 ~ Fig. 7.27) are plotted in Fig. 7.45. The property of multi-channel inversion is beneficial in that frequencies at which there are problems such as ill-conditioning and minima of HRTFs are automatically avoided. On the other hand, with the absence of the cross-over filters, multi-channel systems do not have some of the merits of the "OSD" system.

One of the important advantages is that of the "OSD" system being a multi-way system. The inversion of multi-channel systems ensures that most of the lower frequency signals are distributed to the pair of units with larger span since the condition numbers of the pair are always smaller than the loudspeaker pairs with smaller span at low frequencies. However, some of the higher frequency signals are also distributed to the pairs of units with larger span since there are a number of frequencies for which the larger span gives a smaller condition number due to its periodic nature. This requires the pairs with larger span to produce a very wide frequency range of signals, which is not practical.

Another merit of the "OSD" system, which being a 2-channel system, is also lost in a multi-channel system. Only two independent output signals, hence only two channels of amplifiers, are required for a passive cross-over "OSD" system whereas the same number of channels of amplifiers as number of driver units are always required for a multi-channel system.

7.8 Objective evaluation of the discrete "OSD" system

The main objective here is to realise practical working systems with the OSD principle and investigate its performance. The plant matrix **C** was measured and the practical filter matrix **H** was designed.

7.8.1 Experimental Procedure

The design of a practical OSD system is relatively flexible and the system can be adopted in accordance with particular application. Among a number of practical examples described in Section 7.5, the equally discretised 3-way system and the 2-way system whose discretisation is biased towards high frequencies were realised for the evaluation. The reason for the choice was that these two different systems can be realised without changing the position of driver units.

The three way system was designed to ensure a condition number that is as small as possible over a frequency range that is as wide as possible, as illustrated in Fig. 7.21, Fig. 7.22, and Fig. 7.26. The sub-low region was also covered by the low frequency pair.

The 2-way system is obtained by omitting the lowest frequency units from the 3-way system (Fig. 7.32 ~ Fig. 7.34). The units spanning 32° now cover mid-low frequencies which also covers the sub-low region. For comparison purposes, the "Stereo Dipole" (SD) system described in Chapter 2 was also realised as an example of a conventional 1-way system.

Each driver unit covering a different frequency range was chosen to ensure similar characteristics as far as possible. These drivers were enclosed by closed cabinets mounted on a circular steel frame which was in the horizontal plane including the interaural axis. This ensured the accurate alignment of the units and the listener's head (Fig. 7.46). The distance between the units and the centre of the head (at the intersection of interaural axis and median plane) was set to 1.4m.

Among the choice of cross-over filter types described in Section 7.6.4, passive cross-over networks were used. Their cut-off frequencies were 450Hz/3500Hz for the 3-way system and 3500Hz for the 2-way system.

The plant matrix **C** was obtained using a maximum length sequence (MLS) measurement technique with the KEMAR dummy head microphones with a sampling frequency of 88.2kHz in an anechoic chamber. The data were down sampled to 44.1kHz. The model DB-061 was used for the left pinna and the model DB-065 was used for the right pinna to obtain two sets of plant matrices. However, the data obtained with DB-065 was used for the later evaluation. The free field response of each loudspeaker system was also measured with a free field microphone.

Among a number of methods described in Section 7.6, the inverse filter matrix **H** was designed from a single 2 by 2 plant matrix as in Section 7.6.1. This is because it could exclude a number of additional factors which could potentially affect the performance such as the effect of cross-over filter response, the effect of the driver units' response, and the effect of interaction between two different frequency unit pairs around cross-over frequency regions by including all of them within the plant matrix to be inverted. This may not necessarily be the best way of realising the practical system but would enable the fundamental investigation of these different systems. More detailed descriptions of the experimental procedure can be found in [52].

7.8.2 Results

The frequency responses of the loudspeaker system for the 3-way, the 2-way, and the SD system are shown in Fig. 7.47. In general, each system had a reasonably good response within the operating frequency range. The SD and the 2-way system start loosing their low frequency response earlier than the 3-way system. The slope is the steepest for the SD. For the high frequency end, the 3-way and the 2-way loudspeakers have much better response, more specifically a reasonably good response up to about 16kHz, compared to the SD system whose output is up to about 10kHz. Being a single driver system, it is inevitable that the loudspeakers for the SD have narrower operational frequency range. This is a fundamental but one of the important advantages of the OSD system.

The plant responses including loudspeaker responses and HRTFs of these three systems are shown in Fig. 7.48. This reveals two characteristic notches in the response due to the KEMAR head and pinna around 3.2kHz and 8kHz in addition to the loudspeaker responses. It should be noted that analysis in previous chapters assumes flat (free field) plant response. Therefore, each plant response induces additional load to the system inversion. The frequencies where the plant response is low compared to its peak frequencies could not be expected to show good control performance. The OSD system has advantages over the SD system in this respect as well, having relatively uniform plant response over a wide frequency range.

The practical inverse filter matrix **H** was designed from these plant measurements with a method described in [29] and the elements of the resulting control performance matrix $\mathbf{X} = \mathbf{CH}$ are shown in Fig. 7.49. Even though $\|\mathbf{X}\|$ is ensured to be unity (0dB), the degradation of control performance (independent control of two binaural signals) can clearly be seen by the decrease in diagonal elements (X_{11}, X_{22}) and, especially, by the increase of non-diagonal terms (noise (cross-talk), X_{12}, X_{21}). The frequency range where signal to noise ratio was more than 10dB was between 50Hz and 16kHz for the 3-way OSD system, between 130Hz and 16kHz for the 2-way OSD system, but only between 400Hz and 8kHz with a number of non-controlled frequencies above 4kHz for the SD system. It should be noted that much better cross-talk supression than 10dB is required to control binaural signals correctly where the cross term (noise) input level can be significantly larger than the diagonal term input level at the contra lateral ear in most cases. Bearing this in mind, the frequency range where signal noise ratio was more than 40dB was between 350Hz and 15kHz. for the 3-way OSD system, 700Hz and 15kHz by the 2-way OSD system, but 1kHz and 4kHz for the SD system. It should however be noted that having no control does not mean there are no reproduced signals since $\|\mathbf{X}\|$ is unity at most of the frequencies. It just means that the binaural signals are not fed to each ear independently. It is also noted that the good control performance with the SD is available around the frequency where the frequency-span relationship of the OSD principle is fulfilled, not in the region where the dipole approximation holds as its name suggests.

7.9 Subjective evaluation

A series of subjective evaluations were carried out in order to investigate the performance of the OSD systems. The SD system was used again as a reference. The inverse filters were implemented with a digital signal processor. Twelve young adults who all had normal hearing with no history of hearing problems, served as paid volunteers. The evaluation was performed in an anechoic chamber.

7.9.1 General impression

First, a number of binaural recordings were played to listeners with the 3-way OSD system, the 2-way OSD system, and the SD system. These were recorded with the Aachen Head from Head Acoustics and Neumann KU 100. Because of the wider operational frequency range and greater dynamic range of the OSD systems, the improvement in sound quality was significant compared to the SD system. In addition, the spatial impression was very different. The OSD system was reported to give the stronger impression of being "in" the sound environment. The 3-way OSD system clearly gave the best impression among the three. It is supposed that better and robust control performance over a wider frequency range probably leads to better synthesis of time and frequency domain localisation cues, which then gave the better spatial perception.

7.9.2 Localisation experiment

A systematic localisation test was carried out to investigate the difference in spatial perception further. The 3-way OSD system and the SD system were used for this detailed investigation. Presentation of a single incident sound wave from various directions is investigated as it is the very basic element consisting of a complex sound environment. Pink noise was used as the source signal because of its flat response on a logarithmic frequency scale. The HRTF database measured at MIT Media Lab [38] was used for the binaural filters **A**.

An adjustable chair and a small head-rest were used in order to ensure the head of the listener was positioned correctly regardless the inter-subject difference in body size. It is believed that the subject's head was always within ±10mm of the correct position. The headrest constrained head movement very well, especially the rotational movement which could give false localization cues to subjects. A spherical grid made of thin metal wires surrounded the subject's head in order to give a guide to work out coordinates of perceived directions (Fig. 7.50). The grid is painted in light blue and formed a vertical polar coordinate system with a radius of 1m. The subjects were expected to be more familiar with this coordinate system compared to the interaural polar coordinate system. There were wires every 15° that were labelled with red numbers for azimuth and blue numbers for elevation directions. It was found in preliminary experiments that the subjects can produce a large error when they report a direction without seeing the reference coordinate system. The magnitude of the error in reporting the coordinate is as large as 40° especially when the direction is in the rear hemisphere. The visible coordinate reference reduces this error down to about 5° at the expense of increasing visually related error mainly in the front hemisphere where localisation accuracy is much finer than 5°. A thin black acoustically transparent fabric surrounded the subject supported by the wires in order to minimize the effect of visual information. The subjects could not see anything outside the screen.

A set of 59 stimuli of pink noise with synthesized direction with a duration of 2 seconds each with a gap of 0.5s were presented prior to each set of tests. The directions used were different from those used for the later localization test. The sequence of the stimuli was consistent with vertical polar coordinate system. The purpose of this session is to let subjects become familiar with the sound source signal and sound environment both of which are extremely unusual to them. After a short break, a set of localisation tests were performed.

Each stimulus consisted of a reference signal and a test signal. A reference signal was presented at 0° azimuth and 0° elevation, i.e., directly in front of the listener before each test signal. Both signals had the same sound source signal with durations of 3 seconds for the reference signal and 5 seconds for the test signal with a gap of 3 seconds in between. Directions shown in Fig. 7.51 were chosen for the presentation ensuring equal sampling density from all spherical directions except downwards. They were selected so that each of them is approximately on one of the cones of constant azimuth directions at -80°, -60°, -40°, -20°, ±0°, +20°, +40°, +60°, or +80° in the interaural polar coordinate system. If there were two directions symmetric with respect to the median plane, one of them was omitted to reduce the test duration. Filled circles represent the directions that were used for the localisation tests. The directions that were omitted are denoted by open circles. In order to avoid the effect of presentation order, the order of presentation was randomised. The reference signal not only cancelled the order effect, but also gave subjects prior knowledge of the sound source signal spectrum that is important for the monaural spectral cue.

The subject was instructed to look straight ahead and not to move the head nor body while the stimuli were presented in order to avoid introducing dynamic cues that relate to head movement. The subject's movement was monitored by the experimenter to ensure the instruction was obeyed. The subject's head was not physically fixed but the subjects were instructed to lean against the headrest. The subject was instructed to turn his head after each test stimulus had stopped to evaluate the direction of the sound and state this to the experimenter. The stimuli, a set of reference and test signals, were repeated when subjects had difficulty in making a judgement. The subjects were allowed to choose more than one direction when they perceived two or more separate directions of sound. However, there were only a few cases where such judgement occurred.

7.9.3 Results

The following results use interaural polar coordinates throughout since it coincides with the characteristics of the human auditory function. The cones of constant azimuth roughly represents the cone of confusion where the interaural time difference is constant. This similarity does not hold where the azimuth is very large.

The perceived virtual sound source directions are shown in Fig. 7.52. Filled circles represent the directions that are perceived and its size represents the number of occurrences of the perceived direction. The presented directions are denoted by open circles. Fig. 7.52a shows the results for the OSD system. The responses are evenly spread over various directions except the region above the head. Fig. 7.52b shows the results for the SD system. Contrary to the OSD system, the responses are clustered around the elevated directions at moderate azimuth directions and directly behind the listener. There are significantly more perceptions in front than in the rear.

Fig. 7.53 shows the relation between the presented azimuth direction and the perceived mean azimuth direction for each individual subject (star marker). The overall perceived mean azimuth direction is also plotted (square marker). The dash-dot line shows the relation where presented and perceived directions are the same. The OSD system showed much better azimuth localisation performance for most of the subjects especially towards large azimuth directions. There was little difference in standard deviation between two systems within each subject. The standard deviation among the subjects is very large for the SD system at most of the azimuth directions except 0° azimuth whereas it is very small for the OSD system. This means the OSD system is robust against errors due to individual differences. The error for the OSD system is not much larger than the discrepancy between the cone of constant azimuth and the cone of confusion. The OSD system shows slightly larger azimuth response around 30° azimuth.

Fig. 7.54 shows the relation between presented elevation direction and perceived elevation direction for all subjects. The solid line shows the relation where presented and perceived directions are the same. The dashed line shows a symmetric direction with regard to the frontal plane. Therefore, the responses due to front-back confusion fall around this dashed line. The data with ±80° azimuth directions are omitted because the cone of constant azimuth and the cone of confusion are so different at this high azimuth that there is little relevance in elevation direction. The data on the median plane are also omitted from this plot since they showed a completely different tendency from the other azimuth directions.

There were relatively small differences in performance between two systems. In general, the performance was poor with large random errors. However, the tendency is slightly biased towards the horizontal plane for both systems. One of the many reasons is that when localisation cues for elevation perception are ambiguous, the response tends to fall around the horizontal plane. Nevertheless, this bias towards the horizontal plane was slightly less significant for the OSD system. The SD system also has stronger bias error towards the upper hemisphere. On the contrary, the error for the median plane data showed strong bias towards the horizontal plane rather than random error.

Fig. 7.55 shows the rate of back to front reversals and front to back reversals for each individual subject. The SD shows much more back to front reversals than the OSD. The OSD shows slightly more front to back reversals than the SD. The rate for back to front reversals and front to back reversals are about balanced around 15% average for the OSD system, but for the SD, there are much more back to front reversals (average 23%) than front to back reversals (average 10%). A notable remark is that elevation (including front-back) errors by the SD are large bias errors contrary to smaller random errors by the OSD.

The average angular errors, defined as the angle measured on a great circle between presented and perceived directions, are shown in Fig. 7.56. The OSD system showed smaller errors for all the subjects apart from subject 12. The mean error was 32.4° for the OSD system and 37.7° for the SD system.

7.10 Conclusions

Analysis with a free field model and more realistic plant matrix with head related transfer functions reveals a number of fundamental problems related to multi-channel sound control with system inversion such as binaural synthesis over loudspeakers. A principle of 2-channel (binaural) sound control with loudspeakers is proposed which overcomes the fundamental problems with system inversion by utilizing a variable transducer span. Practical ways to tackle the sub-low frequency region where the frequency-span relationship is remote from the optimal are also described.

The proposed principle has various advantages. No dynamic range loss due to system inversion directly means good signal to noise ratio but also leads to less distortion and longer life of transducers. The robustness to errors has advantages in many respects, e.g. incorrect inverse filters due to restriction of hardware resources, differences between individuals or products, and the misalignments that are inevitable in practical use. The directivity of sound radiation reduces the chance of the 3D effect being destroyed by reflections from surrounding objects. The system inversion does not result in coloration because of the flat response of the inverse filters, and this adds practicality by enabling the listener to enjoy the reproduced sound signals even outside the "sweet region". As a natural consequence of this, the reflections or reverberation of the room are not coloured either.

The practical system can be realised in a number of ways including discretising the theoretical continuously variable transducer span that results in multi-way sound control system. The discretisation enables the use of conventional transducer units and cross-over filter networks. The relationship between the position of a driver unit and the frequency region to be covered can be determined easily. Further developments to realize ideal continuously distributed transducer will be beneficial to improve the performance of such systems.

The objective evaluation of the realised discrete OSD systems revealed many practical advantages as well as confirmed the superiority of the principle itself. Due to wider operational frequency range and much smaller dynamic range loss, the improvement in sound quality was obvious. The subjective evaluation confirmed the superiority of the principle in terms of spatial perception.

Related publications

[A 9] T. Takeuchi and P. A. Nelson, "Optimal Source Distribution", UK Patent Application No.: 0015419.5

[A 10] T. Takeuchi and P. A. Nelson, "Optimal source distribution for virtual acoustic imaging," ISVR Technical Report No.288, University of Southampton (2000).

[A 11] T. Takeuchi and P. A. Nelson, "Optimal source distribution for binaural synthesis over loudspeakers", Acoustic Research Letters Online, 2(1), 7-12 (2001).

[A 12] T. Takeuchi and P. A. Nelson, "Optimal source distribution for virtual acoustic imaging," presented at the 140th Meeting of the Acoustical Society of America and Noise-con 2000, 3-8 December 2001, Newport Beach, California, USA.

[A 13] T. Takeuchi and P. A. Nelson, "Optimal source distribution system for virtual acoustic imaging," 110th AES Convention Preprint 5372 (L8).

[A 14] T. Takeuchi and P. A. Nelson, "Optimal source distribution for binaural synthesis over loudspeakers", to be published in J. Acoust. Soc. Am.

[A 15] T. Takeuchi and P. A. Nelson, "Optimal source distribution for binaural synthesis over loudspeakers (in Japanese)", The proceedings of the 2001 Autumn meeting of the acoustical society of Japan, 573-574 (2001).

[A 16] T. Takeuchi, M. Teschl and P. A. Nelson, "Subjective evaluation of the Optimal Source Distribution system for virtual acoustic imaging," The proceedings of the AES 19th International Conference, 21-24 June, 2001, Schloss Elmau, Germany, 373-385.

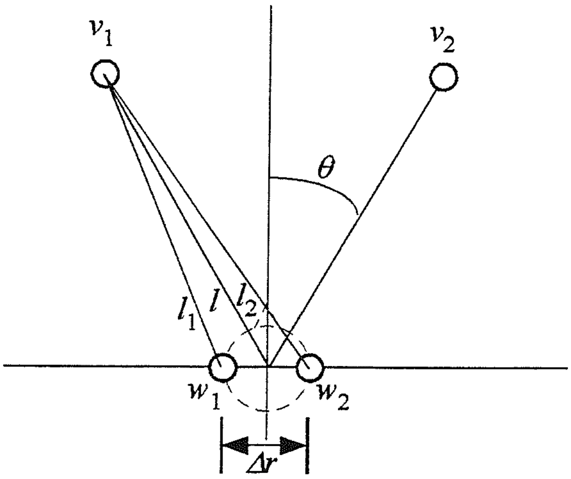


Fig. 7.1 Geometry of a 2-source 2-receiver system under investigation.

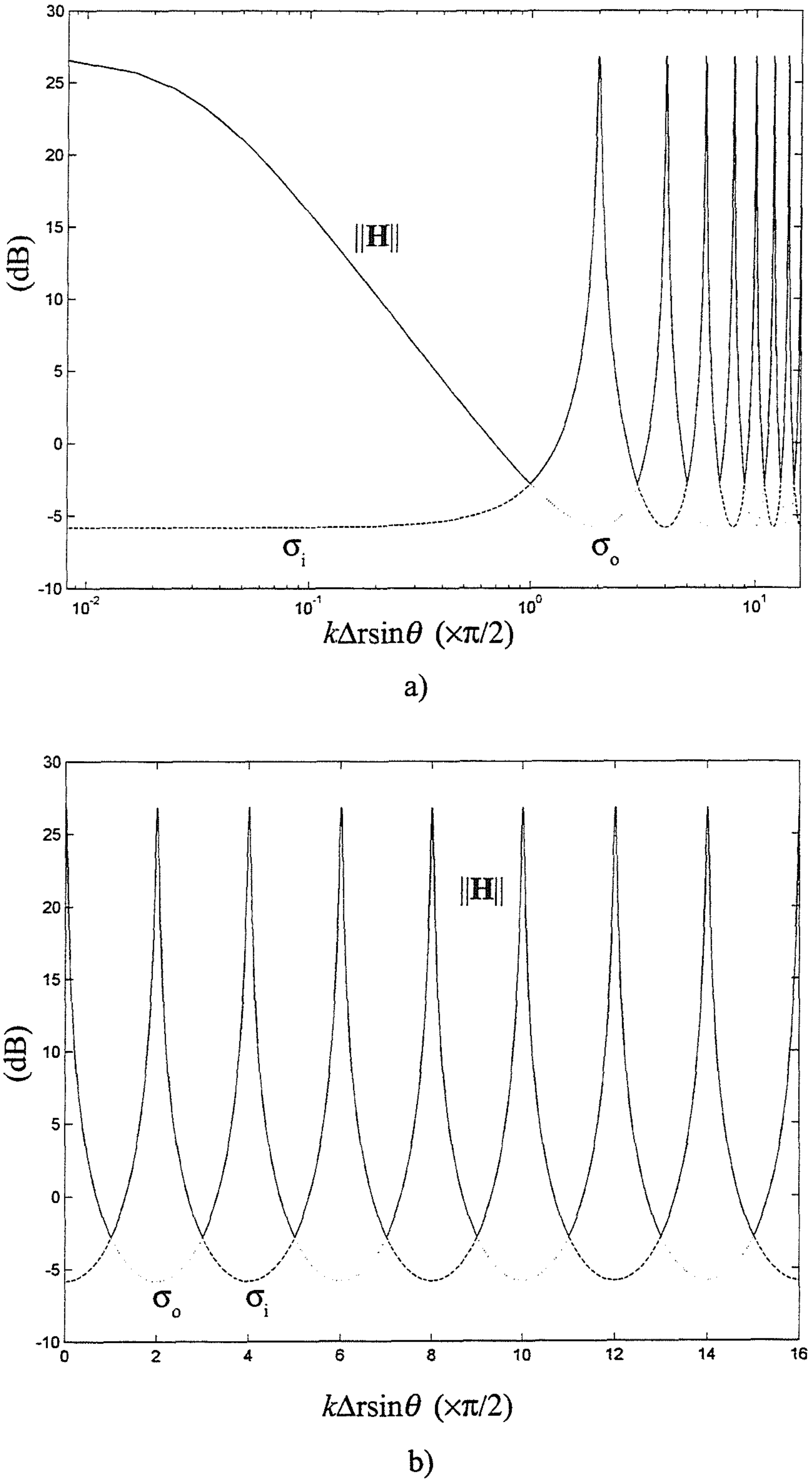


Fig. 7.2 Norm and singular values of the inverse filter matrix **H** as a function of $k\Delta r\sin\theta$. a) Logarithmic scale. b) Linear scale.

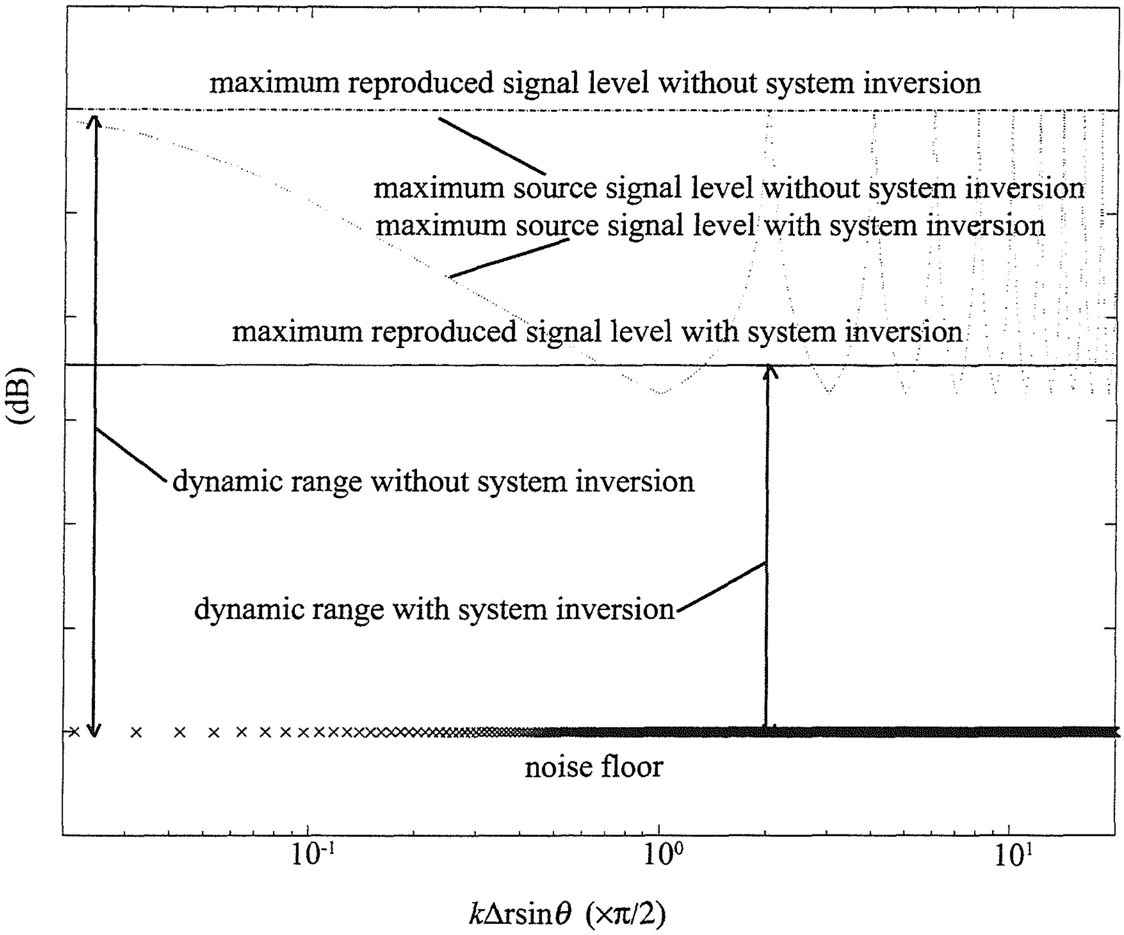


Fig. 7.3 Dynamic range loss due to system inversion.

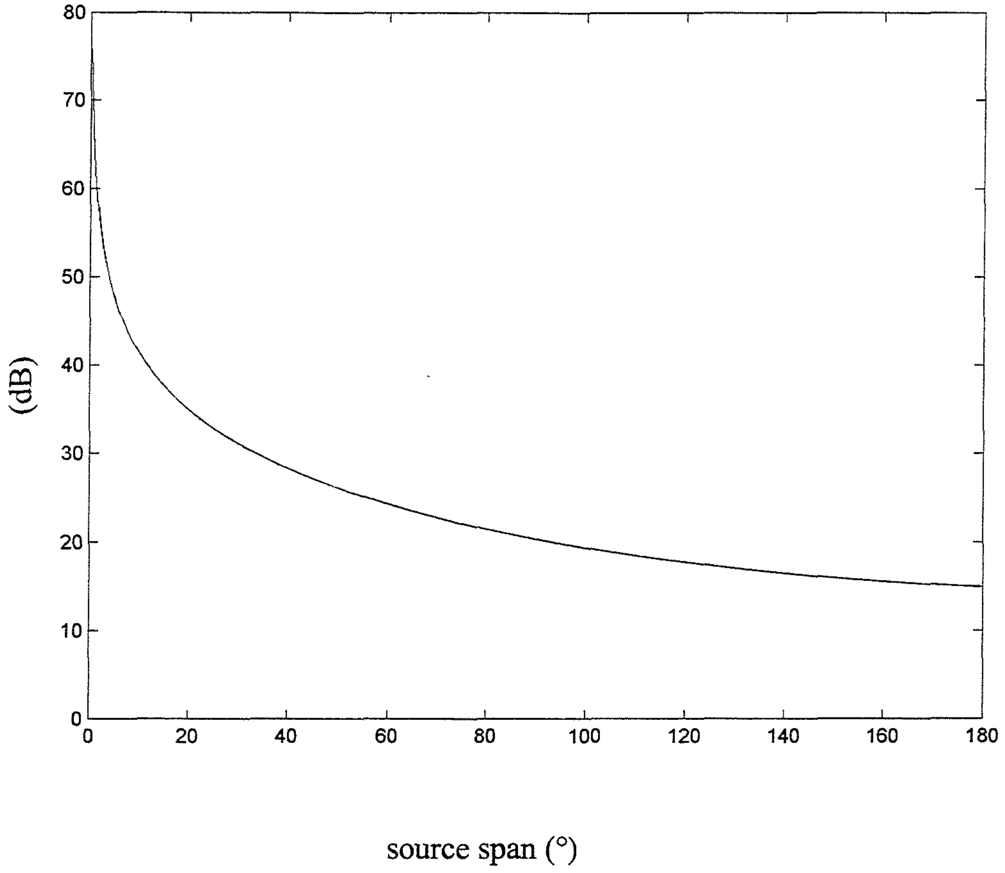
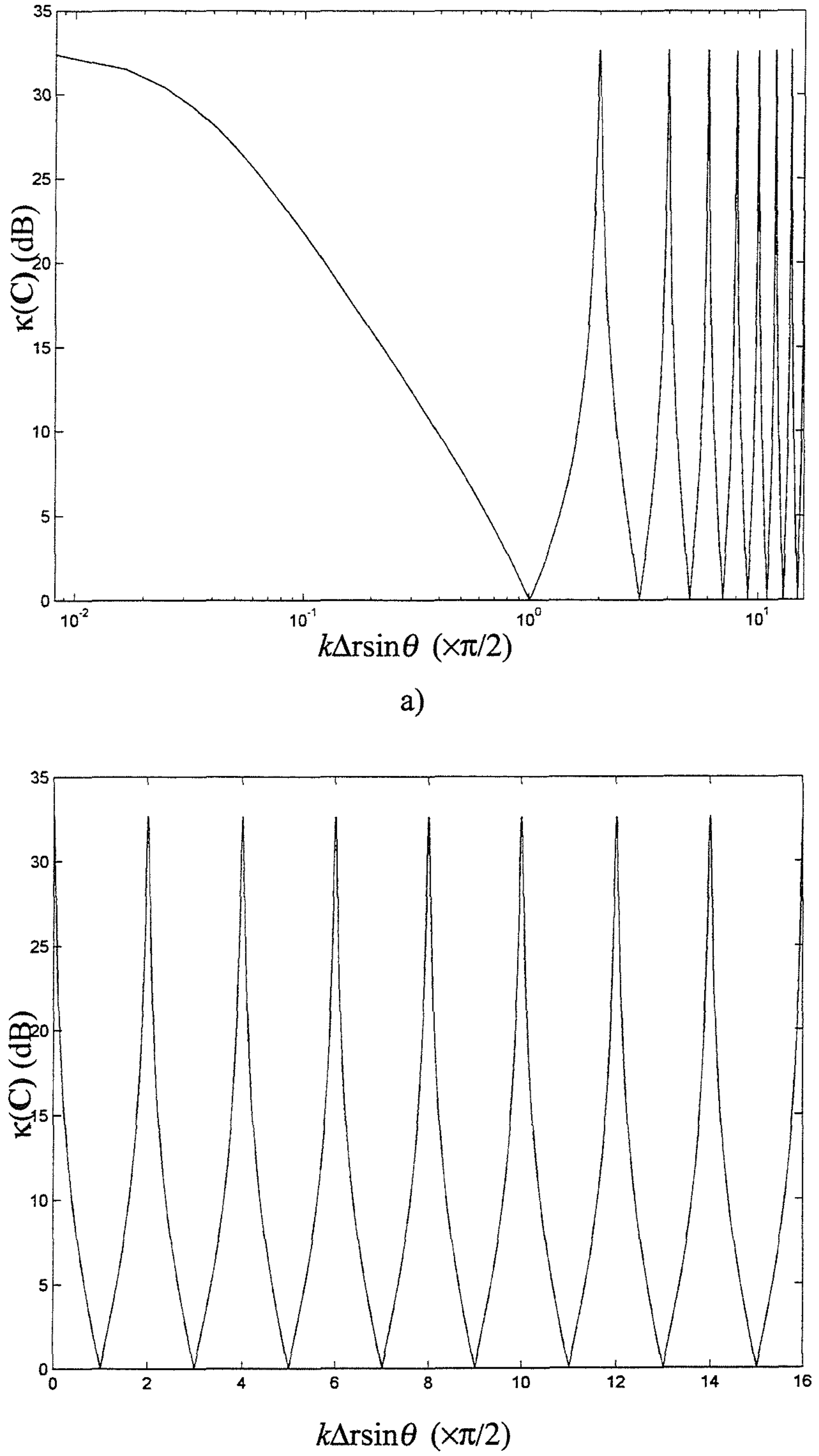


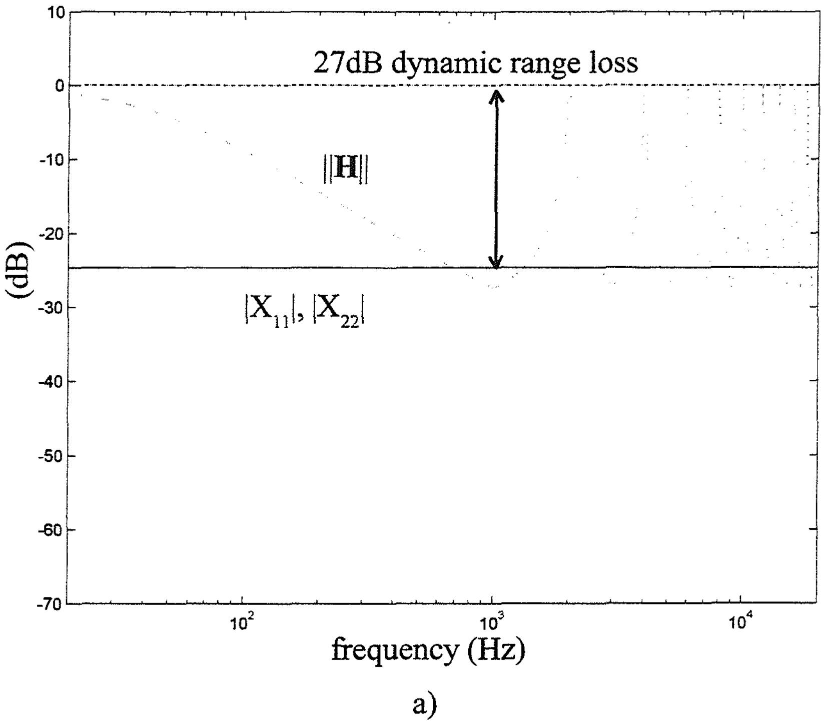
Fig. 7.4 Dynamic range loss as a function of source span.



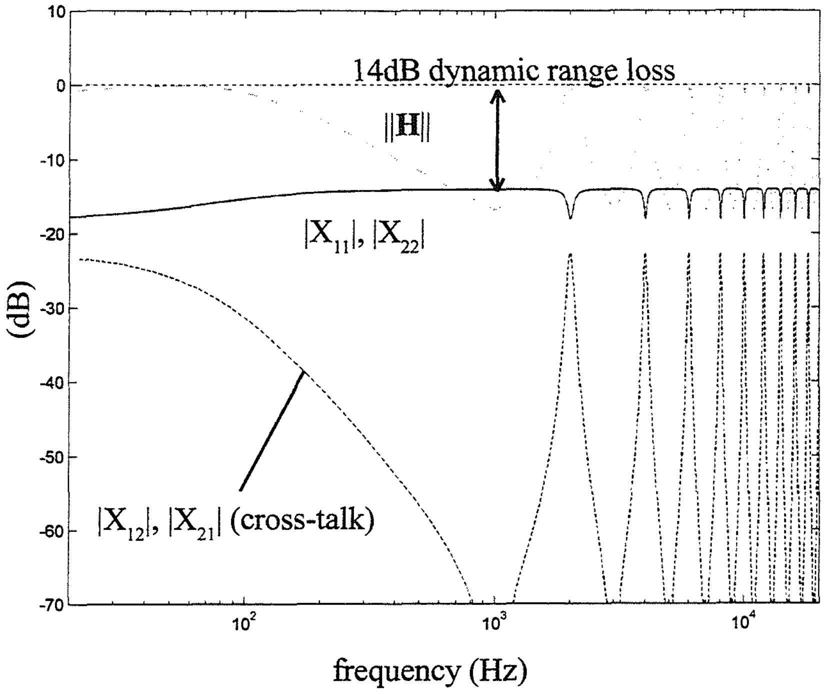
a)

b)

Fig. 7.5 Condition number κ(**C**) as a function of *n*. a) Logarithmic scale. b) Linear scale.

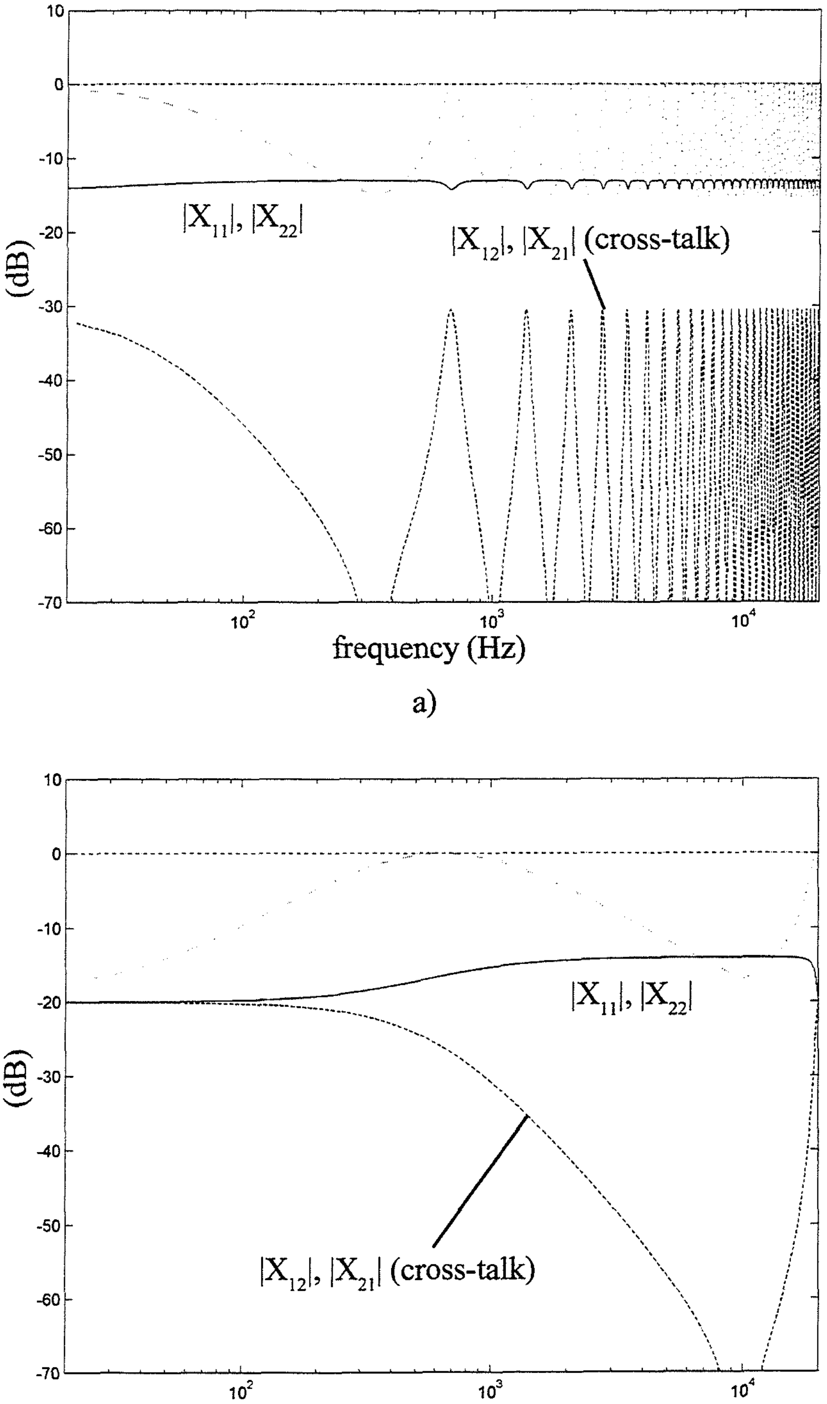


a)



b)

Fig. 7.6 Dynamic range improvement and loss of control performance with regularisation. a) Without regularisation. b) With regularisation



a)

b)

Fig. 7.7 Effect of changing source span. a) Larger source span (180°). b) Smaller source span (7.2°)

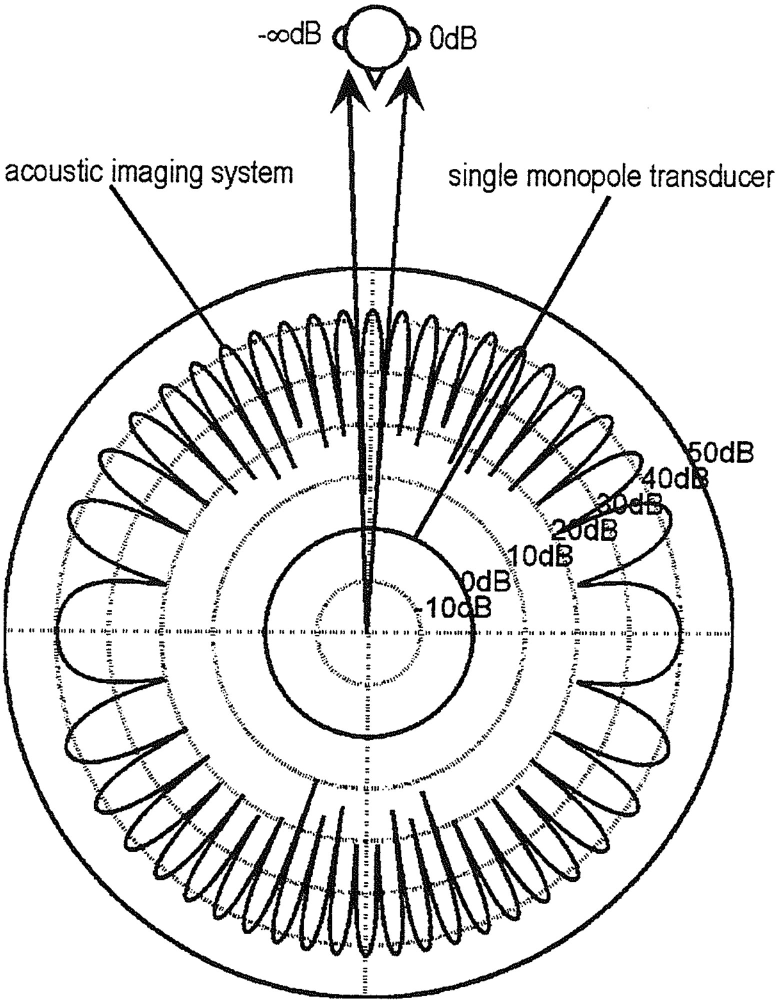


Fig. 7.8 Sound radiation by the control transducer pairs with reference to the receiver directions (0dB and -8dB).

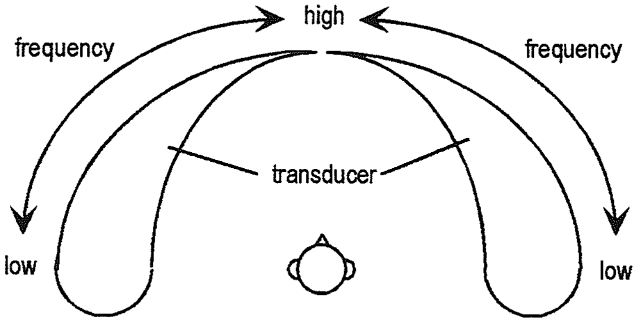


Fig. 7.9 Principle of the "OSD" system.

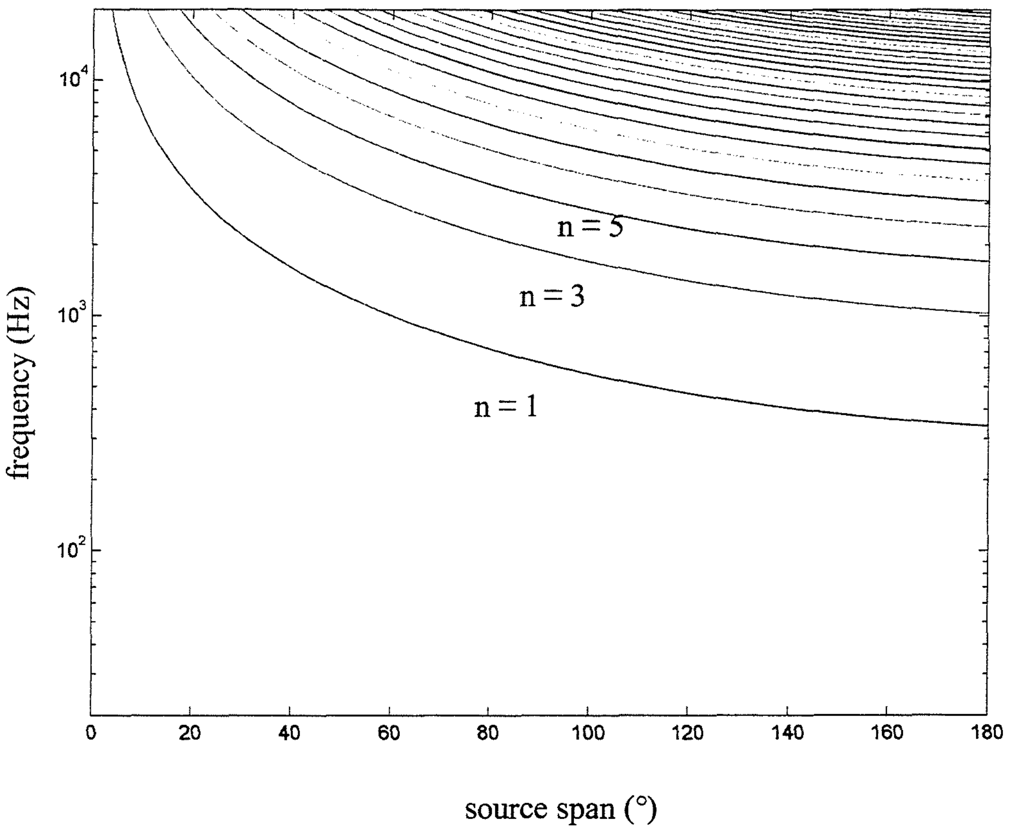


Fig. 7.10 Relationship between source span and frequency for different odd integer number *n*.

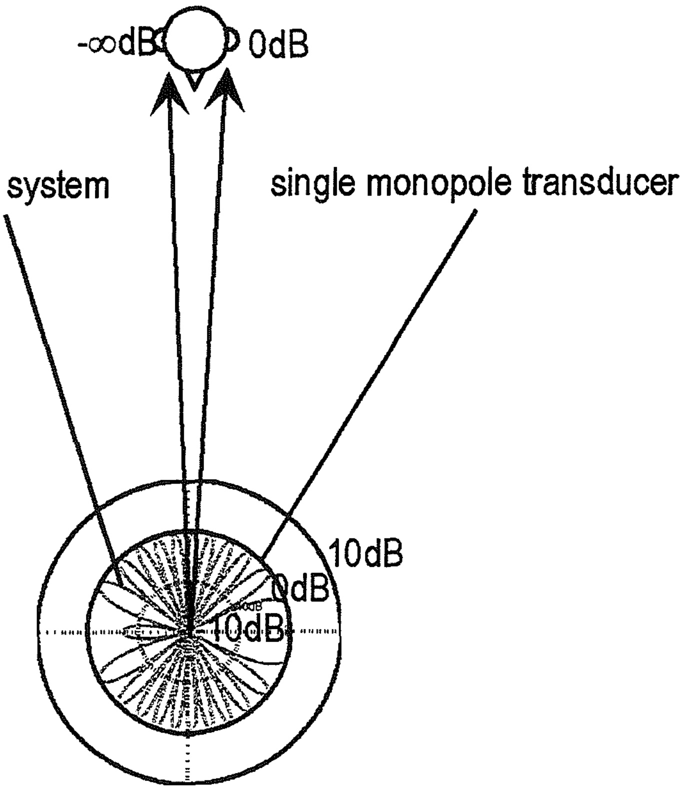


Fig. 7.11 Sound radiation by the "OSD" loudspeakers with reference to the receiver directions (0dB and -8dB).

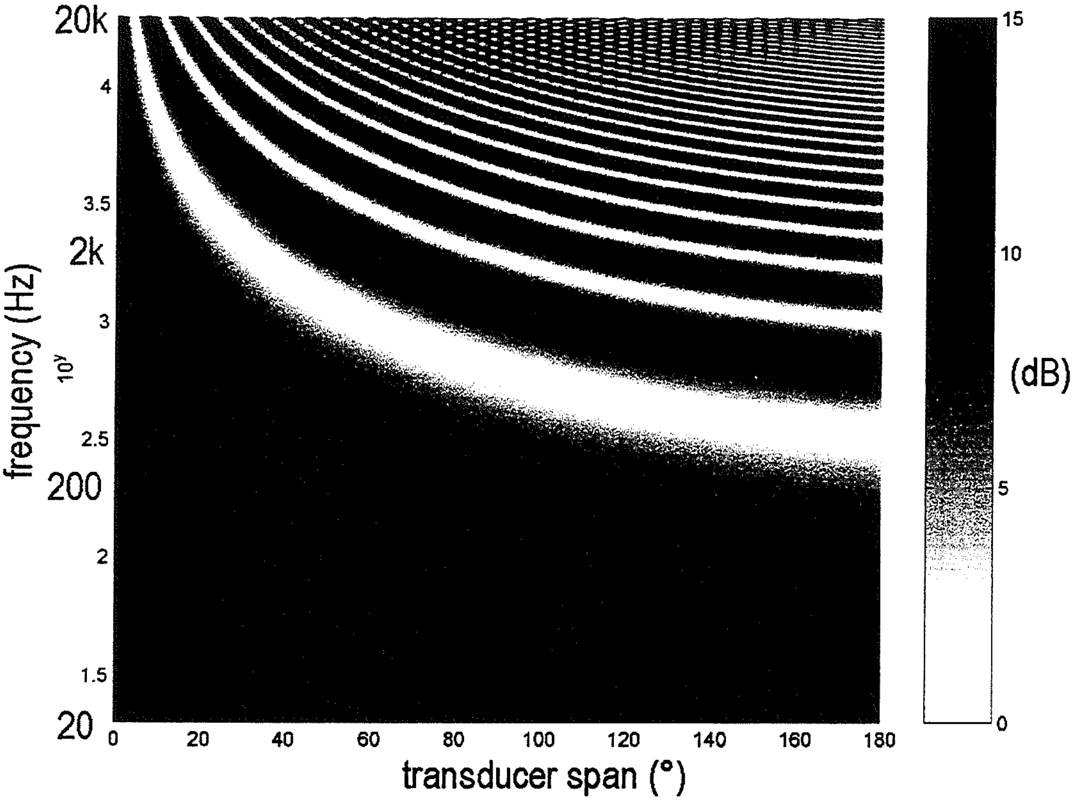


Fig. 7.12 Condition number κ(**C**) of a free field plant matrix **C** as a function of source span and frequency.

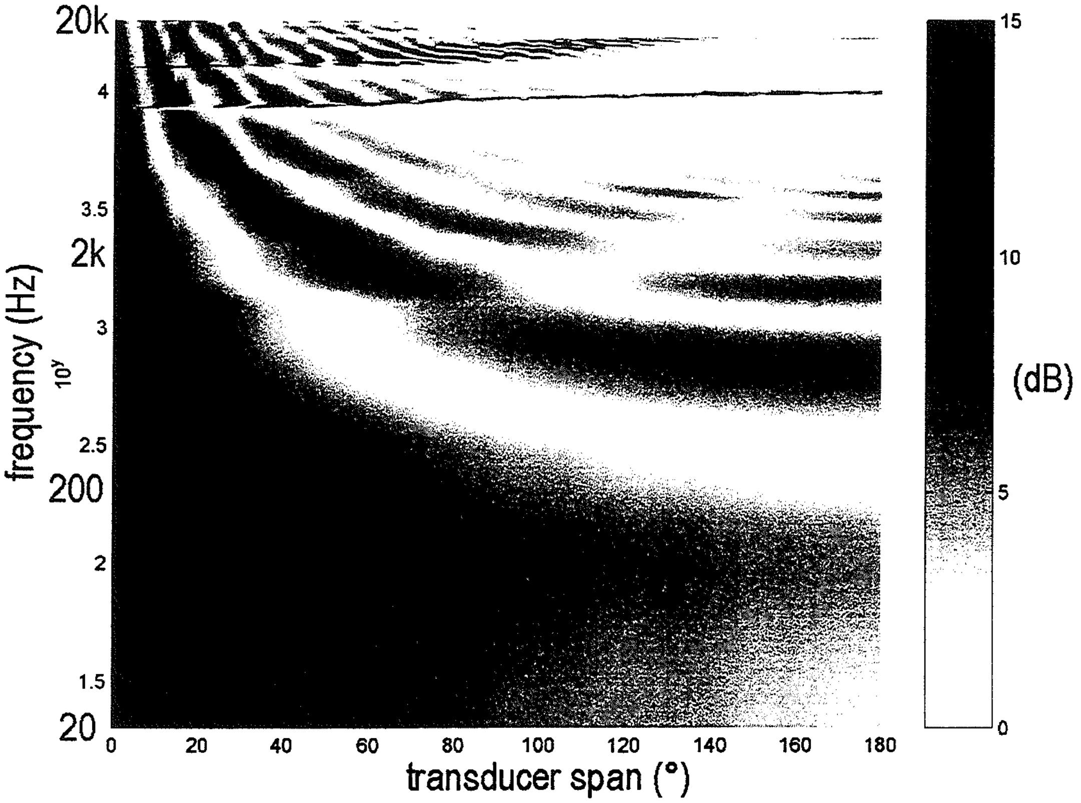


Fig. 7.13 Condition number κ(**C**) of a plant matrix **C** with HRTFs as a function of source span and frequency.

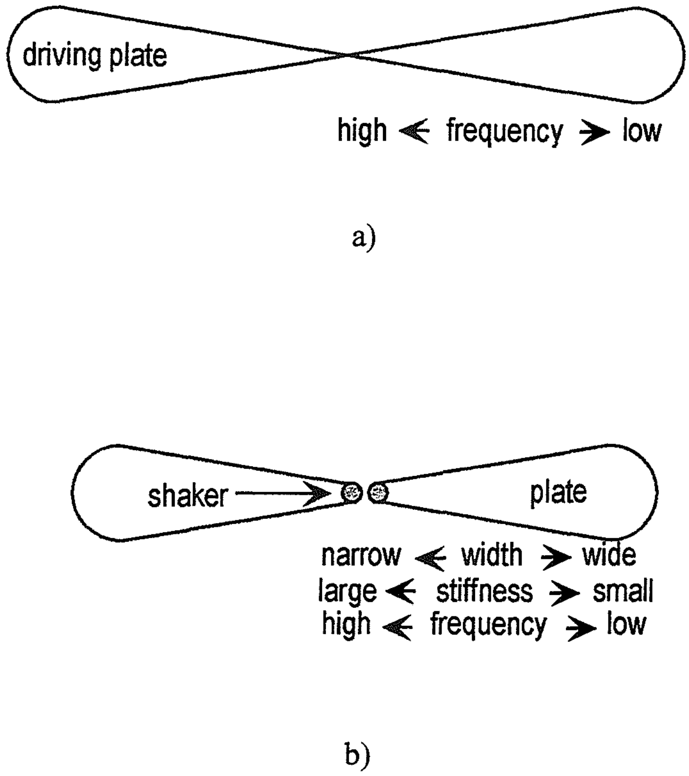


Fig. 7.14 Flat panel transducers. a) individual excitation. b) point excitation.

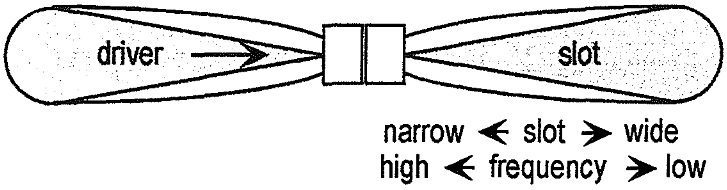


Fig. 7.15 Acoustic waveguide type transducers.

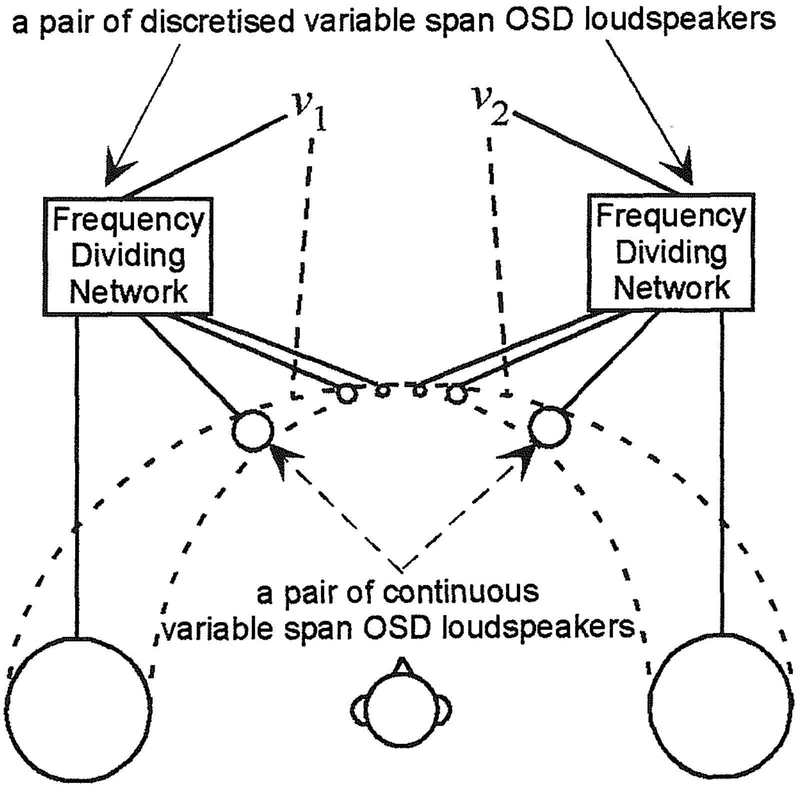

a pair of discretised variable span OSD loudspeakers
v_1
v_2
Frequency Dividing Network
Frequency Dividing Network
a pair of continuous variable span OSD loudspeakers

Fig. 7.16 Discretised variable frequency / span transducers.

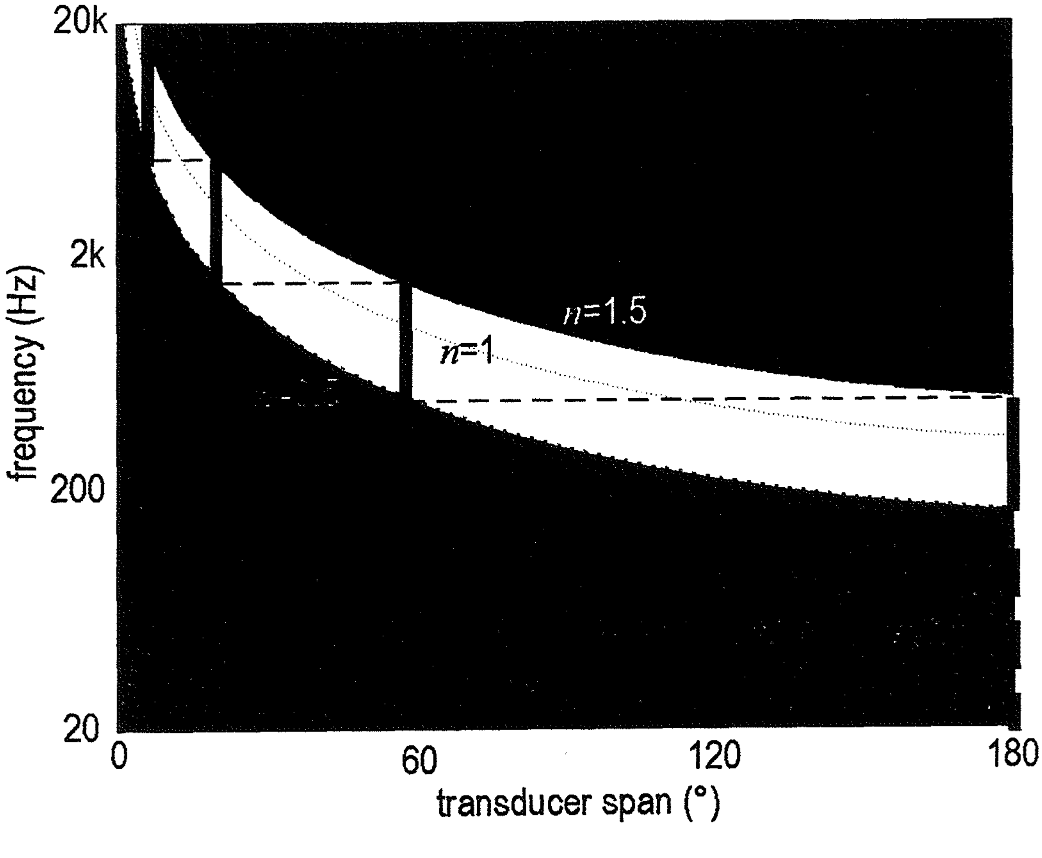


Fig. 7.17 An example of frequency/span region and discretisation

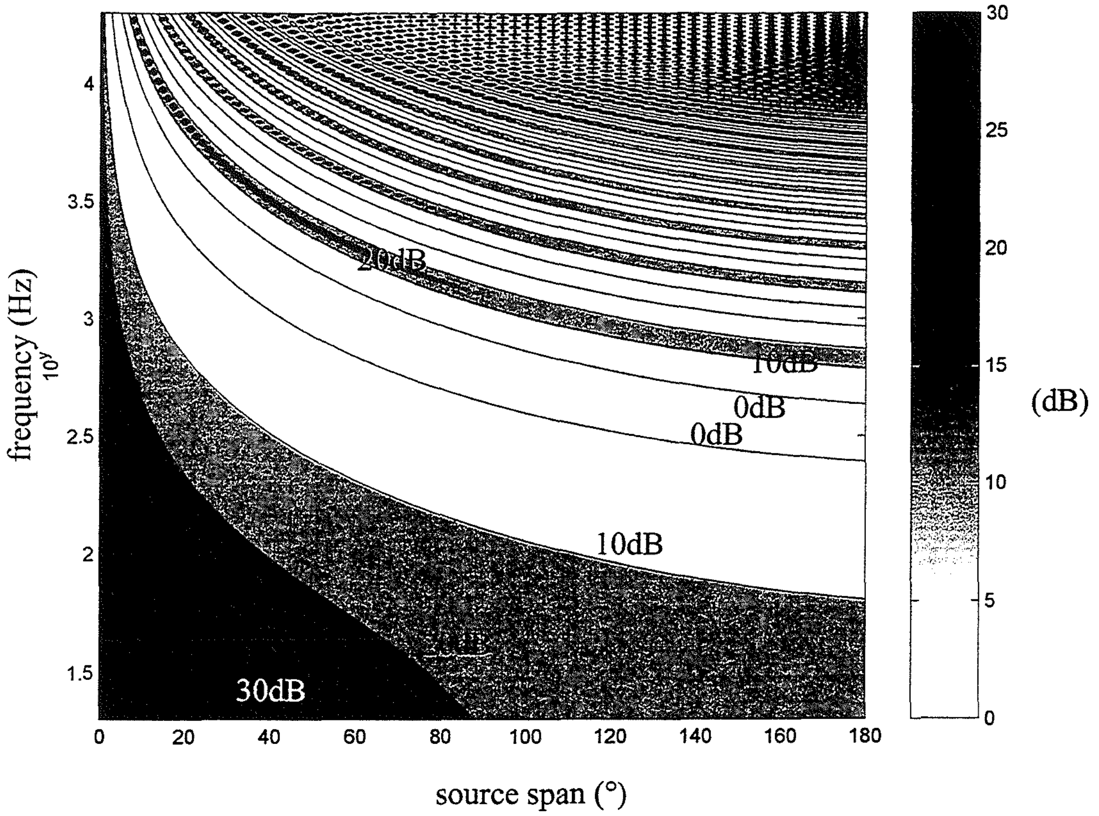


Fig. 7.18 Required dynamic range loss as a function of source span and frequency range.

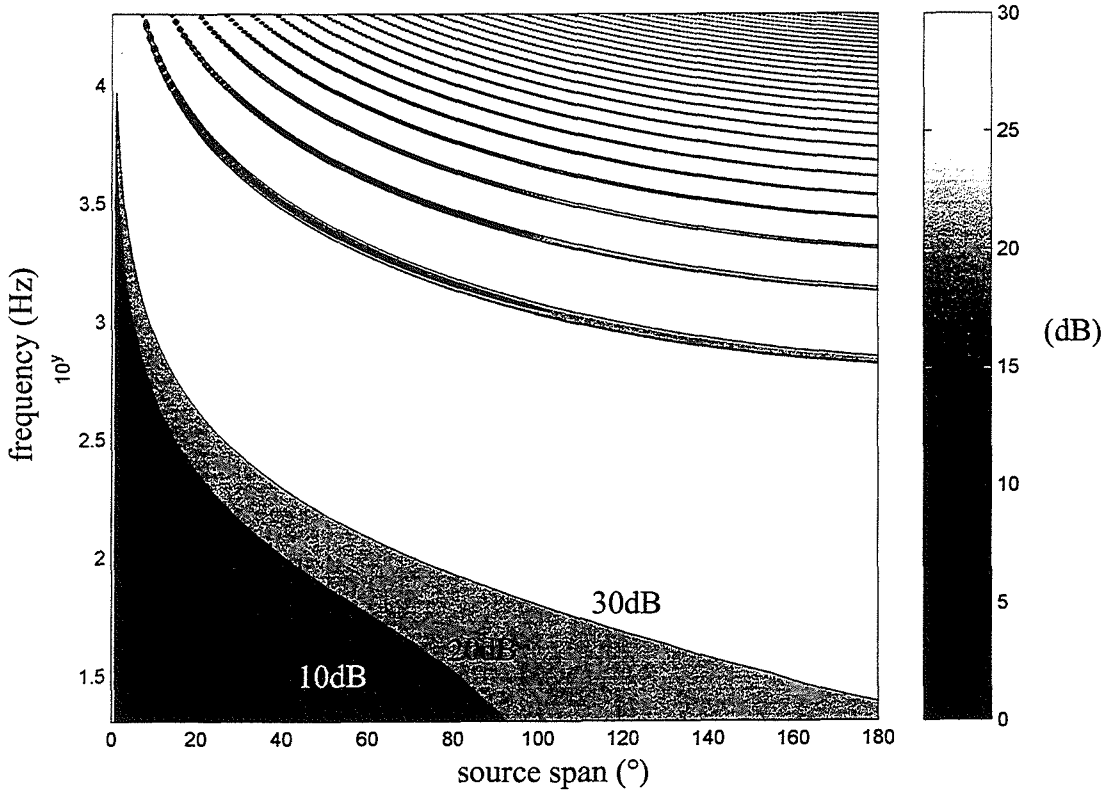


Fig. 7.19 Cross-talk cancellation performance as a function of source span and frequency with 5% of error.

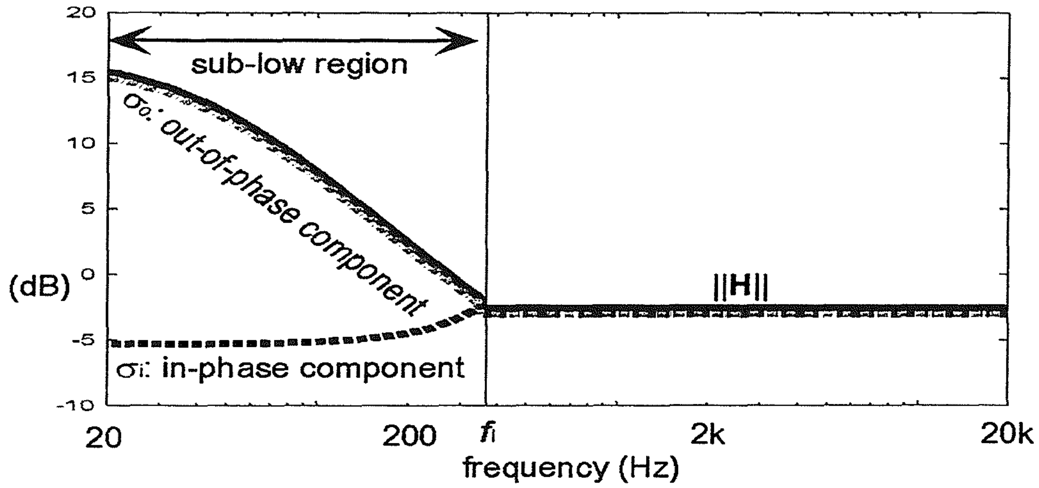


Fig. 7.20 Norm and two singular values of the inverse filter matrix **H** with "OSD" principle.

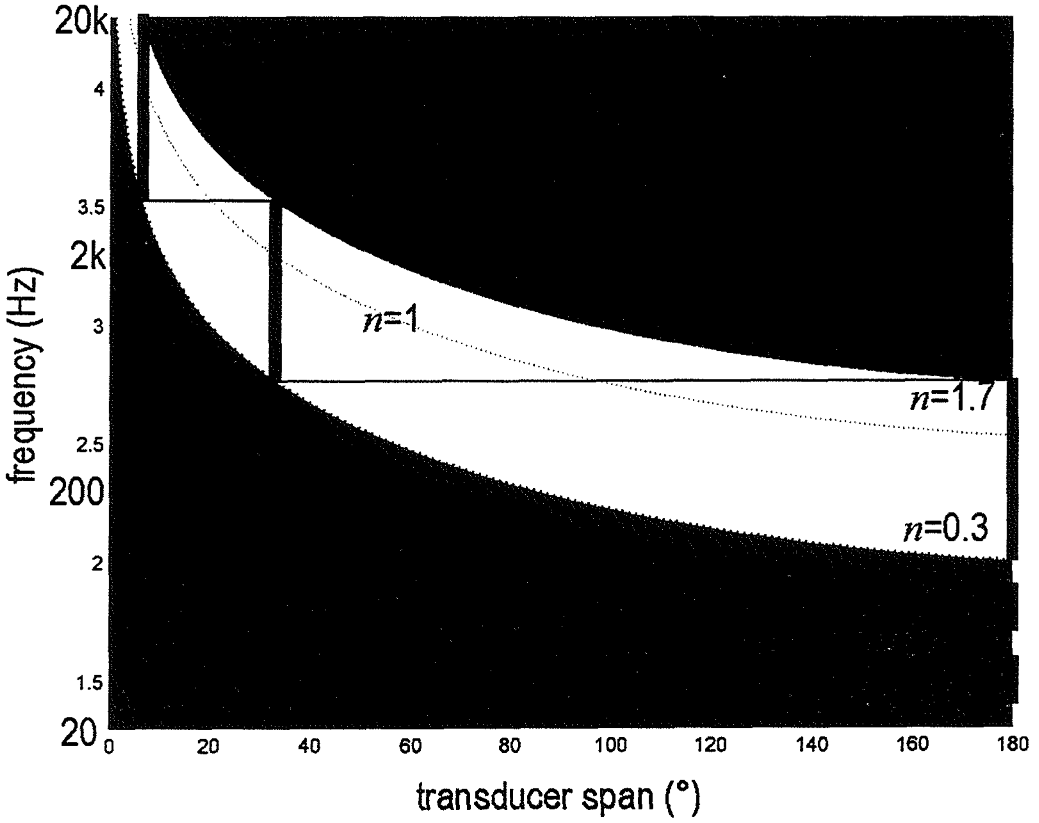


Fig. 7.21 Frequency/span region for systems with $n \gg 1$ and $n = 0.7$, and an example of discretisation for a 3-way system.

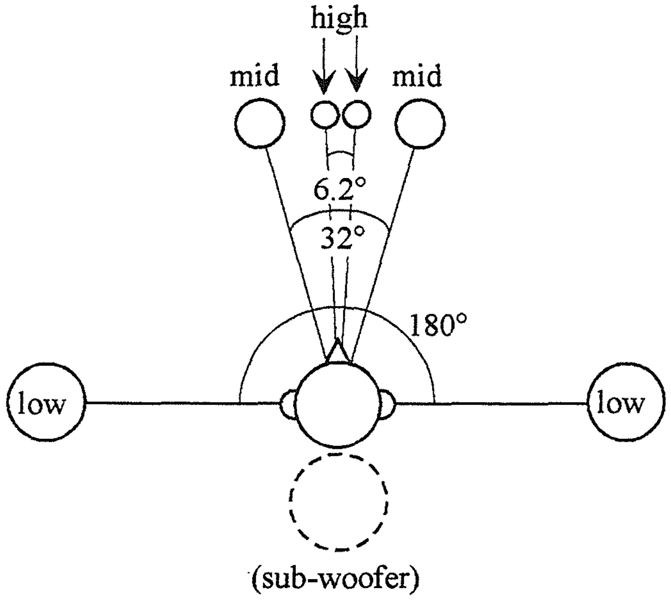


Fig. 7.22 An arrangement of a 3-way system with *n* » 1 and *n* = 0.7.

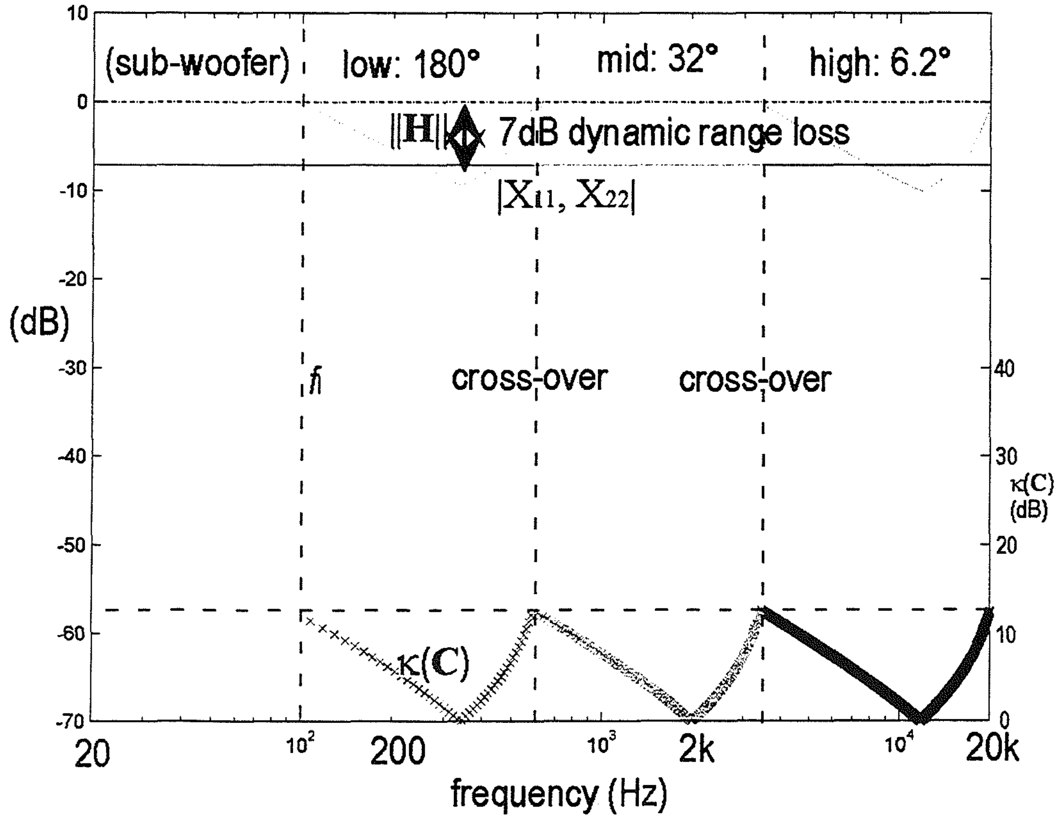
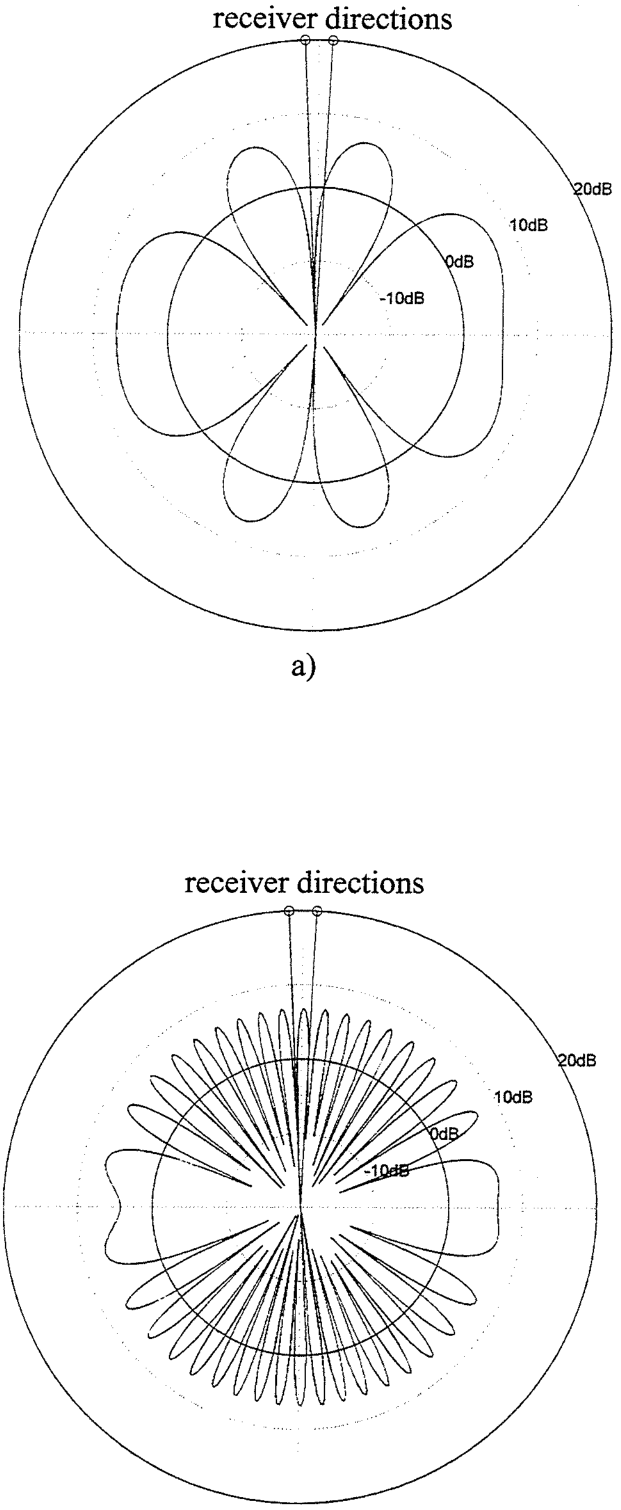


Fig. 7.23 Performance of a 3-way system with *n* » 1 and *n* = 0.7.



a)

b)

Fig. 7.24 Maximum far field sound pressure level produced by the transducer pairs with reference to the receiver directions (0dB and -¥dB). $n = 0.7$. a) $n = 0.3$ b) $n = 1.7$

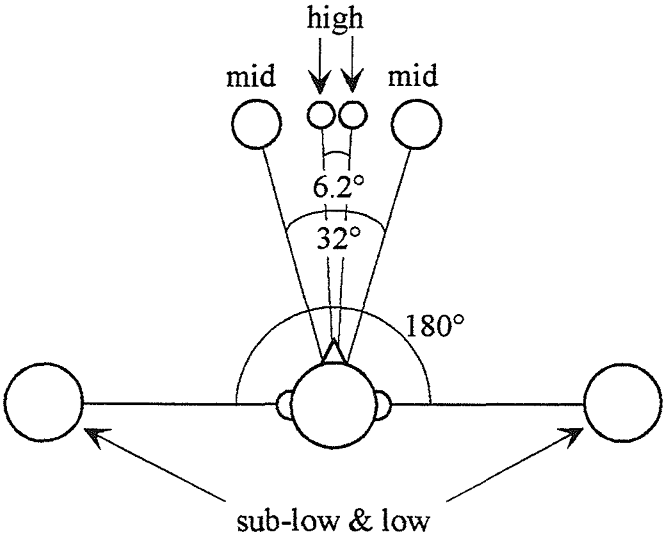


Fig. 7.25 An arrangement of a 3-way system with n » 1 and $n = 0.7$ with the lowest pair covering sub-low region.

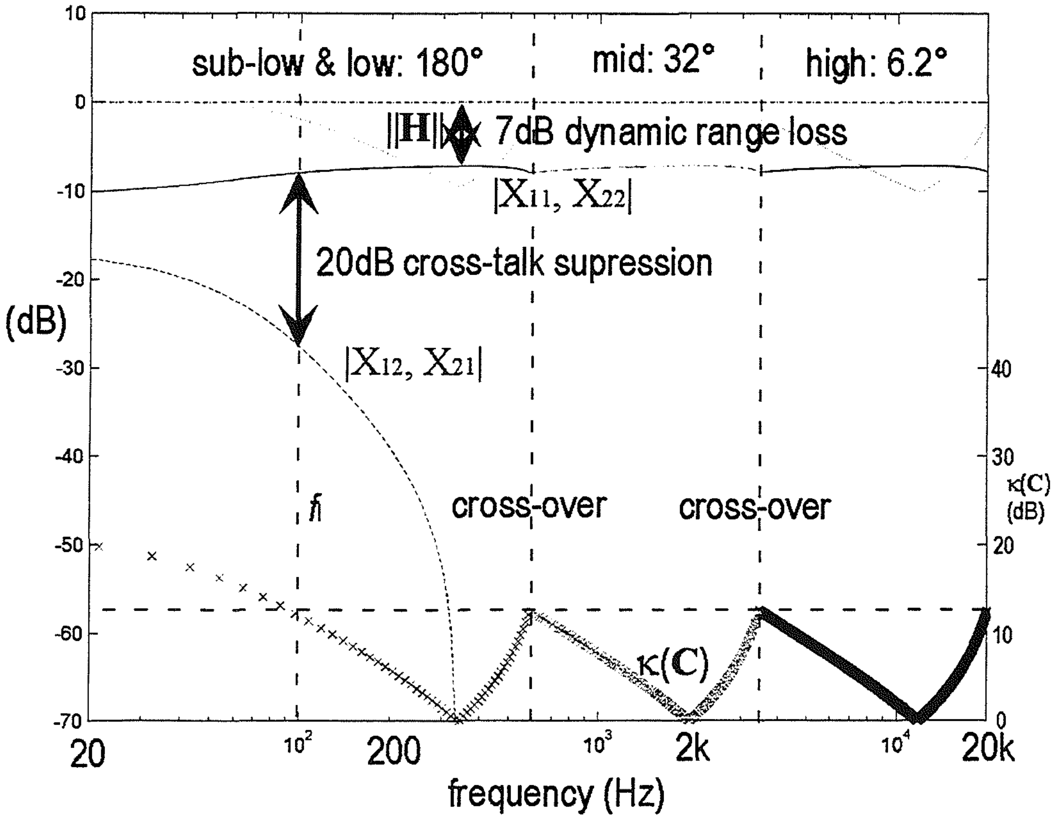


Fig. 7.26 Performance of a 3-way system with the lowest pair covering sub-low region.

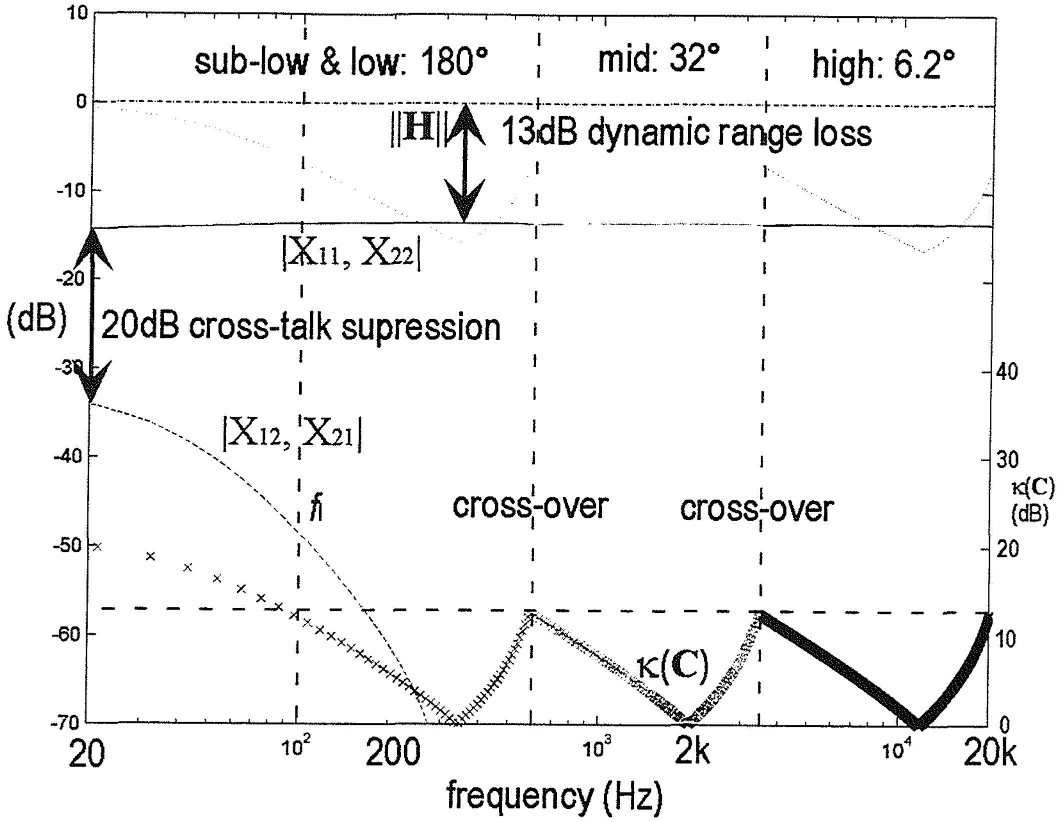


Fig. 7.27 Performance of a 3-way system with regularisation for 13dB dynamic range loss

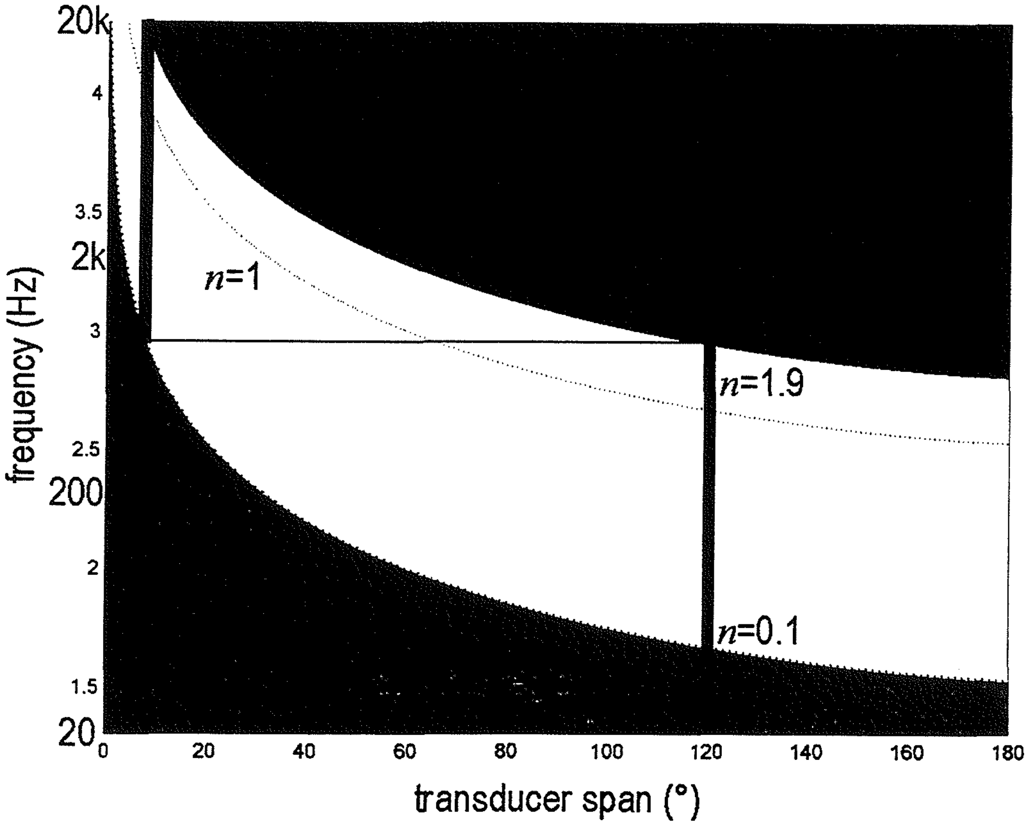


Fig. 7.28 Frequency/span region for systems with $n \gg 1$ and $n = 0.9$, and an example of discretisation for a 2-way system.

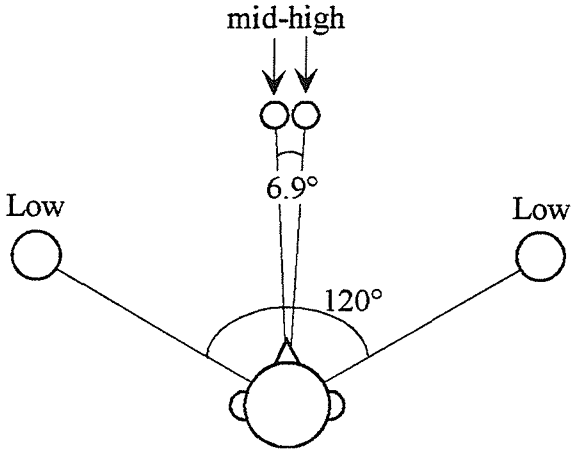


Fig. 7.29 An arrangement of a 2-way system with *n* » 1 and *n* = 0.9.

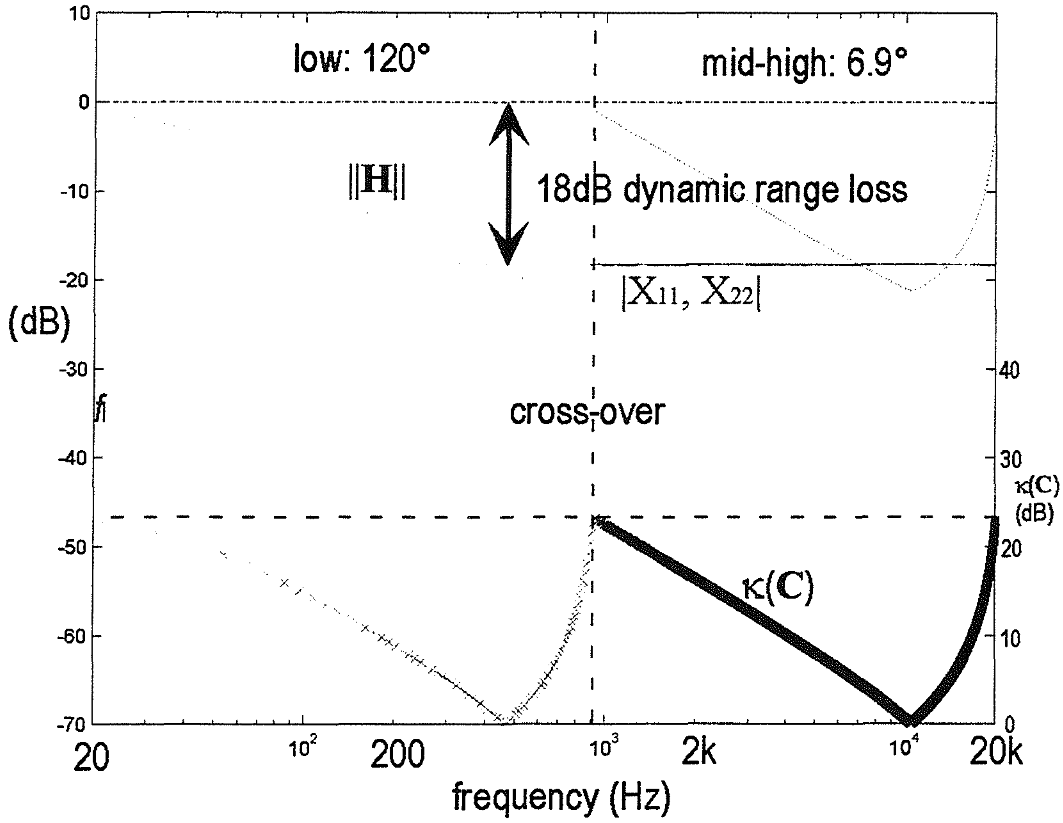
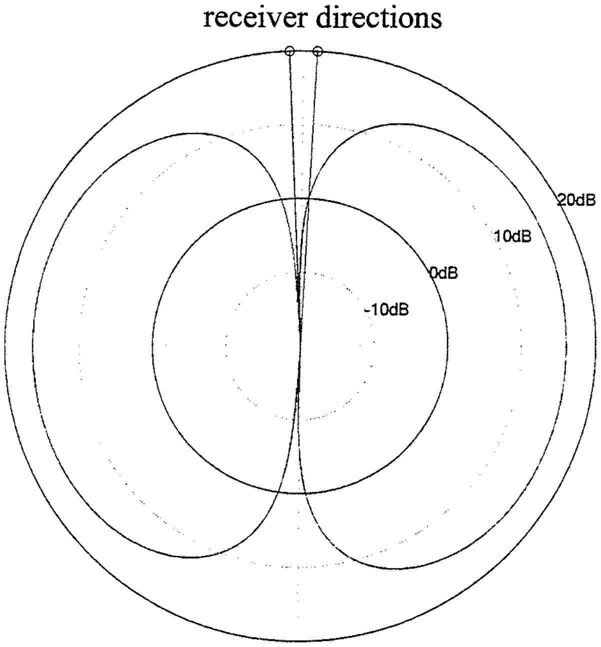
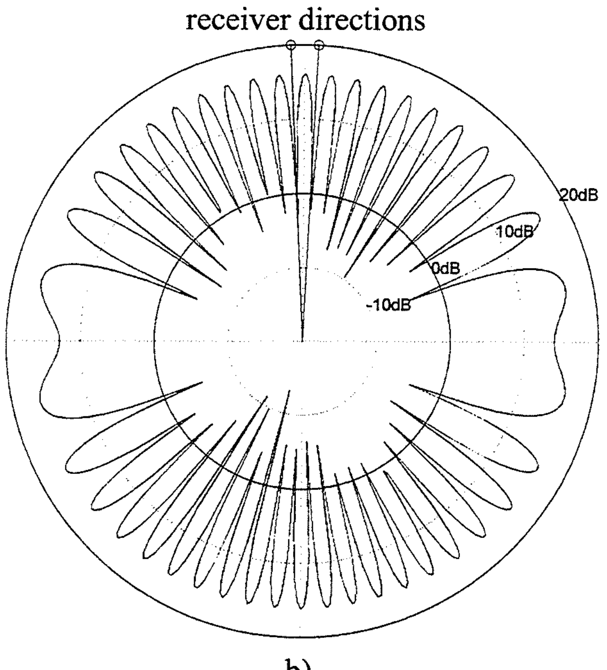


Fig. 7.30 Performance of a 2-way system with *n* » 1 and *n* = 0.9.



a)



b)

Fig. 7.31 Maximum far field sound pressure level produced by the transducer pairs with reference to the receiver directions (0dB and -¥dB). $n = 0.9$. a) $n = 0.1$ b) $n = 1.9$

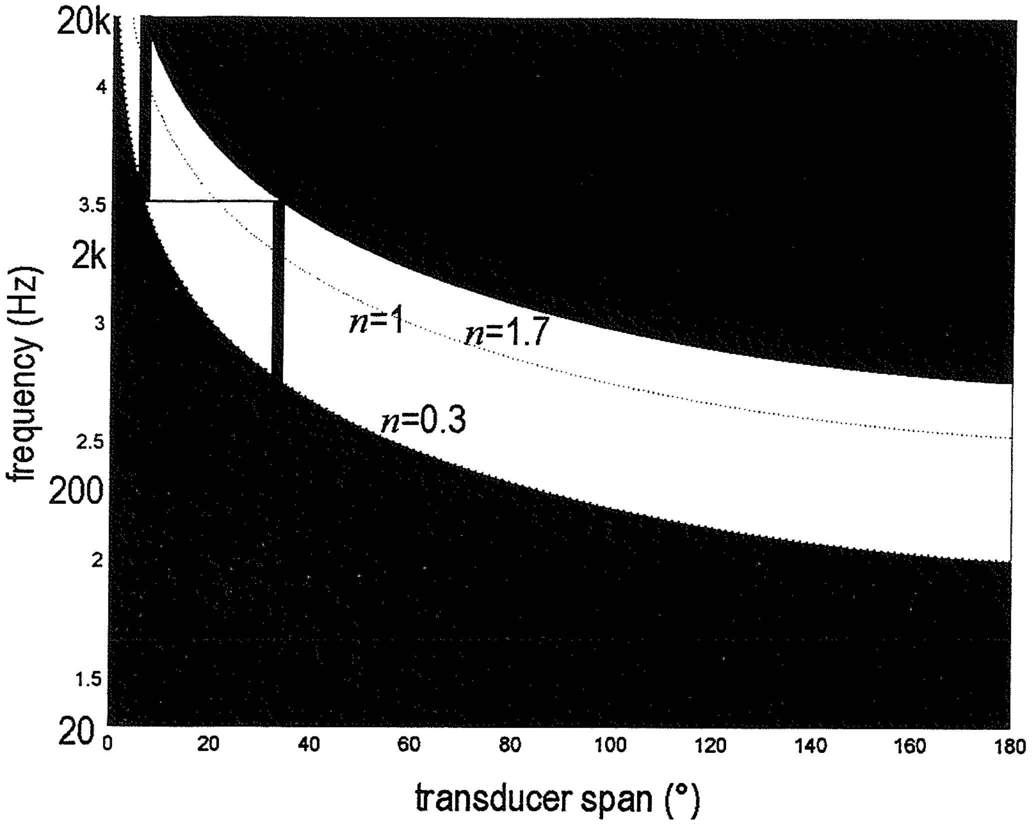


Fig. 7.32 An example of discretisation for 2-way system with $n \gg 1$ and $n = 0.7$ with regularisation for 18dB dynamic range loss.

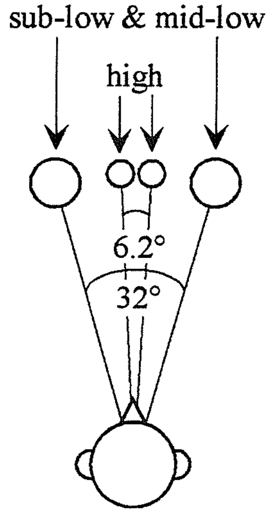


Fig. 7.33 An arrangement of a 2-way system with $n \gg 1$ and $n = 0.7$.

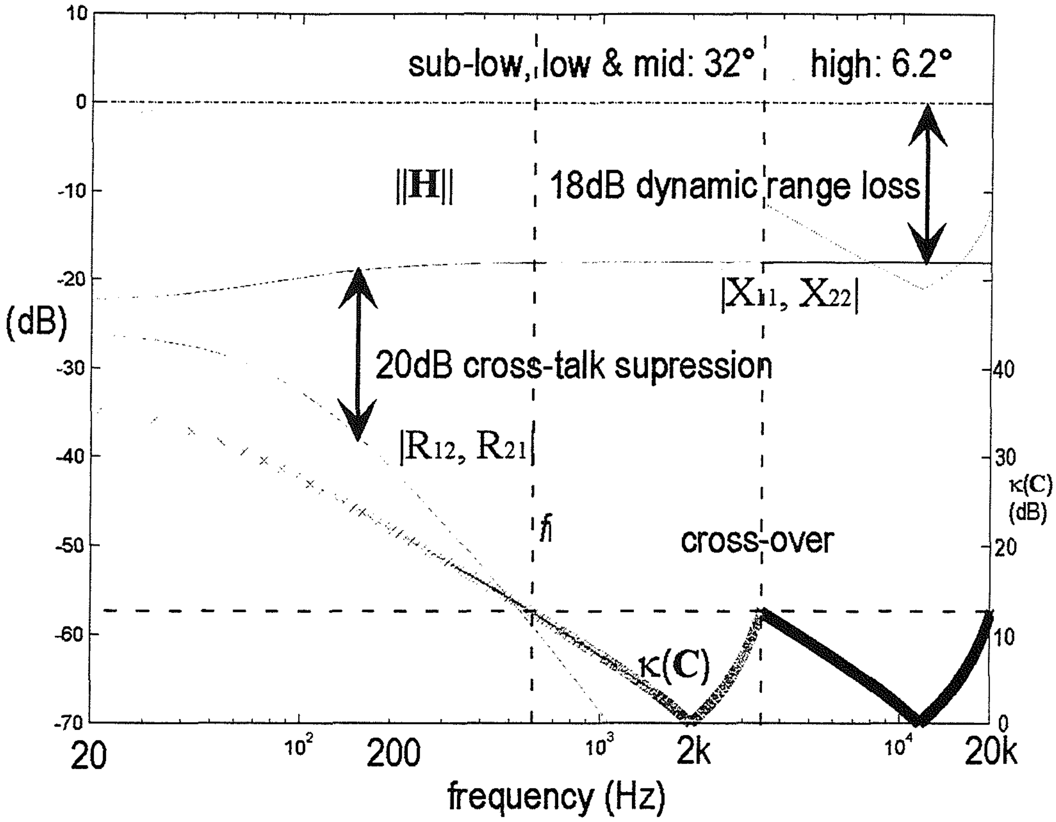


Fig. 7.34 Performance of a 2-way system with *n* » 1 and *n* = 0.7.

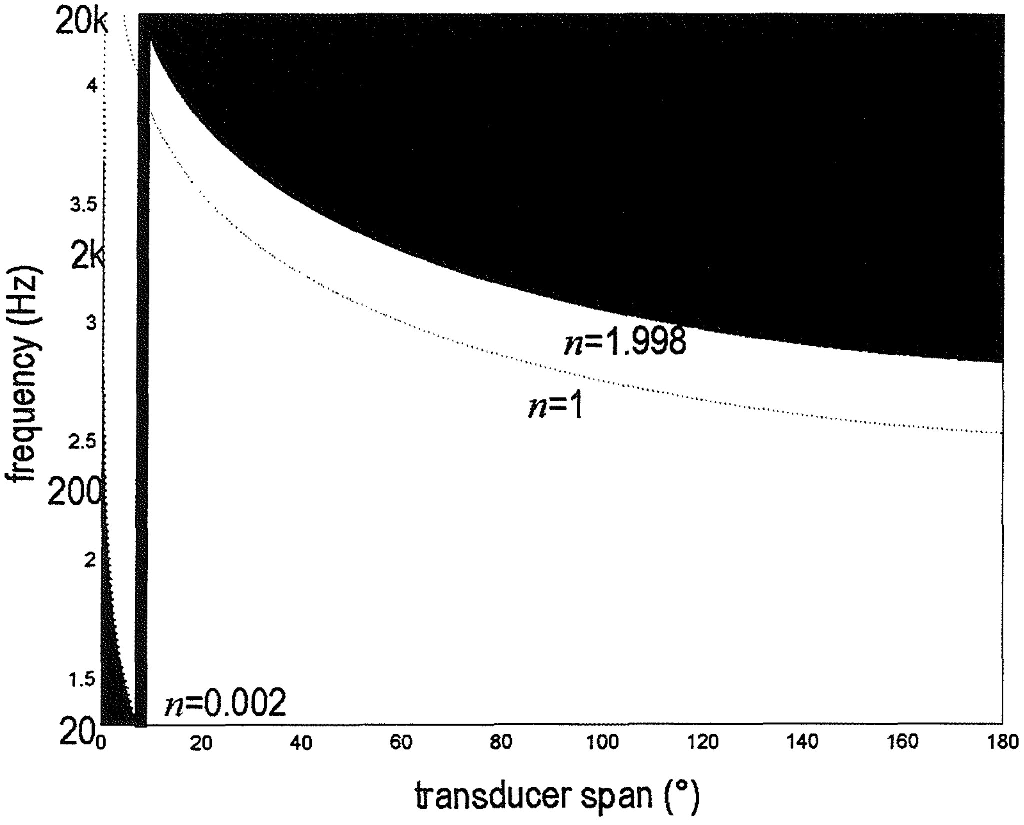


Fig. 7.35 Frequency/span region for systems with *n* » 1 and *n* = 0.998, and an example of discretisation for a 1-way system.

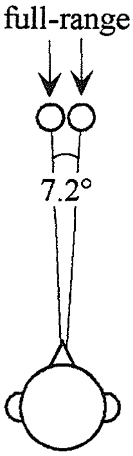


Fig. 7.36 An arrangement of a 1-way system with n » 1 and n = 0.998.

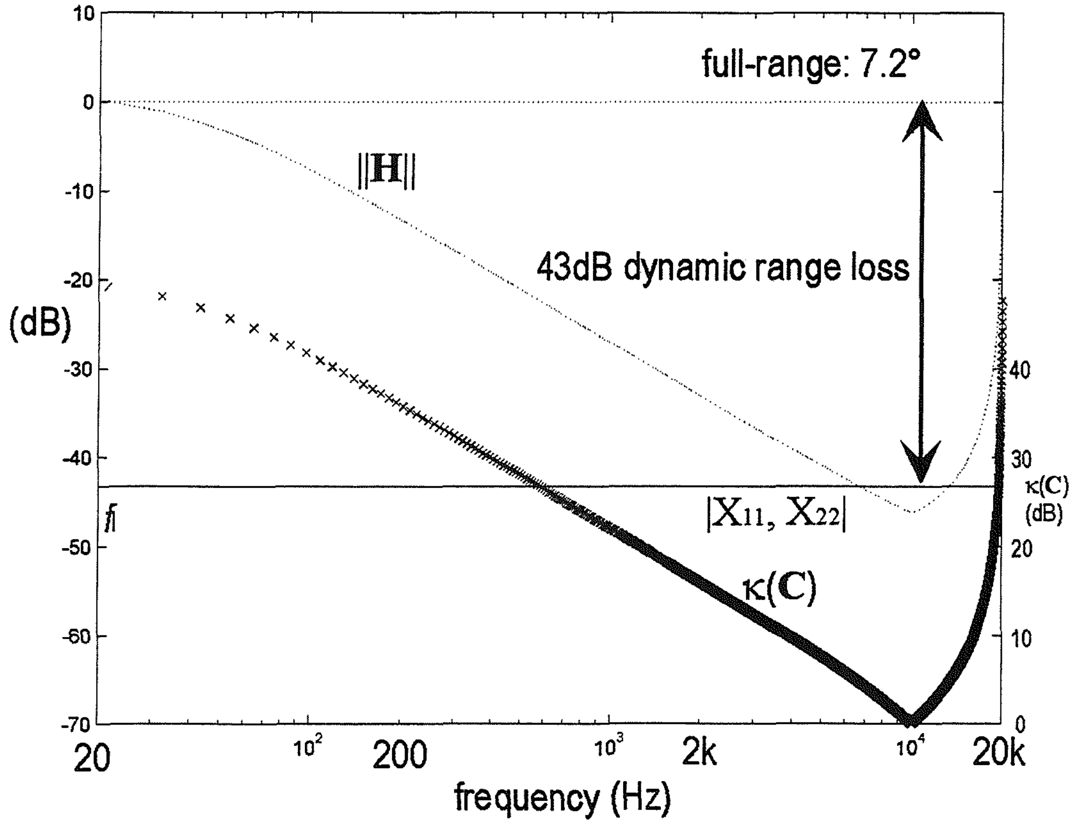


Fig. 7.37 Performance of a 1-way system with n » 1 and n = 0.998.

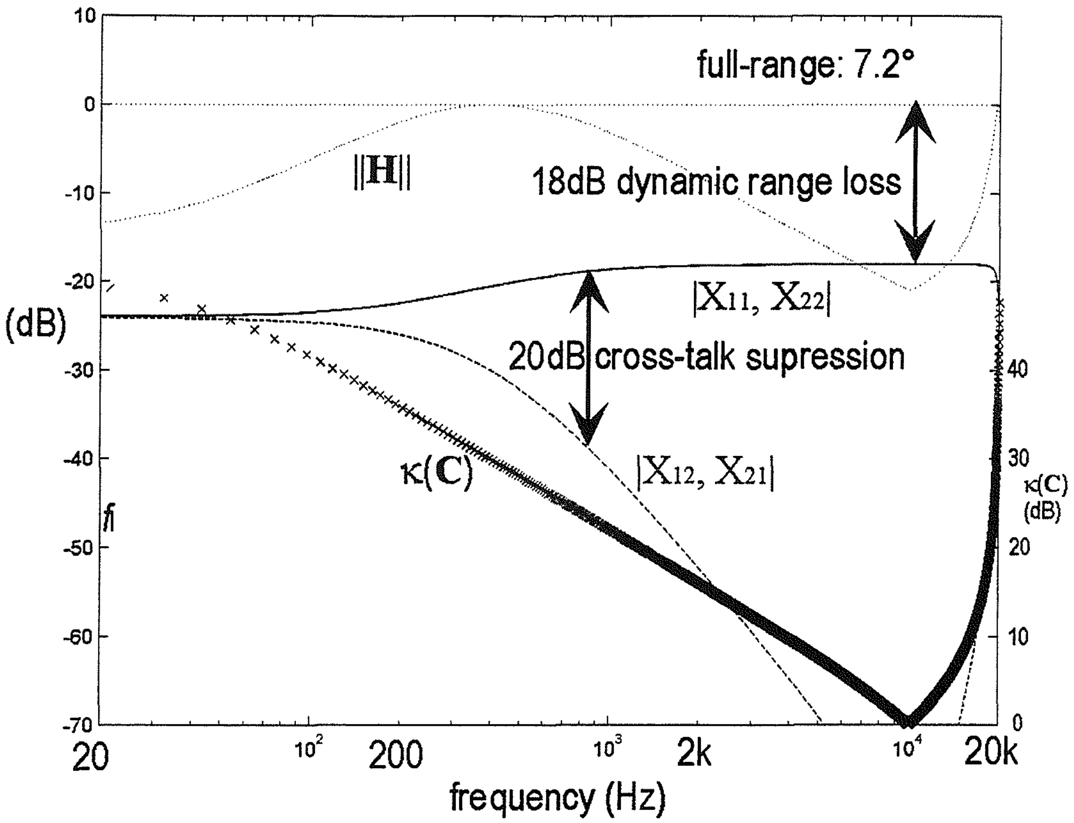


Fig. 7.38 Performance of a 1-way system with *n* » 1 and *n* = 0.998 with regularisation for 18dB dynamic range loss.

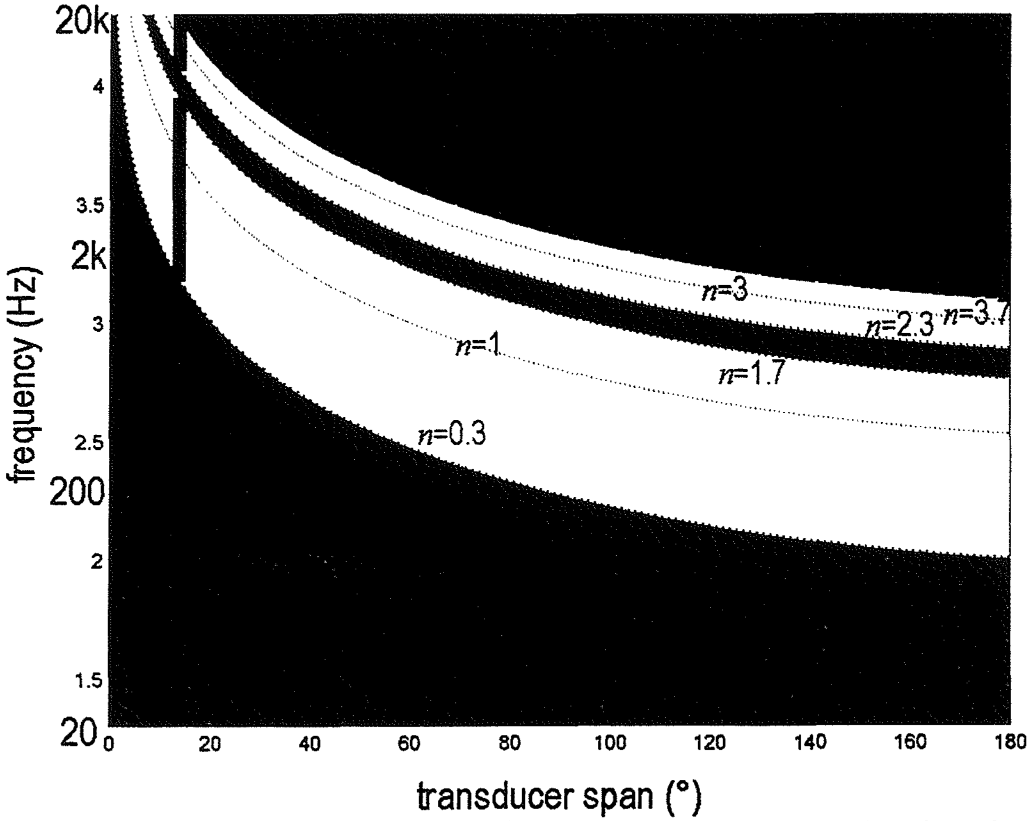


Fig. 7.39 Frequency/span region for a multi-region systems with n » 1 and n » 3 with *n* = 0.7, and an example of discretisation for a 1-way system.

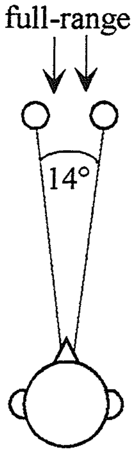


Fig. 7.40 An arrangement of a 1-way multi-region system with *n* » 1 and *n* » 3 with *n* = 0.7.

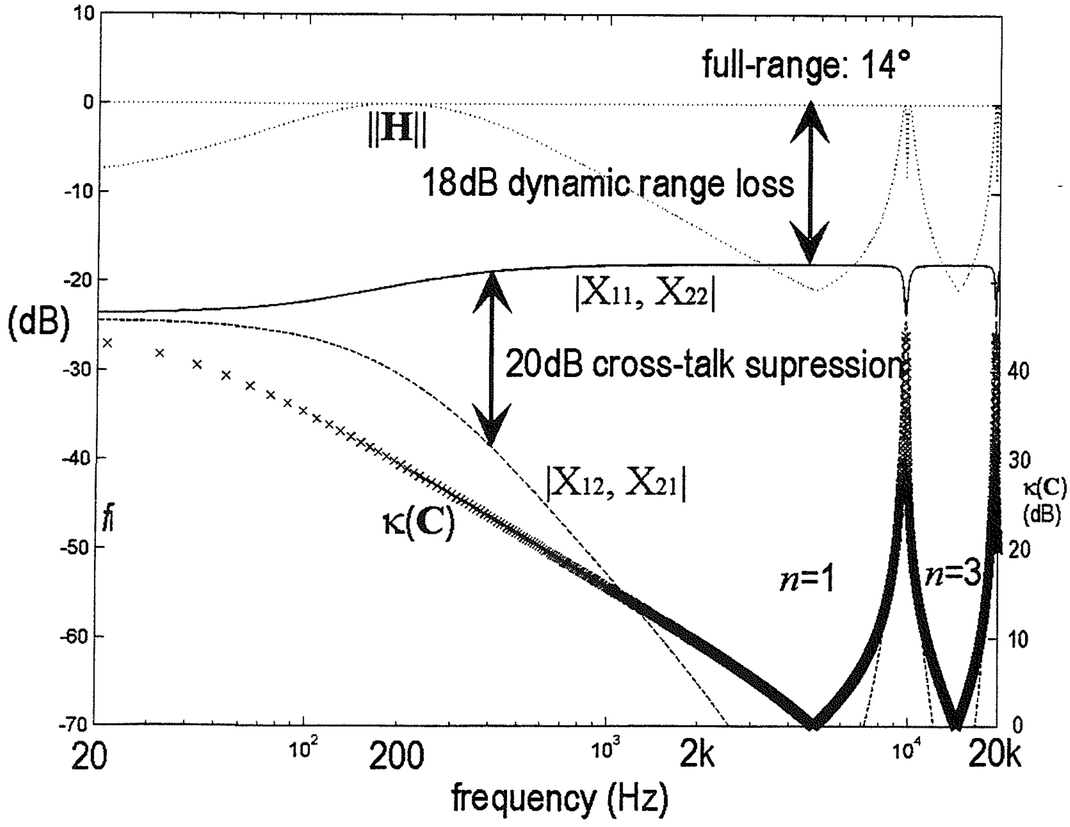


Fig. 7.41 Performance of a 1-way multi-region system with n » 1 and n » 3 with *n* = 0.7, with regularisation for 18dB dynamic range loss.

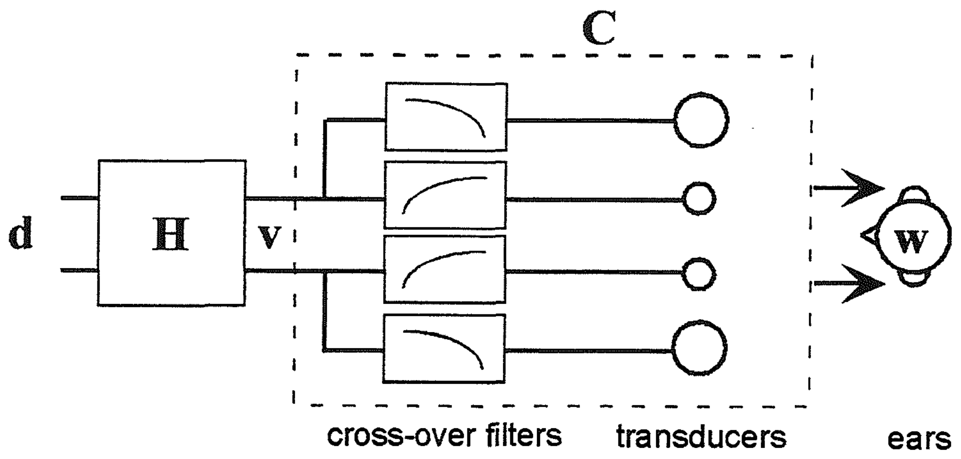


Fig. 7.42 Block diagrams for cross-over filters and inverse filters when a 2 by 2 plant matrix **C** is used to design inverse filters.

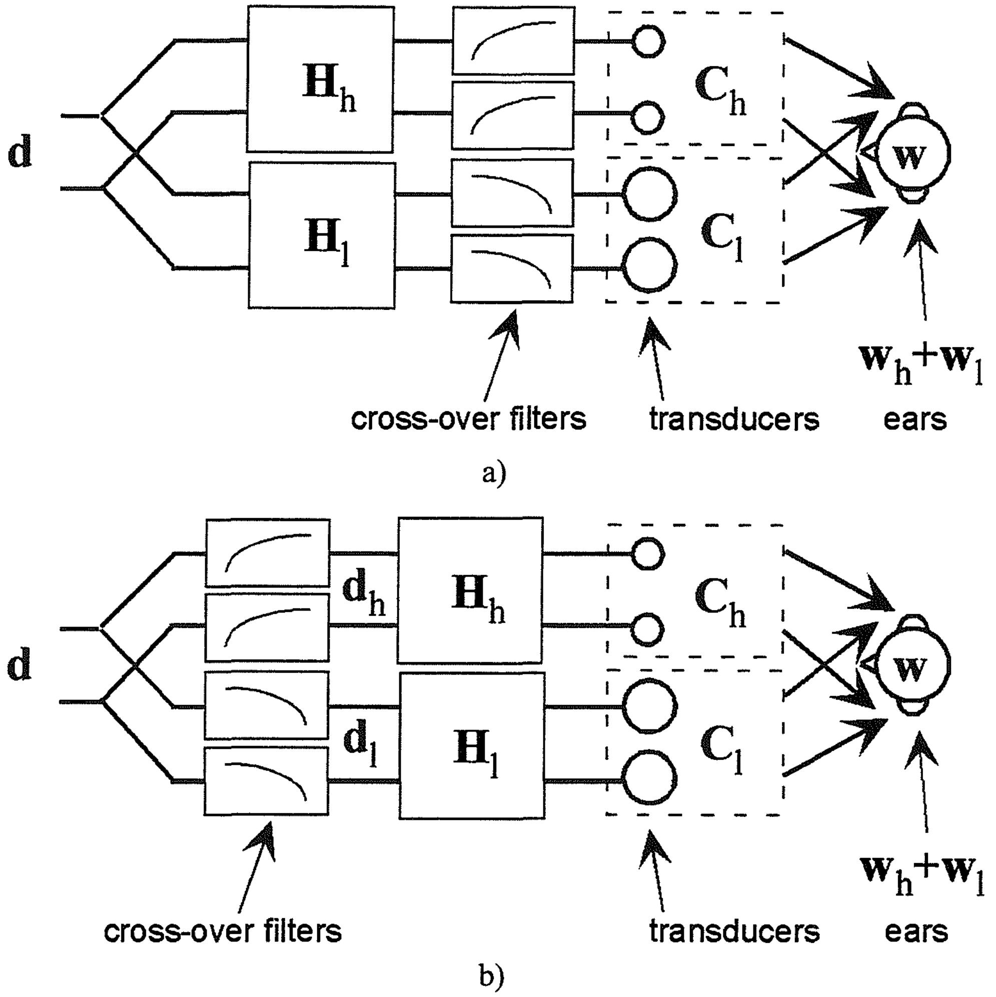


Fig. 7.43 Block diagrams for cross-over filters and inverse filters when *m* (number of driver pairs) of 2 by 2 plant matrices **C** are used separately to design *m* inverse filter matrices. a) Cross-over filters after inverse filters. b) Cross-over filters prior to inverse filters.

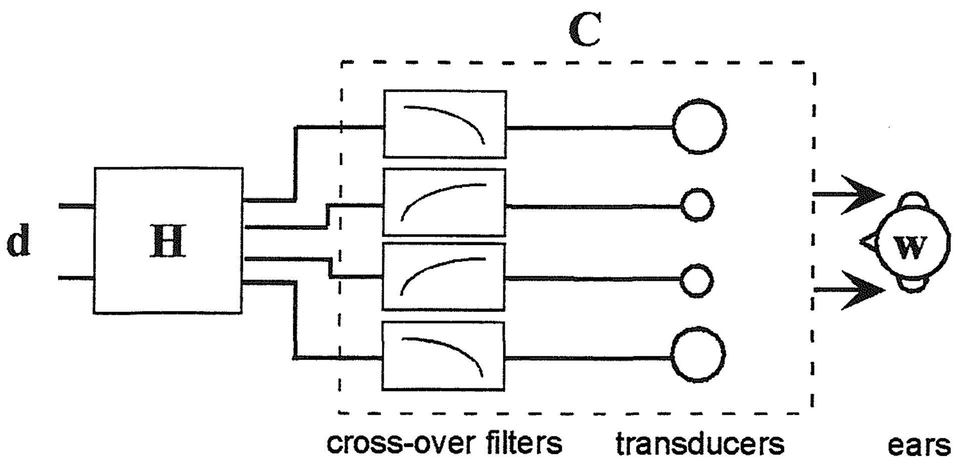


Fig. 7.44 Block diagrams for cross-over filters and inverse filters when a 2 by (2 ′ *m*) plant matrix **C** is used to design inverse filters.

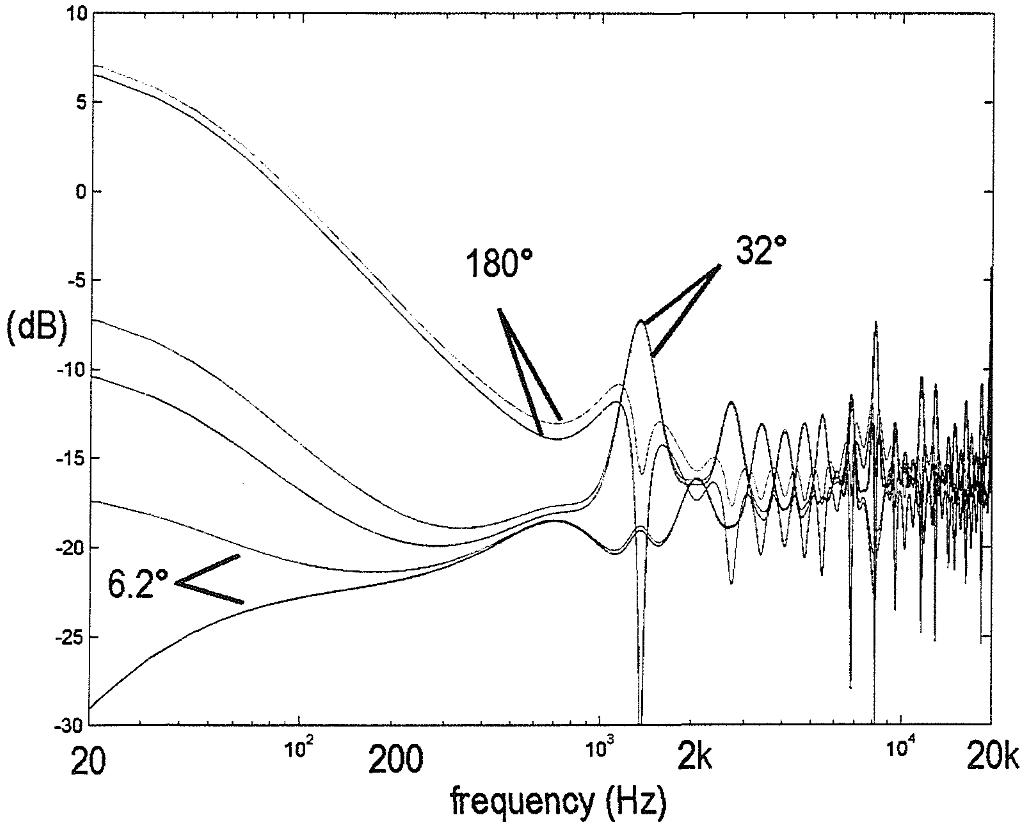


Fig. 7.45 An example of inverse filter responses for a multi-channel system (6 channels).

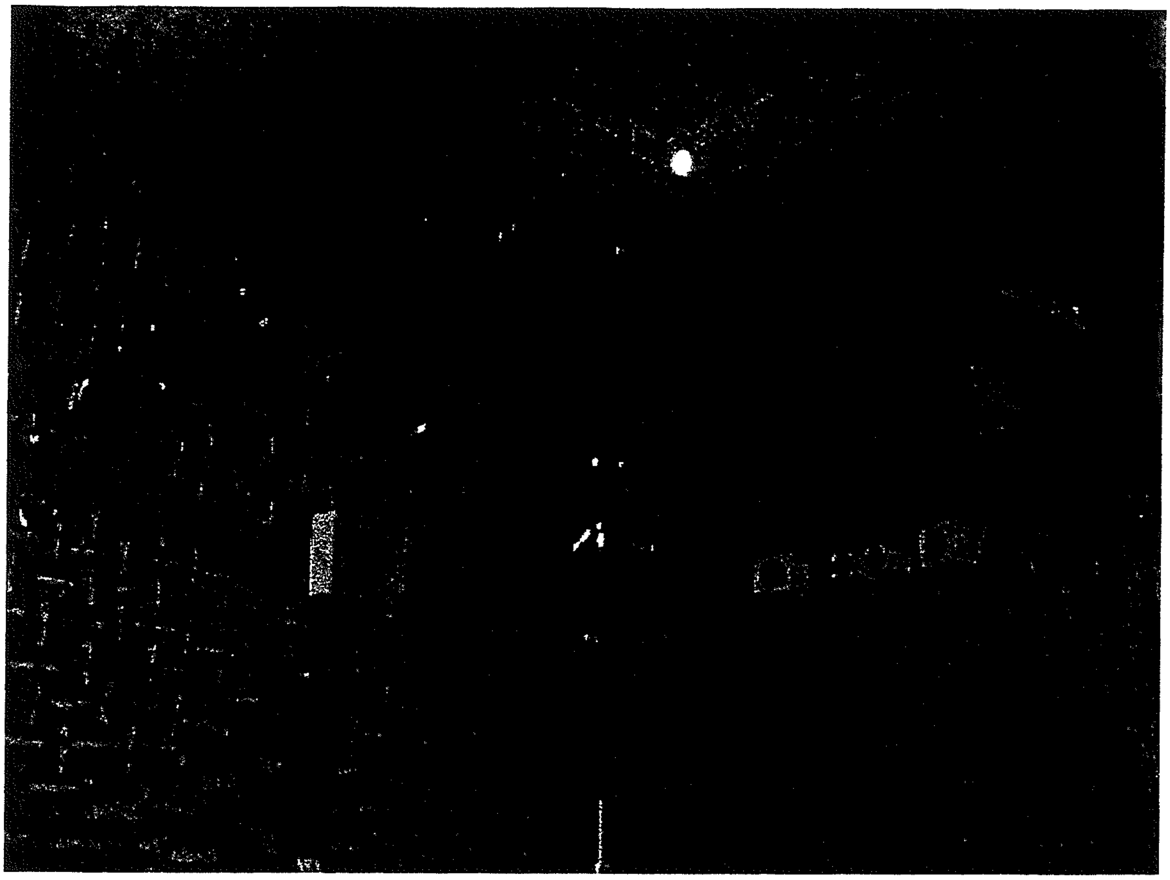


Fig. 7.46 Experimental rig for objective measurement.

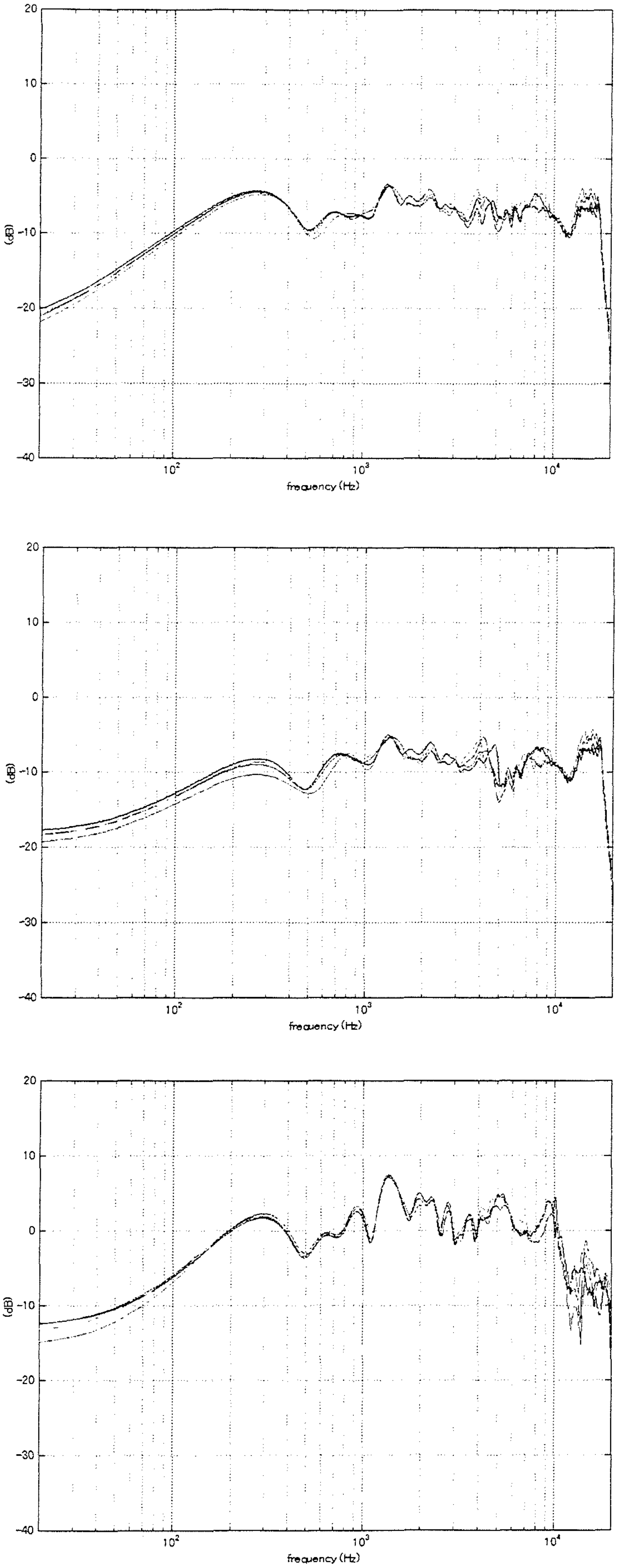


Fig. 7.47 Loudspeaker response. Top) 3-way OSD. Middle) 2-way OSD. Bottom) SD.

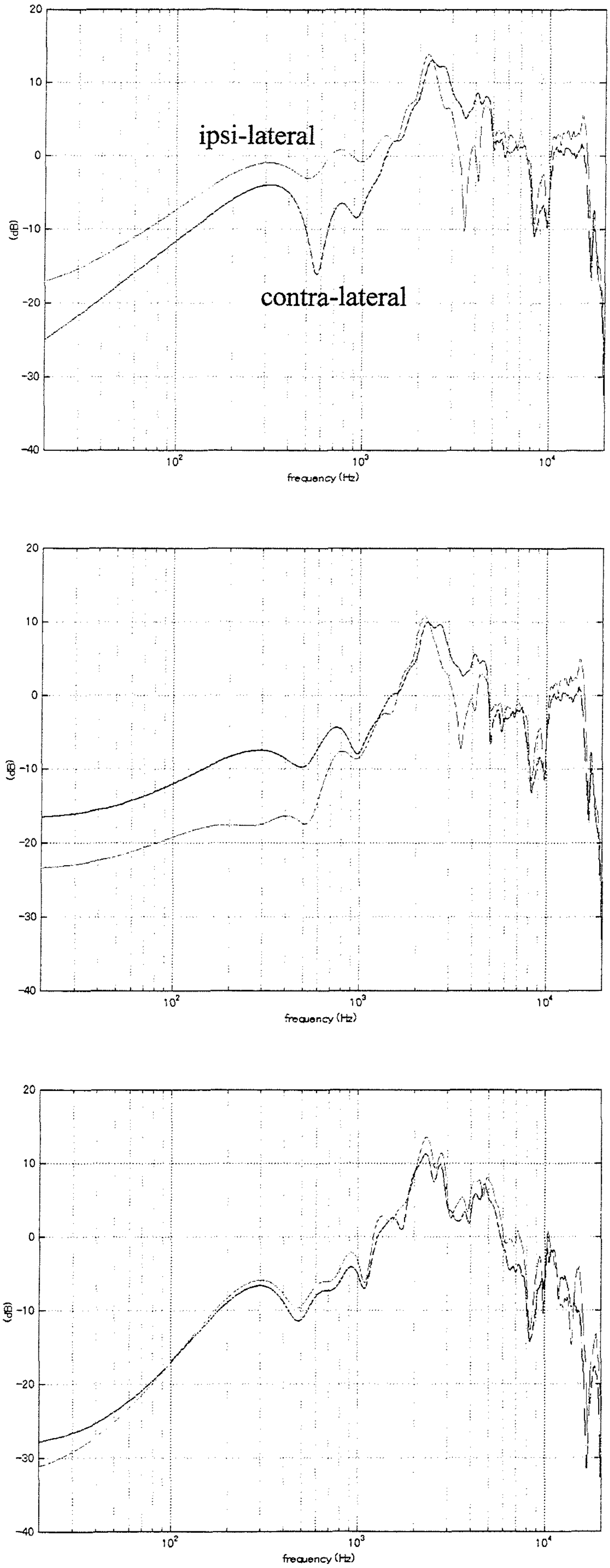


Fig. 7.48 Plant response. Top) 3-way OSD. Middle) 2-way OSD. Bottom) SD.

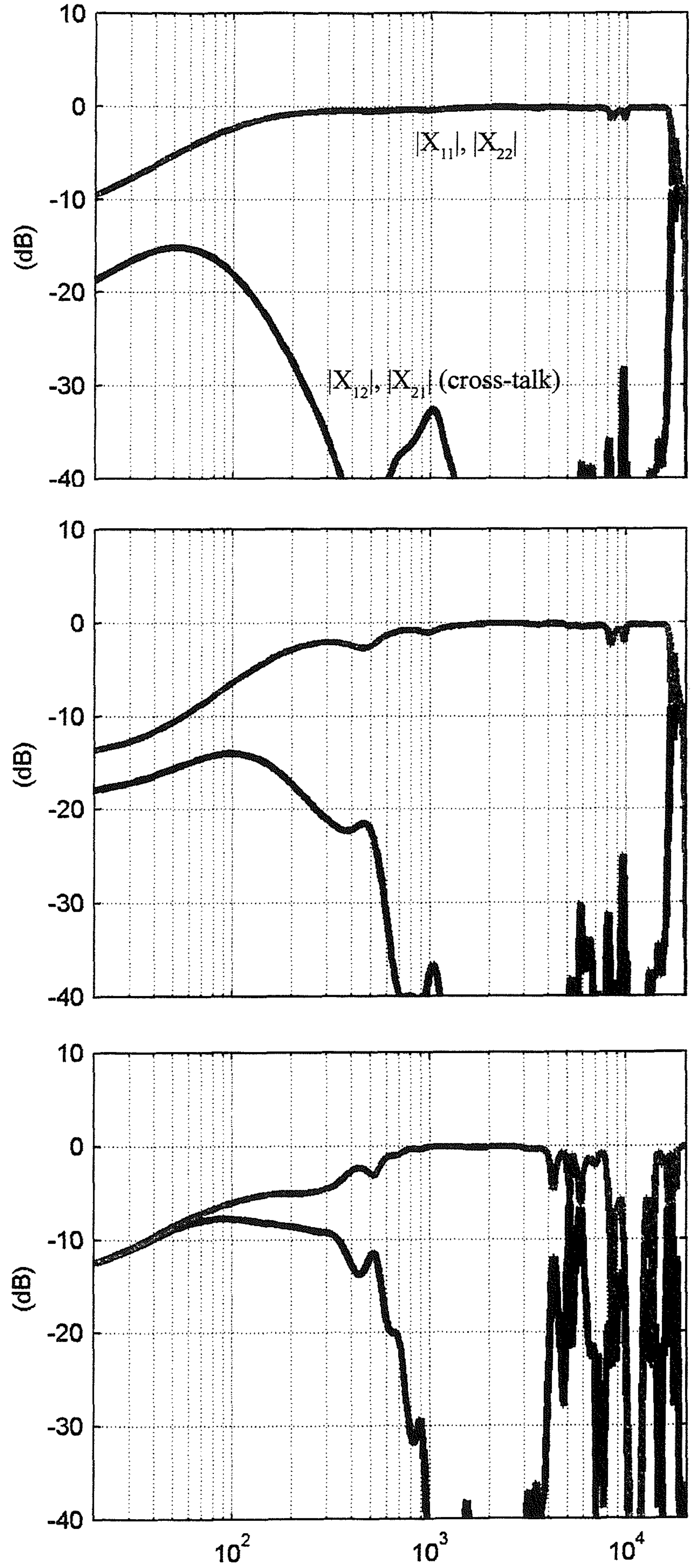


Fig. 7.49 Control performance. Top) 3-way OSD. Middle) 2-way OSD. Bottom) SD.

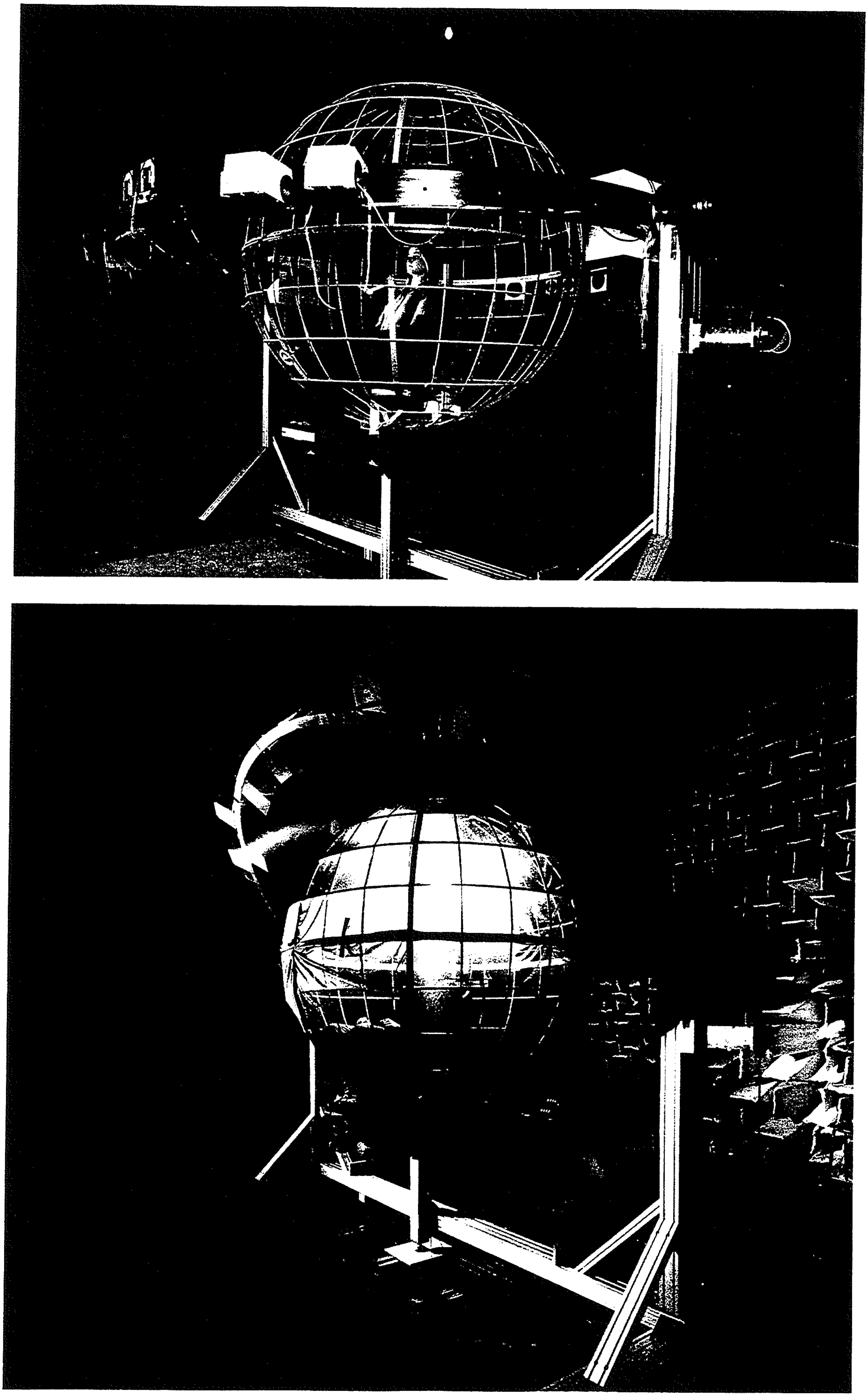
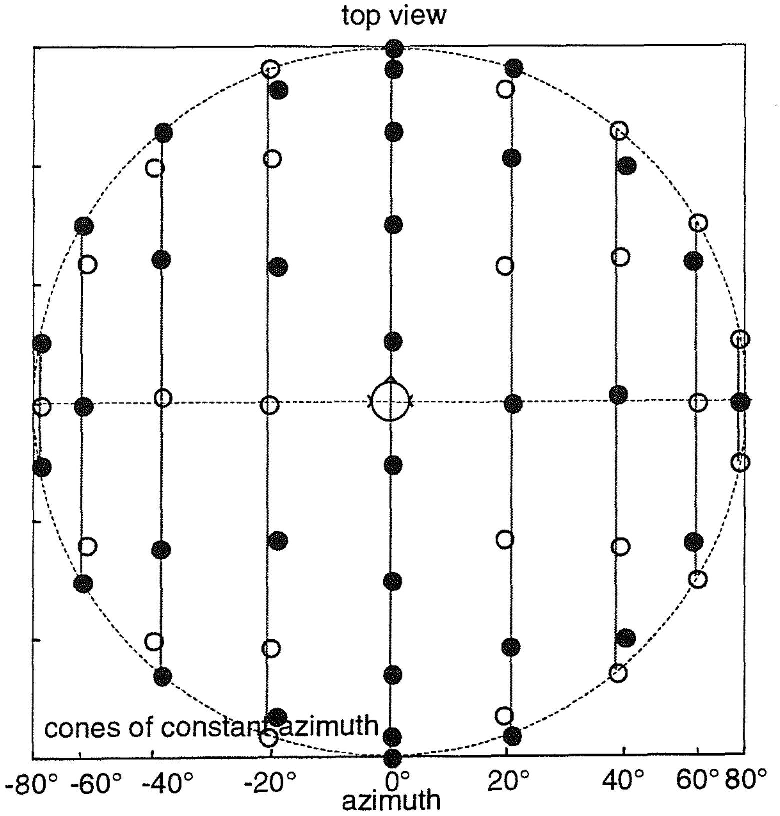


Fig. 7.50 Experimental rig for subjective evaluation.



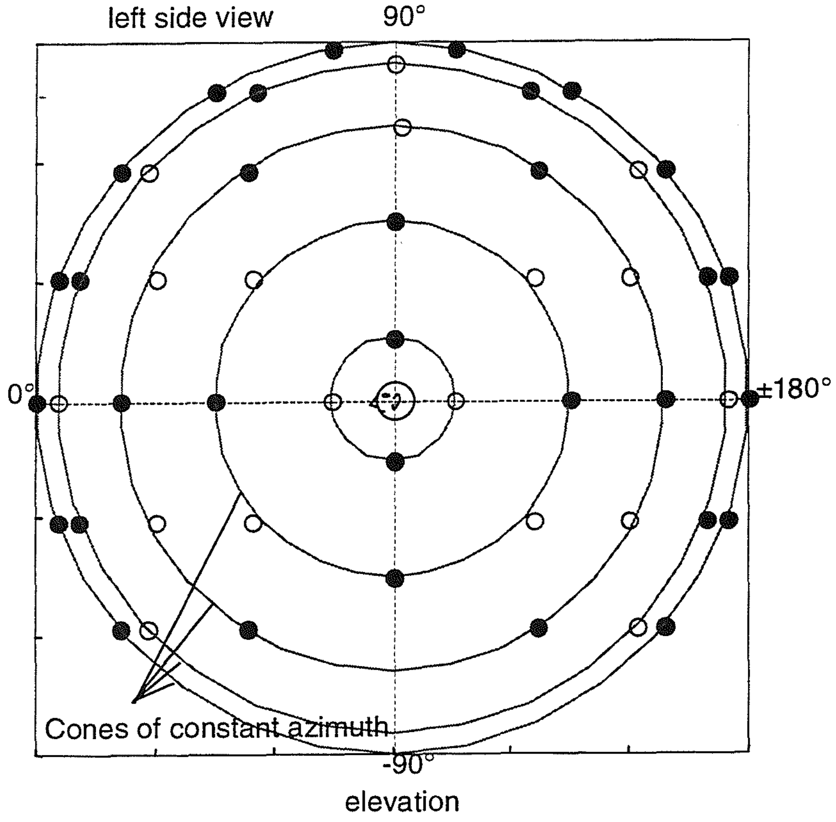


Fig. 7.51 Tested directions Top) top view. Bottom) side view.

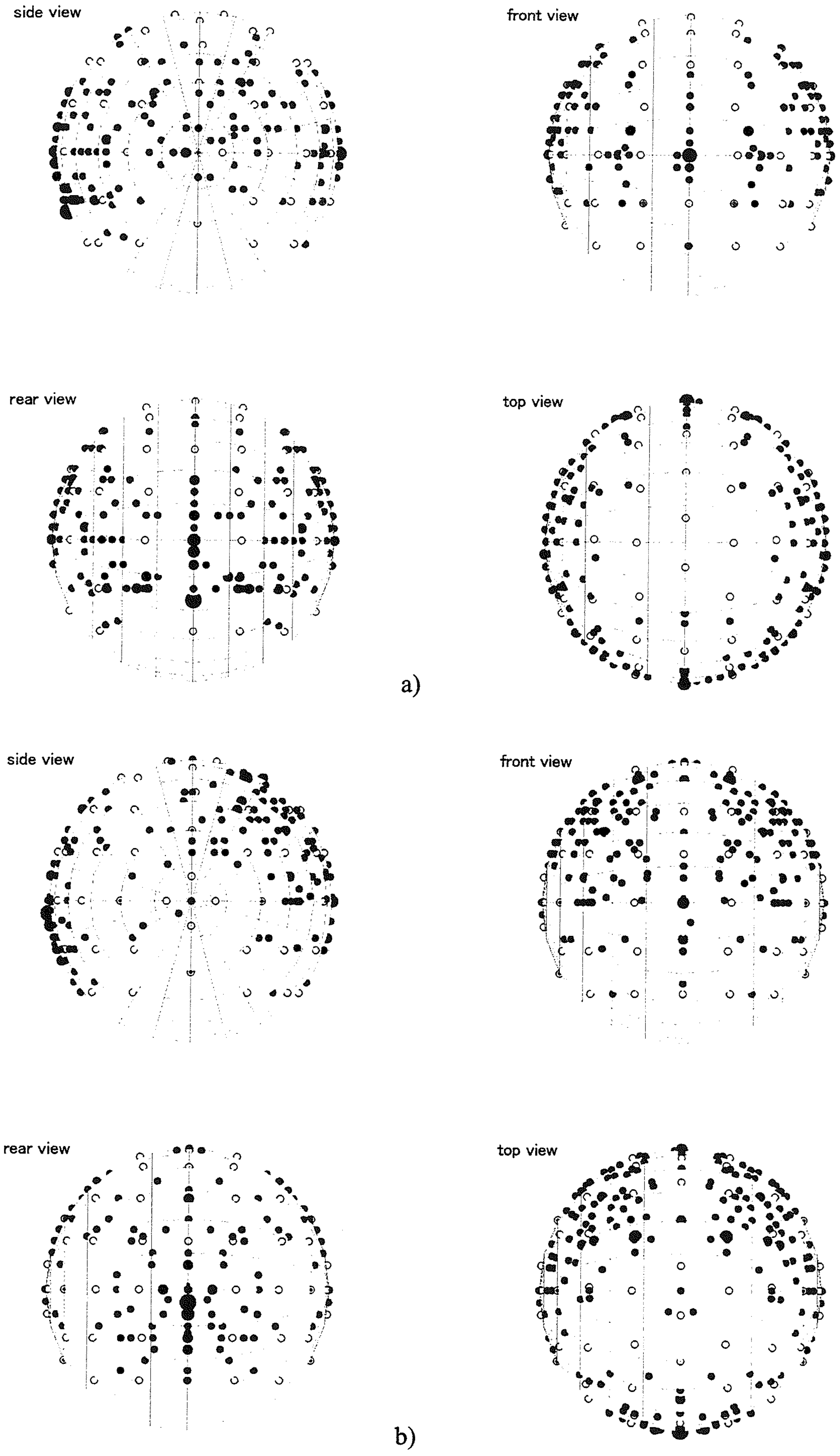
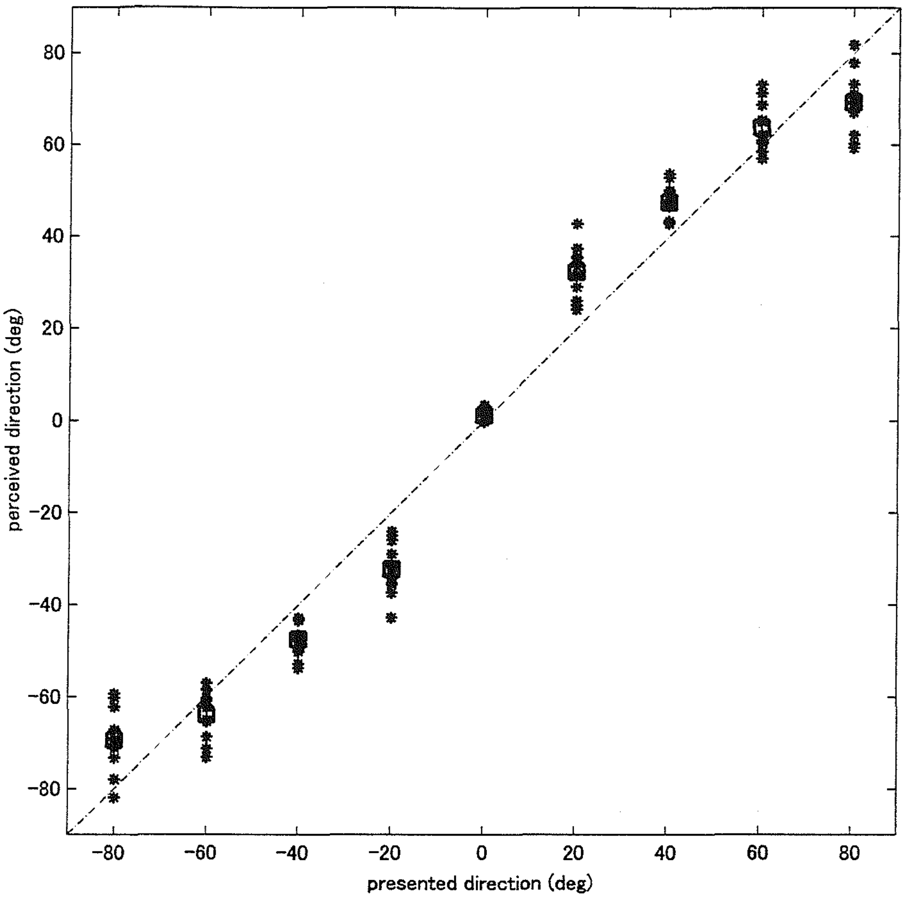


Fig. 7.52 Perceived virtual sound source directions. a) OSD. b) SD.



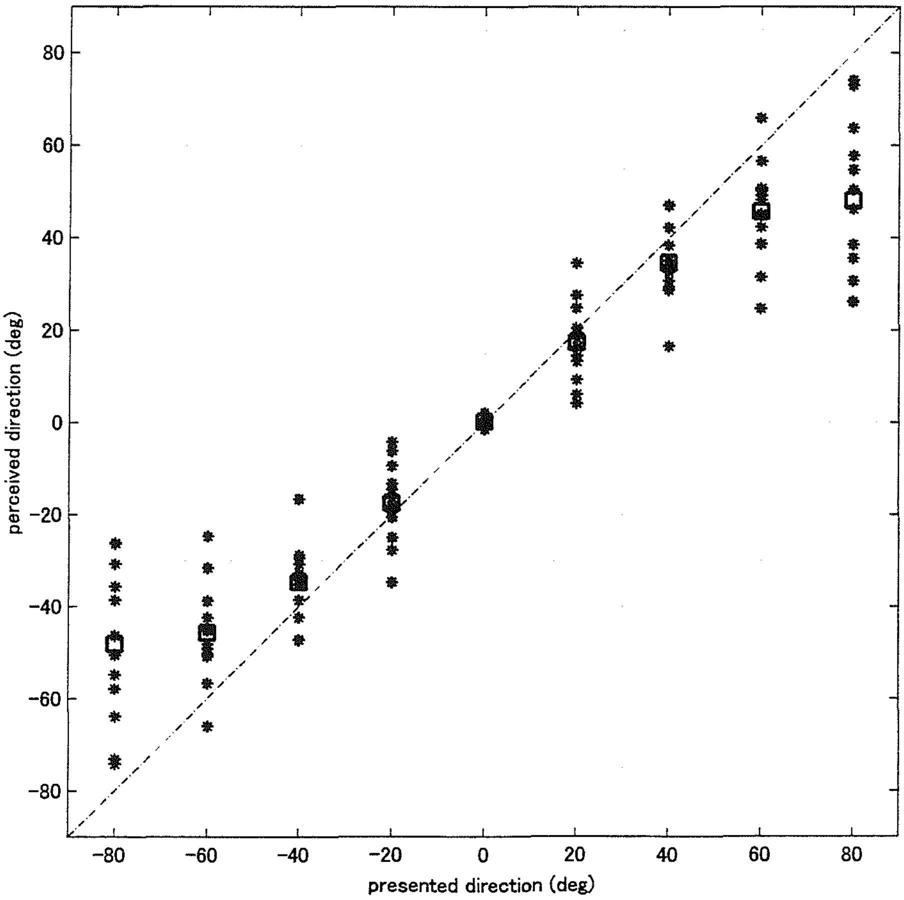


Fig. 7.53 Azimuth localisation performance. Top) OSD. Bottom) SD.

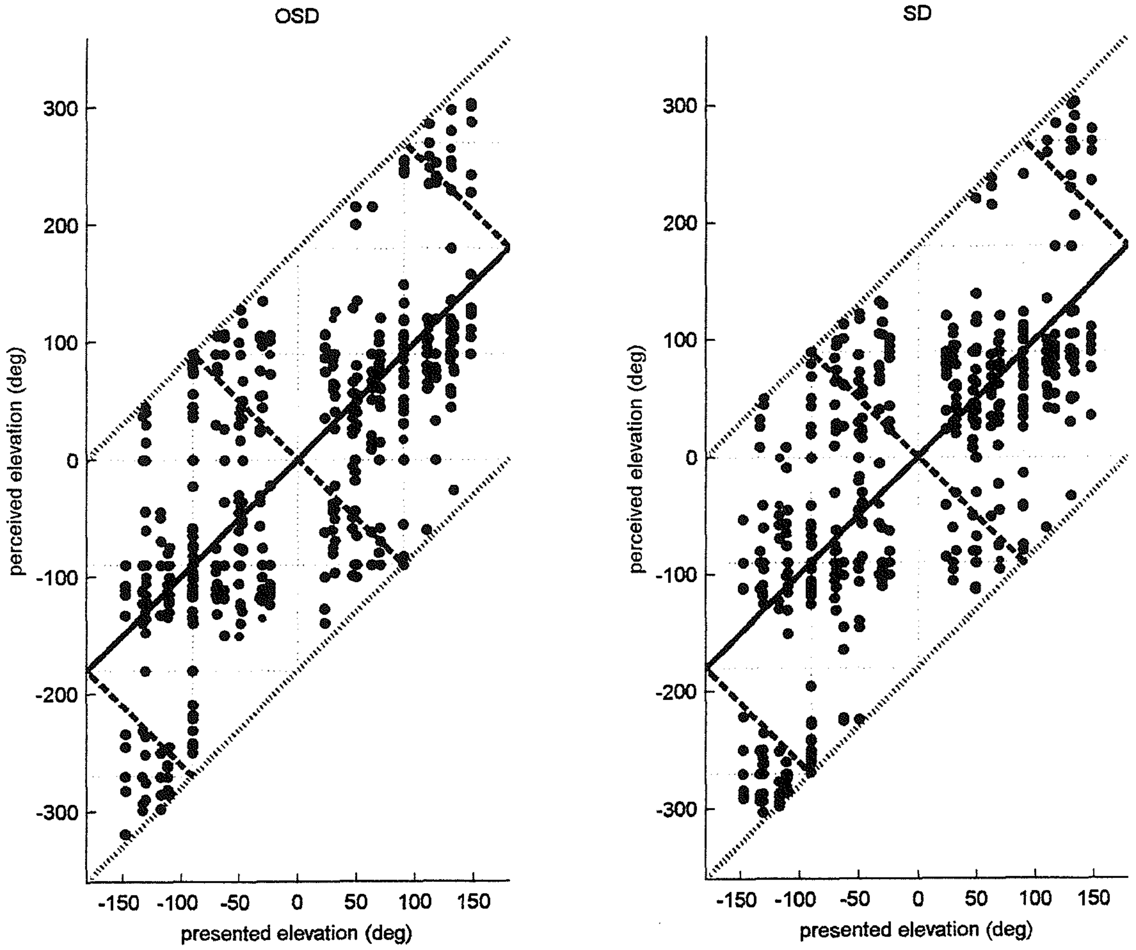
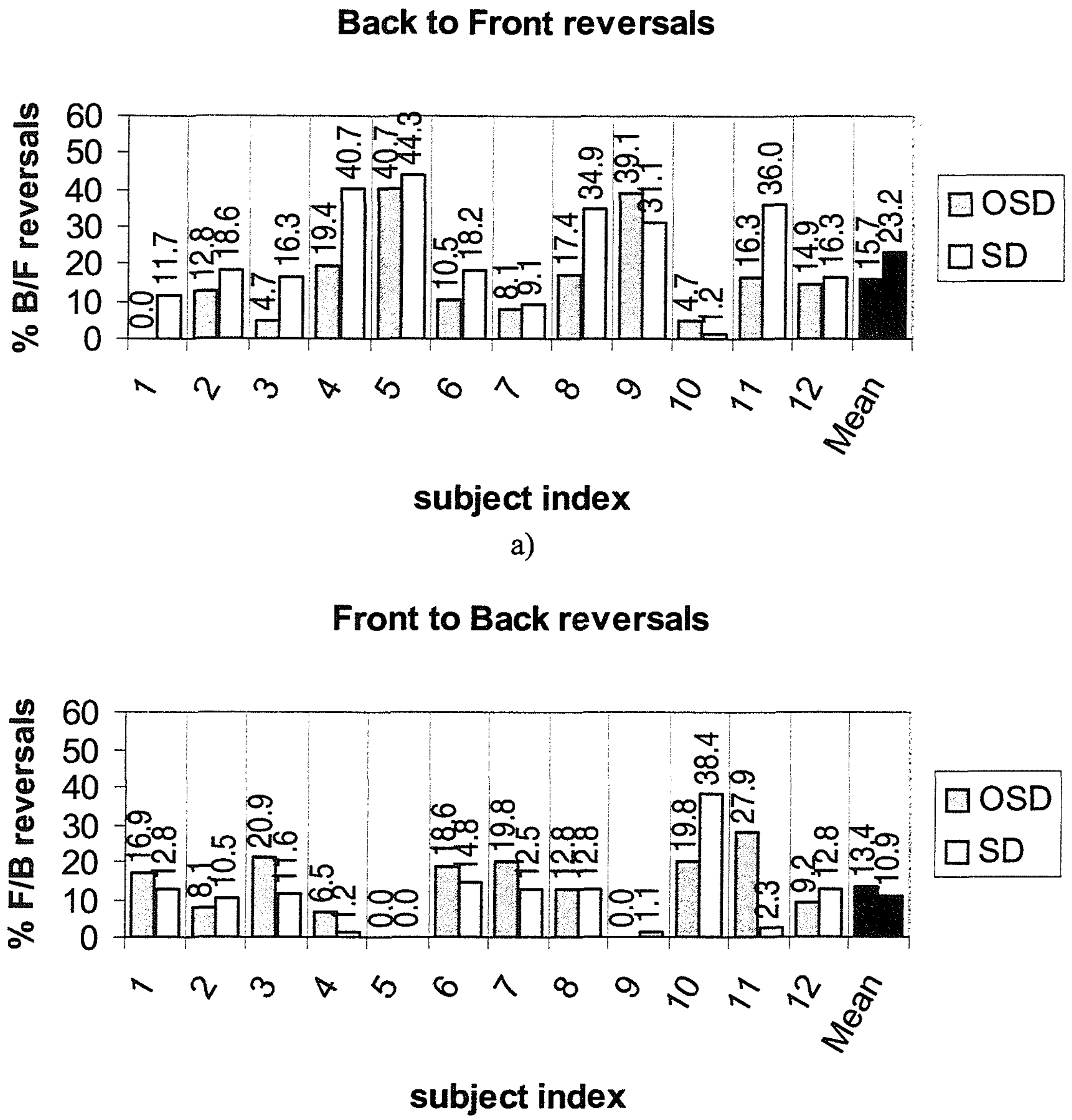


Fig. 7.54 Elevation localisation performance. Left) OSD. Right) SD.



a)

b)

Fig. 7.55 Rate of back to front reversals and front to back reversals for each individual subject and overall mean. a) back to front reversals. b) front to back reversals.

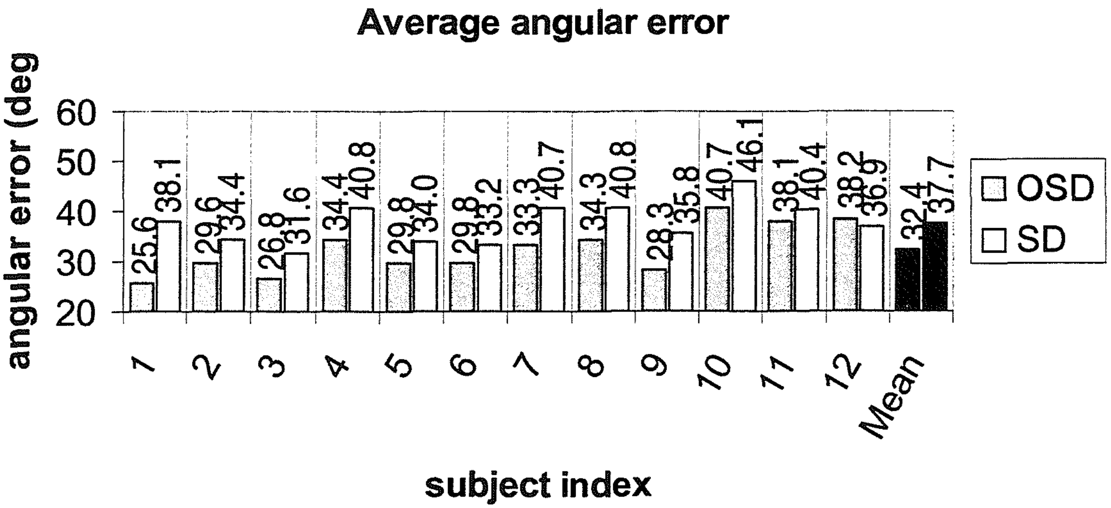


Fig. 7.56 Average angular error for each subject and overall mean.

8 Elevated control transducers

8.1 Introduction

The development of both the "Stereo Dipole" and the "Optimal Source Distribution" virtual acoustic imaging systems dealt with the azimuth location of the control transducers. In the past, the elevation location of transducers for binaural reproduction over loudspeakers has received even less attention than the azimuth location. In most of the past research, the transducers are usually placed on the horizontal plane that includes the listener's head. This convention is probably adapted from conventional stereophony for which the virtual images are perceived at the same elevation as the transducers. Since the majority of the sound sources in everyday life are on the horizontal plane, as a consequence of the fact that most objects are on the ground, placing transducers on the horizontal plane was a natural choice. However, since the binaural technique enables in principle the synthesis of sound waves from any direction, there is no reason to restrict the transducer position to the horizontal plane.

The objective of this Chapter is to point out that binaural synthesis over loudspeakers can also be made to operate remarkably effectively when the control transducers are not in the horizontal plane in front of the listener. It has been shown that the most significant error in binaural reproduction is front-back confusion. In cases of loudspeaker synthesis, this often results in bias error where a rear image is perceived in front, i.e., towards the control transducer direction (Chapter 4). When the transducers are placed around the frontal plane, this bias error is expected to be unlikely, being at the border of the front and rear hemisphere.

In order to find out the characteristics of various elevation positions of the control transducers, an analysis of the spectral cues and dynamic cues has been performed.

Positions in the frontal plane above the listener's head are found to be promising as an alternative control transducer location. A subjective experiment is performed in order to compare between two alternative control transducer locations, 0° elevation and 90° elevation.

8.2 Inversion of the plant

When the plant is inverted, peaks and dips in the plant transfer functions are suppressed or filled by the inverse filters in order to achieve the synthesis of the desired signal spectra. Therefore, a certain amount of dynamic range is lost through this process, i.e. through the compensation of the plant response. In this respect, a flat plant response is preferable to that with significant peaks and dips. Furthermore, it has also been revealed that the mismatch between the individual plant HRTFs and the design plant HRTFs often results in the synthesis of the wrong spectra (Chapter 4). The mismatch is most likely to happen where notches exist whose position can vary considerably among individuals and are hence less likely to be cancelled out properly. There may be some elevation directions where the inversion of the plant is easier than in the other directions. Therefore, the plant responses for the "Optimal Source Distribution" and the "Stereo Dipole" system were measured in order to study this possibility.

The experimental procedure is fundamentally the same as that described in Section 7.8.1, except that the circular steel frame on which the control transducers were mounted was rotated around the interaural axis at 1° increments from –180° elevation to 180° elevation in order to obtain the plants for various elevation directions (Fig. 8.1). There are gaps of 10° in the regular sampling in the directions centred on -85° and 95° due to the required size and shape of the transducers and the ring. The data obtained with DB-065 was used for the evaluation although another set of data was obtained with DB-061. Fig. 8.2 shows the frequency response of the plant HRTFs of the OSD system along the different elevations. There are several distinct dips seen in the frequency response above 5kHz. The frequencies giving the dips goes up as the elevation of the control transducers becomes larger (in the "up" direction) in the front hemisphere, then goes down again as the elevation continuously becomes larger (in the "down" direction). The frequencies associated with these dips are roughly symmetric with respect to the frontal plane, and hence are likely to be a source of the front and back reversals. On the other hand, they are distinctively different in the vertical direction. Therefore, up-down reversal is expected to be much less likely to happen than the front-back reversal from the viewpoint of the similarity of the spectral shapes.

In general, the response is stronger in the front half than in the rear half. The response at the rear bottom quarter has numerous dips and generally is weaker, and therefore, this region seems to be less useful as a control transducer location. On the other hand, the region around 90° elevation (between 60° and 120° elevations) draws attention since the plant has a relatively flat smooth response without any prominent dips. This characteristic of the plant response is an additional and physically supported benefit to just being on the border of the frontal and rear hemispheres. A drawback of the overhead position is that high frequency response above 12kHz is weaker than that for the directions towards the front.

The frequency response of the plant of the SD system along different elevations is shown in Fig. 8.3. The general tendency is the same as the OSD system, except that the control transducers for the SD system have less response above 12kHz. In fact, the elevation dependency of the spectrum shape is relatively steady regardless the azimuth direction.

This can be seen in Fig. 5.18 and Fig. 5.19 showing the response on ±50° azimuth direction as well as the response along the directions on the median plane (Fig. 8.4). The most noticeable azimuth dependency is that the slope formed by the dip in frequency response as the elevation changes becomes shallower as the sound source moves away from the median plane.

The condition numbers for the plant matrix of the OSD and SD system are shown in Fig. 8.5. Fig. 8.5a suggests that the frontal hemisphere is the better location for the control transducers although there may be a consequence that the discretisation of the ideal OSD was optimised for the frontal hemisphere. Fig. 8.5b suggests a similar result although the picture is smeared by the non-controlled region inherent in the SD system around the 10kHz to 12kHz. It is worth noting that this ill-conditioned frequency coincides with the characteristic dips with elevation dependency at around 0° and ±180° elevations (horizontal plane). This may explain the stronger tendency by the SD system of bias error towards the horizontal plane.

8.3 Dynamic cues

It is also known that when the listener has ambiguity in judging whether the sound is from front or from the rear with spectral cues, he may make use of the dynamic change of cues with respect to head movement. Fig. 8.6 shows the interaural time difference (ITD) in conjunction with the yaw rotational movement, which is likely to be used for front-back discrimination. In addition, the yaw rotation is by far the most likely form of movement in the course of the object localisation process by all the senses including vision. The ITD is calculated in the same way as that described in Section 5.2.2. The sound source is on the median plane at the elevations from 0° to 90° with 10° increments.

The ITDs given by the yaw rotation of the head from −180° to 180° are plotted in order to illustrate the ITD change by the sound sources in the upper hemisphere. The ITD change due to the sound sources in the lower hemisphere shows a similar tendency but is not illustrated here. The slopes of the ITD curves show the dynamic change of ITD in accordance with its elevation. Most of the frontal source directions produce negative change and rear directions corresponds to positive change.

When the control transducers are at 0° elevation (in front on the horizontal plane), as has been used in many trials, the yaw rotational movement always produces a negative change of ITDs, more specifically, a negative value corresponding to the frontal source at 0° elevation. This is illustrated in Fig. 8.7a showing an example of the ITD change due to a yaw rotation from -40° to 40°. However, when the control transducers are at 90° elevation (on the frontal plane), the yaw rotational movement does not produce any ITD change (Fig. 8.7b), which is only the case when the sound source is directly above or below the head in the real acoustic environment. Therefore, even though it does not give additional information to resolve the front-back ambiguity, it will not give "wrong" cues that may result in systematic bias error (in this example, bias towards the front).

The head movement should be restricted, in principle, for the synthesis of virtual acoustic environments unless the control filters are adjusted according to the head movement. However, there will often be some uncontrollable head movement or errors in adjusting the control filters in accordance with the head movement, especially in practical conditions. Therefore, placing the control transducers in positions in the frontal plane, especially in the upper hemisphere (above the head), has an advantage over other locations.

8.4 Other considerations

It is a known phenomenon that when a listener has ambiguity in judging the height of the sound source, humans tend to take directions in the horizontal plane as a default, since this is most likely to happen in the real acoustic environment. Therefore, concern about bias perception in the up-down direction would be somewhat relaxed.

When virtual visual information as well as acoustic information is to be presented to the subject, it is preferable to avoid the existence of the transducers in the listener's sight. This is especially important for systems that aim to present virtual visual information over the whole field of vision of the listener. This implies that elevation directions between -90° and 90° are best avoided (Fig. 8.8).

It has been shown that the area where the control is reasonably good extends a relatively large radius along the directions perpendicular to the interaural axis (Chapter 5). This characteristic is sometimes used in order to present to multiple listeners aligned on the median plane. When the head position is displaced from the optimal position along directions perpendicular to the transducer directions on the median plane, it produces errors due to the change in elevation direction. In such a case, the frequency response of the plant changes considerably around 0° whereas it does not around 90° elevation, as seen in Fig. 8.2 ~ Fig. 8.4.

8.5 Subjective experiments

The analysis above strongly suggests that 90° elevation (in the frontal plane in the upper hemisphere) seems to have several advantages and provides a possible alternative to the usual location at 0° elevation (in the horizontal plane in front). A pair of subjective evaluations were carried out in order to confirm this observation. A localisation experiment for the OSD system is carried out for both 0° elevation and 90° elevation. Another localisation experiment with false dynamic information induced by a head rotation is also carried out. The experimental procedure is fundamentally the same as that described in Section 7.9. Three young adults who all had normal hearing with no history of hearing problems, served as paid volunteers.

8.5.1 Localisation performance with the control transducers above the head.

The perceived virtual sound source directions are shown in Fig. 8.9. Filled circles represent the directions that are perceived and its size represents the number of occurrences of the perceived direction. The presented directions are denoted by open circles. Fig. 8.9a shows the results for the 0° elevation control transducer location. The responses are clustered towards the horizontal plane (0° and ±180° elevation). There is little perception in the region above and below the head. Fig. 8.9b shows the results for the 90° elevation control transducer locations. The responses are more evenly spread over various elevations. Nevertheless, a cluster around 80° (near the transducer elevation) is noted as well as that of around -140° (lower rear quarter). There is relatively little perception in the lower front quarter. The characteristics shown by the control transducers at 90° elevation seem particularly suitable for the presentation of a virtual acoustic environment together with visual information. This is because the image presented by the visual system is likely to shift the auditory perception towards the front, therefore reducing the errors.

The perceived directions have been decomposed into the azimuth directions and elevation directions and are shown in Fig. 8.10 and Fig. 8.11. Both of the two elevation transducer locations showed very good azimuth localisation performance. In Fig. 8.10, the median values (the square marker), 25 percentiles and 75 percentiles (the star marker) of the all the responses presented to the azimuth direction are plotted. There is little difference between the two. Conversely, the elevation localisation proved to be much more difficult with both transducer locations. Therefore, all the responses are plotted in Fig. 8.11 and the size of each filled circle represents the number of responses at that direction. The dashed line shows the direction of the control transducers. The cluster around the horizontal plane is noted in Fig. 8.11a which shows the response by the 0° transducer elevation. Fig. 8.11b produced by the 90° transducer elevation shows less biased responses although the results are somewhat scattered.

8.5.2 Effect of the false dynamic cue

Another set of localisation experiments was carried out with false dynamic information induced by listener head rotation. An initial experiment where a yaw rotation of ±3° was continuously induced by the subjects themselves showed little difference in localisation performance. The observation supports the superiority of the spectral cue over the dynamic cue. However, in order to investigate the difference in two different control transducer elevations, the yaw rotation was increased to ±5°.

The perceived virtual sound source directions are shown in Fig. 8.12. In Fig. 8.12a, it is clear that the perception is biased completely towards the front hemisphere, when comparing results with those shown in Fig. 8.9a. There is very little response to the rear of the subject. There is a notable difference of the perception of the virtual sound sources on the median plane compared to the other azimuth directions. There, the virtual sources for all the elevations collapsed not only towards the front but also on to the horizontal plane where the control transducers are located. This is not the case for other azimuth directions for which only the bias error towards the front hemisphere is outstanding. The elevation cues other than the front-back discrimination are more robust here than on the median plane and supports the importance of the binaural spectral shape cue [27]. On the contrary, little change in bias localisation error is observed in Fig. 8.12b.

The results for the azimuth directions and elevation directions are shown in Fig. 8.13 and Fig. 8.14. Again, both of the two alternative transducer locations showed very good azimuth localisation performance. There is little difference between the two. On the contrary, there is a significant difference in elevation localisation between the two different transducer locations. Most of the perceptions are clearly biased towards the control transducer elevation when they are at 0° elevation. However, the bias is not at all as strong when the control transducers are at 90° elevation.

8.6 Conclusions

In order to establish the characteristics of the various elevation positions of the control transducers, the analysis of the spectral cues and dynamic cues as well as a set of subjective experiment has been performed. The frequency response of the plant suggests that promising control transducer positions are in the frontal plane above the listener's head. The condition of the plant matrix shows the disadvantage of locations in the rear hemisphere. An analysis of the dynamic cues induced by unwanted head rotation strongly supports the transducer location on the frontal plane. A subjective experiment is performed in order to compare between two alternative control transducer locations, on the horizontal plane in front of the listener and on the frontal plane above the listener's head. The results without false dynamic cues show that both can perform equally well, with different advantages and disadvantages. However, the control transducer location above the head clearly shows the advantage of discriminating against false dynamic information.

The characteristics of the localisation error support the hypothesis that the transducer location above the head may be especially suitable when visual information is presented at the same time as audio information.

Related publications

[A 17] T. Takeuchi and P. A. Nelson, "Elevated control transducers for virtual acoustic imaging", UK Patent Application

[A 18] T. Takeuchi and P. A. Nelson, "Optimal source distribution for binaural synthesis over loudspeakers (in Japanese)", Technical report of IEICE, EA2001-67 (2001).

[A 19] T. Takeuchi, P. A. Nelson and M. Teschl, "Elevated Control Transducers for Virtual Acoustic Imaging", 112th AES Convention Preprint 5596 (O6).

[A 20] T. Takeuchi and P. A. Nelson, "Control transducer locations for virtual acoustic imaging using binaural principle," presented at the 143th Meeting of the Acoustical Society of America, 3-7 June 2002, Pittsburgh, USA.

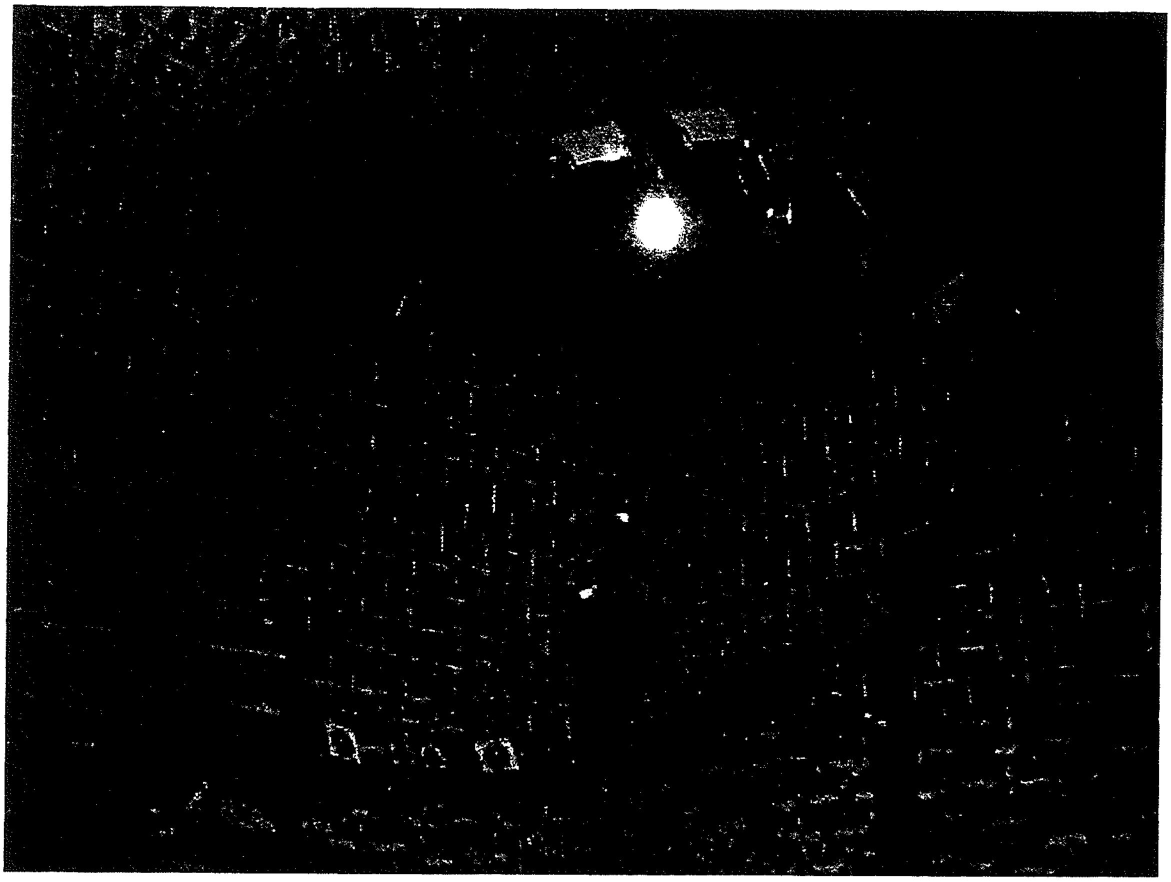
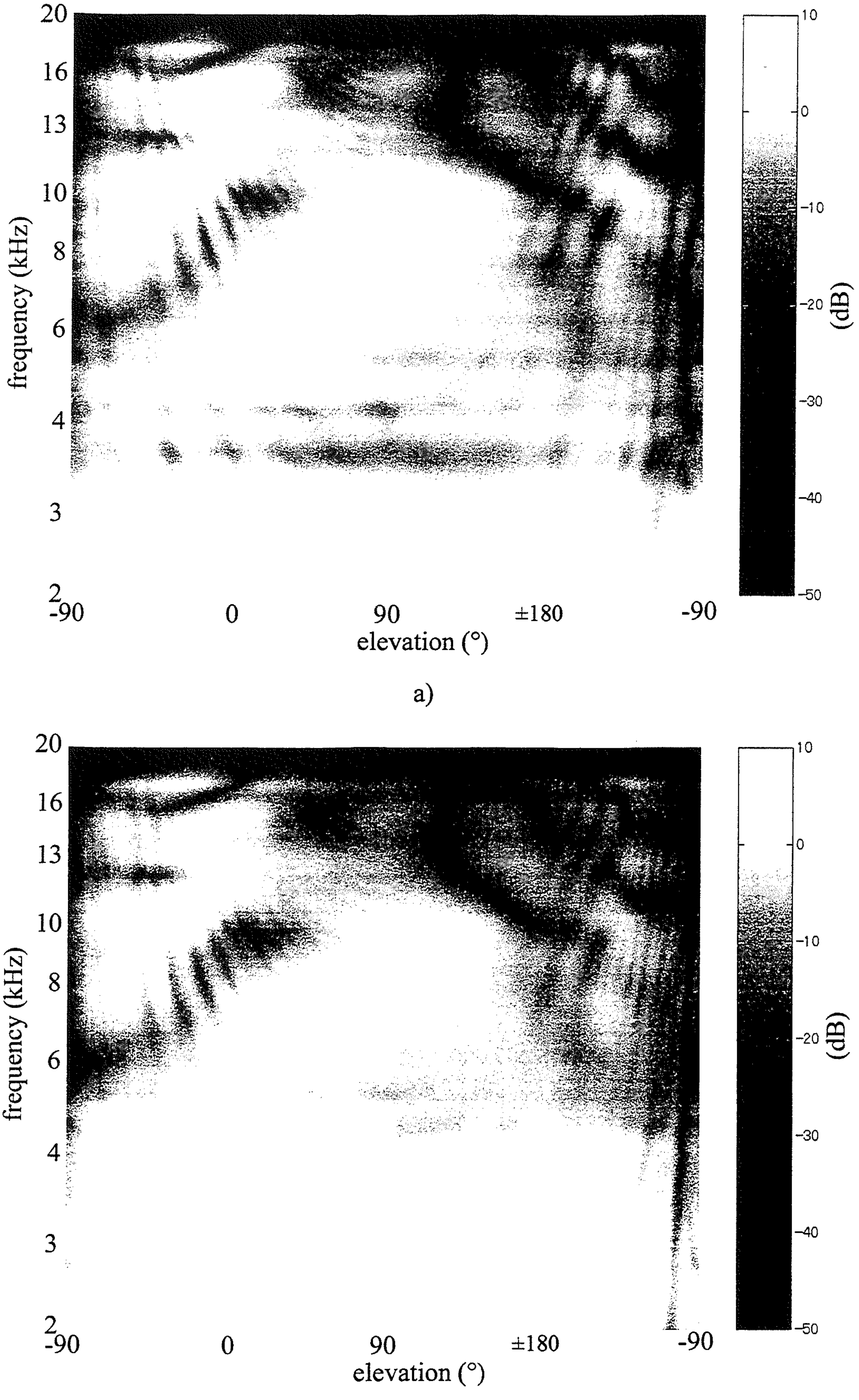


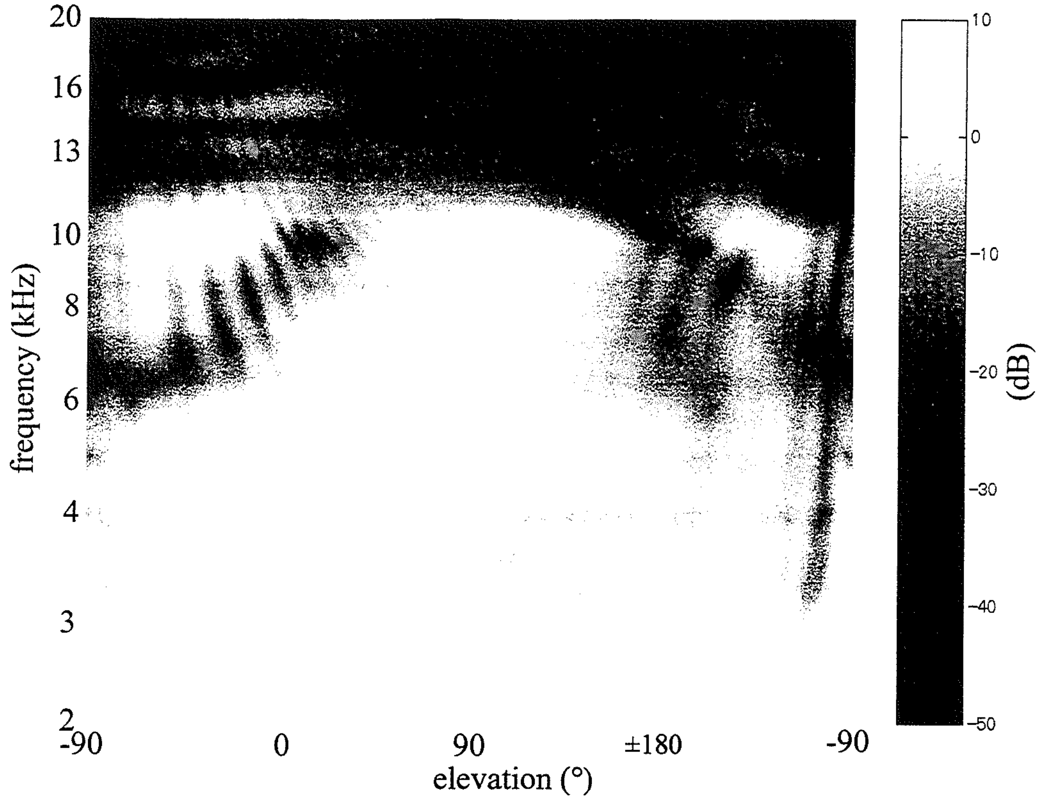
Fig. 8.1 Experimental rig for the plant measurement.



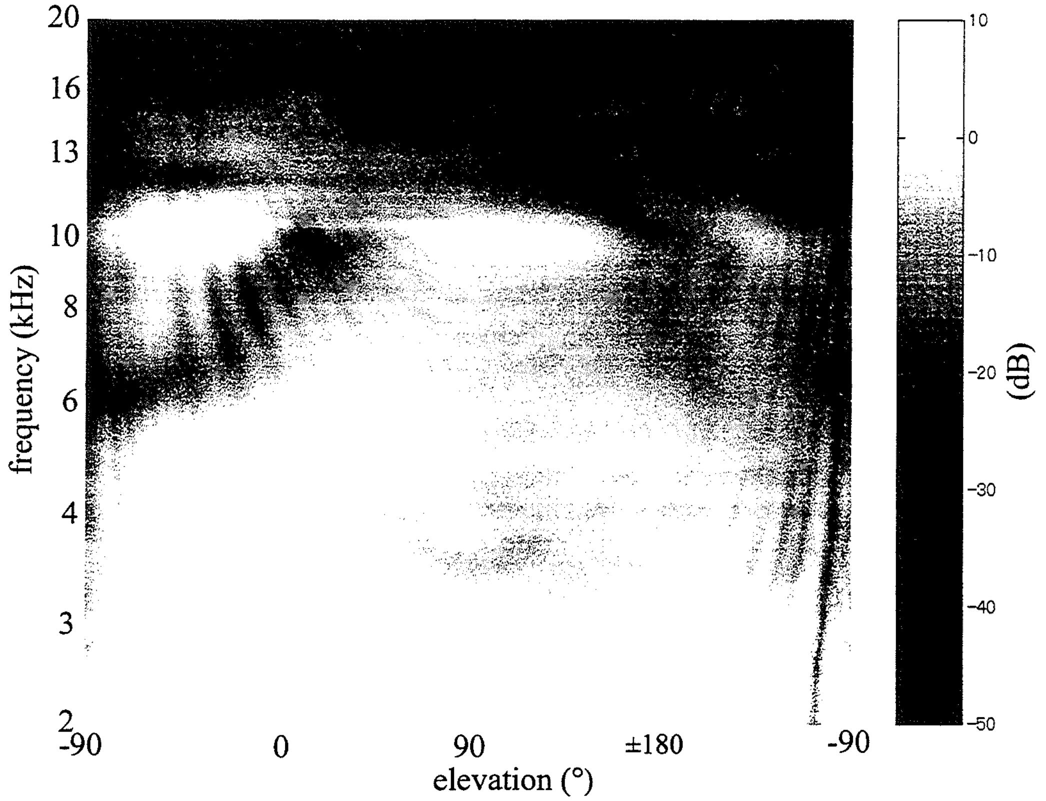
a)

b)

Fig. 8.2 Frequency response of the plant of the OSD system for the various elevations. a) plant for the ipsi-lateral ear. b) plant for the contra-lateral ear.



a)



b)

Fig. 8.3 Frequency response of the plant of the SD system for the various elevations. a) plant for the ipsi-lateral ear. b) plant for the contra-lateral ear.

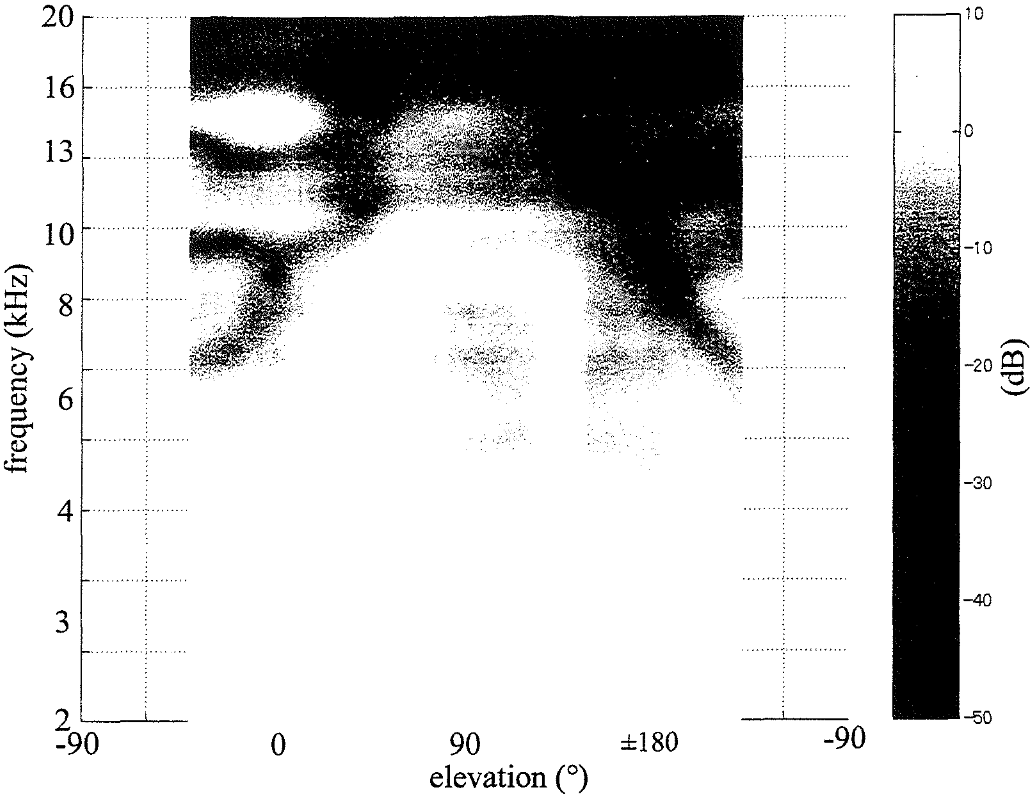
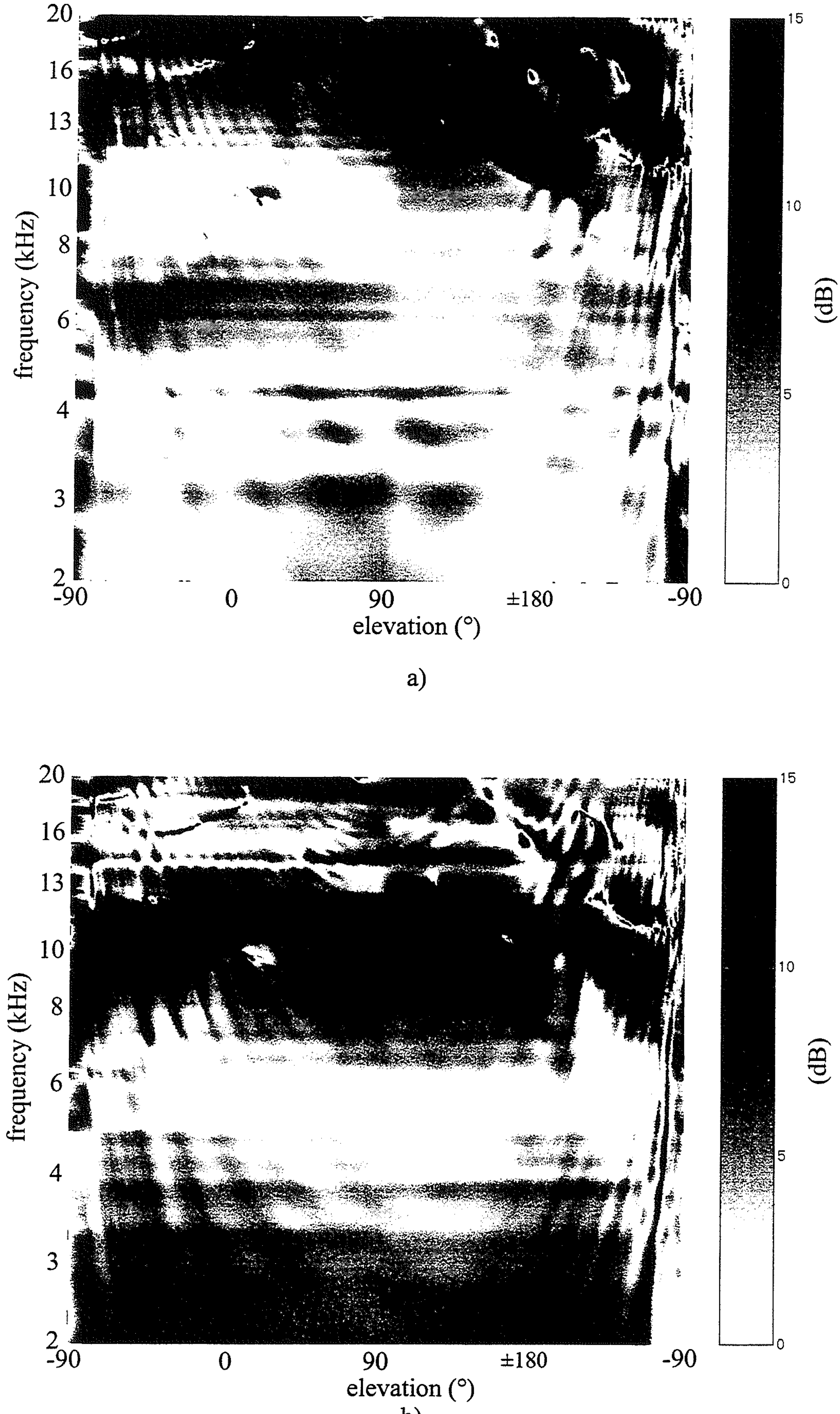


Fig. 8.4 Frequency response of the HRTFs for the directions on the median plane (calculated from the MIT database).



a)

b)

Fig. 8.5 Condition number for the plant matrix. a) OSD system. b) SD system.

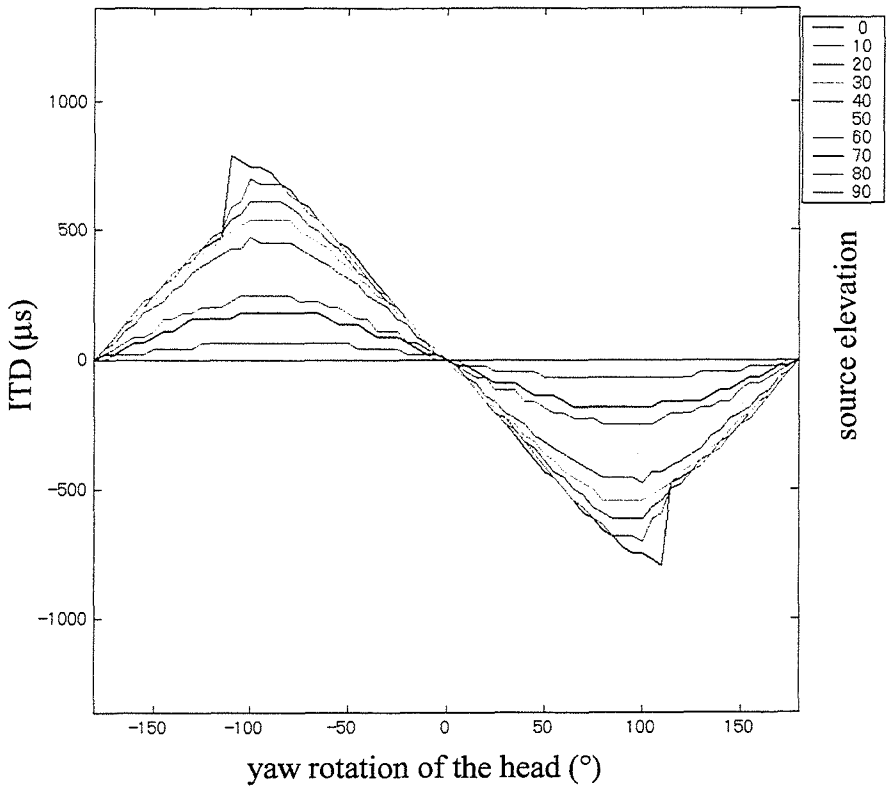
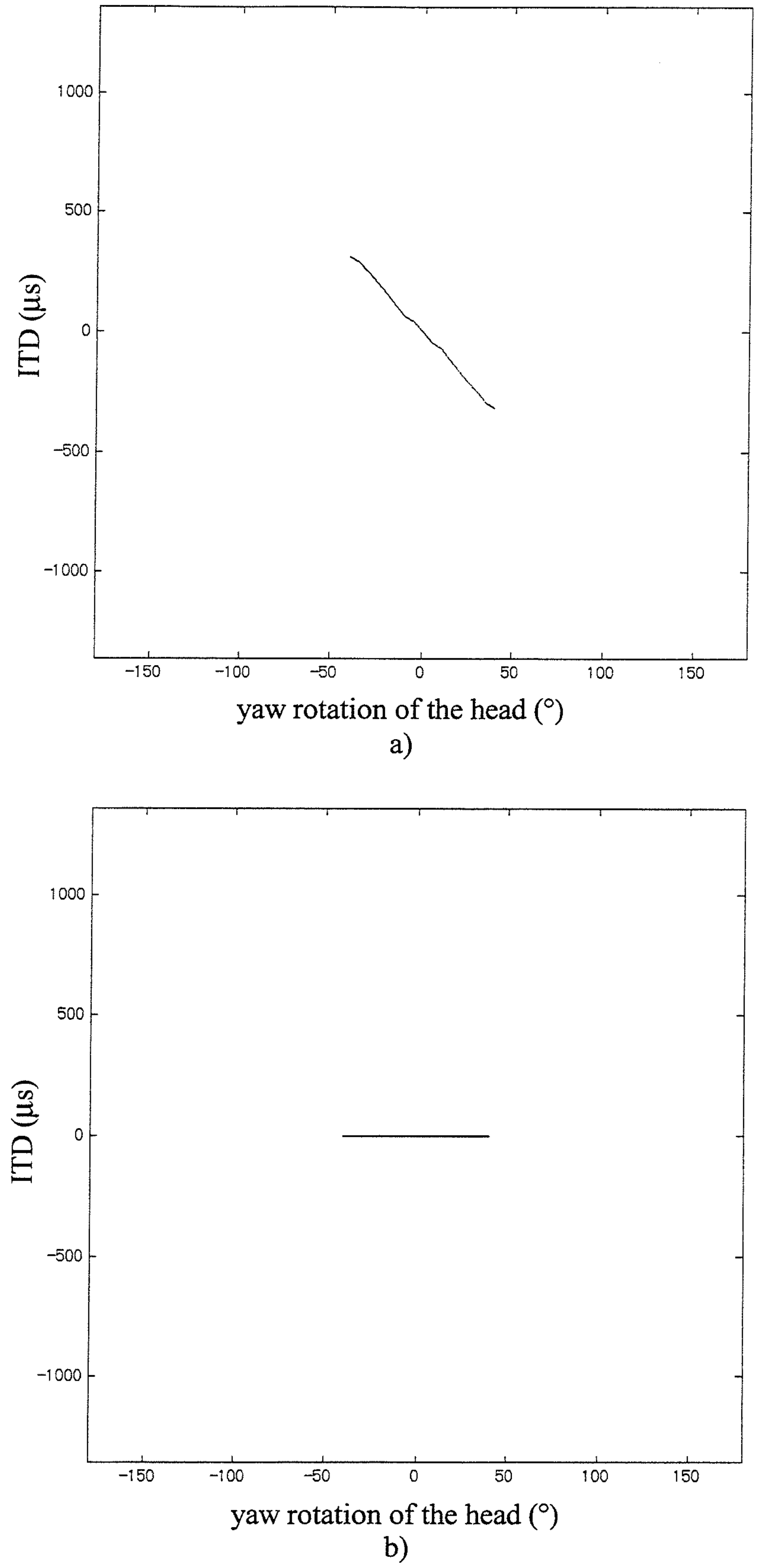


Fig. 8.6 Change of ITD for sound sources at various elevation directions corresponding to the yaw rotational displacements.



a)

b)

Fig. 8.7 Dynamic change of ITD for all virtual source directions produced by the control transducers in conjunction with yaw rotational movement over -40° to 40°. a) 0° elevation (in front on the horizontal plane). b) 90° elevation (on the frontal plane) (Example: the SD system.)

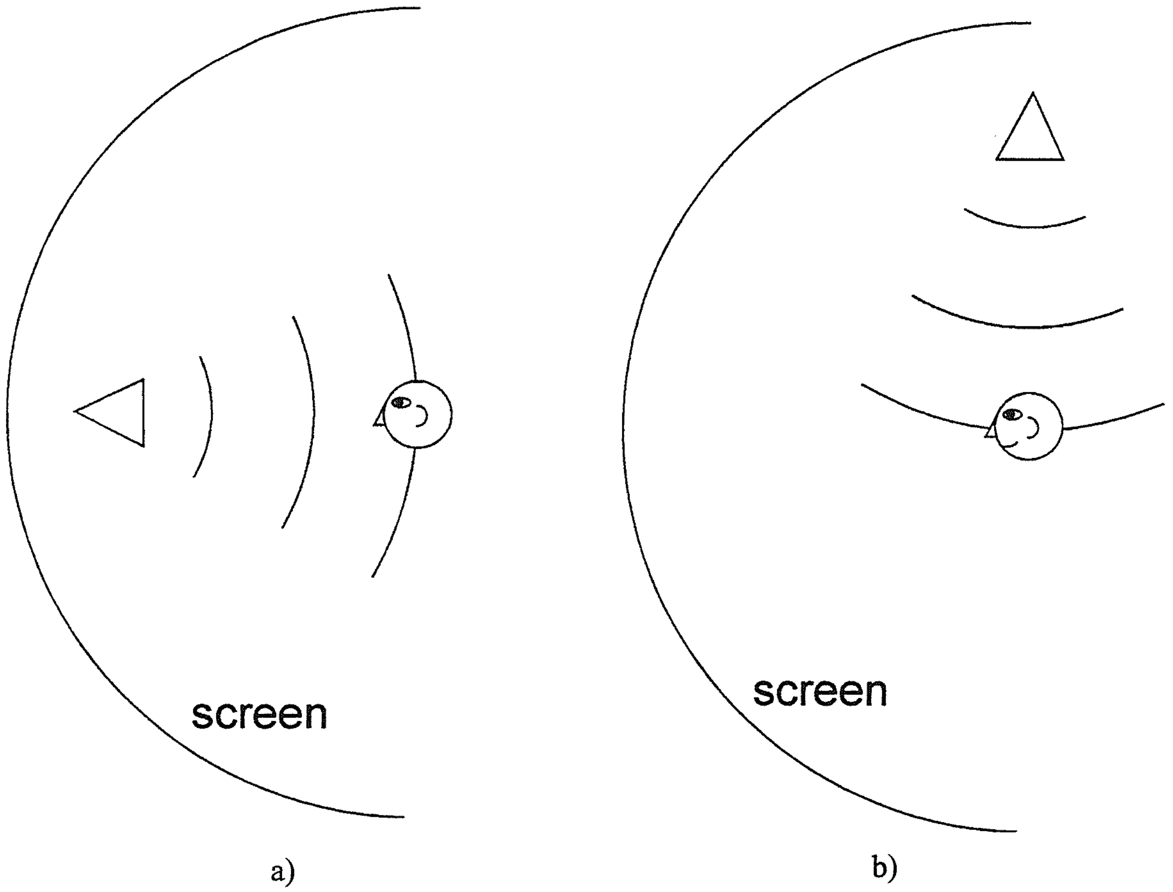
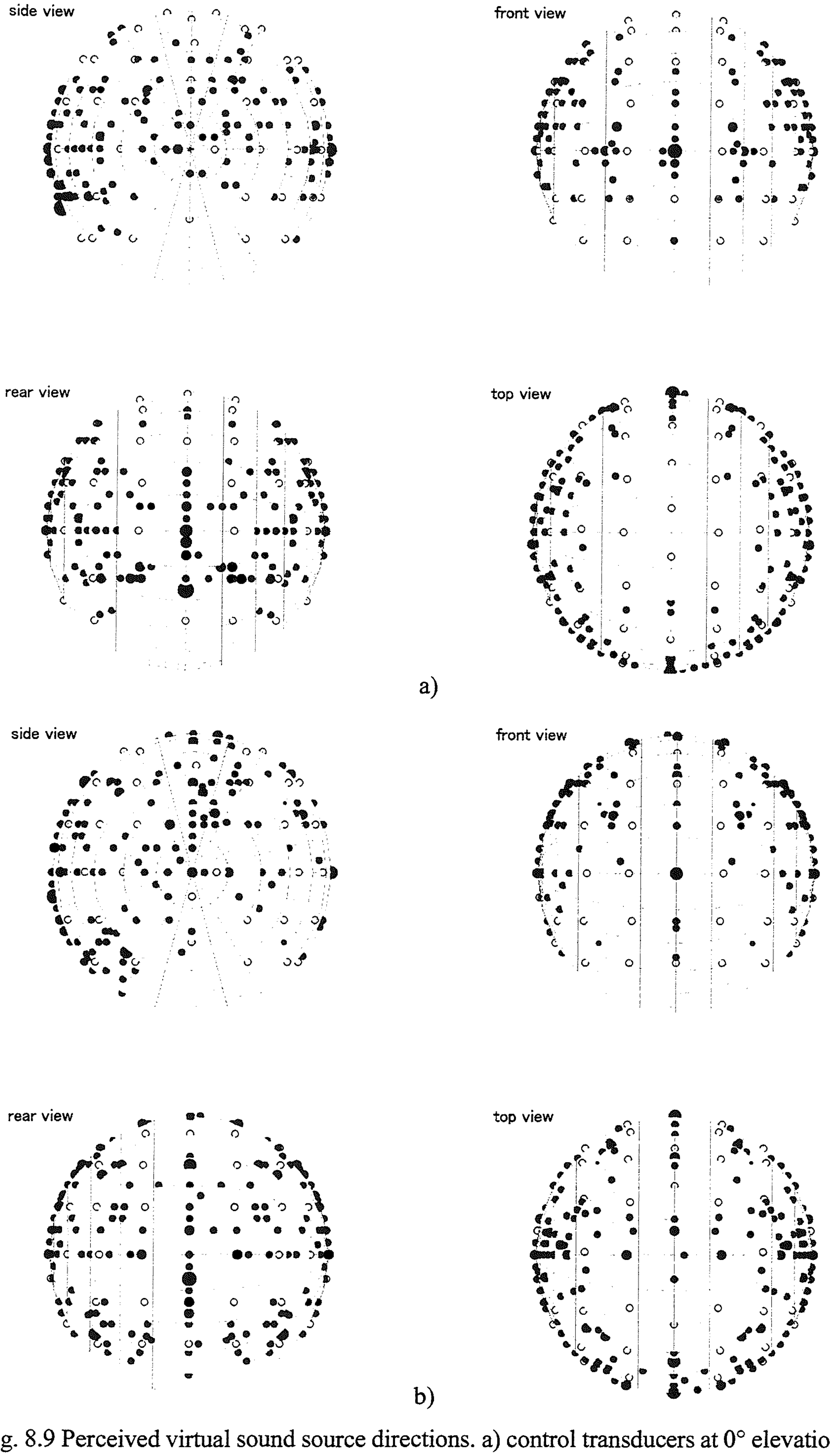


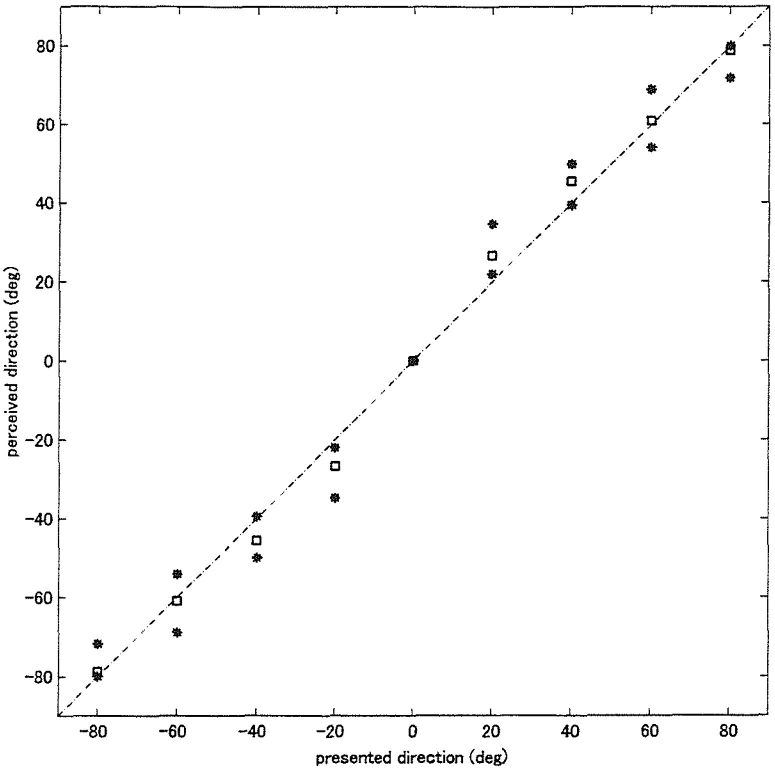
Fig. 8.8 Binaural reproduction over loudspeakers with visual information. a) control transducers around 0° elevation. b) control transducers around 90° elevation.



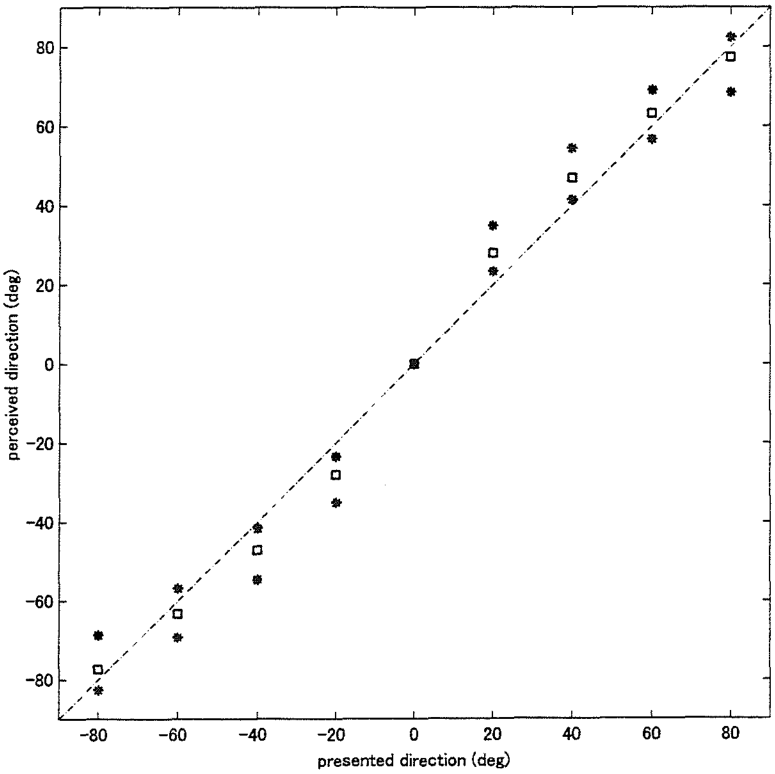
a)

b)

Fig. 8.9 Perceived virtual sound source directions. a) control transducers at 0° elevation. b) control transducers at 90° elevation.

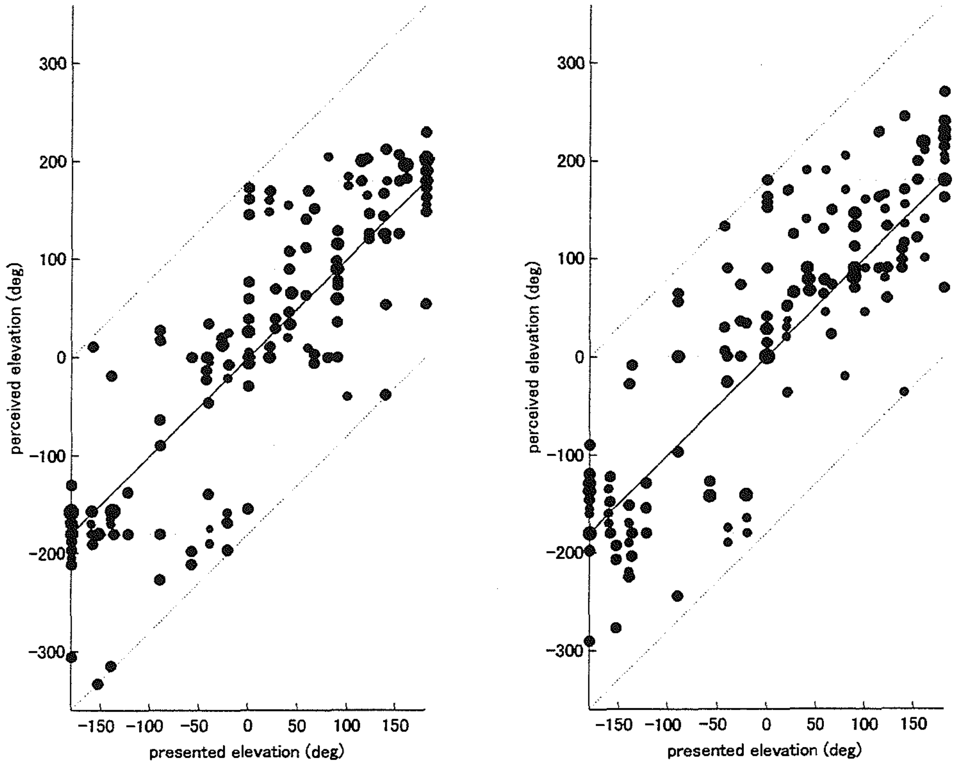


a)



b)

Fig. 8.10 Azimuth localisation performance. The square marker denotes median and the star marker denotes 25 and 75 percentile. a) control transducers at 0° elevation. b) control transducers at 90° elevation.



a) b)

Fig. 8.11 Elevation localisation performance. a) control transducers at 0° elevation. b) control transducers at 90° elevation.

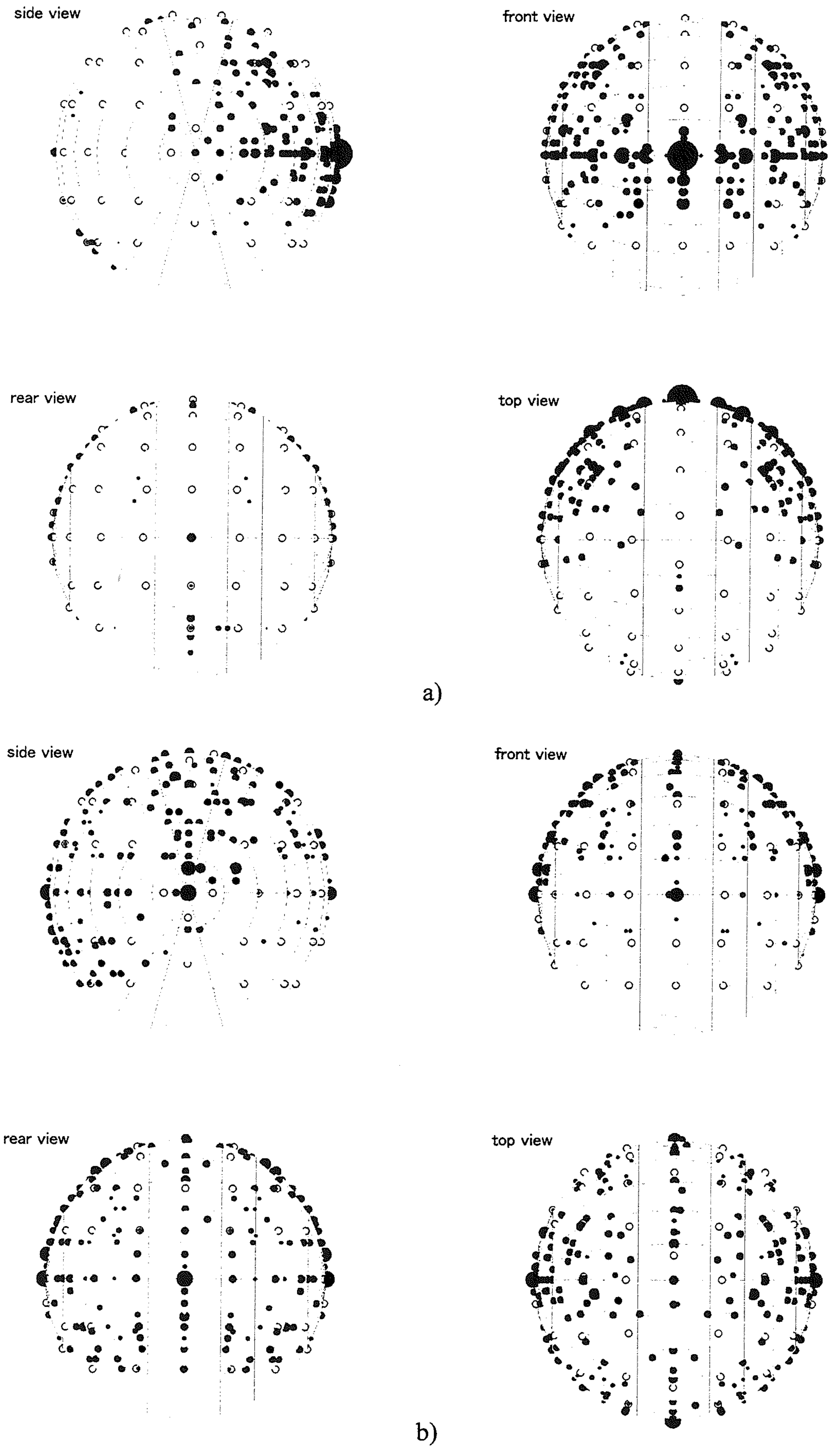


Fig. 8.12 Perceived virtual sound source directions with false dynamic information by yaw head rotation. a) control transducers at 0° elevation. b) control transducers at 90° elevation.

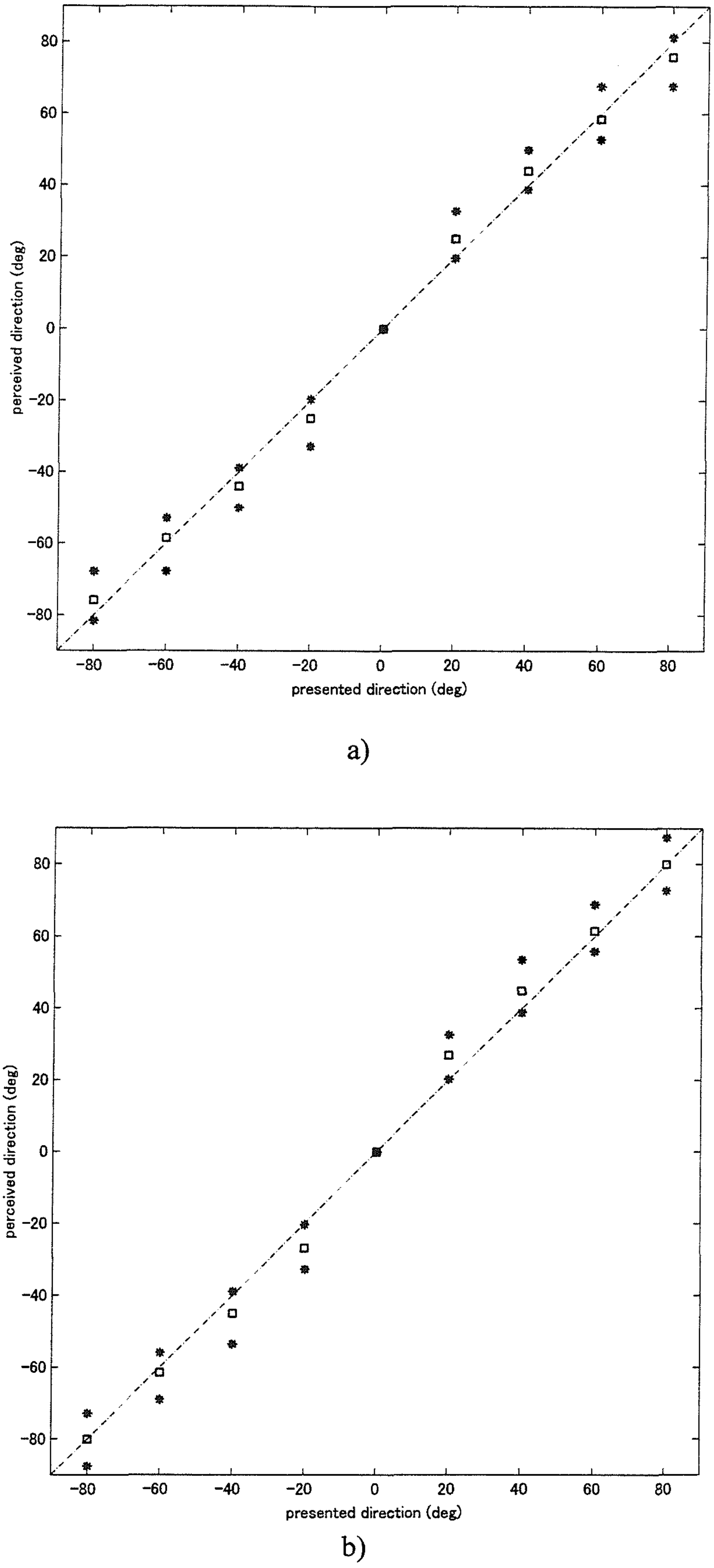


Fig. 8.13 Azimuth localisation performance with false dynamic information by yaw head rotation. The square marker denotes median and the star marker denotes 25 and 75 percentile. a) control transducers at 0° elevation. b) control transducers at 90° elevation.

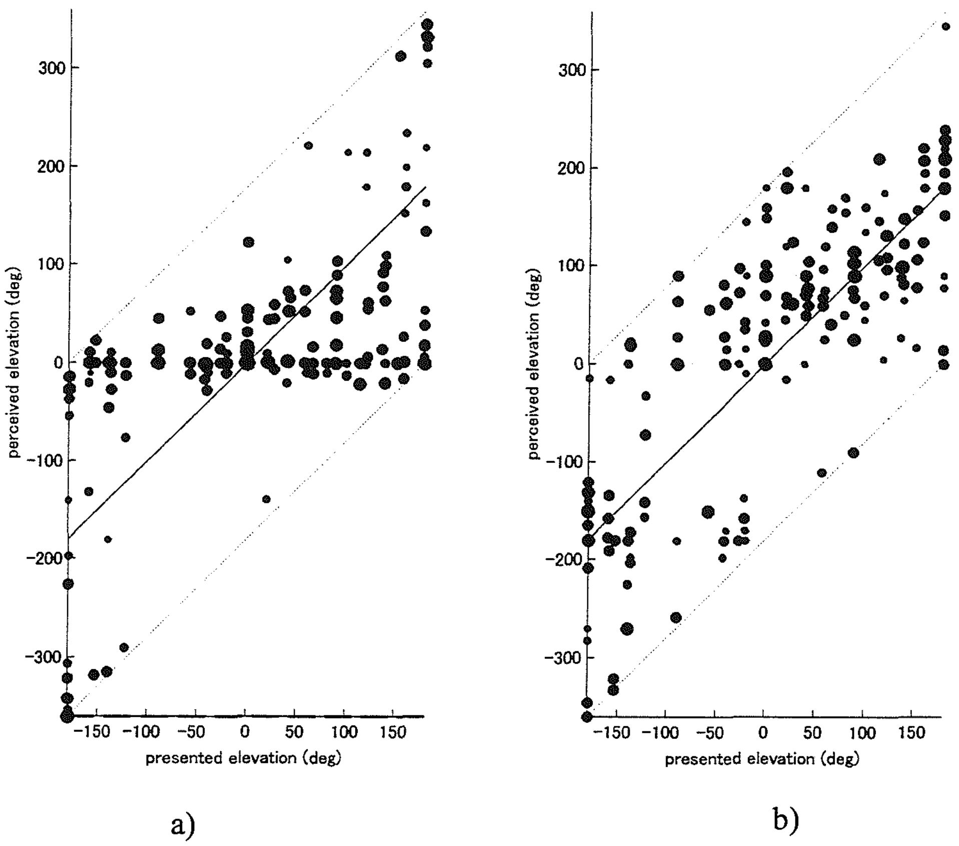


Fig. 8.14 Elevation localisation performance with false dynamic information by yaw head rotation. a) control transducers at 0° elevation. b) control transducers at 90° elevation.

9 Summary

Various aspects of spatial sound reproduction with systems using binaural reproduction over loudspeakers are investigated in this thesis. Investigation of factors contained in the plant that could lead to the deterioration in the performance of such systems is the particular objective. The ability to produce the spatial aspects of the intended sound field has received more attention than the quality of sounds. As a consequence of the investigation, a number of methods to overcome or mitigate the problems associated with such systems have been proposed.

The location of transducers for binaural reproduction over loudspeakers has received little attention so far. Two transducers are usually placed symmetrically in front of a listener subtending an angle of about 60°. A binaural synthesis over loudspeakers can also be made to operate remarkably effectively, and arguably more effectively, by using a pair of loudspeakers that are placed very close together. Such a system is referred to as a "Stereo Dipole". It is shown that it is possible to achieve independent control of the sound signal at two ears with a monopole transducer and a dipole transducer at the same position. When two closely spaced monopole transducers are used, the sound field produced is a good approximation to that produced by a point monopole and a point dipole transducer up to a given frequency. The basic behaviour of the sound fields generated by virtual sound imaging systems are analysed. It is demonstrated that the use of two closely spaced loudspeakers also approximates such a source combination. Subjective experiments are also performed to establish the basic understanding of the performance of virtual spatial sound reproduction systems.

Virtual acoustic imaging systems based on the binaural technique turned out to show differences in performance between individual subjects. To make detailed investigations possible, the individual head related transfer functions (HRTFs) of each subject were measured. Synthesised binaural signals were compared with those specifically designed for each subject. It was found that mismatch of the HRTFs resulted in errors in image location.

Binaural sound presention with two loudspeakers requires the listener's ears to be in the relatively small region which is under control of the system. Misalignment of the head results in inaccurate synthesis of the binaural signals. Consequently, directional information associated with the acoustic signals is inaccurately reproduced. When the two loudspeakers are placed close together, the spatial rate of change of the generated sound field is much smaller than that generated by two loudspeakers spaced apart. Therefore, the performance of such a system is expected to be more robust to misalignment of the listener's head. Robustness of performance is investigated with respect to head displacement in three translational and three rotational directions. A comparison is given between systems consisting of two loudspeakers either placed close together or spaced apart. The extent of effective control with head displacement and the resulting deterioration in directional information is investigated in the temporal and spectral domain by analysing synthesised binaural signals. Subjective localisation experiments are performed for cases in which notable differences in performance are expected from the previous analysis. It is shown that the system comprising two loudspeakers that are close together is very robust to misalignment of the listener's head.

When a virtual acoustic imaging system based on the binaural technique is brought into a reverberant environment, reflections of the original sound give rise to a deterioration in the performance of the system. However, it is expected that the psychological functions of hearing such as the precedence effect may to some extent help preserve directional information. The effect of an infinite uniform reflecting surface on the performance of such systems are examined by computer simulations and subjective experiments. The ability to localise a virtual monopole source in the horizontal plane is investigated. The performance of the system is quite well preserved. This result suggests that such systems may work in normal rooms to some extent.

The system inversion involved in binaural presentation over loudspeakers gives rise to a number of problems such as a loss of dynamic range and a lack of robustness to small errors and room reflections. The amplification required by the system inversion results in loss of dynamic range. The control performance of such a system deteriorates severely due to small errors resulting from, e.g., misalignment of the system and individual differences in the head related transfer functions at certain frequencies. The required large sound radiation results in severe reflection which can also reduce the control performance. A method of overcoming these fundamental problems is proposed. A conceptual monopole transducer is introduced whose position varies continuously as frequency varies. This gives a minimum requirement for the processing of the binaural signals for the control to be achieved and all the above problems either disappear or are minimized. The inverse filters have flat amplitude response and the reproduced sound is not coloured even outside the relatively large "sweet area". A number of practical solutions are suggested for the realization of such optimally distributed transducers. One of them is a discretization that enables the use of conventional transducer units. The results of objective and subjective evaluation confirmed the advantage of the OSD system.

The elevation location of transducers for binaural reproduction over loudspeakers has hitherto received little attention. In most of the past research, the transducers are usually placed on the horizontal plane that includes the listener's head. In order to examine the characteristics of various elevation positions of the control transducers, an analysis is performed of both the spectral and dynamic cues that relate to localisation. The frequency response of the plant that relates transducer outputs to ear pressure signals suggests that control transducer positions will be promising at positions in the frontal plane above the listener's head. The condition of the plant matrix shows the disadvantage of locations in the rear hemisphere. The analysis of the dynamic cues induced by unwanted head rotation strongly supports the use of transducer locations in the frontal plane. As a result of this analysis, transducer positions in the frontal plane above the listener's head are found to be promising as an alternative control transducer location. A subjective experiment is performed in order to compare between two alternative control transducer locations; in the horizontal plane in front of the listener and on the frontal plane above the listener's head. The results without false dynamic cues show that both can perform equally well, with different advantages and disadvantages. However, the control transducer location above the head clearly shows an advantage with respect to the transmission of false dynamic information. The characteristics of the localisation error support that the transducer location above the head is especially suitable when visual information is presented at the same time as audio information.

Appendices

Appendix 1 Spherical interpolation and extrapolation of HRTFs

When spherically sampled data is to be interpolated or extrapolated, it must be dealt with appropriately. Spherical extrapolation poses little problem since it can be achieved with simple spherical attenuation and time delay. It reduces the amount of required information from three dimensions to two dimensions. Interpolation needs to be made after the transfer functions are decomposed into magnitude, phase, and delay. Otherwise the interpolation in the time domain or in the frequency domain produces large errors. Phase need to be corrected since the interpolation in the frequency region where the difference of phase exceeds π results in a distortion of the time response. Spherical interpolation can be achieved with simple algebra by using solid angle as the weighting factor. The block diagram depicting the interpolation and extrapolation scheme is shown in Fig. A. 1. Each of the components of the process depicted in this figure are explained in the following paragraphs.

A.1.1 Extrapolation

When a required sound source is in the far field, the transfer functions corresponding to the source can be obtained by extrapolating the transfer functions (which are obtained by the spherical interpolation described later) on the sampling spherical surface. Thus the extrapolation reduces the amount of required information from three dimensions to two dimensions. The extrapolation only requires spherical attenuation and time delay due to propagation from a point source. However, since the time resolution of a normal digital audio signal and thus the impulse response data (22.7μs/sample if the sampling frequency is 44.1kHz) is much coarser than the minimum audible time resolution (10μs [25]), a non-integer number of sample delays needs to be dealt with by using a fractional delay filter.

A.1.2 Spherical bilinear interpolation

In order to obtain the transfer functions on the sampling spherical surface, it is necessary to interpolate from the far field transfer functions that are sampled two dimensionally on the spherical surface (Fig. A. 2). In the example of two dimensional spherical interpolation for the far field HRTF data, the weighting factor (w_n) associated with a value (x_n) to be interpolated at each vertex is the solid angle made by two other vertices and the point to be interpolated. The values (x_1, x_2, x_3) are, for example, magnitude, phase, or delay decomposed from the HRTFs at the sampled points on the sampling surface. The spherical interpolation can be achieved by using bilinear interpolation algebra with solid angle as the weighting factor (w_1, w_2, w_3). The obtained value (x_i) is given by

$$x_i = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3}{w_1 + w_2 + w_3} \tag{A 1}$$

When the interpolation point falls on the line (great circle) connecting two points on a spherical surface, Eq. (A 1) simply reduces to a linear interpolation using the angle made by the other point and the point to be interpolated as the weighting factor. When near field data is to be interpolated, another dimension of radius need to be incorporated.

A.1.3 Interpolation domain

Interpolation of multiple transfer functions (impulse responses) in the time domain produces large error. A good example of this is in the interpolation of two sinusoidal time series with the same amplitude but one of them out of phase (π) with the other. Instead of another sinusoid with the same amplitude with the phase in between (π/2), a constant value of zero is obtained. The result is the same in the frequency domain when two complex numbers are interpolated. Therefore, the interpolation needs to be made after the transfer functions are decomposed into magnitude, phase, and delay.

The interpolation of magnitude poses little problem. The interpolation of delay decomposed from discrete time series often results in a non-integer number of samples delay. This needs to be dealt in the same way as non-integer number of samples delay resulting from extrapolation. The interpolation of phase requires the most attention due to its cyclic nature. The phase difference between each vertex must not exceed π. This is mostly achieved by extracting an appropriate amount of delay. An example of the resulting phase responses at the vertices and the results of interpolation are shown in Fig. A. 3. Three phase responses of HRTFs (KEMAR) for the right ear sampled at (13° azimuth, -40° elevation), (12° azimuth, -30° elevation), and (18° azimuth, -30° elevation) are plotted together with the phase response interpolated from them. However, frequency regions where the difference of phase exceeds π still exist such that the interpolation results in a distortion of the time response as shown in Fig. A. 4. This is because 2π phase changes occur in certain frequency ranges, and their frequencies are different at every vertex. These 2π phase changes at each vertex at different frequencies result in too large a phase difference in order to be interpolated correctly.

A.1.4 Phase jump detection

In order to detect the 2π phase changes which give rise to the problem described above, a minimum component of the phase response is subtracted from the original phase responses and the resulting all pass components are shown in Fig. A. 5. The minimum phase component can be calculated [53] from the Hilbert transform of the logarithm of the magnitude response (Fig. A. 6) as follows.

$$\arg\left[X\left(e^{j\omega}\right)\right] = \frac{1}{2\pi} P \int_{-\pi}^{\pi} \log\left|X\left(e^{j\theta}\right)\right| \cot\left(\frac{\theta-\omega}{2}\right) d\theta \tag{A 2}$$

where the symbol *P* denotes Cauchy principal value [54] and is shown in Fig. A. 7. In Fig. A. 5, the frequencies where the phase moves through 2π can be clearly seen. Therefore, the original phase can be corrected so that 2π phase jumps are enforced at each resonance, thus each resulting phase response is similar (Fig. A. 8). The corrected all pass components are shown in Fig. A. 9. The resulting time responses are shown in Fig. A. 10 and it can be seen that the distortion of time responses observed in Fig. A. 4 have disappeared. Clearly the delay needs to be adjusted to incorporate these enforced 2π phase jumps accordingly. Exchanging between the 2π phase jumps and one sample delays itself introduce no error. It is just a matter of how the phase is expressed or wrapped. The frequency which is most likely to have interpolation phase error (though much less likely than the conventional method) is around the frequency where the 2π phase jump is enforced. However, this is where the effect of the error is minimal since the magnitude of the transfer functions are minimum and hence the contribution of this frequency region is minimum. It should be noted that this method requires no assumption or approximation about the phase response. Often the phase response is approximated as minimum phase for interpolation since it can be easily calculated from the magnitude response. This method produces the correct phase response including the all pass components of the desired phase response.

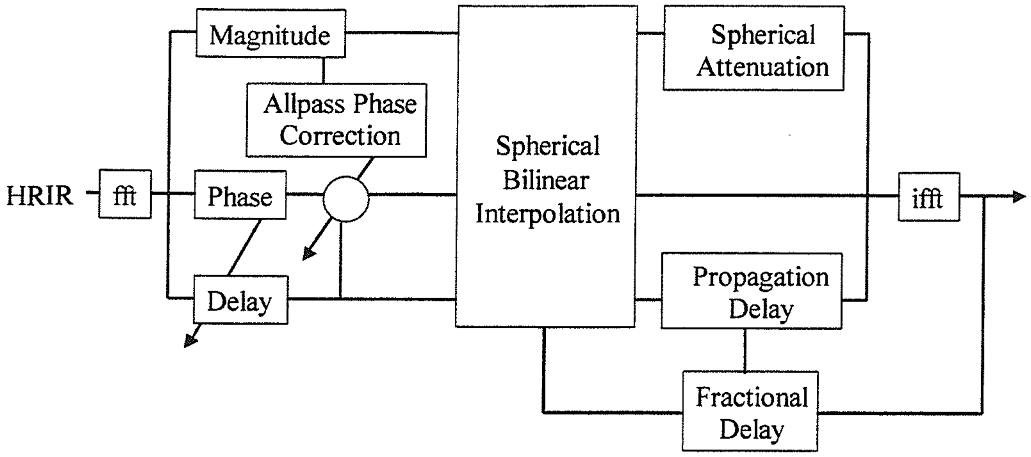


Fig. A.1 Interpolation and extrapolation of transfer functions (impulse responses).

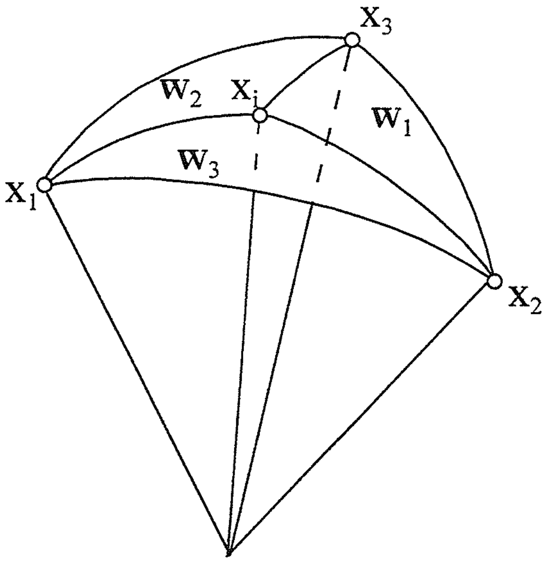


Fig. A.2 Spherical interpolation.

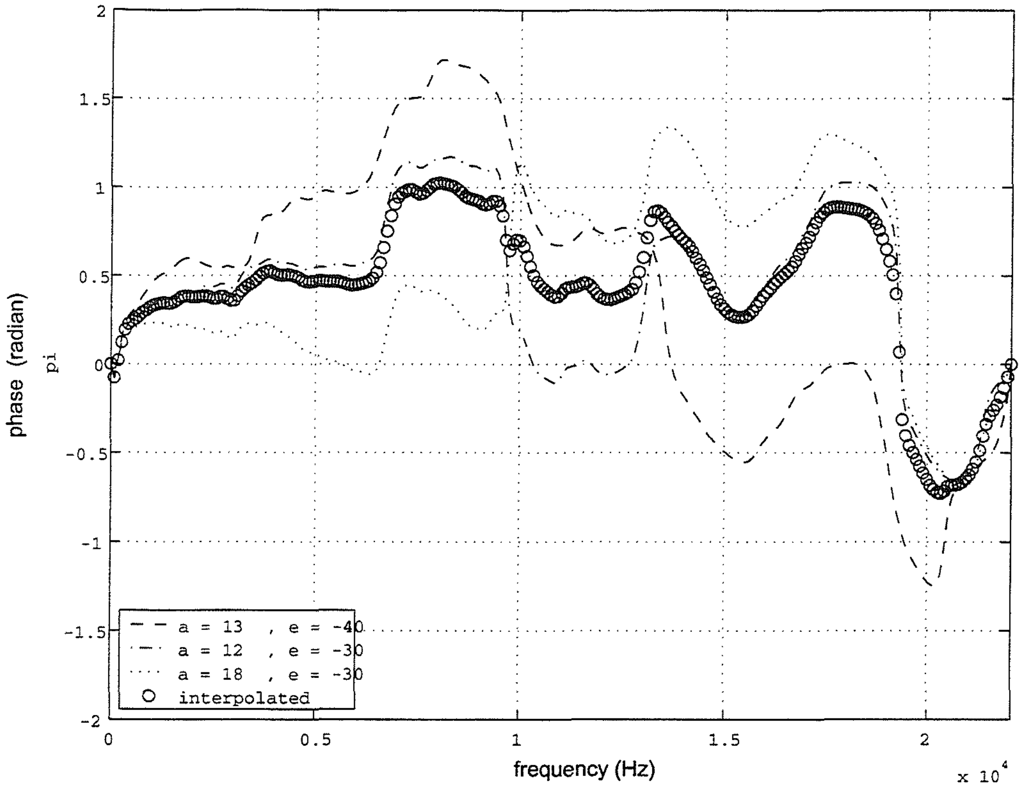


Fig. A.3 Phase response of transfer functions when pure delay is separated.

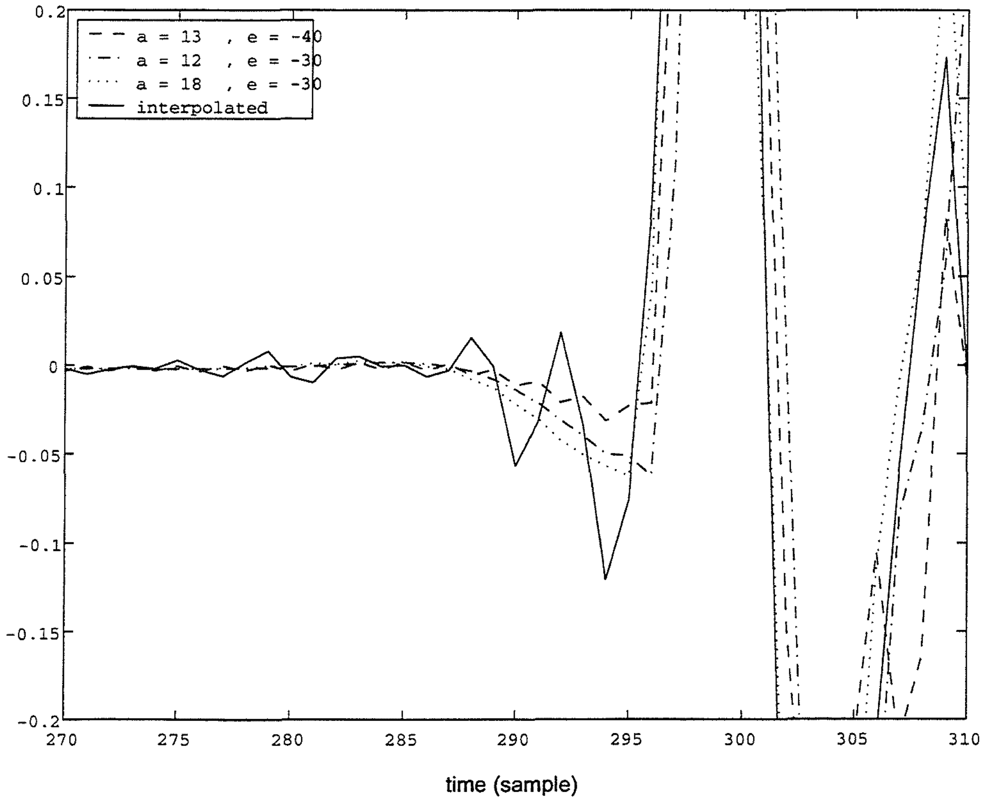


Fig. A.4 Impulse response before and after the interpolation without phase correction.

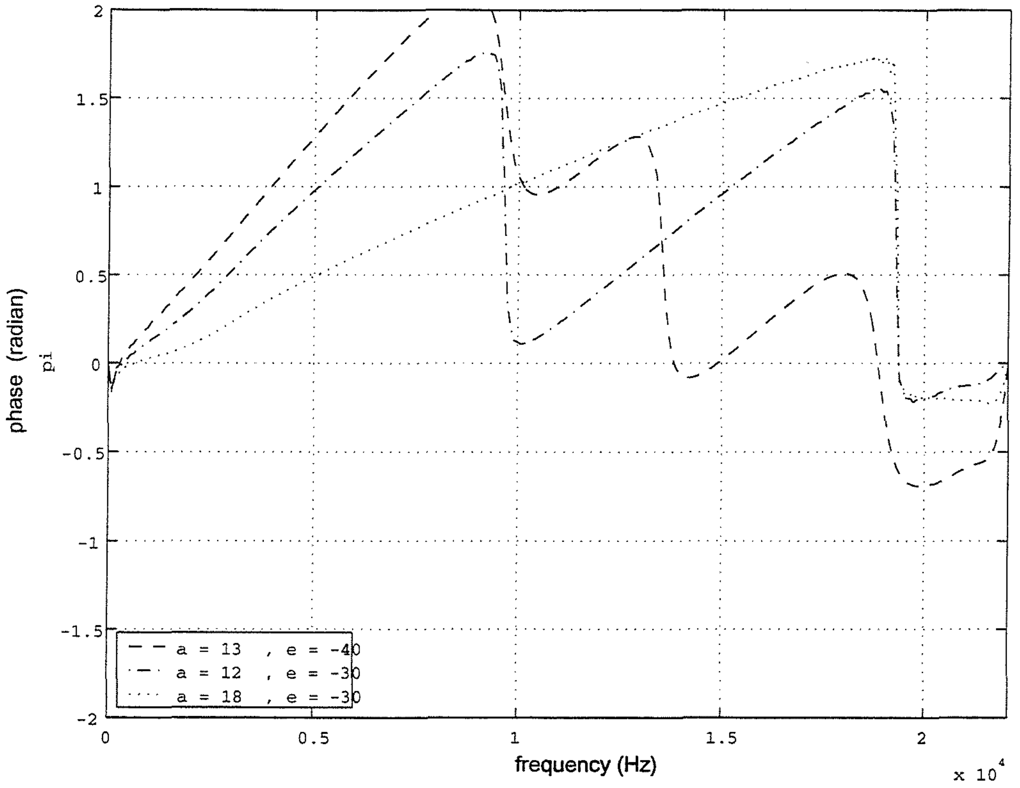


Fig. A.5 Difference of original phase response and minimum phase component (all pass component).

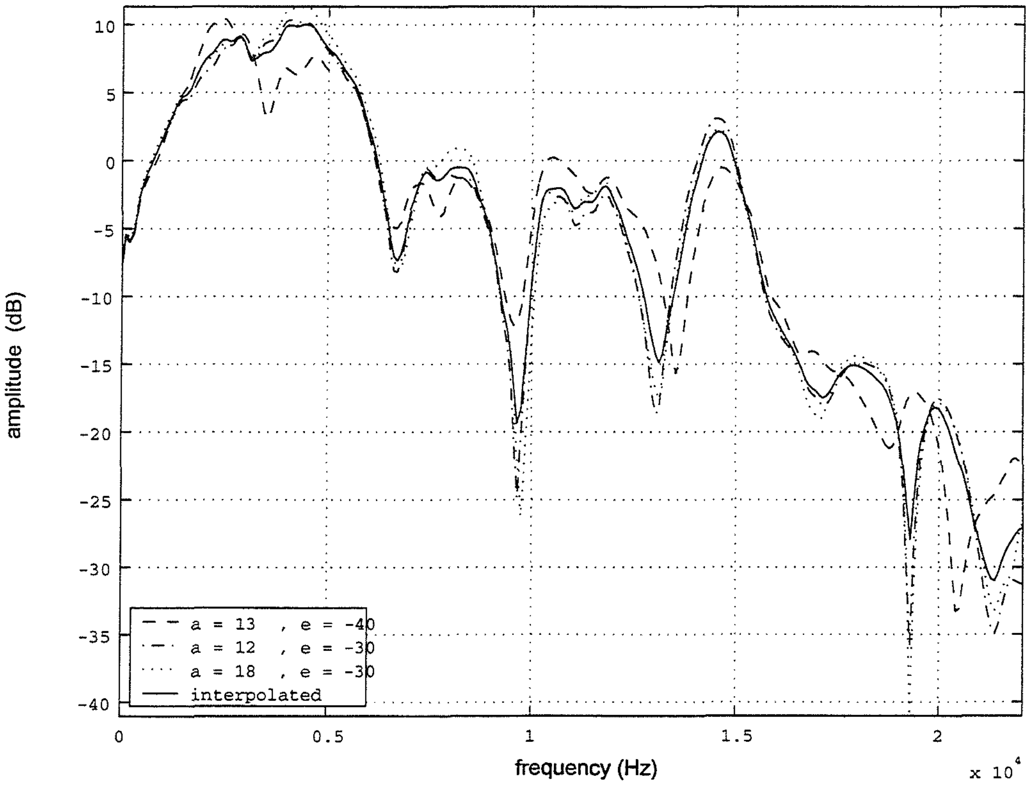


Fig. A.6 Magnitude response of transfer functions.

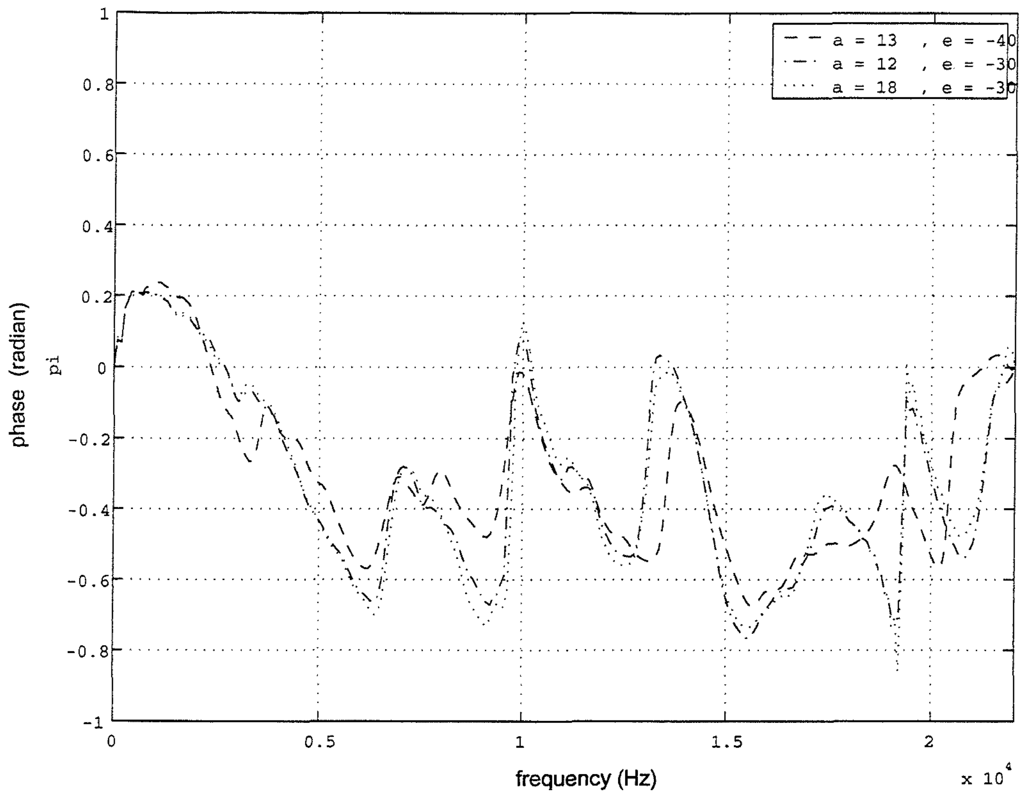


Fig. A.7 Minimum phase component obtained from the magnitude of the transfer functions.

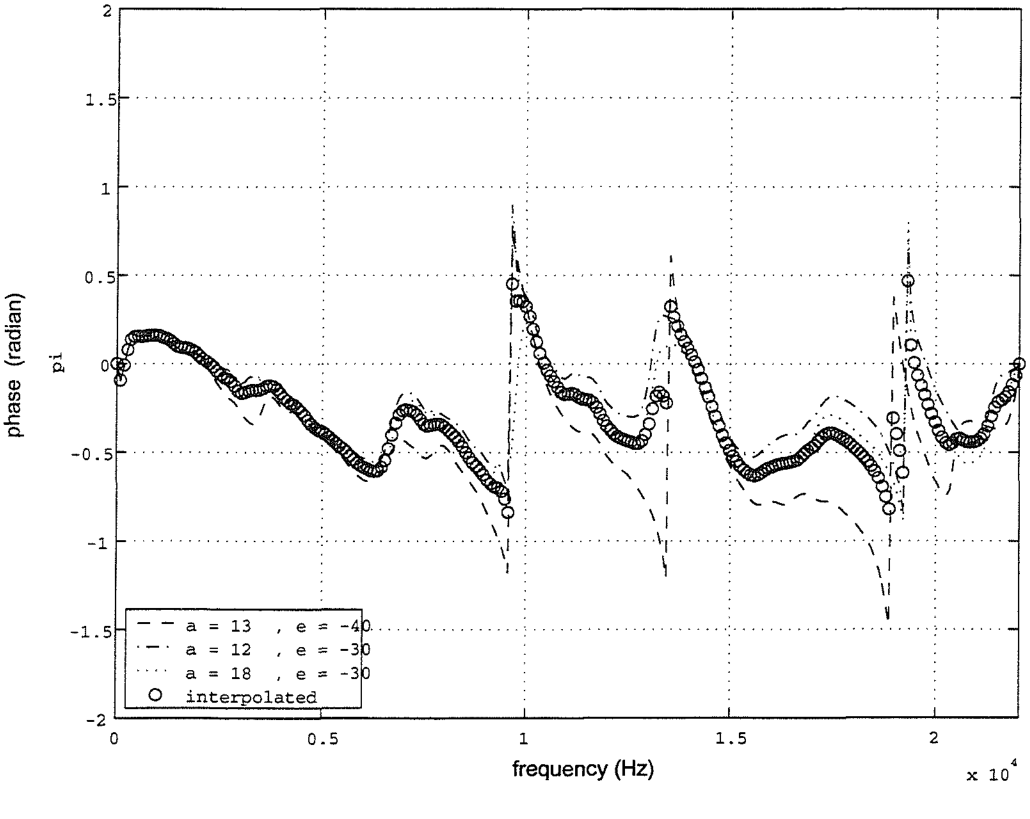


Fig. A.8 Corrected phase response for the interpolation.

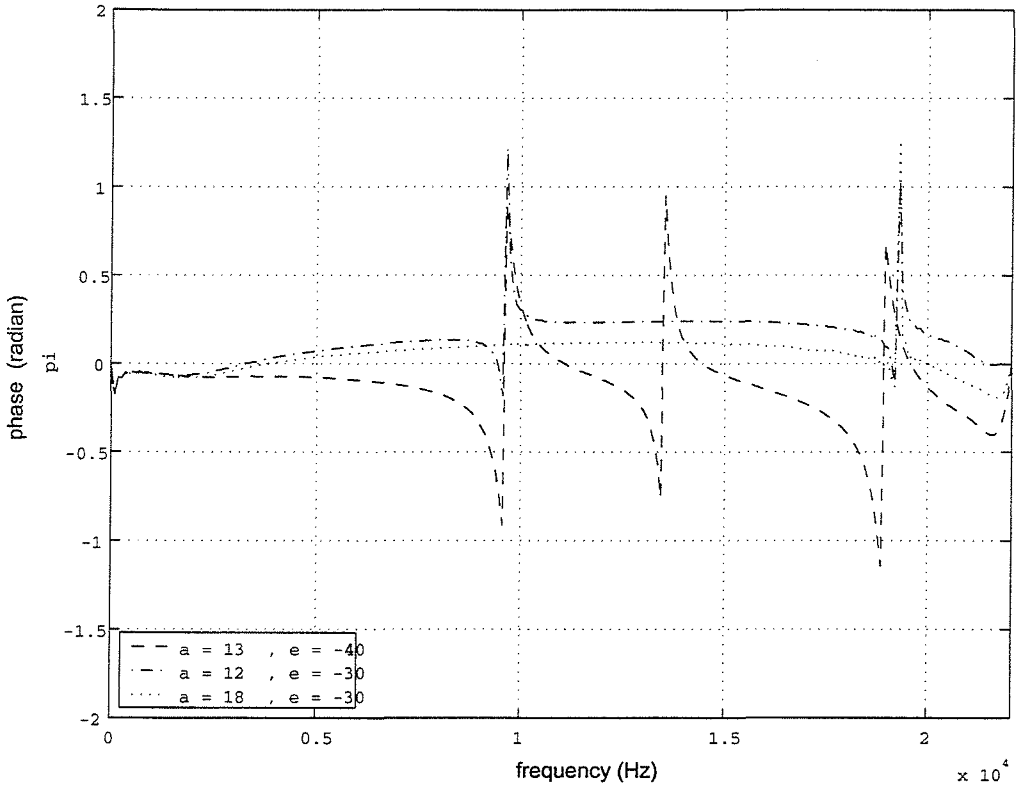


Fig. A.9 Corrected all pass components.

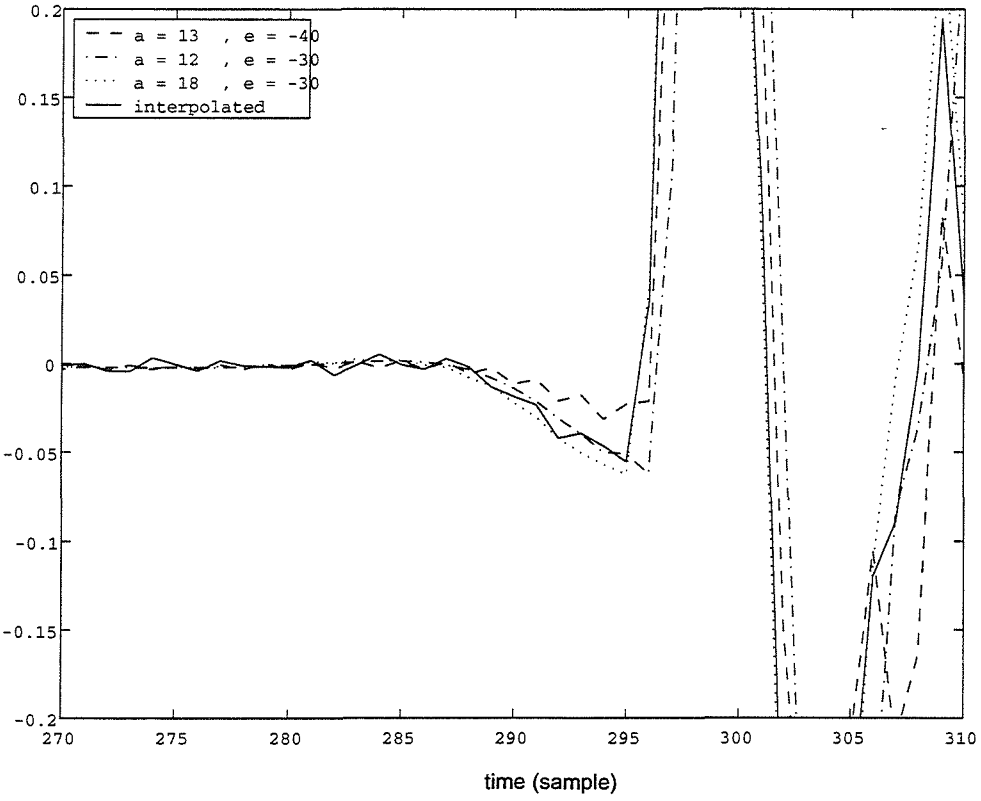


Fig. A.10 Impulse response before and after the interpolation with phase correction.

Appendix 2 Singular value decomposition

When the desired signals are defined as Eq. (7.3), this effectively normalises the plant transfer function matrix **C** with respect to the acoustic pressure signals which would have been produced by the closer of two sound sources to the receiver point. Then this normalised plant transfer function matrix **C** can be written as

$$\mathbf{C}=\begin{bmatrix} 1 & g\,\mathrm{e}^{-\mathrm{j}k\Delta l} \\ g\,\mathrm{e}^{-\mathrm{j}k\Delta l} & 1 \end{bmatrix} \tag{B 3}$$

It is possible to express **C** with unitary matrices **U** and **V** such that

$$\mathbf{C}=\mathbf{U\Sigma V}^{\mathrm{H}} \tag{B 4}$$

where **Σ** is the diagonal matrix whose elements are singular values of **C**, σ_1 and σ_2. The singular values of **C** can be found from the square roots of eigenvalues of $\mathbf{C}^{\mathrm{H}}\mathbf{C}$.

$$\mathbf{C}^{\mathrm{H}}\mathbf{C}=\begin{bmatrix} 1 & g\,\mathrm{e}^{\mathrm{j}k\Delta l} \\ g\,\mathrm{e}^{\mathrm{j}k\Delta l} & 1 \end{bmatrix}\begin{bmatrix} 1 & g\,\mathrm{e}^{-\mathrm{j}k\Delta l} \\ g\,\mathrm{e}^{-\mathrm{j}k\Delta l} & 1 \end{bmatrix}=\begin{bmatrix} 1+g^2 & g(\mathrm{e}^{\mathrm{j}k\Delta l}+\mathrm{e}^{-\mathrm{j}k\Delta l}) \\ g(\mathrm{e}^{\mathrm{j}k\Delta l}+\mathrm{e}^{-\mathrm{j}k\Delta l}) & 1+g^2 \end{bmatrix} \tag{B 5}$$

The eigenvalues of $\mathbf{C}^{\mathrm{H}}\mathbf{C}$ are given by

$$\lambda_{1,2} = (1+g\,\mathrm{e}^{\mathrm{j}k\Delta l})(1+g\,\mathrm{e}^{-\mathrm{j}k\Delta l}),(1-g\,\mathrm{e}^{\mathrm{j}k\Delta l})(1-g\,\mathrm{e}^{-\mathrm{j}k\Delta l})$$

(B 6)

Therefore, the singular values of **C** are given by

$$\sigma_{1,2} = \sqrt{\lambda_{1,2}} = \sqrt{(1+g\,\mathrm{e}^{\mathrm{j}k\Delta l})(1+g\,\mathrm{e}^{-\mathrm{j}k\Delta l})},\sqrt{(1-g\,\mathrm{e}^{\mathrm{j}k\Delta l})(1-g\,\mathrm{e}^{-\mathrm{j}k\Delta l})}$$

(B 7)

It is possible to find an orthonormal set of eigenvectors $\mathbf{x}_i$ such that

$$\mathbf{C}^{\mathrm{H}}\mathbf{C}\mathbf{x}_i = \sigma_i^2\mathbf{x}_i$$

(B 8)

Therefore,

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

(B 9)

The vectors comprising **U** are given by

$$\mathbf{y}_i = \frac{1}{\sigma_i}\mathbf{C}\mathbf{x}_i$$

(B 10)

Hence

$$\mathbf{U}=\frac{1}{\sqrt{2}}\begin{bmatrix}\sqrt{\dfrac{1+g\,\mathrm{e}^{-\mathrm{j}k\Delta l}}{1+g\,\mathrm{e}^{\mathrm{j}k\Delta l}}} & \sqrt{\dfrac{1-g\,\mathrm{e}^{-\mathrm{j}k\Delta l}}{1-g\,\mathrm{e}^{\mathrm{j}k\Delta l}}}\\ \sqrt{\dfrac{1+g\,\mathrm{e}^{-\mathrm{j}k\Delta l}}{1+g\,\mathrm{e}^{\mathrm{j}k\Delta l}}} & -\sqrt{\dfrac{1-g\,\mathrm{e}^{-\mathrm{j}k\Delta l}}{1-g\,\mathrm{e}^{\mathrm{j}k\Delta l}}}\end{bmatrix}$$

(B 11)

The singular value decomposition of **C** may therefore be written as

$$\begin{aligned}\mathbf{C}&=\mathbf{U\Sigma V}^{\mathrm{H}}\\&=\frac{1}{\sqrt{2}}\begin{bmatrix}\sqrt{\dfrac{1+g\mathrm{e}^{-\mathrm{j}k\Delta l}}{1+g\mathrm{e}^{\mathrm{j}k\Delta l}}} & \sqrt{\dfrac{1-g\mathrm{e}^{-\mathrm{j}k\Delta l}}{1-g\mathrm{e}^{\mathrm{j}k\Delta l}}}\\ \sqrt{\dfrac{1+g\mathrm{e}^{-\mathrm{j}k\Delta l}}{1+g\mathrm{e}^{\mathrm{j}k\Delta l}}} & -\sqrt{\dfrac{1-g\mathrm{e}^{-\mathrm{j}k\Delta l}}{1-g\mathrm{e}^{\mathrm{j}k\Delta l}}}\end{bmatrix}\begin{bmatrix}\sqrt{(1+g\,\mathrm{e}^{\mathrm{j}k\Delta l})(1+g\,\mathrm{e}^{-\mathrm{j}k\Delta l})} & 0\\ 0 & \sqrt{(1-g\,\mathrm{e}^{\mathrm{j}k\Delta l})(1-g\,\mathrm{e}^{-\mathrm{j}k\Delta l})}\end{bmatrix}\begin{bmatrix}\dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{2}}\\ \dfrac{1}{\sqrt{2}} & -\dfrac{1}{\sqrt{2}}\end{bmatrix}\end{aligned}$$

(B 12)

Note that

$$\mathbf{C}^{-1}=\left[\mathbf{U\Sigma V}^{\mathrm{H}}\right]^{-1}=\mathbf{V\Sigma}^{-1}\mathbf{U}^{\mathrm{H}}$$

(B 13)

since $\mathbf{V}^{\mathrm{H}}\mathbf{V}=\mathbf{I}$, $\left[\mathbf{V}^{\mathrm{H}}\right]^{-1}=\mathbf{V}$, $\mathbf{U}^{\mathrm{H}}\mathbf{U}=\mathbf{I}$ and $\mathbf{U}^{-1}=\mathbf{U}^{\mathrm{H}}$. Therefore,

$$
\begin{aligned}
\mathbf{H} &= \mathbf{C}^{-1} \\
&= \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^{\mathrm{H}} \\
&= \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}
\begin{bmatrix} \frac{1}{\sqrt{(1+g\,\mathrm{e}^{\mathrm{j}k\Delta l})(1+g\,\mathrm{e}^{-\mathrm{j}k\Delta l})}} & 0 \\ 0 & \frac{1}{\sqrt{(1-g\,\mathrm{e}^{\mathrm{j}k\Delta l})(1-g\,\mathrm{e}^{-\mathrm{j}k\Delta l})}} \end{bmatrix}
\frac{1}{\sqrt{2}}
\begin{bmatrix} \sqrt{\frac{1+g\,\mathrm{e}^{\mathrm{j}k\Delta l}}{1+g\,\mathrm{e}^{-\mathrm{j}k\Delta l}}} & \sqrt{\frac{1+g\,\mathrm{e}^{\mathrm{j}k\Delta l}}{1+g\,\mathrm{e}^{-\mathrm{j}k\Delta l}}} \\ \sqrt{\frac{1-g\,\mathrm{e}^{\mathrm{j}k\Delta l}}{1-g\,\mathrm{e}^{-\mathrm{j}k\Delta l}}} & -\sqrt{\frac{1-g\,\mathrm{e}^{\mathrm{j}k\Delta l}}{1-g\,\mathrm{e}^{-\mathrm{j}k\Delta l}}} \end{bmatrix}
\end{aligned}
\tag{B 14}
$$

Hence

$$
\begin{aligned}
\sigma_i &= \frac{1}{\sqrt{\left(1+g\mathrm{e}^{-\mathrm{j}k\Delta r\sin\theta}\right)\left(1+g\mathrm{e}^{\mathrm{j}k\Delta r\sin\theta}\right)}} \\
\sigma_o &= \frac{1}{\sqrt{\left(1-g\mathrm{e}^{-\mathrm{j}k\Delta r\sin\theta}\right)\left(1-g\mathrm{e}^{\mathrm{j}k\Delta r\sin\theta}\right)}}
\end{aligned}
\tag{B 15}
$$

are the singular values of the inverse filter matrix **H**.

References

[1] B. B. BAUER, 1963, Some techniques toward better stereophonic perspective, Institute of Electrical and Electronics Engineers Transactions on audio, Part I

[2] H. A. M. CLARK, G. F. DUTTON and P. B. VANDENYN, 1958, The stereosonic recording and reproducing system, Journal of the Audio Engineering Society, Vol. 6, No. 2, 102-115

[3] A. D. BLUMLEIN, 1958, British patent specification 394.325, Journal of Audio Engineering Society, Vol. 6, No. 2, 91-130

[4] J. C. STEINBERG and W. B. SNOW, 1934, Auditory perspective - Physical factors, Bell System Technical Journal, 245-258

[5] M. A. Gerzon, "Ambisonics in Multichannel Broadcasting and Video", J. Audio Eng. Soc., **33** (11), 859-871 (1985)

[6] G. STEINKE, W. AHRNERT, P. FELS, W. HOEG and F. STEFFEN,1987,True directional sound system oriented to the original sound and diffuse sound structures - New applications of the Delta Stereophony System (DSS),Journal of Audio Engineering Society, 35 380

[7] P. H. HERINGA, B. H. M. KOK and Y. DEKEYREL, 1987, The acoustics and sound system for hemispherical film projection, Journal of Audio Engineering Society, 35(3), 119-128

[8] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic Control by Wave Field Synthesis," J. Acoust. Soc. Am. **93**, 2764-2778 (1993).

[9] P. BOER and R. VERMEULEN, 1939, Philips techn. Rdsch. 4, 329-332

[10] P. DAMASKE and V. MELLERT, 1969, Acoustica 22, 153-162

[11] J. Blauert, "Spatial Hearing: The Psychophysics of Human Sound Localization," (MIT Press, Cambridge, MA, 1997), Chap. 2, pp. 50-136.

[12] H. Møller, "Fundamentals of Binaural Technology," Appl. Acoust. 36, 171-218 (1992).

[13] D. R. Begault, 3-D Sound for Virtual Reality and Multimedia (AP Professional, Cambridge, MA, 1994).

[14] B. S. Atal and M. R. Schroeder 1962 U.S. Patent 3,236,949. Apparent sound source translator.

[15] M. R. Schroeder, B. S. Atal, "Computer Simulation of Sound Transmission in Rooms," IEEE Intercon. Rec. Pt7, 150-155 (1963).

[16] P. Damaske, "Head-related Two-channel Stereophony with Reproduction," J. Acoust. Soc. Am. **50**, 1109-1115 (1971).

[17] H. Hamada, N. Ikeshoji, Y. Ogura And T. Miura, "Relation between Physical Characteristics of Orthostereophonic System and Horizontal Plane Localisation," Journal of the Acoustical Society of Japan, (E) **6**, 143-154, (1985).

[18] G. Neu, E. Mommerts and A Schmitz 1992 Acoustica 76, 183-192. Investigations of true directional sound reproduction by playing head referred recordings over two loudspeakers: part I

[19] G. Urbach, E. Mommerts and A Schmitz 1992 Acoustica 77, 153-161. Investigations on the directional scattering of sound reflections from the playback of head referred recordings over two loudspeakers: part II

[20] P. A. NELSON, H. HAMADA AND S. J. ELLIOTT, 1992, Adaptive inverse filters for stereophonic sound reproduction, *IEEE Transactions on Signal Processing*, **40** (7), 1621-1632.

[21] L. A. Jeffress, "A Place Theory of Sound Localization," J. Comp. Physiol. Psychol. **41**, 35-39 (1948).

[22] H. S. Colburn, "Theory of Binaural Interaction Based on Auditory-Nerve Data. I. General Strategy and Preliminary Results on Interaural Discrimination," J. Acoust. Soc. Am. **54**, 1458-1470 (1973).

[23] G. B. Hanning, "Detectability of Interaural Delay in High-frequency complex waveforms," J. Acoust. Soc. Am. **55**, 84-90 (1974).

[24] J. C. Middlebrooks, and D. M. Green, "Directional Dependence of Interaural Envelope Delays," J. Acoust. Soc. Am. **87**, 2149-2162 (1990).

[25] R. G. Klump, and H. R. Eady, "Some Measurements of Interaural Time Difference Thresholds," J. Acoust. Soc. Am. **28**, 859-860 (1956).

[26] R. A. Butler, and R. Flannery, "The Spatial Attributes of Stimulus Frequency and Their Role in Monaural Localisation of Sound in the Horizontal Plane," Percept. psychophys. **28**, 449-457 (1980).

[27] C. Lim, and R. O. Duda, "Estimating the Azimuth and Elevation of a Sound Source from the Output of a Cochlea Model," Proc. Twenty-eighth Annual Asilomer Conference on Signals, Systems and Computers (IEEE, Asilomar, CA), 399-403 (1994).

[28] P.A. Nelson, F. Orduna-Bustamante, and H. Hamada, "Inverse Filter Design and Equalisation Zones in Multi-Channel Sound Reproduction," IEEE Trans. Speech Audio Process. **3**(3), 185-192 (1995).

[29] O. Kirkeby, P. A. Nelson, F. Orduna-Bustamante, and H. Hamada, "Local Sound Field Reproduction Using Digital Signal Processing," J. Acoust. Soc. Am. **100**, 1584-1593 (1996).

[30] P. A. NELSON, F. ORDUNA-BUSTAMANTE AND H. HAMADA, 1996, Multi-channel signal processing techniques in the reproduction of sound, *Journal of Audio Engineering Society*, Vol. **44**, No. **11**, 973-989.

[31] O. KIRKEBY, P. A. NELSON, H. HAMADA AND F. ORDUNA-BUSTAMANTE, 1996, Fast deconvolution of multi-channel systems using regularisation, *ISVR Technical Report,* No. **255**, University of Southampton.

[32] P. A. Nelson and S. J. Elliott, Active Control of Sound (Academic Press, 1992)

[33] O. Kirkeby, P. A. Nelson, and H. Hamada, "Stereo Dipole," UK Patent Application, 9603236.2, 1996.

[34] O. Kirkeby, P. A. Nelson, and H. Hamada, "Local Sound Field Reproduction Using Two Closely Spaced Loudspeakers," J. Acoust. Soc. Am. **104** (4), 1973-1981 (1998).

[35] J. L. Bauck and D. H. Cooper, "Generalized Transaural Stereo and Applications," Journal of the Audio Engineering Society **44** (9), 683-705 (1996).

[36] Fr. Heegaard 1958 EBU Rev. pt A – Technical, No.52, 2-6. (Reprinted in Journal of the Audio Engineering Society 40, 692-705, 1992)

[37] U. Burandt, C. Poesselt, S. Ambrozus, M. Hosenfeld and V. Knauff, "Anthropometric contribution to standardising manikins for artificial head microphones and to measuring headphones and ear protectors," Applied Ergonomics, 22.6, 373-378 (1991)

[38] B. Gardner, and K. Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone," MIT Media Lab Perceptual Computing - Technical Report No. 280 (1994).

[39] E. M. Wenzel, M. Arruda, D. J. Kistler and F. L. Wightman, "Localisation using nonindividualized head-related transfer functions," J. Acoust. Soc. Am. 94(1), 111-123 (1993)

[40] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-Related Transfer Functions on Human Subjects," J.Audio Eng. Soc., 43, 300-321 (1995)

[41] H. Moller, M. F. Sorensen, C. B. Jensen and D. Hmmershoi, "Binaural technique: Do we need individual recordings?" J. Audio Eng. Soc., vol. 44, pp.451-469 (1996).

[42] E. VILLCHUR and M. C. LILLION, 1975, Probe-tube microphone assembly, Journal of Acoustic Society of America, Vol. 57, No. 1, January

[43] E J WILLIAMS, 1959, Regression Analysis, Wiley Publications in Statistics, New York

[44] T. C. T. Yin, and J. C. Chan, "Interaural Time Sensitivity in Medial Superior Olive of Cat," J. Neurophysiol. **64**, 465-488 (1990).

[45] T. R. Stanford, S. Kuwada, and R. Batra, "A Comparison of the Interaural Time Sensitivity of Neurones in the Inferior Colliculus and Thalamus of the Unanesthetized Rabbit," J. Neurophysiol. **12**, 3200-3216 (1992).

[46] T. N. Buell, C. Trahiotis, and L. R. Bernstein, “Lateralization of Low-Frequency Tones: Relative Potency of Gating and Ongoing Interaural Delays,” J. Acoust. Soc. Am. **90**, 3077-3085 (1991).

[47] F. L. Wightman, and D. J. Kistler, “The Dominant Role of Low-frequency Interaural Time Differences in Sound Localization,” J. Acoust. Soc. Am. **91**, 1648-1661 (1992).

[48] B C J MOORE, 1995, An Introduction to the Psychology of Hearing, Third Edition, Academic Press, London

[49] S. Barnett, Matrices - Methods and Applications (Oxford University Press, Oxford, 1990), Chap. 8, pp. 218-225.

[50] F. Asano, Y. Suzuki, and T. Sone, “Sound equalization using derivative constraints,” Acustica, 82, 311-320 (1996).

[51] B. Rakerd and W. M. Hartmann, “Localization of sound in rooms, II: The effects of a single reflecting surface,” J. Acoust. Soc. Am 78(2), 524-533 (1985)

[52] M Teschl, “Binaural Sound Reproduction via Distributed Loudspeaker Systems,” Diplomarbeit, Universitaet fuer Musik und darstellende Kunst Graz (2000)

[53] A. V. Oppenheim and R. W. Schafer, “Digital Signal Processing,” (Prentice/Hall International, Inc., London, 1975), Chap. 7, pp. 337-375.

[54] F. B. Hildebrand, “Advanced Calculus with Applications,” (Prrentice-Hall Inc., Englewood Cliffs, N.J.,1962)

[55] D. B. Ward and G W Elko, “Effect of Loudspeaker Position on the Robustness of Acoustic Crosstalk Cancellation”, pp.106-108, IEEE Signal Processing Letters, 6(5), (1999)

[56] J. Bauck, “A simple loudspeaker array and associated crosstalk canceller for improved 3D audio”, J. Audio Eng. Soc., vol. 49, pp.3-13 (2001).

Bibliography

[57] O. KIRKEBY, 1995, Reproduction of acoustic fields, PhD Dissertation, ISVR, University of Southampton, England

[58] F. ORDUNA-BUSTANAMTE, 1995, Digital signal processing for multi-channel sound reproduction., PhD Thesis (University of Southampton)