UNIVERSITY OF SOUTHAMPTON

Visually adaptive virtual acoustic imaging

By

John Frederick William Rose

Doctor of Philosophy

FACULTY OF ENGINEERING AND APPLIED SCIENCE

INSTITUTE OF SOUND AND VIBRATION RESEARCH

March 2004

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND APPLIED SCIENCE

INSTITUTE OF SOUND AND VIBRATION RESEARCH

Doctor of Philosophy

VISUALLY ADAPTIVE VIRTUAL ACOUSTIC IMAGING

By John Frederick William Rose

Virtual acoustic imaging systems give listeners the perception of sound images at locations where no sound sources exist. They achieve this with filters whose design incorporates the paths' transfer functions from the sound sources to the listener's ears. The success of a virtual acoustic imaging system depends greatly on how close the listener is to the intended listener location. The small area of acceptable listener locations is a major limitation of virtual acoustic imaging systems. Adaptable virtual acoustic imaging systems greatly enhance the usefulness of the system and a listener's enjoyment by modifying the intended listener location by adapting the virtual acoustic imaging filters as the listener location changes. The work in this thesis looks at the performance of an adaptable virtual acoustic imaging system that utilises a video head tracking procedure to track the listener's movements. Computer simulations and subjective evaluations consider the size of the 'sweet spot' at a variety of listener locations. Dynamic subjective tests also consider the update rate of the filters with and without incorporation of a simple video head-tracking algorithm. The computer simulations show that due to the change in geometry, listener head locations further from the inter-source axis result in a higher 'sweet spot' frequency range. At asymmetric arrangements, head shadowing decreases the robustness of the system at the contra-lateral ear and increases the robustness at the ipsi-lateral ear. The most important parameter affecting the 'sweet spot' is how closely the virtual acoustic image location corresponds to the system's loudspeakers. Listener movements of about every 3 cm or less seem to require a filter update for image stability. The image processing computation time significantly affects the perceived stability of a virtual acoustic image.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisors, Phillip Nelson and Boaz Rafaely, for their great encouragement and guidance of this work. I would also like to thank Takashi Takeuchi for many valuable insights and suggestions. This thesis is dedicated to my parents.

1. INTRODUCTION

1.1. The Problem

A virtual acoustical imaging system is capable of giving a listener the perception of a sound image at a location where no actual sound source exists. These types of systems may utilise headphones or loudspeakers in order to deliver the sound signals to the listener's ears that approximate the sound signals that would have been produced by a sound source at the perceived virtual acoustic image location. Loudspeaker virtual acoustic imaging systems are less obtrusive to a listener but are limited by the extent of the spatial regions where the system controls the sound field. Should the listener's ears move out of the control regions or alternatively should the listener move out of the 'sweet spot', the desired listener perception of the virtual acoustic image would not be achieved.

Headphone delivery systems have been equipped with magnetic head trackers and a database of pre-designed virtual acoustic imaging filters that are selected in real time in order to adapt for listener movements [1]. Headphones avoid the response of the room affecting the listening experience. They also provide a way to control exactly what the signals are heard by the listener. With loudspeaker systems, both of the listener's ears are capable of hearing the output from all of the system's loudspeakers. An undesirable aspect of headphone systems is their intrusiveness. Magnetic head trackers also require the listener to affix a sensor somewhere on the listener's head.

Garas [2] proposes an adaptive loudspeaker virtual audio system without the need for a head tracking device. The system utilises microphones placed in the listener's ears and adaptive filtering. The spectrum content of the source signal affects the convergence time of the adaptive filters so a separate training signal may be added to the loudspeaker output with a flat spectrum over a wide frequency band. This reduces the convergence time but introduces an unwanted audible sound capable of being heard by the listener. This approach is also intrusive, requiring the listener to place microphones in their ears.

It is possible to avoid the inconvenience of requiring the listeners to equip themselves before undergoing the listening experience with the use of loudspeakers and a video head-tracking scheme. Kyriakakis et al. [3] have developed such a system that tracks listener movement and then modifies the loudspeaker output based on the listener's location. A simple time delay is adjusted to account for the head's location. Time delay is an important sound localisation cue for detecting horizontal location of sounds containing low frequencies. This system is limited by not adjusting for other important sound localisation cues such as sound level difference and spectral cues. These cues are important in localisation of middle and high frequency sounds and spectral cues are important in determining vertical location.

Gardner [4] has developed a visually adaptive loudspeaker virtual acoustic imaging system that adjusts for time difference, level difference, and spectral cues through the use of measured dummy head related transfer functions (HRTFs). The system is set up to track and adjust the acoustics for head rotations. The performance of the system for lateral head motions that position the listener at asymmetrical arrangements are not accounted for.

The work in this thesis is concerned with the development of a visually adaptive virtual acoustic imaging system that utilises two loudspeakers. The system adjusts for lateral head motions. This is an extension of work presented in an MSc dissertation [5]. The system tracks the head unobtrusively with a video camera and a simple image-processing algorithm. The system utilises the head location information to select appropriate pre-designed virtual acoustic imaging filters that correspond to the listener's head location. Figure 1.1 displays a sketch of the overall system.

The main limitation of the system developed for the MSc was inadequate knowledge of the acoustics and psychoacoustics of the system at laterally displaced asymmetric head locations. An emphasis in this thesis is on the acoustics and system performance at non-traditional asymmetric listener locations. It is realised that modification of the acoustic imaging filters for every sample would require extremely powerful

computation ability and this type of continuous adaptation would be impractical. An emphasis in this work is then to consider how frequently the filters must be updated in order to deliver the desired perception to a moving listener. The original contribution of this thesis includes detailed examination of the performance of a virtual acoustic imaging system at asymmetric listener locations. This evaluation is carried out both subjectively and through computer simulations.

Chapter 2 presents a review of important relevant topics such as human sound localisation, binaural technology, and the modelling and measuring of the acoustic path responses. In order to design the filters one needs to know the response of the acoustic paths from the virtual acoustic imaging system's loudspeakers to the listener's ears (plant responses) and the response of the acoustic paths from the virtual sound source location to the listener's ears. The path responses themselves can contain the information about the response of the acoustic environment including reflections off the listener, the listening room, and objects within the room. The approach in this thesis generally assumes an anechoic environment but may include the effect of reflections off the listener. A virtual acoustic imaging system with virtual acoustic imaging filters designed with anechoic path responses still give listeners the perception of virtual acoustic images in non-anechoic environments. This has been commonly noticed but the best listening environment for such a system is in an anechoic environment. The degree of degradation of the performance of the system in non-anechoic environments is not great. The effect of the room acoustics on the performance of the virtual acoustic imaging system is a current area of research and not considered in this work. Here the consideration is a system capable of adapting for listener movements.

Chapter 3 provides a review describing the particular virtual acoustic imaging system that this thesis examines together with a justification for its geometry. Chapter 3 also provides a review of the filter design procedure. The filters are the convolution of the response of the virtual acoustic paths with the inverse of the plant's acoustic path responses. There are different approaches to the design of the inverse filters including a frequency domain technique that utilises regularisation. This method was adopted

for this work. Besides reviewing the system arrangement and filter design technique, chapter 3 also presents some new analysis of the system performance at asymmetric listener locations including different concepts that affect the robustness of the system.

A major focus of the work is in gaining an understanding of the required rate of update of the virtual acoustic imaging filters. To this end, chapter 4 simulates and analyses the information available at the listener's ears at different listener positions and with different sets of virtual acoustic imaging filters. The simulations predict the "sweet spot" size and assess limiting aspects of the system at a range of listener locations. Chapters 5 and 6 present subjective evaluations of the "sweet spot" size and the required update rate of the virtual acoustic imaging filters respectively. Chapter 7 presents another dynamic subjective evaluation of the system including the incorporation of the video head-tracking image processing algorithm.

Figure 1.2 depicts some of the geometric variables used throughout this thesis. The figure shows two (2) sources or loudspeakers and two (2) receivers or ears and denotes them as left and right or correspondingly one (1) and two (2). Also shown is a virtual sound source with dashed lines. The inter-source and the inter-receiver axes run parallel to each other and lie half way between the sources and receivers respectively. These axes are separated by a distance x. The inter-receiver axis to the right of the inter-source axis corresponds to a positive x value. The inter-receiver axis to the left of the inter-source axis corresponds to a negative x value. This distance is the lateral head translation of the listener from the inter-source axis. This is the main type of listener motion movement considered in this thesis. The source span 2θ is the subtended angle between the two sources and the listener when the listener is at the symmetric listener location (i.e. $x = 0$). The figure also shows an angle between the inter-receiver axis and a line connecting a point half way between the receivers and a point half way between the sources θ_{ls}. The virtual acoustic image angle θ_v is the angle between the inter-receiver axis and the location of the virtual acoustic image.

1.2. Contributions of the Thesis

- An investigation has been undertaken of the effect of lateral head translations on the performance of a virtual acoustic imaging system with an emphasis on the robustness of the system to such movements. The different methods used to explore the robustness of the system all resulted in similar conclusions regarding the degree of robustness.

- The relative position of the virtual sound image with respect to the position of the real sound sources of the virtual acoustic imaging system was found to be the most salient factor influencing the robustness of the system. The system was found to be more robust the closer the virtual acoustic image coincided with the position of the real sound sources of the system. This was shown to be the case with computer simulations that considered different localisation cues as well as with subjective evaluations of the system undertaken in a variety of conditions.

- Asymmetric, off-axis, sound-field computer simulations of a virtual acoustic imaging system were undertaken with an emphasis on evaluating the degree of success in delivering localisation cues and on the robustness of the cues to lateral head motions.

- Asymmetric, off-axis, subjective evaluations of a virtual acoustic imaging system were undertaken with an emphasis on evaluating the robustness of the system to lateral head translations.

- Dynamic, off-axis, subjective evaluations of an adaptive virtual acoustic imaging system were undertaken with the objective of finding the frequency with which the audio filters required updating in order to achieve an adequate rate of adaptation of the system to lateral head motion.

- Subjective examinations were undertaken of a visually adaptive virtual acoustic imaging system that utilised a video camera and a video head-tracking algorithm with an emphasis on finding the affect of the video processing on the performance of the dynamic virtual acoustic imaging system. The inclusion of the image-processing algorithm in the dynamic virtual acoustic imaging system was found to degrade the performance of the system. The efficiency of implementation of the head-tracking algorithm had scope for improvement.

- An image processing head-tracking algorithm was developed that combines template matching with the pattern search adaptive algorithm. This approach was found to be successful at tracking lateral head translations under constrained conditions.

1.3. Publications Relevant to Thesis

J. F. W. Rose, "A Visually Adaptive Virtual Sound Imaging System" (MSc. thesis, Institute of Sound and Vibration Research, Southampton, UK, 1999).

J. Rose, "Variance of sweet spot size with head location for virtual audio," 110th Audio Engineering Society Convention Preprint 5391, 2001.

J. Rose, P. A. Nelson, B. Rafaely, and T. Takeuchi, "A study of virtual acoustic imaging systems for asymmetric listener locations," ISVR Technical Report No. 295, November 2001.

J. Rose, P. A. Nelson, B. Rafaely, and T. Takeuchi, "Sweet spot size of virtual acoustic imaging systems at asymmetric listener locations," J. Acoust. Soc. Am. 112, 1992-2002 (2002).

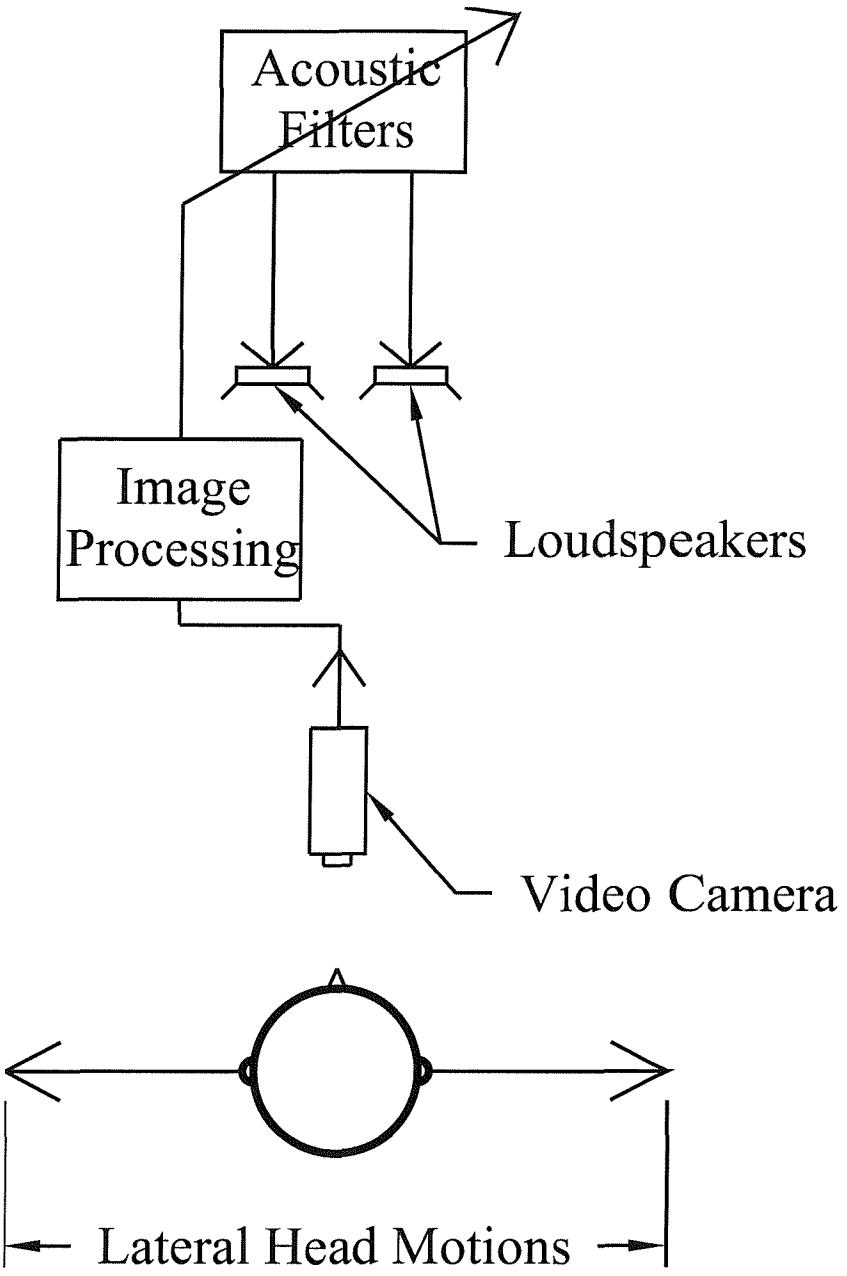


Fig. 1.1 *Depiction of the main elements of the visually adaptive virtual acoustic system that adapts to lateral head movements by the listener.*

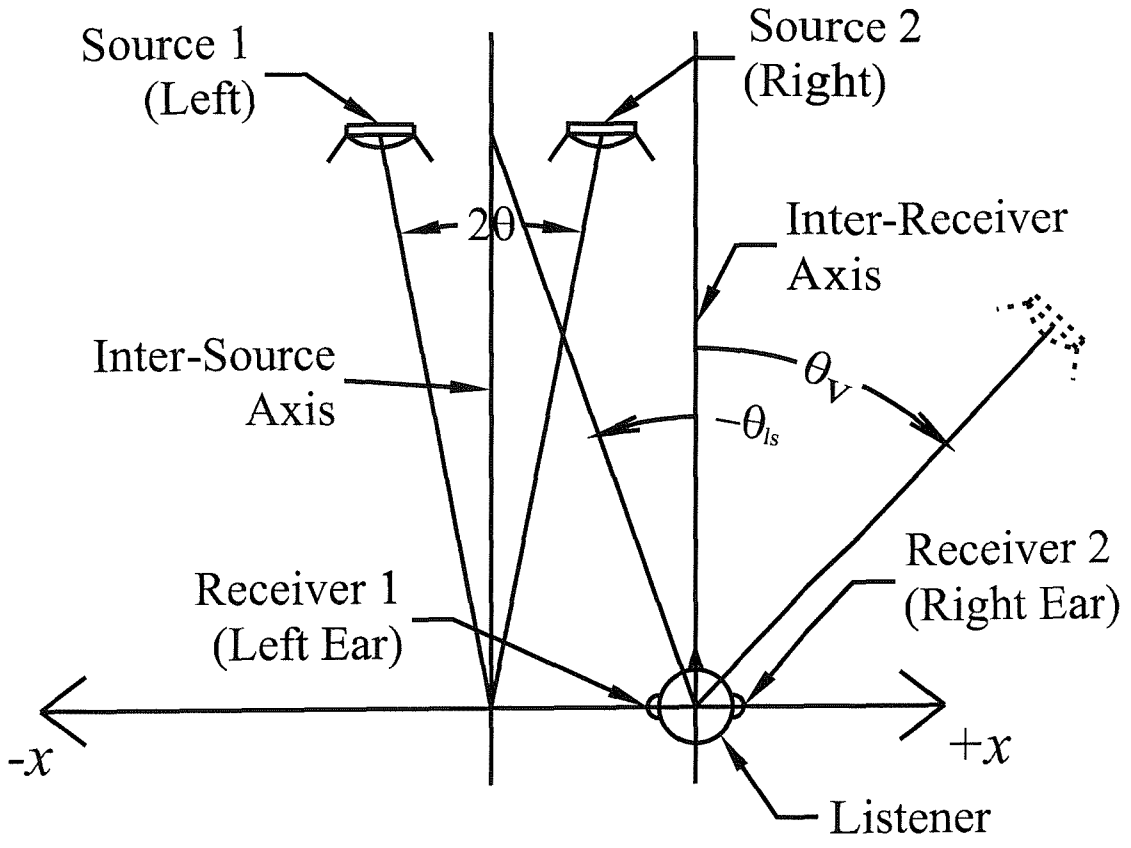


Fig. 1.2 *The definition of geometric variables used throughout thesis.*

2. BACKGROUND

2.1. Introduction

This chapter reviews background concepts relevant to the thesis work. Material relevant to the thesis includes human sound localisation, binaural technology, models of acoustic transfer functions, and the methods used in order to compute sound localisation cues. These concepts serve as a basis for understanding virtual acoustic imaging systems and the reasoning behind the approaches taken in order to simulate and evaluate their performance.

2.2. Human Sound Localisation

This section reviews human sound localisation cues. Threshold detection values provide insight into human sound localisation ability and imply the degree of precision required in a computer model of human sound localisation. Other reviews of sound localisation cues can be found in references [1,6-8].

2.2.1. Interaural cues

A dominant focus of spatial hearing research is on the differences between the two ears' sound signals due to the horizontal source location. A sound source not located on the median plane (the plane halving the listener's head and located an equal distance between the ears) produces a sound signal that arrives at the ipsilateral ear (ear closest to the source) first and then at a short time later a slightly scaled version arrives at the contralateral ear (ear furthest from the source). The name given to the delay of arrival time between the two signals is the interaural time difference (ITD). The name given to the difference in level of the two arriving signals is the interaural level difference (ILD).

Interaural level difference is due to the shadowing effect of the head, which depends very much on the wavelength of the sound. At frequencies that have wavelengths larger than the width of the head, diffraction causes ILD to be very small. Natural sounds below 500 Hz rarely create a large enough ILD that can be utilised in

localisation, but above 3 kHz ILD cues are generally large and reliable [6]. The smallest just noticeable difference in ILD is less than 1 dB and sensitivity to ILD is almost constant as a function of frequency in the band of frequencies from 250 Hz to 10 kHz [9,10].

Humans utilise the ITD cue at frequencies below about 1.5 kHz where the wavelength is greater than the path length difference between the two ears. Above approximately 1.5 kHz, the ITD cue becomes ambiguous because this situation corresponds to a shift of more than a single period of the sound wave. Therefore, the decrease in human sensitivity to ITD above about 1 kHz avoids much confusion and sound localisation errors. For complex signals containing higher frequencies, humans can utilise the ITD of the envelope of the signal [11-14]. There is also evidence that for wide-band stimuli that include low frequencies, the low frequency ITD is the dominant cue utilised in localisation affecting location perception much more than the higher frequency ILD cue [15]. ITD just noticeable difference threshold values for band-passed random noise (150-1700 Hz) and a 1 kHz tone are 9 μs and 11 μs respectively [16].

The "duplex theory" of sound localisation refers to the utilisation of ILD at high frequencies and ITD at low frequencies in order to localise sound sources along the horizontal plane. Lord Rayleigh made the first observations of these cues preceding more than a hundred years of their extensive study [17].

Interaural cues assist in azimuthal localisation of sound sources but do not help to determine elevation. Sources directly in front, behind or exactly above the listener all produce the same interaural cues (i.e. zero (0) ITD and zero (0) ILD). In fact, any source of sound positioned on the median plane produce zero (0) ITD and zero (0) ILD. Sound sources located anywhere on a spatial "cone of confusion" that is concentric with the axis connecting the two listener's ears and has an apex half way in-between the ears produce constant ITD and constant ILD interaural cues. Figure 2.1 displays the median plane and one "cone of confusion". In reality, the locations

that produce constant ITD and ILD are not perfect cones because human heads are not perfectly symmetric. The three-dimensional surfaces that produce ITD and ILD for real human heads are approximately cone-like.

2.2.2. Spectral cues

Spectral cues help to differentiate between sound source locations on cones of confusion. The spectrum of the sound that arrives at the eardrum depends on the sound's path of propagation. This includes reflection paths off the listener's head, pinnae, and torso. Reflections add delayed versions of the signal resulting from the direct path. These delayed versions of the signal may contribute constructively or destructively to the frequency spectrum, which create peaks or notches in the overall signal heard by the listener. The effect of pinna folds is dramatic on frequencies above 4 kHz and is dependent on the sound source position including its elevation. The effect of a pinna is that of a directional dependent linear filter, which is unique to every individual, but most spectral filtering characteristics of pinnae share some general features.

Of particular importance is a pinna notch in the frequency spectrum that increases from approximately 4 kHz to 10 kHz for frontal sound sources increasing in elevation. Reflections of frontal sounds off the posterior concha wall add destructively to the direct sound at the notch's frequency. The shape of the concha results in a higher frequency pinna notch for higher elevated sound sources. There is evidence that the pinna notch is an important cue utilised in determining elevation of a sound source [18].

2.2.3. Dynamic cues

Dynamic cues also help listeners disambiguate the sound source location from the locus of possible locations on a cone of confusion. Head motion causes dynamic changes in the spectral and interaural cues, which can assist in the localisation task given sufficient stimulus duration. Dynamic subjective experiments that examine the importance of dynamic cues for sound localisation tend to compare listeners' abilities

to localise stationary sound sources under stationary and dynamic head conditions. These studies show that localisation of both real [19] and virtual [20] sound sources improve with the presence of dynamic cues.

Dynamic cues are also present for stationary listeners when the sound source is in motion. Most of the research on the detection of sound source motion is restricted to examination of horizontal movements. Asking the usual question of "Is motion present?" under conditions of real moving sources or simulated auditory motion leads to determination of the minimum thresholds in which listeners detect motion. This work largely culminates in finding minimum audible angles (MAA), minimum audible movement angles (MAMA), or occasionally marked endpoint (ME) thresholds. These studies give insight into human auditory motion detection. Physiological studies on animals also provide insight into the possible workings of the auditory system in this regard [21].

The deduction of MAA thresholds employs two real stationary sound sources that simulate motion. Two stationary sound sources can simulate auditory motion by emitting two correlated sound pulses in rapid succession from one source and then the other source. The listener perceives the sound moving from the position of first sound source to the position of the second sound source. The variables varied in these studies are the type and duration of sound signal, the interval of time between the two pulses of sound, and the positions of the sound sources. The MAA is the smallest angle made by the two sources with the subject at the apex in which the subjects still perceive auditory motion. MAAs are dependent on the above-mentioned variables. Keeping all other variables the same and reducing the angle smaller than the MAA will result in the listener perceiving a single stationary sound source. The smallest MAA threshold is about 1° for sounds in front of the subject [22].

The experimental situations in which MAA thresholds are determined are "unnatural" in that they do not normally occur in normal listening situations. Two stationary sound sources do not usually give someone the perception of motion unless done

intentionally with an audio system. In an effort to quantify human auditory motion capabilities under "natural" conditions, the deduction of MAMA thresholds employs one real moving sound source that presents real motion. The sound source rotates about the listener and emits a short burst of sound. Elimination of any noise produced by the rotation, such as air noise, is important so as not to influence the subjects' perception. The MAMA is the smallest angle made by the positions of the sound source at the onset and offset of the pulse of sound with the subject at the apex in which the subject still detects auditory motion. Variables in this study are the speed of rotation, the radius of rotation, the type and duration of the sound signal, and the position of the sound source at onset. The listener not only receives the onset and offset auditory information but also the dynamic information as the sound source moves between the onset and offset source positions. Slow-moving sound sources in front of the subject achieve the smallest MAMA thresholds but not smaller than the best MAA threshold [23]. This suggests that human auditory spatial resolution is superior under static conditions.

To test the importance of the dynamic information, the deduction of ME thresholds employs one real moving sound source that simulates motion. This sound source emits two short bursts of correlated sound in rapid succession while rotating about the listener. The ME is the smallest angle made by the positions of the sound source at onset of the first burst and at offset of the last burst with the listener at the apex, which the listener can still perceive the sensation of auditory motion. This scheme eliminates most of the dynamic information available to the subject. ME thresholds slightly exceed MAMA thresholds for slower targets. This perhaps implies that the human auditory system detects motion by utilising information collected throughout the trajectory of slower moving sound sources and does not just extract spatial samples (snapshots) at the movement endpoints [24].

2.2.4. Cross-modal and cognitive effects

Interaction of audition with other modes of perception such as vision, conduction of sound through the skull, inertial or vestibular cues, tactile sensations, as well as memory, and associations also affect our experience of sound sources. Conflicting

cues can act as a source of noise to our auditory system and change our perception of sound quality, location, or make sounds more or less noticeable. Correlated cues from different perceptual modalities can enhance our experience of simulated environments and increase "realism". This section briefly reviews research on the influence of the presence of stimuli from other modalities on auditory perception. Begault [25] and Blauert [7] overview cross-modal effects more comprehensively.

Presenting an auditory stimulus with a matching visual counterpart in synchrony but with conflicting localisation cues often causes the auditory stimulus to be localised at the position of the visual stimulus. This bias toward the visual stimulus holds for horizontal virtual acoustic image angles θ_v up to about $\pm 15\text{-}25^\circ$ (i.e. $|\theta_v| < \sim 25^\circ$) but auditory information dominates at wider angles. The virtual acoustic image angle θ_v is the angle between the inter-receiver axis and the location of the virtual acoustic image (e.g. see Fig. 1.1). Auditory signals are easier to detect with the presence of a correlated visual stimulus. For example, speech intelligibility in the presence of noise improves when the listener is able to watch the speaker's moving lips.

Oscillations of the skull can conduct sound to the inner ear through the temporal bone. In normal listening situations the signals reaching the inner ear through bone conduction are about 40 dB below signals travelling through the outer and middle ears. Therefore, the effect of these signals is negligible in normal hearing situations. However, bone conduction is important for hearing and localising sound underwater. In addition, when the skull is excited directly or when the listener is wearing earplugs, most of the sound experienced is through bone conduction.

Located next to the middle ear, the vestibular organ is key to our perception of our head's position and orientation or our sense of balance. This organ responds mainly to acceleration, gravity, inertia, and centrifugal forces but loud low frequency sound can stimulate it as well. However, vestibular cues do not seem to be of much direct import on perception of auditory events. Dynamic sound localisation cues depend on an understanding of the head position and so the vestibular organ may aid in sound

localisation indirectly. Simulated environments sometimes simulate motion in order to increase "realism" but the objective is proper correlation between vestibular and visual cues.

The range of frequencies between about 20-150 Hz is potentially both audible and tactile. Audition is usually dominant and our tactile sensation's role in auditory localisation is likely to be negligible. The benefit of simulating vibro-acoustic sensation is in enhancing "realism" of the simulated environment. This is important in headphone applications but when using loudspeakers to deliver the auditory sensations the correlated tactile sensations are already present.

The presence of other modes of perception can act as distractions to accurate sound localisation. If undesirable sensations from other modes are unavoidable, it may be better to provide desirable correlated sensations from the other modes to mask the unwanted sensations. Correlated modes can enhance a person's experience of simulated environments. The visual modality in particular can greatly affect sound localisation. Memories also affect perceptions. An inexperienced or first time user may react to simulated sensations very differently to someone with experience. Consideration of these effects is important when designing subjective experiments and developing systems having applications in uncontrolled and varying situations.

2.3. Binaural Technology and Surround Sound

Virtual acoustic imaging systems utilise binaural technology, which is a different technology to that incorporated in the popular surround sound systems. Confusion between these two types of audio reproduction systems exists and so an explanation of their differences seems appropriate.

Surround sound systems are very common in movie theatres and their popularity for home entertainment is increasing. The so-called "5.1" systems have three front channels (right, left, and centre), two rear channels (left and right), and a channel for

low frequency sounds. These systems attempt to create an immersive high quality audio environment. They move audio images between loudspeakers by panning amplitudes and time delays. Stable images to the side of the listener and images outside of the horizontal plane of the loudspeakers are not possible. A practical problem for most people is the lack of space or disposable income to justify purchasing a cumbersome five or six loudspeaker entertainment system.

The idea of binaural technology is to give listeners the perception of an auditory experience by presenting sound signals at the listener's ears that closely approximate the sound signals that would have been present at the listener's ears had she been in the desired real auditory environment. Binaural technology utilises head related transfer functions (HRTF's), which contain interaural and spectral localisation cues. If the binaural system quickly adapts to head motion then dynamic cues may be present as well. It is possible to deliver the binaural signals with either headphones or loudspeakers. Loudspeaker binaural systems meant for a single listener often use two loudspeakers. The advantage of loudspeaker binaural systems is creating virtual audio images in three-dimensional space with only two loudspeakers. A disadvantage is the colouration of the sound by filtering the signals with HRTFs that might not match the HRTFs of the listener. Another disadvantage is a small "sweet spot" size. Adaptive virtual acoustic systems attempt to correct this problem by changing the filtering for the varying head location so that the listener is always in the "sweet spot" [2,4]. Some televisions now incorporate virtual acoustic imaging systems based on the use of two loudspeakers with appropriate filtering. A product based on the Stereo Dipole [26] is also available as an add-on product to a home gaming system.

2.4. Models of the Acoustic Paths to the Ears

This section reviews three models of the acoustic paths from a sound source to the listener's ears that are used is the computer simulations presented in this thesis. These acoustic paths make it possible to predict the sound arriving at the listener's ears and to design the acoustic filters used in virtual acoustic systems.

It is possible to directly measure the path responses on each individual listener [27]. This approach is impractical if there are many different or unknown people that will use a virtual acoustic imaging system. This section reviews three alternative path models. As the models progress in realism, their complexities also increase. Comparing results obtained using the different models makes it possible to attribute certain aspects of the sound field to specific model characteristics. The free field model provides insight into geometrical effects. In addition to geometrical effects, the spherical head model well approximates the shadowing effects of the listener's head. The dummy head related transfer function (HRTF) includes these effects and the effects of the dummy's pinnae, neck, and torso.

2.4.1. Free field approximation

The simplest approximation of the path's responses is to remove the physical head, replace it with two receivers at the position of the ears, and replace the system's loudspeakers with point monopole sources. The environment for the arrangement is assumed anechoic. This conceptual situation has the advantage of having a simple analytical solution. The acoustic complex pressure p due to a point monopole source at a distance r from the source is [28]

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

where an $e^{j\omega t}$ time dependence is assumed and where k is the wave number ω/c_0, ρ_0 is the density of the medium, c_0 is the sound speed, ω is the angular frequency, and q is the effective complex source strength (volume velocity). By designating complex acoustic pressure p as the output and complex source volume acceleration $j\omega q$ as the input, the frequency domain transfer function $C(j\omega)$ of the path from monopole source to a position at a distance r from the source is

$$C(j\omega) = \frac{\rho_0 e^{-jkr}}{4\pi r}. \qquad (2.2)$$

In the time domain the impulse response is a scaled delta function delayed by the sound's travel time (i.e. r/c_0). The free field computer simulations discussed in this thesis all employ Eq. (2.2). The receivers (ears) are always set at 18 cm apart. The low computation required for this type of simulation provides quick results that provide insight into the basic effects of the geometry on the sound field, especially at low frequencies.

2.4.2. Spherical head approximation

A common approach used to improve the approximation is to model the head as a perfectly rigid sphere where the two ears are on the surface at ends of a diameter. This model yields an analytical solution by again assuming an anechoic environment and modelling the sound sources as point monopole sources. Taking the complex source volume acceleration $j\omega q$ of a point monopole source as the input and acoustic complex pressure p on the surface of the sphere as the output, the total frequency response transfer function $C_t(j\omega)$ is [29]

$$C_t(j\omega) = C_{ff}(j\omega) + C_s(j\omega). \qquad (2.3)$$

The two components of Eq. (2.3) correspond to the incident $C_{ff}(j\omega)$ and scattered $C_s(j\omega)$ sound fields. Equations (2.4-2.6) specify these elements [29].

$$C_{ff}(j\omega) = -\frac{j\rho_0 k}{4\pi}\sum_{m=0}^{\infty}(2m+1)j_m(ka)\times\left[j_m(kr) - jn_m(kr)\right]P_m(\cos\phi) \qquad (2.4)$$

$$C_s(j\omega) = \frac{\rho_0 k}{4\pi}\sum_{m=0}^{\infty} b_m\left[j_m(ka) - jn_m(ka)\right]P_m(\cos\phi) \qquad (2.5)$$

$$b_m = j(2m+1)\frac{j_m(kr) - jn_m(kr)}{1 - \frac{jn_m'(ka)}{j_m'(ka)}} \tag{2.6}$$

In these equations r is the distance from the source to the centre of the sphere, a is the sphere's radius and the angle ϕ is defined in Fig. 2.2. The functions j_m and n_m are mth-order spherical Bessel functions of the first and second kind respectively and P_m is the mth-order Legendre polynomial. By employing the definition of the spherical Hankel function of the second kind $h_m^{(2)}$,

$$h_m^{(2)} = j_m - jn_m, \tag{2.7}$$

the relationship between the spherical Bessel functions and their derivative,

$$j_m'(z) = j_{m-1}(z) - \left(\frac{m+1}{z}\right) j_m(z) \text{ and} \tag{2.8a}$$

$$n_m'(z) = n_{m-1}(z) - \left(\frac{m+1}{z}\right) n_m(z), \tag{2.8b}$$

and the relationship between the spherical Bessel functions (j_m, n_m, and $h_m^{(2)}$) to their corresponding standard (cylindrical) Bessel functions (J_m, N_m, and $H_m^{(2)}$),

$$j_m(z) = \sqrt{\frac{\pi}{2z}} \cdot J_{m+1/2}(z), \tag{2.9}$$

which also holds if j_m and J_m are substituted for by either n_m and N_m or by $h_m^{(2)}$ and $H_m^{(2)}$, one is able to substitute Eqs. (2.4) and (2.5) into Eq. (2.3) and manipulate the result into the form,

$$C_t(j\omega) = \frac{-j\rho_0}{8\sqrt{ra}} \sum_{m=0}^{\infty} (2m+1) H^{(2)}_{m+\frac{1}{2}}(kr) P_m(\cos(\phi)) \left[\frac{J_{m+\frac{1}{2}}(ka) H^{(2)}_{m-\frac{1}{2}}(ka) - J_{m-\frac{1}{2}}(ka) H^{(2)}_{m+\frac{1}{2}}(ka)}{H^{(2)}_{m-\frac{1}{2}}(ka) - \frac{(m+1)}{ka} H^{(2)}_{m+\frac{1}{2}}(ka)} \right]$$

. (2.10)

Equation (2.10) is in a more favourable form for implementation in computer software packages in which the functions J_m and H_m are provided. However, an infinite summation is not possible. The infinite summation is over m, which is related to the order of the functions in the equation. If the frequency is 1 Hz the argument ka is

$$ka = \frac{2\pi f}{c_0} a = \frac{2\pi (1\text{ Hz})}{(344\text{ m/s})} (0.09\text{ m}) \approx 0.0016 . \quad (2.11)$$

for a 9 cm radial sphere, which is a dimension comparable with an adult human head. With this value for the argument the Hankel functions of the second kind in Eq. (2.10) increase approximately exponentially with increasing order. The imaginary part of the term $H_{m+1/2}^{(2)}(ka = 0.0016)$ (i.e. $-N_{m+1/2}(ka = 0.0016)$) is plotted in Fig. 2.3 as a function of increasing m (order). The function reaches such large values at orders above 67.5 that the Matlab computer software has an overload error and is unable to compute the answer. This limits the possible number of terms in the summation in Eq. (2.10) to $m = (0 \ldots 67)$. Figures 2.4 and 2.5 show the calculated magnitude and phase, respectively, of the direct path transfer function to a 9 cm radial sphere at 1 Hz as a function of the number of terms in the summation. From the figures, it appears that 68 terms in the summation is well sufficient to ensure convergence of the calculated

transfer function response. Figure 2.6 shows the time and frequency responses of a calculated transfer function when the angle subtending the point monopole sound source and the receiver point on the surface of the 9 cm radial sphere ϕ is 120° as calculated with terms up to m = (30, 40, and 67). Notice the frequency response is characteristically low-pass. This is true for receiving points on the opposite side of the sound source (i.e. $90° > \phi > 270°$). With m = 20 (dashed line) the frequency response starts to deviate significantly from the frequency response when m = 67 (solid line) at about 9 kHz. With m = 30 (dotted line) the frequency response starts to deviate from m = 67 at about 15 kHz. With m = 40 (dash dot line) the frequency response starts to deviate from m = 67 at about 20 kHz. For the simulations in this thesis, m = (0 ... 67) and the radius of the sphere is always set at 9 cm.

At frequencies near and below 1000 Hz, the spherical head model well approximates the shadowing effects of a human head. The model's predictions of the binaural localization cue, interaural time delay (ITD), agree closely with measurements made on man-shaped dummies [4,30] and on subjects [31]. The spherical head model's prediction of the binaural localisation cue, interaural level difference (ILD), differs by 1-2 dB from measurements made on subjects due to the sphere's lack of a neck [30]. A practical disadvantage of this model is the slow convergence of the Bessel functions making the calculations time consuming.

2.4.3. Dummy head related transfer functions (HRTF)

The previous two models made no effort to model the effects of the listener's pinnae, torso, and neck. The last model reviewed here makes use of direct measurements of a KEMAR (Knowles Electronics Manikin for Acoustic Research) dummy's head related transfer functions [32]. Head related transfer functions (HRTF) refer to transfer functions from sound sources to ears that include the effect of the head and pinnae. The KEMAR dummy has median human adult dimensions including its pinnae. The acoustic behaviour of KEMAR's ear canals and eardrum simulators matches that of real ears. The MIT Media Lab measured HRTFs of a KEMAR dummy in an anechoic chamber at a sampling frequency of 44.1 kHz [33]. This

database is available for downloading [34]. The measurements at 0° elevation at a radial distance of 1.4 m for the full 360° of azimuth sampled in 5° increments are employed in this thesis with the loudspeaker responses deconvolved.

To match the angle of interest some manipulation of the data is necessary. The measured HRTFs are a discrete sampling of a continuous auditory space. Listeners moving into situations where unmeasured HRTF spatial locations are required make interpolation of the measured HRTFs necessary. It is far from obvious which interpolation method is the best for this task. Takeuchi [35] describes a spherical interpolation procedure utilising bilinear interpolation algebra for HRTFs sampled on a spherical surface of the form,

$$x_i = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3}{w_1 + w_2 + w_3} \tag{2.12}$$

where the values (x_1, x_2, x_3) can be magnitude or phase decomposed from the measured HRTFs on the sampling surface and the real valued weighting factors (w_1, w_2, w_3) are the associated solid angles. If the point of interest to be interpolated lies in between two sampled points on a great circle of the spherical surface, then one of the solid angles is zero (0) and Eq. (2.12) reduces to a linear interpolation. For example, if the samples associated with x_1 and x_2 lie on the horizontal plane and we are interested in a point on the horizontal plane in between these two points than w_3 is zero (0) and Eq. (2.12) reduces to,

$$x_i = \frac{w_1 x_1}{w_1 + w_2} + \frac{w_2 x_2}{w_1 + w_2} = W_1 x_1 + W_2 x_2 \tag{2.13}$$

where $W_1 + W_2 = 1$ and both W_1 and W_2 are a combination of the weights w_1 and w_2, i.e.,

$$W_1 = \frac{w_1}{w_1 + w_2}, \qquad W_2 = \frac{w_2}{w_1 + w_2}. \tag{2.14a,b}$$

Throughout this thesis, the only HRTFs utilised are on the horizontal plane and linear interpolation is always the method of interpolation between the measured HRTFs. Some researchers first decompose the HRTFs into magnitude and phase and then carry out the interpolation on these elements separately [35,36]. Others perform interpolation on the complex valued frequency response directly [37]. These two methods are approximately equivalent when the magnitudes of the frequency responses of the two measured HRTFs are approximately equal. Some simple complex algebra helps to reveal this insight. If y_i is the interpolated complex frequency response obtained by linear interpolation on the magnitude and phase of the sampled complex frequency responses (x_1 and x_2), and x_i is the interpolated complex frequency response obtained by linear interpolation of the complex frequency responses (x_1 and x_2) then,

$$x_1 = |x_1| e^{i\phi_{x1}}, \quad x_2 = |x_2| e^{i\phi_{x2}}, \quad x_i = |x_i| e^{i\phi_{xi}}, \; y_i = |y_i| e^{i\phi_{yi}} \tag{2.15a-d}$$

$$|y_i| = W_1 |x_1| + W_2 |x_2|, \tag{2.16}$$

$$e^{i\phi_{yi}} = W_1 e^{i\phi_{x1}} + W_2 e^{i\phi_{x2}}, \text{ and} \tag{2.17}$$

$$x_i = W_1 x_1 + W_2 x_2 = W_1 |x_1| e^{i\phi_{x1}} + W_2 |x_2| e^{i\phi_{x2}} \tag{2.18}$$

y_i is then found by the multiplication of the right hand side of Eq. (2.16) with the right hand side of Eq. (2.17), viz.

$$y_i = |y_i|e^{\phi_{y_i}} = \left(W_1^2|x_1| + W_1W_2|x_2|\right)e^{i\phi_{x_1}} + \left(W_2^2|x_2| + W_1W_2|x_1|\right)e^{i\phi_{x_2}} \tag{2.19}$$

To see under what conditions y_i is equal to x_i one sets the right hand sides of Eq. (2.19) and Eq. (2.18) equal to each other. This yields

$$\left(W_1^2|x_1| + W_1W_2|x_2|\right)e^{i\phi_{x_1}} = W_1|x_1|e^{i\phi_{x_1}} \tag{2.20}$$

and

$$\left(W_2^2|x_2| + W_1W_2|x_1|\right)e^{i\phi_{x_2}} = W_2|x_2|e^{i\phi_{x_2}}\,. \tag{2.21}$$

Both Eqs. (2.20) and (2.21) reduce to

$$W_1 + W_2\left|\frac{x_2}{x_1}\right| = 1\,. \tag{2.22}$$

Since it is known that $W_1 + W_2 = 1$ from Eq. (2.14), the only way Eq. (2.22) will be approximately true or alternately the only way $y_i \approx x_i$ is when

$$\left|\frac{x_2}{x_1}\right| \approx 1\,. \tag{2.23}$$

When the magnitudes of the frequency responses of the two sampled measured HRTFs are approximately equal, the linear interpolation of the complex frequency response is approximately equivalent to linear interpolation of the magnitude and phase separately. In this thesis, linear interpolation is performed on the complex

frequency response without first decomposing it into magnitude and phase. Figures 2.7-2.9 show an example of a result of this kind of interpolation in the magnitude of the frequency response (Fig. 2.7), phase of the frequency response (Fig. 2.8), and a portion of the time domain response (Fig. 2.9) for a HRTF to the listener's right ear from a sound source 37.5° to the listener's front left. The solid line is the interpolated result from the measured HRTFs at 40° (dashed line) and 35° (dotted line) to the listener's front left. The interpolated results seem to be reasonably close to the two measured HRTFs. Note the dip in the magnitude of the frequency responses at about 8 kHz. This dip is the pinna notch discussed in section 2.2.2. The magnitude, phase and time response of the interpolated data (Figs. 2.7-2.9) all lay in between the measured HRTF responses for nearly all of the data shown. Interpolation between HRTFs is a current topic of debate [36-41] and is important for the application of adaptive virtual acoustic imaging systems. This thesis does not concentrate on the effect of the interpolation method but just utilises the simple method of linear interpolation in the complex frequency response as a reasonably accurate and simple approach.

The interpolation just discussed was to correct for the angle of the HRTF in the horizontal plane. The distance of the HRTF to the listener's ear could also be different from the measurement distance of 1.4 m. Correcting for distance involves modifying both the amplitude and time delay of the HRTF. Applying the far-field spherical radiation approximation leads to a simple scaling of the amplitude by the inverse of the propagation distance. For example, if the impulse response of a HRTF is $c(n)$ measured at a distance of 1.4 m and one wishes to correct this for a distance of r, than one can multiply the response $c(n)$ by 1.4 m and divide it by r to correct the amplitude of the HRTF for the new distance r, i.e.

$$c(n)\big|_r = \frac{(1.4\text{ m})c(n)\big|_{1.4\text{ m}}}{r}. \qquad (2.24)$$

The far-field approximation is appropriate for the frequency range of about 100 Hz to 10 kHz for a distance of 1.4 m from the source to ear location and with a maximum

source dimension of 4-in. [42]. Note that the loudspeaker used in the measurements had a 4-in. woofer. The discrete sampling of the HRTFs in time makes correcting for the propagation time delay more difficult than the above correction for amplitude. This is because the time delay will not generally correspond to an integer number of samples. If the amount of delay does correspond to an integer number of samples, the HRTF time response is simply shifted in time by the appropriate number of samples. This process can be thought of as filtering the signal with a digital unit impulse response that is delayed by the propagation delay time. For propagation time delay not equal to an integer number of samples, the optimal method is to first reconstruct the continuous signal from the sampled signal, shift the signal with the appropriate amount of delay, and then resample the signal. Alternatively, this can be considered a filtering of the signal with a linear phase all-pass filter with unity magnitude and with constant group delay equal to the propagation time. The ideal fractional delay interpolation filter $h_{id}(n)$ is a sampled sinc function of the form [43],

$$h_{id}(n) = \frac{\sin[\pi(n-D)]}{\pi(n-D)} = \mathrm{sinc}(n-D), \tag{2.25}$$

where D is the delay time in samples and n is the integer discrete time index. This filter will perfectly delay the signal without distortion for any integer or noninteger delay time. If the delay time D is an integer than $h_{id}(n)$ has a value of one (1) at $n = D$ and values of zero (0) elsewhere, that is, $h_{id}(n)$ is a digital unit impulse response with delay D. For noninteger values of D, the ideal filter $h_{id}(n)$ is a shifted, infinitely long, noncausal, sampled sinc function. To make this digital filter a realisable finite impulse response (FIR) filter it is often truncated, making the finite interpolation filter $h(n)$,

$$h(n) = \begin{cases} \mathrm{sinc}(n-D) & \text{for } M \le n \le M+N \\ 0 & \text{otherwise} \end{cases} \tag{2.26}$$

where N is the filter order and M is the integer time index of the first nonzero value of the impulse response $h(n)$. The optimal value of M (value which produces the least

error in the interpolation) is such that D is placed at the middle of the truncated impulse response $h(n)$. This truncation introduces error that is concentrated around the Nyquist frequency (22.05 kHz) due to the Gibbs phenomenon. Figures 2.10 and 2.11 show two examples of this type of fractional delay filter. The delays in the figures are 64.25 samples (Fig. 2.10) and 64.4 samples (Fig. 2.11). The time responses (Figs. 2.10a and 2.11a) show the sampled sinc function, which is centred on the delay time D. The magnitudes of the frequency responses (Figs. 2.10b and 2.11b) are constant and near one (1) at most frequencies but start significantly deviating from one (1) at frequencies above about 20 kHz. The group delays of the filters (Figs. 2.10b and 2.11b) are also very constant and as desired at low frequencies but start to deviate from the desired delay at frequencies above about 20 kHz. This error above 20 kHz is introduced by the truncation of the sinc interpolation filter. The number of coefficients in the FIR filters is 128 (i.e. $N = 127$). The application of the fractional delay filters shown in Figs. 2.10 and 2.11 to the HRTF shown in Fig. 2.12 yields the delayed HRTFs shown in Figs. 2.13 and 2.14 respectively. The magnitude of the HRTF does not change much due to the fractional delay filtering but the phase of the filtered HRTFs decreases much more rapidly with increasing frequency. In this thesis, filtering the KEMAR dummy HRTFs with a shifted, truncated, digital sinc function with 128 samples compensates for propagation time. The actual time delay added in this thesis is determined by assuming a 344 m/s sound speed.

Many of the computer simulations and the design of most of the virtual acoustic imaging filters used in the subjective experiments throughout this thesis utilise KEMAR dummy HRTFs. Dummy HRTFs have the advantage over the previous two transfer function models of not only containing appropriate interaural information but also spectral information. However, the spectral characteristics of KEMAR do not exactly match those of individual subjects. This mismatch can affect localization performance [44]. Interpolation for horizontal angle and distance introduce some slight error into the responses.

2.5. Conclusions

Virtual acoustic imaging systems attempt to give the listener the impression of a virtual auditory environment by delivering the appropriate interaural and spectral sound localisation cues. Although these cues are not the only factors determining human sound localisation they are the most important. In order to replicate sound signals at positions in space with loudspeakers one must model the responses of acoustic paths. The free field approximation provides a simple model that well approximates time responses at low frequencies. The spherical head model well approximates responses at low frequencies as well as the at middle range frequencies. Measured dummy HRTFs provide realistic approximations of eardrum sound signals throughout most of the audible frequency range. Calculation of low frequency ITD cues may employ any of the three models. Calculation of ILD may incorporate the spherical head model or the dummy HRTFs. Spectral cues are present in only the dummy HRTFs.

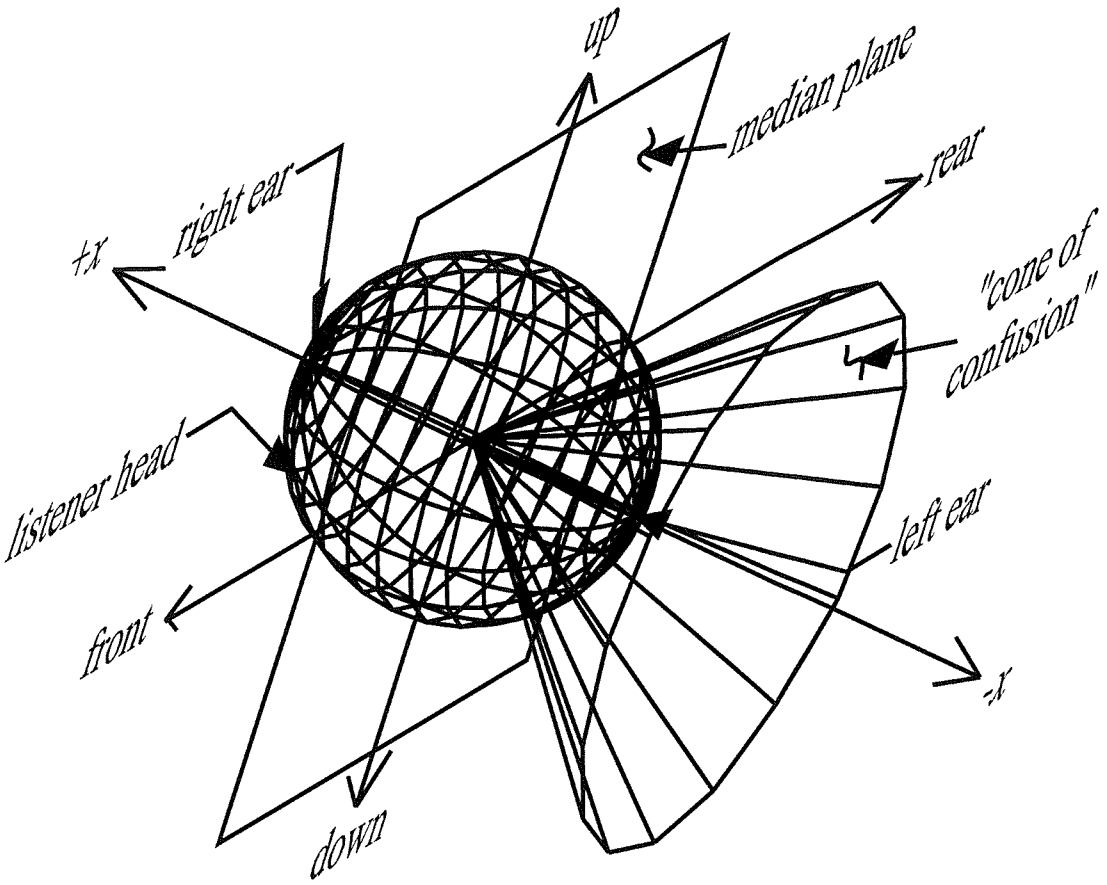


Fig. 2.1 *Three-dimensional image showing the median plane and a "cone of confusion".*

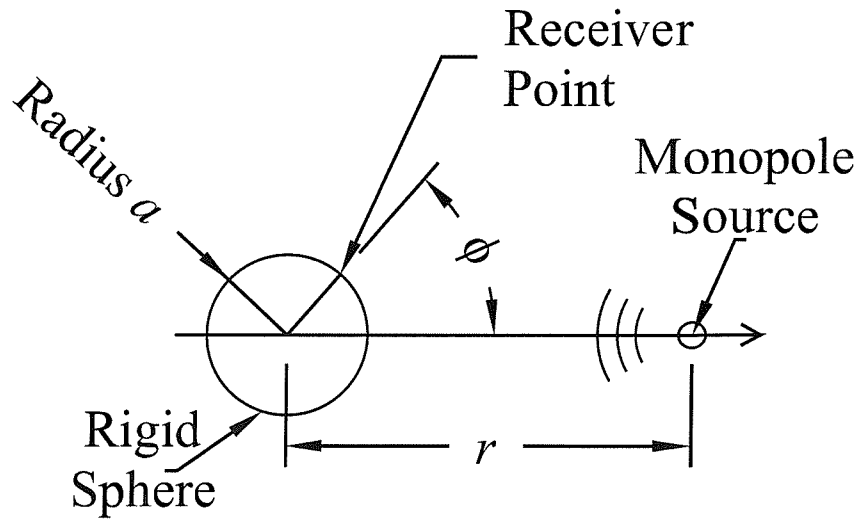


Fig. 2.2 *Variables used when calculating the sound field scattered from a rigid sphere.*

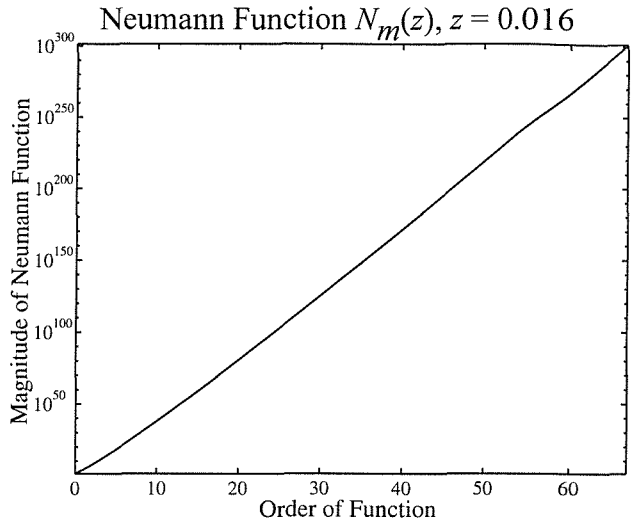


Fig. 2.3 *Imaginary part of the Hankel function* $H_m^{(2)}(z = 0.0016)$ *or alternatively the negative of the Neumman function* $-N_m(z = 0.0016)$ *with the argument z constant and equal to 0.0016 as a function of the function's order m. Beyond order 67.5 the software has an overload error and higher orders are beyond the software's capabilities with* $z = 0.0016$.

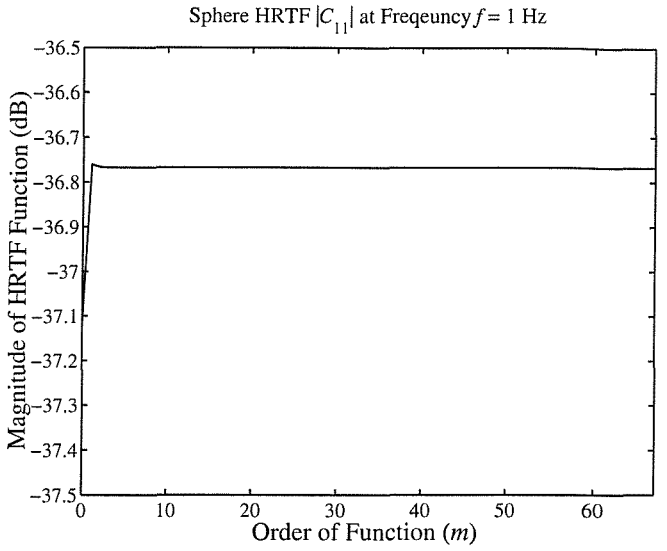


Fig. 2.4 *Calculated magnitude of the direct path spherical transfer function for a 9 cm radial sphere at 1 Hz as a function of the order m. Beyond order 67.5 the software has an overload error and higher orders are beyond the software's capabilities.*

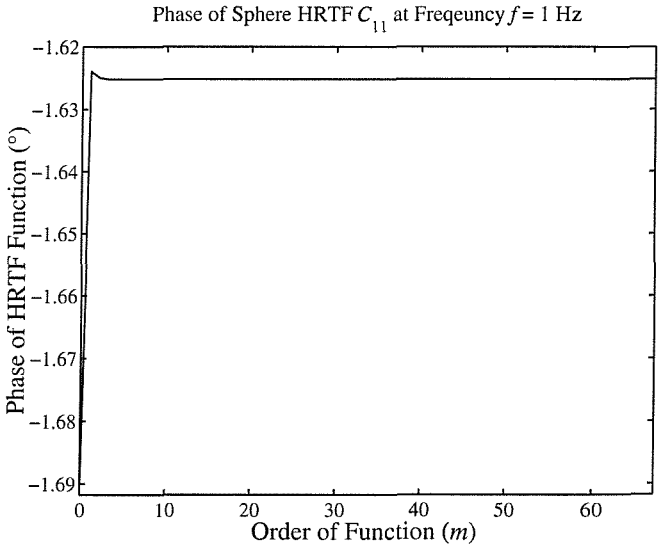


Fig. 2.5 *Calculated phase of the direct path spherical transfer function for a 9 cm radial sphere at 1 Hz as a function of the order m. Beyond order 67.5 the software has an overload error and higher orders are beyond the software's capabilities.*

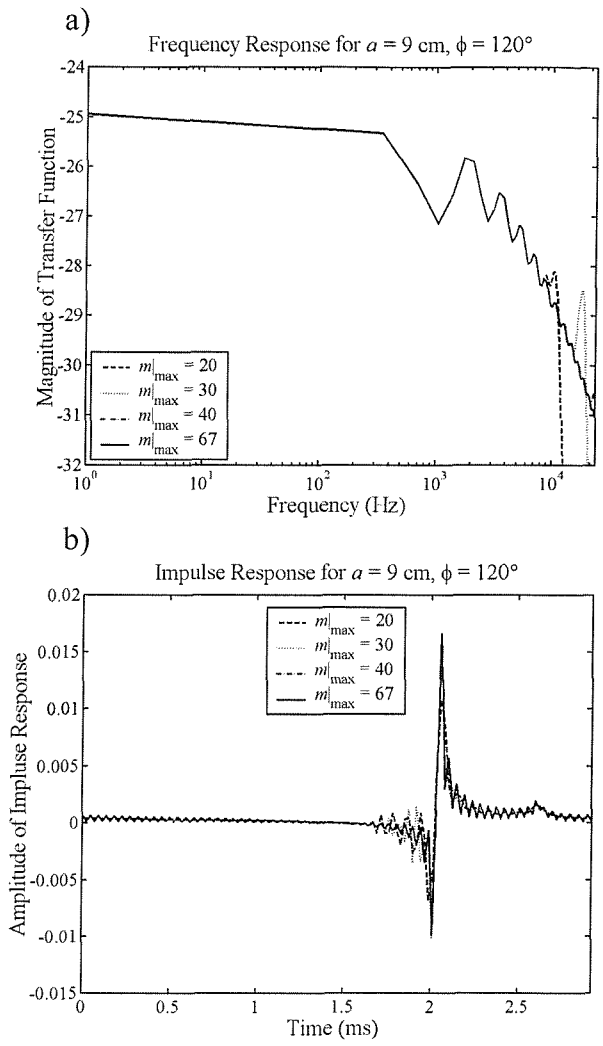


Fig. 2.6 a) Magnitude of the *frequency response and b) time impulse response of a spherical transfer function calculated up to and including 20 (dashed lines), 30 (dotted lines), 40 (dash dot lines), and 67 (solid lines) values of m in the summation. The receiver is on the surface of the 9 cm radial sphere and subtends an angle of 120° with the sound source location.*

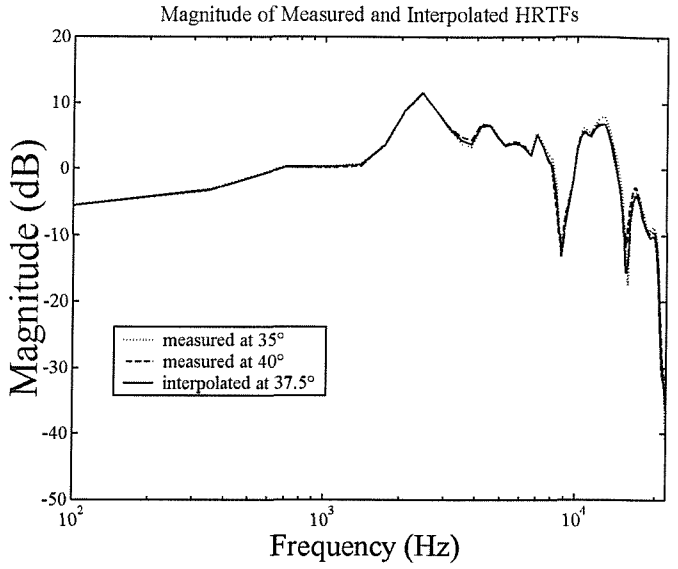


Fig. 2.7 *Magnitude of measured HRTFs to the right ear of KEMAR dummy from 35° to KEMAR's front left (dotted line), 40° to KEMAR's front left (dashed line), and the resulting magnitude of the frequency response of an interpolation between the two measured HRTFs to 37.5° to KEMAR's front left (solid line). The linear interpolation was undertaken on the complex frequency response.*

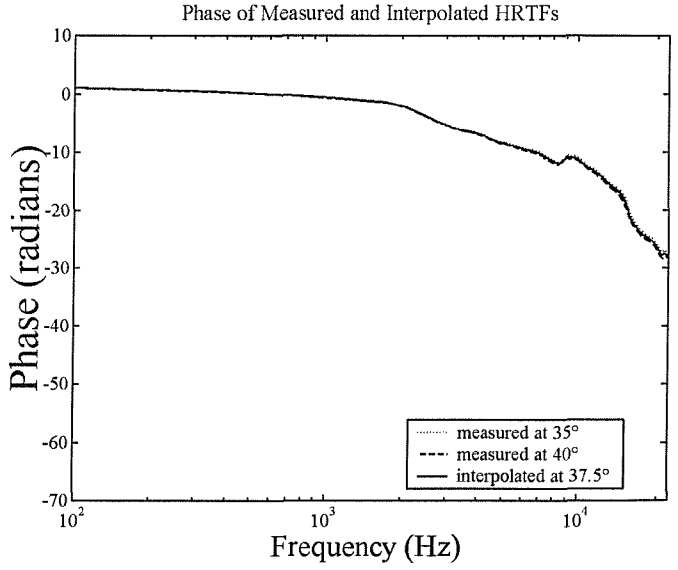


Fig. 2.8 *Phase of measured HRTFs to the right ear of KEMAR dummy from 35° to KEMAR's front left (dotted line), 40° to KEMAR's front left (dashed line), and the resulting phase of the frequency response of an interpolation between the two measured HRTFs to 37.5° to KEMAR's front left (solid line). The linear interpolation was undertaken on the complex frequency response.*

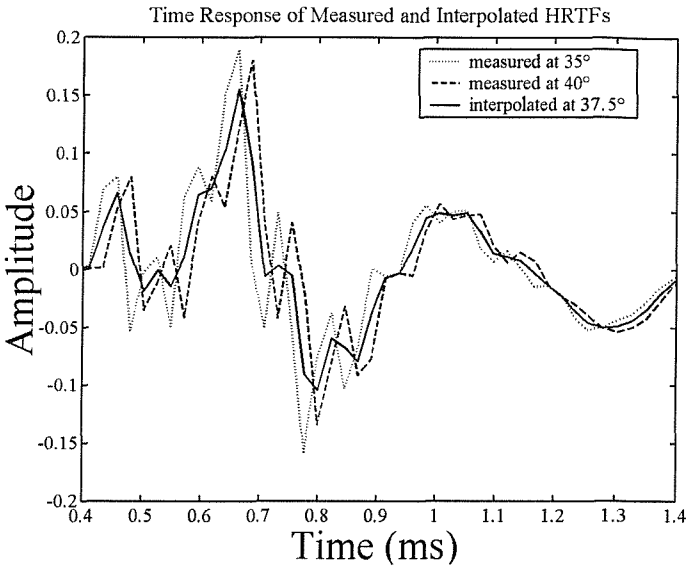


Fig. 2.9 *Time response of measured HRTFs to the right ear of KEMAR dummy from an angle of 35° to KEMAR's front left (dotted line), 40° to KEMAR's front left (dashed line), and the resulting time response of an interpolation between the two measured HRTFs from an angle of 37.5° to KEMAR's front left (solid line). The linear interpolation was done on the complex frequency response.*

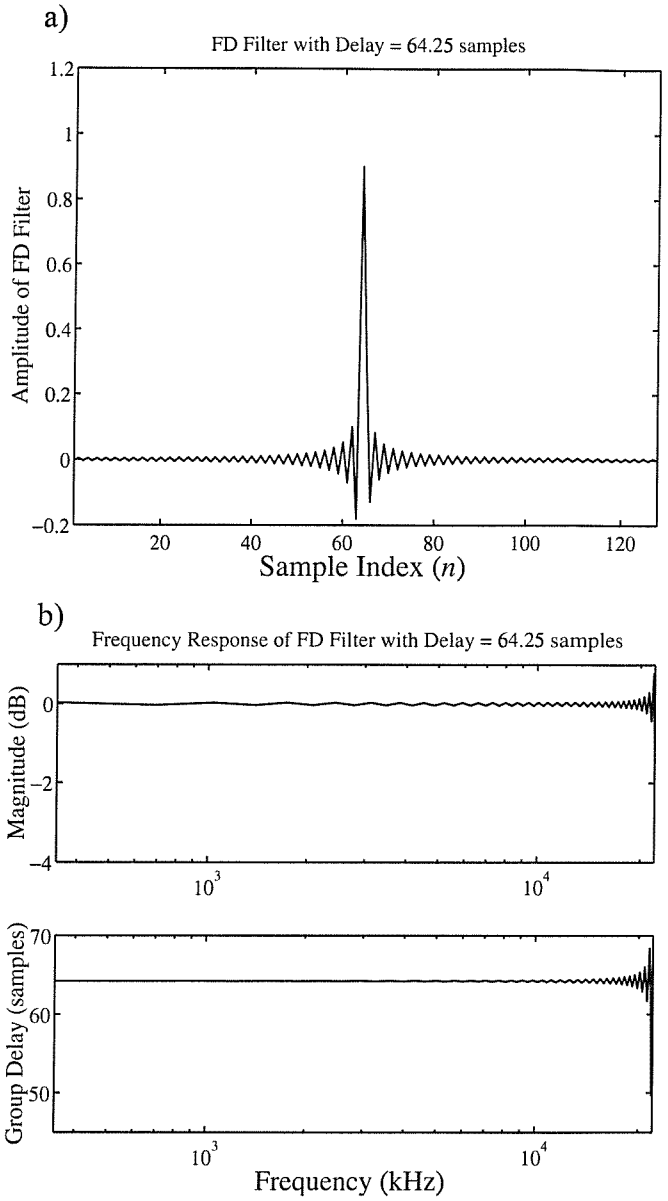


Fig. 2.10 *Time and frequency response of a fractional delay filter with a delay of 64.25 samples. The time response of the filter a) is a truncated sinc function with 128 coefficients. The truncation introduces error in both the b) magnitude of the frequency response and group delay at high frequencies.*

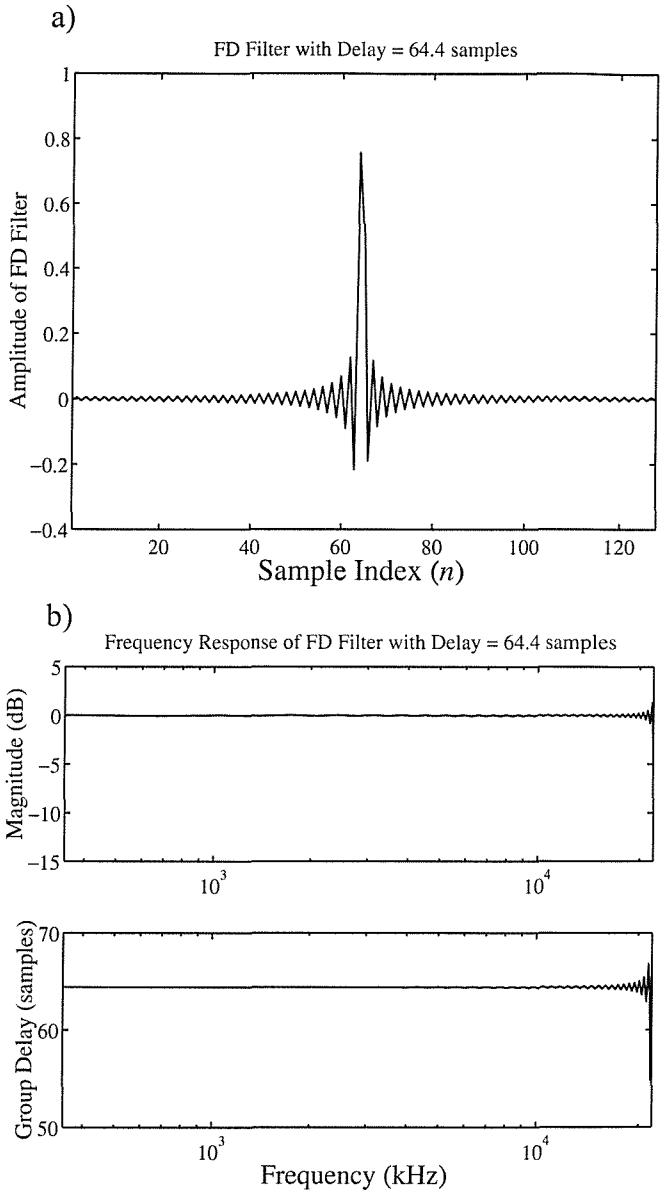


Fig. 2.11 *Time and frequency response of a fractional delay filter with a delay of 64.4 samples. The time response of the filter a) is a truncated sinc function with 128 coefficients. The truncation introduces error in both the b) magnitude of the frequency response and group delay at high frequencies.*

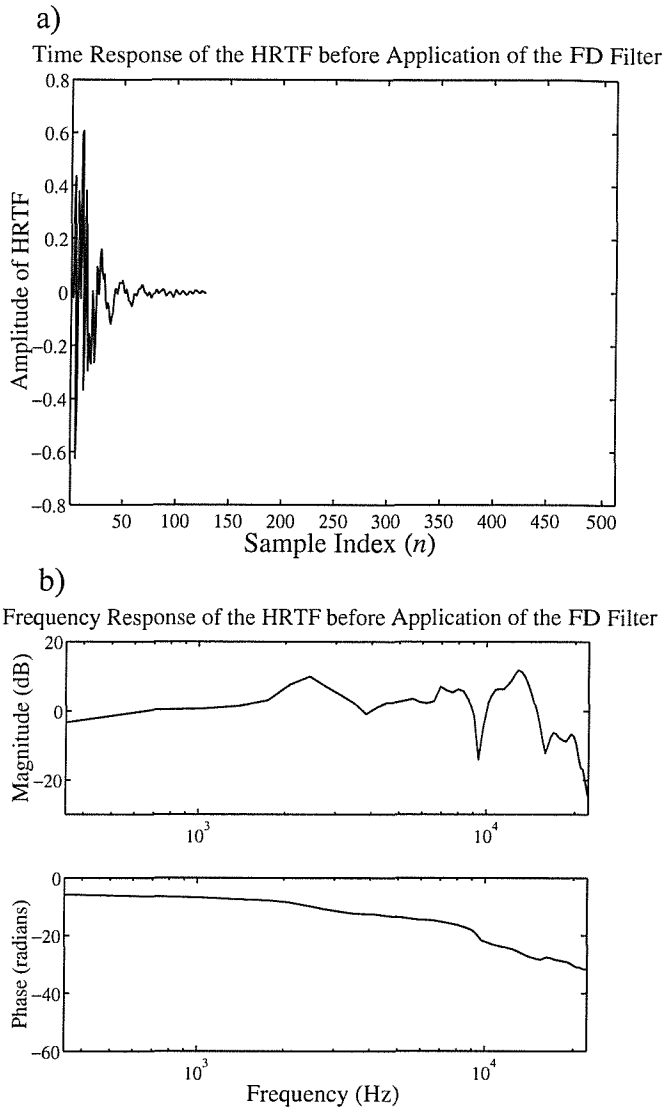


Fig. 2.12 *Time and frequency response of a measured KEMAR dummy HRTF with 128 coefficients. This is for the path from a sound source located directly at KEMAR's right side to his right ear.*

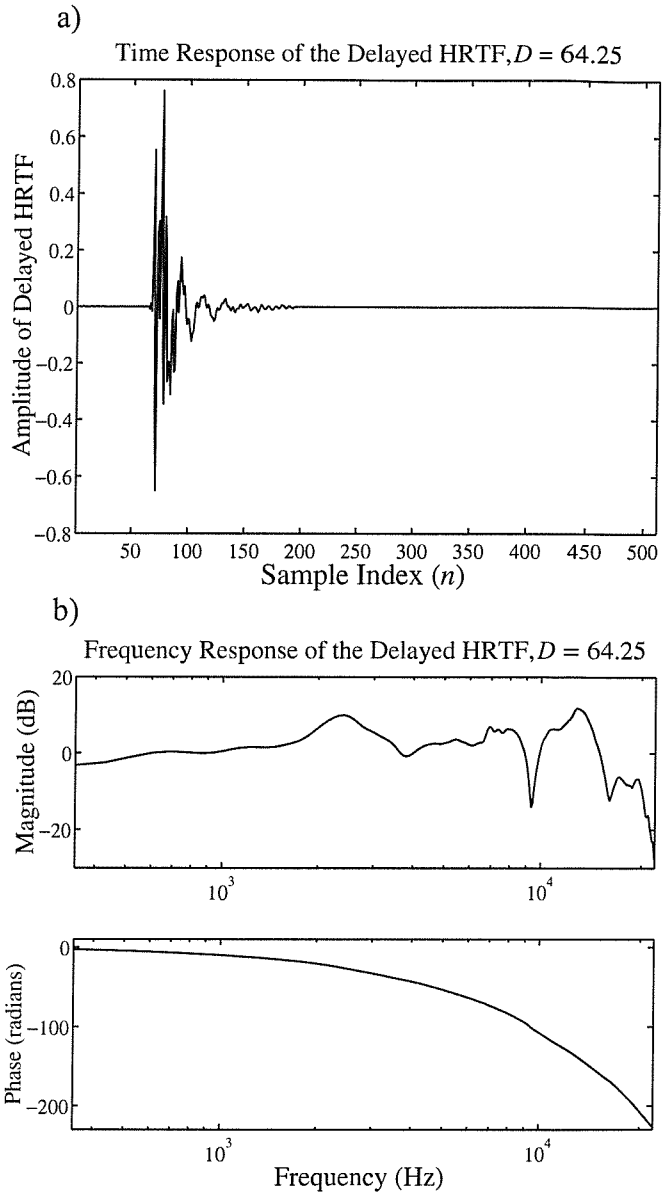


Fig. 2.13 *Time and frequency response of the HRTF shown in Fig. 2.12 after being filtered with the fractional delay filter shown in Fig. 2.10 that has a delay of 64.25 samples.*

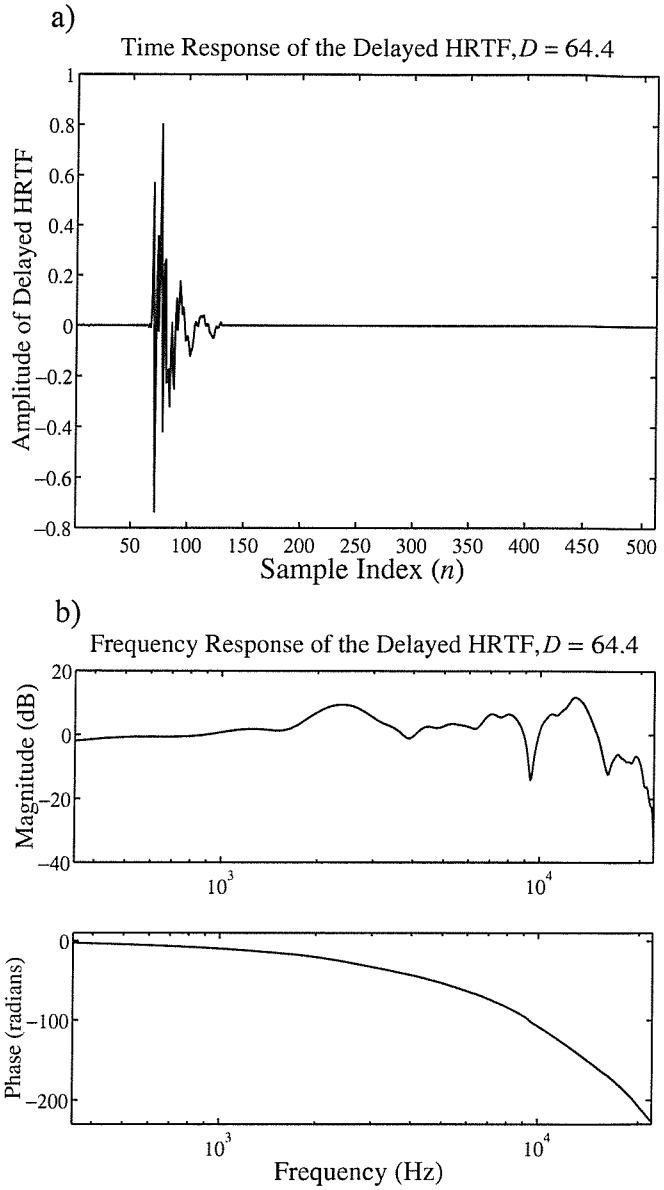


Fig. 2.14 *Time and frequency response of the HRTF shown in Fig. 2.12 after being filtered with the fractional delay filter shown in Fig. 2.11 that has a delay of 64.4 samples.*

3. VIRTUAL ACOUSTIC IMAGING

3.1. Introduction

Two of the most important parameters affecting the performance of a virtual acoustic imaging system are the system's geometry and the design of the virtual acoustic imaging filters. This chapter discusses both of these topics, first considering the effect of the geometry on the system's performance and then describing the design procedure of the virtual acoustic imaging filters. Important concepts affecting performance are the "ringing" frequency and the condition number of the matrix of HRTFs that must be inverted. The relationship between these concepts and the geometry of the virtual acoustic imaging system explains the rationale behind the arrangement of the Stereo Dipole virtual acoustic imaging system used in this thesis.

3.2. System Performance

This section considers some of the physics of the sound field for virtual acoustic imaging systems that attempt to deliver binaural signals to listeners located on the inter-source axis and off-axis. Figure 3.1 shows a plan view of a virtual acoustic imaging system and depicts the acoustic paths from the two real loudspeakers to the listener's ears and the virtual acoustic paths from a virtual sound source to the two ears. Figure 3.2 shows the corresponding system block diagram [45]. The plant matrix $\mathbf{C}(z)$ represents the transfer functions of the paths from the two real sound sources to the two ears.

$$\mathbf{C}(z) = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \tag{3.1}$$

The goal of the system is to produce, at the ears of the listener, a vector of desired binaural signals $\mathbf{d}(z)$. One acquires the desired signals $\mathbf{d}(z)$ by filtering the source

signal $S(z)$ with vector of virtual acoustic paths $\mathbf{a}(z)$ that represent the transfer functions from a virtual sound source to the listener's ears.

$$\mathbf{a}(z) = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \tag{3.2}$$

The method of presenting the listener with the sound through two loudspeakers ensures a modification of $\mathbf{d}(z)$ by the plant matrix $\mathbf{C}(z)$ (Eq. 3.1). The problem is to design a matrix of filters $\mathbf{X}(z)$ that will cancel the effect of $\mathbf{C}(z)$ i.e.,

$$\mathbf{X}(z) = \mathbf{C}(z)^{-1}. \tag{3.3}$$

Generally, the plant matrix $\mathbf{C}(z)$ is non-minimum phase so that its inverse is non-causal making the filters non-realisable. Employing a modelling delay Δ is an approach to this problem. The idea is to shift the filters' impulse responses forward in time by an amount Δ so that most of the filters' energy is contained in the causal time domain. Also the poles of $\mathbf{X}(z)$ (the zeros of $\mathbf{C}(z)$) can have magnitudes very near to unity. This makes the rate of decay of the filters very slow so that long duration filters with many coefficients are required. The technique of regularisation introduces some error to the inversion process but increases the rate of decay of the inverse filters. The error introduced by regularisation is concentrated around peaks in the inversion, which are the most ill-conditioned frequencies. Alternatively, regularisation can be thought of reducing the required "effort" of the system, as an example shows below. So to help with the problems of long-duration high-output filters and non-causality, the criterion in Eq. (3.3) is relaxed by allowing for some error and by employing a modelling delay Δ so that Eq. (3.3) becomes,

$$\mathbf{C}(z)\mathbf{X}(z) \approx z^{-\Delta}\mathbf{I}, \tag{3.4}$$

where **I** is the identity matrix of order two (2). The filter matrix $\mathbf{X}(z)$ is the cross-talk cancellation filter matrix, and its job is to cancel the effect of the plant matrix $\mathbf{C}(z)$. The vector of virtual acoustic imaging filters $\mathbf{h}(z)$ is given by the matrix multiplication of the cross-talk cancellation filters $\mathbf{X}(z)$ and the virtual path vector $\mathbf{a}(z)$, i.e.

$$\mathbf{h}(z) = \mathbf{X}(z)\mathbf{a}(z). \tag{3.5}$$

From Fig. 3.2 the source outputs $\mathbf{v}(z)$ are

$$\mathbf{v}(z) = \mathbf{X}(z)\mathbf{d}(z) \tag{3.6}$$

and the reproduced signals $\mathbf{w}(z)$ are

$$\mathbf{w}(z) = \mathbf{C}(z)\mathbf{v}(z) \tag{3.7}$$

This section utilises the above relationships to consider some properties of the plant matrix **C** and its affect on the performance of the system.

3.2.1. Condition number

It is possible to factor **C** or in fact any m by n matrix into two orthogonal matrices **U** (m by m) and **V** (n by n) and a diagonal matrix **Σ** (m by n) such that,

$$\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{H} \tag{3.8}$$

where H denotes the conjugate transpose operation. In our application, the matrix **C** is complex and so the orthogonal matrices **U** and **V** are actually unitary matrices. The columns of **U** are eigenvectors of $\mathbf{C}\mathbf{C}^{H}$, and the columns of **V** are eigenvectors of

$\mathbf{C}^H\mathbf{C}$. The values $\sigma_1, \sigma_2, \ldots, \sigma_r$, where r is the rank of $\mathbf{C}$, fill the first r places on the main diagonal of $\mathbf{\Sigma}$. These values are the square roots of the nonzero eigenvalues of both $\mathbf{C}\mathbf{C}^H$ and $\mathbf{C}^H\mathbf{C}$ or simply the singular values of the matrix $\mathbf{C}$.

The condition number κ with respect to matrix inversion is the ratio of the matrix's maximum to minimum singular value i.e.,

$$\kappa(\mathbf{C}) = \frac{\sigma_{\max}}{\sigma_{\min}} \tag{3.9}$$

This is a measure (for a system of linear equations) of the solution's sensitivity or "vulnerability" to small perturbations or errors. A well-conditioned matrix has a condition number equal or close to unity. An ill-conditioned matrix has a large condition number. The interest here is in the sensitivity to error of system inversion of an assumed plant matrix $\mathbf{C}$. A source of error in the plant matrix $\mathbf{C}$ is in differences of the HRTF model used and the actual listener's HRTF. These differences include head displacements from the assumed listener location. Therefore, $\mathbf{C}$'s condition number reflects the system's robustness to head displacements [46].

With respect to the path lengths shown in Fig. 3.1, Eq. (3.10) defines the path length difference Δl as

$$\Delta l = \frac{1}{2}\left(l_{12} + l_{21} - l_{11} - l_{22}\right). \tag{3.10}$$

At symmetric listener positions, $l_{11} = l_{22} = l_1$ and $l_{12} = l_{21} = l_2$ so that Eq. (3.10) becomes

$$\Delta l = l_2 - l_1. \tag{3.11}$$

Equation (3.1) is modified to become Eq. (3.12) by modelling the acoustic paths shown in Fig. 3.1 with the free field approximation of Eq. (2.1), assuming the symmetric listener location, and by normalising the plant matrix by $\frac{4\pi l_1}{\rho_0 e^{-jkl_1}}$ so that the normalised plant matrix $\mathbf{C}_N$ at the symmetric on-axis free field location is [47]

$$\mathbf{C}_N(j\omega) = \frac{4\pi l_1}{\rho_0 e^{-jkl_1}}\mathbf{C}(j\omega) = \begin{bmatrix} 1 & ge^{-jk\Delta l} \\ ge^{-jk\Delta l} & 1 \end{bmatrix} \tag{3.12}$$

where $g = l_1/l_2$. The in-phase and out-of-phase singular values of this matrix are [35]

$$\sigma_{\text{in}} = \sqrt{\left(1 + ge^{jk\Delta l}\right)\left(1 + ge^{-jk\Delta l}\right)} \tag{3.13}$$

and

$$\sigma_{\text{out}} = \sqrt{\left(1 - ge^{jk\Delta l}\right)\left(1 - ge^{-jk\Delta l}\right)}. \tag{3.14}$$

Figure 3.3 shows the behaviour of the singular values as a function of the wavenumber times the path length difference $k\Delta l$. Recall from Eq. (3.9) that condition number $\kappa(\mathbf{C})$ of the system is the ratio of the maximum to minimum singular values. From Fig. 3.3 it is seen that large condition numbers occur at frequencies where $k\Delta l$ = 0,π,2π,3π,4π,5π,... (ill-conditioned frequencies) or where the path length difference is equal to integer numbers of half wavelengths (i.e. $\Delta l = 0, \frac{\lambda}{2}, \lambda, \frac{3\lambda}{2}, 2\lambda, \frac{5\lambda}{2}, \ldots$ where λ is the acoustic wavelength).

Figure 3.3 shows peaks in σ_{in} occurring at every other ill-conditioned frequency where the path length difference is equal to integer numbers of wavelengths (i.e.

$\Delta l = 0, \lambda, 2\lambda, \ldots$). At these frequencies, the vector of normalised pressure at the receiver points

$$\mathbf{p}_N = \begin{bmatrix} p_{N1} \\ p_{N2} \end{bmatrix} = \mathbf{C}_N \mathbf{v} = \mathbf{C}_N \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \tag{3.15}$$

is related to the source volume accelerations **v** by [47]

$$\tilde{p}_{N\text{in}} = (p_{N1} + p_{N2}) = (1+g)(v_1 + v_2) = (1+g)\tilde{v}_{\text{in}} \tag{3.16}$$

or

$$\tilde{p}_{N\text{out}} = (p_{N1} - p_{N2}) = (1-g)(v_1 - v_2) = (1-g)\tilde{v}_{\text{out}} \tag{3.17}$$

where $\tilde{p}_{N\text{in}}$ and $\tilde{p}_{N\text{out}}$ are the in-phase and out-of-phase pressure components at the receiver positions respectively and $\tilde{v}_{\text{in}}$ and $\tilde{v}_{\text{out}}$ are the in-phase and out-of-phase components of the source volume accelerations respectively. The in-phase pressure component is well coupled (by 1+g) to the in-phase component of the volume accelerations and the out-of-phase pressure component is only weakly coupled (by 1-g) to the out-of-phase component of the volume accelerations at these frequencies. Also at these frequencies, if the cross-talk cancellation matrix **X** is a perfect inverse of the plant **C** then the in-phase and out-of-phase components of the volume accelerations relate to the in-phase component of the desired signals $\tilde{d}_{\text{in}} = d_1 + d_2$ and the out-of-phase component of the desired signals $\tilde{d}_{\text{out}} = d_1 - d_2$ by [47],

$$\tilde{v}_{\text{in}} = (1+g)^{-1}\tilde{d}_{\text{in}} \tag{3.18}$$

$$\tilde{v}_{\text{out}} = (1-g)^{-1}\tilde{d}_{\text{out}} \tag{3.19}$$

The in-phase component of the volume accelerations is weakly coupled to the in-phase component of the desired signals and the out-of-phase component of the volume accelerations is well coupled to the out-of-phase component of the desired signals. These couplings compensate for the strong and weak couplings of the in-phase and out-of-phase pressure components to the in-phase and out-of-phase components of the volume accelerations respectively. The inverse filters maximally amplify the out-of-phase component of the desired signals at frequencies where the path length difference is equal to integer numbers of wavelengths (i.e. $\Delta l = 0, \lambda, 2\lambda, \ldots$).

Figure 3.3 shows peaks in σ_{out} occurring at other ill-conditioned frequencies where the path length difference is equal to odd integer numbers of half wavelengths (i.e. $\Delta l = \frac{\lambda}{2}, \frac{3\lambda}{2}, \frac{5\lambda}{2}, \ldots$). At these frequencies [47],

$$\tilde{p}_{N\text{in}} = (1-g)\tilde{v}_{\text{in}}, \tag{3.20}$$

$$\tilde{p}_{N\text{out}} = (1+g)\tilde{v}_{\text{out}} \tag{3.21}$$

and if we assume $\mathbf{X} = \mathbf{C}^{-1}$,

$$\tilde{v}_{\text{in}} = (1-g)^{-1}\tilde{d}_{\text{in}} \tag{3.22}$$

$$\tilde{v}_{\text{out}} = (1+g)^{-1}\tilde{d}_{\text{out}} \tag{3.23}$$

The in-phase pressure component is weakly coupled to the in-phase component of the volume accelerations, the out-of-phase pressure component is well coupled to the out-of-phase component of the volume accelerations. To compensate for these couplings, the in-phase component of the volume accelerations is well coupled to the in-phase component of the desired signals, and the out-of-phase component of the volume accelerations is weakly coupled to the out-of-phase component of the desired signals at these frequencies. The inverse filters maximally amplify only the in-phase component of the desired signals at these frequencies. Therefore, from Eqs. (3.18), (3.19), (3.22), and (3.23) the ill-conditioned frequencies (i.e. $\Delta l = 0, \frac{\lambda}{2}, \lambda, \frac{3\lambda}{2}, 2\lambda, \frac{5\lambda}{2}, \ldots$) require the system to work with maximal "effort" and are associated with dominance of either the in-phase or out-of-phase components of the source volume accelerations.

Figure 3.3 also shows that the smallest condition numbers occur at frequencies where $k\Delta l$ = $\pi/2, 3\pi/2, 5\pi/2, 7\pi/2, 9\pi/2, 11\pi/2, \ldots$ (well-conditioned frequencies) or where the path length difference is equal to odd integer numbers of quarter wavelengths (i.e. $\Delta l = \frac{\lambda}{4}, \frac{3\lambda}{4}, \frac{5\lambda}{4}, \frac{7\lambda}{4}, \frac{9\lambda}{4}, \frac{11\lambda}{4}, \ldots$). At every other one of these well-conditioned frequencies where $\Delta l = \frac{\lambda}{4}, \frac{5\lambda}{4}, \frac{9\lambda}{4}, \ldots$ [47],

$$\tilde{v}_{\text{in}} = (1 - jg)^{-1} \tilde{d}_{\text{in}} \tag{3.24}$$

$$\tilde{v}_{\text{out}} = (1 + jg)^{-1} \tilde{d}_{\text{out}} . \tag{3.25}$$

At the other well-conditioned frequencies where $\Delta l = \frac{3\lambda}{4}, \frac{7\lambda}{4}, \frac{11\lambda}{4}, \ldots,$

$$\tilde{v}_{\text{in}} = (1 + jg)^{-1} \tilde{d}_{\text{in}} \tag{3.26}$$

$$\tilde{v}_{\text{out}} = (1 - jg)^{-1} \tilde{d}_{\text{out}} . \qquad (3.27)$$

In Eqs. (3.24-3.27) the moduli of the coupling factor are all equal. Neither the in-phase nor out-of-phase components of the source volume accelerations dominate the solution. The inverse filters amplify both the in-phase and out-of-phase components of the desired signals equally making the "effort" required from the system minimal at frequencies where the path length difference is equal to odd integer numbers of quarter wavelengths (i.e. $\Delta l = \frac{\lambda}{4}, \frac{3\lambda}{4}, \frac{5\lambda}{4}, \frac{7\lambda}{4}, \frac{9\lambda}{4}, \frac{11\lambda}{4}, \ldots$).

The geometry of the virtual acoustic imaging system affects the value of the path length difference Δl, which is associated with ill-conditioned and well-conditioned frequencies and the required "effort" of the system as seen above. More generally, the system's geometry is an important influential factor that is crucial to the system's overall success. Figures 3.4 and 3.5 show the system's condition number as a function of both frequency and source span 2θ or the subtended angle between the two loudspeakers and the listener's head. For example, in Fig. 3.1 the system's source span 2θ is 10°. In Figs. 3.4 and 3.5, the horizontal axes represent the source span in degrees and the vertical axes represent frequency on a logarithmic scale in kilohertz. The grey scale represents the condition number in decibels (dB) with black and white corresponding to high and low condition numbers respectively. The free field model was utilised in the calculations that resulted in both of these figures.

The black bold vertical line in Fig. 3.4 corresponds to a 10° source span over a frequency range of approximately 300 Hz to 8 kHz. These important audio frequencies are well-conditioned (light coloured region) at this source span. A 10° source span corresponds to the geometry of the Stereo Dipole virtual acoustic imaging system [26,48,49]. The Stereo Dipole places the loudspeakers closely together to take advantage of good conditioning over a broad and important audio frequency range [45].

The idea behind the optimal source distribution (OSD) virtual sound imaging system is to stay in the well-conditioned region over the broadest possible frequency range [50,51]. One way to do this is by using more pairs of transducers at different source spans to emit different ranges of frequencies. The black vertical lines in Fig. 3.5 show the frequency ranges of an OSD system with three transducer pairs at source spans of 6°, 32°, and 180°. The OSD binaural system exhibits excellent control of the sound field over almost the entire audible frequency range [52].

The work in this thesis concentrates on the Stereo Dipole virtual acoustic imaging system and considers its performance at off-axis asymmetric listener locations. Figures 3.6-3.8 show the condition number of **C** for the Stereo Dipole as a function of frequency for listener locations between ±1.5 m off-axis using the free field approximation (Fig. 3.6), spherical head model (Fig. 3.7), and the KEMAR dummy HRTFs (Fig. 3.8). The vertical axis in these figures represents frequency in kilohertz. Figures 3.6-3.8 show the results on both a linear and logarithmic frequency scale in the top (Figs. 3.6a, 3.7a, and 3.8a) and bottom (Figs. 3.6b, 3.7b, and 3.8b) figures respectively. The horizontal axes in all of these figures represent displacement of the listener location from the inter-source axis in centimetres.

The most notable feature in Fig. 3.6 is a narrow band of ill-conditioning that approximately defines a parabola as a function of off-axis listener locations. The ill-conditioning occurs at about 11 kHz for on-axis (i.e. $x = 0$) and increases off-axis until at about 1 m off-axis the ill-conditioning occurs at frequencies greater than the audible frequency range. Fig. 3.7 also shows this feature when the transfer functions are computed using the spherical head model. The result is a slightly smeared version of the free field model. The logarithmic frequency scale in Figs. 3.6b and 3.7b reveals that the system is also ill-conditioned at very low frequencies, which have a similar approximate parabola shape as a function of off-axis listener locations. These figures show that the well-conditioned frequency range shifts up in frequency as the listener location moves further off-axis. The frequencies which are the most robust to head displacement increase as the intended listener location moves away from the inter-source axis. Figure 3.8 shows how the dummy's pinnae responses complicate matters.

One can still see the general trend noticed in Figs. 3.6 and 3.7 but this is far less clear. The loudspeakers used to measure the HRTF's were of limited bandwidth and the results below about 200 Hz in Fig. 3.8 reflect this limitation. The pinnae resonances seem to cause the ill-conditioned frequencies to oscillate. The KEMAR dummy head matrix of HRTFs is also ill-conditioned around 8 kHz for head positions greater than 50 cm off-axis. This is likely to be due to the small response at the pinna notch in the HRTFs. When including the effects of the pinnae and torso, the system appears to be robust over a broad frequency range at listener positions between ±50 cm off-axis.

Figures 3.9-3.11 show condition numbers for the traditional 60° source span 2θ stereo arrangement as calculated with the free field model (Fig. 3.9), spherical head model (Fig. 3.10), and KEMAR dummy HRTFs (Fig. 3.11). With this geometry, the loudspeakers are further away from each other and the path length difference Δl is greater than with the Stereo Dipole's geometry. At on-axis for the Stereo Dipole the ill-conditioned frequency where $\Delta l = \lambda/2$ occurs was seen to be at about 11 kHz (e.g. see Fig. 2.8a) but for the 60° arrangement this ill-conditioned frequency occurs at about 2 kHz for the on-axis listener location as seen in Figs. 3.10 and 3.11. Frequencies close to 2 kHz are important in many applications such as reproduction of speech or music. The harmonics of this frequency are also ill-conditioned and in the case of the 60° arrangement many of these harmonics occur within the audible frequency range as seen in Figs. 3.9 and 3.10. There is a much more narrow range of well-conditioned frequencies for the 60° loudspeaker arrangement than for the closely spaced loudspeaker arrangement of the Stereo Dipole.

3.2.2. Ringing frequency

Equation (3.1) becomes Eq. (3.28) by modelling the acoustic paths shown in Fig. 3.1 with the free field approximation of Eq. (2.1).

$$\mathbf{C}(j\omega)=\frac{\rho_0}{4\pi}\begin{bmatrix}\dfrac{e^{-jkl_{11}}}{l_{11}} & \dfrac{e^{-jkl_{12}}}{l_{12}}\\ \dfrac{e^{-jkl_{21}}}{l_{21}} & \dfrac{e^{-jkl_{22}}}{l_{22}}\end{bmatrix} \tag{3.28}$$

where Fig. 3.1 shows the distances l_{11}, l_{12}, l_{21}, and l_{22}. In this case, the inverse of the plant matrix $\mathbf{C}^{-1}$ has a simple analytical solution and the cross-talk cancellation matrix $\mathbf{X}$ (Eq. (3.24)) becomes,

$$\mathbf{X}(j\omega)=\mathbf{C}(j\omega)^{-1}=\frac{4\pi}{\rho_0}\left(\frac{1}{\dfrac{e^{-jk(l_{11}+l_{22})}}{l_{11}l_{22}}-\dfrac{e^{-jk(l_{12}+l_{21})}}{l_{12}l_{21}}}\right)\begin{bmatrix}\dfrac{e^{-jkl_{22}}}{l_{22}} & -\dfrac{e^{-jkl_{12}}}{l_{12}}\\ -\dfrac{e^{-jkl_{21}}}{l_{21}} & \dfrac{e^{-jkl_{11}}}{l_{11}}\end{bmatrix}. \tag{3.29}$$

Rearranging Eq. (3.29) by multiplying the top and bottom of the expression in the brackets by $l_{11}l_{22}e^{jk(l_{11}+l_{22})}$ yields Eq. (3.30),

$$\mathbf{X}(j\omega)=\frac{4\pi}{\rho_0}\left(\frac{1}{1-\dfrac{l_{11}l_{22}}{l_{12}l_{21}}e^{-jk(l_{12}+l_{21}-l_{11}-l_{22})}}\right)\begin{bmatrix}l_{11}e^{jkl_{11}} & -\dfrac{l_{11}l_{22}}{l_{12}}e^{-jk(l_{12}-l_{11}-l_{22})}\\ -\dfrac{l_{11}l_{22}}{l_{21}}e^{-jk(l_{21}-l_{11}-l_{22})} & l_{22}e^{jkl_{22}}\end{bmatrix} \tag{3.30}$$

The expression in the round brackets of Eq. (3.30),

$$\frac{1}{1-\dfrac{l_{11}l_{22}}{l_{12}l_{21}}e^{-jk(l_{12}+l_{21}-l_{11}-l_{22})}}, \tag{3.31}$$

approaches infinity at frequencies $f = 0, f_r, 2f_r, 3f_r, \cdots$ for $\frac{l_{11}l_{22}}{l_{21}l_{12}} \approx 1$, where f_r is known as the "ringing" frequency and related to the path length differences by

$$f_r = \frac{c_0}{l_{12} + l_{21} - l_{11} - l_{22}} \tag{3.32}$$

Using Eq. (3.10) for the definition of the path length difference Δl one finds that the "ringing" frequency f_r corresponds to the first ill-conditioned frequency (besides zero hertz (0 Hz)) when $\Delta l = \lambda / 2$. Employing the relationship for a geometric series of the type

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \qquad \text{(for } |x|<1\text{)} \tag{3.33}$$

to Eq. (3.30) yields,

$$\mathbf{X}(j\omega) = \frac{4\pi}{\rho_0} \begin{bmatrix} l_{11} e^{jkl_{11}} & -\frac{l_{11}l_{22}}{l_{12}} e^{-jk(l_{12}-l_{11}-l_{22})} \\ -\frac{l_{11}l_{22}}{l_{21}} e^{-jk(l_{21}-l_{11}-l_{22})} & l_{22} e^{jkl_{22}} \end{bmatrix} \sum_{n=0}^{\infty} \left(\frac{l_{11}l_{22}}{l_{12}l_{21}} \right)^n e^{-jk(l_{12}+l_{21}-l_{11}-l_{22})n} \tag{3.34}$$

Equation (3.34) reveals cross-talk cancellation as inherently a recursive process with the associated "ringing" frequency f_r. Figure 3.12 shows the "ringing" frequency for listener positions between ±1.5 m off-axis for the 10° Stereo Dipole source span arrangement shown in Fig. 3.1. At the on-axis location, f_r is at its minimum of about 11 kHz. As the listener moves off-axis, the "ringing" frequency increases monotonically. At about 1 m off-axis f_r passes beyond the audible frequency range.

This is exactly the behaviour of the ill-conditioned frequencies of the free field model in Fig. 3.6.

3.2.3. Time domain solution

Figure 3.2 shows a block diagram of the implementation of the virtual acoustic imaging system [45]. Along with the plant and cross-talk cancellation filter matrices (i.e. **C** and **X**), the figure shows the vector of virtual paths **a**(z) (in the z-domain). The figure also shows the vector of desired binaural signals **d** that one wishes to approximately reproduce at the listener's ears, the vector of source output signals **v** that are outputted from the sound sources and the vector of synthesised binaural signals **w** actually delivered to the listener that hopefully closely approximate the desired signals **d**.

Solving for the vector of source outputs **v** in Fig. 3.2 yields Eq. (3.35)

$$\mathbf{v} = [V_1(j\omega) \quad V_2(j\omega)]^{\mathrm{T}} = \mathbf{X}\mathbf{d} \tag{3.35}$$

where the superscript T is the transpose operator. Substituting Eq. (3.33) into Eq. (3.35) and setting $\mathbf{d} = \left[0 \quad \frac{\rho_0}{4\pi} D(j\omega)\right]^{\mathrm{T}}$ so that no signal is desired at the left ear and a signal with frequency response $D(j\omega)$ scaled by $\rho_0/4\pi$ is desired at the right ear yields the source outputs $V_1(j\omega)$ and $V_2(j\omega)$

$$V_1(j\omega) = \frac{-l_{11}l_{22}}{l_{12}} D(j\omega) e^{-jk(l_{12}-l_{11}-l_{22})} \sum_{n=0}^{\infty} g^{2n} e^{-jk\Delta l 2n}, \tag{3.36a}$$

$$V_2(j\omega) = l_{22} D(j\omega) e^{-jkl_{22}} \sum_{n=0}^{\infty} g^{2n} e^{-jk\Delta l 2n}. \tag{3.36b}$$

where Eq. (3.10) defines the path length difference Δl and $g = \sqrt{\frac{l_{11}l_{22}}{l_{12}l_{21}}}$. The inverse Fourier transform converts the source outputs of Eq. (3.36) into their time domain counterparts $v_1(t)$ and $v_2(t)$. This turns out to be an easy task by employing the "shifting property" of the inverse Fourier transform $\mathcal{F}^{-1}\{\}$, viz.

$$\mathcal{F}^{-1}\{ e^{-j2\pi ft}X(f) \} = x(t-\tau) \tag{3.37}$$

and by noting the property of the delta function $\delta(t)$ (unit impulse function)

$$\int_{-\infty}^{\infty} e^{\pm 2\pi jtf}\,\mathrm{d}f = \delta(\tau). \tag{3.38}$$

Converting the source outputs of Eq. (3.36) into their time domain counterparts yields,

$$v_1(t) = -\frac{l_{11}l_{22}}{l_{12}}\, d\left(t + \frac{l_{11} + l_{22} - l_{12}}{c_0} \right) * \left[1 + g^2\delta(t-2\tau) + g^4\delta(t-4\tau) + \cdots \right] \tag{3.39a}$$

$$v_2(t) = l_{22}\, d\left(t + \frac{l_{22}}{c_0} \right) * \left[1 + g^2\delta(t-2\tau) + g^4\delta(t-4\tau) + \cdots \right] \tag{3.39b}$$

where $*$ is the convolution operator, $d(t)$ is the time domain response of $D(j\omega)$, and $\tau = \Delta l/c_0 = 1/2f_r$, half of the "ringing" frequency's period. The second source first emits a positive pulse and then at $(l_{12} - l_{11})/c_0$ time later the first source emits a scaled negative version of the first pulse and then the second source emits another positive pulse scaled further at $(l_{12} + 2l_{21} - l_{11} - 2l_{22})/c_0$ time later. This pattern repeats forever. If g is significantly smaller than one (1) the pulses die out quickly. Each source emits a

progressively smaller pulse every 2τ. The time between pulses emitted between sources depends on the listener location. In symmetric arrangements, the time between the two sources emitting pulses is always τ. In asymmetric listener arrangements, the time between source one emitting a pulse and source two emitting a pulse is different from the time between source two emitting a pulse and source one emitting a pulse. The "ringing" frequency is always associated with the recursive solution of each source output separately.

Employing Eq. (3.39) to simulate a sound field for the 10° source span and a symmetric listener location with $d(t)$ equal to the Gaussian pulse shown in Fig. 3.13 yields Fig. 3.14. The spectrum of the pulse avoids the "ringing" frequency at about 11 kHz as shown in Fig. 3.13. Figure 3.14 shows plan views of the simulated sound field with white representing acoustic pressure values greater than one (1), black representing values less than negative one (-1), and shades of grey represent in-between values. The larger white circles are the two-monopole acoustic sources and the smaller white circles are the positions of the listener's ears in free space. The simulations utilise the free field approximation and so do not take into account reflections off the listener. The eighteen (18) frames sample the sound field with the same amount of time passing between each snapshot. The sequence of the snapshots starts at the top left frame and ends at the bottom right frame. The position of the listener is exactly on the inter-source axis. The sound field is simple and constrained in time, mainly consisting of the desired pulse. Figure 3.15 shows the output of source one (1), which appears to be quite constrained in time.

Setting $d(t)$ equal to the Gaussian pulse shown in Fig. 3.16 with the same bandwidth as the previous pulse but now centred on the "ringing" frequency yields the simulated sound field of Fig. 3.17. After the initial desired pulse passes the listener's ear the sources continue to "ring" for a long time afterward. This "ringing" occurs at the "ringing" frequency f_r with large source outputs. Figure 3.18 shows the output of source one (1) in this case. The sound field of Fig. 3.16 looks not to tolerate much head movement at any time before causing the listener to hear the loud amplitude "ringing" frequency. This illustrates the important point that avoiding the ill-

conditioned "ringing" frequency avoids long, high amplitude, undesirable, "ringing" pulses and its associated complicated sound field.

3.3. Inverse Filter Design

This section presents an overview of well-established material, which considers the design of virtual acoustic imaging filters that deliver sound signals to the listener's ears that closely approximate the signals from a virtual sound source located somewhere in 3D space [4,5,26,45,48,49,53]. Performing many of the calculations in the frequency domain greatly reduces the filter computation time. The application of the regularisation technique minimises the problems associated with ill-conditioned frequencies. Figure 3.1 shows the geometrical situation and Fig. 3.2 shows the block diagram of the design problem.

The error $\mathbf{e}(z)$ between the desired signals $\mathbf{d}(z)$ and the reproduced signals $\mathbf{w}(z)$ in Fig. 3.2 is defined

$$\mathbf{e}(z) = \mathbf{d}(z) - \mathbf{w}(z) = \mathbf{d}(z) - \mathbf{C}(z)\mathbf{v}(z). \tag{3.40}$$

To design the filters of Eq. (3.5) with a consideration of the performance of the system at delivering the desired signals to the listener's ears and the "effort" of the system to achieve this task one defines a frequency domain cost function $J(j\omega)$ that is the sum of a "performance term" $\mathbf{e}(j\omega)^{\mathrm{H}}\mathbf{e}(j\omega)$, which is a measure of how well the system reproduces the desired signals at the listener's ears and an "effort" penalty term $\beta\mathbf{v}(j\omega)^{\mathrm{H}}\mathbf{v}(j\omega)$, which is a proportional to the power that is input into the transducers.

$$J(j\omega) = \mathbf{e}^{\mathrm{H}}(j\omega)\mathbf{e}(j\omega) + \beta\mathbf{v}^{\mathrm{H}}(j\omega)\mathbf{v}(j\omega) \tag{3.41}$$

Here $e^{j\omega}$ has been substituted for z where ω is the angular frequency in radians. The regularisation parameter β is a constant positive real value that weights the "effort" that the system uses in order to minimise the sum of squared errors $\mathbf{e}(j\omega)^{\mathrm{H}}\mathbf{e}(j\omega)$. If β is zero (0) then the minimum of the cost function exactly produces the desired signals at the listener's ears no matter how much "effort" is required of the system to achieve this. Sometimes solutions of this type are beyond a given system's capabilities in practice. Increasing the regularisation parameter β reduces the solution's required "effort" at the expense of the accuracy of the reproduced sound signals. This trade-off is considered further below. The optimal source inputs $\mathbf{v}_0(j\omega)$ that minimise the cost function of Eq. (3.41) $J(j\omega)$ are

$$\mathbf{v}_0(j\omega) = \left[\mathbf{C}^{\mathrm{H}}(j\omega)\mathbf{C}(j\omega) + \beta\mathbf{I}\right]^{-1}\mathbf{C}^{\mathrm{H}}(j\omega)\mathbf{d}(j\omega). \tag{3.42}$$

By comparing Eq. (3.6) with Eq. (3.42) one finds the optimal frequency domain cross-talk cancellation filter matrix $\mathbf{X}_0(j\omega)$ that minimises the cost function $J(j\omega)$, viz.

$$\mathbf{X}_0(j\omega) = \left[\mathbf{C}^{\mathrm{H}}(j\omega)\mathbf{C}(j\omega) + \beta\mathbf{I}\right]^{-1}\mathbf{C}^{\mathrm{H}}(j\omega). \tag{3.43}$$

From Eqs. (3.5) and (3.43) the optimal frequency domain vector of virtual acoustic imaging filters $\mathbf{h}_0(j\omega)$ that minimises the cost function $J(j\omega)$ is the matrix multiplication of $\mathbf{X}_0(j\omega)$ and $\mathbf{a}(j\omega)$,

$$\mathbf{h}_0(j\omega) = \left[\mathbf{C}^{\mathrm{H}}(j\omega)\mathbf{C}(j\omega) + \beta\mathbf{I}\right]^{-1}\mathbf{C}^{\mathrm{H}}(j\omega)\mathbf{a}(j\omega). \tag{3.44}$$

Equations (3.45) and (3.46) carry out the practical application of Eqs. (3.43) and (3.44) in the discrete frequency domain.

$$\mathbf{X}_o(k) = \left[\mathbf{C}^H(k)\mathbf{C}(k) + \beta\mathbf{I}\right]^{-1}\mathbf{C}^H(k) \tag{3.45}$$

$$\mathbf{h}_o(k) = \left[\mathbf{C}^H(k)\mathbf{C}(k) + \beta\mathbf{I}\right]^{-1}\mathbf{C}^H(k)\mathbf{a}(k) \tag{3.46}$$

In these equations, k denotes discrete frequency in integer samples. In order to design a set of virtual acoustic imaging filters with this method, one needs both the virtual path responses $\mathbf{a}(n)$ and the plant responses $\mathbf{C}(n)$ by either measurements or modelling. These responses should be transformed into the discrete frequency domain with the discrete Fourier transform (DFT) and then Eq. (3.46) is applied to yield the discrete frequency responses of the virtual acoustic imaging filters $\mathbf{h}_o(k)$. This vector of filters should be then transformed into the discrete time domain by the inverse discrete Fourier transform (IDFT). To implement the modelling delay Δ of Eq. (3.4) a cyclic shift is carried out on the time responses of the filters so that the amount of modelling delay is equal to half of the IDFT size. This is the method of fast deconvolution with regularisation and is very efficient [54]. The filter designs in this thesis all employ this method.

The trade-off between "effort" and performance of the system for different values of the regularisation parameter β is considered with help of some examples. Figure 3.19 shows measured KEMAR dummy time and frequency responses of direct C_{11} and cross-talk paths C_{12} of the plant matrix $\mathbf{C}(k)$ for the Stereo Dipole acoustic imaging system with the listener at the traditional symmetric on-axis location. Note that at the symmetric on-axis location the two direct paths, C_{11} and C_{22}, are equivalent to each other as are the two cross-talk paths, C_{12} and C_{21} (see Fig. 3.1). The locations of the Stereo Dipole's two loudspeakers subtend an angle of 10° with the listener so that the sound sources are spaced relatively close together. Figure 3.19 shows that the two different paths' responses are very similar because of the proximity of the sources. The pinna notch at about 8 kHz is readily seen in the frequency responses (Figs. 3.19c,d).

Placing the plant matrix $\mathbf{C}(k)$ that is shown in Fig. 3.19 into Eq. (3.45) with values of the regularisation parameter β equal to 10^{-6}, 10^{-4}, 10^{-2}, and 1 yields the cross-talk cancellation filter matrices $\mathbf{X}_o(k)$ shown in Figs. 3.20-3.23 respectively. The smaller the value of β the slower the filters' impulse responses decay in time and the greater is number of coefficients required in order to adequately represent the filters. In addition, increasing β decreases the amplitude of the filters' impulse responses. In the frequency domain, this translates to a decrease in the magnitude of the filters' frequency response at the ill-conditioned low frequencies until β is large enough that it affects the whole frequency range as in Fig. 3.23 when $\beta = 1$. This reduction in the filters' response at low frequencies is useful for loudspeakers that have a limited response at low frequencies. The "effort" of the system is associated with the amplitude and length of the filters and decreases with increasing regularisation. To help evaluate cross-talk cancellation performance a control performance matrix $\mathbf{R}(k)$ is defined [45].

$$\mathbf{R}(k) = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = \mathbf{C}(k)\mathbf{X}(k) \tag{3.47}$$

With perfect cross-talk cancellation the magnitudes of the frequency responses of the direct paths, R_{11} and R_{22}, is unity (1) or zero decibels (0 dB) and the magnitudes of the frequency responses of the cross-talk paths, R_{12} and R_{21}, is zero (0) or negative infinity decibels ($-\infty$ dB) for the entire frequency range. In the time domain this corresponds to the direct paths, r_{11} and r_{22}, being unit impulse responses (i.e. delta functions, $\delta(n-\Delta)$) and the cross-talk paths r_{12} and r_{21} being zero (0) at all time. Figures 3.24-3.27 show the impulse and frequency responses of $\mathbf{R}(k)$ when the plant matrix $\mathbf{C}(k)$ is as shown in Fig. 3.19 and the cross-talk cancellation matrices $\mathbf{X}_o(k)$ are as shown in Figs. 3.20-3.23 respectively.

With $\beta = 10^{-6}$, Fig. 3.24a shows that the impulse response of the direct path r_{11} is almost a perfect shifted delta function (i.e. $\approx \delta(n-\Delta)$). The corresponding frequency response of the direct path R_{11} (Fig. 3.24c) appears to be very near zero decibels (0

dB) for the entire audible frequency range with next to no variability. The cross-talk path's impulse response r_{12} (Fig. 3.24b) appears to be very near to zero (0) for its entire length. Its corresponding frequency response R_{12} (Fig. 3.24d) contains much variation but is below approximately –30 dB for all of the frequency range shown. The system achieves at least 30 dB cross-talk cancellation for the entire audible frequency range when $\beta = 10^{-6}$.

From Fig. 3.25 the system achieves at least 30 dB cross-talk cancellation for frequencies above about 200 Hz when $\beta = 10^{-4}$. Below 200 Hz not much cross-talk cancellation is achieved. The impulse response of the direct path r_{11} in Fig. 3.25a appears to be a slightly distorted delta function. The direct path frequency response is around –5 dB at frequencies below about 100 Hz. The cross-talk path's impulse response r_{12} (Fig. 3.25b) contains a very small amount of energy around 20 ms.

The direct path's response is more distorted and the cross-talk path contains more energy when the regularisation parameter is increased to $\beta = 10^{-2}$ as seen in Fig. 3.26. At this value of β the system achieves at least 30 dB cross-talk cancellation at frequencies above about 1500 Hz. However, this level of cross-talk cancellation performance is not achieved around the peak at about 8 kHz. Therefore, with this amount of regularisation there is a rather limited range of frequencies where the system achieves significant cross-talk cancellation.

When $\beta = 1$ Fig. 3.27 shows that the system achieves very little, if any, cross-talk cancellation at all of the frequencies shown. The peak amplitude of the direct path's impulse response r_{11} (Fig. 3.27a) is about 40% of peak amplitude with $\beta = 10^{-6}$ (Fig. 3.24a). The cross-talk path's impulse response r_{12} (Fig. 3.27b) is starting to look increasingly like the direct path's impulse response at these higher values of β.

Clearly from Figs. 3.26 and 3.27 the values of $\beta = 10^{-2}$ and $\beta = 1$ achieve poor cross-talk cancellation. From Fig. 3.20 the value of $\beta = 10^{-6}$ results in long filter lengths and the system has to work very hard especially at low frequencies. The value of $\beta = 10^{-4}$

might then be selected as a decent compromise that will achieve reasonably good control of the sound field at the receiver points without requiring a great amount of system "effort" or large filter lengths.

Figure 3.28 shows measured KEMAR dummy HRTFs from a sound source located 45° to the front right (i.e. $\theta_v = 45°$) to both ears of KEMAR. One can take these path responses to be the virtual paths $\mathbf{a}(k)$ and use them in Eq. (3.47) in order to create virtual acoustic imaging filters that attempt to give the listener the perception of a virtual acoustic image 45° to their front right. This results in the virtual acoustic imaging filters $\mathbf{h}_o(k)$ shown in Figs. 3.29-3.32 when the plant matrix $\mathbf{C}(k)$ is that shown in Fig. 3.19 and values of β are equal to 10^{-6}, 10^{-4}, 10^{-2}, and 1 respectively.

3.4. Conclusions

The method of fast deconvolution with regularisation is a quick way to design causal, finite impulse response (FIR), inverse, virtual acoustic imaging filters. Regularisation makes the inversion process much easier, reduces the required length of the filters, and lessens the "effort" of the virtual acoustic imaging system especially at difficult to produce and ill-conditioned low frequencies.

The geometry of the virtual acoustic imaging system ultimately has great affect on the system's performance. The 10° source span is a good choice of source span for systems using a single pair of loudspeakers. The choice of a 10° source span means that within the audible frequency range the only ill-conditioned frequencies are around 0 Hz and the "ringing" frequency f_r at about 11 kHz which corresponds to a path length difference equal to half an acoustic wavelength (i.e. $\Delta l = \lambda / 2$). This is because the next ill-conditioned frequency is the first harmonic of f_r at about 22 kHz, which is out of the human auditory range.

The ill-conditioned frequencies are associated with dominance of either the in-phase or out-of-phase components of the source volume accelerations, a smaller spatial

control region or "sweet spot", and larger and longer source outputs. The loudspeakers work the hardest at ill-conditioned frequencies and control smaller spatial regions. The system is the least robust to head movement at these frequencies. Generally, it is best to avoid ill-conditioned frequencies. In contrast, the system is the most robust to head movements at well-conditioned frequencies, which require smaller and shorter source outputs.

At off–axis asymmetric listener locations the range of frequencies considered well-conditioned shifts to higher frequencies. Improving the system's control at higher frequencies by moving off-axis comes at the expense of a loss of control at lower frequencies. The range of frequencies of interest is application dependent but fairly low frequencies are particularly important for sound localisation. Thus, virtual acoustic image localisation performance might degrade if the listener moves far enough away from the inter-source axis.

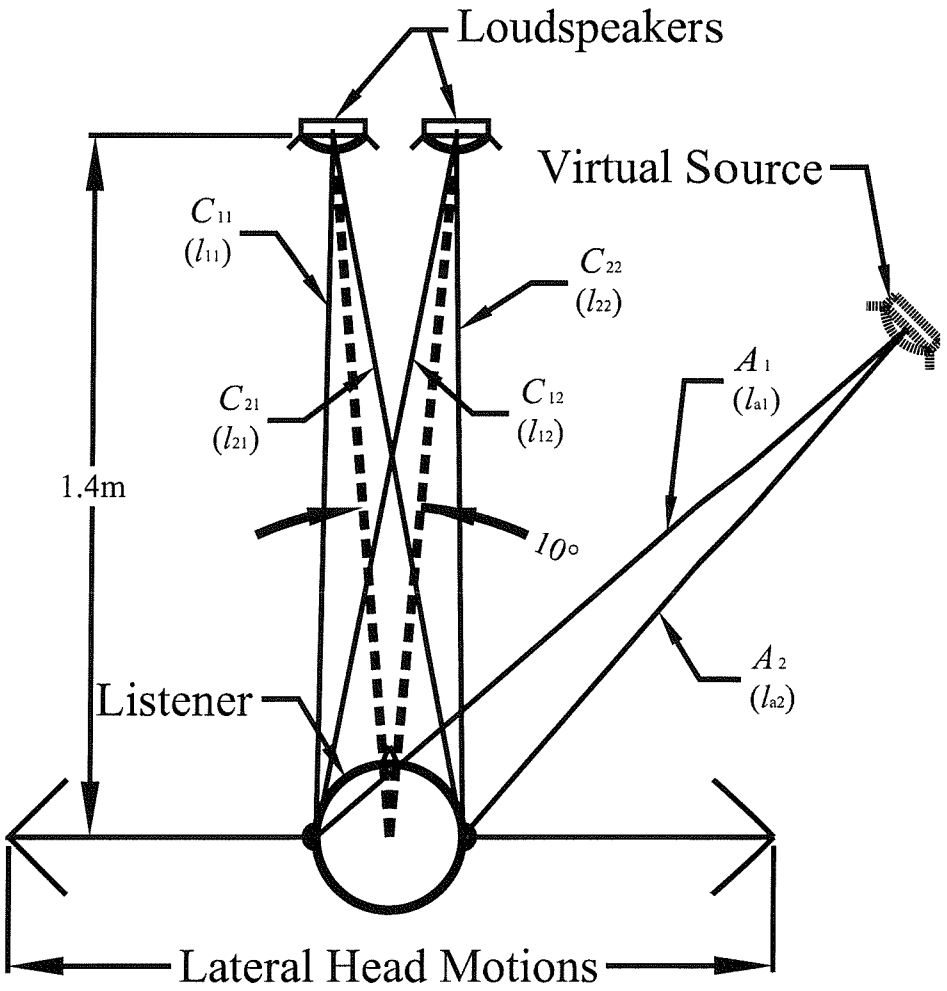


Fig. 3.1 *Plan view depicting lateral head translations and the Stereo Dipole geometrical arrangement. The elements of the electroacoustic path plant matrix* **C** *and the virtual paths* **a** *are shown as well as the corresponding lengths.*

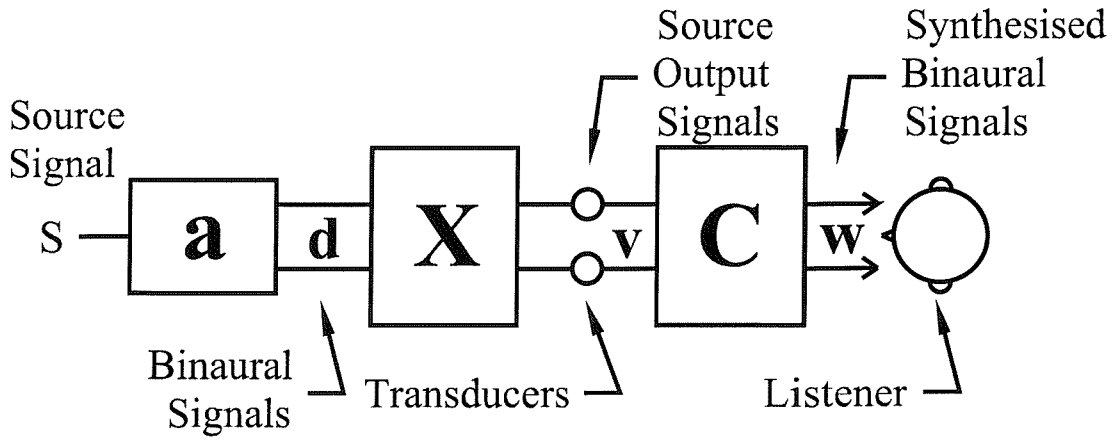


Fig. 3.2 *Block diagram of the implemented virtual acoustic imaging system.*

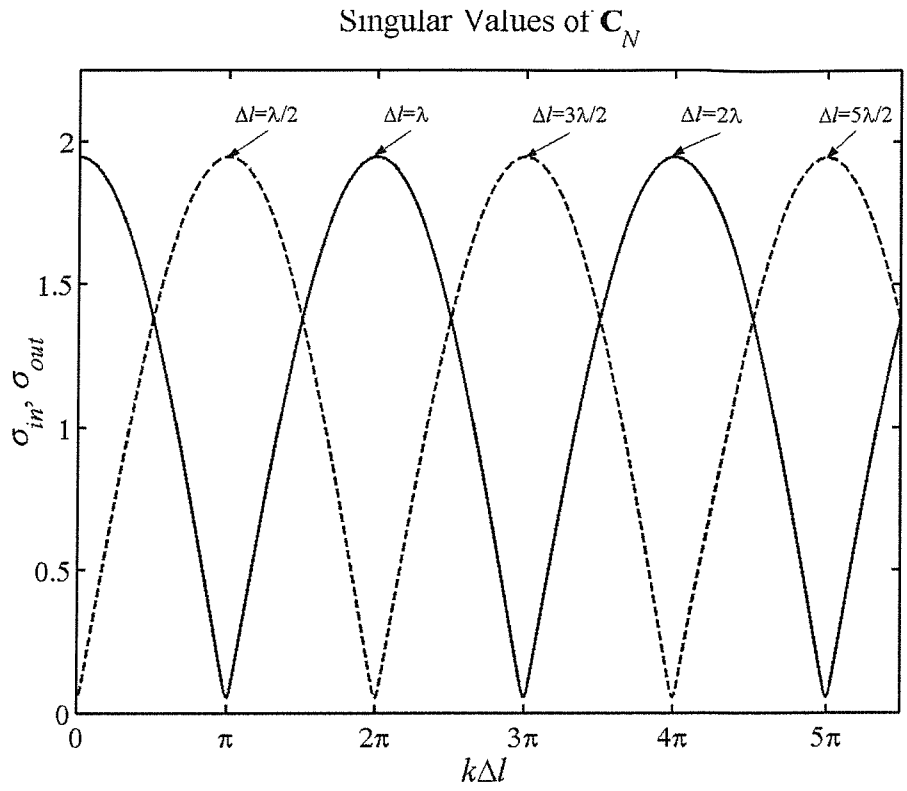


Fig. 3.3 *The singular values* σ_{in} (———) *and* σ_{out} (– – –) *of the normalised plant matrix* $\mathbf{C}_N$ *as a function of the wavenumber-multiplied by the path length difference* $k\Delta l$.

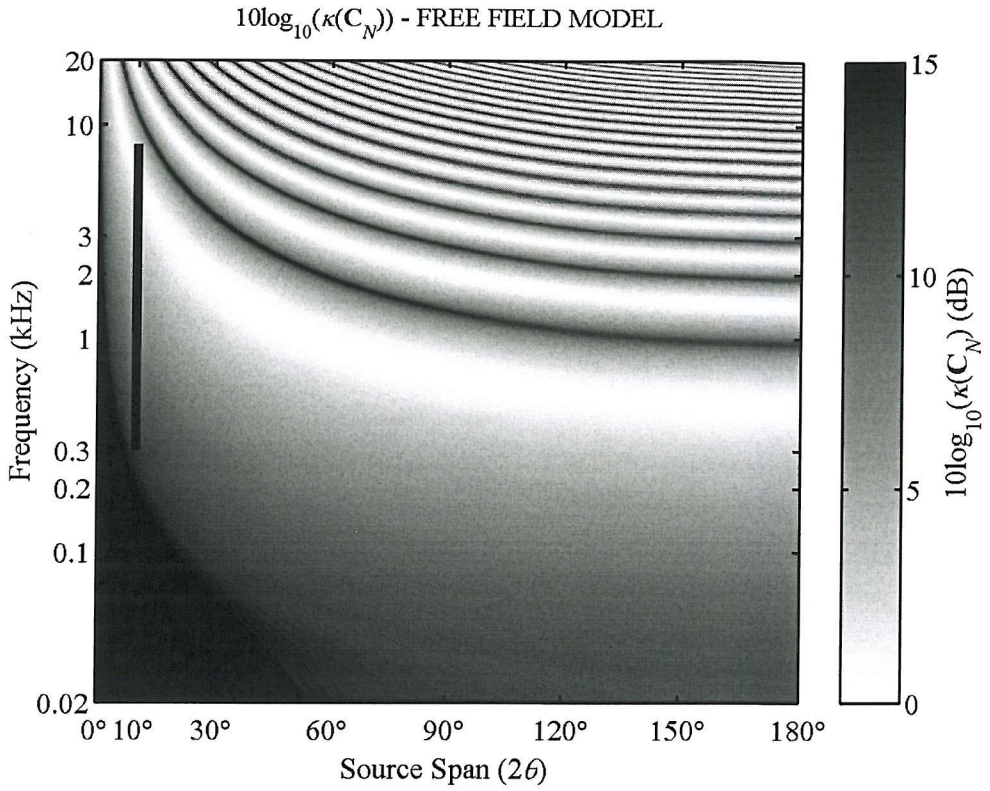


Fig. 3.4 *The dependence of condition number on the source span. The vertical lines show the choice of source span for the Stereo Dipole system.*

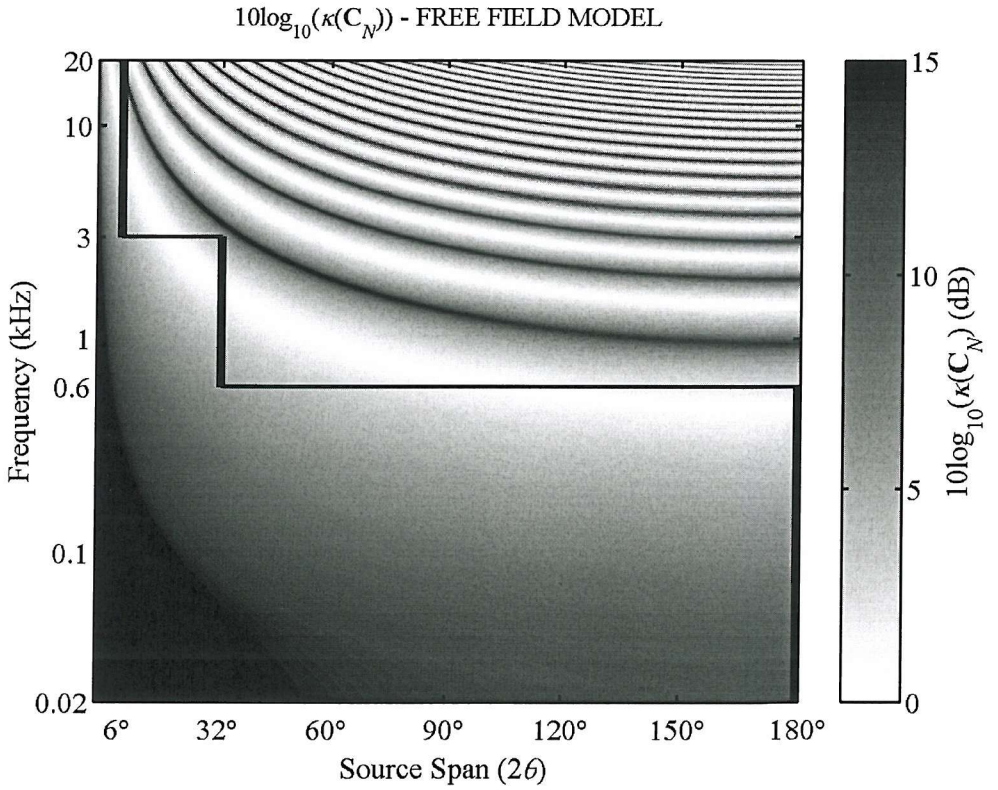


Fig. 3.5 *The dependence of condition number on the source span. The vertical lines show the choice of source span for the Optimal Source Distribution (OSD) system.*

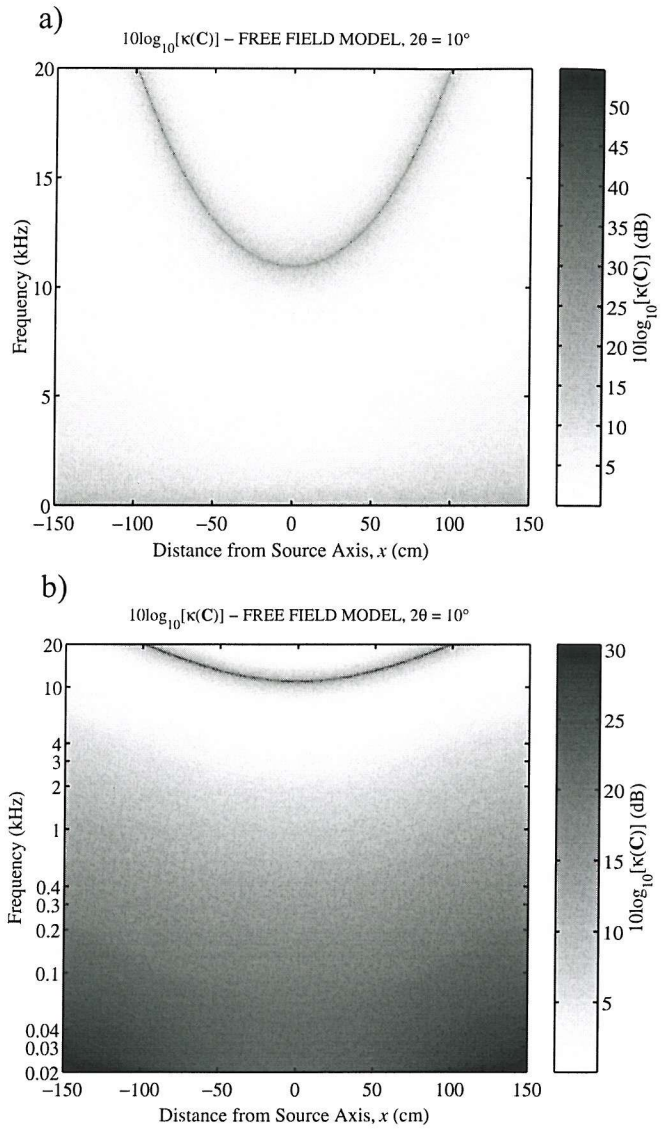


Fig. 3.6 *The dependence of the plant matrix condition number κ(***C***) on the listener location for the Stereo Dipole system based on the free field approximation shown with a) a linear frequency scale and b) a logarithmic frequency scale.*

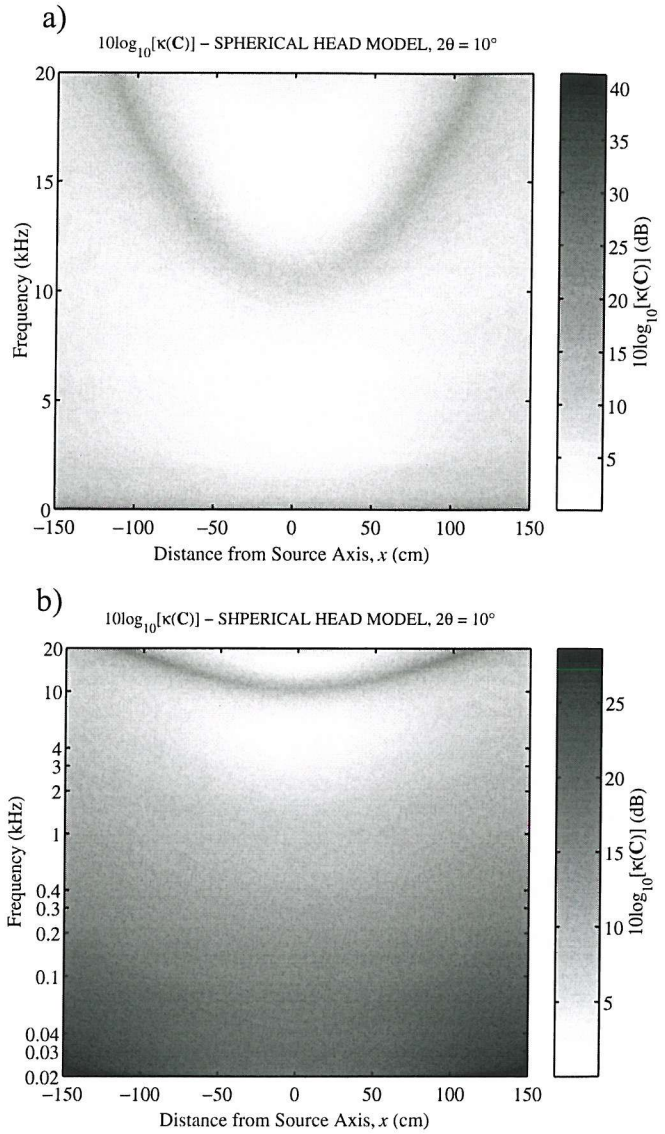


Fig. 3.7 *The dependence of the plant matrix condition number κ(***C***) on the listener location for the Stereo Dipole system based on the spherical head model shown with a) a linear frequency scale and b) a logarithmic frequency scale.*

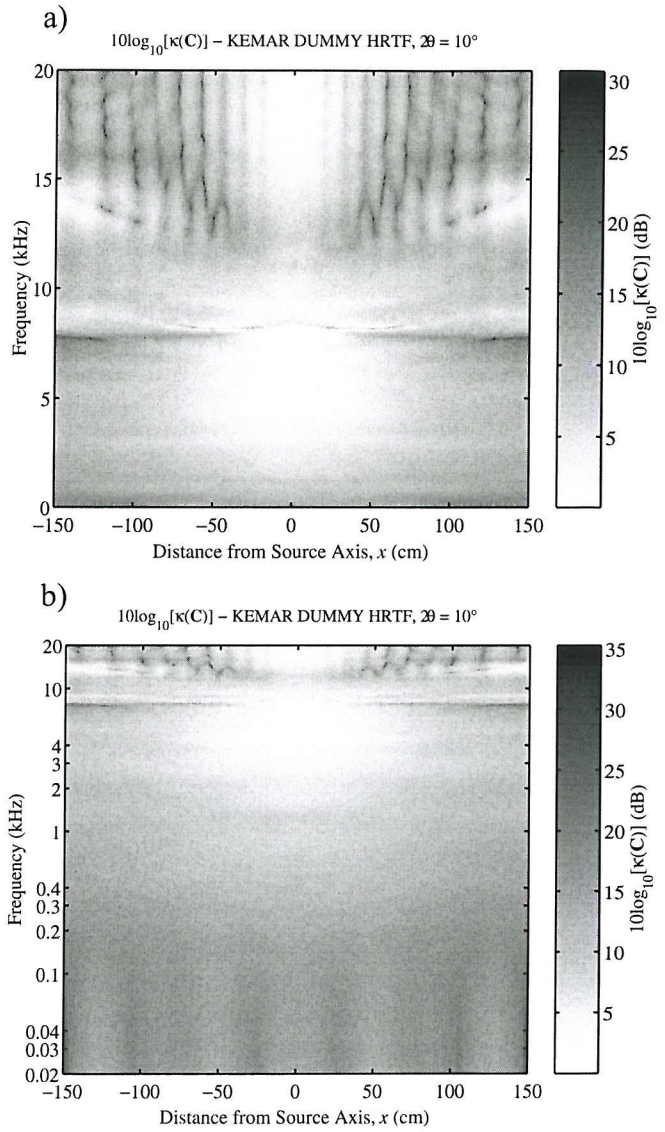


Fig. 3.8 *The dependence of the plant matrix condition number κ(**C**) on the listener location for the Stereo Dipole system based on the KEMAR dummy HRTFs shown with a) a linear frequency scale and b) a logarithmic frequency scale.*

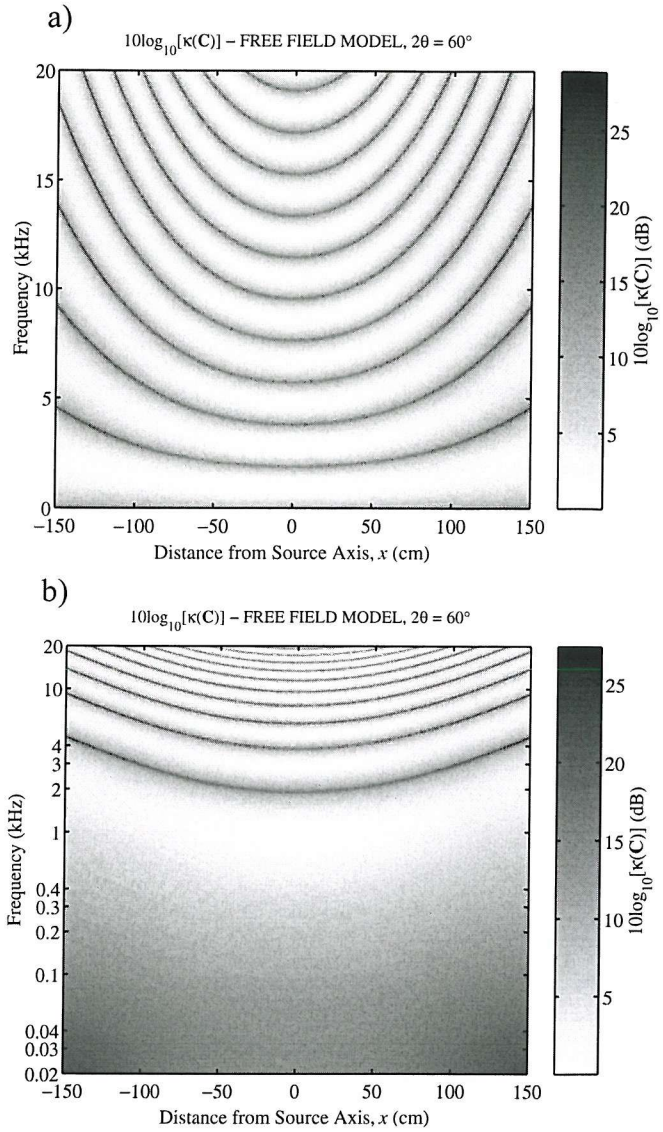


Fig. 3.9 *The dependence of the plant matrix condition number κ(***C***) on the listener location for a system with a traditional 60° source span 2θ arrangement based on the free field approximation shown with a) a linear frequency scale and b) a logarithmic frequency scale.*

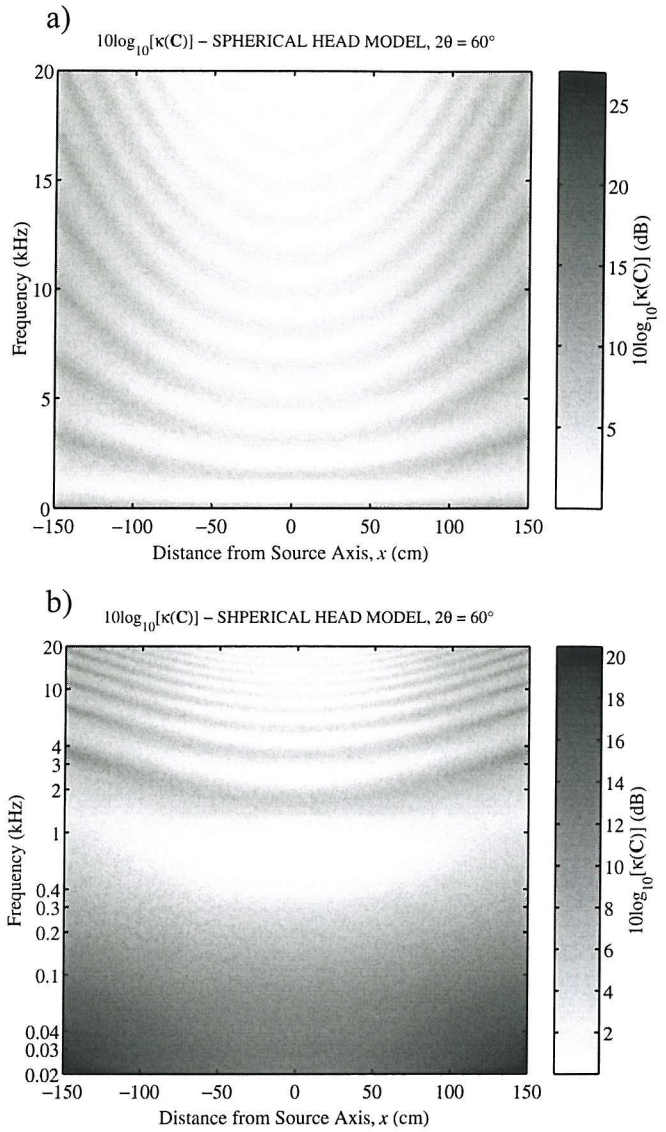


Fig. 3.10 *The dependence of the plant matrix condition number κ(**C**) on the listener location for a system with a traditional 60° source span 2θ arrangement based on the spherical head model shown with a) a linear frequency scale and b) a logarithmic frequency scale.*

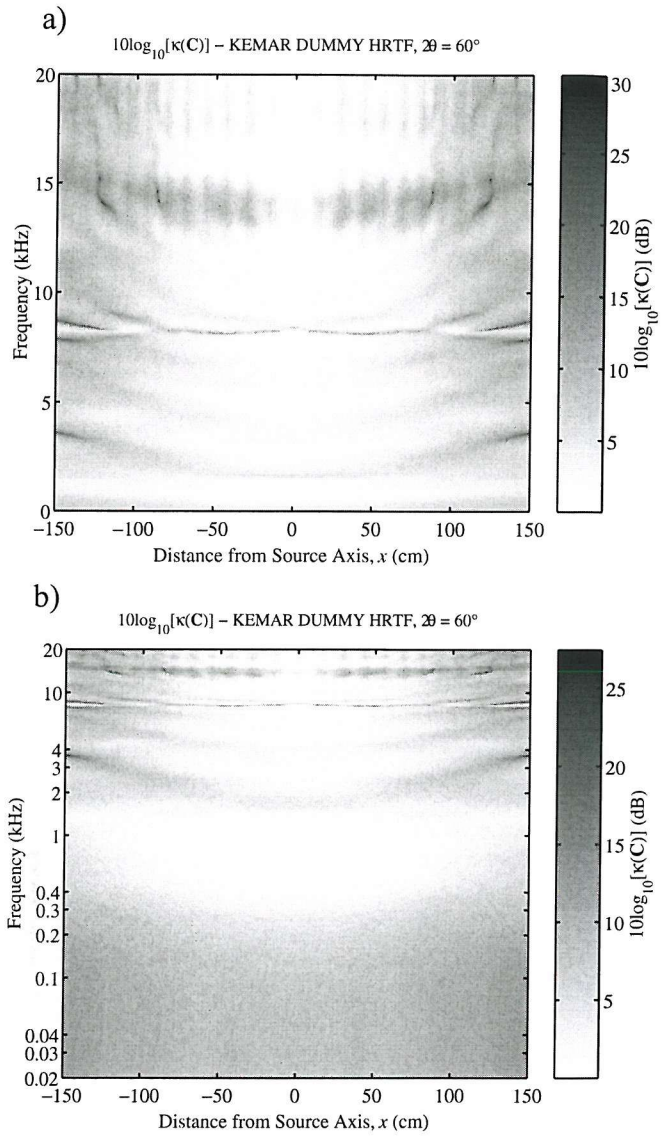


Fig. 3.11 *The dependence of the plant matrix condition number κ(**C**) on the listener location for a system with a traditional 60° source span 2θ arrangement based on KEMAR dummy HRTFs shown with a) a linear frequency scale and b) a logarithmic frequency scale.*

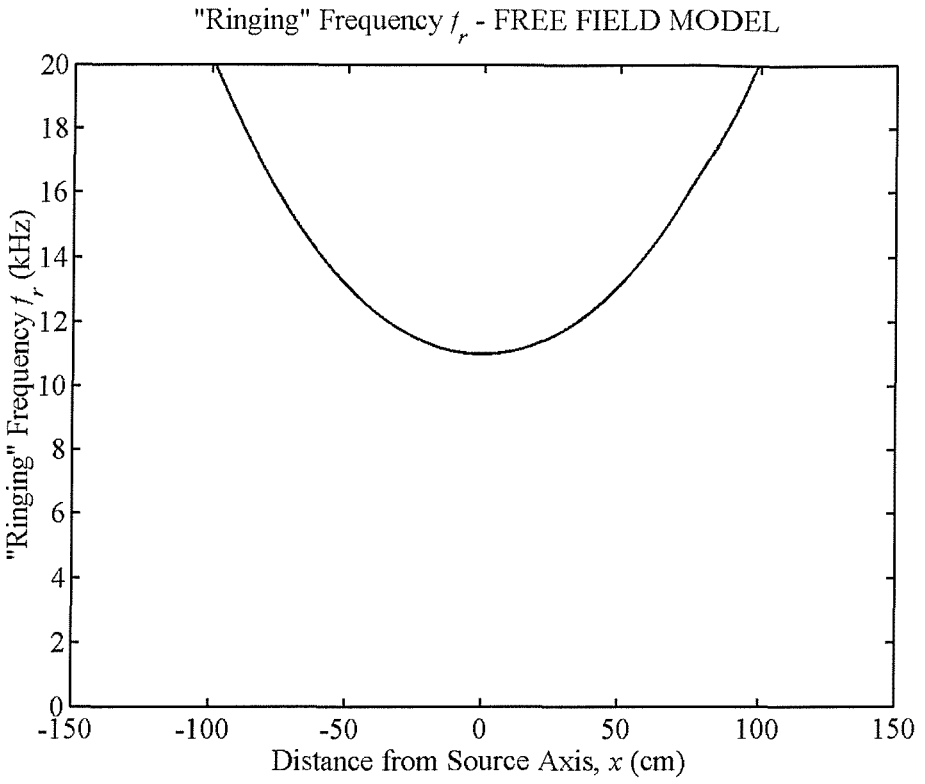


Fig. 3.12 *The dependence of the "ringing" frequency on the listener location for the Stereo Dipole system as based on the free field approximation.*

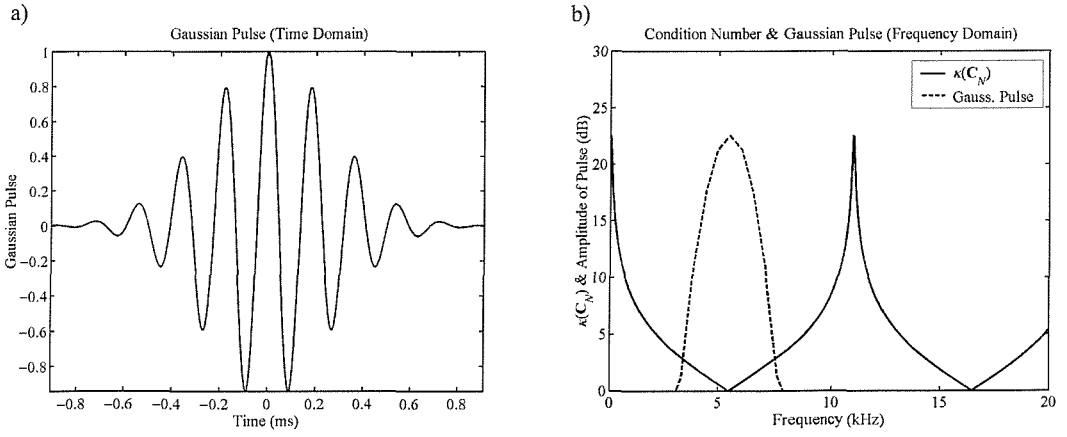


Fig. 3.13 *Time and frequency response of the Gaussian pulse used in the sound filed simulation of Fig. 3.14. The frequency plot also shows the condition number as a function of frequency for the system.*

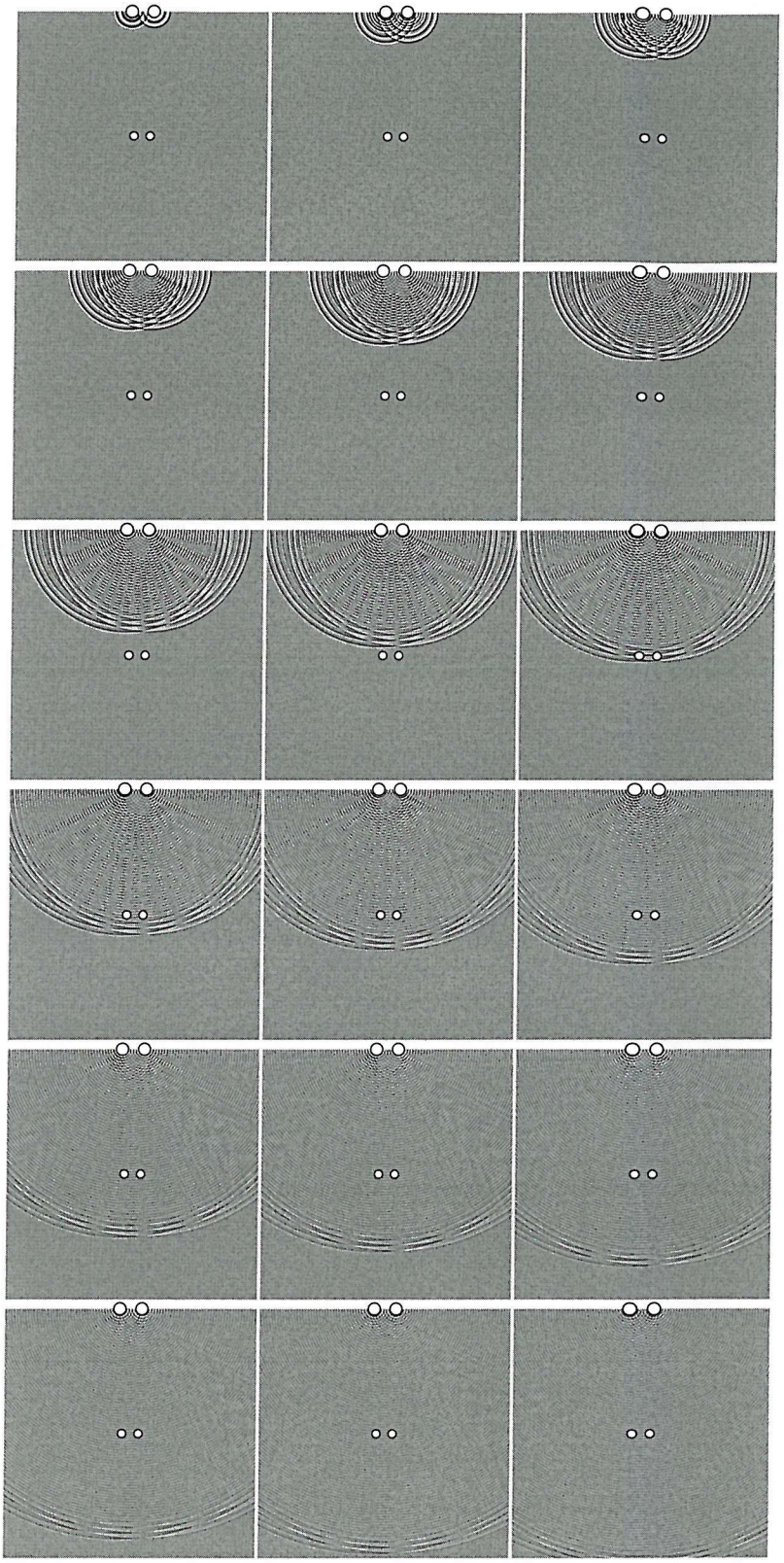


Fig. 3.14 *Sound field produced by the Stereo Dipole system when producing the Gaussian pulse of Fig. 3.13 at the left receiver (receiver one) and a zero signal at the right receiver (receiver two). The large and small white circles denote the sources and receivers respectively.*

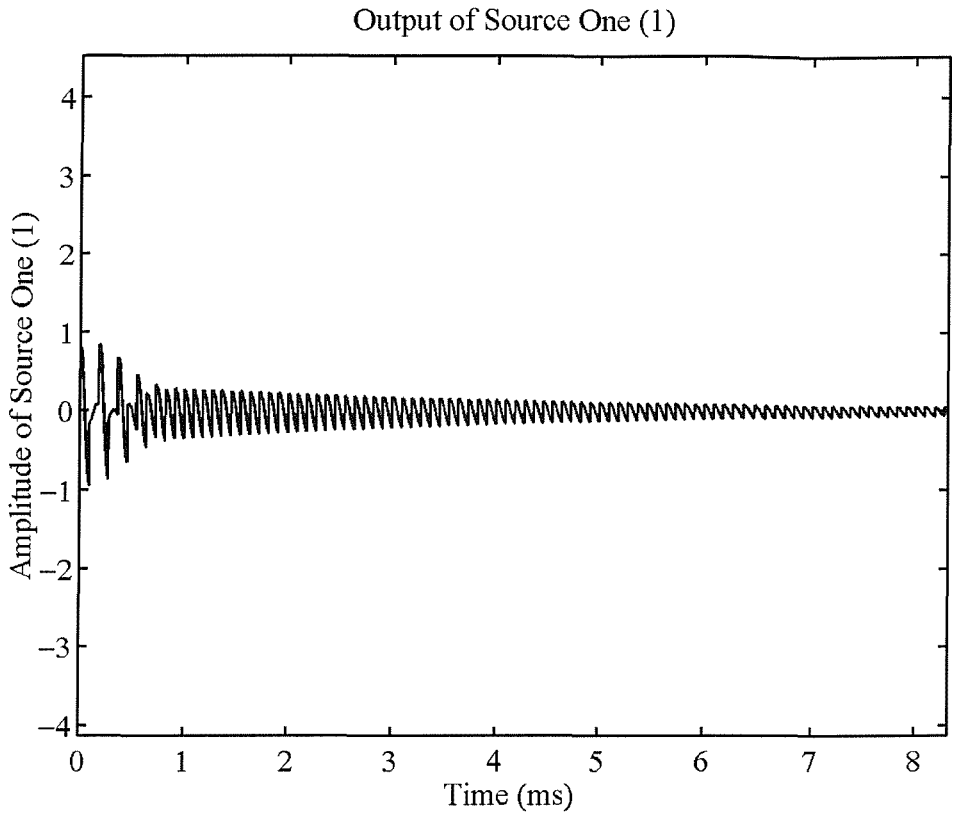


Fig. 3.15 *Output of the left source (source one) for producing the sound field shown in Fig. 3.14.*

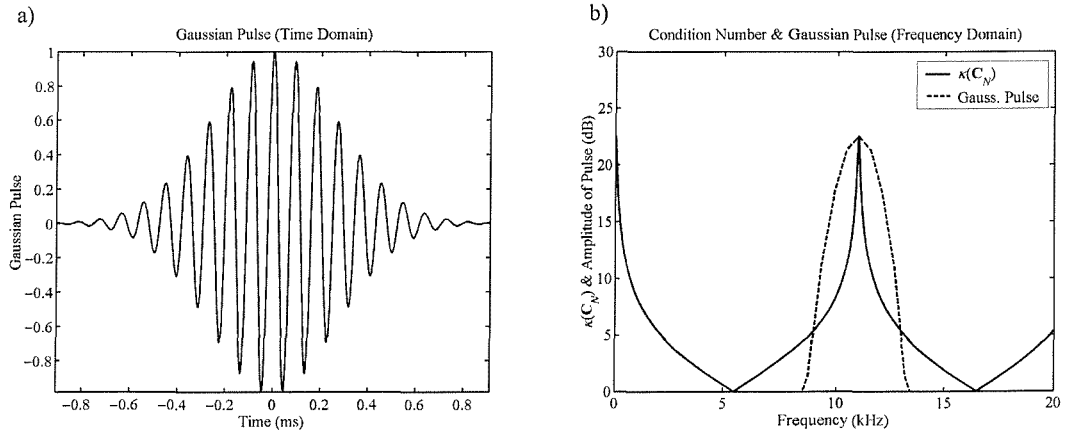


Fig. 3.16 *Time and frequency response of the Gaussian pulse used in the sound filed simulation of Fig. 3.17. The frequency plot also shows the condition number as a function of frequency for the system.*

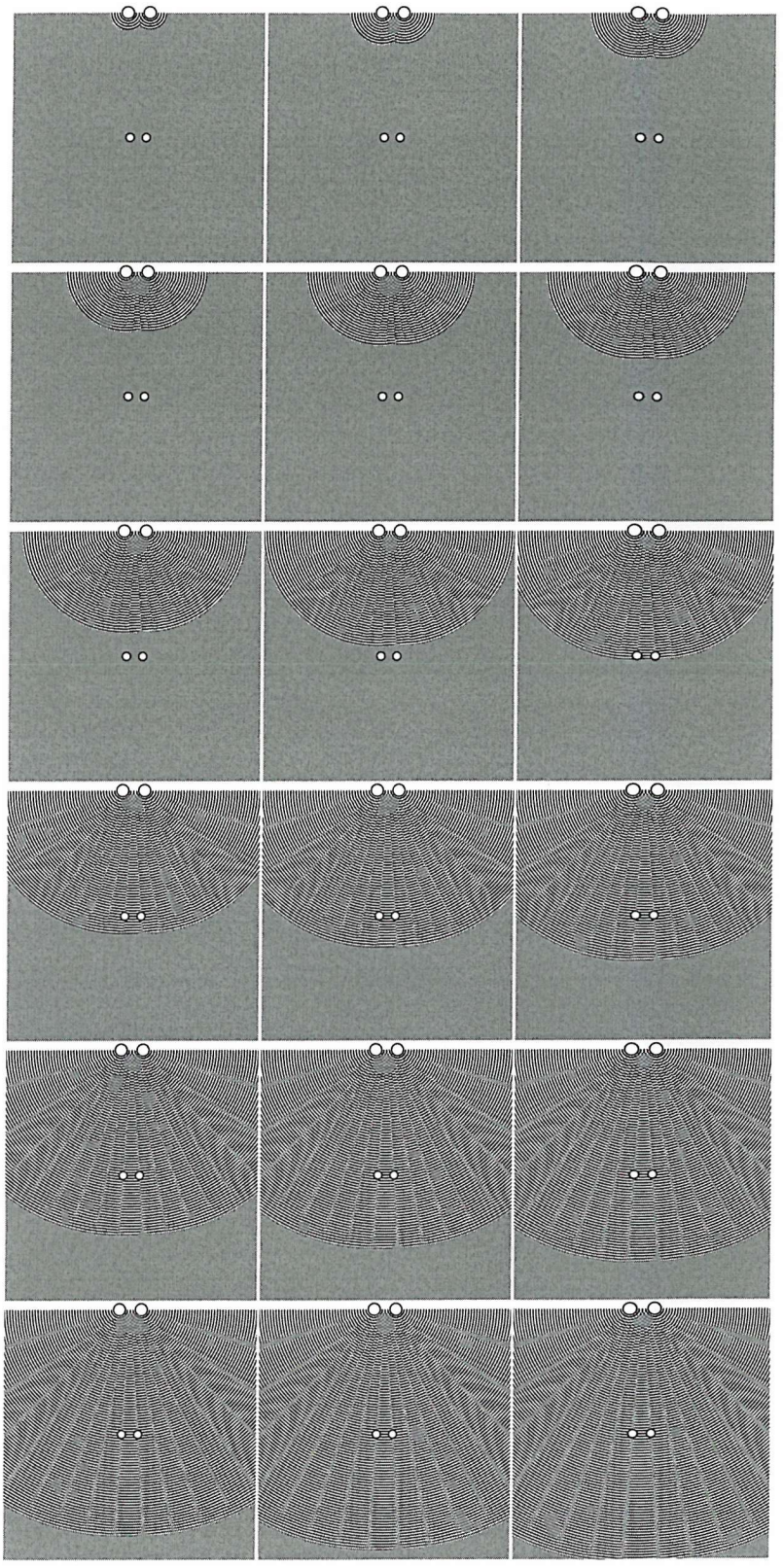


Fig. 3.17 *Sound field produced by the Stereo Dipole system when producing the Gaussian pulse of Fig. 3.16 at the left receiver (receiver one) and a zero signal at the right receiver (receiver two). The large and small white circles denote the sources and receivers respectively.*

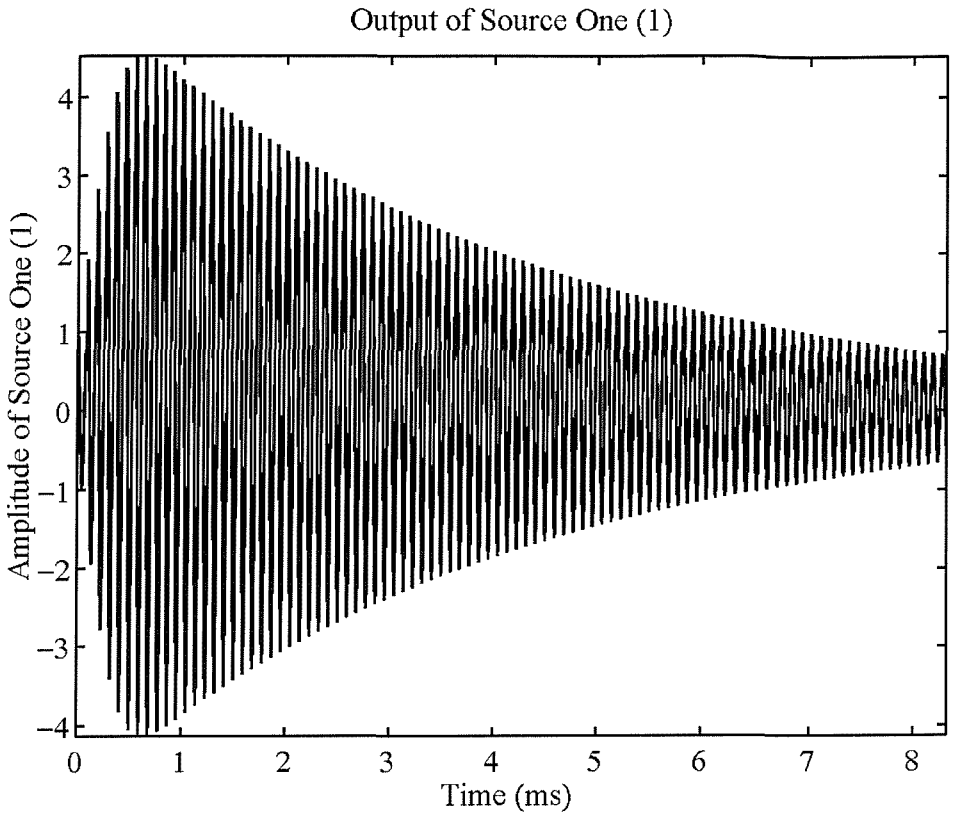


Fig. 3.18 *Output of the left source (source one) for producing the sound field shown in Fig. 3.17.*

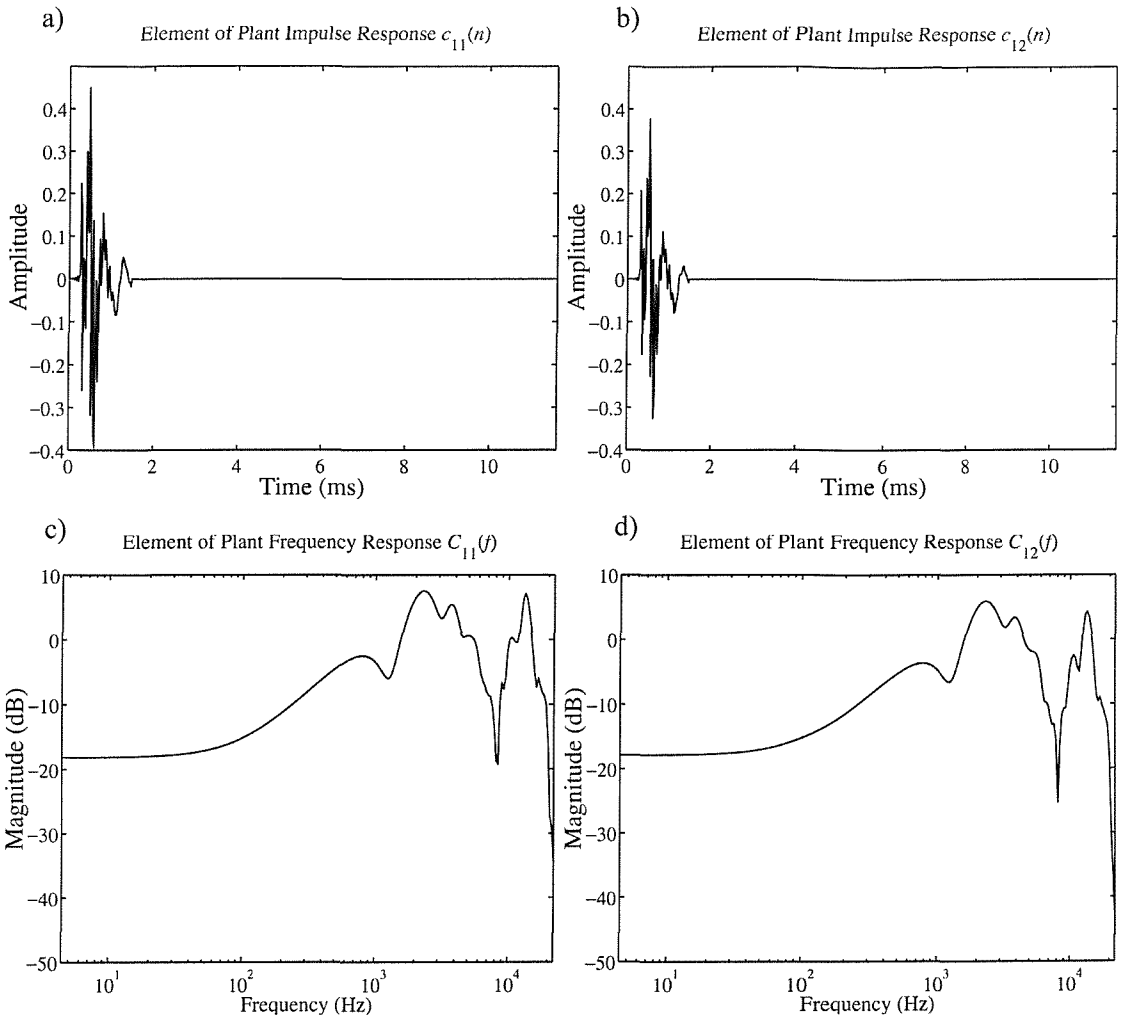


Fig. 3.19 *Plant transfer function responses for KEMAR located on-axis of the Stereo Dipole. a) impulse response of the direct paths c_{11} and c_{22}, b) impulse response of the cross-talk paths c_{12} and c_{21}, c) magnitude of the frequency response of the direct paths C_{11} and C_{22}, and d) magnitude of the frequency response of the cross-talk paths C_{12} and C_{21}.*

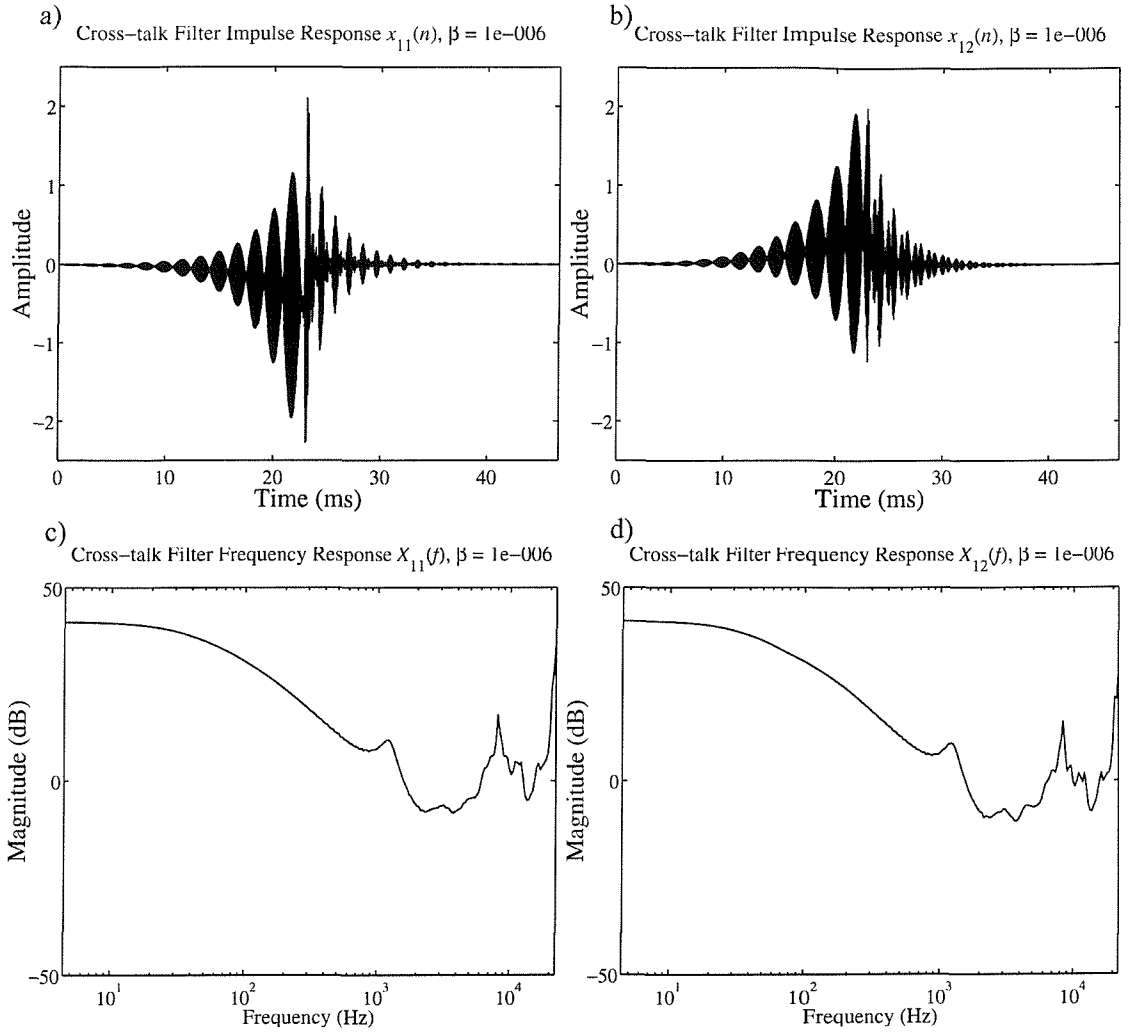


Fig. 3.20 *Cross-talk cancellation filter responses that invert the plant responses shown in Fig. 3.19 with regularisation* $\beta = 10^{-6}$. *a) impulse response of the direct path filters* x_{11} *and* x_{22}, *b) impulse response of the cross-talk path filters* x_{12} *and* x_{21}, *c) magnitude of the frequency response of the direct path filters* X_{11} *and* X_{22}, *and d) magnitude of the frequency response of the cross-talk path filters* X_{12} *and* X_{21}.

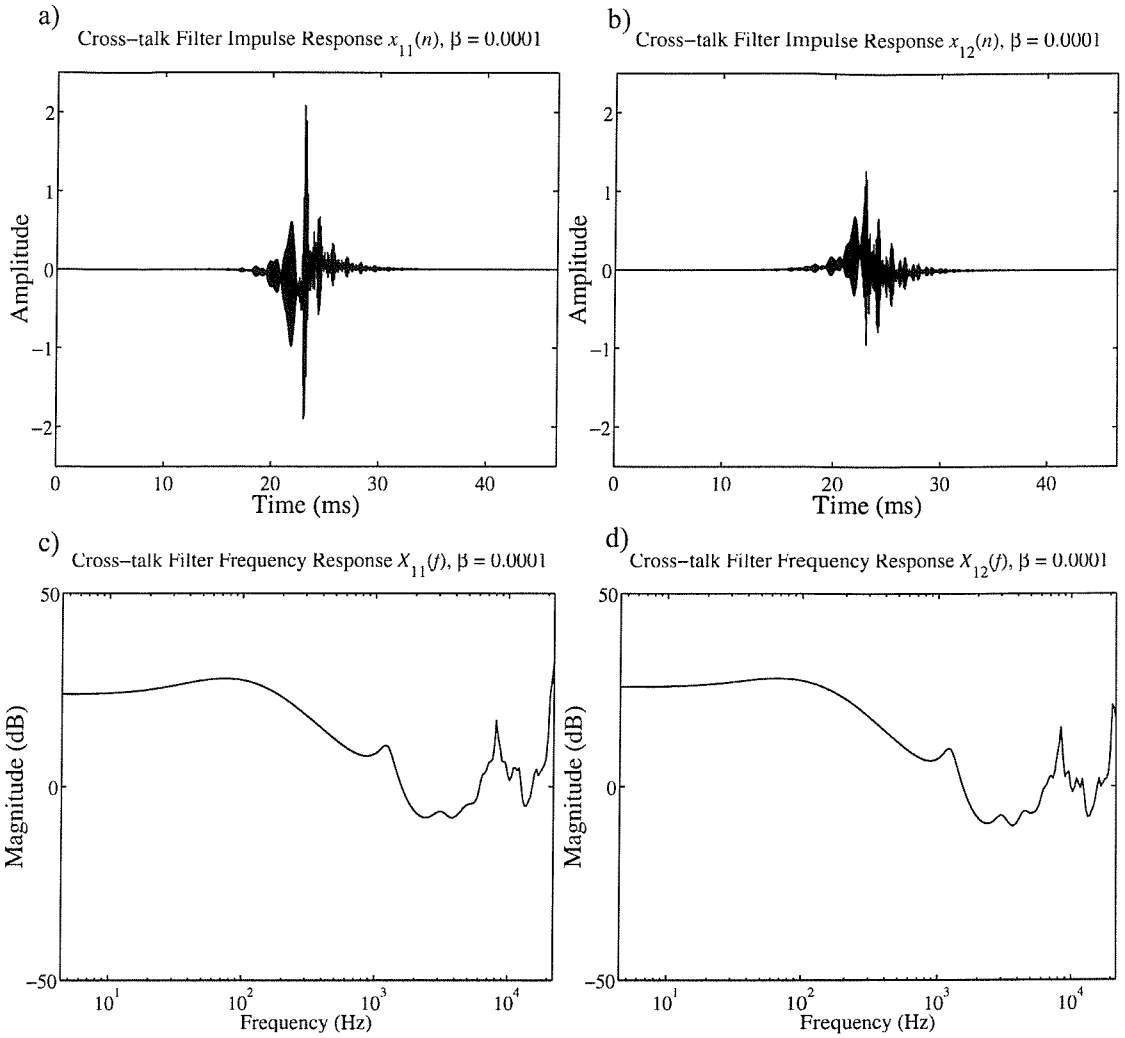


Fig. 3.21 *Cross-talk cancellation filter responses that invert the plant responses shown in Fig. 3.19 with regularisation* $\beta = 10^{-4}$. *a) impulse response of the direct path filters* x_{11} *and* x_{22}, *b) impulse response of the cross-talk path filters* x_{12} *and* x_{21}, *c) magnitude of the frequency response of the direct path filters* X_{11} *and* X_{22}, *and d) magnitude of the frequency response of the cross-talk path filters* X_{12} *and* X_{21}.

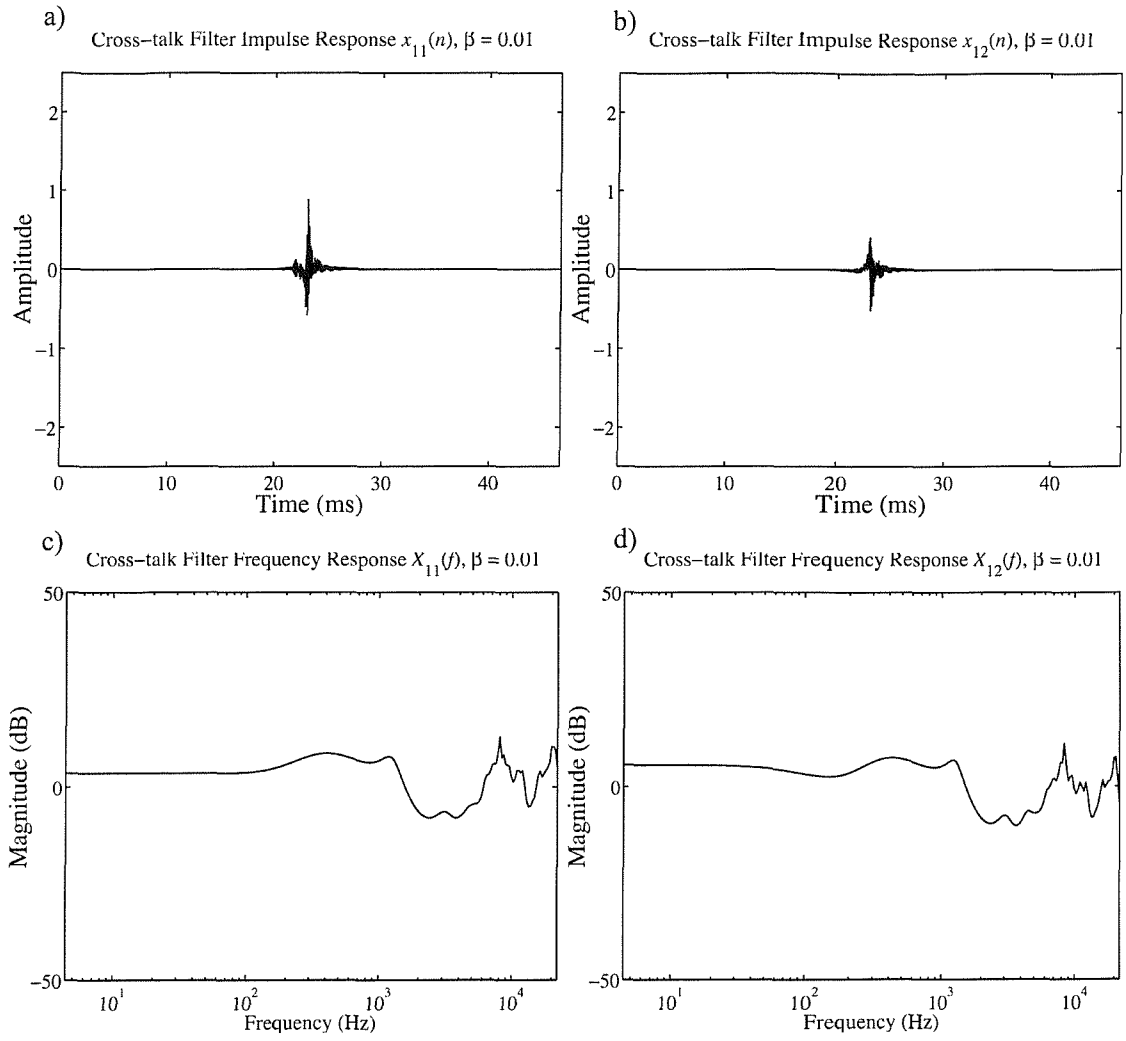


Fig. 3.22 *Cross-talk cancellation filter responses that invert the plant responses shown in Fig. 3.19 with regularisation* $\beta = 10^{-2}$. *a) impulse response of the direct path filters* x_{11} *and* x_{22}, *b) impulse response of the cross-talk path filters* x_{12} *and* x_{21}, *c) magnitude of the frequency response of the direct path filters* X_{11} *and* X_{22}, *and d) magnitude of the frequency response of the cross-talk path filters* X_{12} *and* X_{21}.

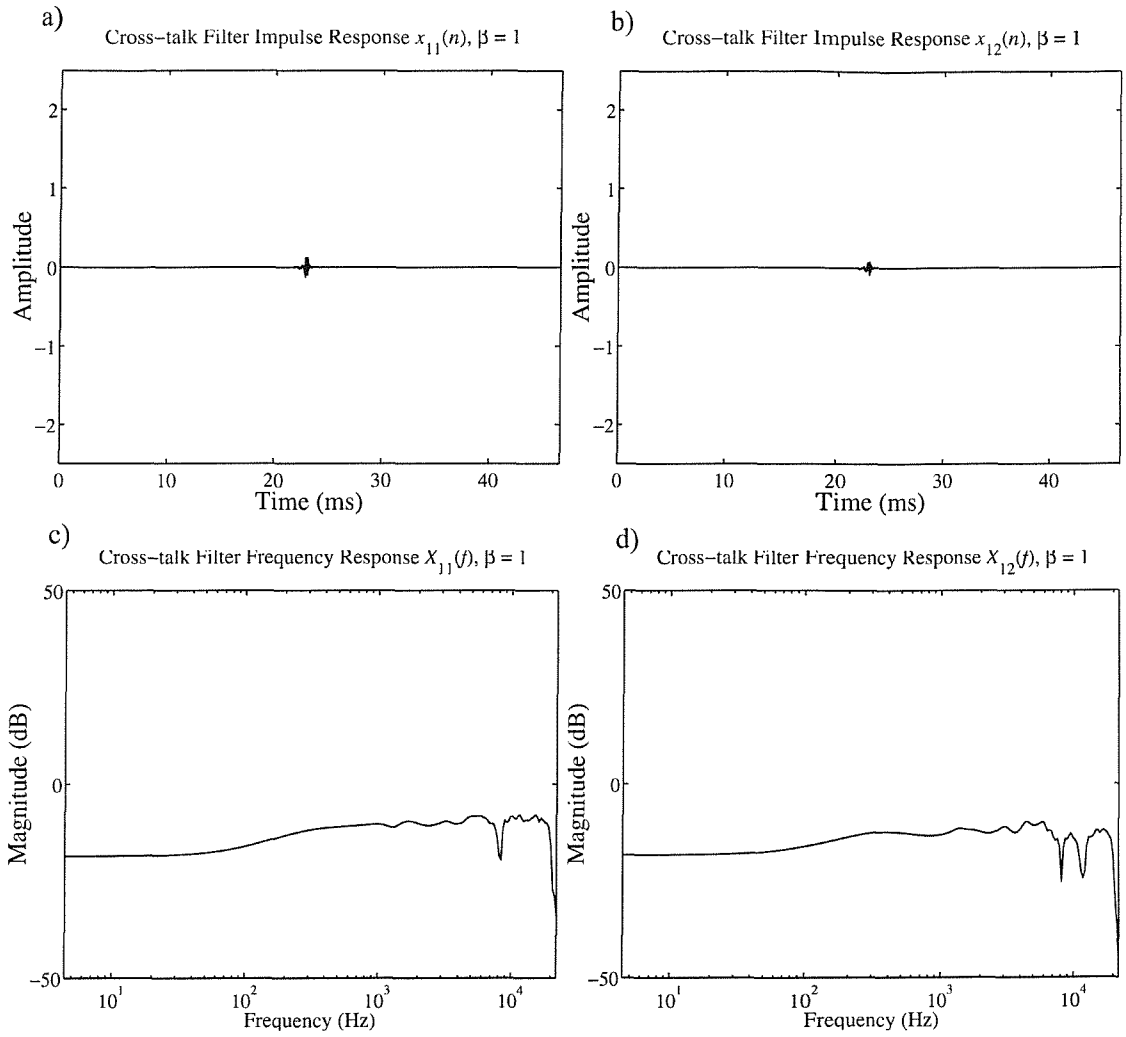


Fig. 3.23 *Cross-talk cancellation filter responses that invert the plant responses shown in Fig. 3.19 with regularisation β = 1. a) impulse response of the direct path filters x_{11} and x_{22}, b) impulse response of the cross-talk path filters x_{12} and x_{21}, c) magnitude of the frequency response of the direct path filters X_{11} and X_{22}, and d) magnitude of the frequency response of the cross-talk path filters X_{12} and X_{21}.*

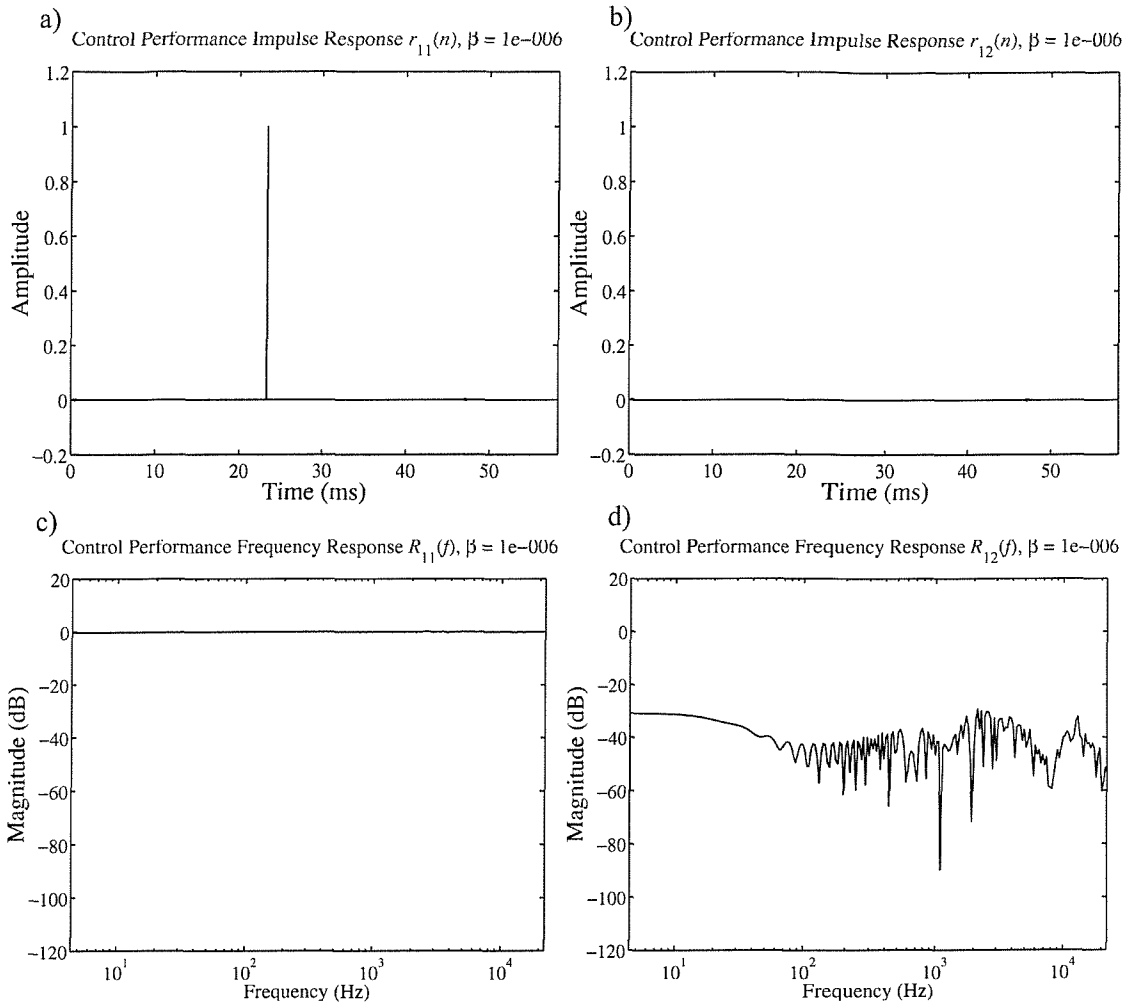


Fig. 3.24 *Control performance matrix responses for the cross-talk filters shown in Fig. 3.20 with regularisation* $\beta = 10^{-6}$. *a) impulse response of the direct paths* r_{11} *and* r_{22}, *b) impulse response of the cross-talk paths* r_{12} *and* r_{21}, *c) magnitude of the frequency response of the direct paths* R_{11} *and* R_{22}, *and d) magnitude of the frequency response of the cross-talk paths* R_{12} *and* R_{21}.

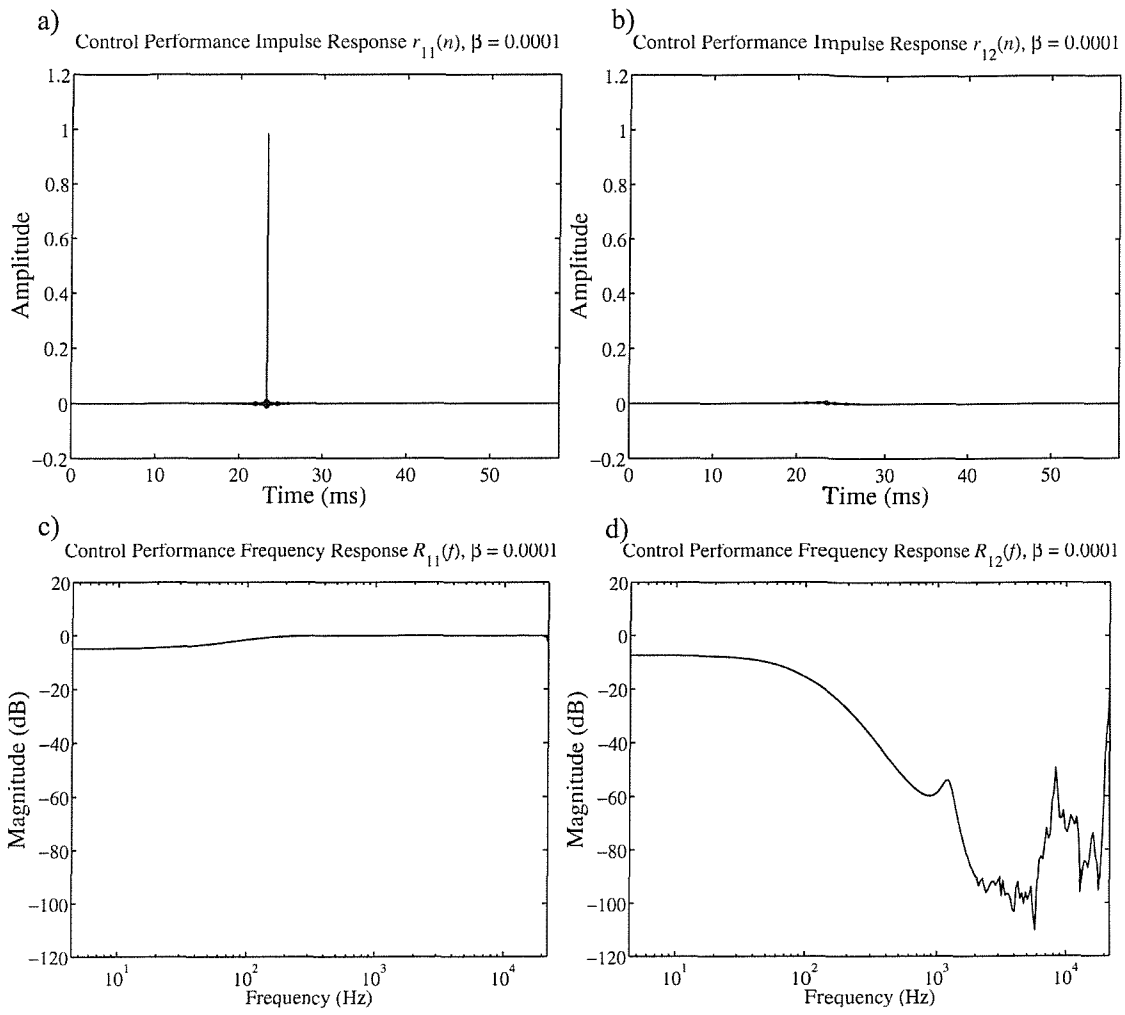


Fig. 3.25 *Control performance matrix responses for the cross-talk filters shown in Fig. 3.21 with regularisation* $\beta = 10^{-4}$. *a) impulse response of the direct paths* r_{11} *and* r_{22}, *b) impulse response of the cross-talk paths* r_{12} *and* r_{21}, *c) magnitude of the frequency response of the direct paths* R_{11} *and* R_{22}, *and d) magnitude of the frequency response of the cross-talk paths* R_{12} *and* R_{21}.

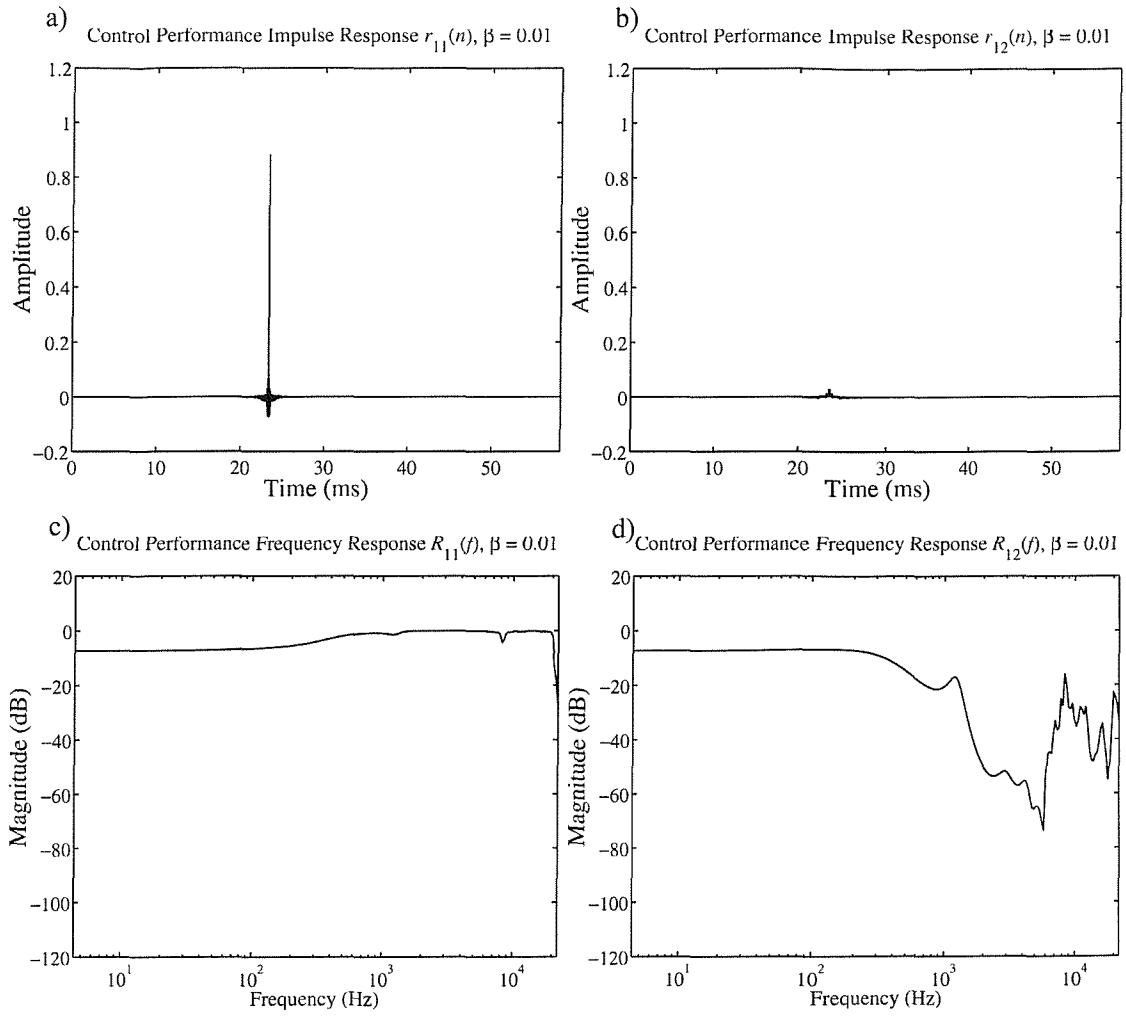


Fig. 3.26 *Control performance matrix responses for the cross-talk filters shown in Fig. 3.22 with regularisation* $\beta = 10^{-2}$. *a) impulse response of the direct paths* r_{11} *and* r_{22}, *b) impulse response of the cross-talk paths* r_{12} *and* r_{21}, *c) magnitude of the frequency response of the direct paths* R_{11} *and* R_{22}, *and d) magnitude of the frequency response of the cross-talk paths* R_{12} *and* R_{21}.

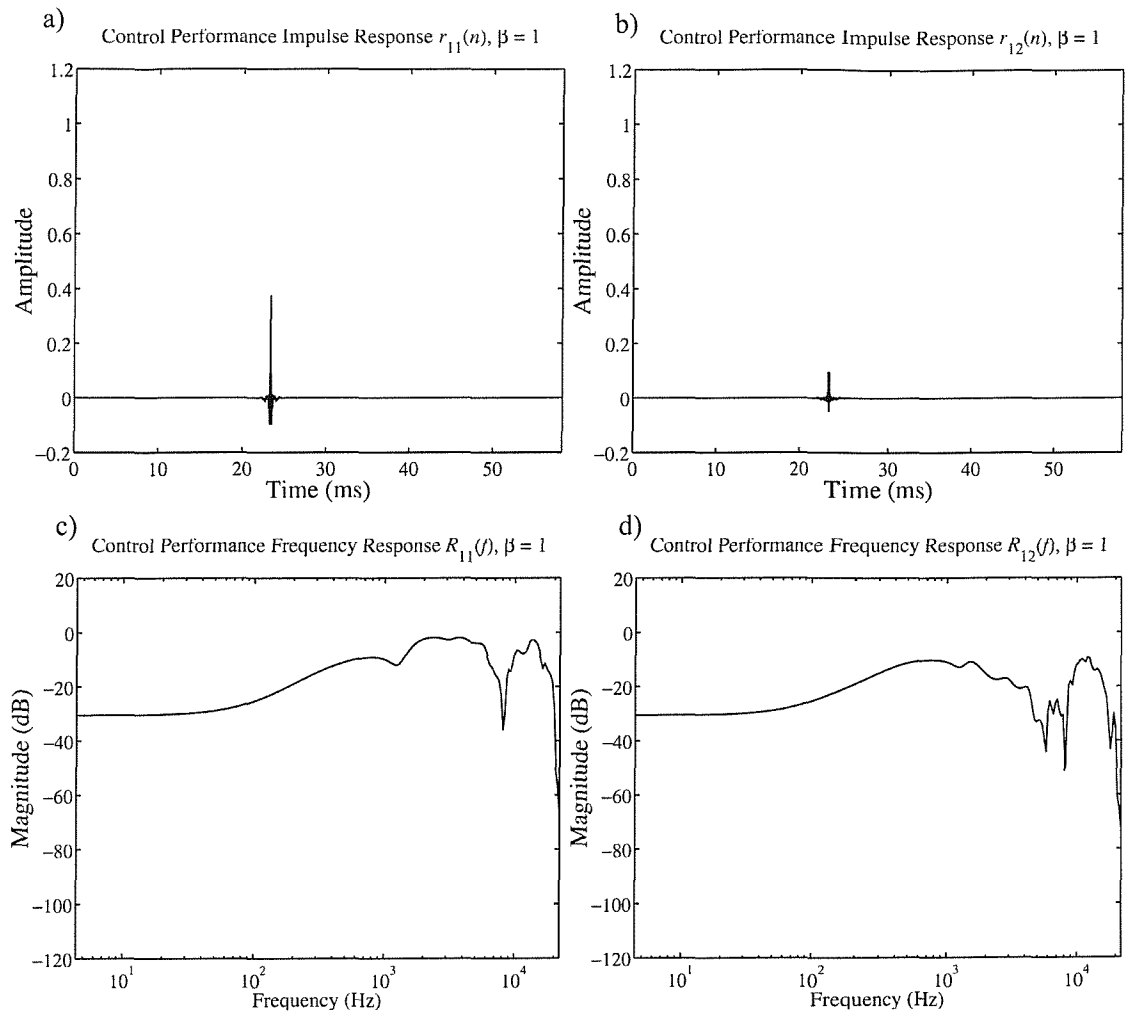


Fig. 3.27 *Control performance matrix responses for the cross-talk filters shown in Fig. 3.23 with regularisation β = 1. a) impulse response of the direct paths r_{11} and r_{22}, b) impulse response of the cross-talk paths r_{12} and r_{21}, c) magnitude of the frequency response of the direct paths R_{11} and R_{22}, and d) magnitude of the frequency response of the cross-talk paths R_{12} and R_{21}.*

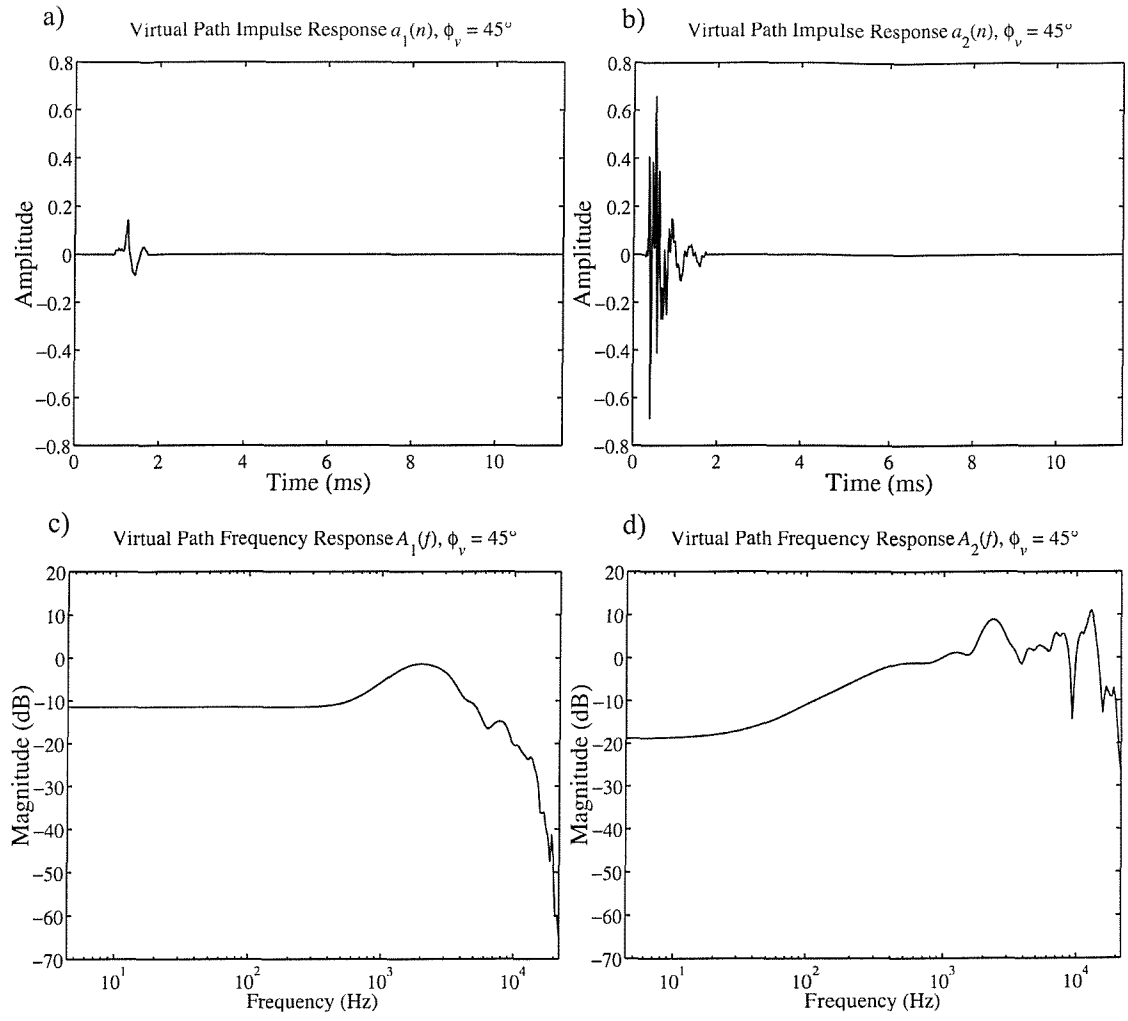


Fig. 3.28 *Virtual path transfer function responses from a sound source located 45° to the front right of KEMAR to KEMAR's ears. a) impulse response of the path to KEMAR's left ear a_1 b) impulse response of the path to KEMAR's right ear a_2, c) magnitude of the frequency response of the path to KEMAR's left ear A_1, and d) magnitude of the frequency response of the path to KEMAR's right ear A_2.*

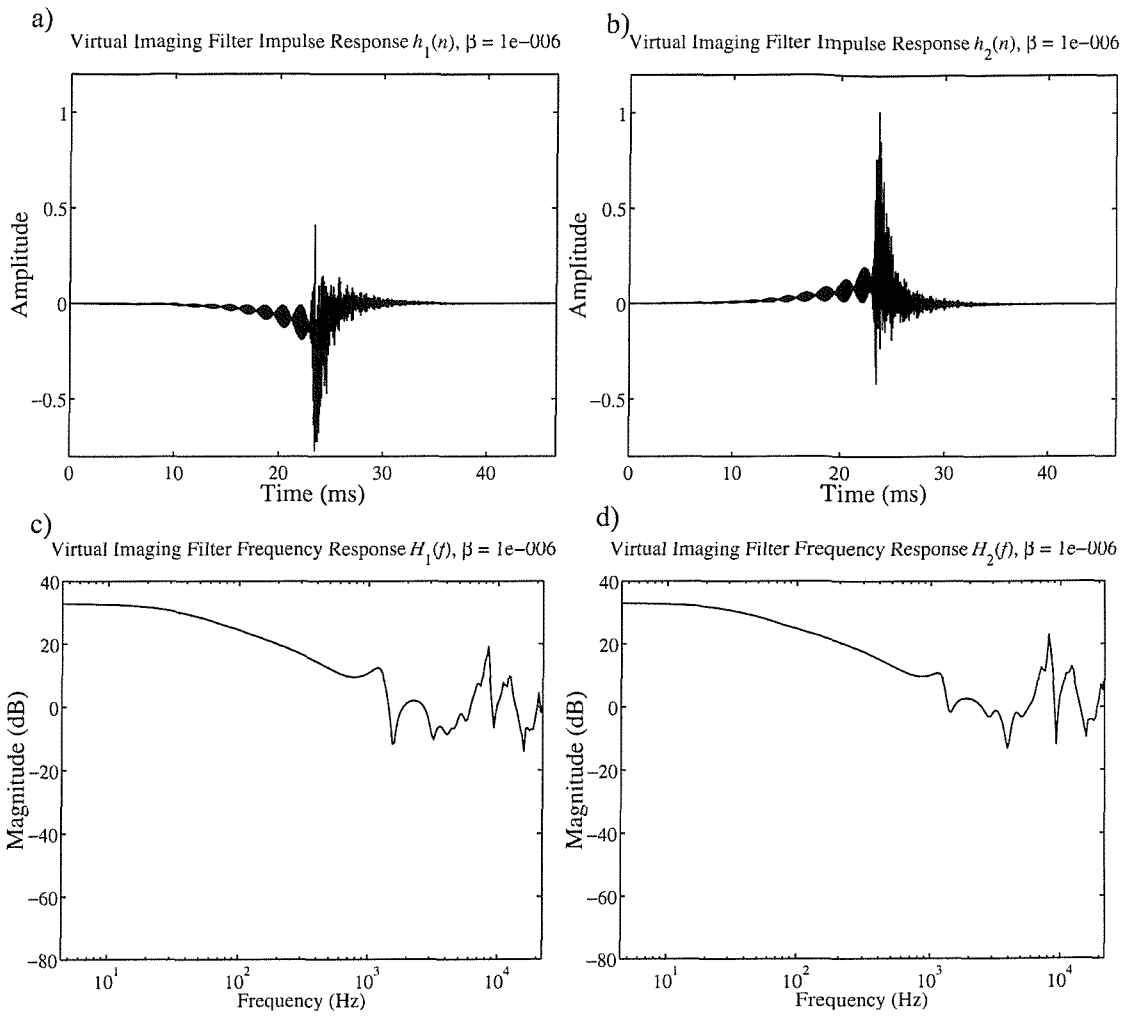


Fig. 3.29 *Virtual acoustic imaging filter responses designed from the virtual paths shown in Fig. 3.28 and the plant responses shown in Fig. 3.19 with regularisation* $\beta = 10^{-6}$. *a) impulse response of the direct path filters* h_{11} *and* h_{22}, *b) impulse response of the cross-talk path filters* h_{12} *and* h_{21}, *c) magnitude of the frequency response of the direct path filters* H_{11} *and* H_{22}, *and d) magnitude of the frequency response of the cross-talk path filters* H_{12} *and* H_{21}.

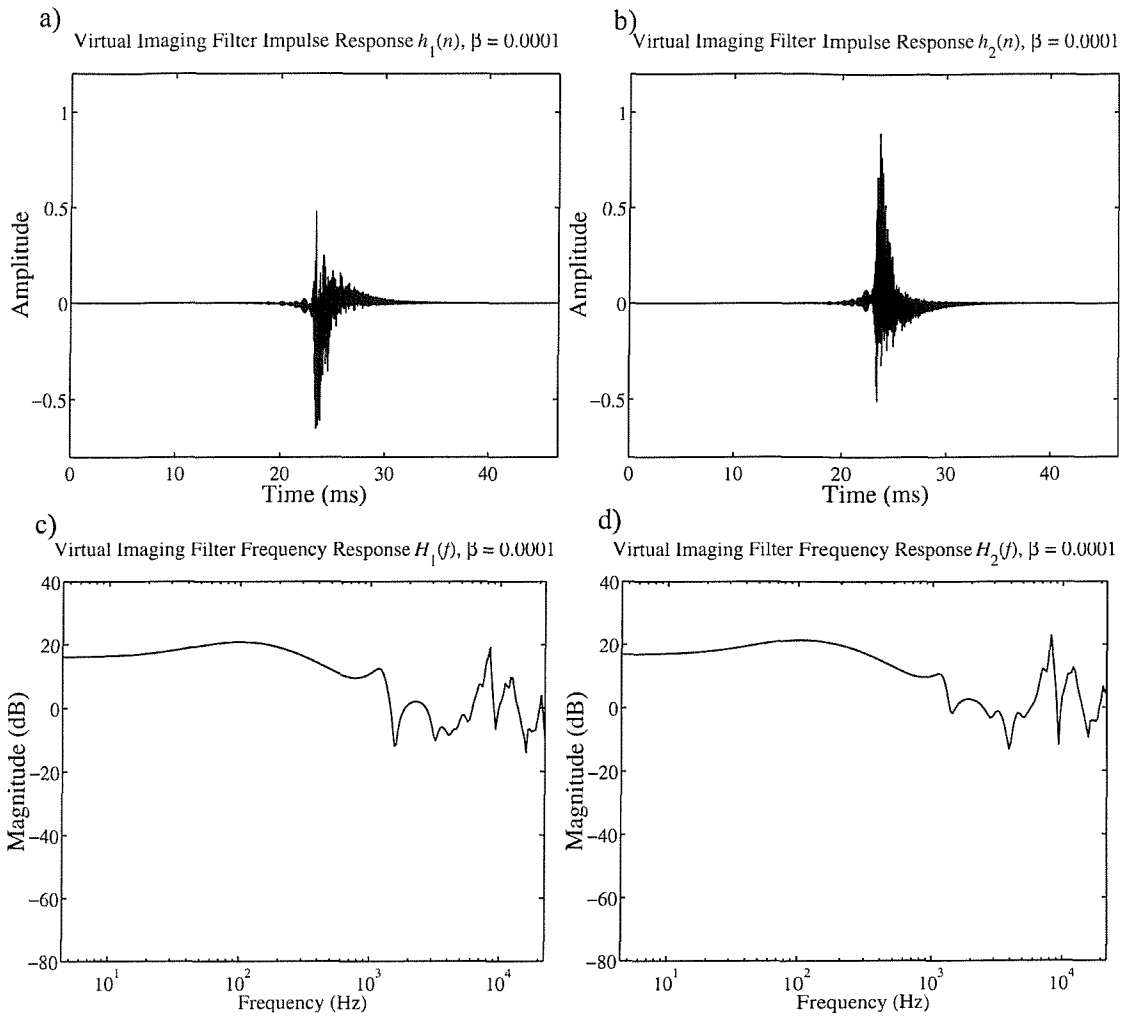


Fig. 3.30 *Virtual acoustic imaging filter responses designed from the virtual paths shown in Fig. 3.28 and the plant responses shown in Fig. 3.19 with regularisation* $\beta = 10^{-4}$. *a) impulse response of the direct path filters* h_{11} *and* h_{22}, *b) impulse response of the cross-talk path filters* h_{12} *and* h_{21}, *c) magnitude of the frequency response of the direct path filters* H_{11} *and* H_{22}, *and d) magnitude of the frequency response of the cross-talk path filters* H_{12} *and* H_{21}.

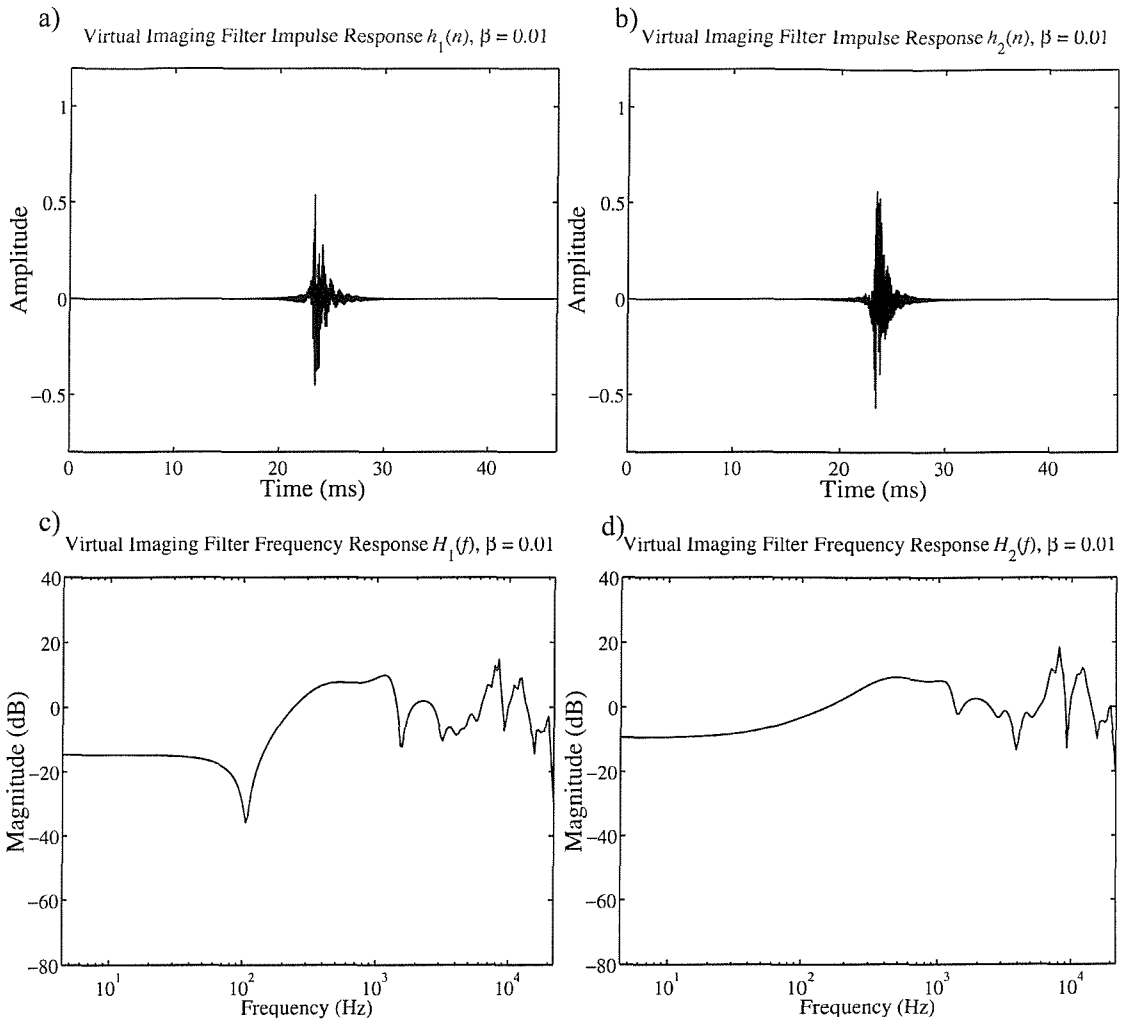


Fig. 3.31 *Virtual acoustic imaging filter responses designed from the virtual paths shown in Fig. 3.28 and the plant responses shown in Fig. 3.19 with regularisation* $\beta = 10^{-2}$. *a) impulse response of the direct path filters* h_{11} *and* h_{22}, *b) impulse response of the cross-talk path filters* h_{12} *and* h_{21}, *c) magnitude of the frequency response of the direct path filters* H_{11} *and* H_{22}, *and d) magnitude of the frequency response of the cross-talk path filters* H_{12} *and* H_{21}.

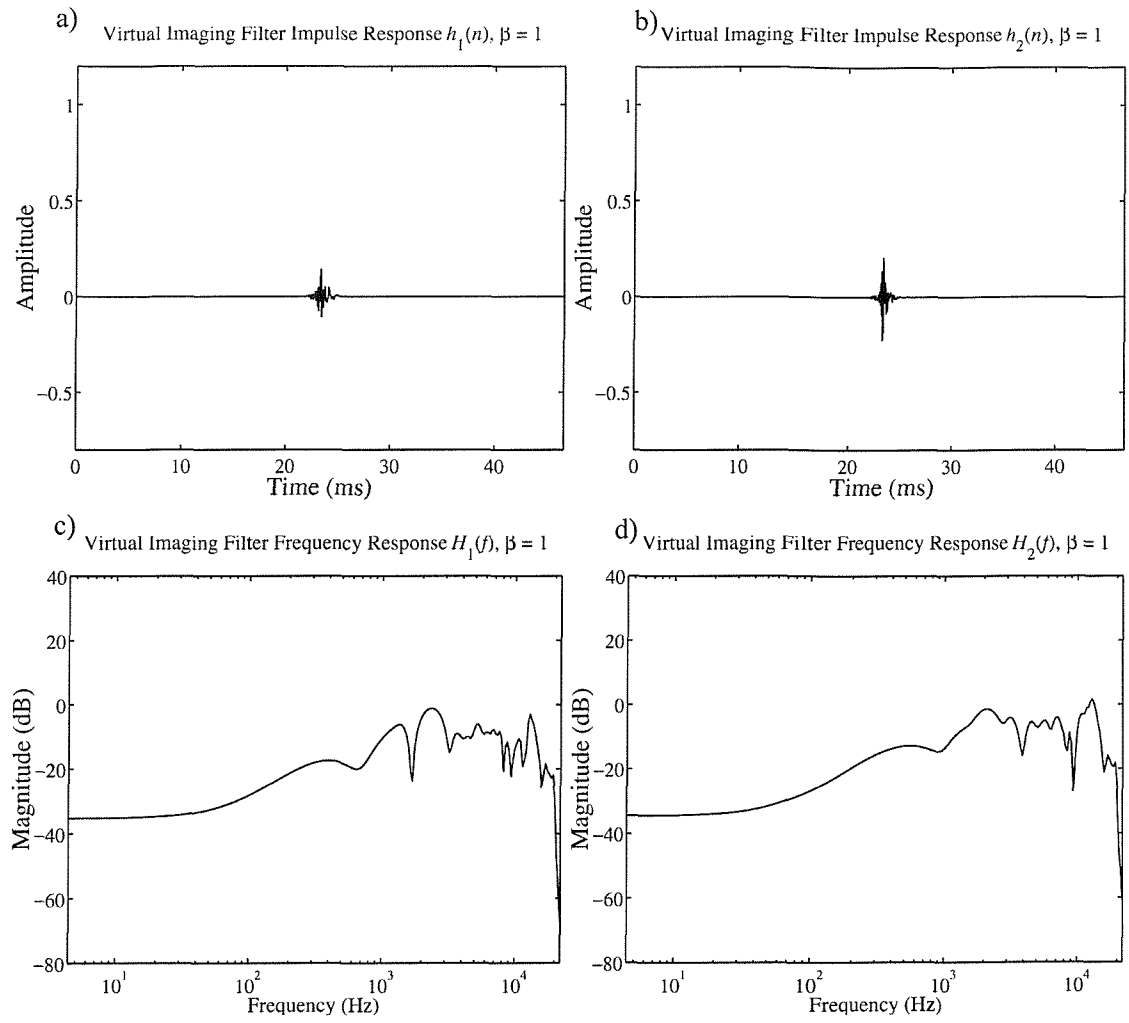


Fig. 3.32 *Virtual acoustic imaging filter responses designed from the virtual paths shown in Fig. 3.28 and the plant responses shown in Fig. 3.19 with regularisation β = 1. a) impulse response of the direct path filters h_{11} and h_{22}, b) impulse response of the cross-talk path filters h_{12} and h_{21}, c) magnitude of the frequency response of the direct path filters H_{11} and H_{22}, and d) magnitude of the frequency response of the cross-talk path filters H_{12} and H_{21}.*

4. COMPUTER SIMULATIONS OF VIRTUAL IMAGING PERFORMANCE AT ASYMMETRIC LISTENER LOCATIONS

4.1. Introduction

This section presents computer simulations considering the basic physics of the sound field in order to gain some understanding of the size and nature of the "sweet spot" at off-axis locations throughout the audible frequency range. The computer simulations utilise all three acoustic path approximations described in section 2.3. The previous chapter considered the concepts of the "ringing" frequency and condition number and their relation to geometry and robustness of the virtual acoustic imaging system. This chapter extends this analysis with several different measures of performance of simulated sound fields created by a 10° source span virtual acoustic imaging system. Section 4.2 considers the performance of cross-talk cancellation at different head positions and uses a performance criterion to derive a "sweet spot" size. Section 4.3 simulates the ILD change for non-optimal head locations and its dependence on intended listener location and virtual sound source angle. Section 4.4 uses an ITD criterion to derive a "sweet spot" size. Section 4.5 simulates sound fields that produce virtual sound images for off-axis listeners. The relative position of the real sound sources with respect to the virtual sound image position is found to have a great effect on the robustness of the virtual acoustic imaging.

4.2. Cross-talk Cancellation Effectiveness

The effect of the listener's location on the system's cross-talk cancellation performance is now considered. A cross-talk cancellation performance criterion provides a useful basis for estimating the "sweet spot" size that is independent of virtual acoustic image direction. Recall the matrix $\mathbf{R} = \mathbf{CX}$ defined by Eq. (3.47) in section 3.2. Perfect cross-talk cancellation occurs when the diagonal terms of $\mathbf{R}$, or the direct paths, are time shifted delta functions in the time domain (e.g. $\delta(t - \tau)$) and the off-diagonal terms of $\mathbf{R}$, or the cross-talk paths, are zero (0). Channel separation

refers to the ratio of the magnitude of the frequency responses of the two paths from the two sources to the same receiver (i.e. $|R_{12}/R_{11}|$ or $|R_{21}/R_{22}|$) [4]. This is essentially a measure of the effectiveness of the cross-talk cancellation process.

Figures 4.1-4.3 show the magnitudes of the frequency responses at the left ear R_{11} and R_{12} for head displacements 0.1 mm and 5 cm from optimal locations on-axis and 1 m off-axis with the free field, spherical head, and KEMAR dummy transfer function approximations respectively. Optimally, $|R_{11}|$ should be zero decibels (0 dB) over the whole frequency range and $|R_{12}|$ should be minus infinity (-∞ dB). However, the results of such a calculation are constrained by the round off error of the computer program so that $|R_{12}|$ appears to be about –200 dB and not -∞ dB for software that rounds off to sixteen (16) significant figures. The choice of a 0.1 mm head displacement approximates the results for the optimal head location and avoids the problem of the results just reflecting the round off error of the computer software.

The top two panels in each of these figures (i.e. 4.1a,b, 4.2a,b, and 4.3a,b) show the results for the symmetric on-axis optimal location. For a 0.1 mm head displacement (Figs. 4.1a, 4.2a, and 4.3a) $|R_{11}|$ is very flat over the whole frequency range and $|R_{12}|$ is below –30 dB at frequencies above 100 Hz for the free field approximation (Fig. 4.1a) and the spherical head approximation (Fig. 4.2a) and at frequencies above 200 Hz for the KEMAR dummy HRTF results. A peak in $|R_{12}|$ occurs at the "ringing" frequency at about 11 kHz. A head displacement of 5 cm greatly degrades performance. At this head displacement, Figs. 4.1b, 4.2b, and 4.3b show about 10 dB cross-talk cancellation below 5 kHz but poor performance above this frequency.

Figures 4.1c,d, 4.2c,d, and 4.3c,d show the results for the 1 m off-axis optimal location. The "ringing" frequency here is about 20 kHz due to the small path length difference Δl at this listener location. At this listener location, the frequency range of cross-talk cancellation performance has shifted to higher frequencies. The "sweet spot" shifts to higher frequencies at off-axis listener locations because of a smaller path length difference Δl.

The free field responses are the sharpest with the greatest amplitude range. The dummy's HRTF responses are the most variable due to pinna effects. The results for the dummy head show poor cross-talk cancellation performance above 20 kHz due to the error introduced by the fractional delay interpolation filter (see section 2.4.3).

Figures 4.4-4.6 show the cross-talk cancellation performance at both ears for filters designed for the on-axis head location and at 50 cm, 1 m, and 2 m off-axis. In these figures, the horizontal axes represent the head position relative to the assumed design location, the vertical axes represent frequency, and the grey scale represents the ratio of the cross-talk path to the direct path in decibels. The white areas represent good cross-talk cancellation. The bases of the calculations for Figs. 4.4, 4.5, and 4.6 are the free field approximation, spherical head model, and the KEMAR dummy measurements respectively. In these figures, the left and middle columns show the results for the left and right ear respectively. Depictions of the corresponding geometrical arrangements are provided in the right-most column.

The basis of the calculations for Fig. 4.4 is the free field approximation. The vertical black lines at head displacements of ±18 cm correspond to situations where an ear is at the intended location for the other ear. Consideration of Figs. 4.4a,b reveals that at on-axis the system tolerates head displacements of about 4-5 cm for cross-talk cancellation of at least 10 dB at low frequencies. Around 11 kHz (the "ringing" frequency), the on-axis system tolerates practically no head displacement. Above the "ringing" frequency, the system tolerates about 2-3 cm displacements from the optimum position. Figures 4.4c,d show the cross-talk cancellation performance for filters designed at 50 cm off-axis. The performance is similar to that at the on-axis position except that the "ringing" frequency is at about 13 kHz. At 1 m off-axis (Figs. 4.4e,f), the "ringing" frequency is at about 20 kHz. At 2 m off-axis (Figs. 4.4g,h), the "ringing" frequency is beyond the audio frequency range. These results imply that an expansion of the frequency range where the system is the most robust occurs as the result of a higher "ringing" frequency for off-axis systems. The higher "ringing" frequency results from a smaller path length difference Δl as discussed in previous chapter 3.

Figure 4.5 shows results using the spherical head model. This model reveals the same behaviour of the "ringing" frequency. However, when including the shadowing effects of the sphere, the cross-talk cancellation at the right ear (right column) appears to be much more robust than the left ear (left column). The left ear appears to have the most robust cross-talk cancellation on-axis. As filters are designed for listener locations farther off-axis to the listener's right, the "sweet spot's" frequency range increases but its width decreases for the left ear and increases for the right. The limiting case might be the smaller of the two widths and so the expectation might be that the "sweet spot" size decreases in width for off-axis listening.

Figure 4.6 shows results from the dummy HRTFs. This figure displays the same types of behaviour due to the "ringing" frequency and head shadowing. The results are obviously more complex in this case with the dip in the HRTFs at about 8 kHz having an effect. The results above 20 kHz reflect error introduced by the fractional delay interpolation filter due to the Gibbs phenomenon (e.g. see Fig. 2.10b and 2.11b).

Figures 4.7-4.9 show cross-talk cancellation performance under the same conditions as Figs. 4.4-4.6 but on a log frequency axis. It is evident that at listener positions further away from the inter-source axis, cross-talk cancellation performance at low frequencies decreases.

Designating at least 10 dB cross-talk cancellation performance in the frequency range 0.3-3 kHz at the position of the listener's ears as the criterion for the boundary of the "sweet spot" resulted in Fig. 4.10 for calculations based on the three (3) different models of transfer functions. This range of frequencies corresponds with the band-limited white noise stimulus that was presented to subjects during the experiments discussed in chapters 5-7. The choice of 10 dB cross-talk cancellation was made as a reasonable estimate of the amount of cross-talk cancellation needed in order to ensure delivery of a desired subjective perception. This choice is the result of some informal subjective observations of the virtual acoustic imaging system with different cross-talk cancellation filters. In Fig. 4.10, the abscissas represent the intended design filter

location relative to on-axis. The ordinates represent the head movements from the intended location. The solid lines give the "sweet spot" size for the range of intended listener locations between ±2 m off-axis as determined by an achievement of at least 10 dB cross-talk cancellation performance. Figure 4.10 also displays two vertical dashed lines provided to facilitate comparisons between the computed "sweet spot" size here and subjective results of chapter 5. Chapter 5 discusses a subjective examination of the designed filter locations within the dashed lines of Fig. 4.10 and compares the subjective results with the calculated "sweet spot" size shown within the dashed lines. The plots in the left and right columns are the results for the listener's left and right ears respectively. The results for the left ear are the mirror image of the results for the right ear due to the similar geometries.

The basis for Figs. 4.10a,b is the free field model at the left and right ears. The "sweet spot" size for the ipsi-lateral ear (ear closer to the sources) is almost constant between about ±4 cm to ±6 cm from the optimal position. The "sweet spot" size for the contra-lateral ear (ear farthest from the sources) tolerates a little more movement away from the sources at farther off-axis positions. For example, at positions greater than 1 m off-axis the contra-lateral ear has a "sweet spot" size of 10 cm away from the sources and 5 cm toward the sources.

The basis for Figs. 4.10c,d is the spherical head model at the left and right ear. The effect of head shadowing greatly increases the robustness of cross-talk cancellation at the contra-lateral ear, especially at farther off-axis positions. The ipsi-lateral ear becomes less robust at farther off-axis positions, until it reaches about ±3 cm at positions farther than 50 cm off-axis. The limiting case might be the ipsi-lateral ear, giving the expectation of a decrease in "sweet spot" size for off-axis listening.

The basis for Figs. 4.10e,f is KEMAR dummy HRTFs at the left and right ears. The "sweet spot" has the same general shape for this model as for the spherical head model. The measurements of the dummy with the inherent complexities of its shape introduce some randomness to the "sweet spot" and generally reduce its size by about 1 cm - 2 cm throughout.

Combining the results of Fig. 4.10 for both ears separately, results in Fig. 4.11. This figure shows the boundaries of the "sweet spot" that achieves at least 10 dB cross-talk cancellation at both ears based on the free field model (Fig. 4.11a), spherical head model (Fig. 4.11b), and KEMAR dummy head HRTFs (Fig. 4.11c). The system tolerates more movement toward the inter-source axis than away from the inter-source axis. This is true for all of the transfer function models. The KEMAR dummy HRTF results also suggest that listener locations further off-axis than about 50 cm are not very robust to head movements at all.

The "sweet spot's" frequency range increases at off-axis listening positions because of a smaller path length difference at off-axis positions. Head shadowing counters this benefit by decreasing the "sweet spot" size for the ipsi-lateral ear (ear closer to the sources). However, head shadowing increases the "sweet spot" size for the contra-lateral ear (ear farthest from the sources) at off-axis listener locations. In practice, the expansion of the "sweet spot's" frequency range to include higher frequencies might not matter if the sound includes low frequencies (below 1.5 kHz) due to the dominance of the low frequency interaural time difference cue for sound localization [15]. Within the dashed lines in Figs. 4.10 and 4.11, the "sweet spot" size is fairly constant and allows for head movements up to about ±3 cm - ±5 cm. Chapter 5 compares this calculated result to subjective impressions at the corresponding range of listener locations. The size of the "sweet spot" as calculated by the 10 dB cross-talk cancellation performance criteria at both ears of the KEMAR dummy HRTF is very small at listener locations further than about 50 cm off the inter-source axis. This may be a fundamental limitation of the system for off-axis listening.

4.3. Change of ILD

After computing the sound signals at a listener's eardrums one must then model the brain's extraction of information from the signals in order to attempt prediction of a listener's auditory perception. This section considers the effect of the virtual source location and intended listener location on the robustness to the ILD localisation cue. After simulation of the sound field at the receiver points, when two (2) unit impulses

are input the system at the same time, an overall ILD cue in decibels (dB) is calculated by Eq. (4.1) from the sampled sound pressure signals at the listener's eardrums $p_1(n)$ and $p_2(n)$ by

$$\mathrm{ILD} = 10 \log \frac{\sum_{n=0}^{N} p_1^2(n)}{\sum_{n=0}^{N} p_2^2(n)} \tag{4.1}$$

where n is the sample index in integer number of samples and N is the length in number of samples of the digital sound pressure signals at the listener's eardrums $p_1(n)$ and $p_2(n)$. Note that this method calculates ILD by including the entire audio frequency range. This was done in the interest of completeness even though ILD is not the most salient localisation cue below approximately 1.5 kHz and above about 3 kHz. When the listener is at the optimal listener location a properly designed virtual acoustic imaging system delivers an ILD cue to the listener that closely matches the ILD produced by a real acoustic source at the desired virtual acoustic image position. Figure 4.12 show simulation results of the ILD degradation as the listener moves away from the optimal intended listener location.

The basis of the calculations for the results of Fig. 4.12 is KEMAR dummy HRTFs. The left column shows the position of the optimal listener location with respect to the system's sound sources for the results in the right columns. The left column figures also show the angle between the listener and a point half way in between the two sources θ_{ls}. The right column shows the difference of ILD from the ILD at the optimal intended listener location as a function of virtual acoustic image angle θ_v (see Fig. 1.1) and listener displacement from the intended listener location. The vertical axis represents the virtual acoustic image angle θ_v in degrees and is varied between ±90°. Therefore, the desired perception is an image location varied from horizontal angle locations in front of the listener between exactly to the left of the subject (θ_v = +90°) to exactly to the listener's right (θ_v = -90°). The horizontal axis represents listener displacement from the optimal intended listener location in centimetres (cm) varied

between ±10 cm. The grey scale represents change of ILD from the intended ILD with white representing low values and black representing high values. The figure shows results for optimal intended listener locations 40 cm to the left of the inter-source axis, 20 cm to the left of the inter-source axis, exactly on the inter-source axis, and 90 cm to the right of the inter-source axis (i.e. x = -40 cm, -20 cm, 0, 90 cm). The ILD cue appears to be the most robust when the virtual acoustic image location coincides with the position between the two sources (i.e. $\theta_v = \theta_{ls}$).

4.4. Calculation of the "Sweet Spot" Boundary from Just Noticeable ITD

This section considers the effect of head displacements away form the intended listener location on the ITD cue at a range of listener locations. There is convincing evidence that the human brain extracts ITD by performing a running cross-correlation of the two ear signals [55-58]. Physiological studies on dogs [59], cats [60], rabbits [61], kangaroo rats [62], and barn owls [63] have located delay lines and coincidence detectors in brains that together perform a cross correlation analysis, which encodes ITD to an anatomical "place". Evidence suggests that the medial superior olive is the centre for interaural time analysis [64].

Computer models of the human auditory system generally first implement a band-pass filter bank to simulate the frequency selectivity of the cochlea and then perform an interaural cross-correlation (IACC) in each of the frequency bands [57,65,66]. The time lag at which the peaks in the lACC functions occur is then the estimate of ITD. Therefore, ITD is generally a function of critical frequency bandwidth.

For calculations of ITD in this section, a simple low-pass filtering of the two ear signals replaces the filter bank. A single cross-correlation calculation then follows. The reasoning behind this scheme is based on the evidence [6,7] that humans utilise ITD at only low frequencies and ITD is fairly constant as a function of frequency [15]. The low-pass filter has 512 coefficients and a cut-off frequency of 4 kHz. This

has an effect of averaging the results for the critical frequency bandwidths below 4 kHz.

For left and right ear low-passed sampled sound pressure signals $p_{1L}(n)$ and $p_{2L}(n)$, the digital interaural cross-correlation function $\Psi(q)$ is,

$$\Psi(q) = \sum_{n=0}^{N-q-1} p_{1L}(n) p_{2L}(n+q) \qquad 0 \le q < N \tag{4.2a}$$

$$\Psi(-q) = \sum_{n=0}^{N-q-1} p_{1L}(n+q) p_{2L}(n) \qquad 0 \le q < N \tag{4.2b}$$

where N is the number of samples in the sound pressure signals $p_{1L}(n)$ and $p_{2L}(n)$ and q is lag in number of samples. The estimate of ITD is the amount of lag where the maximum value of the IACC function occurs [45] i.e.,

$$\mathrm{ITD} = \left(q\big|_{\max(\Psi)}\right)\frac{1}{f_s}, \tag{4.3}$$

where f_s is the sampling frequency.

Klumpp and Eady have suggested that the just noticeable ITD for humans is about 10 μs [16]. The following designations of "sweet spot" boundary correspond to head displacements from the intended location that introduces this amount of ITD as found by the above method. It is reasonable to think that head displacements less than the size of the "sweet spot" as defined by this method should not introduce enough ITD to change the listener's perception of virtual image location.

Figures 4.13-4.15 show the virtual source locations and the "sweet spot" size as predicted by the three HRTF models. These figures include results for head positions through the range ±2 m off-axis. Figure 4.13 shows the "sweet spot" size when only considering the ITD shift of a single virtual source 45° to the listener's right. Two striking peaks in the "sweet spot" size occur at about 1.9 m and 1 m to the left of the inter-source axis. The peaks correspond to the listener moving to a position where the angle of the real sources θ_{ls} is equal to the virtual source angle θ_v (viz. x = -1.4 m when θ_v = +45°). The vertical dashed lines in Fig. 4.13 are provided to help aid in comparisons of this calculated "sweet spot" size with the results of a subjective evaluation of the "sweet spot" size in chapter 5. The subjective evaluation only considers the range of listener location between inside of the dashed lines.

When considering more than one virtual source these peaks disappear as shown in Figs. 4.14, and 4.15. These figures were calculated by introducing more virtual sources so that the desired listener perception is of hearing multiple virtual acoustic images at one time. The ITD was calculated separately for each individual source and the amount of head displacement from the intended listener location that introduces an ITD shift of more than 10 μs in any of the desired virtual acoustic images was designated as the boundary of the "sweet spot". Adding more virtual sources into the calculations has the effect of decreasing the "sweet spot" size.

Generally, the free field model predicts a larger "sweet spot" for farther off-axis locations while the sphere and dummy models predict a constant "sweet spot" size. The free field model predicts a size of about ±4 cm for locations close to on-axis and about ±10 cm farther off-axis. The dummy model predicts the smallest size "sweet spot" of about ±2 cm. The sphere model predicts a size of about ±4 cm.

For the range of listener locations examined subjectively in chapter 5, Fig. 4.13 shows that the free field model predicts a constant "sweet spot" size of about ±3.5 cm. The shadowing effect of the sphere causes the size to vary within the dashed lines but the dummy head results do predict a fairly constant size of about ±3 cm. It appears that

the 10 μs ITD criterion is a more stringent condition than the 10 dB cross-talk cancellation performance criterion of section 4.2 that found a "sweet spot" size of about ±5 cm at these range of listener locations. These results are compared in chapter 5.

4.5. Sound Field Simulations

This section presents some computer simulations of the sound field produced under comparable conditions to the subjective experimental conditions that will be cited in chapter 5. The goal of the system is to use two sound sources to produce, at two positions in space (the listener's ears), the sound signals that would have been produced by a sound source 45° to the front right of the listener (θ_V = 45°). Figure 4.16 shows the time and frequency responses of the Hanning and Gaussian pulses, which are the desired signals reproduced at the ear positions in the simulations.

Figures 4.17-4.22 show plan views of the simulated free field sound fields with white representing acoustic pressure values greater than one, black representing acoustic pressure values less than negative one, and shades of grey represent values in-between these two. The larger white circles are the two-monopole acoustic sources and the smaller white circles are the positions of the listener's ears in free space. The simulations do not take into account reflections off the listener. A white star shape depicts the position of the virtual acoustic image. The nine (9) frames sample the sound field with the same amount of time passing between each snapshot. The sequence of the snapshots starts at the top left frame and ends at the bottom right frame.

The desired signal is the Hanning pulse in Figs. 4.17-4.19. The desired signal is the Gaussian pulse in Figs. 4.20-4.22. The position of the listener is 35 cm to the left of the inter-source axis in Figs. 4.17 and 4.20. The position of the listener is on the inter-source axis in Figs. 4.18 and 4.21. The position of the listener is 35 cm to the right of the inter-source axis in Figs. 4.19 and 4.22. The two-monopole sources are closest to

the virtual acoustic image in Figs. 4.17 and 4.20 and furthest from the virtual acoustic image in Figs. 4.19 and 4.21.

To produce a virtual sound image to the listener's right the pulse must first reach the listener's right ear and then a slightly scaled version must reach the left ear a short time later. This occurs in all of the simulations but the sound field is more complicated when the monopole sources are further from the virtual acoustic image location (Figs. 4.19 and 4.21). The complexity of the sound field affects the tolerable head movement of the system before the desired perceptions degrade. The sound field is less complicated and therefore more robust to head movement when the sources' are close to the virtual acoustic image location.

Figures 4.23-4.25 show plan views of the simulated sound fields with a 9 cm radial perfectly rigid sphere modelling the listener's head in free space. The eighteen frames sample the sound field with the same amount of time passing between each snapshot. The sequence of the snapshots starts at the top left frame and ends at the bottom right frame.

The position of the listener is 35 cm to the left of the inter-source axis in Fig. 4.23 so that both the system's sound sources and the intended virtual image location are to the listener's front right. The position of the listener is on the inter-source axis in Fig. 4.24 so that the system's sound sources directly in front of the listener while the intended virtual image location is to the listener's front right. The position of the listener is 35 cm to the right of on-axis in Fig. 4.25 so that the system's sound sources are to the listener's front left while the intended virtual image location remains to the listener's front right. The two-monopole sources are closest to the virtual acoustic image in Fig. 4.23 and furthest from the virtual acoustic image in Fig. 4.25. The time delay between the two pulses is greater when the monopole sources are further from the virtual acoustic source location (Fig. 4.25).

Cross correlating the signals at the two ears during these simulations as a function of lateral head position relative to the intended listener location yields Fig. 4.26. The left column represents the interaural cross correlation function (IACC) on the grey scale, the horizontal axes represent the head displacement from the intended optimal location, and the vertical axes represent the time lag between the two ears. The right column shows the peak of the IACC function, which is an estimate of the interaural time difference (ITD) sound localisation cue. The top row of the Fig. 4.26 corresponds to the situation of Fig. 4.23 when the intended optimal listener location is 35 cm to the left of the inter-source axis. The middle row of the Fig. 4.26 corresponds to the situation of Fig. 4.24 when the intended optimal listener location is on the inter-source axis. The bottom row of the Fig. 4.26 corresponds to the situation of Fig. 4.25 when the intended optimal listener location is 35 cm to the right of the inter-source axis.

The ITD function has a slight slope that causes the virtual image to gradually shift further to the listener's right as the listener moves in the negative direction (the listener's left) away from the intended design location and gradually shift in front of the listener when the listener moves in the positive direction (the listener's right). There is an abrupt shift in ITD when the listener moves far enough to the right. This shift in ITD causes the virtual image location to abruptly shift to a location close to directly in front of the listener (i.e. ITD $\approx$ 0). This abrupt shift occurs when the listener is at about 7 cm to the right of the intended location 35 cm to the left of on-axis (Figs. 4.26a,b) and at about 5 cm to the right of the intended location 35 cm to the right of on-axis (Figs. 4.26e,f). Therefore, for a virtual image to the right of the listener, intended listener locations to the left of the inter-source axis are more robust to lateral movements to the listener's right than intended listener locations to the right of the inter-source axis. The more robust case corresponds to the situation where the virtual image is close to the sound sources and the sound field in Fig. 4.25.

Figure 4.27 shows the simulated sound field with the rigid sphere displaced 8 cm to the right of the intended location 35 cm to the right of the inter-source axis. This amount of displacement corresponds to a location where the system fails to deliver

ITD for a virtual sound image at 45° but delivers an ITD closer to zero (0) and as seen in Fig. 4.26f. The null of the first pulse no longer passes through the position of the listener's left ear with 8 cm head displacement so that both the left and right ear experience a positive pulse of sound during the first wave front.

Figure 4.28 shows the lACC function and the ITD for the spherical head approximation sound fields at intended listener locations x = –35 cm, 0, and +35 cm from the inter-source axis with θ_V = 45° as a function of azimuthal head rotation. The slope of the function corresponds to the rate of change of ITD that one would experience with a real sound source at 45° to the listener's front right so that when the head is rotated +45° the ITD function shows zero (0) ITD. However, an unnatural jump in the ITD function is seen at large head rotations. This jump for is at about 75° for x = -35 cm (Fig. 4.28b), 70° for x = 0 (Fig. 4.28d), and 55° for x = +35 cm (Fig. 4.28f). Then the listener location where the virtual source location is closest to the location of real sound sources appears to be the most robust to head rotations.

The position of the sound sources relative to the virtual acoustic image location affects the complexity of sound field and so affects the required filter update movement increment that is capable of achieving a stable virtual acoustic image. These simulations show the need for smaller filter update movement increments when the sound sources are positioned further away from the virtual sound image location.

4.6. Conclusions

The simulations presented show that the degradation of the interaural cues with listener movements away from the optimal location depends on relative location of the virtual sound image with respect to the location of the real sound sources. For production of a single virtual acoustic image, the simulations predict a "sweet spot" size of about ±3 cm. However, if the sources are close to the virtual acoustic image position the "sweet spot" size is more likely a little greater. Both the ITD and cross-talk cancellation simulations predict close to the same size "sweet spot" making it difficult to define the limiting factor of the "sweet spot" size. Farther off-axis listening

has an advantage of a higher "ringing" frequency due to smaller path length differences. Head-shadowing effects decreasing the "sweet spot" size at low frequencies offset this advantage. In addition, when including the complexities of the KEMAR dummy, cross-talk cancellation is not very robust to head motions at off-axis listener locations greater than about 50 cm.

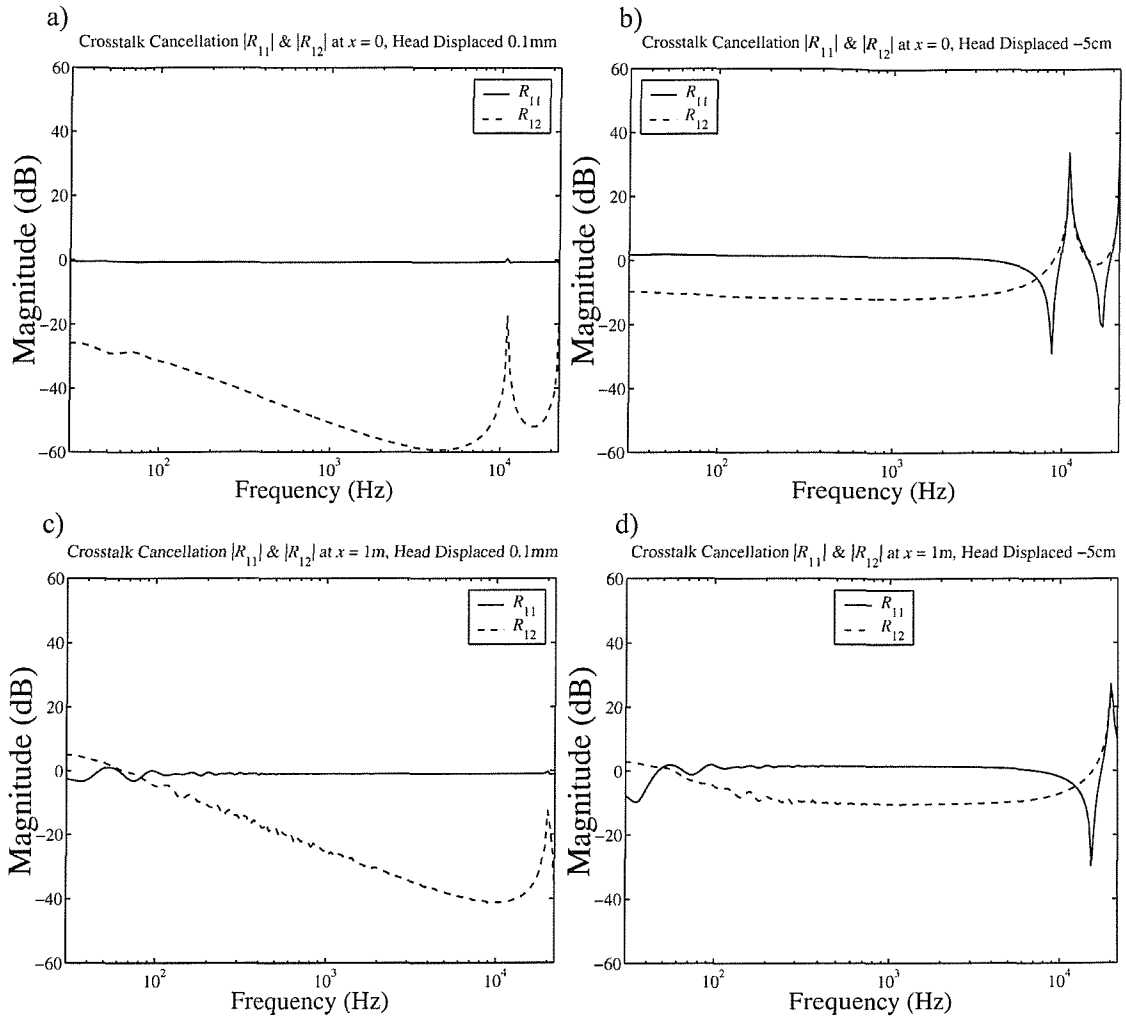


Fig. 4.1 *Elements $|R_{11}|$ and $|R_{12}|$ of the control performance matrix* **R** *for filters designed with the free field approximation. The left column shows results when the head is displaced 0.1 mm from optimum. The right column shows results when the head is displaced 5 cm from optimum. The intended design locations of the listener's head are on the inter-source axis for the top row (i.e. a and b) and 1 m to the right of the inter-source axis for the bottom row (i.e. c and d).*

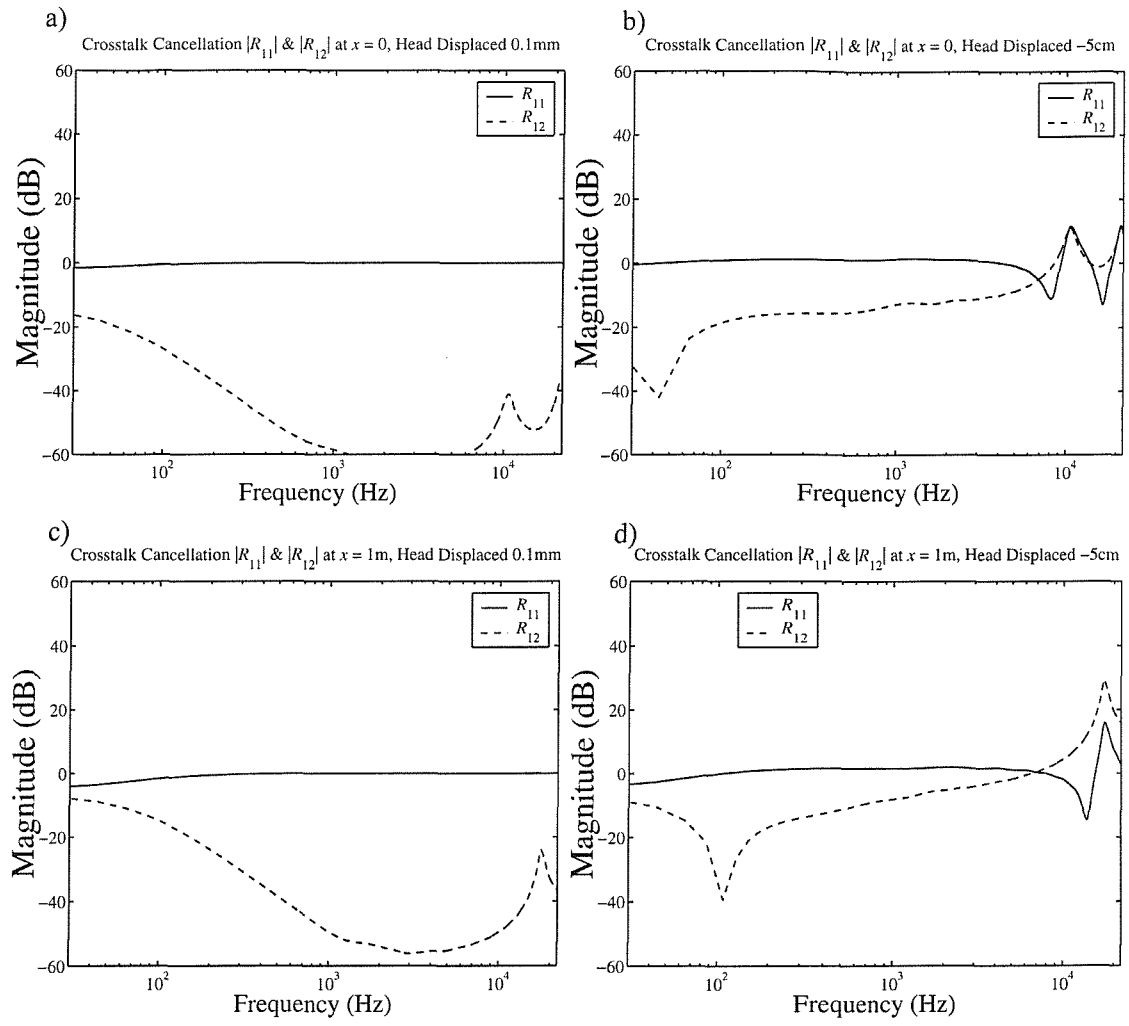


Fig. 4.2 *Elements $|R_{11}|$ and $|R_{12}|$ of the control performance matrix* **R** *for filters designed with the spherical head model. The left column shows results when the head is displaced 0.1 mm from optimum. The right column shows results when the head is displaced 5 cm from optimum. The intended design locations of the listener's head are on the inter-source axis for the top row (i.e. a and b) and 1 m to the right of the inter-source axis for the bottom row (i.e. c and d).*

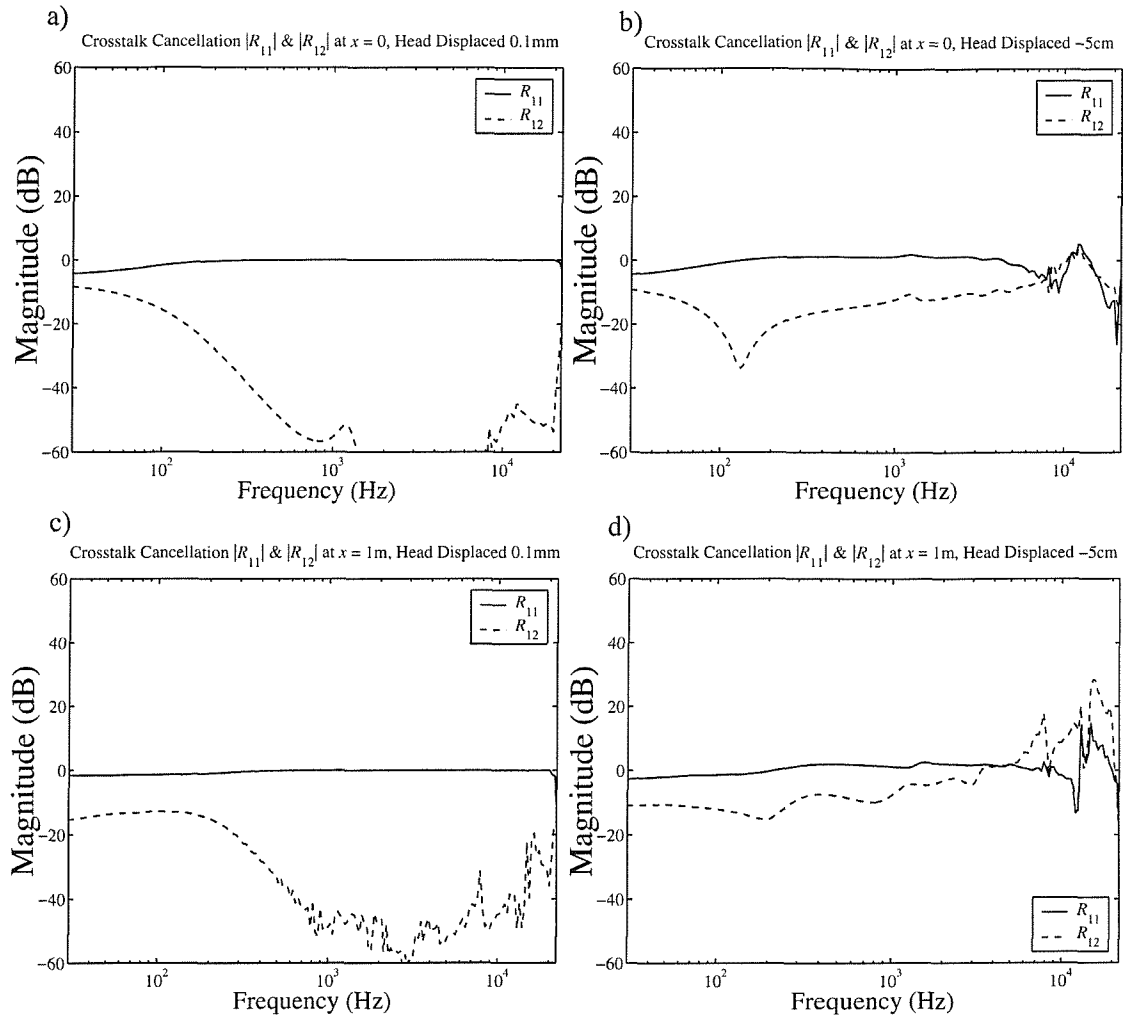


Fig. 4.3 *Elements $|R_{11}|$ and $|R_{12}|$ of the control performance matrix* **R** *for filters designed with dummy HRTF measurements. The left column shows results when the head is displaced 0.1 mm from optimum. The right column shows results when the head is displaced 5 cm from optimum. The intended design locations of the listener's head are on the inter-source axis for the top row (i.e. a and b) and 1 m to the right of the inter-source axis for the bottom row (i.e. c and d).*

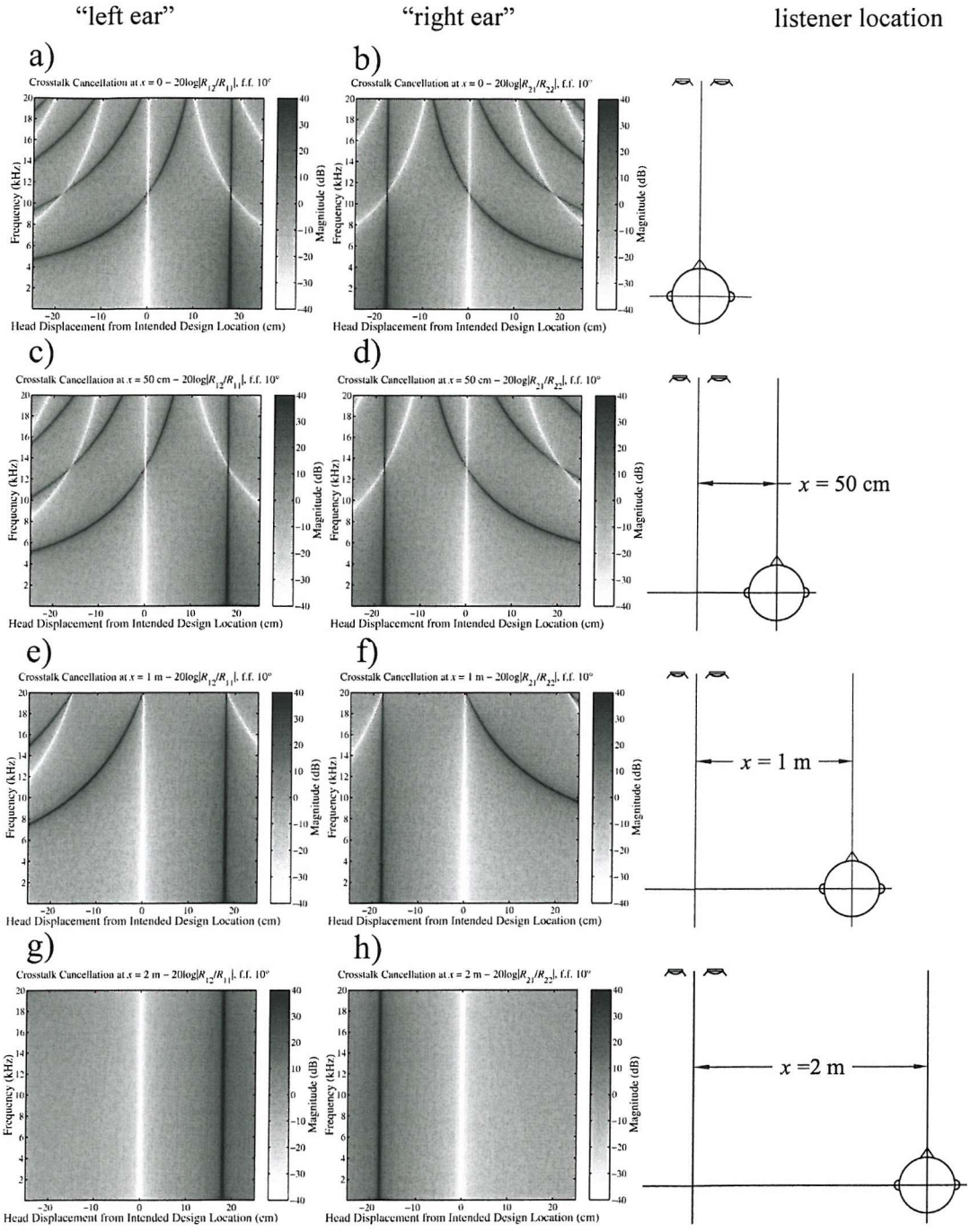


Fig. 4.4 *Cross-talk cancellation effectiveness for the left ear (left column) and right ear (right column) at intended design listener locations a-b) on the inter-source axis, c-d) 50 cm to the right of the inter-source axis, e-f) 1 m to the right of the inter-source axis, and g-h) 2 m to the right of the inter-source axis. The horizontal axes represent the head position relative to the intended design listener location, the vertical axes represent frequency and the grey scale represents the ratio of the cross-talk path to the direct path in decibels. Good cross-talk cancellation performance is achieved in the light areas. The calculations are based on the free field approximation.*

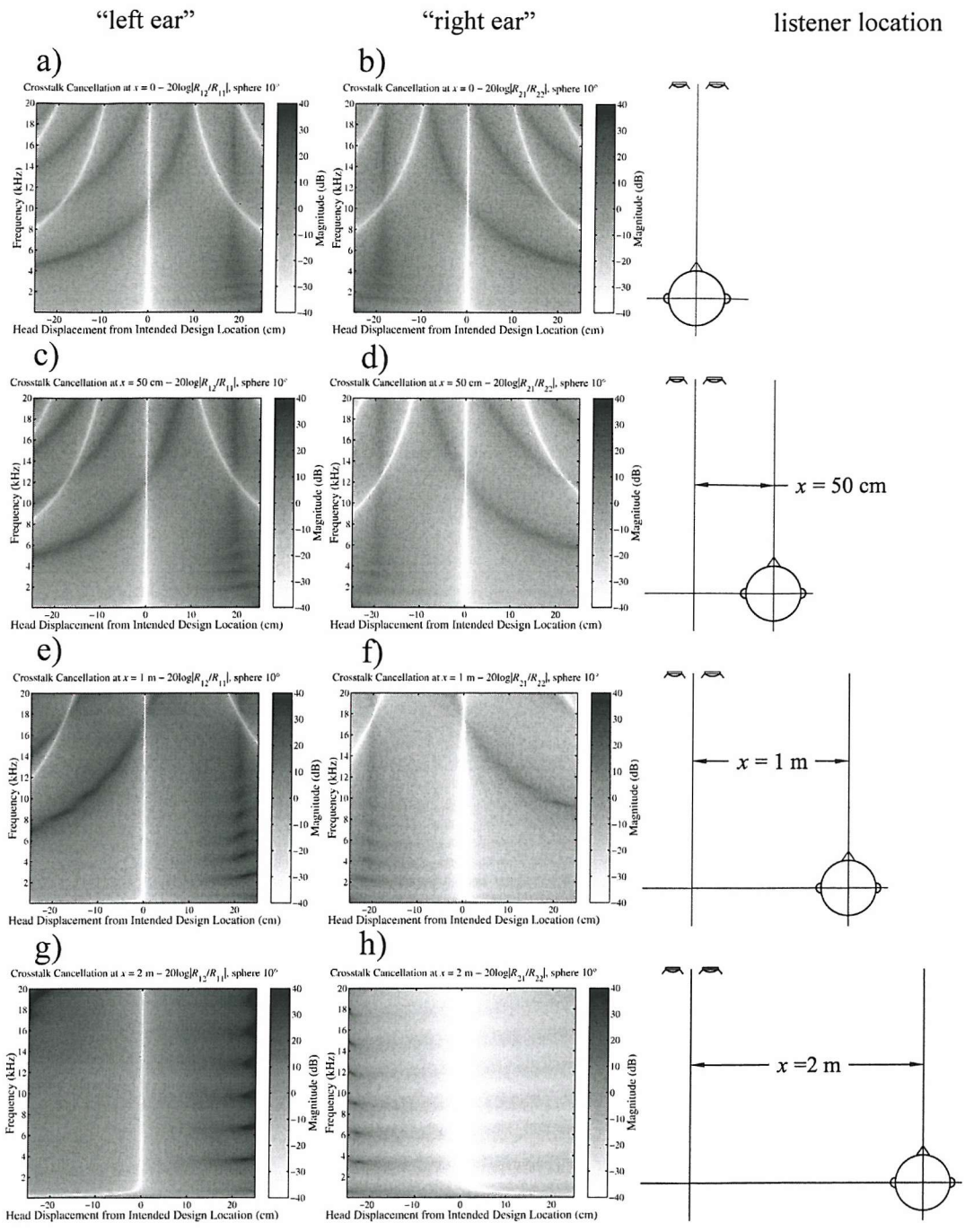


Fig. 4.5 *Cross-talk cancellation effectiveness for the left ear (left column) and right ear (right column) at intended design listener locations a-b) on the inter-source axis, c-d) 50 cm to the right of the inter-source axis, e-f) 1 m to the right of the inter-source axis, and g-h) 2 m to the right of the inter-source axis. The horizontal axes represent the head position relative to the intended design listener location, the vertical axes represent frequency and the grey scale represents the ratio of the cross-talk path to the direct path in decibels. Good cross-talk cancellation performance is achieved in the light areas. The calculations are based on the spherical head model.*

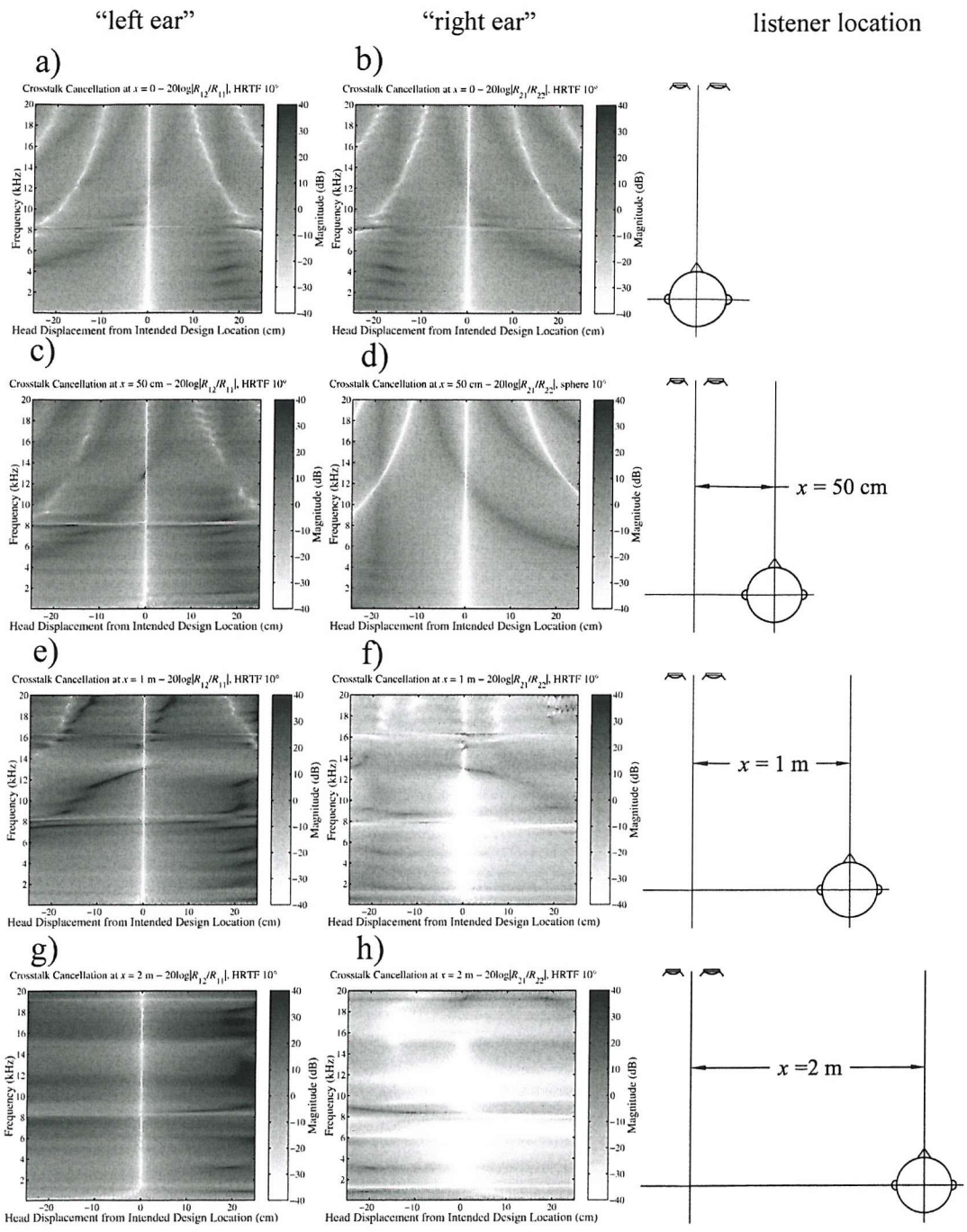


Fig. 4.6 *Cross-talk cancellation effectiveness for the left ear (left column) and right ear (right column) at intended design listener locations a-b) on the inter-source axis, c-d) 50 cm to the right of the inter-source axis, e-f) 1 m to the right of the inter-source axis, and g-h) 2 m to the right of the inter-source axis. The horizontal axes represent the head position relative to the intended design listener location, the vertical axes represent frequency and the grey scale represents the ratio of the cross-talk path to the direct path in decibels. Good cross-talk cancellation performance is achieved in the light areas. The calculations are based on dummy HRTFs.*

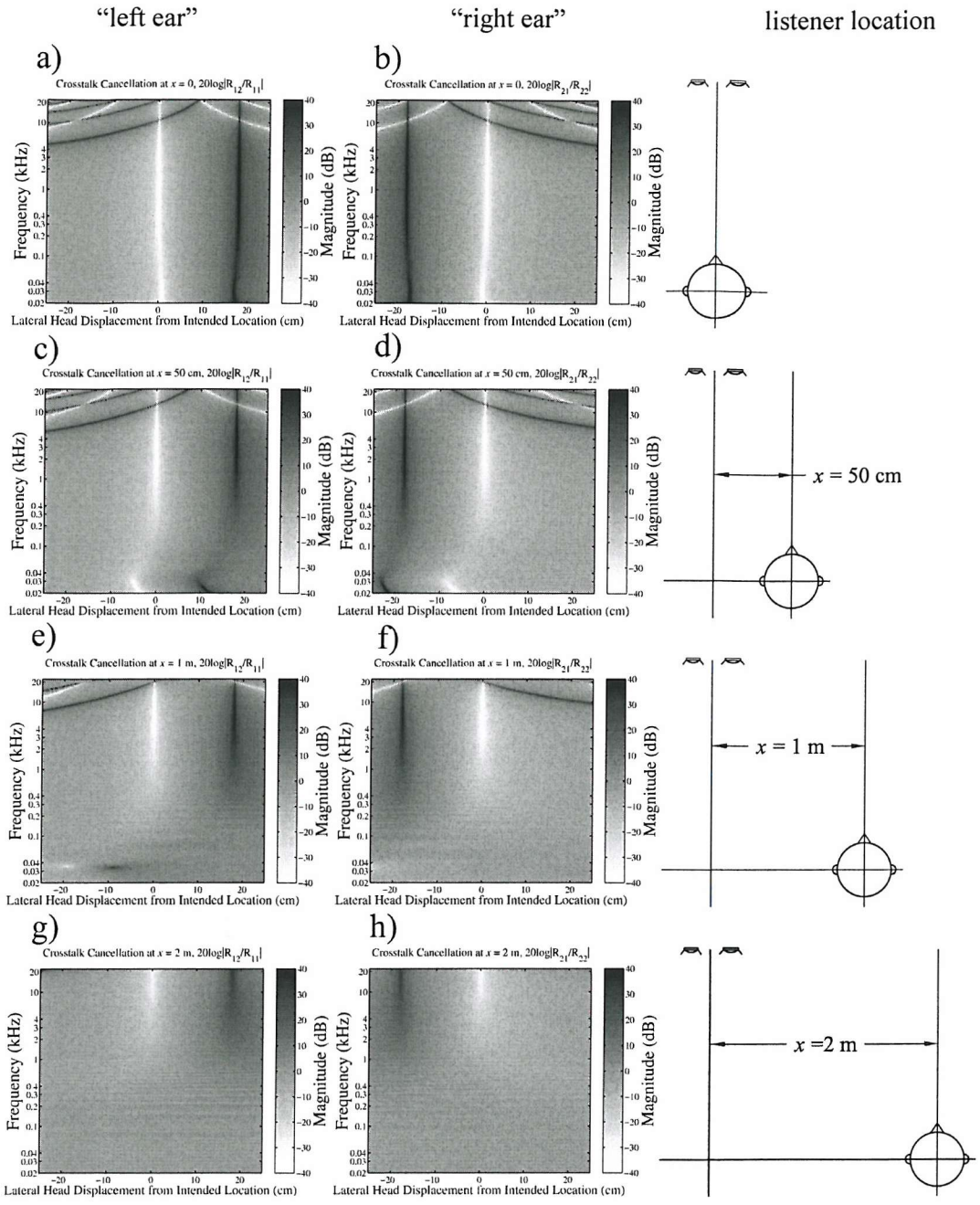


Fig. 4.7 *Cross-talk cancellation effectiveness for the left ear (left column) and right ear (right column) at intended design listener locations a-b) on the inter-source axis, c-d) 50 cm to the right of the inter-source axis, e-f) 1 m to the right of the inter-source axis, and g-h) 2 m to the right of the inter-source axis. The horizontal axes represent the head position relative to the intended design listener location, the vertical axes represent frequency and the grey scale represents the ratio of the cross-talk path to the direct path in decibels. Good cross-talk cancellation performance is achieved in the light areas. The calculations are based on the free field approximation.*

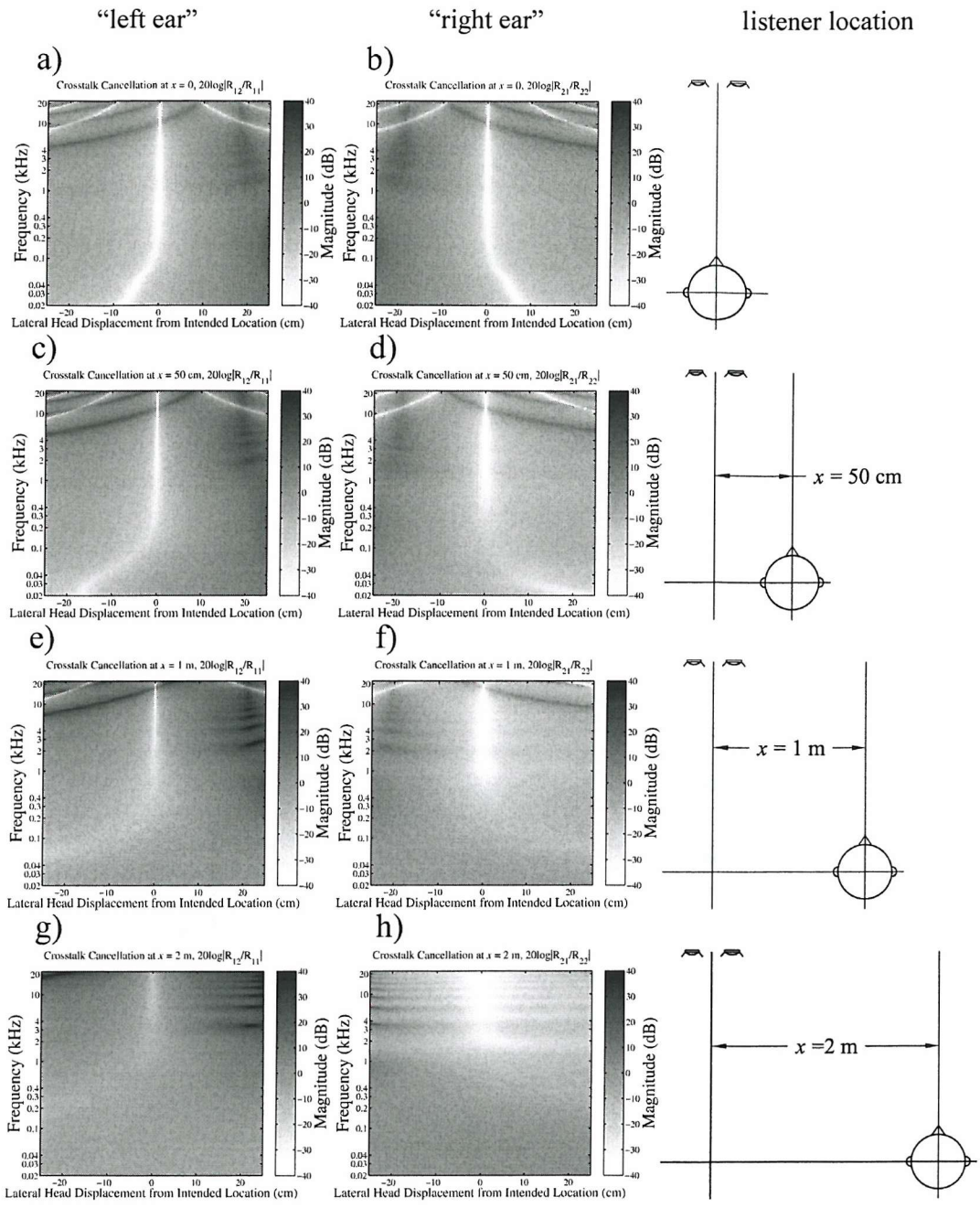


Fig. 4.8 *Cross-talk cancellation effectiveness for the left ear (left column) and right ear (right column) at intended design listener locations a-b) on the inter-source axis, c-d) 50 cm to the right of the inter-source axis, e-f) 1 m to the right of the inter-source axis, and g-h) 2 m to the right of the inter-source axis. The horizontal axes represent the head position relative to the intended design listener location, the vertical axes represent frequency and the grey scale represents the ratio of the cross-talk path to the direct path in decibels. Good cross-talk cancellation performance is achieved in the light areas. The calculations are based on the spherical head model.*

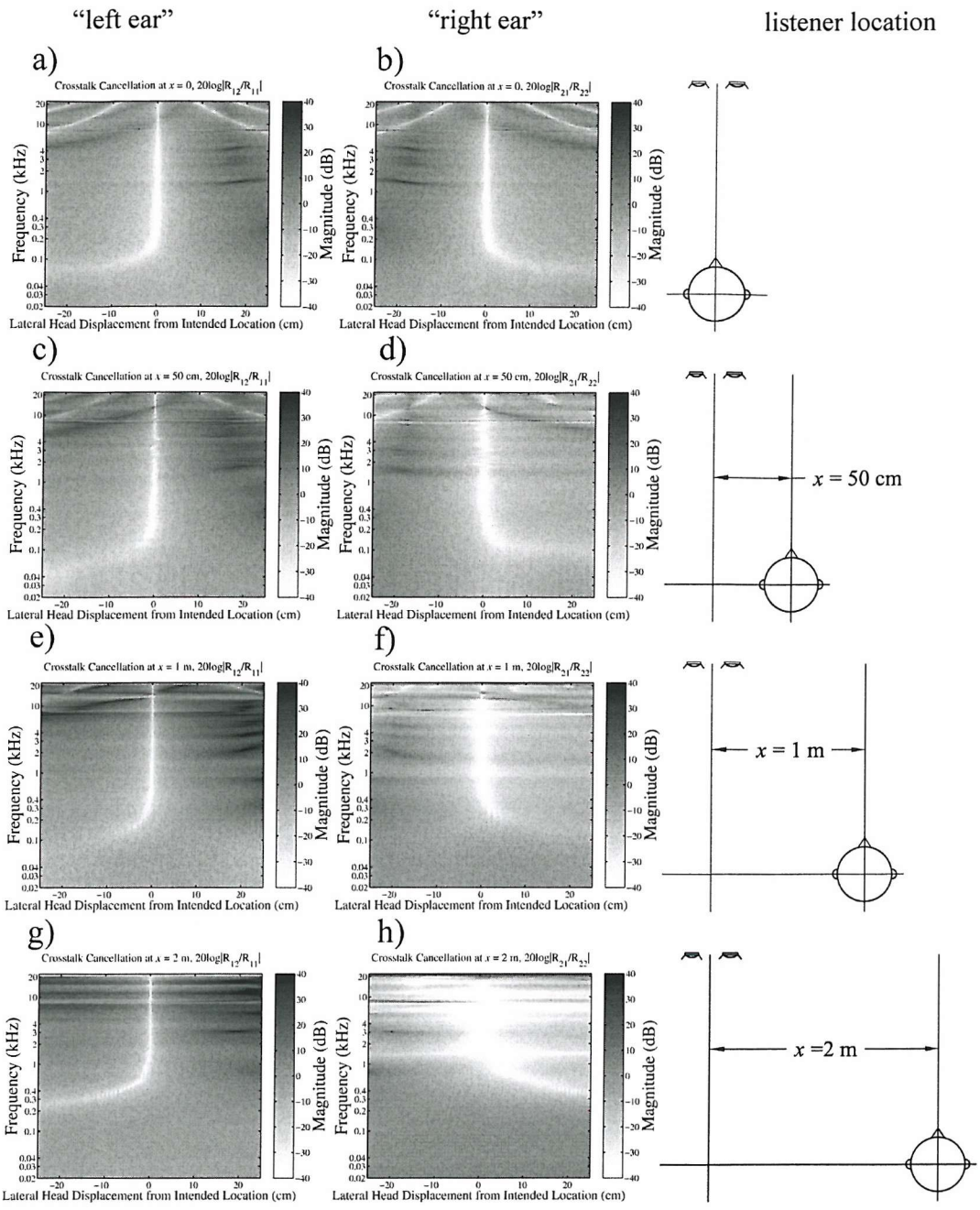


Fig. 4.9 *Cross-talk cancellation effectiveness for the left ear (left column) and right ear (right column) at intended design listener locations a-b) on the inter-source axis, c-d) 50 cm to the right of the inter-source axis, e-f) 1 m to the right of the inter-source axis, and g-h) 2 m to the right of the inter-source axis. The horizontal axes represent the head position relative to the intended design listener location, the vertical axes represent frequency and the grey scale represents the ratio of the cross-talk path to the direct path in decibels. Good cross-talk cancellation performance is achieved in the light areas. The calculations are based on dummy HRTFs.*

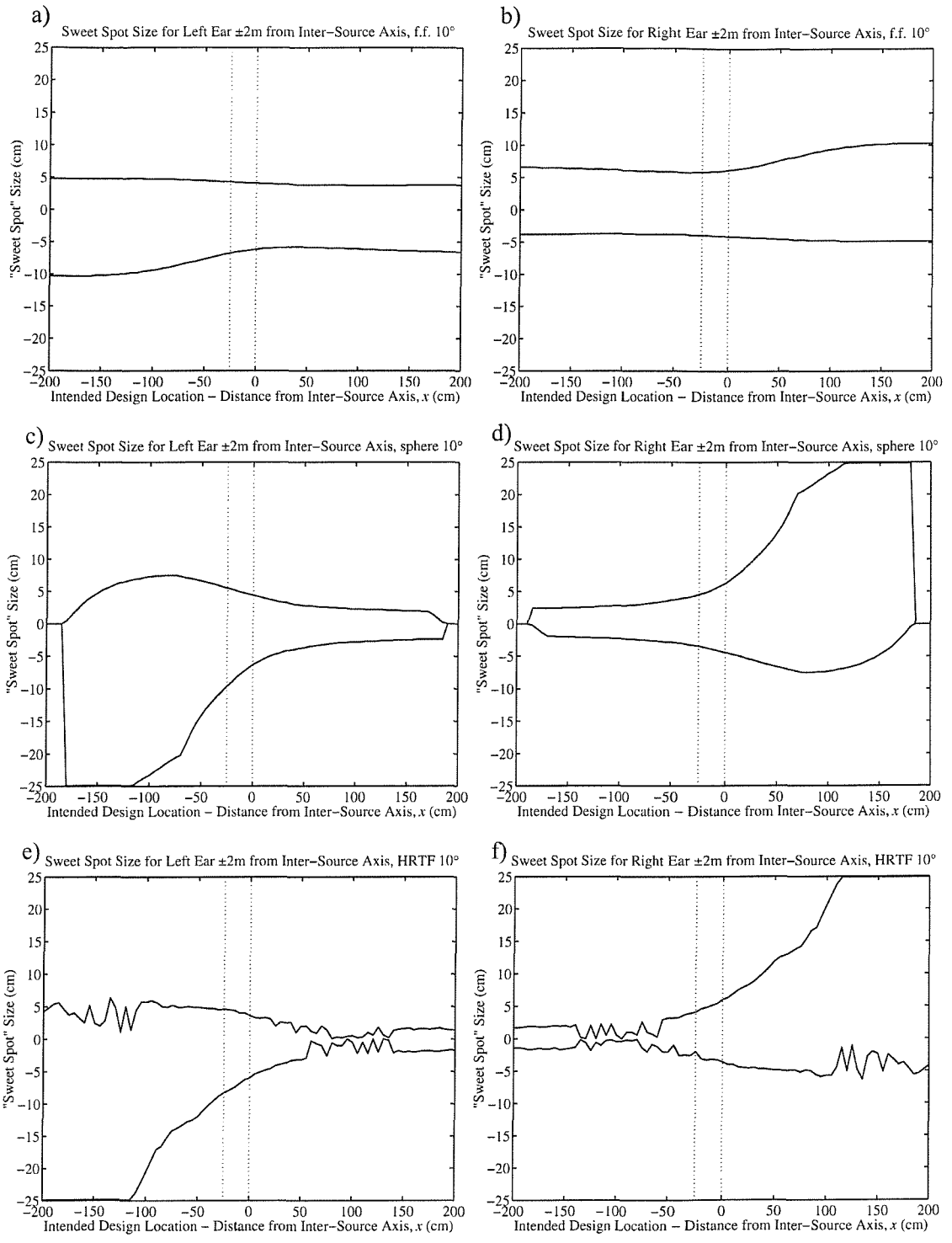


Fig. 4.10 *Boundaries of the sweet spot for a range of head locations between ±2 m off the inter-source axis as calculated from a 10 dB cross-talk cancellation performance criterion in the frequency range 0.3-3 kHz when using a) the free field approximation at the left ear, b) the free field approximation at the right ear, c) spherical head model at the left ear, d) spherical head model at the right ear, e) dummy HRTFs at the left ear, and f) dummy HRTFs at the right ear. Dashed lines show the range of intended design filter head locations examined subjectively in chapter 5.*

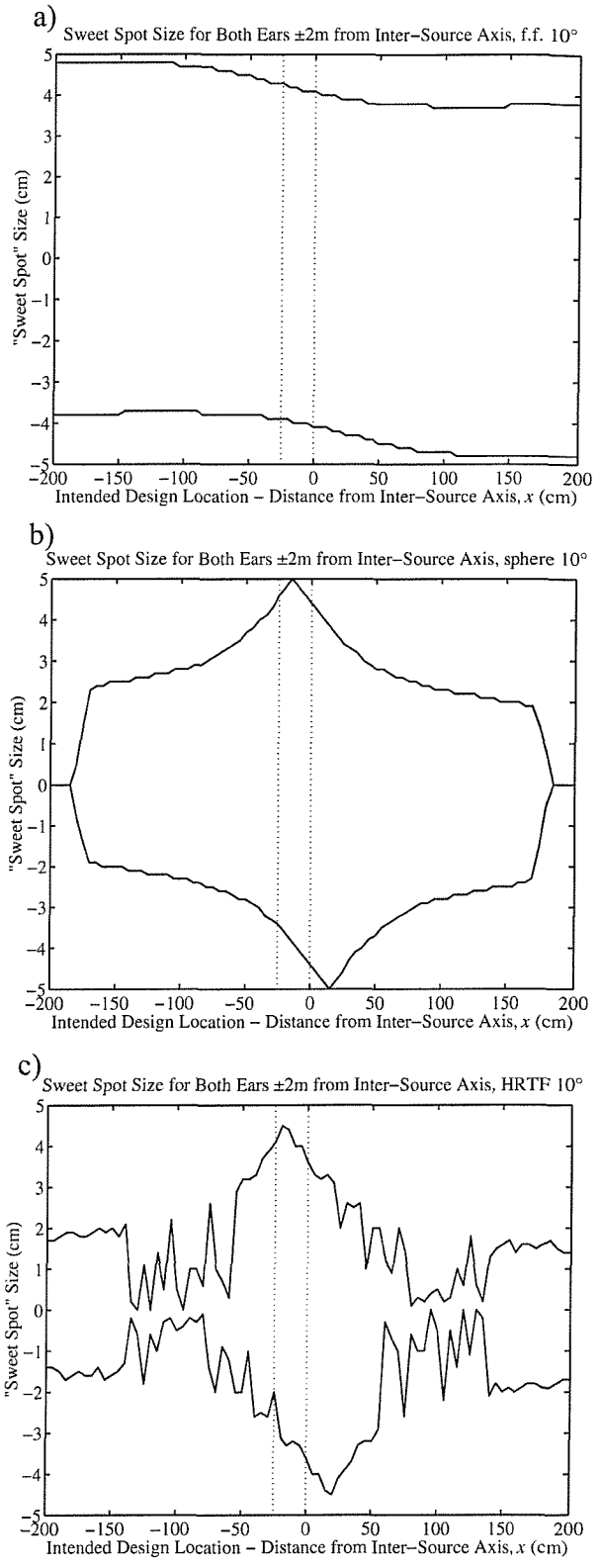


Fig. 4.11 *Boundaries of the sweet spot for a range of head locations between ±2 m off the inter-source axis as calculated from a 10 dB cross-talk cancellation performance criterion in the frequency range 0.3-3 kHz when using a) the free field approximation at both ears, b) spherical head model at both ears, and c) dummy HRTFs at both ears. Dashed lines show the range of intended design filter head locations examined subjectively in chapter 5.*

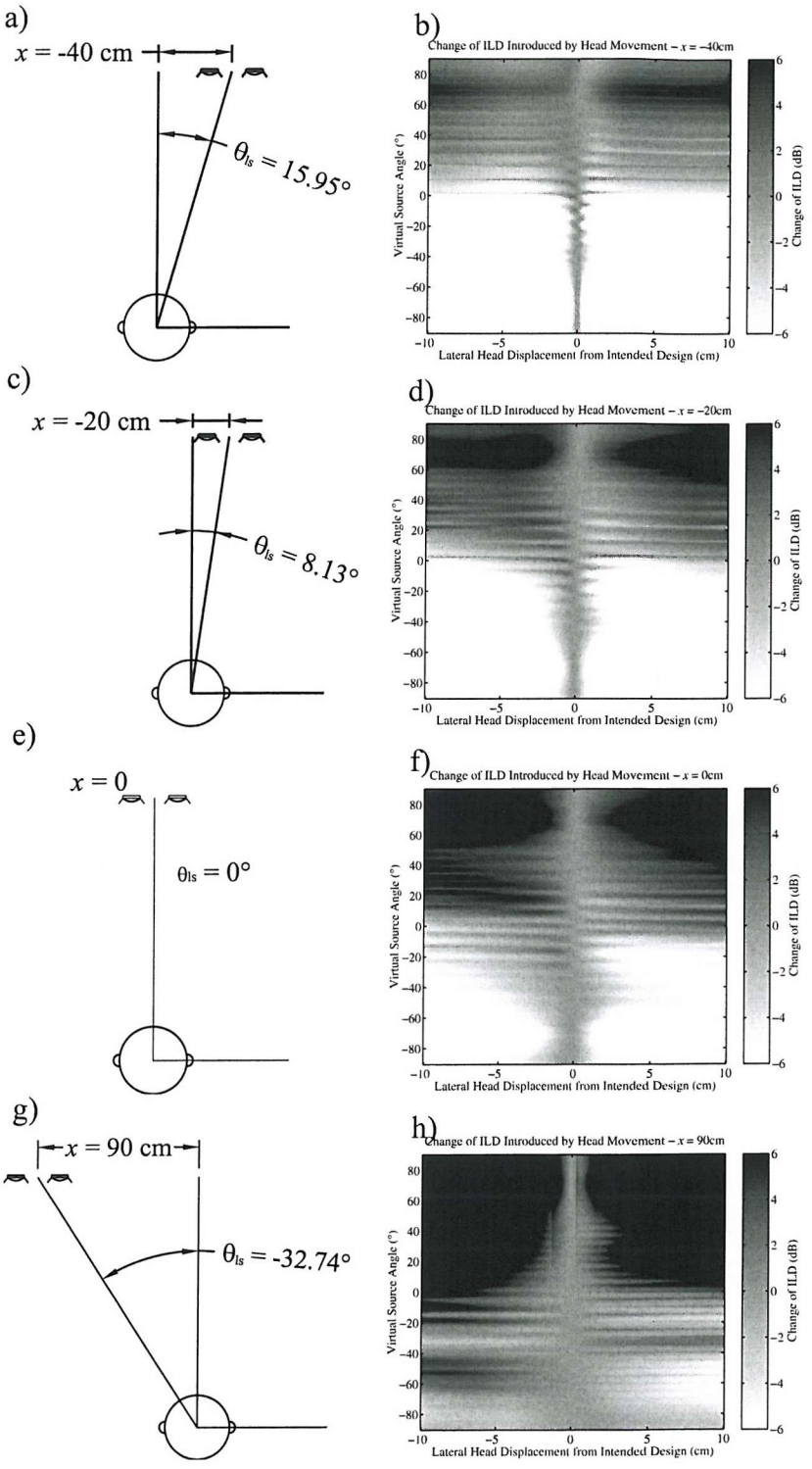


Fig. 4.12 *Degradation of ILD as the listener moves away from the intended listener position is shown in the right column for the intended positions shown in the left column. The vertical axis is the virtual source angle, the horizontal axis is listener displacement from the intended design location and the gray scale represents difference in ILD from the ILD at the intended design location. Calculations are based on dummy HRTFs.*

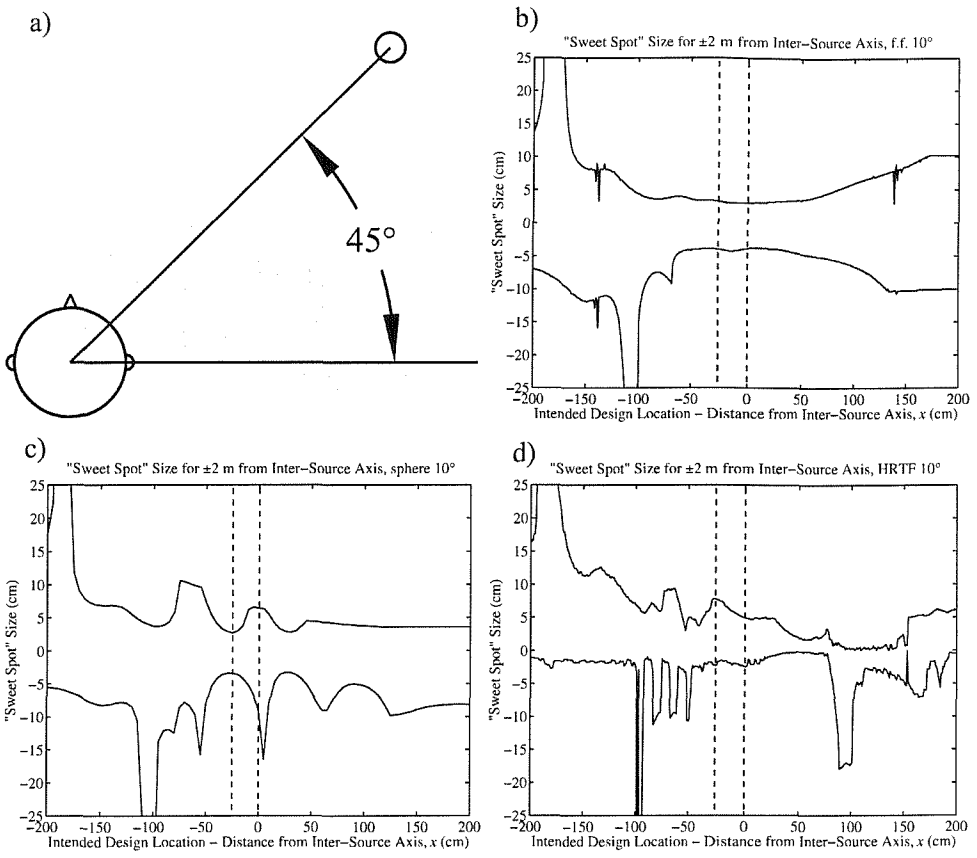


Fig. 4.13 *Boundaries of the sweet spot for a range of head locations between ±2 m off the inter-source axis as calculated from a 10 μs shift in ITD of the single virtual source represented by the small circle in a) when using b) the free field approximation, c) spherical head model, and d) dummy HRTFs. Dashed lines show the range of intended listener locations examined subjectively in chapter 5.*

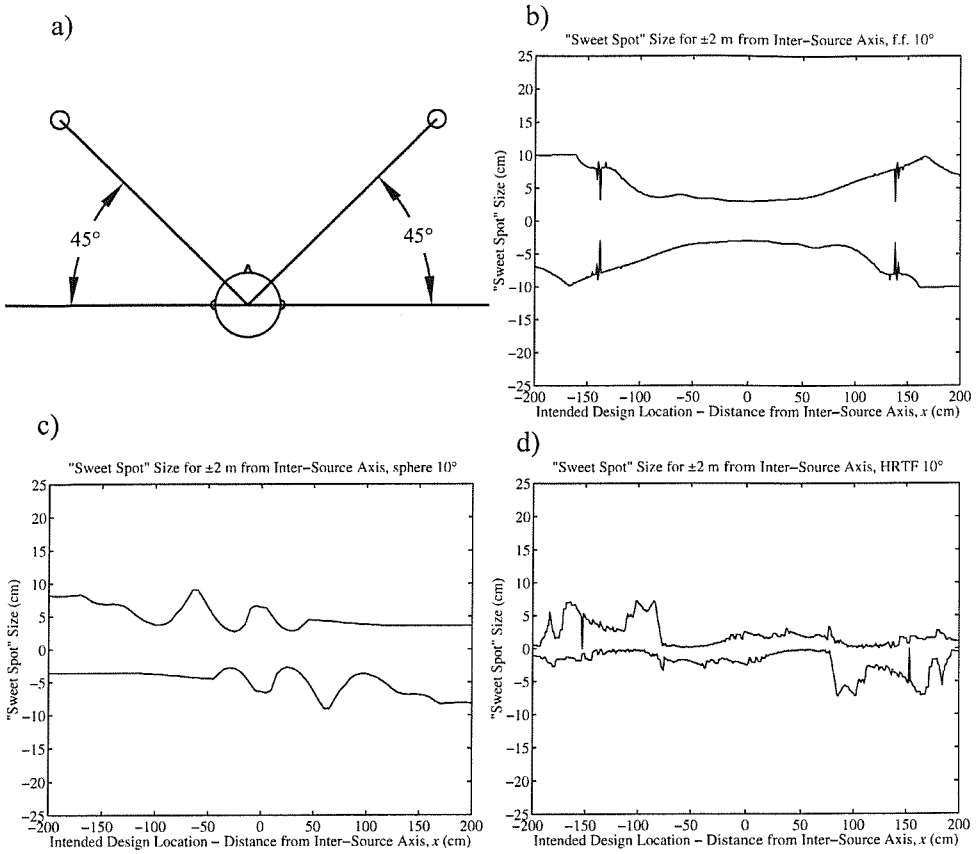


Fig. 4.14 *Boundaries of the sweet spot for a range of head locations between ±2 m off the inter-source axis as calculated from a 10 μs shift in ITD of either of the two virtual sources represented by the small circles in a) when using b) the free field approximation, c) spherical head model, and d) dummy HRTFs.*

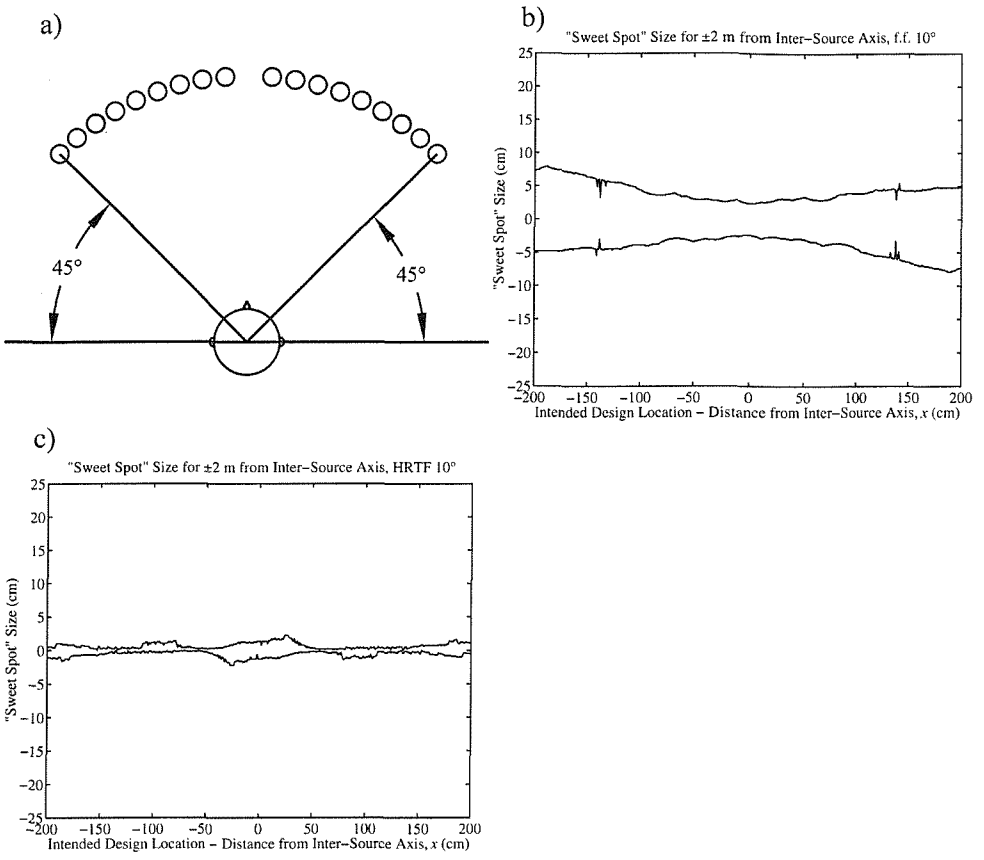


Fig. 4.15 *Boundaries of the sweet spot for a range of head locations between ±2 m off the inter-source axis as calculated from a 10 μs shift in ITD of any of the 18 virtual sources represented by the small circles in a) when using b) the free field approximation and c) dummy HRTFs.*

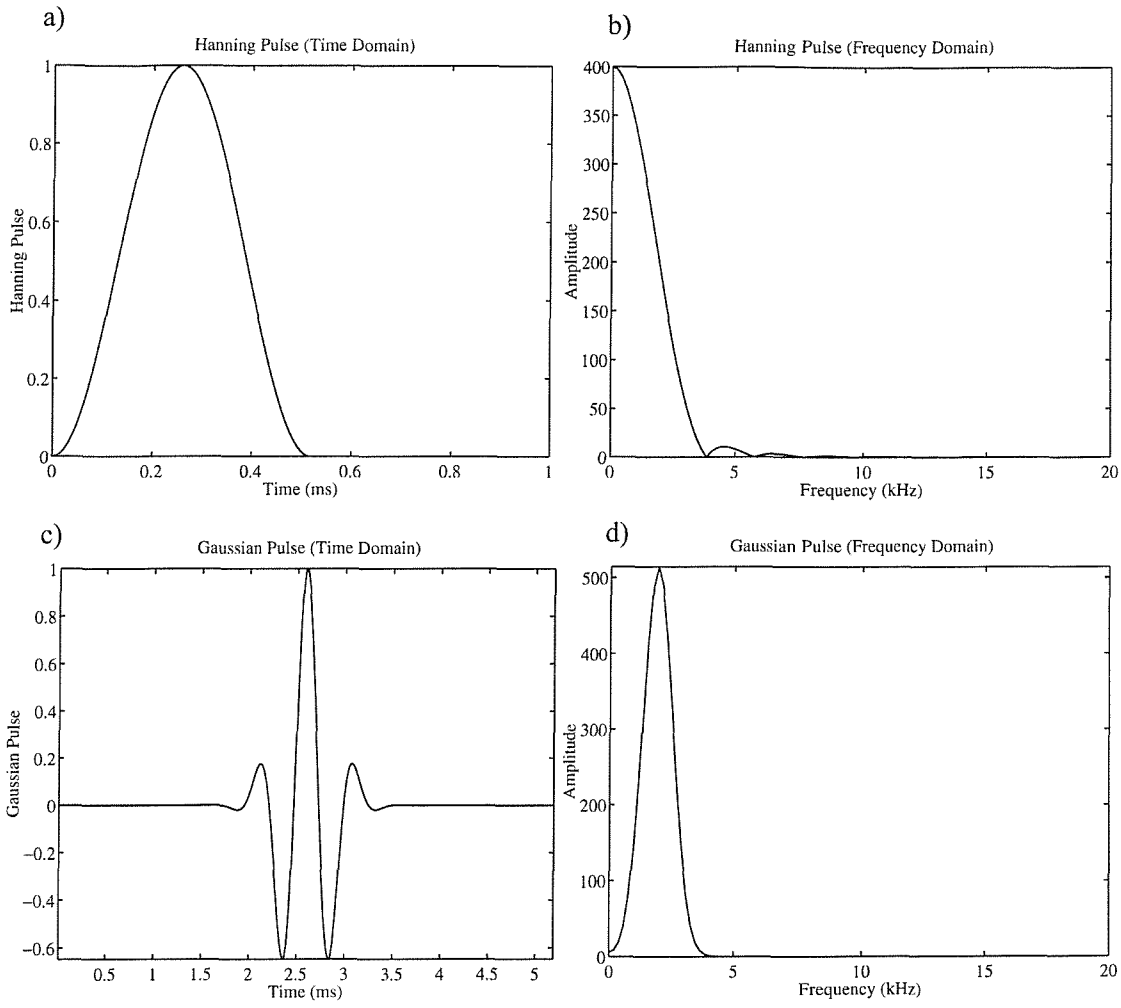


Fig. 4.16 *Time and frequency responses of the Hanning and Gaussian pulses used as desired signals in the sound field simulations of Figs. 4.17-4.28.*

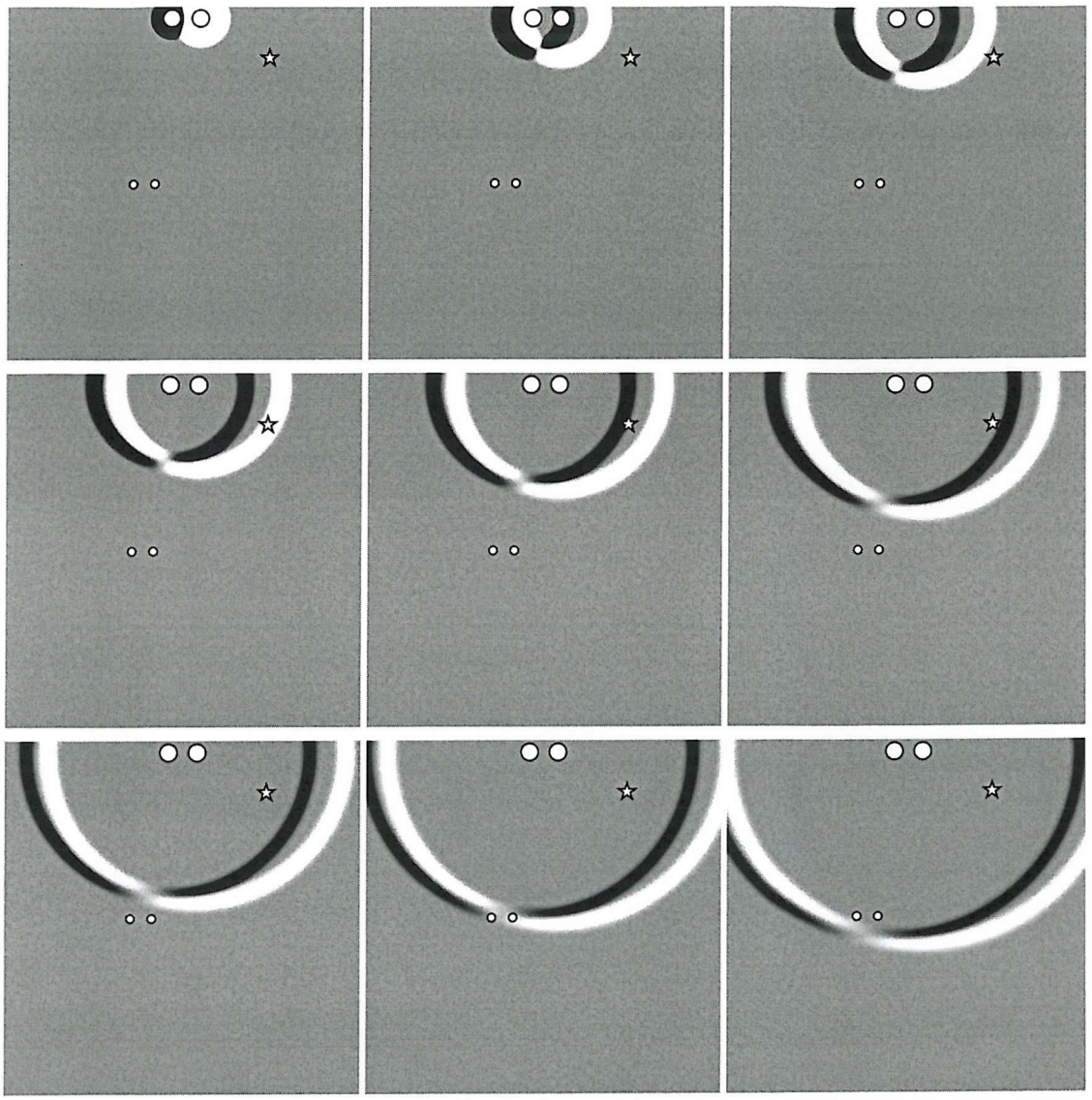


Fig. 4.17 *Sound field produced with the listener 35 cm to the left of the inter-source axis when creating a virtual image at the position of the star shape. The desired signal is the Hanning pulse.*

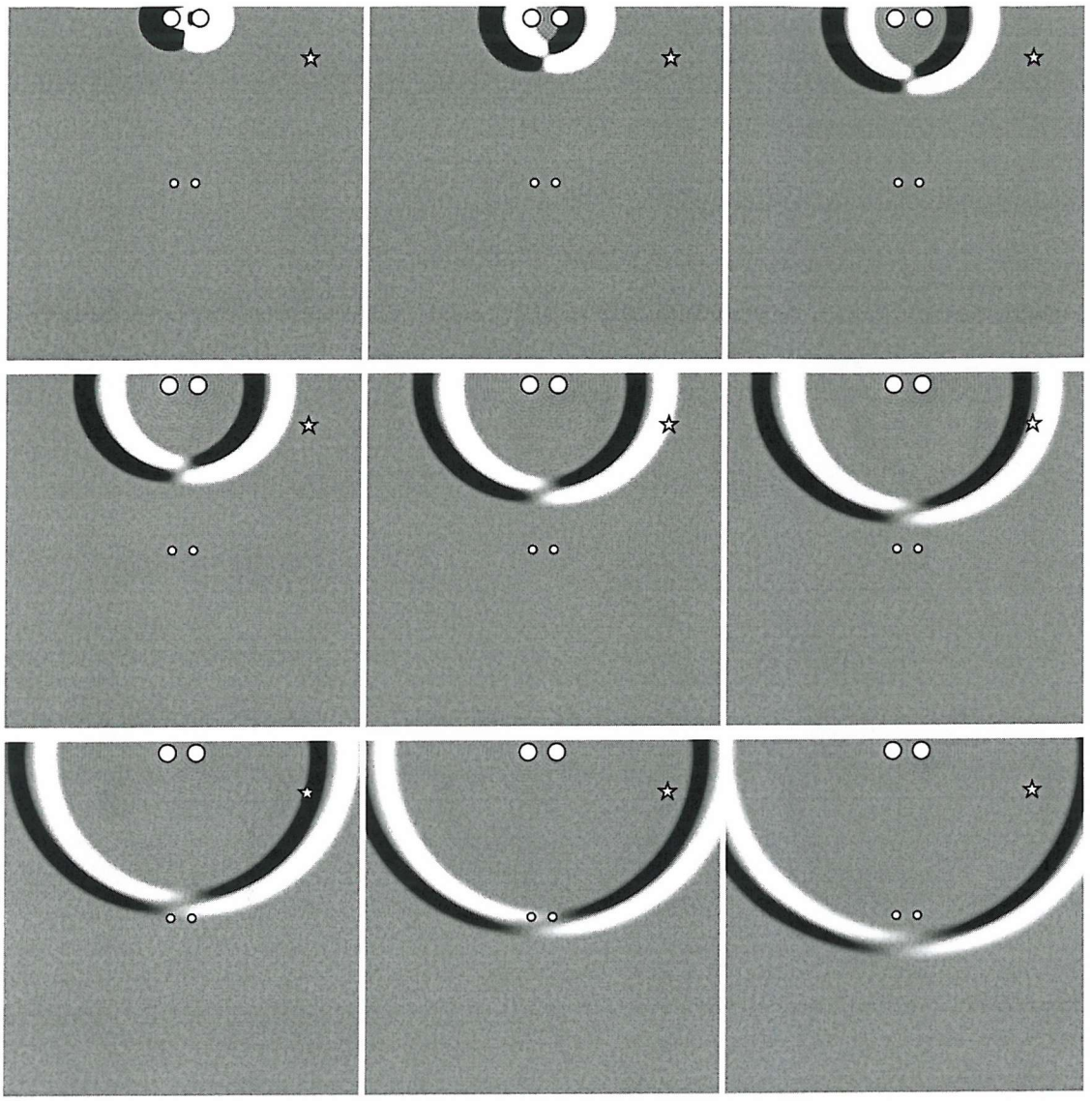


Fig. 4.18 *Sound field produced with the listener on the inter-source axis when creating a virtual image at the position of the star shape. The desired signal is the Hanning pulse.*

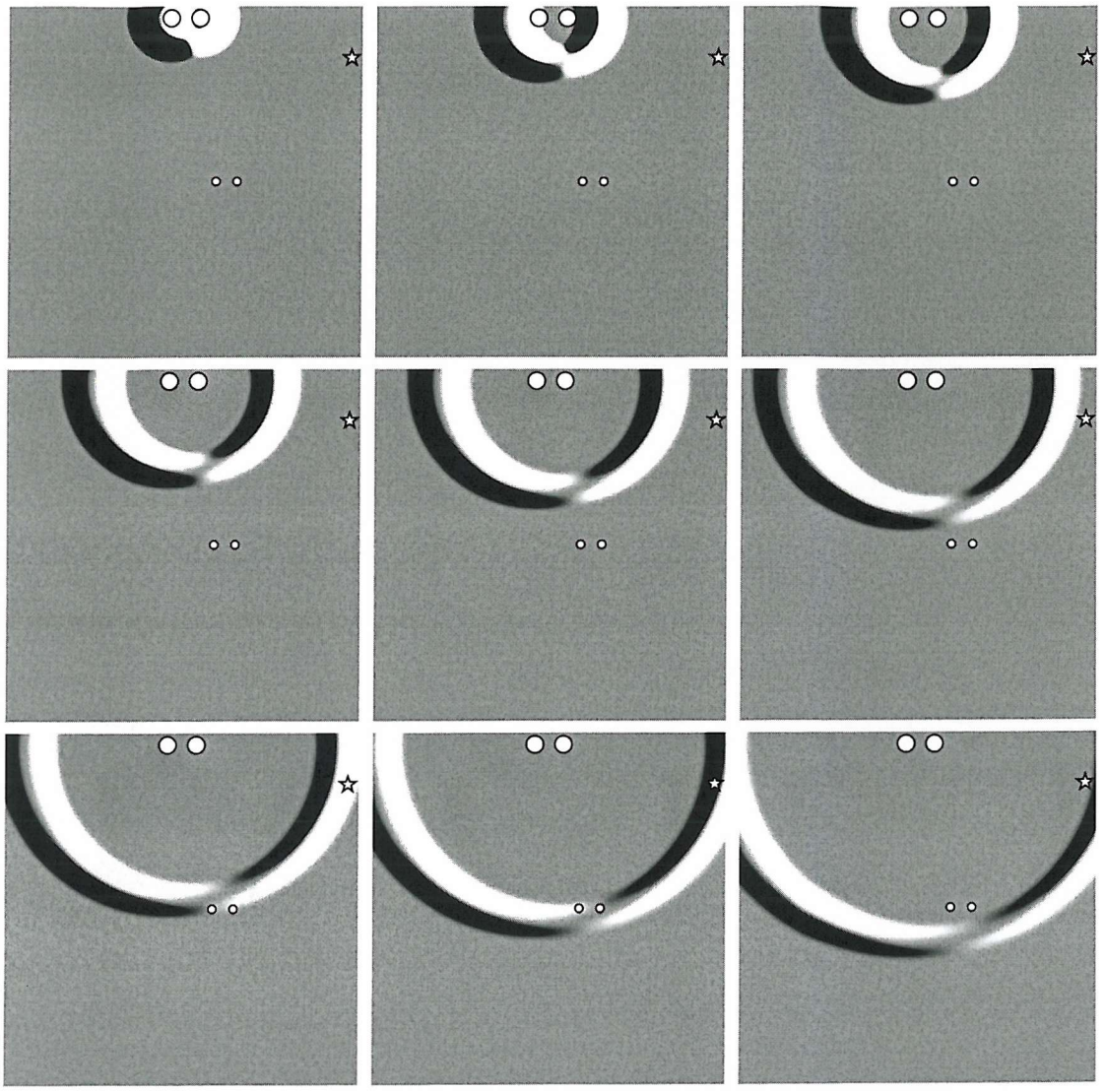


Fig. 4.19 *Sound field produced with the listener 35 cm to the right of the inter-source axis when creating a virtual image at the position of the star shape. The desired signal is the Hanning pulse.*

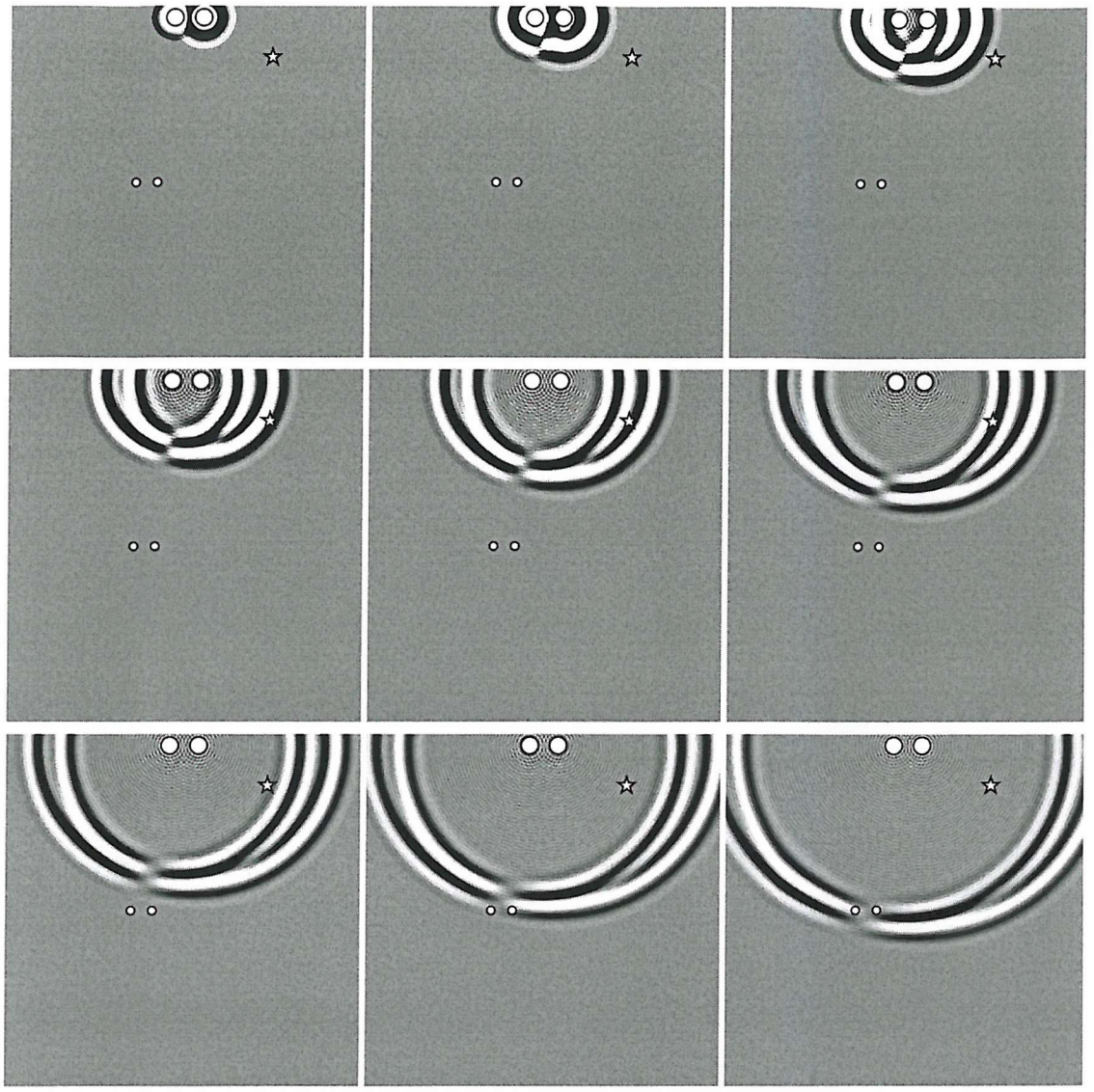


Fig. 4.20 *Sound field produced with the listener 35 cm to the left of the inter-source axis when creating a virtual image at the position of the star shape. The desired signal is the Gaussian pulse. Calculations based on the free field approximation.*

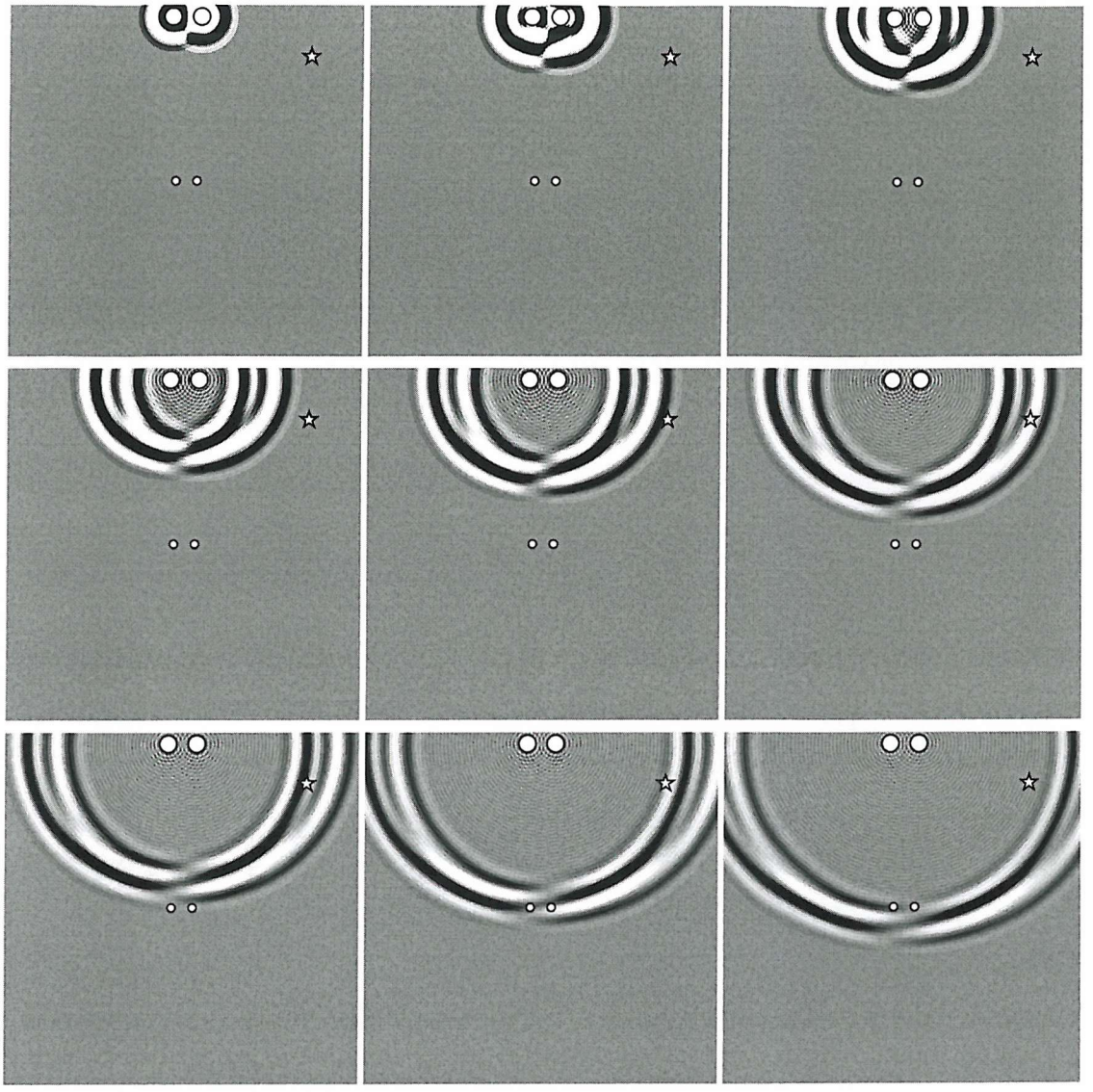


Fig. 4.21 *Sound field produced with the listener on the inter-source axis when creating a virtual image at the position of the star shape. The desired signal is the Gaussian pulse. Calculations based on the free field approximation.*

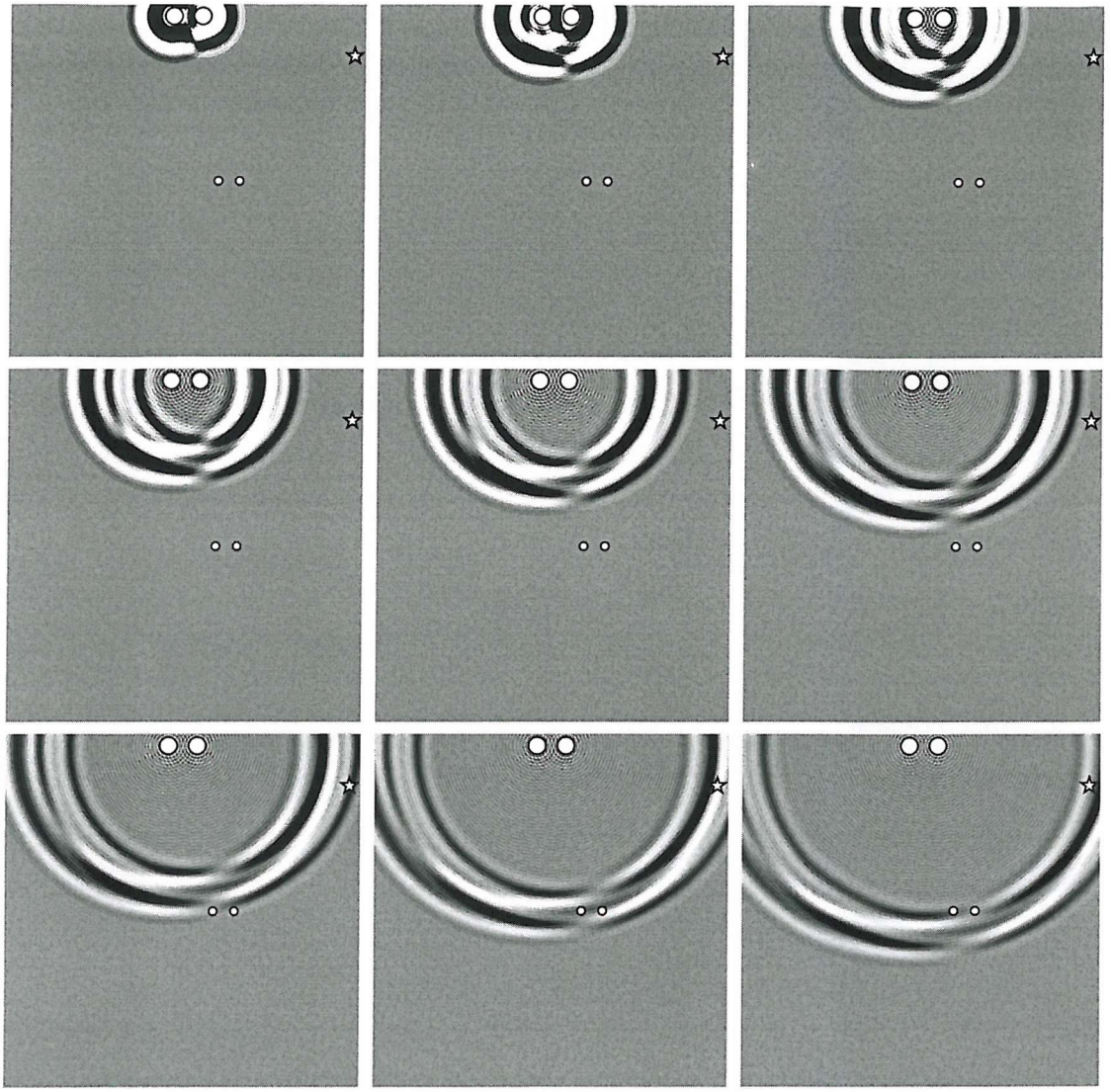


Fig. 4.22 *Sound field produced with the listener 35 cm to the right of the inter-source axis when creating a virtual image at the position of the star shape. The desired signal is the Gaussian pulse. Calculations based on the free field approximation.*

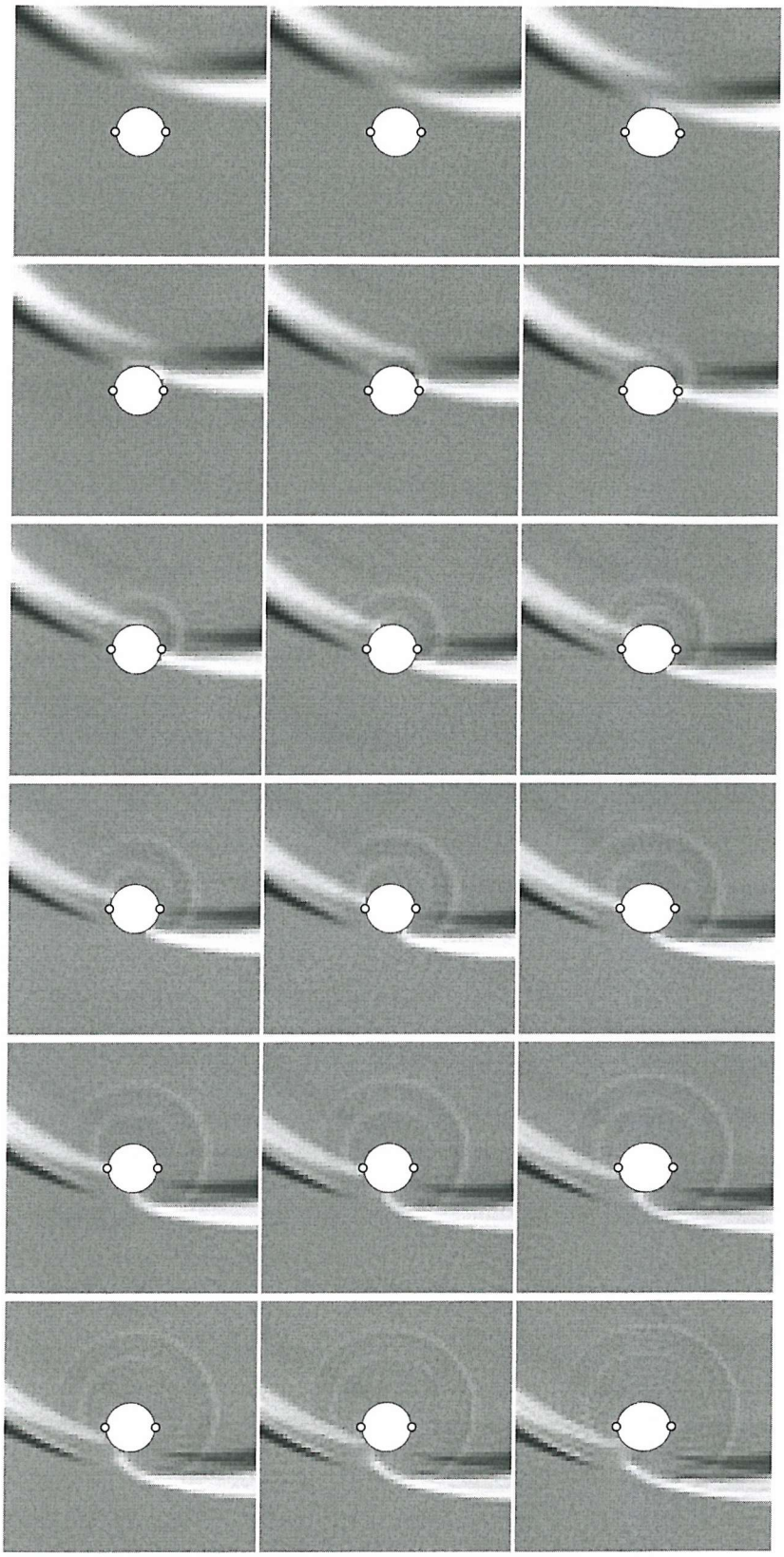


Fig. 4.23 *Sound field produced with the listener 35 cm to the left of the inter-source axis when creating a virtual image 45° to the listener's front right. The desired signal is the Hanning pulse. Calculations based on the spherical head approximation.*

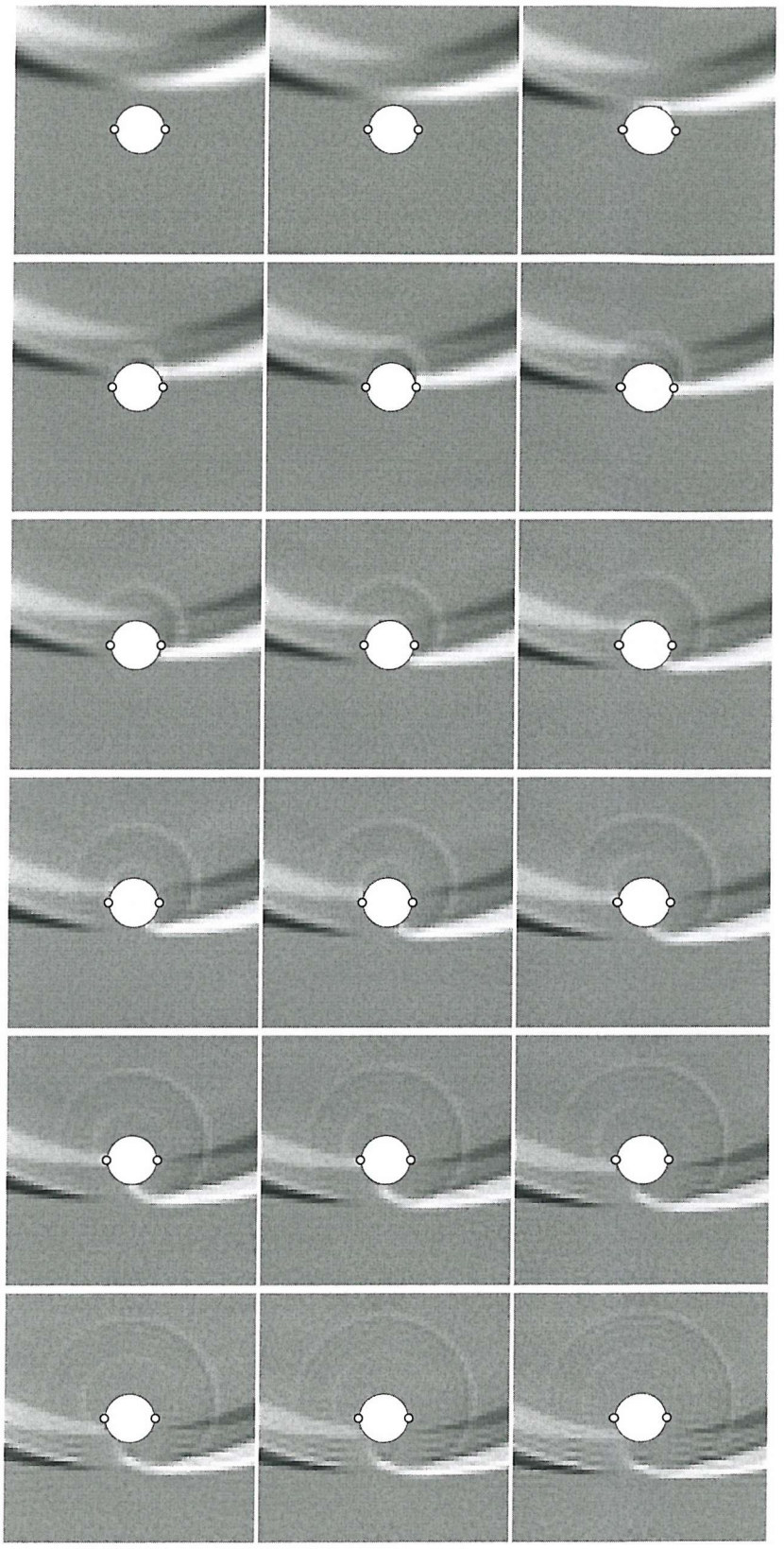


Fig. 4.24 *Sound field produced with the listener on the inter-source axis when creating a virtual image 45° to the listener's front right. The desired signal is the Hanning pulse. Calculations based on the spherical head approximation.*

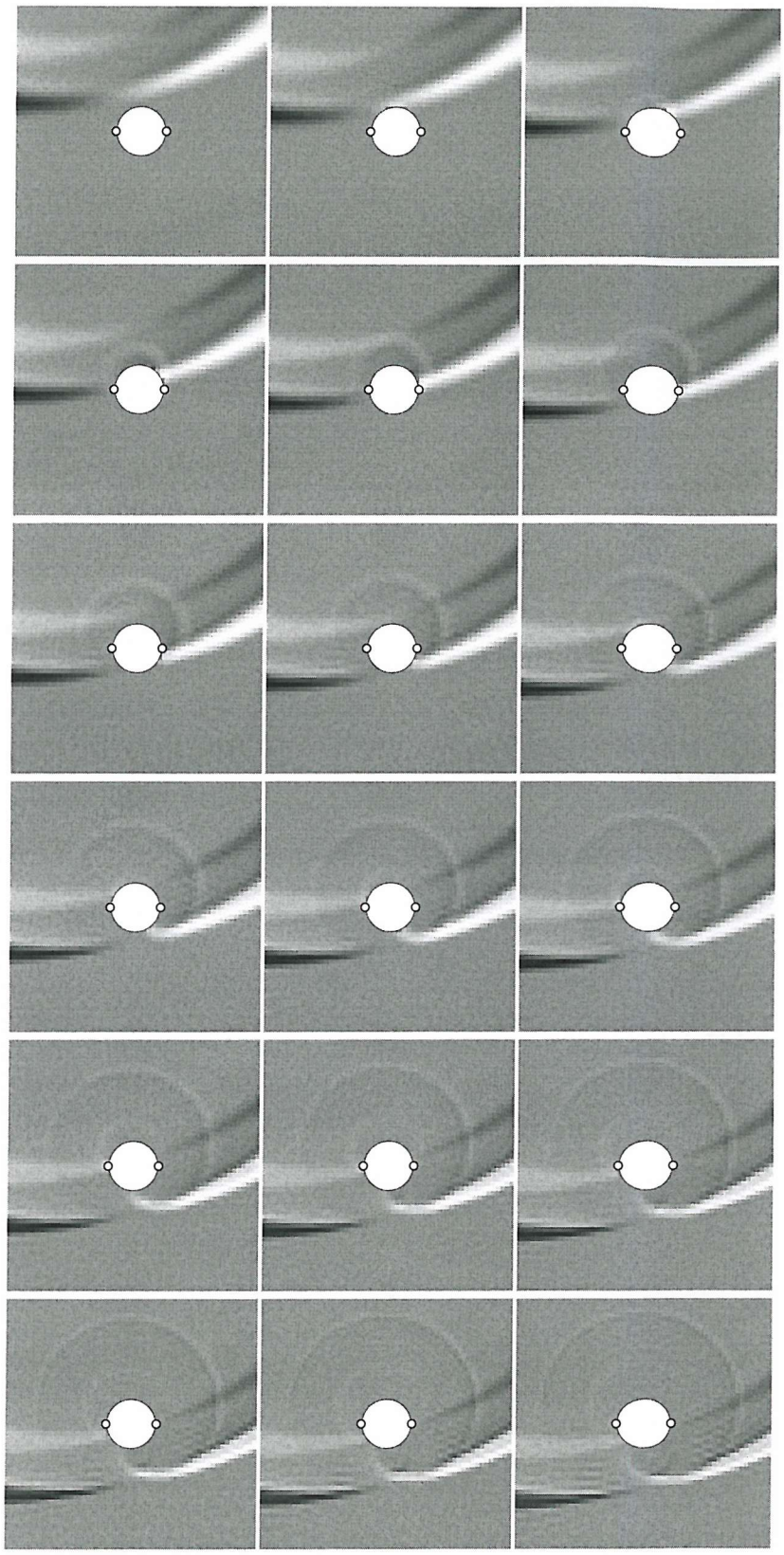


Fig. 4.25 *Sound field produced with the listener 35 cm to the right of the inter-source axis when creating a virtual image 45° to the listener's front right. The desired signal is the Hanning pulse. Calculations based on the spherical head approximation.*

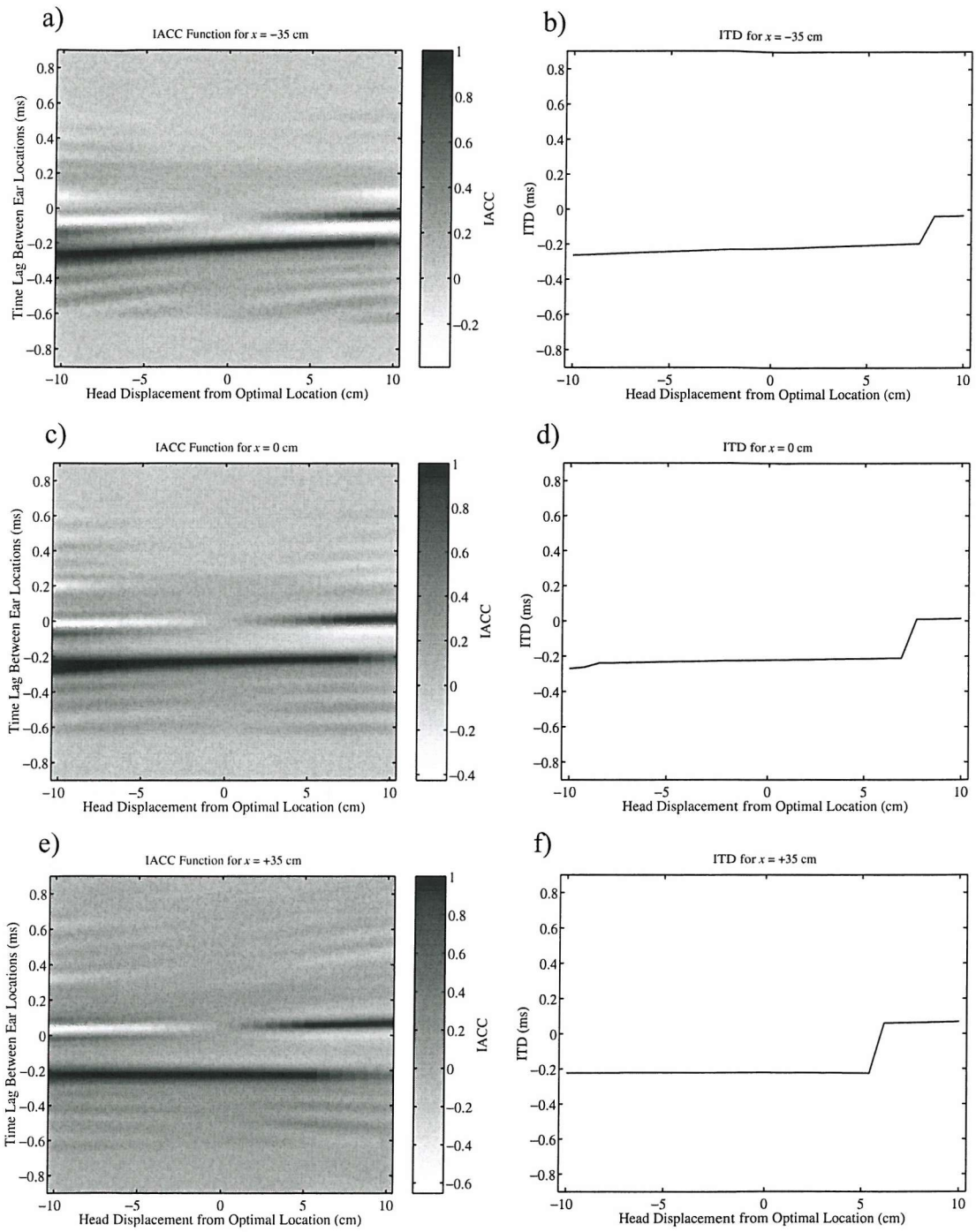


Fig. 4.26 *The IACC function and the maximum of IACC function (ITD) as a function of lateral head displacement from the intended listener location. The calculations correspond to the situations in Figs. 4.23-4.25 for a desired Hanning pulse with* $\theta_v = 45°$ *and calculated with the spherical head approximation. The intended listener location is 35 cm to the left of on-axis for a) and b), on the inter-source axis for c) and d), and 35 cm to the right of the inter-source axis for e) and f).*

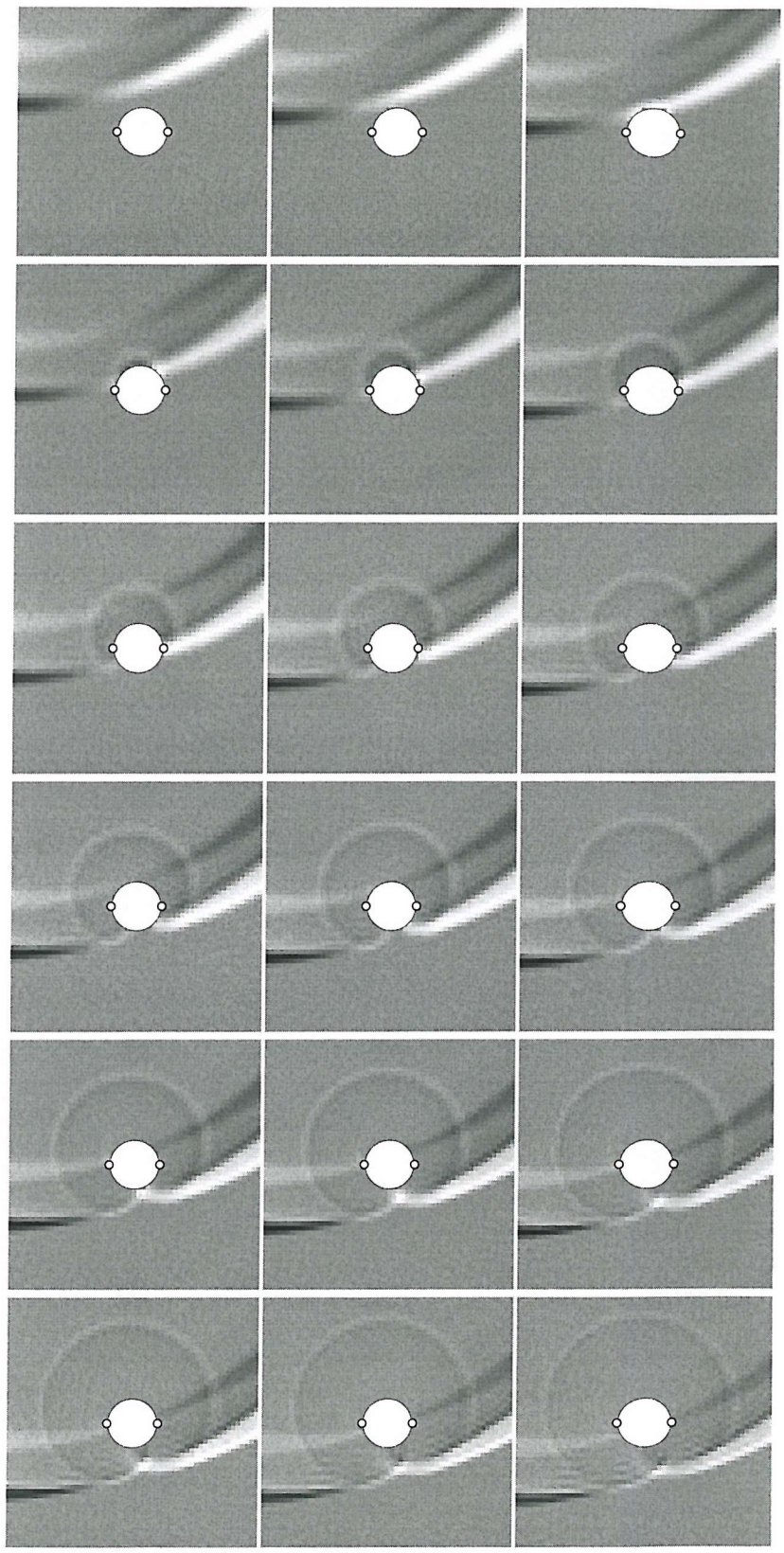


Fig. 4.27 *Sound field produced with the listener displaced 8 cm to the right of the intended listener location 35 cm to the right of the inter-source axis. The system is attempting to create a virtual image 45° to the listener's front right. The desired signal is the Hanning pulse. Calculations based on the spherical head approximation.*

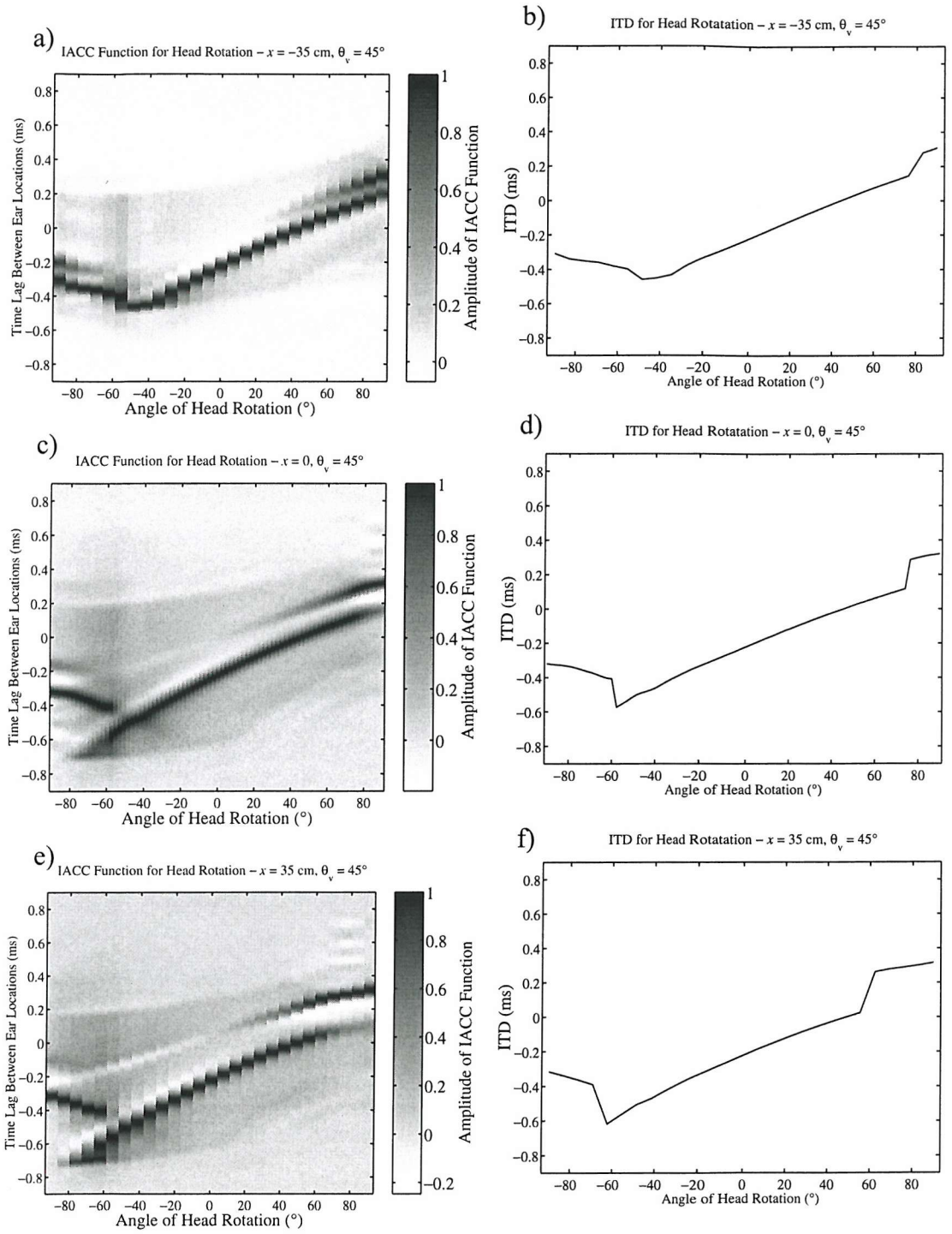


Fig. 4.28 *The IACC function and the maximum of IACC function (ITD) as a function of azimuthal head rotation from the intended listener location. The calculations correspond to the situations in Figs. 4.23-4.25 for a desired Hanning pulse with $\theta_v = 45°$ and calculated with the spherical head approximation. The intended listener location is 35 cm to the left of on-axis for a) and b), on the inter-source axis for c) and d), and 35 cm to the right of the inter-source axis for e) and f).*

5. SUBJECTIVE EXPERIMENT AT ASYMMETRIC LISTENER LOCATIONS

5.1. Introduction

This section describes a subjective experiment undertaken in an anechoic chamber to investigate the "sweet spot" size at a range of head positions in a static case. The subjective results are compared with the results of the calculated "sweet spot" size of the previous chapter that utilised criteria based on cross-talk cancellation performance and just noticeable difference in ITD.

5.2. Procedure

Band passed white noise with a pass-band of 0.3-3 kHz was used as the source signal. This band of frequencies was chosen to include frequencies which humans localise predominately by utilising ITD (below about 1.5 kHz [6-8]) and also to include the first ear canal resonance at about 2 kHz [7] but not include pinnae effects.

The subjects sat inside a 2 m-diameter sphere covered with a black curtain to hide the position of the loudspeakers. Figure 5.1 is a picture of this sphere inside the anechoic chamber where the experiment took place. A small headrest supported the subject's head and the subject was asked to limit their movement as much as possible. The two loudspeakers were mounted on a moveable slide outside of the sphere, 1.4 m away from the subject's head. The two loudspeakers subtend an angle of 10° with the subject's head when the loudspeakers are arranged symmetrically about the subject. The motion of the slide and thereby the loudspeakers was controlled via a computer in the anechoic chamber. Figure 5.2 shows that the loudspeaker's motion was limited to the horizontal direction parallel to a line joining the subject's ears. When asked, the subjects voiced their perception of the horizontal direction of the noise. Figure 5.2 also shows horizontal plane angle locations that are marked inside the sphere.

The virtual acoustic imaging filters were designed off line with the path responses approximated by monopole sources radiating sound onto a perfectly rigid sphere. The spherical approximation was used because it is can be calculated exactly for each unique position. Using the KEMAR dummy measurements database would make interpolation for distance and angle necessary, which introduce some error into the transfer functions. The intended optimal listener locations examined were when the inter-source axis exactly coincides with the inter-receiver axis and when it is offset 5 cm, 10 cm, 15 cm, 20 cm, and 25 cm to the right (i.e. x = 0, –5 cm, -10 cm, -15 cm, -20 cm, -25 cm). Figure 5.3 depicts these six (6) positions with small circles denoting the centre of the two loudspeakers. The intended location of the virtual acoustic image was 45° to the front right of the subject in all cases. The regularisation parameter β was set to 5×10^{-6}. This value is lower for the spherical head transfer functions than that suggested in chapter 3 for designs utilising KEMAR dummy measurements (i.e. 10^{-4}). Figures 5.4-5.9 show the time and frequency responses of the six (6) pairs of virtual acoustic imaging filters. These six (6) sets of virtual acoustic imaging filters were subjectively examined at loudspeaker locations as far as 15 cm from optimal locations. There were fourteen (14) different subjects each with normal hearing between ages 20-34 that took part in the experiment. Each subject evaluated at least two (2) different pairs of virtual acoustic imaging filters meant for two (2) different listener/loudspeaker arrangements, although most of the subjects listened to three (3). On average then, each set of virtual acoustic imaging filters were examined by seven (7) different subjects.

The experiment began by choosing one set of virtual acoustic imaging filters and moving the loudspeakers to the corresponding intended location. This set of filters was then left unchanged while the loudspeakers were displaced incrementally to the left of the position for which the filters were intended. The farthest point examined was about 15 cm away from the intended design location of the filters. The loudspeakers were then displaced incrementally to the right going back to the intended design location. The loudspeakers were then displaced incrementally to the right of the intended design location and then finally moved incrementally left going back to the intended design location. After every increment, the subject was given a

short burst of the stimulus and then asked the azimuthal angle where they perceive the location of the noise source. The loudspeaker increment distance was 1 cm.

5.3. Results

Figure 5.10 shows the subjects' responses for the virtual acoustic imaging filters intended for the design location 5 cm off the inter-source axis. The results for this listener location give an example of the types of results obtained at all six (6) positions examined. Figure 5.10 shows the difference in the angle perceived from the angle perceived when the loudspeakers are at the optimal location (ordinate), plotted against the displacement of the loudspeakers from the intended 5 cm off-axis optimal location (abscissa). The numbers 1-14 designate the different subjects and different symbols represent their individual responses in the plots. Solid lines in these graphs show the mean responses. In addition, error bars show the standard deviation of the data. The standard deviation seems considerable at points but the mean values show clear trends. Generally, the standard deviation is greater for the greater loudspeaker displacements from the intended design position. Front-back confusions are resolved in Fig. 5.10. That is, the plots show the listeners' perceptions of the white noise originating behind them at the corresponding mirrored angles in front. For the 5 cm off-axis listener location of Fig. 5.10, 6.8% of the responses were resolved in this fashion. Table A.1 in the appendix 1 displays the percentages of front-back reversals for all six (6) positions. Appendix 1 also provides figures displaying the percentages of front-back confusions for all locations (Figs. A.1 and A.2) and for the six (6) individual locations separately (Figs. A.3-A.8) as a function of loudspeaker displacement from the intended design location. The results for the other five (5) positions are given in appendix 1 in a similar fashion to Fig. 5.10 in Figs. A.9-A.13.

Figure 5.11 shows the mean responses for all six (6) of the examined designed filter positions in a similar fashion to Fig. 5.10. The six (6) designed filter positions yield similar results. At the optimal listener locations, the subjects generally tended to perceive the direction to be slightly in front of the target location of 45° (e.g. 30°). Other evaluations of virtual sound systems have also found this to occur [53]. Front-

back confusions are resolved in Fig. 5.11. Over all subject responses at the six (6) intended locations for this experiment, 12.2% of the subjects' perceptions were heard from behind them and were corrected for front-back confusions.

Figures 5.10a,b, and 5.11a,b show the results with the loudspeakers moving incrementally away from the intended design location. That is, the loudspeakers started out at the intended designed filter listener location and then were progressively displaced further and further away from the intended design location. Figures 5.10c,d and 5.11c,d show the results with the loudspeakers moving incrementally toward the intended designed filter listener location. Comparisons between the results when the loudspeakers were progressing toward and away from the intended design location reveal discrepancies between responses given with the loudspeakers at the exact same locations. This exhibition of hysteresis suggests that the direction of the subject's last perception affects the subject's next perceived direction. The subject was more likely to localise the sound from the last perceived direction.

Figures 5.10a,c, and 5.11a,c show the results with the loudspeakers left of the intended design location while Figs. 5.10b,d, and 5.11b,d show the responses with the loudspeakers to the right of the intended design location. Comparisons between the results with the loudspeakers on the right or on the left of the intended design location suggest that the virtual sound image maintains its intended location more easily with the loudspeakers to the right. Loudspeaker displacements away from the intended design location to the right correspond to the loudspeakers moving closer to the intended virtual sound image position. This implies that the system is more robust for head displacements away from the virtual source or loudspeaker movements toward the virtual source.

One also might interpret these figures to imply a "sweet spot" size of about 3 cm to one side of the designed filter listener location (6 cm overall). The virtual image collapses somewhat gradually to the spot directly in front of the listener as the listener

moves away from the optimal listening position. As a result, one must decide on how much error is tolerable in order to state the size of the "sweet spot".

5.4. Limitations of Results

While the results of the experiment are useful and provide insight into the problem, this subsection discusses some observations made of the experimental procedure and some associated possible limitations.

1. The virtual acoustic imaging filters were designed with the path responses approximated by monopole sources radiating sound that impinged on a perfectly rigid sphere. Therefore, the subjects' pinnae were not taken into consideration in the filter design. During the experiment, subjects often reported elevation movements. Many subjects localised the source behind them at times (i.e. front-back confusion). These types of responses may have been avoided if the design of the filters had incorporated the pinnae responses.
2. As was noted above, the subjects' responses show some bias toward the last direction that they perceived. An improvement to the experiment might be to present the different designed filters and loudspeaker displacements randomly. Although, this would make the experiment more time consuming.
3. Toward the end of the experiment, some of the subjects expressed a feeling of fatigue in their right ear. This is because the intended virtual image location was invariably 45° to the subject's right. It would make for a more comfortable and interesting experiment by including some variation in the virtual image location. This would complicate matters due to a dependence of "sweet spot" size on the number and locations of virtual sources as shown by the ITD simulations presented in section 4.4.

5.5. Conclusions

There is a steady fall in the perceived location of the virtual acoustic image as the loudspeakers are displaced away from the intended location. The system appears to be more robust when the loudspeakers are close to the virtual acoustic image location.

These results are in broad agreement with the simulations of chapter 4. Front-back confusions occur more commonly at increasingly large displacements from the intended optimal design location. These results should be qualified by the uncertainty in the experiment that is produced as a result of the simplified procedures used in getting the data.

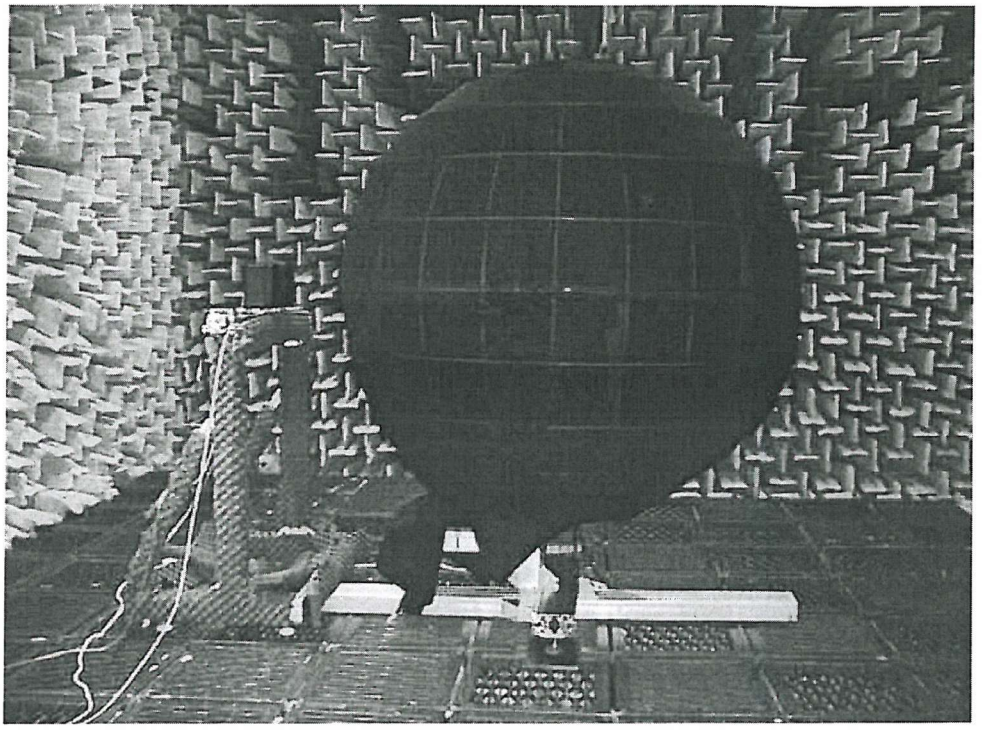


Fig. 5.1 *Set-up for subjective experiment inside anechoic chamber.*

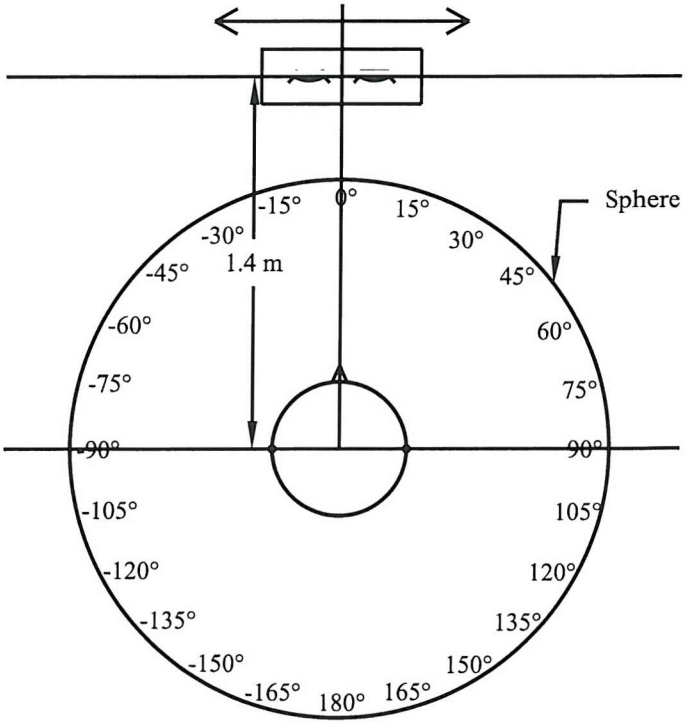


Fig. 5.2 *Plan view of the experimental arrangement with the subject inside the sphere, angles marked, and loudspeakers on a moveable slide.*

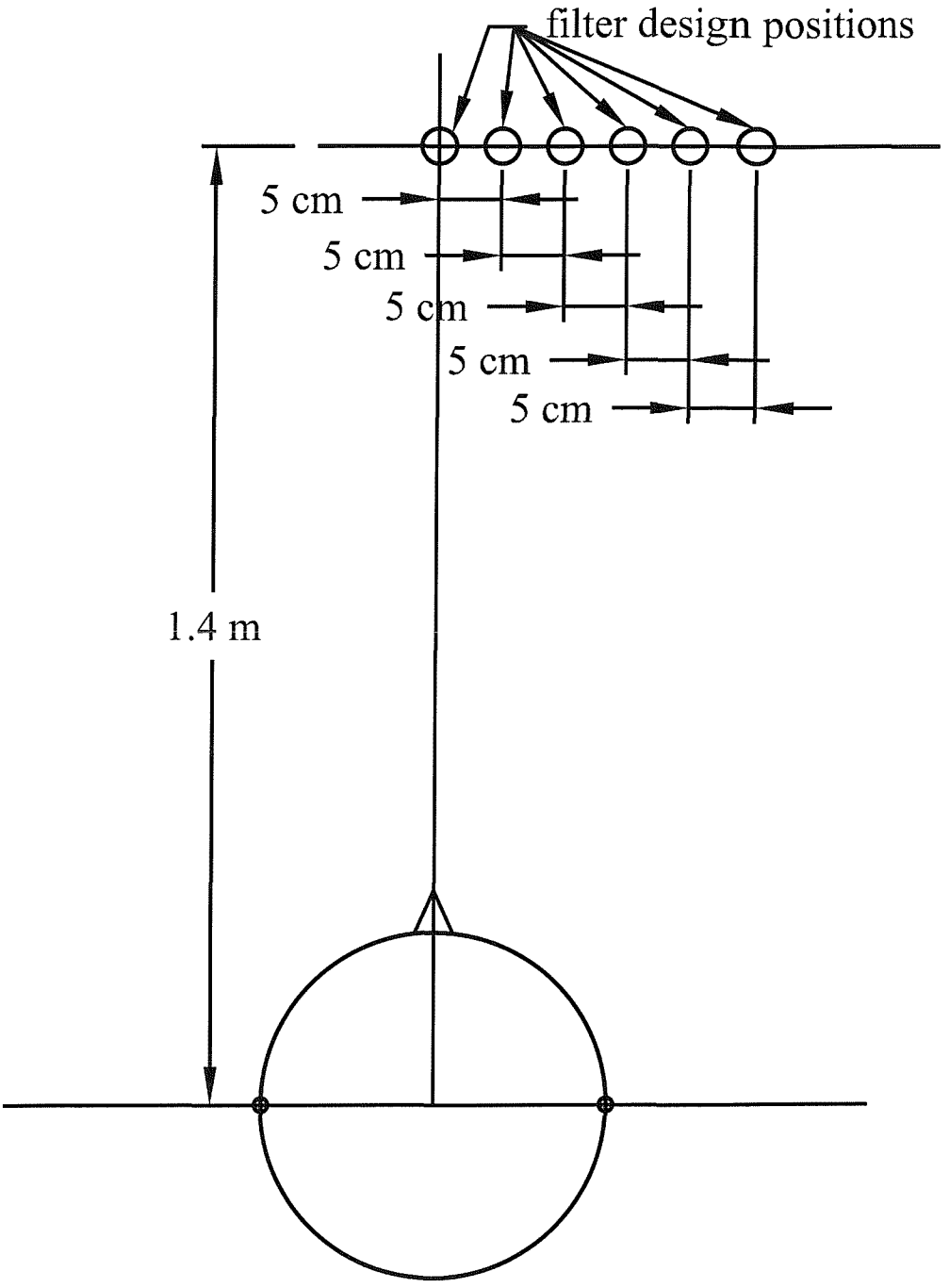


Fig. 5.3 *Plan view depicting the six (6) intended filter design positions that were examined subjectively. The small circles represent the point at the middle of the two loudspeakers for the six (6) positions.*

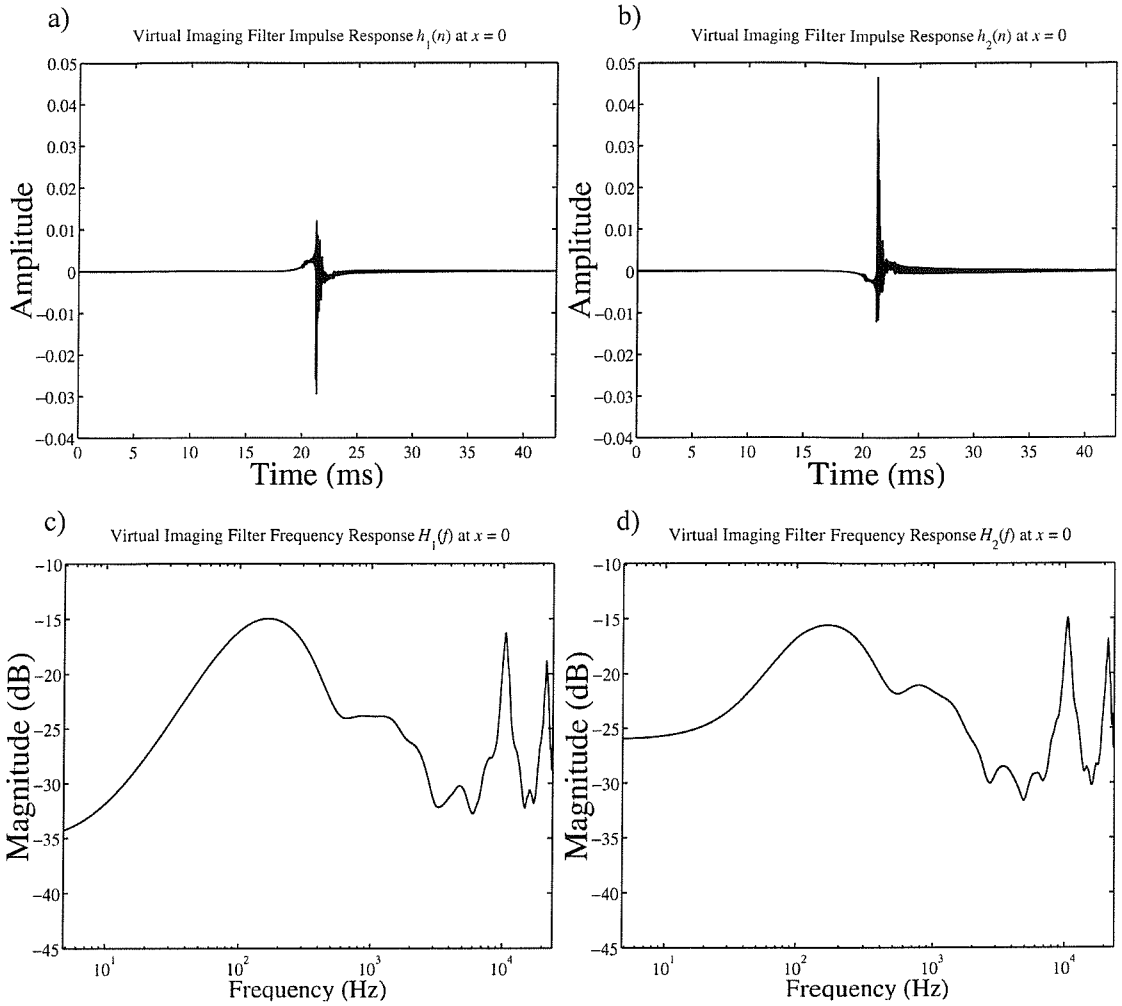


Fig. 5.4 *Virtual acoustic imaging filters for the on-axis symmetric location with the virtual acoustic image angle θ_v 45° to the listener's front right. The design is based on the spherical head approximation with the regularisation parameter β equal to 5×10^{-6}.*

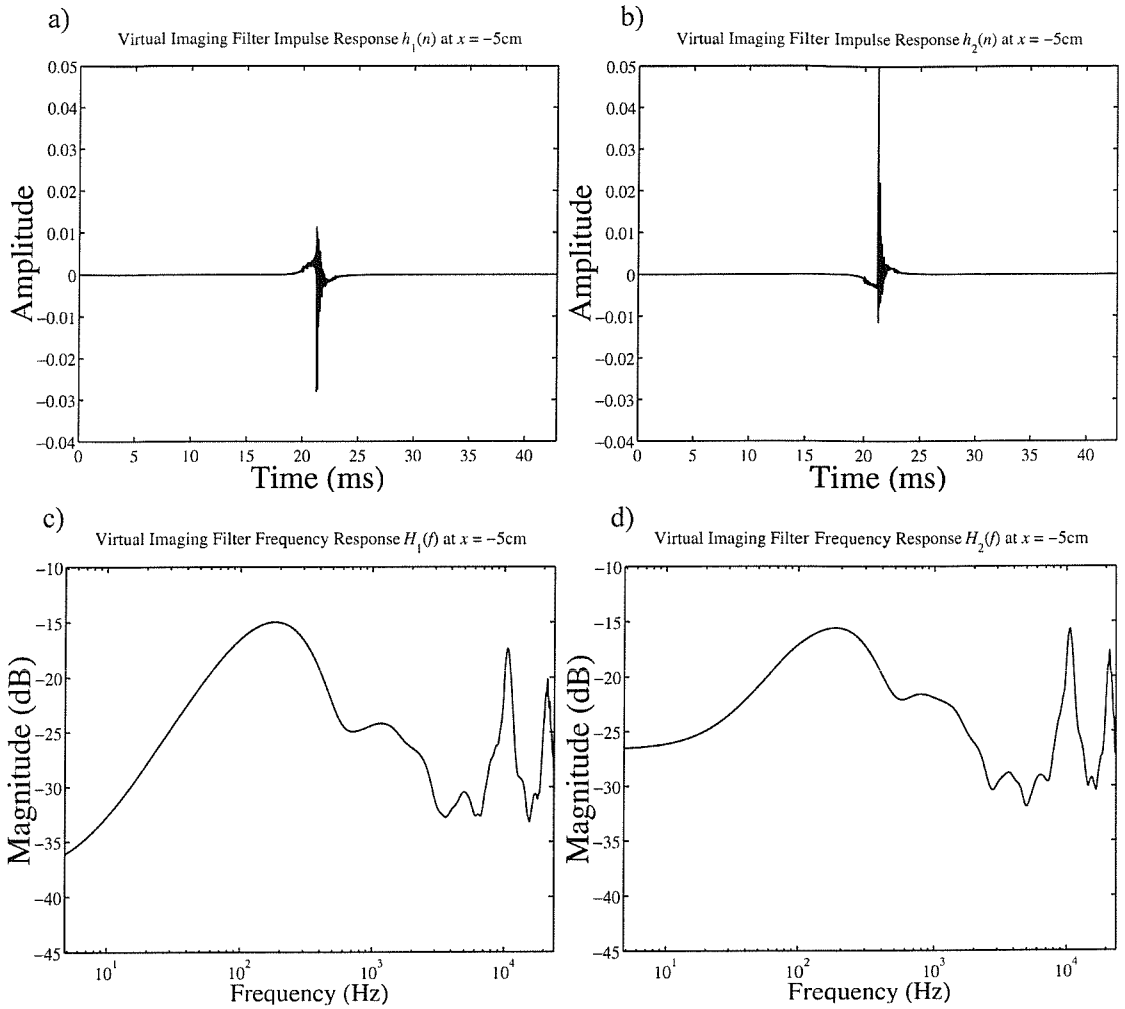


Fig. 5.5 *Virtual acoustic imaging filters for a listener located 5 cm to the left of the inter-source axis (i.e. x = 5 cm) with the virtual acoustic image angle θ_v 45° to the listener's front right. The design is based on the spherical head approximation with the regularisation parameter β equal to 5×10^{-6}.*

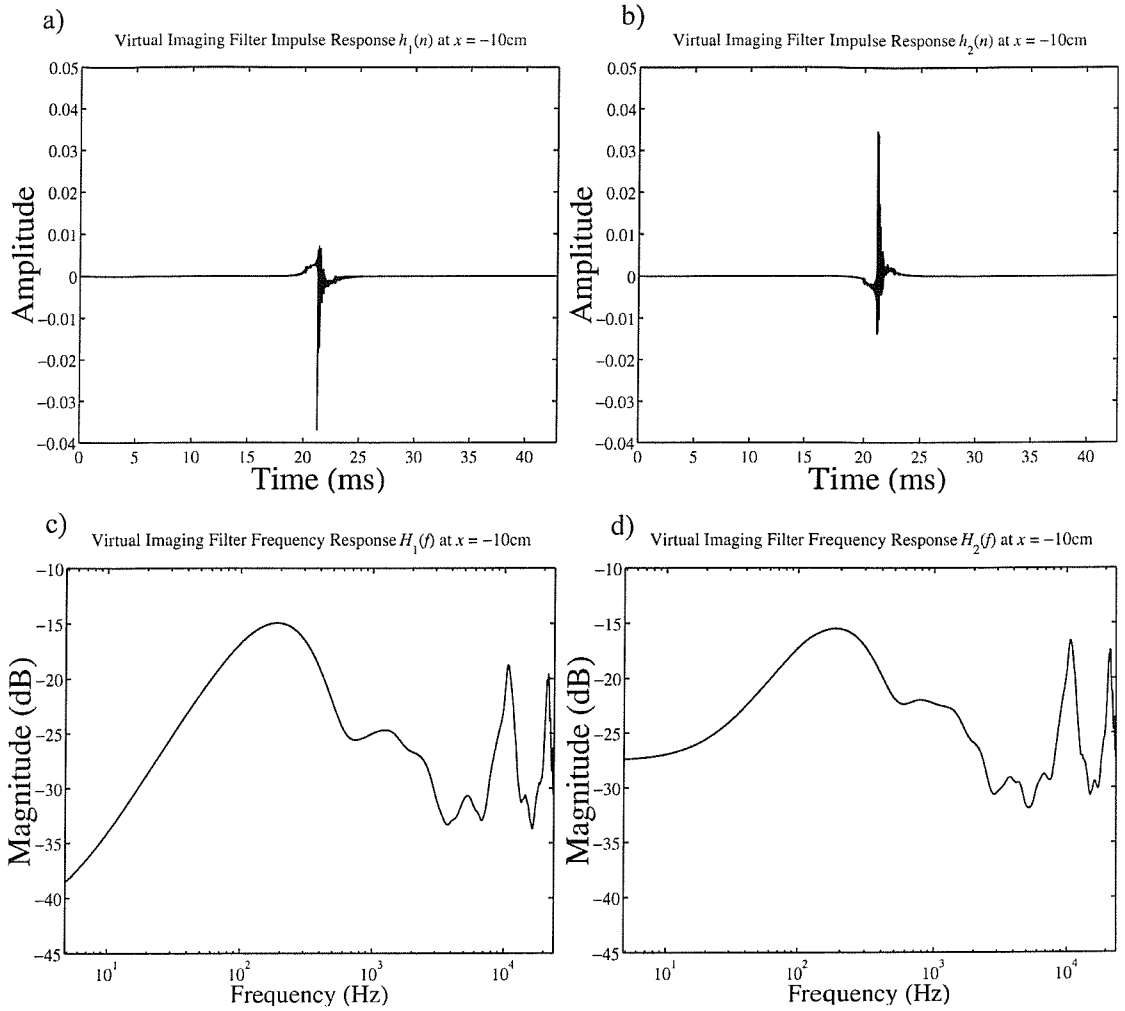


Fig. 5.6 *Virtual acoustic imaging filters for a listener located 10 cm to the left of the inter-source axis (i.e. x = 10 cm) with the virtual acoustic image angle θ_v 45° to the listener's front right. The design is based on the spherical head approximation with the regularisation parameter β equal to 5×10^{-6}.*

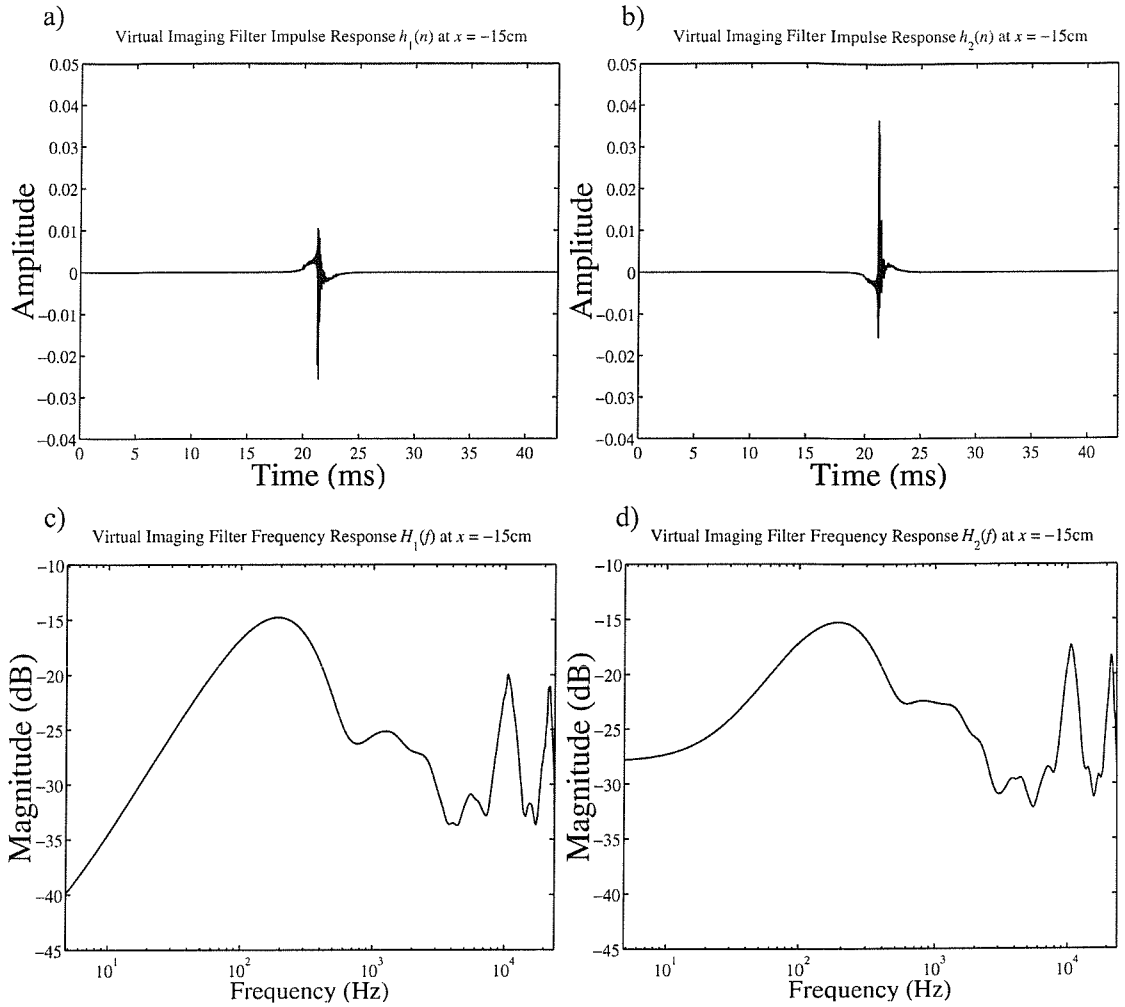


Fig. 5.7 *Virtual acoustic imaging filters for a listener located 15 cm to the left of the inter-source axis (i.e. x = 15 cm) with the virtual acoustic image angle θ_v 45° to the listener's front right. The design is based on the spherical head approximation with the regularisation parameter β equal to 5×10^{-6}.*

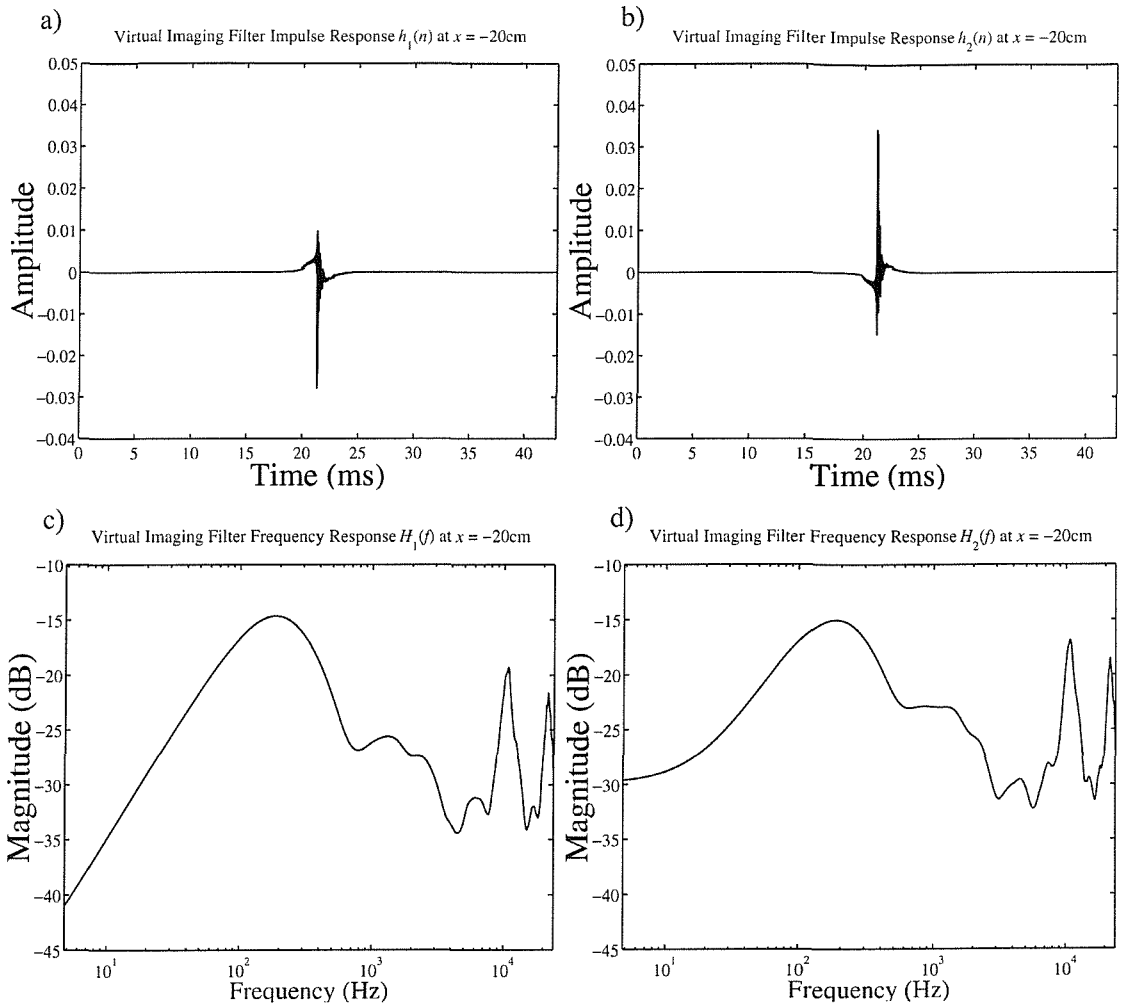


Fig. 5.8 *Virtual acoustic imaging filters for a listener located 20 cm to the left of the inter-source axis (i.e. x = 20 cm) with the virtual acoustic image angle θ_v 45° to the listener's front right. The design is based on the spherical head approximation with the regularisation parameter β equal to 5×10^{-6}.*

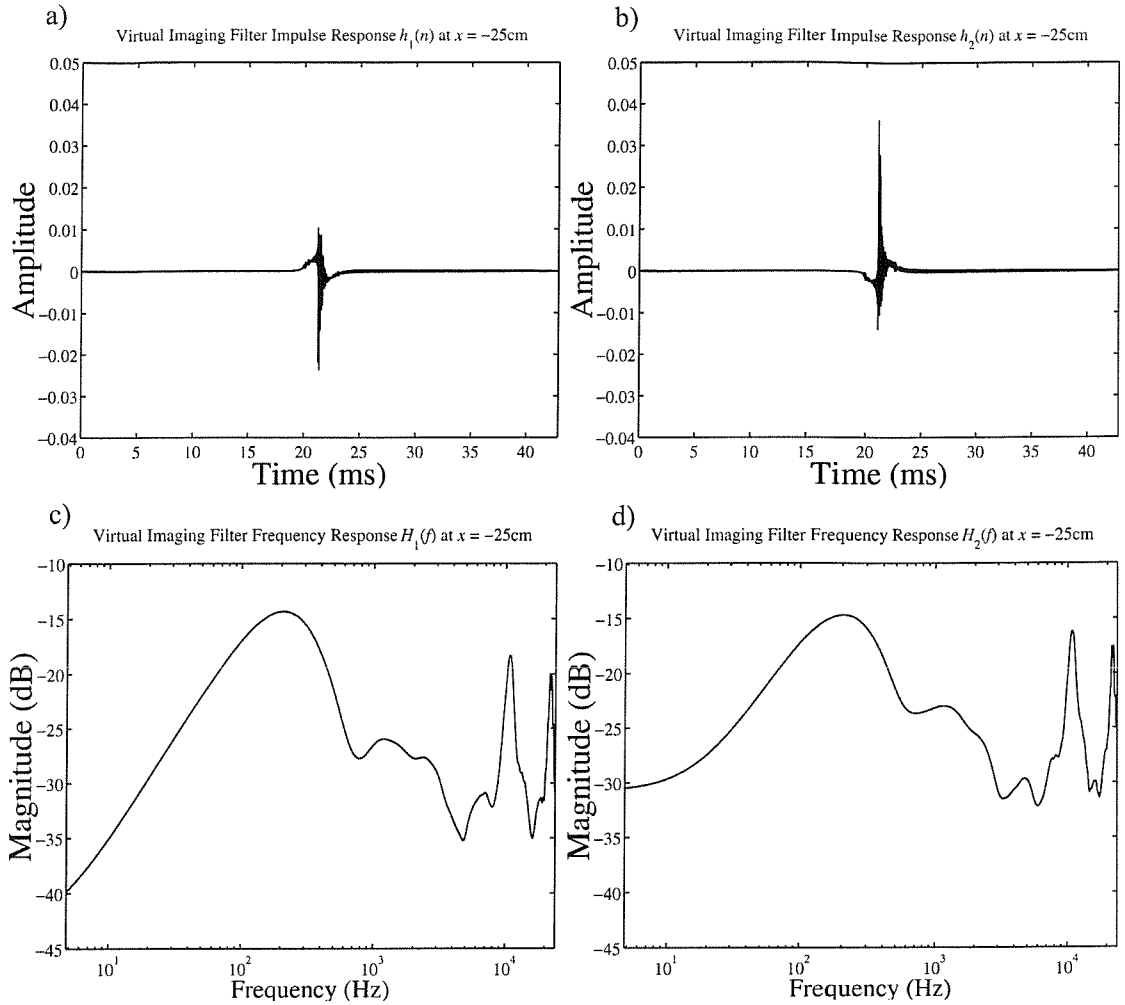


Fig. 5.9 *Virtual acoustic imaging filters for a listener located 25 cm to the left of the inter-source axis (i.e. x = 25 cm) with the virtual acoustic image angle θ_v 45° to the listener's front right. The design is based on the spherical head approximation with the regularisation parameter β equal to 5×10^{-6}.*

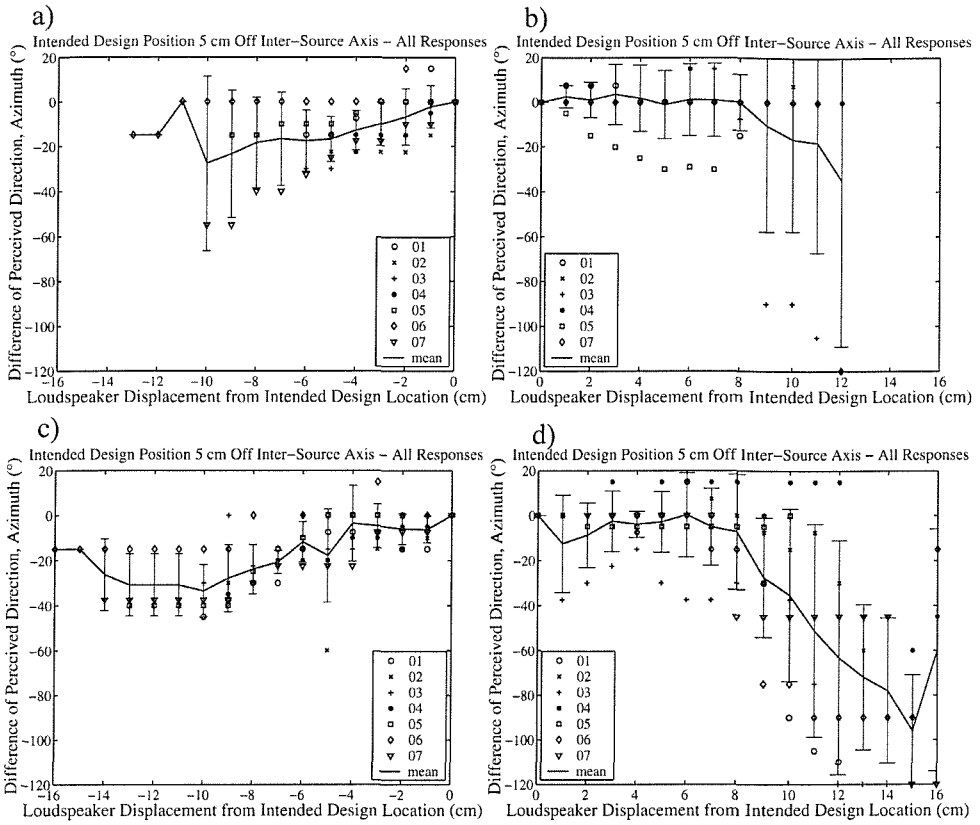


Fig. 5.10 *Subjects' responses for the 5 cm off-axis intended filter design position with front-back confusions resolved. Different symbols denote different subjects. Responses are shown with a) loudspeakers moving away from intended filter design location, to the left, b) loudspeakers moving away from intended design location, to the right, c) loudspeakers moving toward the intended design location, from the left, and d) loudspeakers moving toward the intended design location, from the right.*

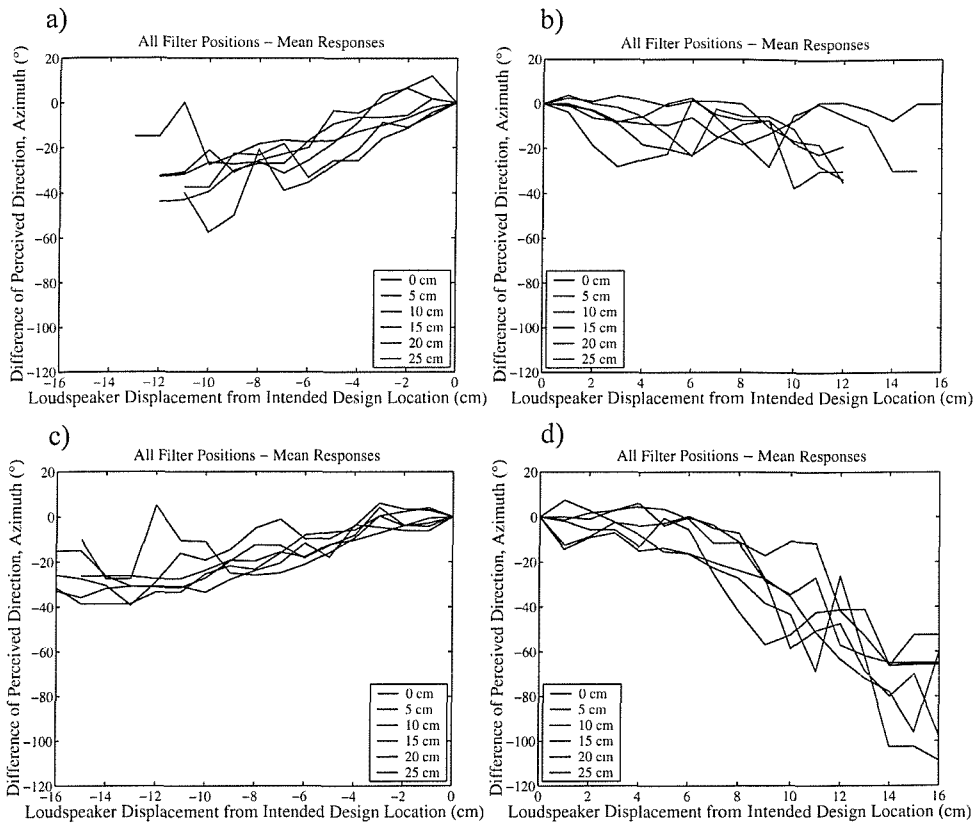


Fig. 5.11 *Mean subject responses for all six (6) different intended filter design positions 0-25 cm off the inter-source axis with front-back confusions resolved. Different lines represent different intended filter design positions. Responses are shown with a) loudspeakers moving away from intended filter design location, to the left, b) loudspeakers moving away from intended filter design location, to the right, c) loudspeakers moving toward the intended filter design location, from the left, and d) loudspeakers moving toward the intended filter design location, from the right.*

6. DYNAMIC SUBJECTIVE EVALUATION

6.1. Introduction

The effectiveness of a virtual acoustic imaging system depends on the listener's head being near the position assumed in the filter design. Should the listener move out of this immediate area, the filters need to be updated for the new arrangement. The last two chapters established the approximate size of the area of listener locations that do not require a filter update. This chapter describes a dynamic subjective evaluation of this update procedure. The parameters are types of stimuli, paths of movement, and frequency with which the filters are updated (which will be referred to as the "filter update movement increment"). The objective of this experiment is to find the threshold parameters necessary for two real moving sound sources to produce a stationary virtual sound image. The goal here is to understand the requirements for designing an effective adaptive virtual acoustic imaging system [5]. The subjects in the experiment were asked to respond to both the location and the stability of the virtual acoustic image.

6.2. Procedure

The experiment examined the Stereo Dipole virtual acoustic imaging system, which has an operational bandwidth that includes frequencies up to about 11 kHz (see chapter 3). Figure 6.1 shows a plan view of the experimental arrangement. The subject remained stationary while a step motor rotated a leadscrew, which moved the loudspeakers along a framework. The loudspeakers only moved in a horizontal direction as shown. A computer synchronised the loudspeaker motion with selection of appropriate virtual acoustic imaging filters from a DSP buffer. A 2 m-diameter sphere veiled with an acoustically transparent black cloth enclosed the seated subjects so that they could not see the loudspeakers. A headrest helped to limit head movement. This arrangement was assembled within an anechoic chamber.

The finite impulse response (FIR), virtual acoustic imaging filter designs employed KEMAR dummy HRTF measurements. Filtering with a fractional delay filter [43] and scaling with the inverse of distance corrected these HRTFs for distance differences between the current experimental arrangement and the HRTF measurement conditions (see the interpolation procedures described in section 2.4.3). The procedure for the design of the virtual acoustic imaging filters was the method of fast deconvolution with regularisation [54] as described in section 3.1. The regularisation parameter β was 10^{-4} for all of the filters designed in this experiment. Preparing for the experiment involved the design of the virtual acoustic imaging filters off-line. Figures 6.2-6.13 display some examples of the filters' time and frequency responses used in the experiment for listener locations x = -28 cm, -20 cm, -6 cm, 5 cm, 14 cm, and 35 cm off-axis with virtual acoustic image angles θ_V of both +45° and –45° in the horizontal plane. Note that the time responses of filters for locations farther away from the inter-source axis decay at a slower rate than filters for locations closer to the inter-source axis. The audio DSP card on the Lake DSP Huron audio convolution station stored the virtual acoustic imaging filters in a buffer. Huron's software allowed for continuous selection of the without interpolation between the stored filters. The lag between the time a "change filters" instruction is communicated and the new filters being selected is less than 0.01 seconds. Huron inputs connected to a communication card that controlled the loudspeaker motion.

Twenty (20) different subjects aged 20-34 with normal hearing each listened to twenty-eight (28) different trials with various parameter configurations that were presented in random order. The variables were two (2) different sound source signals, four (4) different loudspeaker paths or stationary loudspeakers with two (2) different virtual acoustic image angles θ_V, and six (6) different filter update movement increments. A session of twenty-eight (28) trials took about 40 min.

Unfiltered women's speech and band-passed white noise with a pass band of 0.3-3 kHz were the two different sound source signals. The virtual acoustic image angle θ_V was either 45° to the right front or 45° to the left front of the subject in the horizontal plane. Figure 6.14 depicts four (4) different loudspeaker path-virtual acoustic image

angle θ_v combinations that subjects examined in the trials. The loudspeakers emitted the sound source signal only while moving along the paths. The maximum rate of the loudspeakers' movement was 5 cm/s and the loudspeakers traversed a total of 70 cm so that the signal's presentation time was about 14 seconds for each trial. The subjects examined six (6) different filter update movement increments (2 cm - 7 cm). The subjects also examined 14 s reference trials with the loudspeakers stationary and a single pair of virtual acoustic imaging filters selected.

With the loudspeakers motionless at the symmetric position directly in front of the subject, the loudspeakers emitted a 4 s unfiltered reference stimulus before each trial. This was to orient the subject and familiarise them with the sound stimulus. The subject then listened to a single trial with the same sound stimulus and other selected variables (virtual acoustic image angle θ_v, loudspeaker path, filter update movement increment). The subject could read angle locations from markings inside the sphere. Before the experiment began, the experimenter instructed the subject to limit their responses to the horizontal angle locations.

After presentation of each trial, the subject wrote down their answers to two questions stated on a clipboard held by the subject. Depending on the type of sound source signal for the trial the first question was either "At what angle did you predominantly hear the white noise source in degrees (please estimate angles between markers)?" or "At what angle did you predominantly hear the speech source in degrees (please estimate angles between markers)?". This first question tested the subject's ability to localise the virtual acoustic image under the current parameter conditions. The second question was either "What percentage of time did you hear the white noise source from this location (to nearest 10%)?" or "What percentage of time did you hear the speech source from this location (to nearest 10%)?". This last question was in an effort to quantify the subject's perception of the stability of the virtual sound image. The subject could have asked to have a trial repeated if they wished. The presentation of the unfiltered stationary reference signal always preceded a trial presentation.

6.3. Results

Front-back confusions are resolved in all of the results below. The percentage of front-back confusions was 23.9% for all 560 responses of perception of image location in this experiment. This is a larger percentage than was found in the static experiment discussed in chapter 5 with filters based on the sphere HRTFs.

The difference in mean responses under different conditions may be due to the different experimental conditions or due to a general variation in the responses that is common to all sets of experimental parameters. The versatile statistical technique, analysis of variance (ANOVA) [67], separates out the sources of differences in mean responses and allows confidence statements about the effect of the different experimental conditions on the mean response. By assuming normal distributions of the data, this powerful numerical tool takes advantage of the additive property of variance or specifically that the variance for a whole system is equal to the sum of the component variances of subdivided data-sets. The ANOVA technique involves computation of the variance of component factors and assessment of the relative importance of the various components.

Table B.1 in appendix 2 gives an ANOVA table of the perceived angle difference comparing the responses given to the speech and white noise stimuli. A detailed explanation of this table provides an example of ANOVA tables. One can designate all of the subjects' responses to the speech stimulus as $y_{11}, y_{12}, y_{13}, \cdots, y_{1N_1}$ and all of the subjects' responses to the white noise stimulus as, $y_{21}, y_{22}, y_{23}, \cdots, y_{2N_2}$ where N_1 and N_2 are the number of responses to the speech and white noise stimuli respectively (here both equal to 280). The second column in Table B.1 is the sum of the squares of the standard deviations from mean responses. Between the stimuli the sum of squares is concerned with deviation of the mean responses of the different stimuli $\bar{y}_i$ from the overall mean response $\bar{y}$ and is calculated from

$$\sum_{i=1}^{K} N_i \left(\bar{y}_i - \bar{y}\right)^2 \qquad (6.1)$$

where K is equal to the number of conditions being compared (here $K = 2$, because two different stimuli are being compared i.e. white noise and speech). Within the stimuli the sum of squares is concerned with the deviations of the individual responses y_{ij} from the associated mean response within the stimuli $\bar{y}_i$ and is calculated from

$$\sum_{i=1}^{K} \sum_{j=1}^{N_i} \left(y_{ij} - \bar{y}_i\right)^2 \qquad (6.2)$$

and the total sum of squares is concerned with the deviations of the individual responses y_{ij} from the overall mean response $\bar{y}$ and calculated from

$$\sum_{i=1}^{K} \sum_{j=1}^{N_i} \left(y_{ij} - \bar{y}\right)^2 . \qquad (6.3)$$

For the example in Table B.1 the grand mean over all responses is $\bar{y}$ = 13.41°, the mean response for speech is $\bar{y}_1$ = 11.45° and for white noise is $\bar{y}_2$ = 15.36°, as shown in Table B.2, and the number of conditions being compared is $K = 2$ stimuli.

The third column in Table B.1 is the number of degrees of freedom (df) associated with the sum of squares. Between stimuli, the number of degrees of freedom is the number of stimuli minus one (i.e. $K - 1$, here equal to 1). Within stimuli, the number of degrees of freedom is the sum of the number of degrees of freedom in each stimulus (i.e. $\sum_{i=1}^{K} N_i - K$, here equal to 558). The total number of degrees of freedom

is the total number of samples minus one (i.e. $\sum_{i=1}^{K} N_i - 1$, here equal to 559). Note both values in the last row of Table B.1 are equal to the addition of the values in the top two rows.

The fourth column in Table B.1 is the mean square and is equal to the sum of squares divided by the number of degrees of freedom (i.e. the value in the second column divided by the value in the third column). The fifth column shows the F-value (viz. 4.112), which is the ratio of the mean square between stimuli to the mean square within stimuli. The F-distribution with d.f. = (1, 558) and a F-ratio of 4.112 corresponds to a significance of 0.043, which is value in the last column of Table B.1. The F-distribution assumes a normal distribution of the sampled data. Assuming the data is distributed normally one can now say with a 100 × (1 - 0.043) = 95.7% level of confidence that the mean responses under the two different stimuli (i.e. $\bar{y}_1$ = 11.45° and $\bar{y}_2$ = 15.36°) are significantly different and their difference is not the result of chance. Appendix 2 gives tables of the mean and standard deviations of the responses under varying conditions as well as more analysis of variance (ANOVA) tables calculated in the same fashion.

The first question tests the subject's localisation of the virtual acoustic image. The first dependent analysed here is the difference between the intended virtual acoustic image angle θ_v and the subject's perceived angle location in degrees. Negative values represent perceptions of the virtual acoustic image closer to the subject's front than the intended location of 45° to the side front of the subject. Positive values represent perceptions of the virtual acoustic image farther to the subject's side than the intended location of 45° to the side front of the subject. The second question tests the subjects' perception of the stability of the virtual acoustic image. The second dependent variable analysed here is the percentage of trial time that the subject perceives a stable virtual sound image.

6.3.1. Stimuli comparison

Figure 6.15 compares the mean localisation responses for the two different stimuli at the six (6) examined filter update movement increments and stationary reference trials. The responses for white noise tend to be a few degrees more to the listener's side than the speech. The mean responses over all trials for the speech and white noise are 11.5° and 15.4° to the side of the intended virtual sound source location (i.e. 56.5° and 60.4° to the subject's side) respectively. The analysis of variance (ANOVA) significance test shows that the stimuli mean responses are significantly different with a confidence level of 95.7% as discussed above (Table B.1). The subjects localised the white noise more to their side than the speech signal.

Figures 6.16 and 6.17 show localisation box-plots [68] for the different source signals during the stationary reference trials and the trials with a 5 cm filter update movement increment respectively. A box-plot is a type of graphic display convenient for data-set comparisons. The shaded boxes' upper and lower bounds are the third and first quartiles respectively, so the vertical length of the box is the interquartile range. Fifty percent (50%) of all of the data-set observations lie within this range. A horizontal line within the box indicates the median response. The vertical lines extending from the boxes, known as whiskers, show the range of the rest of the data excluding outliers. Outliers lay at least one and a half box lengths away from the box edges. All calculations operate on entire data-sets including outliers. The numbers along the abscissa, labelled *N*, show the size of the data-sets or the number of trials each box represents.

The data displayed in Figs. 6.16 and 6.17 all have medians of 15°. The means for the speech and white noise during the stationary reference trials are 11.8° and 18° respectively. During the 5 cm filter update movement increment trials the means are 9° and 15.7°. The interquartile range for the white noise source signal tends to include more points further to the subject's side while the interquartile range for the speech tends to include more points in front of the listener. This reflects the listeners' tendency to hear white noise more to their side than the speech signal. A comparison of Figs. 6.16 and 6.17 shows that the filter update procedure affects the perceived

angle location of white noise more than it affects the perceived angle location of speech.

The speech signal is not low-passed filtered and so contains higher frequencies than the band-passed white noise signal which is low-passed filtered at 3 kHz. Localising higher frequencies at angles more to the listener's front or sometimes actually from the system's two frontal loudspeakers is thought to be due to the smaller spatial control region or "sweet spot" of the imaging system at shorter sound wavelengths [69].

Figures 6.18 and 6.19 compare image stability for the two different sound source signals. Figure 6.18 compares box-plots of the stability of the different source signals during the stationary reference trials. The median response is a stable virtual sound image 100% of the time during the trial for both the speech and white noise source signals. The interquartile ranges of the speech and white noise are 7.50% and 8.75% respectively. The mean responses of the speech and white noise are 96.5% and 95.4% respectively. ANOVA analysis on the results for the stationary reference trials finds no significant difference in the mean stability responses for the two different stimuli (Table B.4). The type of stimuli appears not to affect stability of the virtual sound image during the stationary trials.

Figure 6.19 shows the mean stability responses for both speech and white noise at all the examined filter update movement increments. Increasing the filter update movement increment decreases the virtual acoustic image stability. After 5 cm, the stability of the virtual acoustic image decreases rapidly especially for the speech.

6.3.2. Loudspeaker path comparison

Figure 6.20 shows the mean localisation responses given for the four different loudspeaker paths examined. Figure 6.14 depicts the four different paths and designates them "toward-away" (*ta*), "away-toward" (*at*), "toward-toward" (*tt*), and

"away-away" (*aa*). There is little difference between the responses given during paths "toward-toward", "away-away", and "away-toward" at filter update movement increments below 6 cm. The path "toward-away" tends to yield responses more to the side of the subject than the other paths.

Figure 6.21 shows a box-plot comparing the localisation responses for all trials including the stationary references (designated *nm* "no-movement"). The mean localisation responses over all trials for movement types "away-away", "away-toward", "no-movement", "toward-away", and "toward-toward" are 9.8°, 11.2°, 14.9°, 19.3°, and 12.3° to the side of the intended virtual source location (i.e. 54.8°, 56.2°, 59.9°, 64.3°, and 57.3° to the subject's side) respectively. The median response for path "toward-away" is 25° further to the side of the subject than the intended virtual acoustic image angle θ_v (i.e. at 70° rather than the intended 45°). The median response for all other path types is 15° further to the side of the subject than the intended virtual acoustic image angle θ_v (i.e. at 60° rather than the intended 45°). ANOVA significance tests show that the movement type mean responses are significantly different with a confidence level of 98.6% (Table B.6). There is a low probability (1.4%) that the difference in the mean perceived sound image locations for the five different loudspeaker paths is the result of chance. There is a high confidence level (98.6%) in saying that the path of movement of the loudspeakers affects the mean localisation of the sound image. The responses for the path "toward-away" deviate from the other paths. Throughout the trials using this path, the loudspeakers are on the same side of the subject as the virtual acoustic image. The "toward-away" movement type seems to push the virtual acoustic image angle θ_v further off axis than other movement types. It is not clear why this happens.

Figure 6.22 shows the mean stability responses given for the four different loudspeaker paths. The most successful path at creating stable virtual sound images is "toward-away". The "toward-away" trials are the only trials with the loudspeakers and virtual acoustic image on the same side of the subject throughout the entire trial. The virtual acoustic image is always close to the loudspeakers in these trials.

The two paths that traverse both sides of the subject ("toward-toward" and "away-away") both perform worse at creating a stable virtual acoustic image than the two paths that traverse only one side of the subject ("toward-away" and "away-toward"). The loudspeakers travel a range twice as much in these cases. In addition, the system uses more filters and selects them each a single time. This is different from the movement paths when the loudspeakers travel back and forth on one side of the subject (*ta* and *at*) and the system selects most of the filters twice and uses only about half the number of different filters. The system requires more filters and has more difficultly creating a stable virtual sound image with movement spanning greater ranges.

Figure 6.23 shows a box-plot of all the trials comparing stability of the movement types including the stationary reference trials. Under dynamic conditions, the system is most successful at creating a stable virtual acoustic image with the loudspeakers close to the virtual source ("toward-away") and traversing small distances ("toward-away" and "away-toward").

6.3.3. The effect of filter update movement increment

Figure 6.24 is a box-plot comparing the localisation responses at the different filter update movement increments. Figure 6.25 shows the mean responses at the different filter update movement increments with error bars showing the standard deviation of the data. The median response over all of the data is 15° to the side of the subject further than the intended virtual acoustic image angle θ_v of 45°. Therefore, the most common response is 60° to the side of the subject. The two filter update movement increments that do not have a 15° median response are 2 cm and 6 cm, which have median responses of 25° and 10° (i.e. 70° and 55° to the subject's side) respectively. The mean responses over all trials for the stationary reference trial and filter update movement increments of 2 cm, 3 cm, 4 cm, 5 cm, 6 cm, and 7 cm are 14.9°, 25.5°, 12.4°, 9.6°, 12.4°, 9.3°, and 14.3° to the side of the intended virtual source angle θ_v (i.e. 59.9°, 70.5°, 57.4°, 54.6°, 57.4°, 54.3°, and 59.3° to the subject's side) respectively. ANOVA significance tests show that the mean responses associated with the filter update movement increment are significantly different with a confidence

level of 99.9% (Table B.9). There is a small probability (0.01%) that the difference in the mean perceived sound image angle for the seven different filter update movement increments is the result of chance. There is a high confidence level (99.9%) in saying that the filter update movement increment affects the subject's mean localisation of the sound image. However, by excluding the 2 cm filter update movement increment trials and repeating the ANOVA analysis one finds that the mean responses given for all of the other trials are not significantly different. There is a low confidence level (46.5%) in saying the filter update movement increments 3-7 cm affect the subject's mean localisation of the sound image (Table B.10). The experiment did not find that the filter update movement increment affects the localisation of the sound image except for a filter update movement increment of 2 cm. The reason why a 2 cm filter update movement increment affects the localisation of the sound image is not clear.

Figures 6.26 and 6.27 show a box-plot and graph of the mean stability responses comparing the different filter update movement increments. The clear trend is that the stability of the virtual acoustic image decreases at larger filter update movement increments. The interquartile range abruptly increases between 3 cm and 4 cm and again between the 5 cm and 6 cm filter update increments. Except for the stationary references trials, 3 cm appears to create the most stable virtual acoustic images of the six (6) filter update movement increments examined. ANOVA significance tests show that the filter update movement increment mean responses are significantly different with a confidence level of nearly 100% (Table B.12). There is very little probability that the differences in mean responses from the seven different filter update movement increments are due to chance. The filter update movement increment definitely affects the stability of the virtual acoustic image. Interestingly the stability of the image does not decrease below 50% of the time for the greatest filter update movement increment of 7 cm (Fig. 6.27). From Figs. 6.24 and 6.25 and Table B.10 it is known that the filter update movement increment did not significantly affect the perceived location of the sound image. It could be expected that had the stability of the image decreased below 50% that the perceived location of the image would start to significantly vary.

6.4. Conclusions

The primary objective of the experiment was determination of a filter update movement increment that produces the subjective impression of a stable virtual acoustic image as movement occurs. The results show that increments less than 3 cm achieve the most stable virtual acoustic image of the examined increments. Above this filter update movement increment, there is a steady deterioration of the stability of the virtual sound image. The different increments did not significantly affect the location of the virtual acoustic image especially for the speech source signal. Subjects tended to hear band-limited white noise from a broader spatial extent and more to their side than the unfiltered speech stimulus. Differences in the perceptions of stability by the subjects when using speech or white noise as a source signal were not significant for filter update movement increments below 5 cm. Virtual acoustic image locations far from the loudspeakers require a smaller filter update movement increment than virtual acoustic image locations close to the loudspeakers. The relative position of the system's loudspeakers to the virtual image location is an influential parameter on the required filter update movement increment that produces a stable virtual sound image. Movements traversing greater ranges require more unique virtual acoustic imaging filters and have more difficulty maintaining a stable virtual sound image.

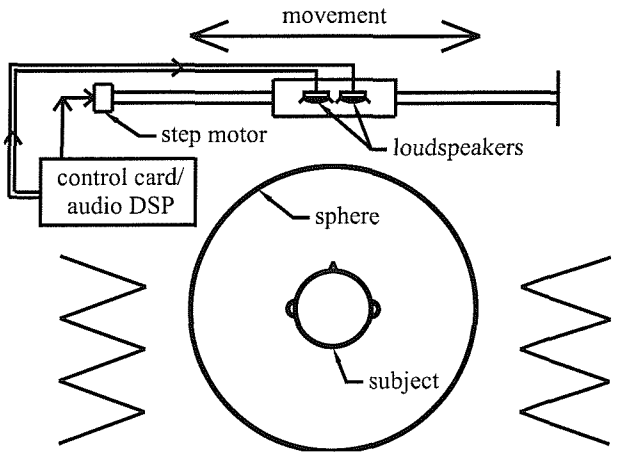


Fig. 6.1 *Plan view of experimental set up.*

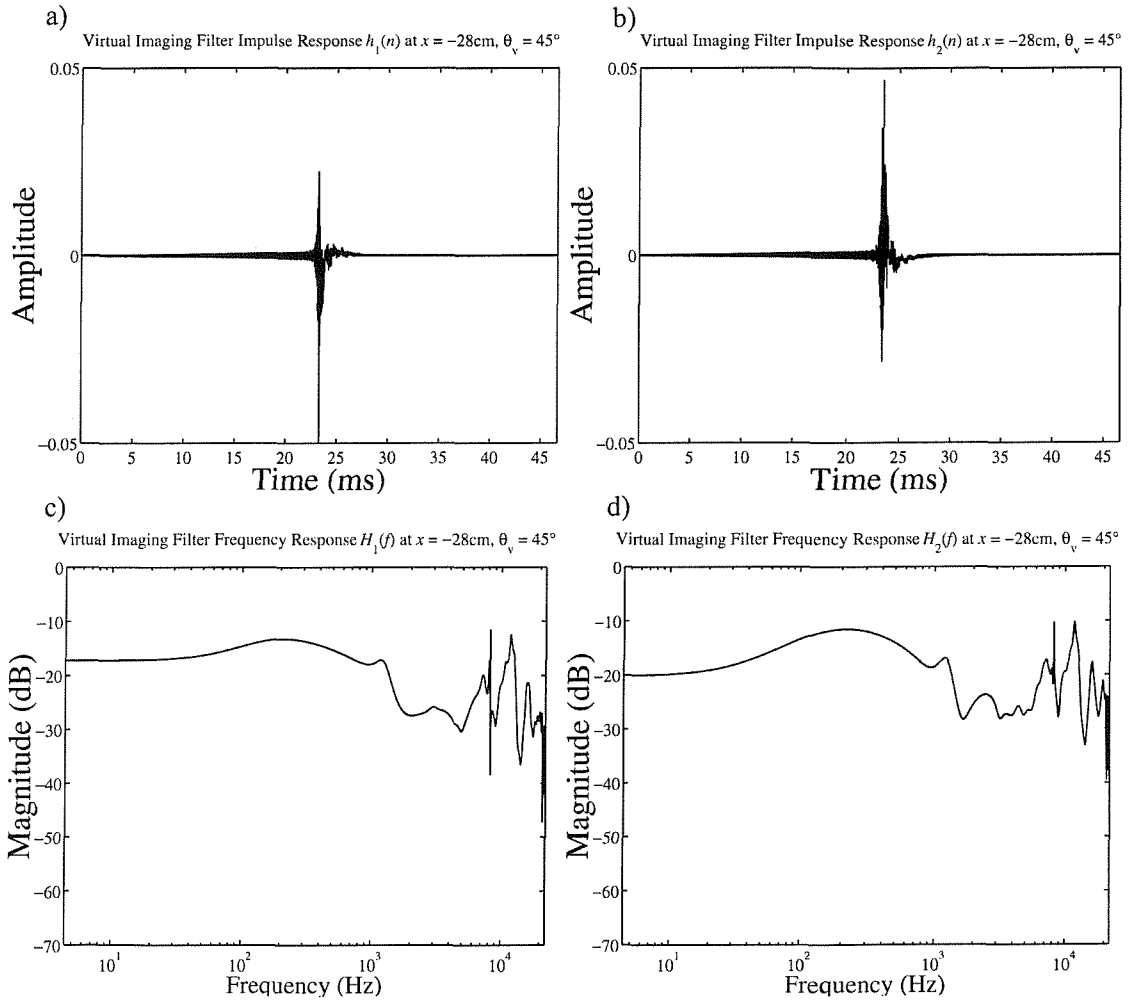


Fig. 6.2 *Virtual acoustic imaging filters for the off-axis asymmetric location at x = -28 cm with the virtual acoustic image angle θ_v 45° to the listener's front right. The design is based on KEMAR dummy HRTFs with the regularisation parameter β equal to 10^{-4}.*

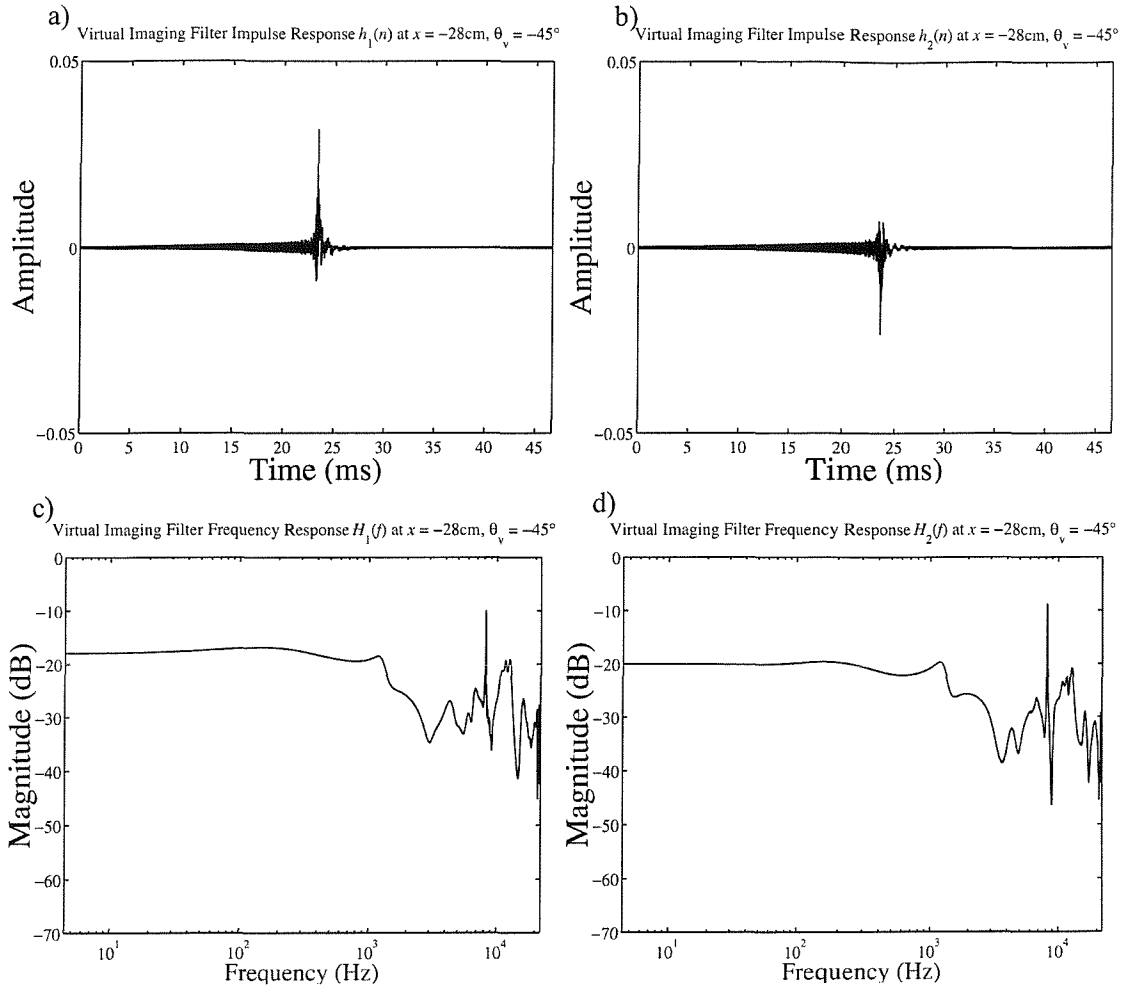


Fig. 6.3 *Virtual acoustic imaging filters for the off-axis asymmetric location at x = -38 cm with the virtual acoustic image angle θ_v 45° to the listener's front left. The design is based on KEMAR dummy HRTFs with the regularisation parameter β equal to 10^{-4}.*

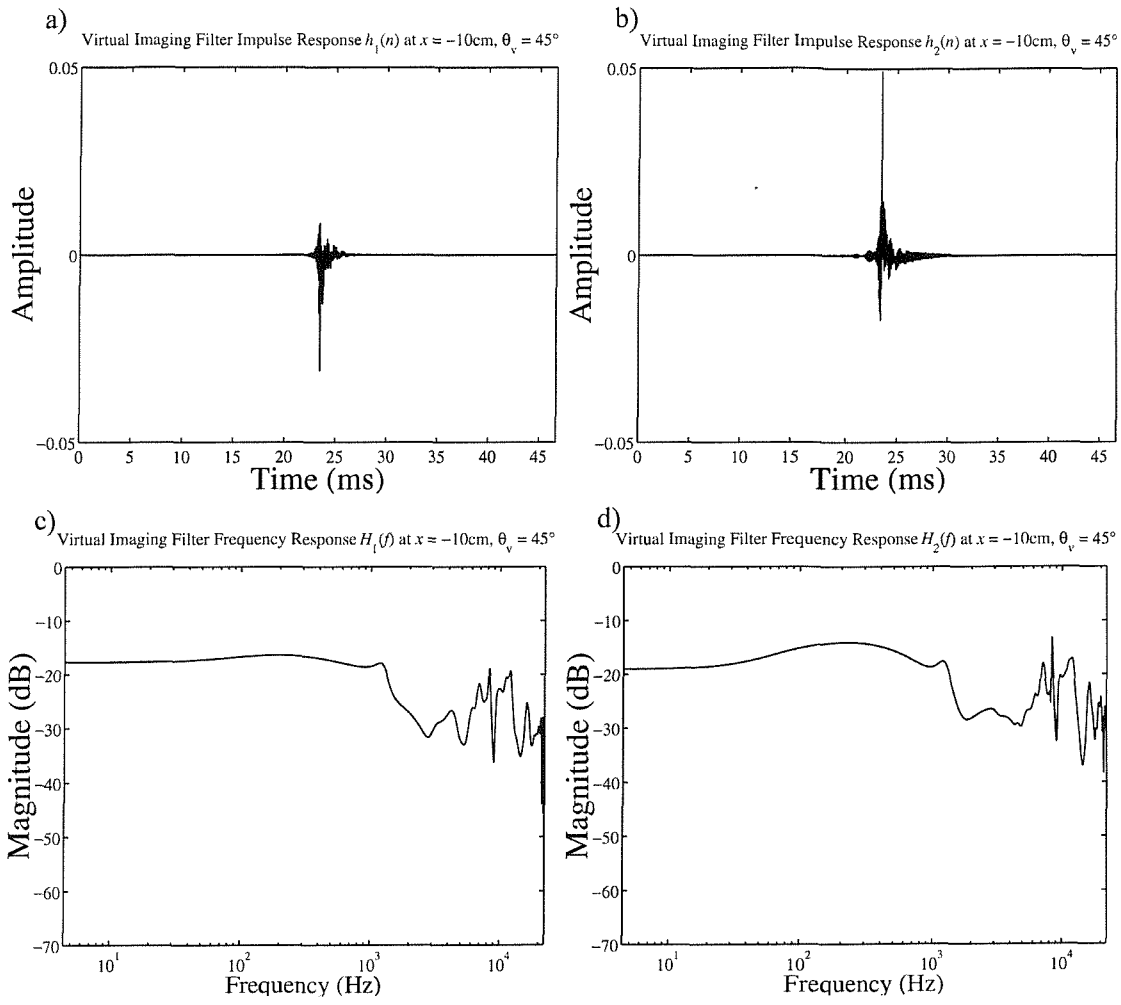


Fig. 6.4 *Virtual acoustic imaging filters for the off-axis asymmetric location at x = -10 cm with the virtual acoustic image angle θ_v 45° to the listener's front right. The design is based on KEMAR dummy HRTFs with the regularisation parameter β equal to 10^{-4}.*

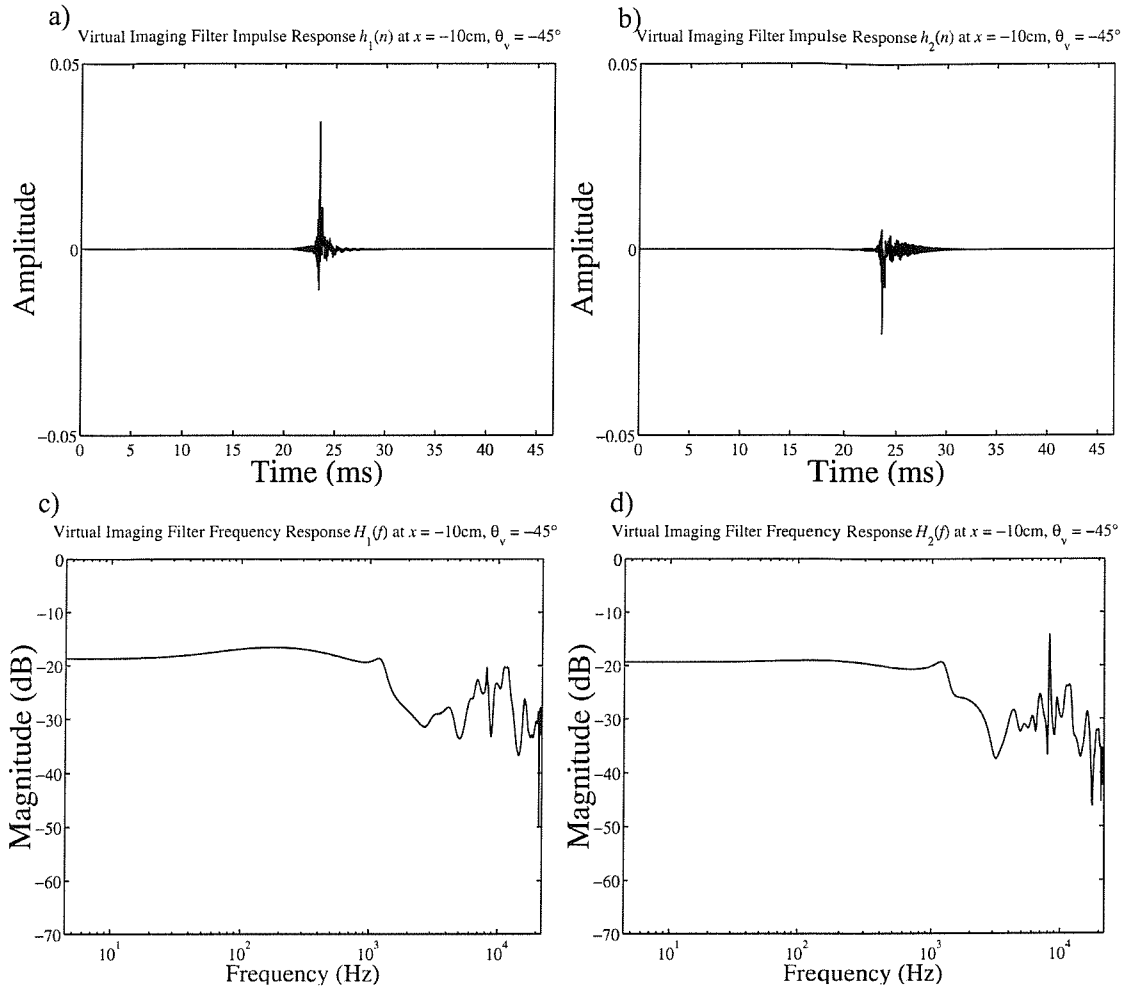


Fig. 6.5 *Virtual acoustic imaging filters for the off-axis asymmetric location at x = -10 cm with the virtual acoustic image angle θ_v 45° to the listener's front left. The design is based on KEMAR dummy HRTFs with the regularisation parameter β equal to 10^{-4}.*

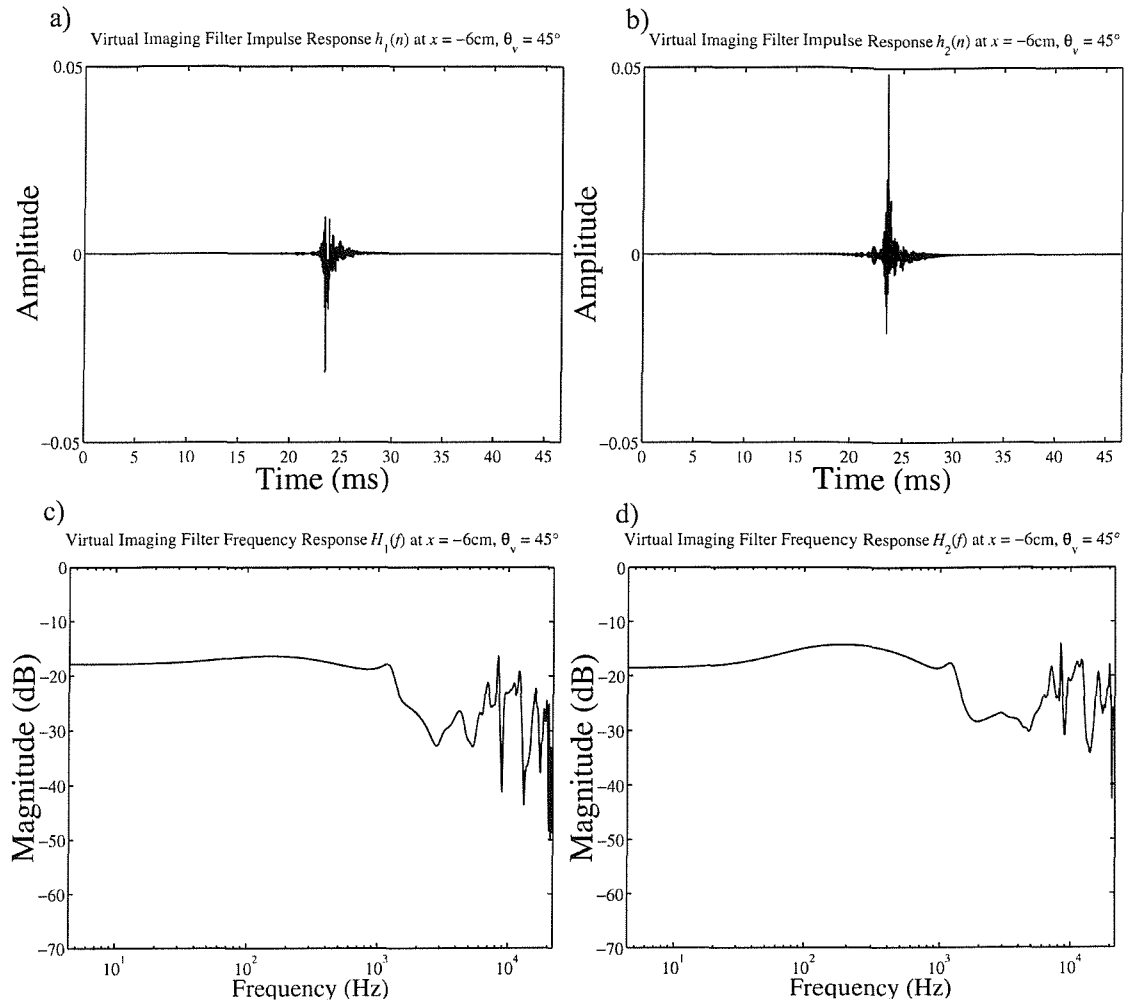


Fig. 6.6 *Virtual acoustic imaging filters for the off-axis asymmetric location at x = -6 cm with the virtual acoustic image angle θ_v 45° to the listener's front right. The design is based on KEMAR dummy HRTFs with the regularisation parameter β equal to 10^{-4}.*

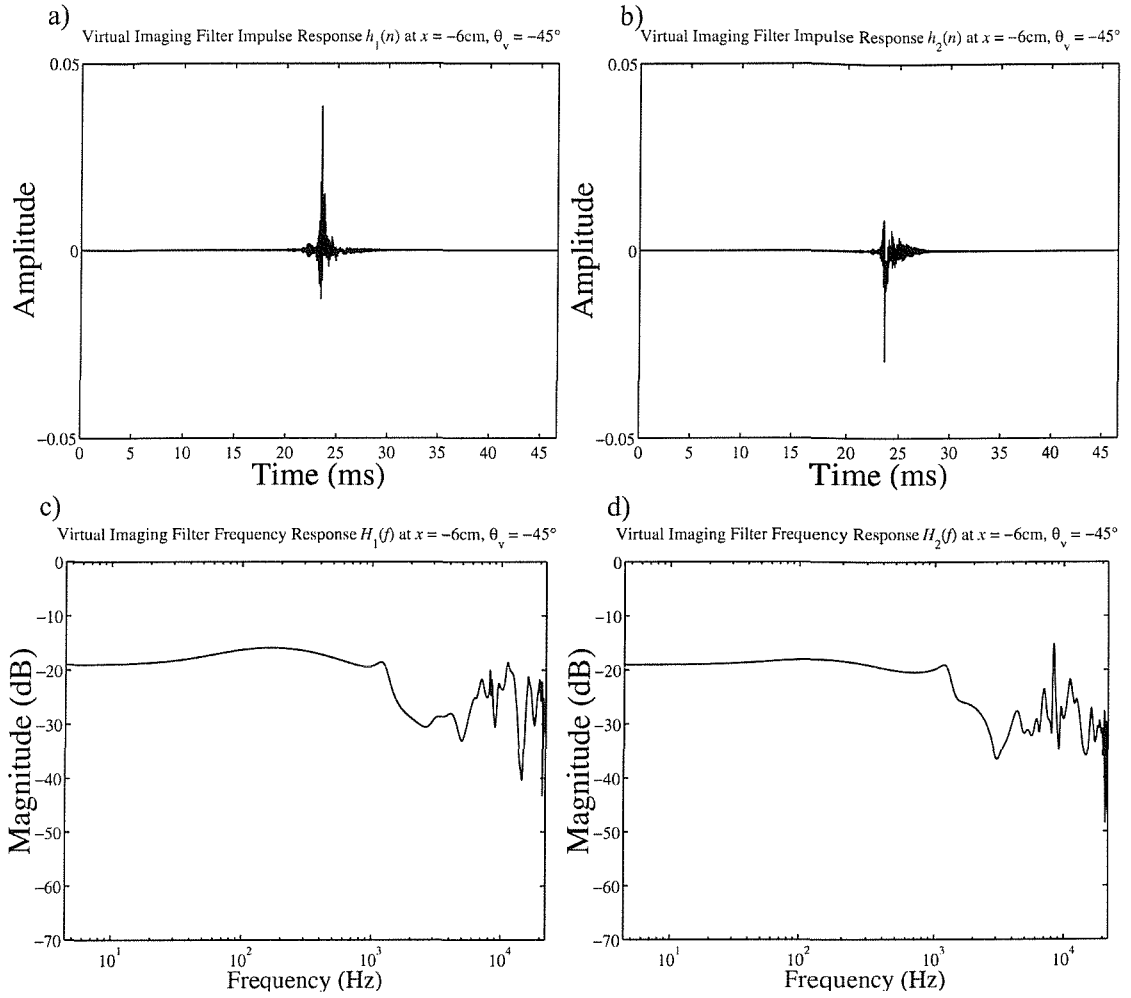


Fig. 6.7 *Virtual acoustic imaging filters for the off-axis asymmetric location at x = -6 cm with the virtual acoustic image angle θ_v 45° to the listener's front left. The design is based on KEMAR dummy HRTFs with the regularisation parameter β equal to 10^{-4}.*

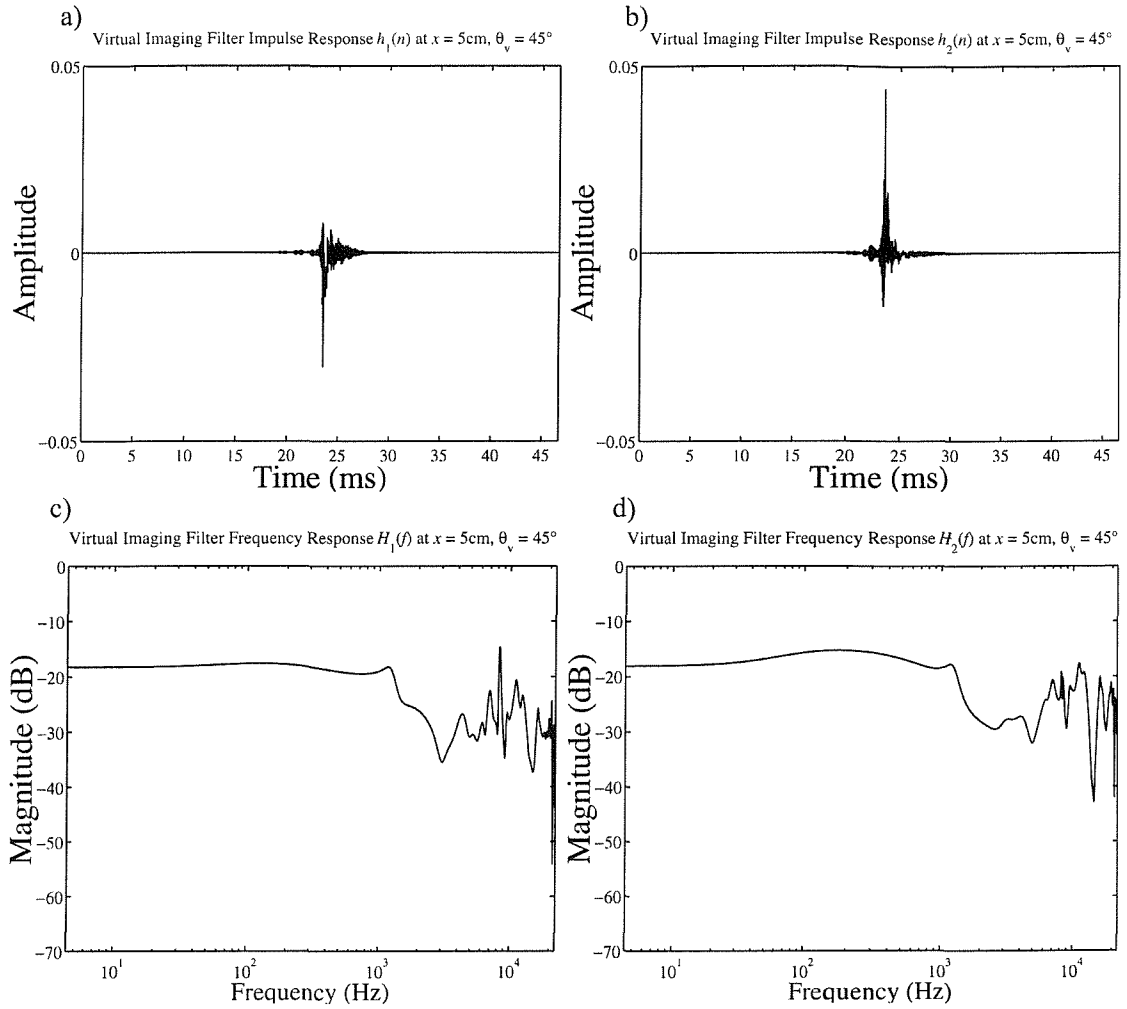


Fig. 6.8 *Virtual acoustic imaging filters for the off-axis asymmetric location at x = 5 cm with the virtual acoustic image angle θ_v 45° to the listener's front right. The design is based on KEMAR dummy HRTFs with the regularisation parameter β equal to 10^{-4}.*

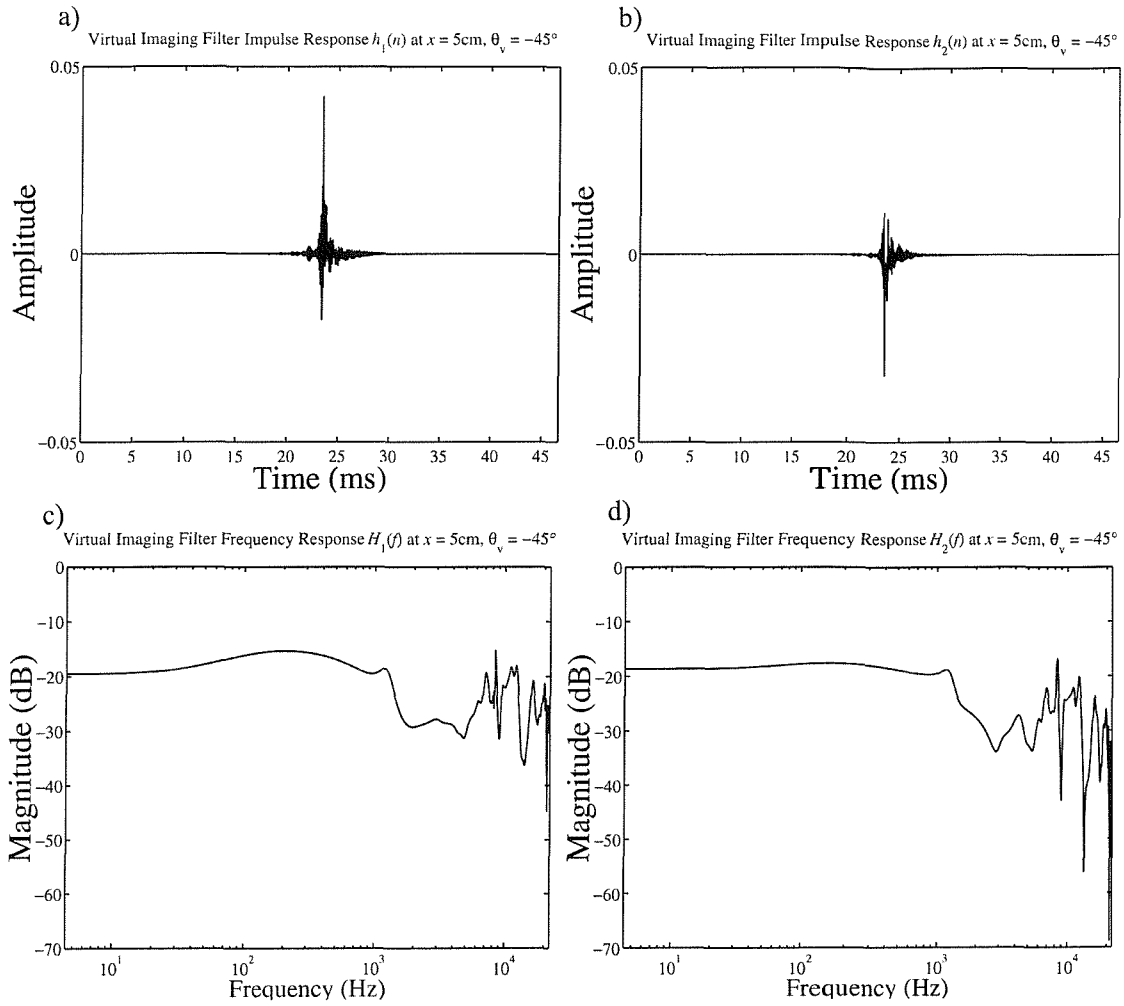


Fig. 6.9 *Virtual acoustic imaging filters for the off-axis asymmetric location at x = 5 cm with the virtual acoustic image angle θ_v 45° to the listener's front left. The design is based on KEMAR dummy HRTFs with the regularisation parameter β equal to 10^{-4}.*

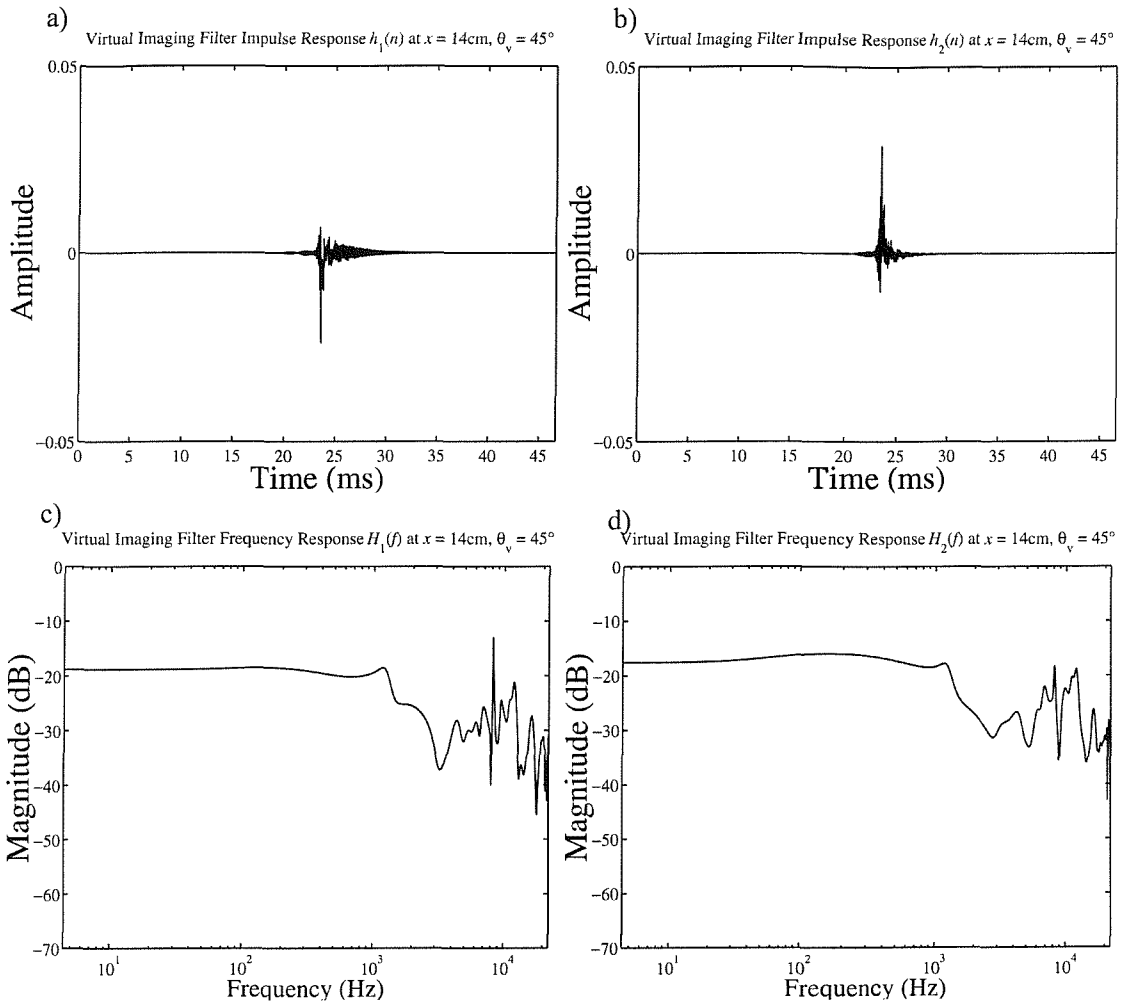


Fig. 6.10 *Virtual acoustic imaging filters for the off-axis asymmetric location at x = 14 cm with the virtual acoustic image angle θ_v 45° to the listener's front right. The design is based on KEMAR dummy HRTFs with the regularisation parameter β equal to 10^{-1}.*

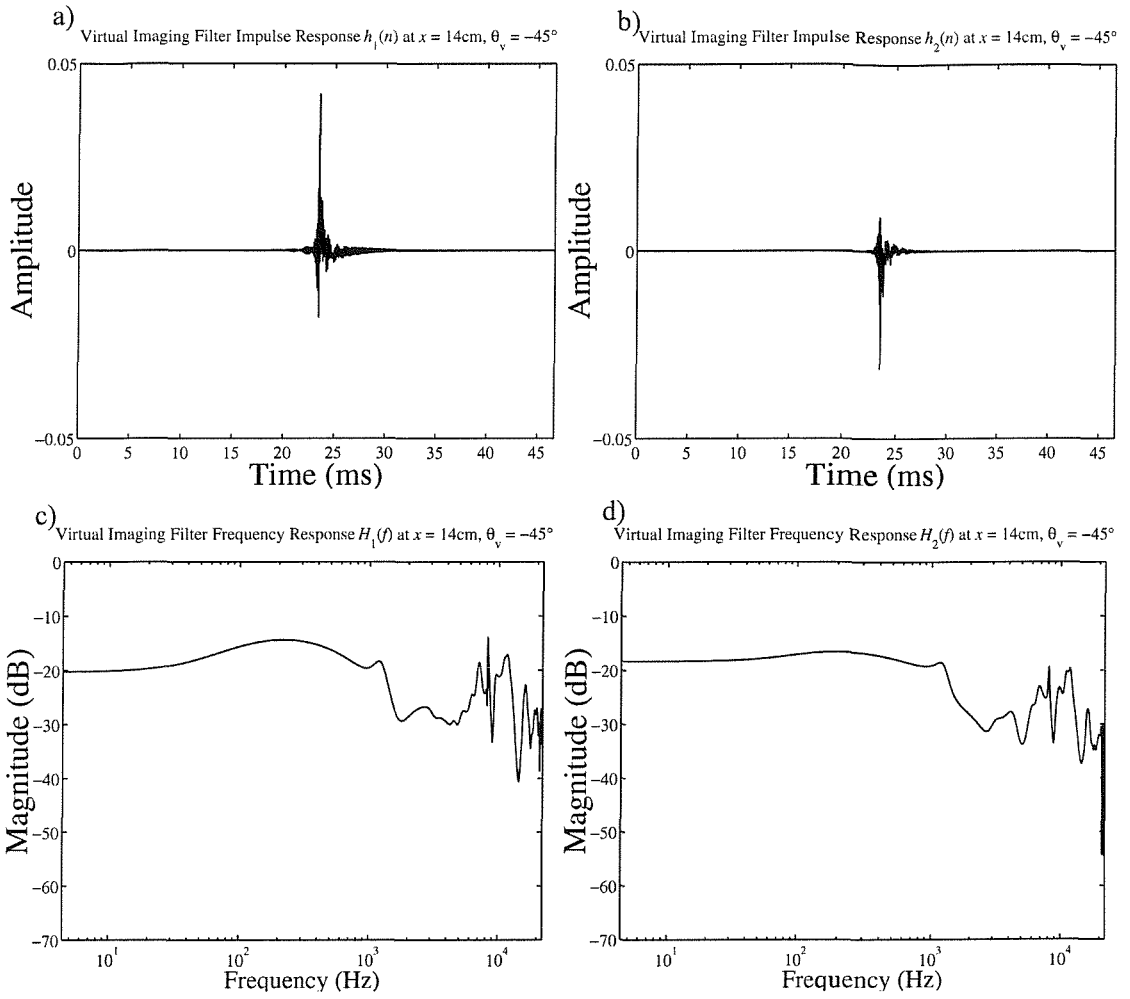


Fig. 6.11 *Virtual acoustic imaging filters for the off-axis asymmetric location at x = 14 cm with the virtual acoustic image angle θ_v 45° to the listener's front left. The design is based on KEMAR dummy HRTFs with the regularisation parameter β equal to 10^{-4}.*

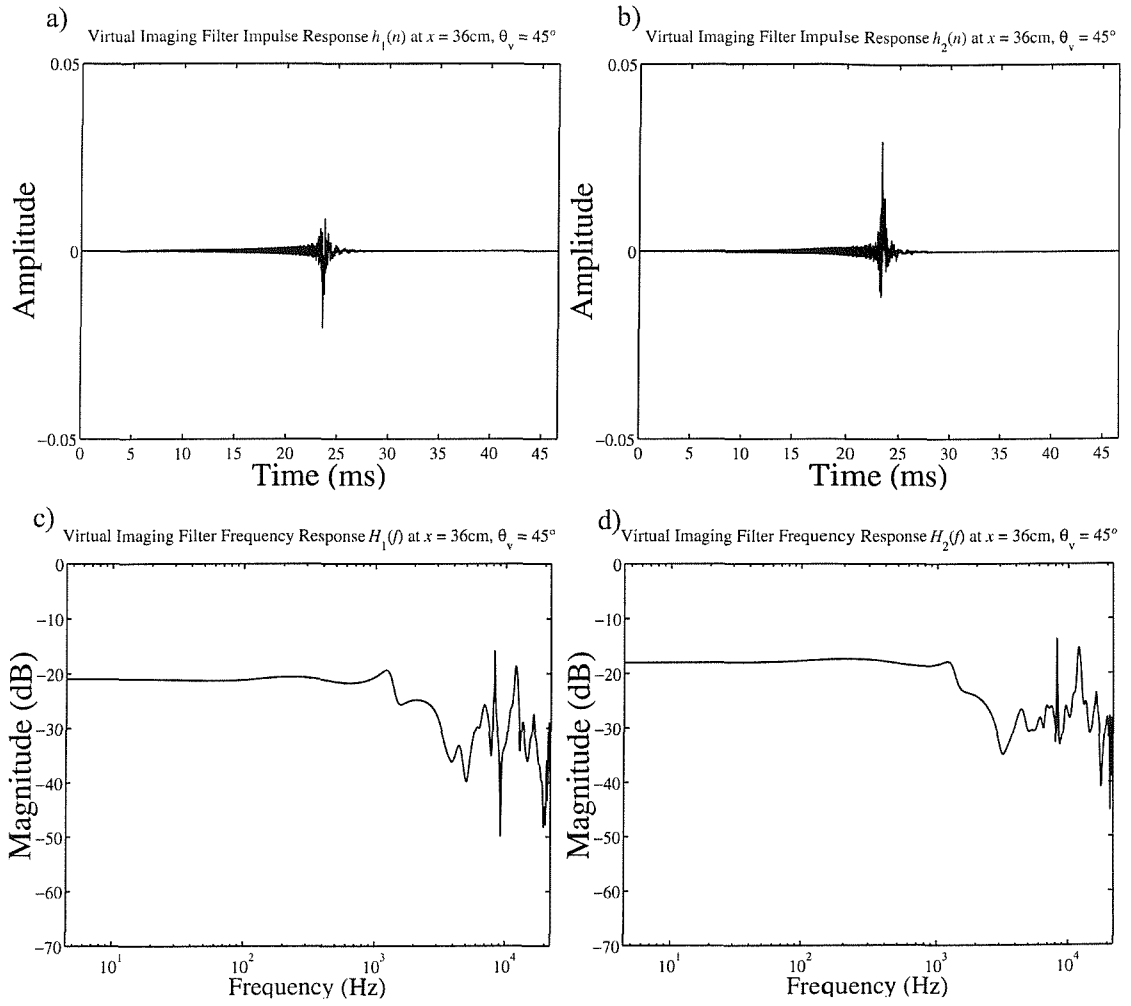


Fig. 6.12 *Virtual acoustic imaging filters for the off-axis asymmetric location at x = 36 cm with the virtual acoustic image angle θ_v 45° to the listener's front right. The design is based on KEMAR dummy HRTFs with the regularisation parameter β equal to 10^{-4}.*

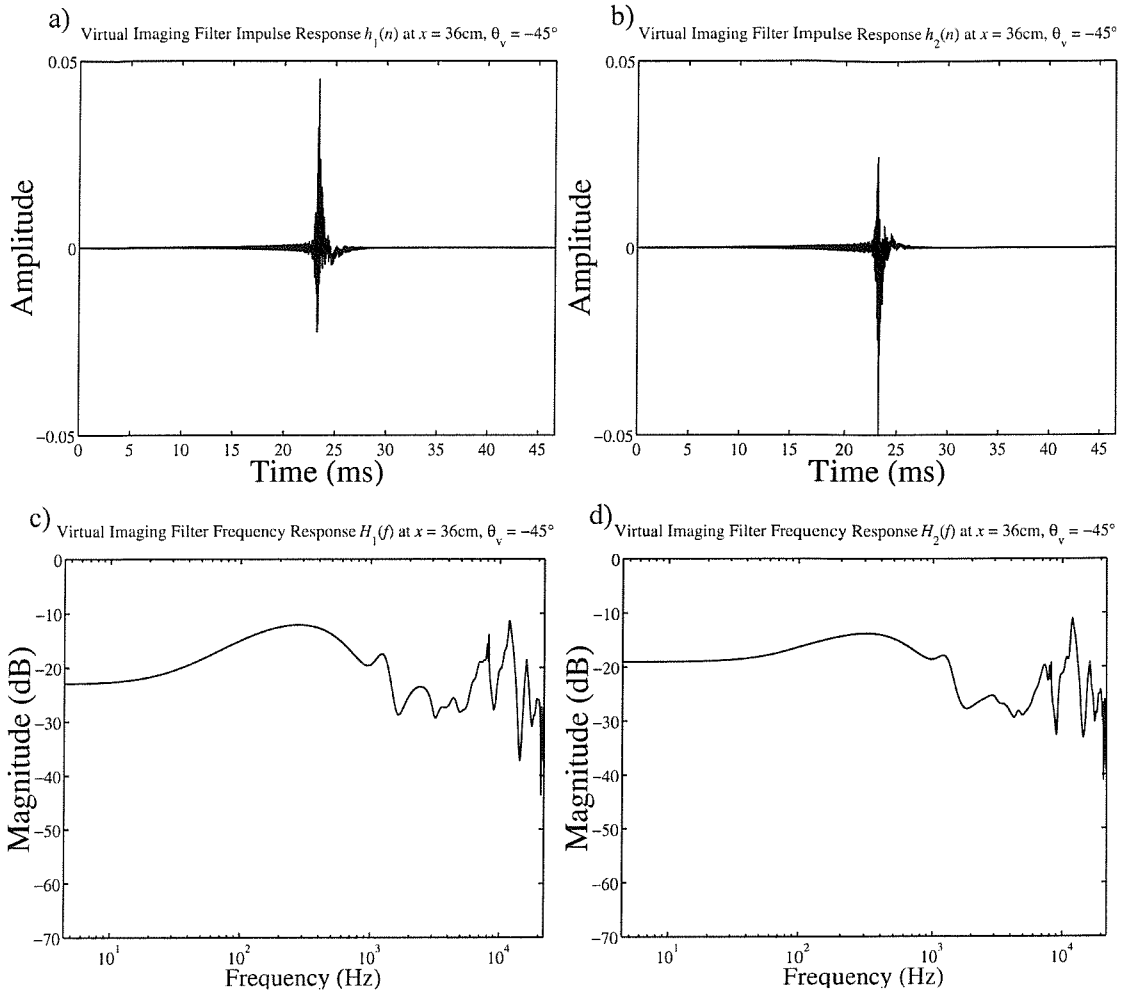


Fig. 6.13 *Virtual acoustic imaging filters for the off-axis asymmetric location at x = 36 cm with the virtual acoustic image angle θ_v 45° to the listener's front left. The design is based on KEMAR dummy HRTFs with the regularisation parameter β equal to 10^{-4}.*

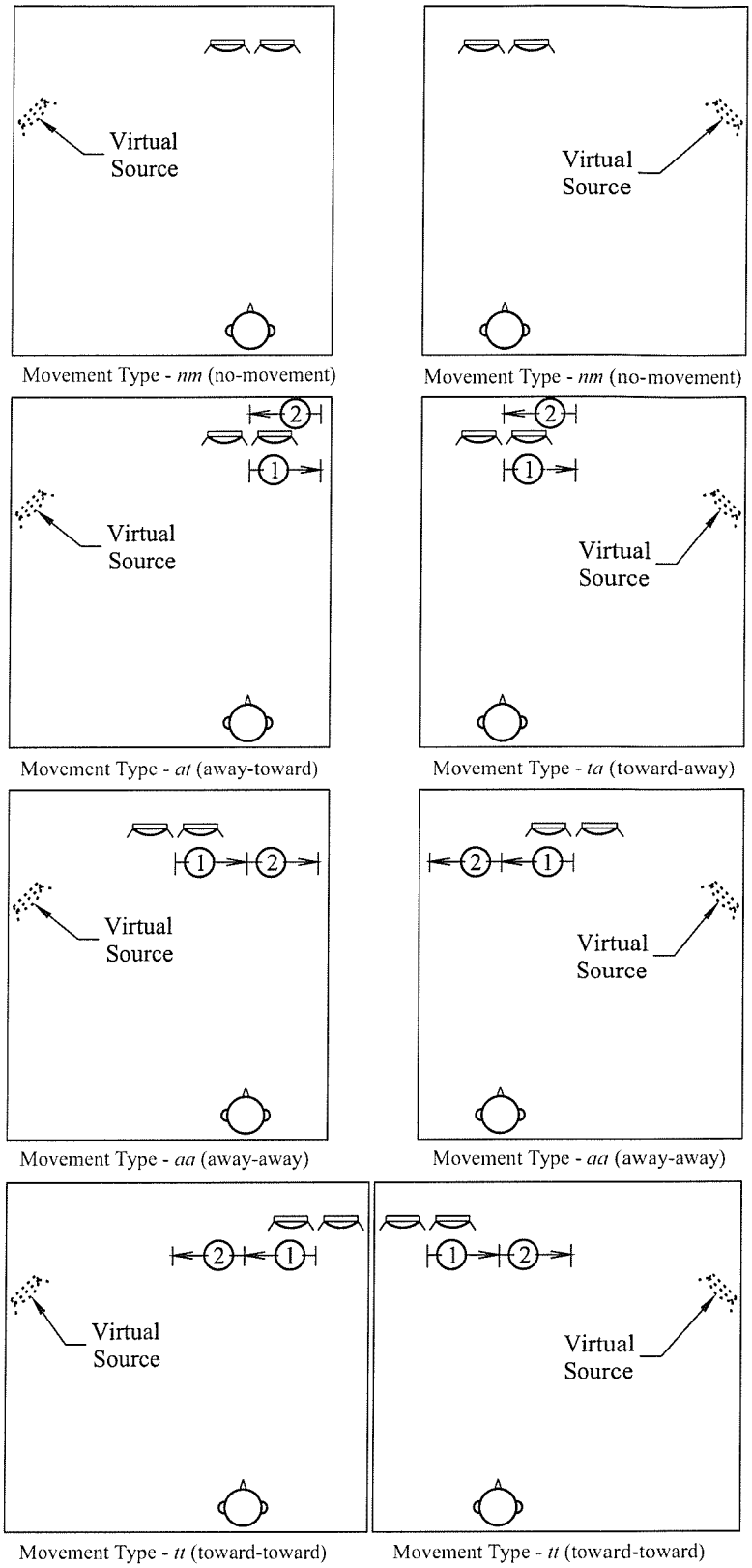


Fig. 6.14 *Depictions of the five (5) different loudspeaker path\virtual source location combinations, no-movement (nm), away-toward (at), toward-away (ta), away-away (aa), and toward-toward (tt).*

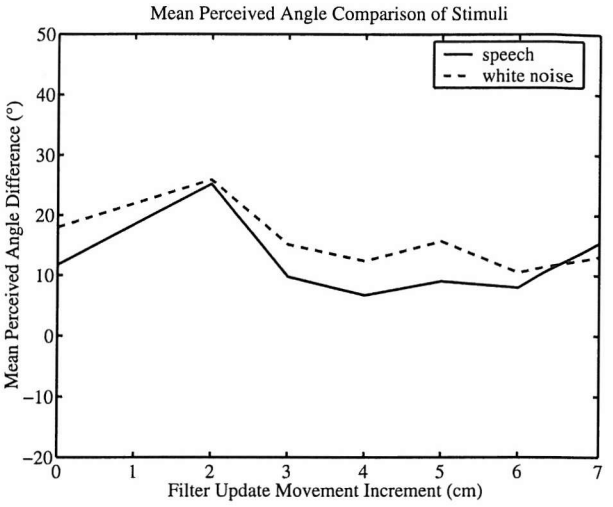


Fig. 6.15 *Mean differences in between the angle perceived and the intended virtual source location as a function of filter update movement increment for speech and white sound sources. Positive values reflect a perception of the source at locations further to the side of the subject than the intended virtual source location. Negative values reflect a perception of the source at locations closer to the front of the subject than the intended virtual source location. A value of zero reflects perceptions of the source at the intended virtual source location.*

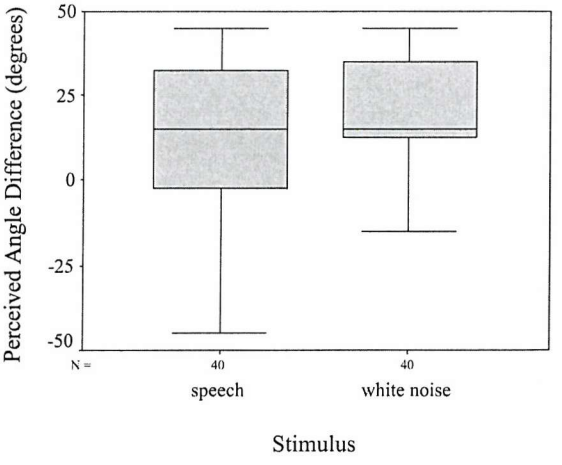


Fig. 6.16 *Box-plot comparing the difference between the angle perceived and the intended virtual source location during the stationary reference trials for the speech and white noise sources. The median for both signals is 15°. The interquartile ranges for the speech and white noise are 37.5° and 23.8° respectively. The means for speech and white noise are 11.8° and 18.0° respectively. The standard deviations for speech and white noise are 25.8° and 20.9° respectively. The size of both data-sets is 40 trials.*

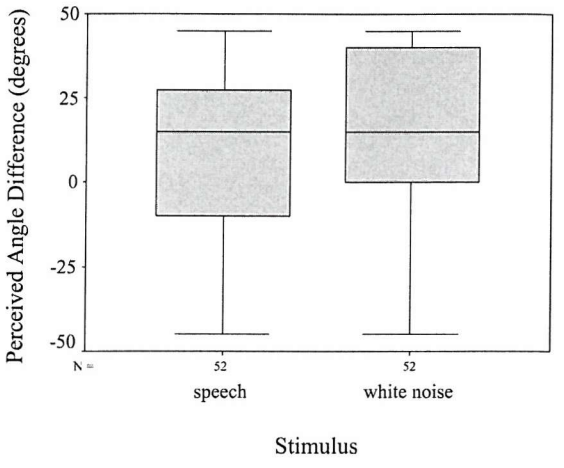


Fig. 6.17 *Box-plot comparing the difference between the angle perceived and the intended virtual source location during the 5 cm filter update movement increment trials for the speech and white noise sources. The median for both signals is 15°. The interquartile ranges for the speech and white noise are 41.3° and 40.0° respectively. The means for speech and white noise are 9.0° and 15.7° respectively. The standard deviations for speech and white noise are 24.1° and 24.5° respectively. The size of both data-sets is 52 trials.*

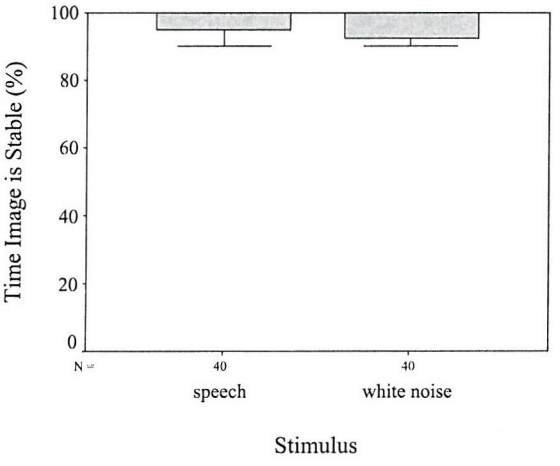


Fig. 6.18 *Box-plot comparing the perceived percentages of time the virtual sound image is stable for the speech and white noise sources during the stationary reference trials. The speech and white noise source signals both have medians of 100%, interquartile ranges of 7.5% and 8.8%, means of 96.5% and 95.4%, and standard deviations of 7.0% and 9.6% respectively. Both data-sets contain 40 trials.*

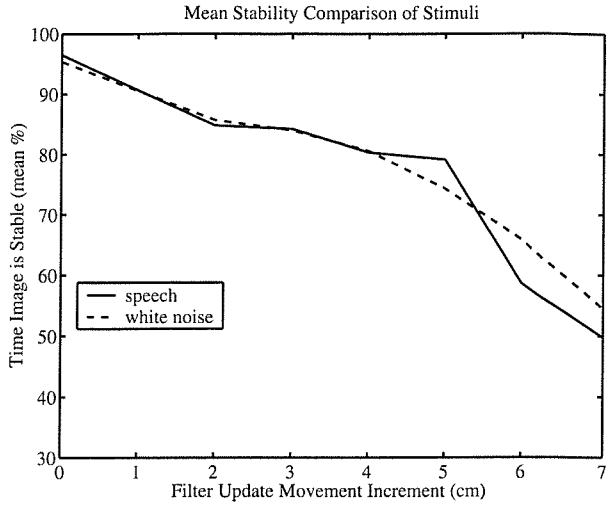


Fig. 6.19 *Mean perceived percentages of time the virtual sound image is stable as a function of filter update movement increment for speech and white sound sources.*

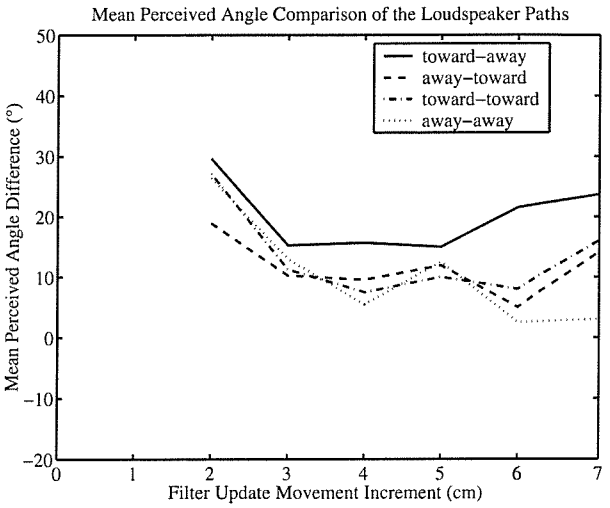


Fig. 6.20 *Mean differences between the angle perceived and the intended virtual source location as a function of filter update movement increment for the four different loudspeaker paths.*

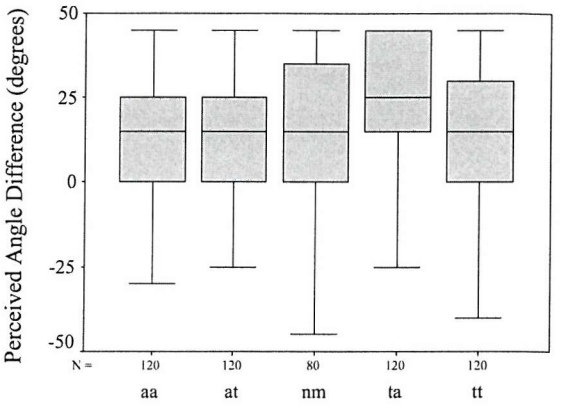


Fig. 6.21 *Box-plot comparing the difference between the angle perceived and the intended virtual source location during all trials for the different types of loudspeaker movement. The medians for all the movement types are 15° except for toward-away which is 25°. The interquartile ranges for aa, at, nm, ta, and tt are 25°, 25°, 35°, 30°, and 30° respectively. The means for aa, at, nm, ta, and tt are 9.8°, 11.2°, 14.9°, 19.3°, and 12.3° respectively. The standard deviations for aa, at, nm, ta, and tt are 22.4°, 18.4°, 23.6°, 25.0°, and 23.0° respectively. The size of all of data-sets is 120 trials except for the stationary reference trial data-set, which contains 80 trials.*

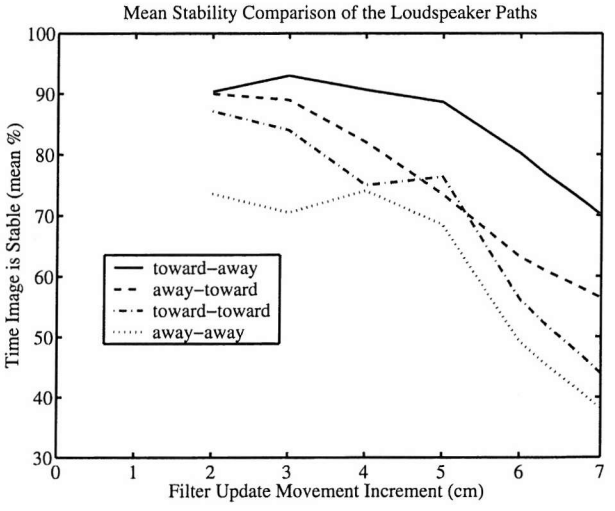


Fig. 6.22 *Mean perceived percentages of time the virtual sound image is stable as a function of filter update movement increment for the four different loudspeaker paths.*

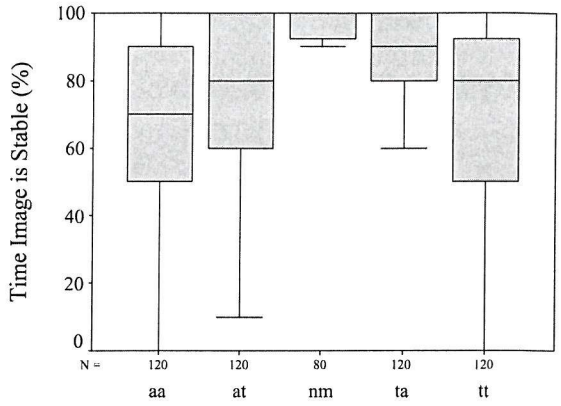


Fig. 6.23 *Box-plot comparing the perceived percentages of time the virtual sound image is stable for the different loudspeaker paths during all trials. The loudspeaker paths aa, at, nm, ta, and tt have medians of 70%, 80%, 100%, 90%, 80%, interquartile ranges of 40%, 40%, 8.8%, 20%, and 43.8%, means of 62.7%, 75.3%, 95.9%, 85.8%, and 70.4%, and standard deviations of 29.2%, 25.0%, 8.3%, 20.0%, and 28.7% respectively. The size of all of data-sets is 120 trials except for the stationary reference trial data-set, which contains 80 trials.*

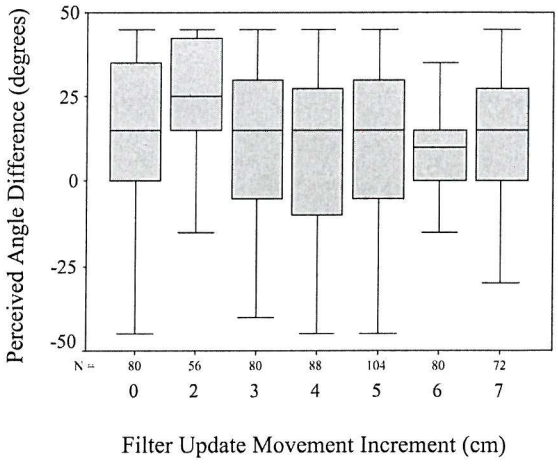


Fig. 6.24 *Box-plot comparing the difference between the angle perceived and the intended virtual source location during all trials for the different filter update movement increments. The stationary reference trial, 2, 3, 4, 5, 6, and 7 cm filter update movement increments have medians of 15°, 25°, 15°, 15°, 15°, 10°, 15°, interquartile ranges of 35°, 28.8°, 35°, 41.3°, 35°, 15°, 28.8°, means of 14.9°, 25.5°, 12.4°, 9.6°, 12.4°, 9.3°, 14.3°, and standard deviations of 23.6°, 17.3°, 22.5°, 25.1°, 24.4°, 21.1°, 20.9° with data-set sizes of 80, 56, 80, 88, 104, 80, 72 respectively.*

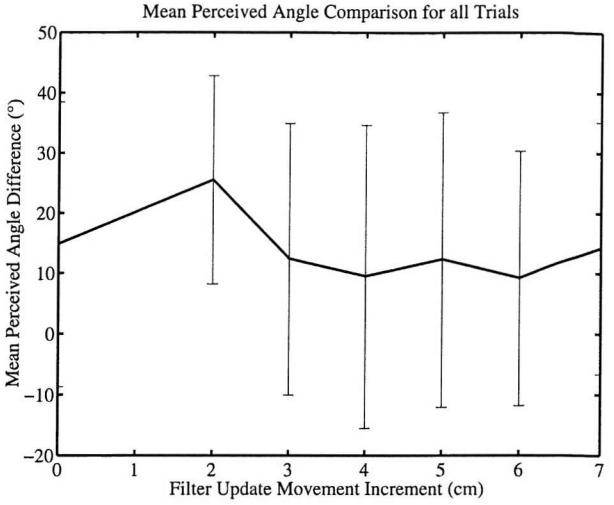


Fig. 6.25 *Mean differences between the angle perceived and the intended virtual source location as a function of filter update movement increment for all trials. Error bars show the standard deviations.*

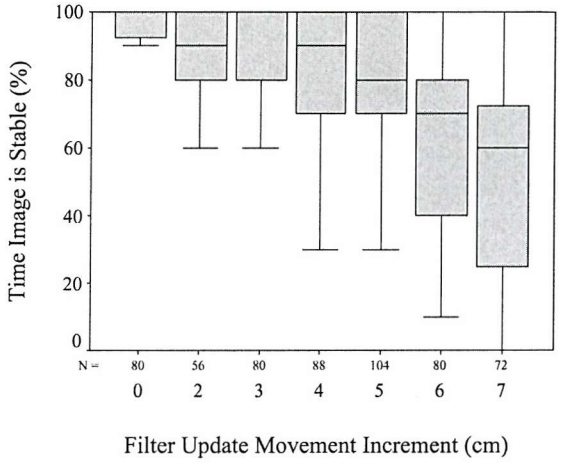


Fig. 6.26 *Box-plot comparing the perceived percentages of time the virtual sound image is stable for the different filter update movement increments during all trials. The stationary reference trial, 2, 3, 4, 5, 6, and 7 cm filter update movement increments have medians of 100%, 90%, 100%, 90%, 80%, 70%, 60%, interquartile ranges of 8.8%, 20%, 20%, 30%, 30%, 40%, 48.8%, means of 95.9%, 85.3%, 84.1%, 80.5%, 76.7%, 62.3%, and 52.1%, and standard deviations of 8.3%, 18.1%, 23.3%, 24.3%, 24.4%, 25.2%, and 30.5% with data-set sizes of 80, 56, 80, 88, 104, 80, 72 respectively.*

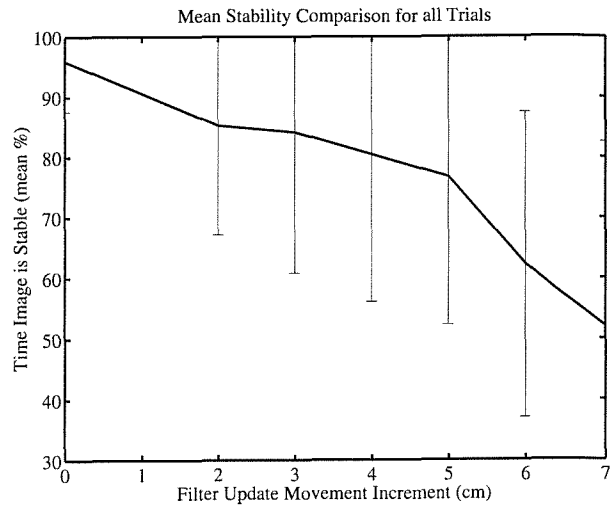


Fig. 6.27 *Mean perceived percentages of time the virtual sound image is stable as a function of filter update movement increment for all trials. Error bars show the standard deviations.*

7. VIDEO HEAD TRACKING

7.1. Introduction

Chapter 6 described a dynamic subjective evaluation of an adaptive virtual acoustic imaging system where the filters were updated by knowing the programmed motion of the loudspeakers. The motion of the loudspeakers was programmed before beginning the experimental trials and the time appropriate to switch the filters was predetermined from the knowledge of where the loudspeakers position would be as time progressed. Generally, an adaptive virtual audio system will have to change filters for undetermined arbitrary listener motions. It is proposed that an unobtrusive way in which to determine the listener's position as time progresses is to utilise a video camera and an image processing head tracking algorithm. In this chapter, a head-tracking algorithm is described and the dynamic subjective experiment described in chapter 6 is repeated but now utilising the adaptive virtual sound imaging system modified to incorporate the video head tracking procedure. The video head tracking is seen to add a lag in the system that significantly affects the subjects' impressions of the stability of a virtual acoustic image.

7.2. Image Processing

The head tracking approach technique taken is to template match with pattern search optimisation. These techniques are described in detail in Rose [5] but explanations are also provided in this section in the interest of completion. This approach may be considered simple or intuitively obvious. The more advanced head tracking approaches briefly discussed in this section give us a glimpse of what potentially could be achieved for more general situations. The increasing use of the Internet has intensified the production of many new products such as personal computer (PC) video cameras. These video cameras are used widely for video conferencing and creating personal movies. They are relatively cheap and simple to use on newer PCs. The prevalence of these types of video cameras and their ease of use with PCs were the primary reasons that this type of video camera was chosen for this thesis.

7.2.1. Head tracking methods

Video pictures, photographs, or even many paintings are two-dimensional representations of the three-dimensional world. In fact, the three-dimensional world is also projected onto the human retina. Cameras project three-dimensional scenes onto a two-dimensional image plane. The image plane is distributed with intensity values that together represent the visual scene. Optical flow is a two-dimensional velocity field defined over the image plane [70] also referred to as visual motion, image flow, or the image velocity field. Optical flow depends on the geometry of the scene and the motion of surfaces in the scene. Determination of optical flow usually involves calculating the time and spatial derivatives of the intensity values in the image. Recovering the three-dimensional scene from optical flow is an inherently ill-posed or underdetermined problem. Humans perceive shape and motion in three-dimensions with help from two sensors (eyes) and past experiences. A sophisticated approach is required to solve this problem computationally and usually assumptions are made about the scene such as the rigidity of the objects.

A head-tracking scheme is presented in reference [71], which is capable of tracking the six (6) rigid motions of a head (three (3) translational and three (3) rotational). A three-dimensional ellipsoidal is automatically fitted to the head image in the first image frame, with the person looking directly at the camera. The ellipsoidal is used as a model for the head shape. The surface of the ellipsoid is sampled, thus creating an array of 3D points. Optical flow is computed, and the rigid motion of the 3D head model that best accounts for the observed flow is interpreted as the motion of the head. The mapping of the optical flow to 3D motion of the ellipsoidal model is done by finding a local optimum with the "simplex" gradient descent technique. Head tracking with this approach was performed for a variety of head shapes, over several frames, and in the presence of background movements.

Optical flow methods (as in reference [70]) were described in reference [72] as requiring massive computational power. A more computationally efficient method is presented in reference [72] that will track rigid head motions. The user chooses the points on the head that are to be tracked by clicking the screen image. These points

within the two-dimensional image are mapped to a three-dimensional head model. The points' co-ordinates in three dimensions are estimated through an iterative procedure consisting of two stages. The first stage is to estimate the depth of the chosen points. The second stage is to estimate the motion parameters by minimising a squared error function with the singular value decomposition technique. These two stages are repeated until the solution converges. This procedure is implemented under the situation of two image frames only and the error accumulates in subsequent frames. Problems also occur in the cases when the tracked points are obscured from the camera's view by an object (including hands) or if the points move out of the sight of the camera. A relatively small number of points can be chosen so that the computation is not too great and the algorithm does run in real time.

In comparing the head-tracking methods presented in references [71] and [72], it is evident that there exists a trade-off between the robustness of a particular algorithm and its computational requirements. Head tracking methods such as those described in [3,71-73] are very sophisticated. The work in this thesis focused on only lateral head movements, where knowledge of the two-dimensional scene is sufficient. Head rotations are another common type of movement, which do affect performance. These movements are more difficult to track and require a more sophisticated tracking algorithm than the one taken here.

7.2.2. Template matching

The head-tracking method utilised in this thesis was template matching. Discrete images consist of individual picture elements or pixels. Colour images are often represented in RGB (red-green-blue) format with pixels that each have three (3) numerical values. The three values signify the amount of red, green, and blue in each pixel. These colour images can be considered three-dimensional matrices with one dimension equal to three (3). Combining the three (3) values at each pixel in the colour image creates a grey scale image. In grey scale images, each pixel is assigned an intensity value. For example, intensity values of zero (0) and one (1) might correspond to completely black and perfectly white pixels respectively. Shades of grey are attributed intensity values between the extreme values given to black and

white pixels. These values give rise to two-dimensional matrices representing images that we may manipulate mathematically. The approach described below uses grey scale images.

Reference [74] presents template matching as it applies to object detection. The idea of template matching can be explained with the help of Figures 7.1 and 7.2. $\mathbf{M}_i$ represents the ith matrix of intensity values of an image that is the ith image of a video sequence. A template **T** (Fig. 7.2) is found by cropping out some identifiable features on the moving object in the first image of the sequence $\mathbf{M}_{i=1}$ (Fig. 7.1a). This is done in practice by positioning the person to track so that their face is centred in the video image. Once the person is in place, a rectangular image centred on the image of a smaller height and width of the original image is taken to be the template **T**. The template **T** is just an image of some features of interest with smaller dimensions (R by S) than the full images $\mathbf{M}_i$ (P by Q, where $R \leq P$ and $Q \leq S$). **T** is extracted from the first image $\mathbf{M}_{i=1}$. As new images are acquired (e.g. i = 2, 3, 4, ...), the template **T** is shifted over the new image and compared to different regions of the new image (e.g. Fig. 7.1b). When the position where the template best matches the new image is found then this position is regarded as the new position of the object.

The position of the template **T** is described by a single point called the template base point and denoted (x_t, y_t). The template base point is initially positioned on the first image (i.e. $\mathbf{M}_{i=1}$) at the centre of the image, i.e.

$$x_t\big|_{i=1} = \begin{cases} \dfrac{P+1}{2} & \text{for } P \text{ odd} \\ \dfrac{P}{2} & \text{for } P \text{ even} \end{cases} \qquad (7.1a)$$

$$y_t\big|_{i=1} = \begin{cases} \dfrac{Q+1}{2} & \text{for } Q \text{ odd} \\ \dfrac{Q}{2} & \text{for } Q \text{ even} \end{cases} . \qquad (7.1b)$$

The template **T** is then taken to be a rectangular image centred on the base point, i.e.,

$$\mathbf{T} = \mathbf{M}_{i=1}\left(\mathbf{x}_t\big|_{x_t}, \mathbf{y}_t\big|_{y_t}\right) \tag{7.2}$$

where the template position vectors $\mathbf{x}_t\big|_{x_t}$ and $\mathbf{y}_t\big|_{y_t}$ are

$$\mathbf{x}_t\big|_{x_t} = \begin{cases} x_t - \dfrac{R+1}{2}+1,\ x_t - \dfrac{R+1}{2}+2,\ \cdots,\ x_t + \dfrac{R+1}{2} & \text{for } R \text{ odd} \\ x_t - \dfrac{R}{2}+1,\ x_t - \dfrac{R}{2}+2,\ \cdots,\ x_t + \dfrac{R}{2} & \text{for } R \text{ even} \end{cases} \tag{7.3a}$$

$$\mathbf{y}_t\big|_{y_t} = \begin{cases} y_t - \dfrac{S+1}{2}+1,\ y_t - \dfrac{S+1}{2}+2,\ \cdots,\ y_t + \dfrac{S+1}{2} & \text{for } S \text{ odd} \\ y_t - \dfrac{S}{2}+1,\ y_t - \dfrac{S}{2}+2,\ \cdots,\ y_t + \dfrac{S}{2} & \text{for } S \text{ even} \end{cases}, \tag{7.3b}$$

and the template base point (x_t, y_t) is defined in Eq. (7.1).

The template **T** never changes for an entire image sequence after first extracting the template **T** from the first image of the sequence $\mathbf{M}_{i=1}$. The template base point can change over a single image and over subsequent images. The template base point (x_t, y_t) moves over an image $\mathbf{M}_i$ so that a cost function is minimised. For this thesis, the best match was found by comparing a cost function J of squared error, i.e.

$$J = \sum_{r=1}^{R}\sum_{s=1}^{S} e_{rs}^2 \tag{7.4}$$

where

$$e_{rs} = \mathbf{M}_i[\mathbf{x}_t(r)|_{x_t}, \mathbf{y}_t(s)|_{y_t}] - \mathbf{T}(r,s), \tag{7.5}$$

and the template position vectors $\mathbf{x}_t|_{x_t}$ and $\mathbf{y}_t|_{y_t}$ are defined in Eq. (7.3) but with a varying template base point (x_t, y_t). Note that the values of the matrix $\mathbf{M}_i(\mathbf{x}_t|_{x_t}, \mathbf{y}_t|_{y_t})$ depend on the choice of template centre position (x_t, y_t) and that the values in the template matrix **T** do not depend on the template position or the image sequence number *i*. The problem is then seen to become a least squares minimisation problem, for each *i* the function to be minimised is

$$J(x_t, y_t)|_i = \sum_{r=1}^{R}\sum_{s=1}^{S}\left\{\mathbf{M}_i\left[\mathbf{x}_t(r)|_{x_t}, \mathbf{y}_t(s)|_{y_t}\right] - \mathbf{T}(r,s)\right\}^2. \tag{7.6}$$

Over the course of this work, a number of different methods were considered that would seek minimal solutions to this function. These methods include variations on the method of steepest descent and considering the problem as a feedback system with integral control. Of the approaches tried, the most success was achieved with the pattern search optimisation approach.

Template matching does have the disadvantage of working well with only plain backgrounds that provide a contrast between the background and the person's head. Templates consisting of an image of the person's head do not match well with the background alone. The plain background was introduced in this experiment as a means of simplifying the task at hand but is not present in generalised situations. For example, if the background was a crowd of people the template might match well to other faces in the image. Using template matching to track one person under these circumstances would obviously be much more difficult.

7.2.3. Pattern search

Pattern search was first introduced by R. Hooke and T. A. Jeeves as one of their "direct search" strategies [75]. A review of the technique with an illustrated example can be found in reference [76]. The pattern search optimisation routine is an alternative to the classical steepest descent algorithm and is well suited for implementation on computers due to the simple logic and repeated arithmetic operations that are involved. The pattern search can be generalised for an arbitrary number of dimensions but will be discussed here with only two degrees of freedom as is applicable to our case of two-dimensional images. Used in conjunction with template matching, the pattern search can be envisioned as a series of template movements across the current video image. The sizes and directions of the moves are dictated by the pattern search algorithm and the moves will hopefully stop when the template is best fitted to the image. This example is also provided in Rose [5].

Fig. 7.3 shows flow charts of the pattern matching algorithm. There are two different types of moves made in the pattern search routine (Fig. 7.3a). The algorithm starts with a set of exploratory moves (Fig. 7.3b). Exploratory moves seek to find information around the immediate area of the current position of the template. A move is a success if the cost function J, at the position after the move, is less than the cost function before the move. Before a set of exploratory moves is made, the cost at the current template position (the base point $x_t\mathbf{i} + y_t\mathbf{j} = \mathbf{b}_1$) is evaluated. The set of exploratory moves then begin with a move in the positive x-direction. The size of this move is called the exploratory step size in the x-direction Δ_x. The cost is then evaluated at this new position and compared to the cost at the base point $\mathbf{b}_1$. If the cost is less at the new position then the new position becomes the temporary new base point of the template $\mathbf{b}_2$. If the cost at this position is greater than that at $\mathbf{b}_1$ then the cost is evaluated in the negative x-direction. That is, the cost $J(\mathbf{b}_1 - \Delta_x\mathbf{i})$ is evaluated. If this move is successful then this position becomes $\mathbf{b}_2$. If unsuccessful, the initial base point $\mathbf{b}_1$ is retained (i.e. $\mathbf{b}_2 = \mathbf{b}_1$). The procedure is then repeated for the y-direction using $\mathbf{b}_2$ as the base point. That is $J(\mathbf{b}_2 + \Delta_y\mathbf{j})$ is evaluated and compared to $J(\mathbf{b}_2)$. If $J(\mathbf{b}_2 + \Delta_y\mathbf{j})$ is less than $J(\mathbf{b}_2)$ then $\mathbf{b}_2 + \Delta_y\mathbf{j}$ becomes the new base point $\mathbf{b}_2$. If not then $\mathbf{b}_2 - \Delta_y\mathbf{j}$ is tried. One can see that a set of exploratory moves result in a new

base point $\mathbf{b}_2$ whose position has less cost than the previous base point $\mathbf{b}_1$. However, the set of exploratory moves do have the potential of all failing in which case the new base point would be the same as the previous base point (i.e. $\mathbf{b}_2 = \mathbf{b}_1$). A flow chart of this procedure is shown in Fig. 7.3b. Note that exploratory movements in the negative direction require the calculation of more cost functions. This causes the method to find a solution more slowly when there is movement in a negative direction.

The pattern search begins with a set of exploratory moves from the first base of the algorithm (Fig. 7.3). After successful initial exploratory moves in both the x and y directions, a pattern move is made in a vector direction and length equal to the set of successful exploratory moves. Another set of exploratory moves are then made from this new point. After the successful exploratory moves are made, a pattern move is made which is equal to the last pattern move plus the last set of exploratory moves. This routine of a pattern move followed by a set of exploratory moves is repeated until a pattern move is unsuccessful. Pattern moves are based on previously successful moves and will increase in size with a series of successes, thereby increasing the speed of finding a solution. The pattern moves also have the potential to follow valleys in the error surface. Exploratory moves can be seen to continually revise the pattern. If a pattern move is unsuccessful, the algorithm will establish a new pattern from the last successful point as was done from the initial starting point.

Initially the size of the exploratory moves (Δ_x, Δ_y) is set arbitrarily. However, when establishing a new pattern, if the exploratory moves in all four (4) directions ($\pm x$ and $\pm y$) are failures then the exploratory step size will be decreased by a specified amount. For example, in the flow diagrams Fig. 7.3 the initial size of the exploratory moves was chosen to be two (2) pixels (i.e. $\Delta_x = 2$ and $\Delta_y = 2$). If the set of exploratory moves are all failures then the step size decreases to one (1) pixel. If the set of exploratory moves fails once more with this new step size then the algorithm will terminate and assume the current position to be the solution.

Figure 7.4 shows a detailed example of pattern search movements. In Fig. 7.4 the template base points are designated by $\mathbf{b}_m$ where $m = 1,2,3\cdots7$ and illustrated by solid circles. The temporary base points are designated by $\mathbf{t}_m$ where $m = 1,2,3\cdots7$. Figure 7.4 shows exploratory moves with dashed lines and pattern moves with solid arrows. Dashed circles denote successful exploratory moves, while dashed X's denote unsuccessful exploratory moves. Successful moves occur when the cost at the position after the move is less than the cost at the position before the move.

The search began at the base point $\mathbf{b}_0$ (80, 60). Two consecutive successful exploratory moves in the positive x- and y-directions resulted in the new base point $\mathbf{b}_1$ at (81, 61). Note that the initial exploratory step sizes were both set at one (1) pixel (i.e. $\Delta_x = 1$, $\Delta_y = 1$). The first pattern move, equal to $2\mathbf{b}_1 - \mathbf{b}_0$, resulted in the temporary base point $\mathbf{t}_1$ at pixel (82, 62). Two consecutive successful exploratory moves about $\mathbf{t}_1$ resulted in the new base point $\mathbf{b}_2$ at (83, 63). The pattern move from this base point, equal to $2\mathbf{b}_2 - \mathbf{b}_1$, resulted in the temporary base point $\mathbf{t}_2$ at (85, 65). Notice that the pattern moves increased in size due to the series of successes in the same direction. The exploratory moves about $\mathbf{t}_2$ were successful in the positive x-direction and the negative y-direction. Note that this set of exploratory moves required computation of more cost functions than when the exploratory moves in the positive x- and y-directions were both successful. The resulting template position was the new base point $\mathbf{b}_3$ at (86, 64). The pattern move from this base point, equal to $2\mathbf{b}_3 - \mathbf{b}_2$, resulted in the temporary base point $\mathbf{t}_3$ at (89, 65). This pattern move was again larger than the last pattern move in the x-direction but was less than the previous pattern move in the y-direction. This shows how the exploratory moves modify the pattern. The next set of exploratory moves resulted in a new template base point at pixel (90, 64) $\mathbf{b}_4$. The pattern move from $\mathbf{b}_4$, equal to $2\mathbf{b}_4 - \mathbf{b}_3$, is solely in the positive x-direction. Specifically, this pattern move was four (4) pixels in the positive x-direction to the temporary base point $\mathbf{t}_4$ at pixel (94, 64). This is due to all four (4) of the exploratory moves in the positive x-direction thus far having been successful while two (2) exploratory moves in the positive y-direction having been successful as well as two (2) exploratory moves in the negative y-direction. Exploring about $\mathbf{t}_4$ resulted in successes in the negative x- and negative y-directions and the new base point $\mathbf{b}_5$ at

pixel (93, 63). The next pattern move $2\mathbf{b}_5 - \mathbf{b}_4$ resulted in the temporary template base point $\mathbf{t}_5$ (96, 62). This point has greater cost than at $\mathbf{b}_5$. All exploratory moves about $\mathbf{t}_5$ failed and so exploratory moves were undertaken about the last successful base point, namely $\mathbf{b}_5$. In this way, the previous developing pattern was cancelled and the exploratory moves about $\mathbf{b}_5$ were in an attempt to form a new pattern. The only successful exploratory move about $\mathbf{b}_5$ was in the negative y-direction to $\mathbf{b}_6$ (93, 62). The pattern move from $\mathbf{b}_6$ was equal to $2\mathbf{b}_6 - \mathbf{b}_5$ to $\mathbf{t}_6$ at pixel (93, 61). Again, the only successful exploratory move about $\mathbf{t}_6$ was in the negative y-direction to $\mathbf{b}_7$ at pixel (93, 60). A pattern move of $2\mathbf{b}_7 - \mathbf{b}_6$ was tried to $\mathbf{t}_7$ (93, 58) but was unsuccessful. Exploratory moves about $\mathbf{b}_7$ were then attempted without any success. At this point normally, the exploratory step sizes would have been decreased but in this example the step sizes were already at the smallest possible, that is, one (1) pixel. So, the routine terminated with the solution of pixel (93, 60). The pattern search method found that the image moved thirteen (13) pixels in the positive x-direction $\mathbf{b}_7 - \mathbf{b}_0$.

Expansion of this approach to a longer sequence of images is relatively simple. After finding a solution for the second frame, a third image is grabbed and the pattern search routine is again implemented on the new image starting from the previously found solution base point. This procedure is repeated for as long as one wishes to track a person's motion. Frame grabbing occurs at irregular intervals depending on how long the pattern search routine takes to find a solution. For the adaptive virtual acoustic imaging system, further latency is added by the virtual acoustic imaging filter selection procedure after the head position has been located. This means that a different pair of virtual acoustic imaging filters could be selected after every image frame grabbed, depending on how much the person is moving. However, the extraction of the template image is done only once, immediately after the first frame is grabbed. That is, the template image is extracted from the first image and used throughout the rest of the video sequence.

7.2.4. Practical considerations

The software developed in this project, within the Matlab computing environment, does all the image processing involved in implementing the head-tracking algorithm.

There are a few steps required in order to get the images into a form that Matlab can manipulate easily.

The program, Vision for Matlab (VFM), perhaps realises the most important step. The author of this software, Farzad Pezeshkpour, provides a website [77] where anyone may download VFM. VFM performs frame grabbing directly from the video camera source into Matlab in the RGB (red-green-blue) image format. The RGB format consists of a three-dimensional array. The first two dimensions correspond to the image size. For example, an image that is 160 pixels in the horizontal direction and 120 pixels in the vertical direction has the first two dimensions of the RGB image of 120 and 160. The third dimension in the RGB format is always three (3). One can think of an RGB image as three overlaying matrices. The three matrices have values representing the amount of red, green, and blue in the image.

In order to reduce computation time the number of dimensions were reduced by converting the RGB image into a common grey-scale image with the Matlab command **rgb2gray** (see the Matlab script trackme6.m in appendix 4). The grey-scale format has only two dimensions and does not retain the original colour values, but consists of grey-scale intensity values.

The next step is to convert the variable type, of the values within the image matrix, from integer precision to double precision. The double variable type provides more manipulability computationally. This conversion is realised by the Matlab command **Im2double** (see the Matlab script trackme6.m in appendix 4). These steps were performed on all acquired images before any further calculations.

A relationship between the real world lateral position in centimetres (cm) X and the position in the image plane in pixels in the x-direction had to be determined. This was done by looking at an image of a measuring stick 1.4 metres away from the camera positioned in the horizontal plane and perpendicular to the camera's lens. The image frame was one hundred and sixty (160) pixels in the lateral direction and the image of

the measuring stick showed that the corresponding lateral distance is one hundred and four centimetres (104cm). Therefore, the scaling factor used throughout the simulations was 160/104 ≅ 1.5 pixels per centimetre. The mapping of an object's movement as seen in the two-dimensional image plane to movement in the three-dimensional real world is, in general, a non-trivial task. The simple relationship shown above was used throughout the subjective experiment discussed in this chapter. This relationship is based on two assumptions. The first assumption is that the tracked object is a distance of 1.4 metres from the camera when it is directly in front of the camera and moves in a purely parallel direction to the image plane. The other assumption made was a perspective projection of the camera.

7.3. Procedure

Figure 7.5 shows a block diagram and a plan view of the system's experimental arrangement. The modification from the experimental arrangement of chapter 6 includes the addition of the video camera and the image processing head tracking algorithm. The output of the head-tracking algorithm is the position of the listener. This output is used to determine the automatic selection of the appropriate set of virtual acoustic imaging filters for the subject's location. The loudspeakers and the video camera were mounted on the moveable slide, which movement was controlled by a programmable control card. Thereby, the system's loudspeakers and video camera were moved while the listener was stationary. The movement tracked by the head-tracking algorithm was the relative motion between the stationary subject and the moving video camera. The black cloth enclosing the sphere and the upper half of the sphere that were used in the experiment in chapter 6 were removed for this experiment. This ensured that there was a line of sight between the video camera and the listener. The subjects were asked to shut their eyes during the experimental trials so that the visual location of the loudspeakers did not affect their subjective responses of the virtual sound image location. Azimuthal angles remained marked on the inside of the hemisphere framework. A black cloth was draped behind the subject to provide a plain background that aids in the video head-tracking performance. Figure 7.6 shows a picture of this set-up including the black cloth background, metal framework lower hemisphere, the two (2) loudspeakers, and the subjects' seat all inside an

anechoic chamber. The wooden box in the right of the picture is an acoustic enclosure that reduces the noise of the motor that controls the loudspeaker\video camera motion. The loudspeakers subtend an angle of 10° with the subject and are a distance of 1.4 m away from the subject. The virtual acoustic imaging filters were identical to those designed for the experiment described in chapter 6 that utilised KEMAR dummy HRTFs and regularisation with $\beta = 10^{-4}$.

Fourteen (14) subjects took part in the experiment between the ages of 23-38 with normal hearing. Each subject responded to twenty-four (24) trials, twelve (12) with unfiltered woman's speech as the stimulus signal, and twelve (12) with band-passed white noise as the stimulus signal. The pass band of the white noise is 0.3-3 kHz. These were the same stimuli used in the experiment described in chapter 6. The virtual image location θ_v was either at 45° to the front right of the subject or at 45° to the front left of the subject (i.e. $\theta_v = +45°$ or $-45°$).

The filter update movement increments were either always 3 cm, always 5 cm, or varied between 3 and 5 cm during the trial depending on the location of the loudspeakers with respect to the virtual image location. During the variable filter update movement increment trials, the filters are updated every 3 cm when the loudspeakers are on the opposite side of the listener as the virtual sound image and updated every 5 cm when the loudspeakers are on the same side as the virtual sound image. These variable filter update movement increments are hereafter designated as 3\5 cm. These three (3) different filter update movement increments (i.e. 3 cm, 5 cm, and 3\5 cm) were chosen after consideration of the results of the experiment described in chapter 6.

The paths traversed by the loudspeakers in this experiment were the same as the paths used in the experiment described in chapter 6 and depicted in Fig. 6.14. The maximum speed of the loudspeakers\video camera motion was 5 cm/s and the maximum off-axis listener location was 35 cm from the inter-source axis. The amount of time of each trial was approximately 14 s. The combinations of the variables were

presented randomly. Before each trial, a stationary unfiltered reference signal was presented to the subject to orient the subject and give them an example of the stimulus.

The subjects responded to both the perceived horizontal location of the virtual acoustic image and the perceived percentage of time the image is at the perceived horizontal location. The subjects could have asked to repeat a trial is they had wished.

7.4. Results

Out of 373 total trials in this experiment, 99 of the subjects' responses localised the virtual sound image at angle locations greater than 90° or from behind them (i.e. $|\theta_v| > 90°$). Therefore, the percentage of front-back reversals was 26.5%, which is very similar to the degree of front-back reversals for the dynamic subjective experiment without video tracking described in chapter 6 (i.e. 23.9%). The following results have all been corrected for front-back reversals. The two responses considered are the difference in the perceived angle location from the intended angle location in degrees and the percentage of time the virtual sound image is perceived at the perceived angle location. Appendix 3 provides tables of the mean and standard deviations of the results and ANOVA tables.

7.4.1. Stimuli comparison

Figure 7.7 shows the mean perceived angle difference responses given for both the speech and white noise stimuli as a function of filter update movement increment. The localisation of the image is not significantly different under the two different stimuli as shown by the ANOVA Table C.2 in appendix 3. This is different from the result obtained in chapter 6 without employing the video head-tracking algorithm. Although, the mean perceived angle location was only about 4° different with the different stimuli in chapter 6, this small difference was shown to be significantly different. With video head tracking, the difference in the mean perceived angle location is about 2° but this small difference cannot be said to be a result of difference in the stimuli.

Similarly, the stability of the image is not seen to be significantly different because of different stimuli as seen in Fig. 7.8 and the ANOVA Table C.4 in appendix 3. Figure 7.8 shows the mean percentage of time the image is stable for both the speech and white noise stimuli as a function of filter update movement increment. Included are the results for the stationary trials (*nm*). This figure shows that the speech and white noise are very similarly stable at all filter update movement increments and is less stable at greater increments. The one exception may possibly be the 3\5 cm filter update movement increment, which appears to be more stable for the white noise than the speech. ANOVA analysis on this filter update movement increment finds that the two mean responses for the different stimuli not significantly different (Table C.5). This is in agreement with the results of chapter 6, which found that the stabilities of the two (2) stimuli are not significantly different. It is interesting that by varying the filter update movement increment between 3 cm and 5 cm the image is more stable than if the filter update movement increment is left at 5 cm constantly for the white noise stimulus, although, this trend does not hold for the speech stimulus and is hardly dramatic for the white noise stimulus. There was little effect on the stability of the image by varying the filter update movement increment between 3 cm and 5 cm depending on if the loudspeakers were on the opposite or same side of the listener as the virtual acoustic image.

7.4.2. Loudspeaker path comparison

Figure 7.9 shows the perceived angle difference for the four (4) different loudspeaker paths as a function of filter update movement increment. The ANOVA table C.7 shows that the results for the different paths are not significantly different. Although, for the 5 cm filter update movement increment it can be said with a 85.6% confidence level that the mean perceived angle difference for the different loudspeaker paths are significantly different (Table C.8). This confidence level is rather low but from Fig. 7.9 the "toward-away" path is generally perceived a little more to the listener's side than the other paths. This was also noticed in chapter 6 without the video head tracking.

Figure 7.10 shows the perceived percentage of time the image is stable for the different loudspeaker paths as a function of filter update movement increment. The "toward-away" loudspeaker path is seen to be a little more stable than the other paths. This again is a trend noticed from the results of the experiment described in chapter 6 without the video head-tracking procedure.

7.4.3. The effect of filter update movement increment

Figure 7.11 shows the difference in the perceived angle location of the virtual acoustic image for all trials as a function of filter update movement increment. Also, shown are the results for the stationary trails (*nm*). The error bars show the standard deviations of the data. The perceived angle location remains at approximately the same location for all of the filter update movement increments and the stationary trials. This statement is confirmed by ANOVA analysis, which shows that the mean responses for the different filter update movement increments are not significantly different (Table C.11). This was also noticed in chapter 6 without the tracking procedure for all of the filter update movement increments with the exception of the significantly different mean perceived angle obtained with a filter update movement increment of 2 cm. The frequency of updating the filters does not significantly effect the perceived location of the virtual sound image at the filter update movement increments considered.

Figure 7.12 shows the percentage of time the image is perceived as stable as a function of the filter update movement increment for all of the trials. Also, shown are the results for the stationary trails (*nm*). The error bars show the standard deviations of the data. The stability of the image decreases as the filter update increment increases. The means are significantly different with a confidence level of 100.0% as seen in the ANOVA Table C.13 in appendix 3. The mean stability response of filter update movement increment 3\5 cm is almost equal to the mean stability response of filter update movement increment 5 cm. The variation of the filter update movement increment between 3 cm and 5 cm depending on if the loudspeakers were on the opposite or on the same side of the listener as the virtual acoustic image, did little or nothing to effect the stability of the image.

Figure 7.13 compares the stability of the virtual sound image at filter update movement increments 3 cm and 5 cm and at the stationary trials (*nm*) as found with this experiment that incorporates a video camera and head tracking algorithm with the results found from the experiment described in chapter 6 without the video camera. The mean stability results for the stationary trials are approximately the same with and without the video processing, as one would expect. When the system is moving and adapting to the movement the better stability is achieved without the video camera and head tracking procedure. The image processing adds some delay into the procedure making the filters adaptation sluggish enough to affect the listeners' perceptions of the stability of the virtual sound image.

7.4.4. Tracking performance

Figures 7.14-7.16 show some examples of the tracking algorithm's results for the some trials of this experiment. The plots in this figure show both the motion of the loudspeakers with a solid line and the selected filter design location with a dashed line. The horizontal axes represent time elapsed during the trial in seconds. The vertical axes represent the relative lateral listener position with respect to the inter-source axis in centimetres (x). Shown in the figures are some examples of different paths of motion and different filter update movement increments. The filter selection is seen to lag the motion by a variable amount between about 0.3-1.5 seconds.

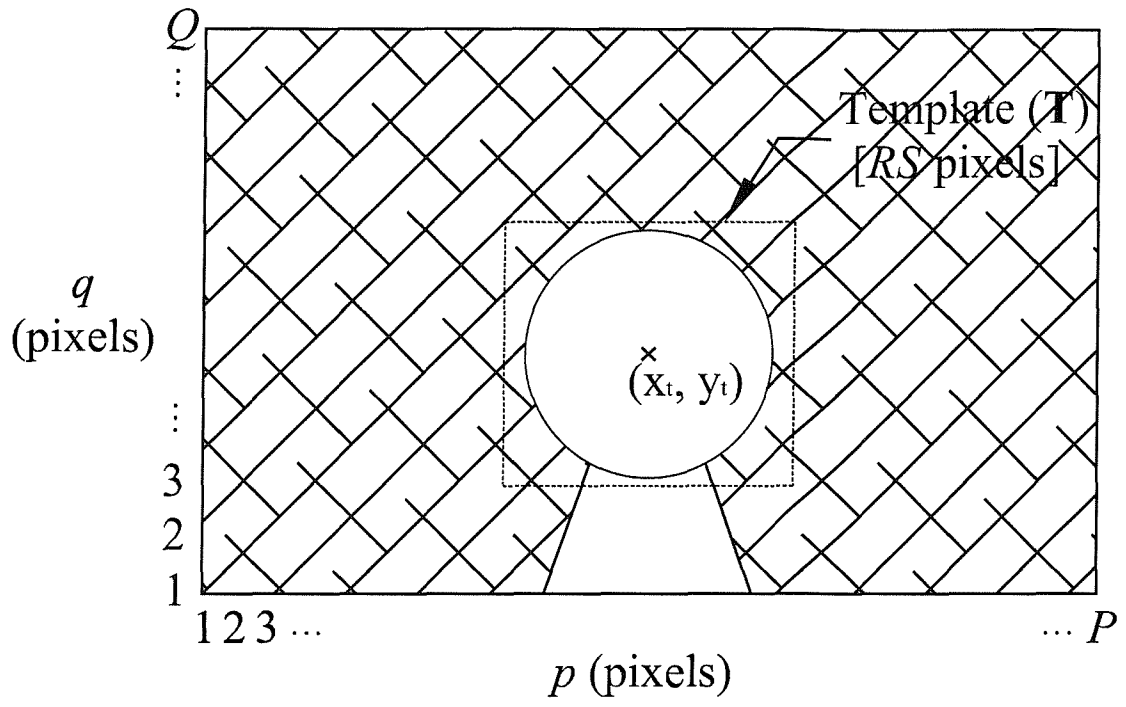
7.5. Conclusions

The disadvantages of the simple video tracking algorithm used was that it depends on knowledge of the listener's distance from the system and requires the listener to move in only a plane perpendicular to the inter-source axis. The algorithm also requires a plain background to contrast with the image of the listener. These conditions severely limit the algorithm usefulness for general use. However, some success was achieved in tracking the motion and giving the subject the desired perception of a stable virtual acoustic image. The lag in the image processing (about 0.3 s) was enough to affect the listener's perception. This is partly due to the algorithm's implementation in Matlab and not on another more efficient program or on a DSP chip. As it was, the

introduction of the video head tracking greatly affected the stability of the virtual acoustic image.

The two different stimuli were not seen to be significantly different in location or stability. The "toward-away" loudspeaker path was seen to be the most stable and slightly pushed the location of the image further to the subjects' sides than the other loudspeaker paths. This is the only path with the loudspeakers on the same side of the subject as the virtual sound image throughout the whole path. These trends were also noticed the experiment without video head tracking in chapter 6. The location of the image was not affected by different filter update movement increments as also found in chapter 6. However, the filter update movement increment did greatly affect the stability of the virtual acoustic image as also was found in chapter 6. The variable 3\5 cm filter update movement increment trials were comparably stable to the 5 cm filter update movement increment trials and so not found to benefit the performance of the system.

a)



b)

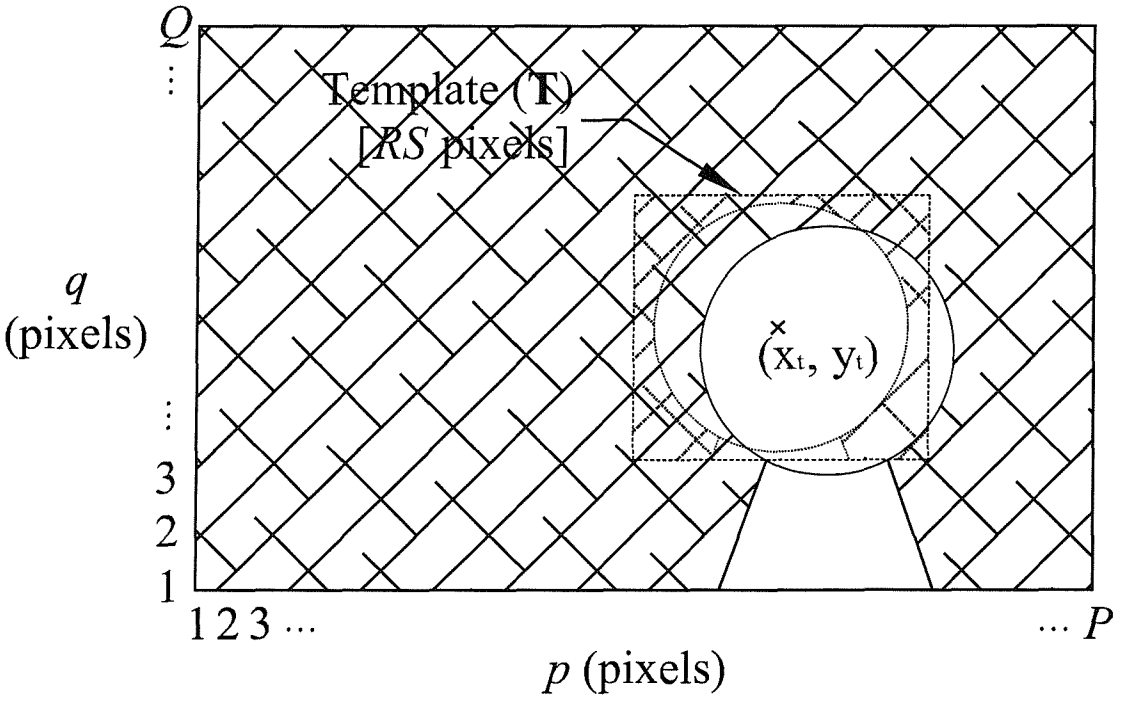


Fig. 7.1 *Two captured images of a sequence a) first captured frame* $\mathbf{M}_{i=1}$ *in which the template* **T** *is extracted from and b) a captured frame* $\mathbf{M}_i$ *after movement has occurred to which the template* **T** *is fitted. These images have P pixels in the horizontal direction and Q pixels in the vertical direction.*

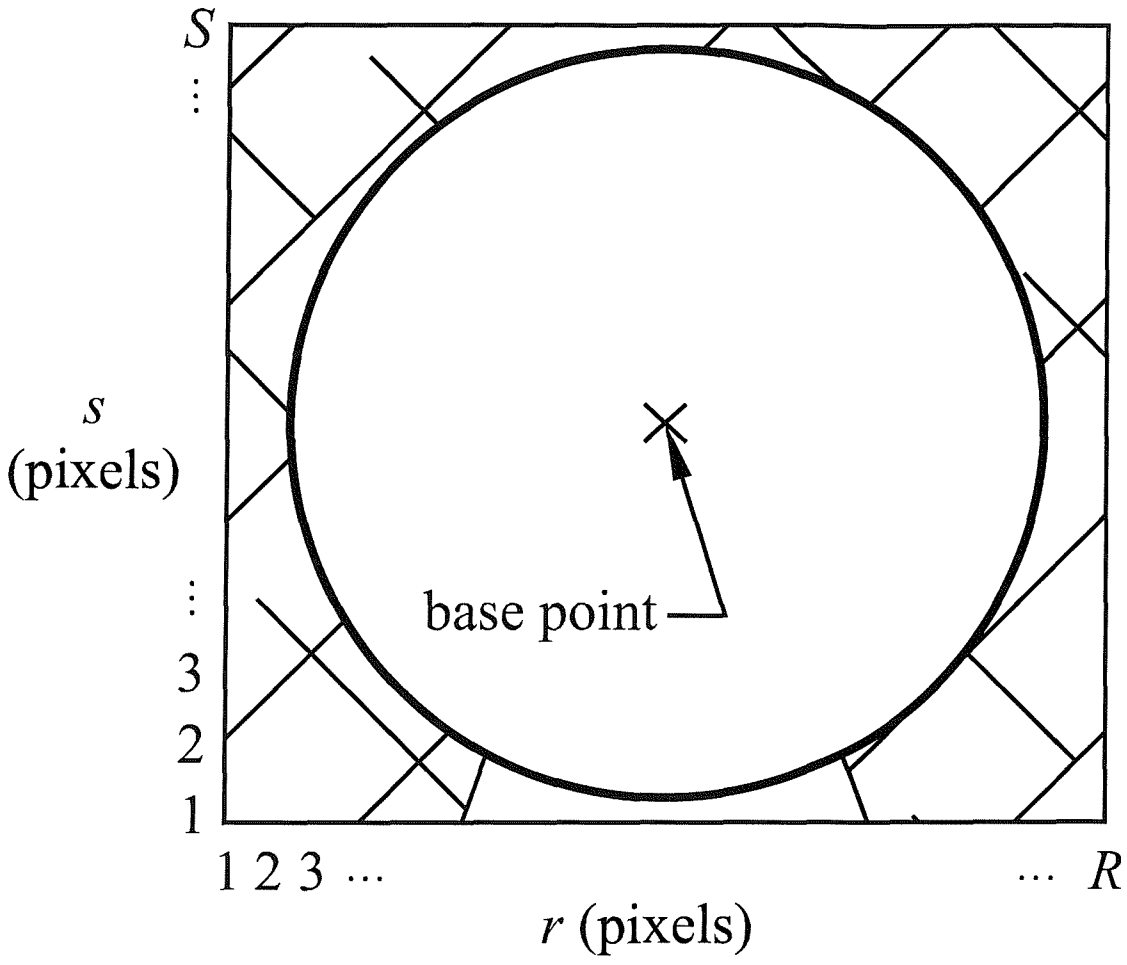
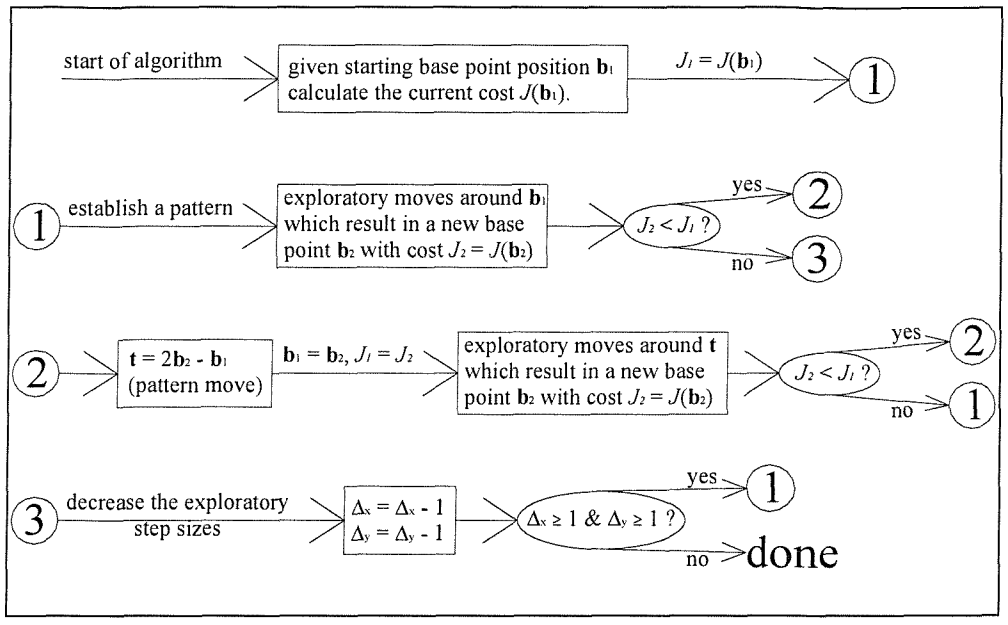
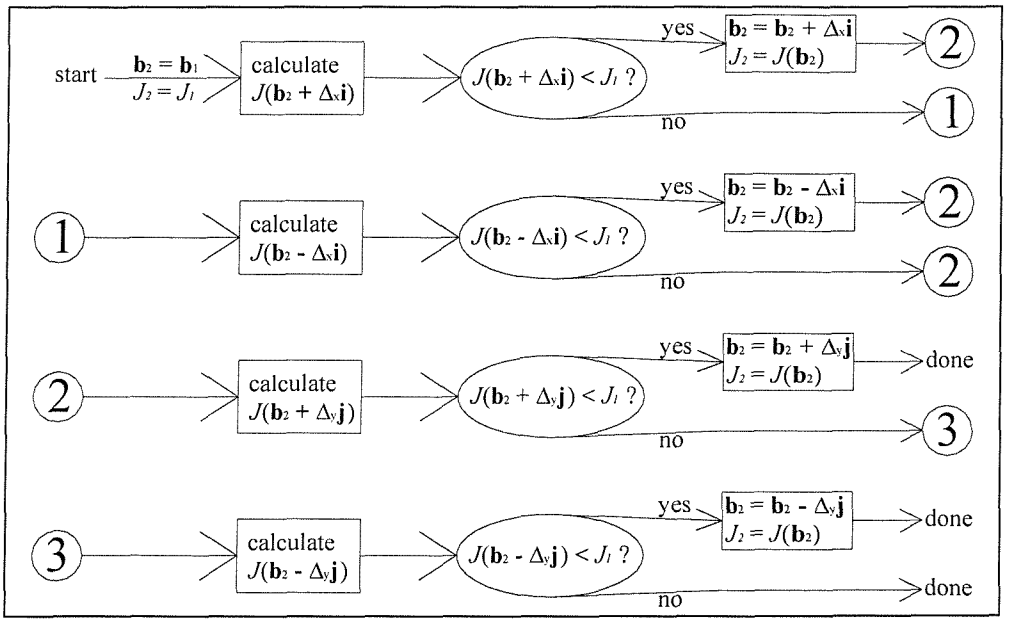


Fig. 7.2 *Close up of the image template* **T** *for the example shown in figure 7.1. This image has R pixels in the horizontal direction and S pixels in the vertical direction.*

a)



b)



c)

$J(\mathbf{q})$	= cost at position $\mathbf{q}$
$\mathbf{b}_1$	= previous template base point position
$\mathbf{b}_2$	= current template base point position
$\mathbf{t}$	= temporary template position
J_1	= cost at position $\mathbf{b}_1$
J_2	= cost at position $\mathbf{b}_2$
Δ_x	= exploratory step size in x-direction
Δ_y	= exploratory step size in y-direction
$\mathbf{i}$	= unit normal vector in positive x-direction
$\mathbf{j}$	= unit normal vector in positive y-direction

Fig. 7.3 *The pattern search algorithm showing a) flow chart of the overall pattern search method b) flow chart for the exploratory moves c) explanation for the notation used in a) and b).*

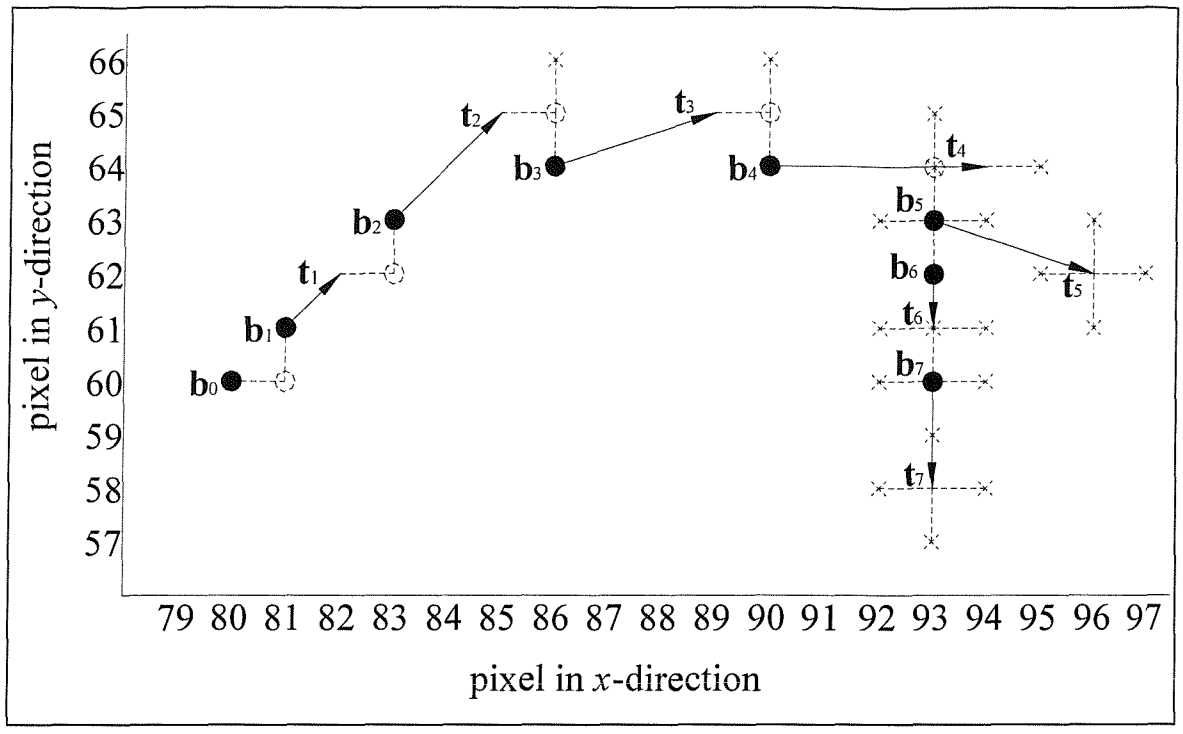


Fig. 7.4 *Example of detailed movements undertaken in a pattern search. The search began at pixel* (*80,60*) $\mathbf{b}_0$ *and stopped at pixel* (*97, 60*) $\mathbf{b}_7$.

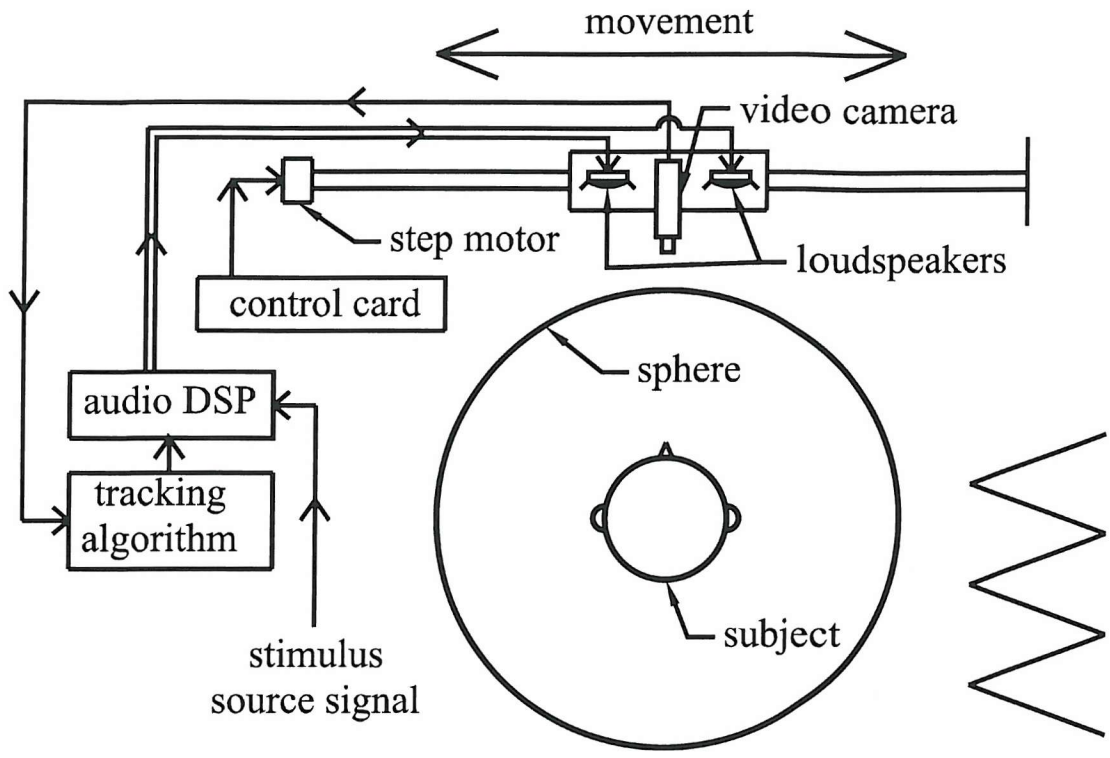


Fig. 7.5 *Plan view of experimental set-up.*

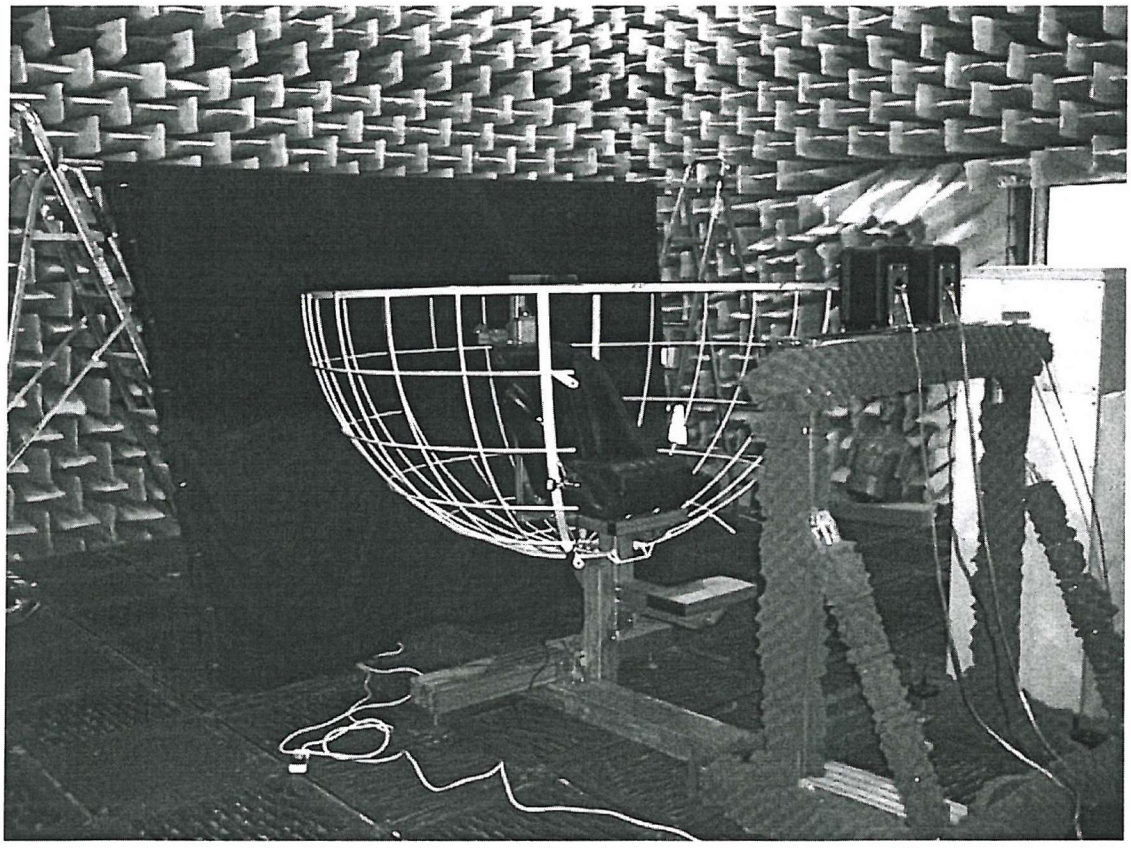


Fig. 7.6 *Picture of experimental rig.*

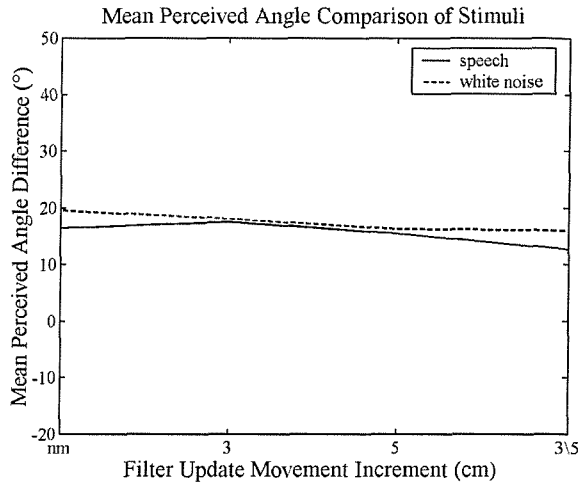


Fig. 7.7 *Mean differences in between the angle perceived and the intended virtual source location as a function of filter update movement increment for speech and white sound sources. Positive values reflect a perception of the source at locations further to the side of the subject than the intended virtual source location. Negative values reflect a perception of the source at locations closer to the front of the subject than the intended virtual source location. A value of zero reflects perceptions of the source at the intended virtual source location.*

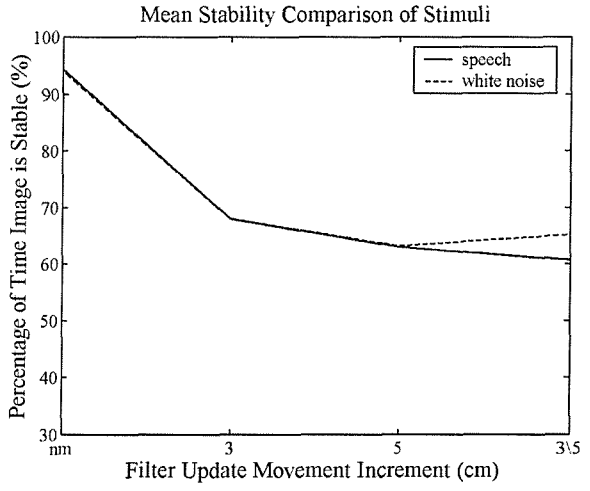


Fig. 7.8 *Mean perceived percentages of time the virtual sound image is stable as a function of filter update movement increment for speech and white sound sources.*

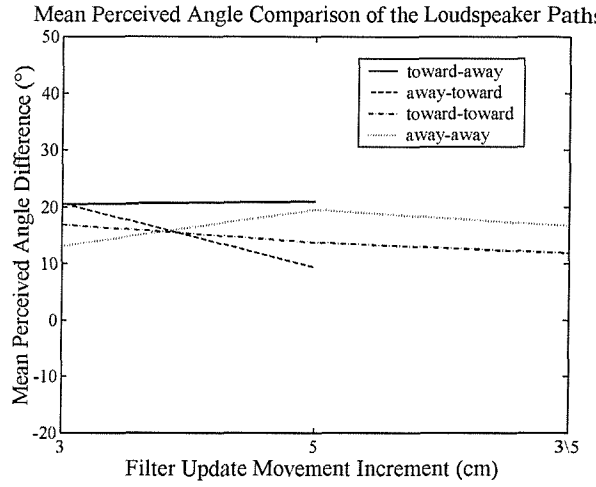


Fig. 7.9 *Mean differences between the angle perceived and the intended virtual source location as a function of filter update movement increment for the four different loudspeaker paths.*

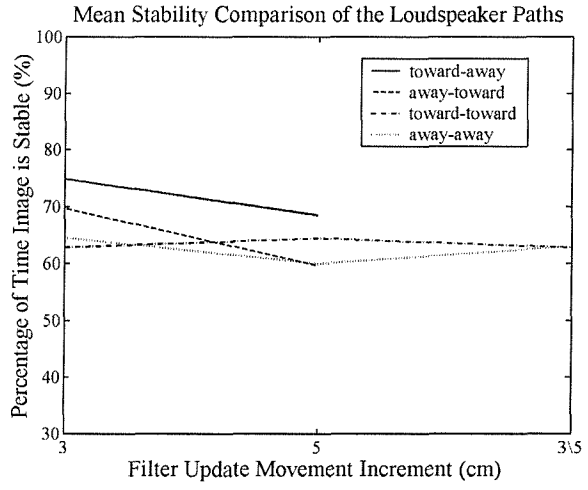


Fig. 7.10 *Mean perceived percentages of time the virtual sound image is stable as a function of filter update movement increment for the four different loudspeaker paths.*

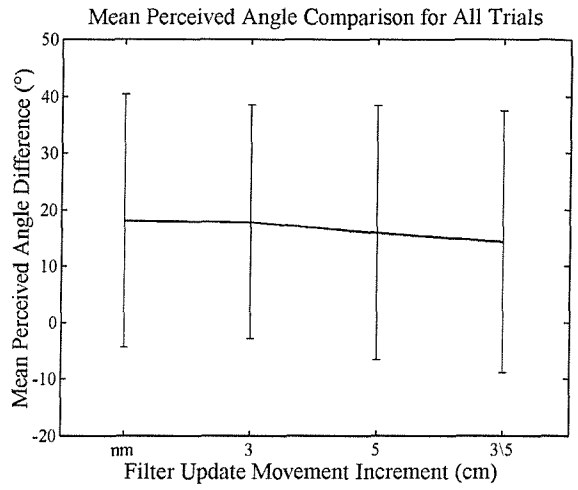


Fig. 7.11 *Mean differences between the angle perceived and the intended virtual source location as a function of filter update movement increment for all trials. Error bars show the standard deviations.*

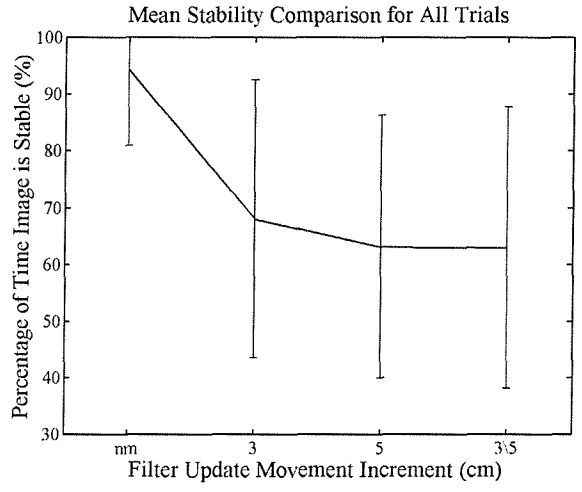


Fig. 7.12 *Mean perceived percentages of time the virtual sound image is stable as a function of filter update movement increment for all trials. Error bars show the standard deviations.*

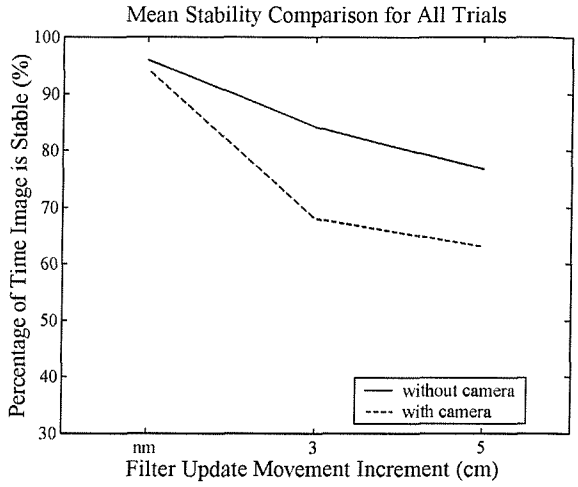


Fig. 7.13 *Mean perceived percentages of time the virtual sound image is stable during the stationary trials (nm) and at filter update movement increments 3 cm and 5 cm for all trials. Solid line is the results for the experiment described in chapter 6 without the video camera and head-tracking algorithm and the dashed line is for the experiment described in chapter 7 with the video camera and head-tracking algorithm.*

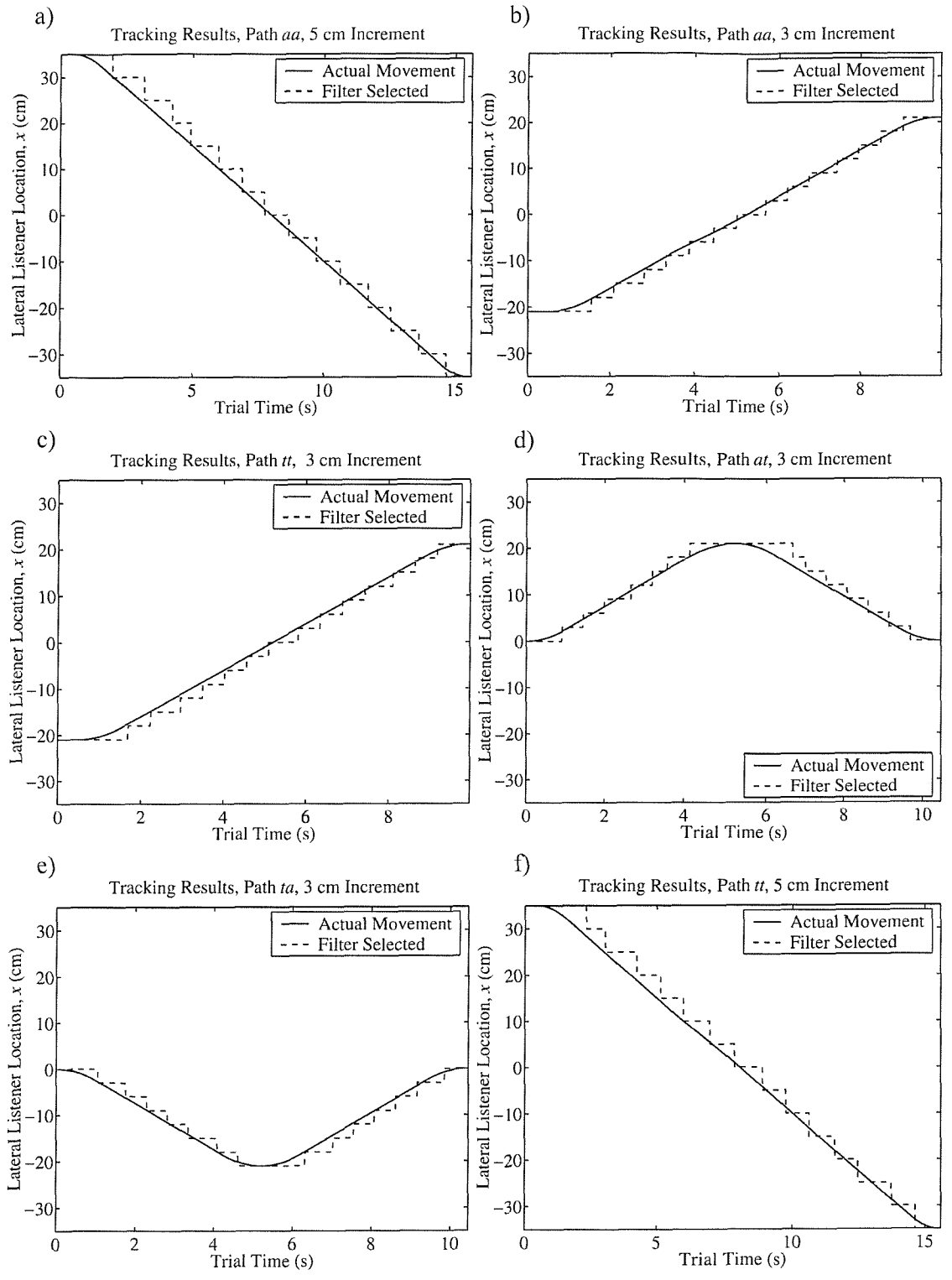


Fig. 7.14 *Motion of loudspeakers and filter selection as a result of the image processing head-tracking output during selected experimental trials.*

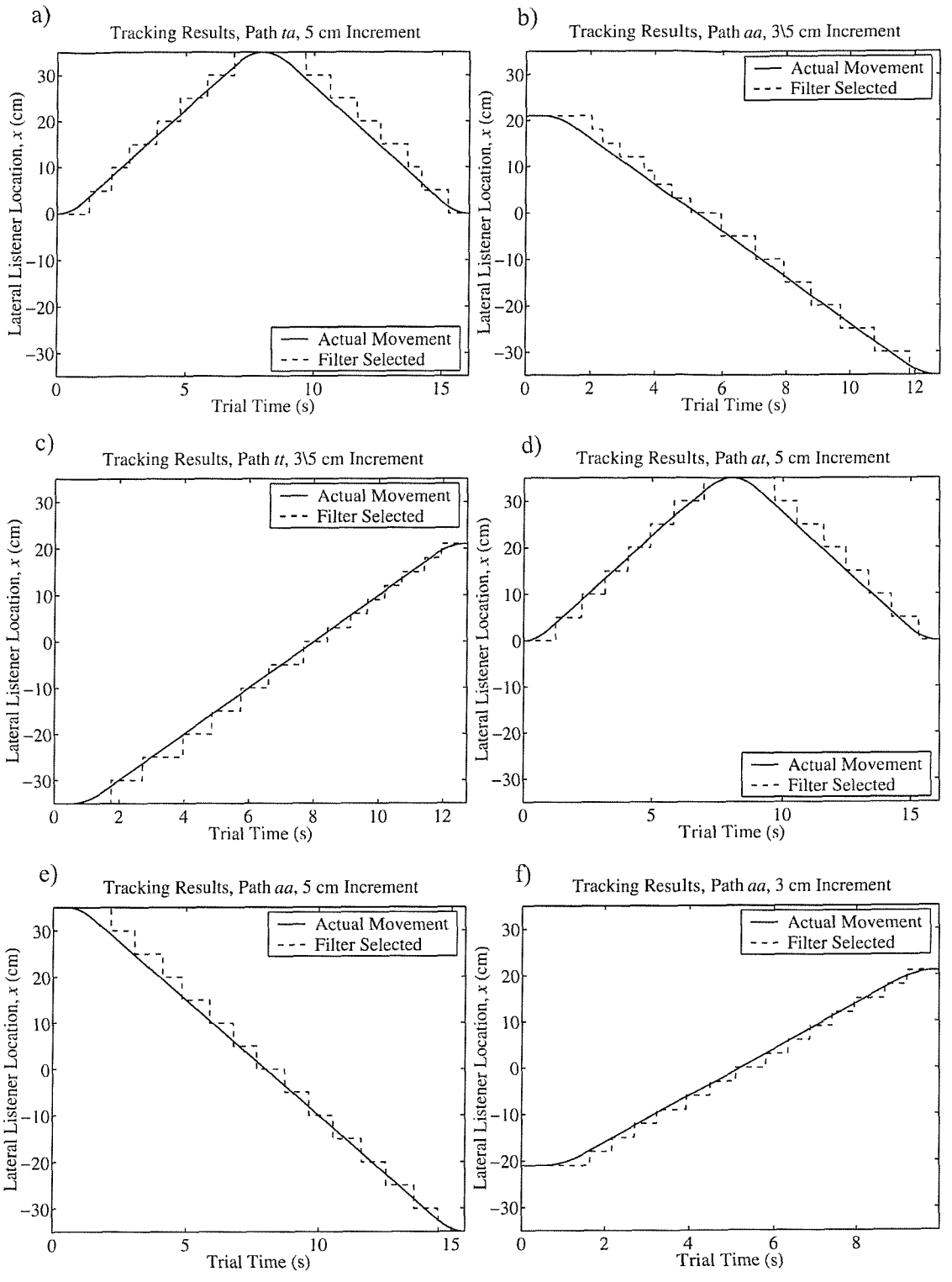


Fig. 7.15 *Motion of loudspeakers and filter selection as a result of the image processing head-tracking output during selected experimental trials.*

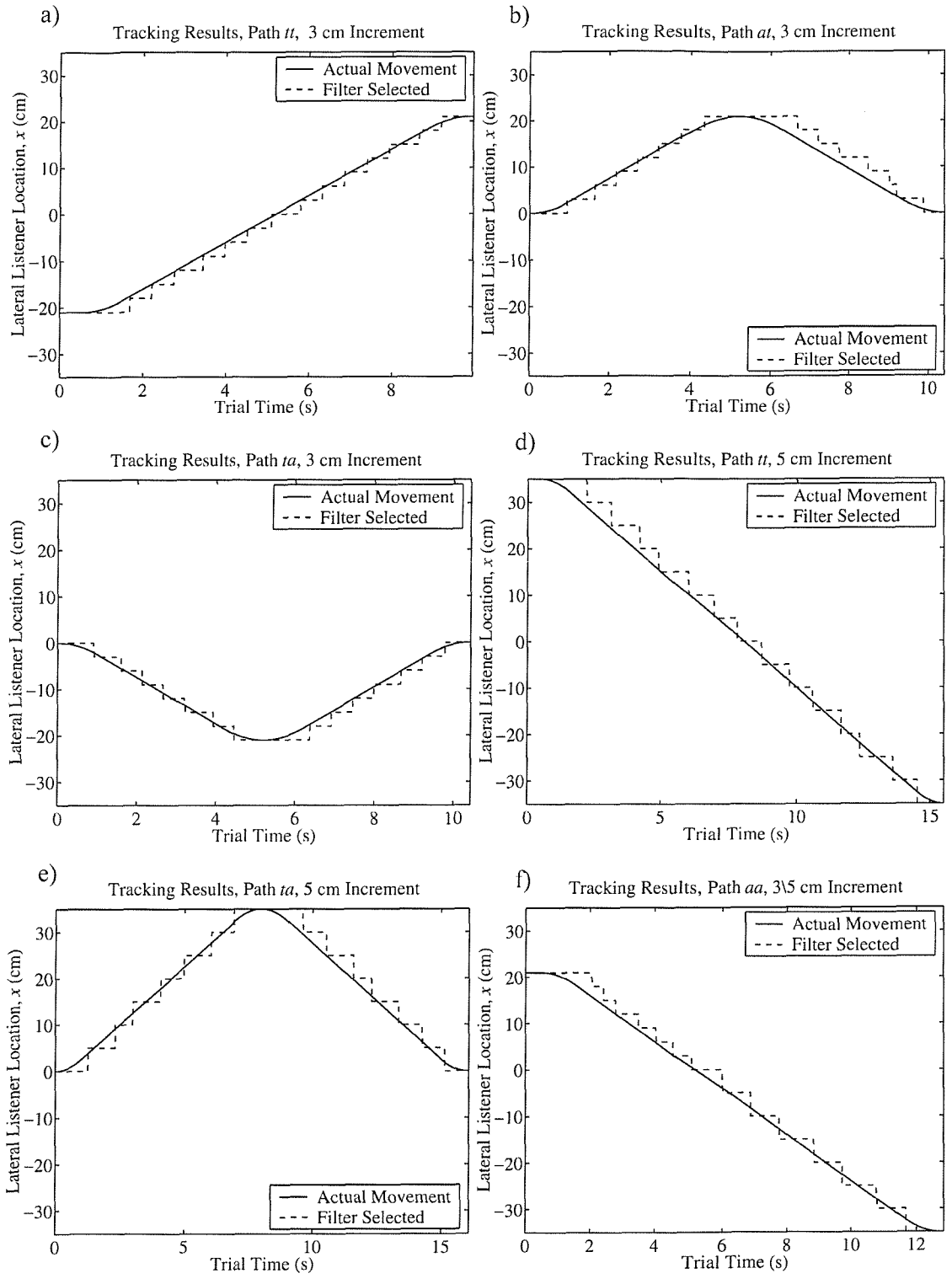


Fig. 7.16 *Motion of loudspeakers and filter selection as a result of the image processing head-tracking output during selected experimental trials.*

8. CONCLUSION

The novel work of this thesis constituted development of an unobtrusive visually adaptive virtual acoustic imaging system capable adaptation for lateral listener movements and an increase of knowledge concerning the performance of the system. The system, utilising two loudspeakers that are a distance of 1.4 m away from the listener when directly in front of the listener, was able to give listeners the impression of a stable virtual acoustic image with lateral movements spanning a distance of 35 cm away from the inter-source axis. The degree of success at achieving a stable virtual acoustic image depends on several factors, the main of which are the degradation of human sound localisation cues, the update rate of the virtual acoustic imaging filters, and the relative position of the virtual acoustic image with respect to the position of the loudspeakers. The rate of degradation of the sound localisation cues with respect to listener displacement depends on the geometry of the system. The results below all pertain to the previously developed Stereo Dipole virtual acoustic imaging system geometric arrangement, which is known to be particularly robust to listener motion. Possible further improvements or modifications of the system are also briefly mentioned below.

Both interaural time difference (ITD) and interaural level difference (ILD) sound localisation cues degrade at similar rates for head displacements from the system's optimal intended design locations. Their comparable rate of degradation makes it problematical to separate the affects of the localisation cues on the system's robustness to listener location and make a statement about the relative importance of the cues on the system's performance. Taken separately the two cues both predict a similar size of "sweet spot". The frequency ranges where these two cues are important are different and so the cues are more or less important depending on the frequency range of the source signal. The similarity in the rate of degradation of the two cues meant that two different source signals with different frequency spectra achieved similar subjective results from the system.

The frequency range that is the most robust to head motion shifts up to higher frequencies at listener locations further away from the inter-source axis. The low frequencies contain salient localisation information and so if the source signal includes low frequencies the robustness of the system's performance at asymmetric listener locations is less than the robustness of the system at the traditional symmetric on-axis location. At asymmetric listener locations head shadowing causes the robustness of ILD at the two ears to be dramatically different. The sound field of the contra-lateral ear is more robust than that at the ipsi-lateral ear.

The most important factor in determining the size of the "sweet spot" is the position of the sound sources relative to the virtual image location. A quicker filter update rate is required when the virtual acoustic image and the real sound sources are farther away from each other. The computer simulations and the subjective evaluations both suggest a filter update movement increment of less than about 3 cm in circumstances when the virtual source location is 45° to the front side of the listener and the loudspeakers on the opposite side of the listener. With the virtual acoustic image at 45° to the front side of the listener and the real sound sources on the same side of the listener, both the computer simulations and the subjective evaluations suggest a filter update movement increment of less than about 5 cm. Changing the filter update movement increment between either 3 cm or 5 cm depending on whether or not the loudspeakers are on the opposite or same side of the listener as the 45° virtual acoustic image respectively, resulted in comparable system performance of that achieved by keeping the filter update movement increment constantly at 5 cm. Varying the filter update movement increment depending on the relative location of the virtual acoustic image and the location of the loudspeakers more effectively, most likely requires a more sophisticated approach that more gradually varies the filter update movement increment.

The adaptive virtual acoustic imaging system performance was significantly worse when the system was modified to include the video head-tracking algorithm. A substantial improvement to the system would be to implement the image processing on a DSP chip to speed up the processing time and reduce the adaptation latency in

the system, which is currently around 0.3-1.5s. The image processing DSP card would have an input connected to the video signal, implement the tracking algorithm, and then output the resulting calculated position of the listener to the audio DSP card for appropriate virtual acoustic imaging filter selection. The image processing time is a major restriction of the performance of the system as it is currently implemented on a personal computer with the user friendly but relatively inefficient Matlab software package.

More sophisticated head-tracking algorithms have yet to be tried. It would be helpful to improve the image-processing head tracking algorithm so that the system could be used in more general situations without the constraints of a plain background and limited to the only one possible type of listener movement.

The video camera might also be used for detecting and classifying the shape of a listener's pinnae to select appropriate virtual acoustic imaging filters that have been calculated with HRTFs with similarly shaped pinnae. This would require several different types of virtual acoustic imaging filters designed with different HRTFs all of which would be stored in a database available for selection. This has been done by Kyriakakis et al. [3].

It would be an improvement if the system were capable of adapting to the listening environment. Even though systems with virtual acoustic imaging filters based on head related transfer functions (HRTFs) measured in an anechoic environment perform reasonably well in non-anechoic environments, the system's performance does noticeably degrade to some extent in these listening environments.

The degree of front-back confusions by listeners may be reduced by placing the loudspeakers above the listener as suggested by Takeuchi [35]. In addition, the optimal source distribution (OSD) loudspeaker arrangement can help increase the robustness and performance of a virtual acoustic imaging system. Although, these

types of geometrical arrangements may be more inconvenient than the Stereo Dipole virtual acoustic imaging system for a typical home entertainment application.

The robustness of the Stereo Dipole at non-traditional listener locations for all types of listener movements would be a helpful area of work. Most important in this proposed work would be finding the required frequency of filter updates in order to deliver the desired virtual acoustic image perception to a moving listener for all the possible different types of head motion. This required update rate will depend on the relative location of the virtual acoustic image to the position of the system's loudspeakers. Revealing this dependence would be very valuable to the design of adaptive virtual acoustic imaging systems.

APPENDICES

Appendix 1

Table A.1 *Percentage of listener front-back reversals for the static subjective experiment discussed in chapter 5.*

Intended Design Location	Number of Presentations	Loudspeakers Moving Toward	Loudspeakers Moving Away	Total
$x = 0$	354	13.2 %	7.4 %	10.7 %
$x = 5$ cm	322	11.6 %	0.7 %	6.8 %
$x = 10$ cm	298	9.0 %	3.1 %	6.4 %
$x = 15$ cm	373	20.4 %	16.0 %	18.5 %
$x = 20$ cm	361	15.4 %	10.6 %	13.3 %
$x = 25$ cm	250	19.1 %	5.5 %	13.2 %
Total	1878	14.9 %	8.3 %	12.2 %

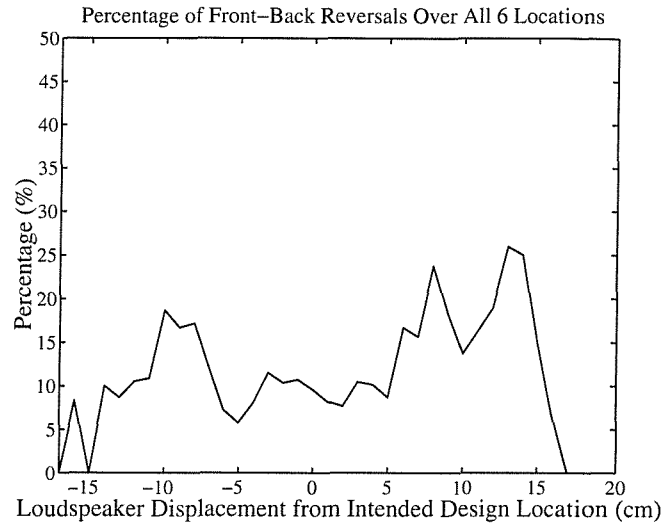


Fig. A.1 *Percentage of front-back reversals for all six (6) intended filter design positions as a function of the loudspeaker displacement from the intended location during the static subjective experiment discussed in chapter 5.*

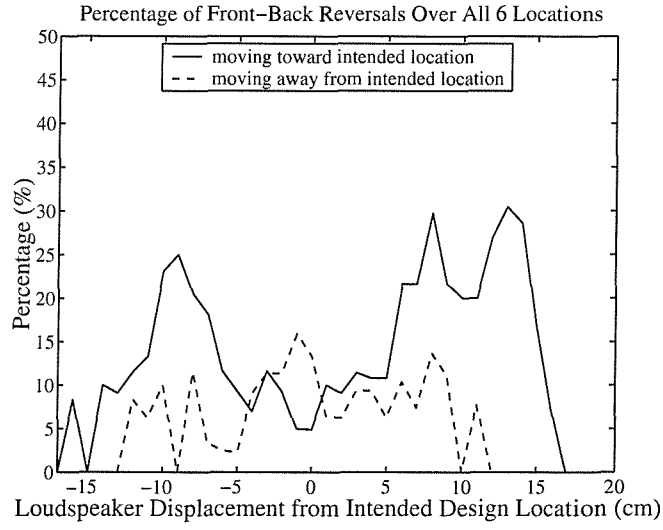


Fig. A.2 *Percentage of front-back reversals for all six (6) intended filter design positions as a function of the loudspeaker displacement from the intended location during the static subjective experiment discussed in chapter 5. Both the results when the loudspeakers were incrementally moved toward the intended location (solid line) and when the loudspeakers were moved incrementally away from the intended filter design location (dashed line) are shown.*

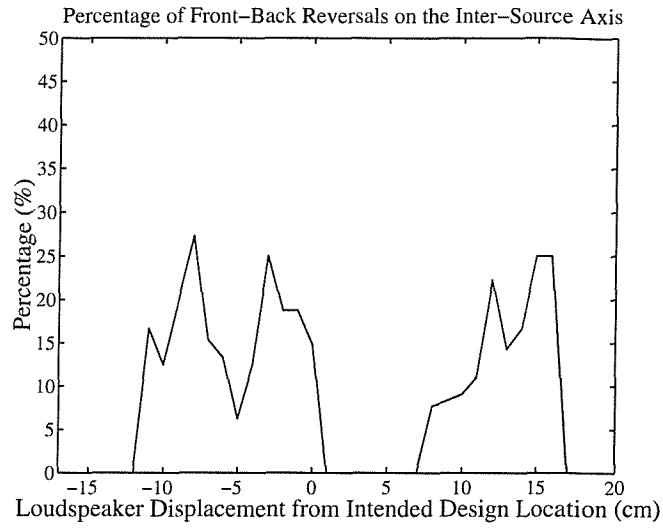


Fig. A.3 *Percentage of front-back reversals for the symmetric on-axis intended filter design position as a function of the loudspeaker displacement from the intended on-axis location during the static subjective experiment discussed in chapter 5.*

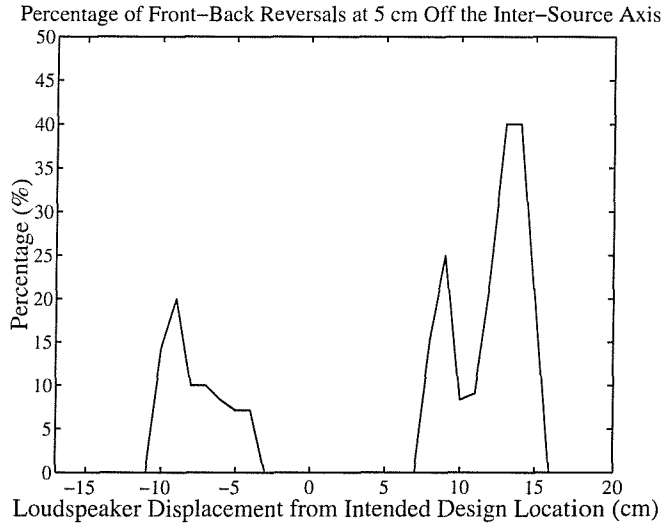


Fig. A.4 *Percentage of front-back reversals for the 5 cm off-axis intended filter design position as a function of the loudspeaker displacement from the intended 5 cm off-axis location during the static subjective experiment discussed in chapter 5.*

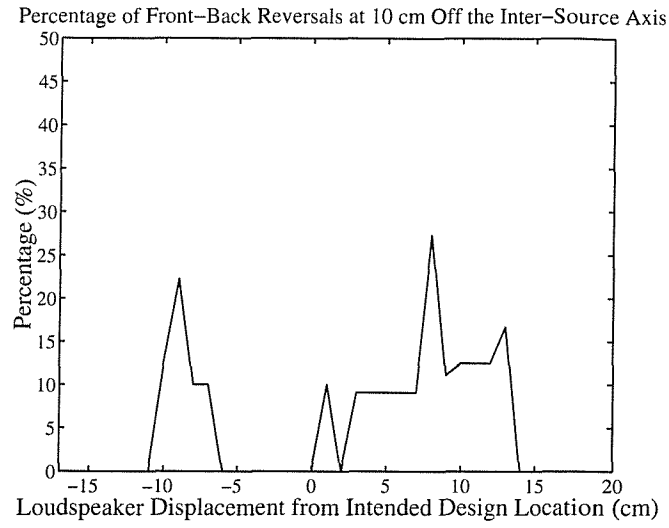


Fig. A.5 *Percentage of front-back reversals for the 10 cm off-axis intended filter design position as a function of the loudspeaker displacement from the intended 10 cm off-axis location during the static subjective experiment discussed in chapter 5.*

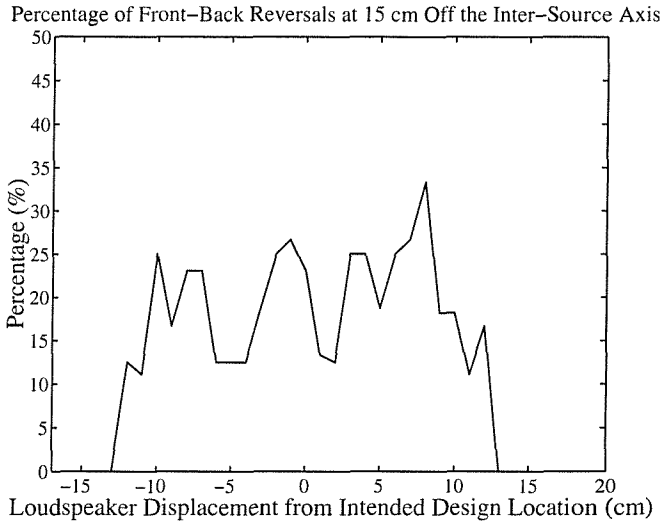


Fig. A.6 *Percentage of front-back reversals for the 15 cm off-axis intended filter design position as a function of the loudspeaker displacement from the intended 15 cm off-axis location during the static subjective experiment discussed in chapter 5.*

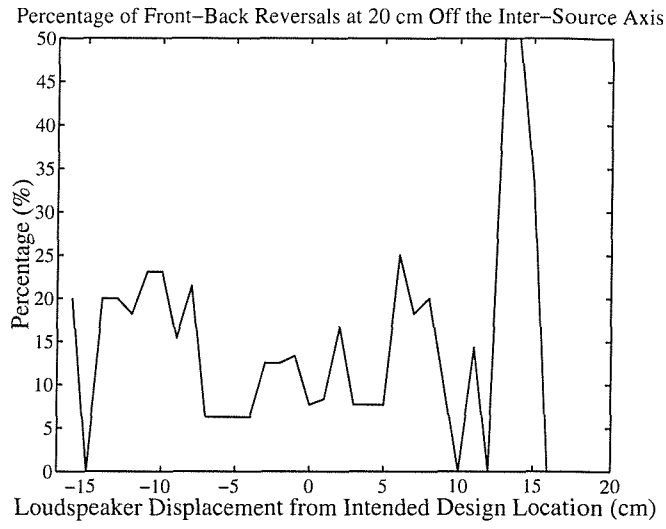


Fig. A.7 *Percentage of front-back reversals for the 20 cm off-axis intended filter design position as a function of the loudspeaker displacement from the intended 20 cm off-axis location during the static subjective experiment discussed in chapter 5.*

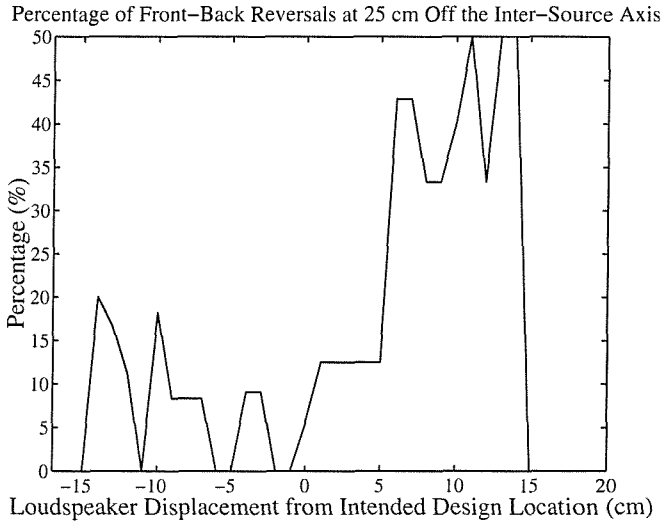


Fig. A.8 *Percentage of front-back reversals for the 25 cm off-axis intended filter design position as a function of the loudspeaker displacement from the intended 25 cm off-axis location during the static subjective experiment discussed in chapter 5.*

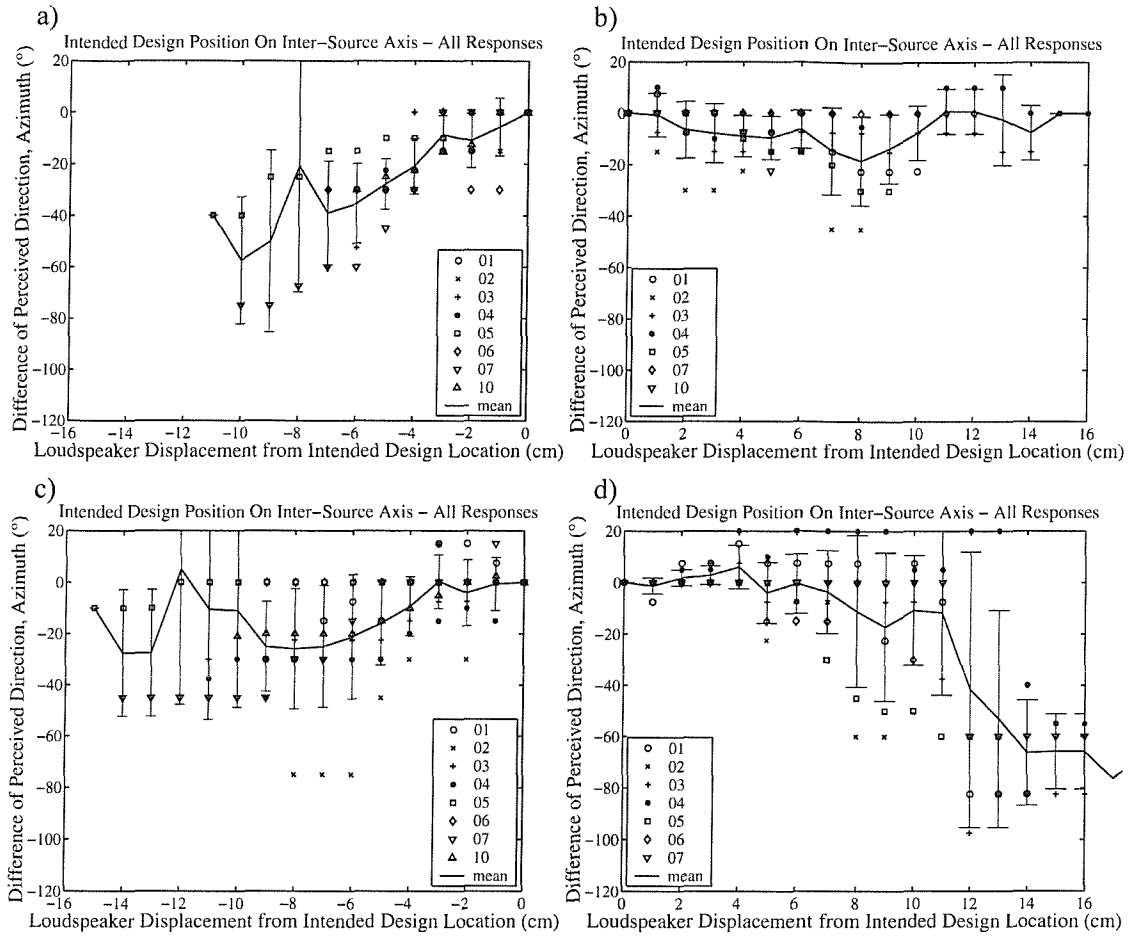


Fig. A.9 *Subjects' responses for the symmetric on-axis intended filter design position with front-back confusions resolved during the static subjective experiment discussed in chapter 5. Different symbols denote different subjects. Responses are shown with a) loudspeakers moving away from intended filter design location, to the left, b) loudspeakers moving away from intended filter design location, to the right, c) loudspeakers moving toward the intended filter design location, from the left, and d) loudspeakers moving toward the intended filter design location, from the right.*

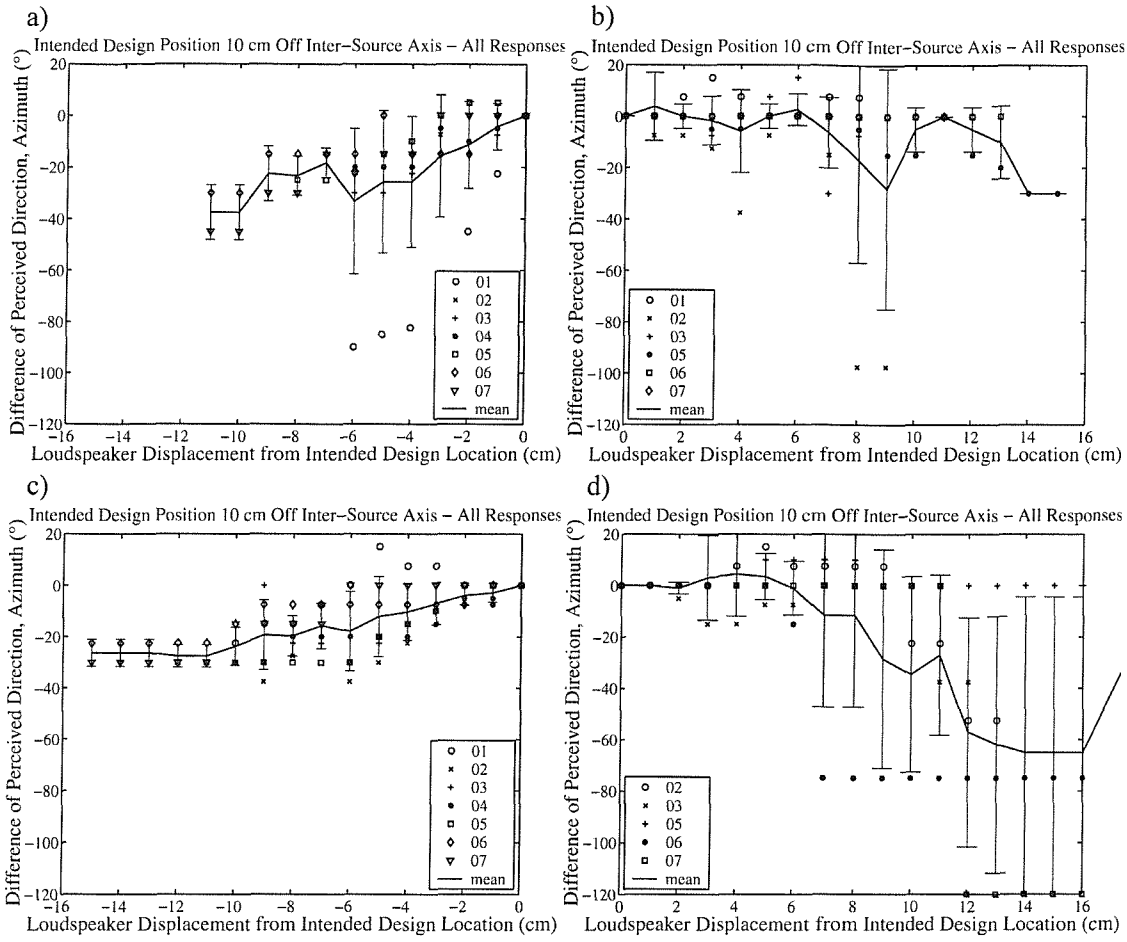


Fig. A.10 *Subjects' responses for the 10 cm off-axis intended filter design position with front-back confusions resolved during the static subjective experiment discussed in chapter 5. Different symbols denote different subjects. Responses are shown with a) loudspeakers moving away from intended filter design location, to the left, b) loudspeakers moving away from intended filter design location, to the right, c) loudspeakers moving toward the intended filter design location, from the left, and d) loudspeakers moving toward the intended filter design location, from the right.*

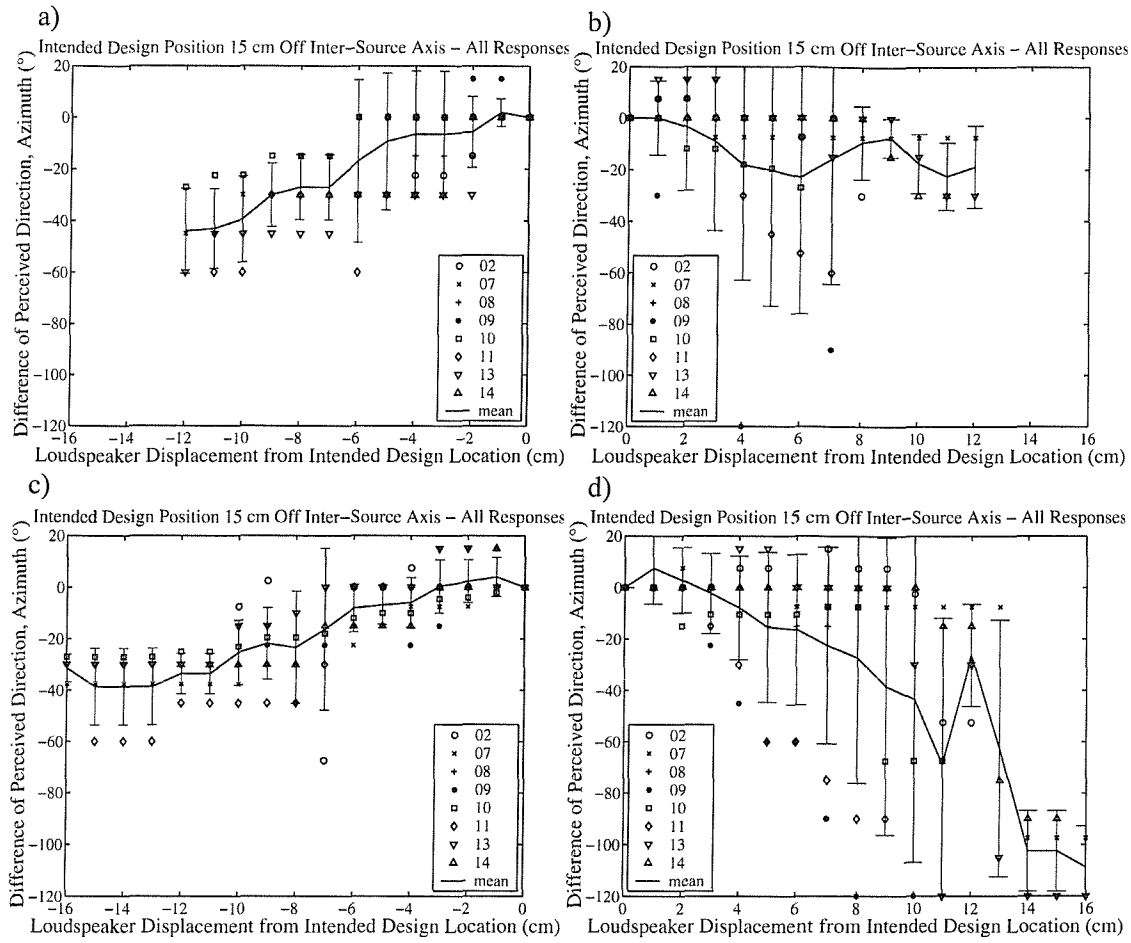


Fig. A.11 *Subjects' responses for the 15 cm off-axis intended filter design position with front-back confusions resolved during the static subjective experiment discussed in chapter 5. Different symbols denote different subjects. Responses are shown with a) loudspeakers moving away from intended filter design location, to the left, b) loudspeakers moving away from intended filter design location, to the right, c) loudspeakers moving toward the intended filter design location, from the left, and d) loudspeakers moving toward the intended filter design location, from the right.*

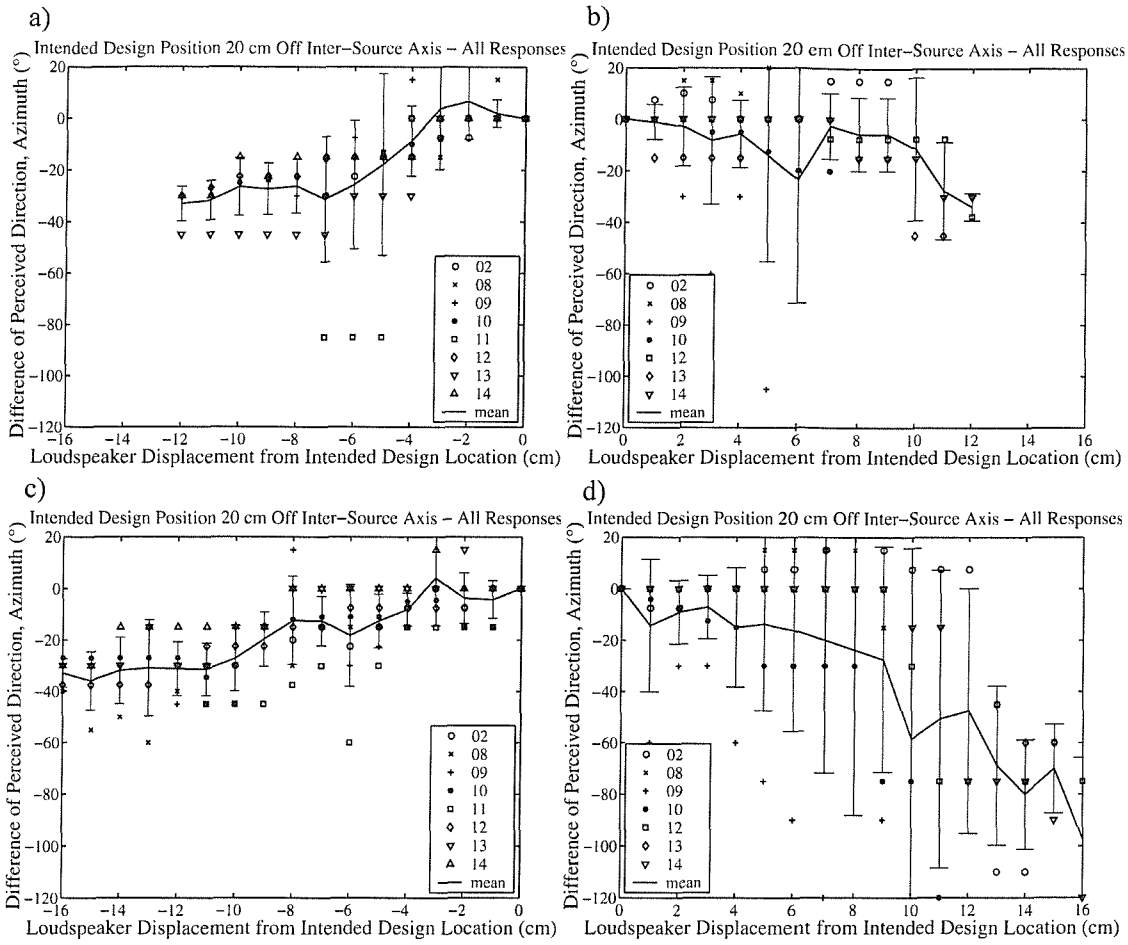


Fig. A.12 *Subjects' responses for the 20 cm off-axis intended filter design position with front-back confusions resolved during the static subjective experiment discussed in chapter 5. Different symbols denote different subjects. Responses are shown with a) loudspeakers moving away from intended filter design location, to the left, b) loudspeakers moving away from intended filter design location, to the right, c) loudspeakers moving toward the intended filter design location, from the left, and d) loudspeakers moving toward the intended filter design location, from the right.*

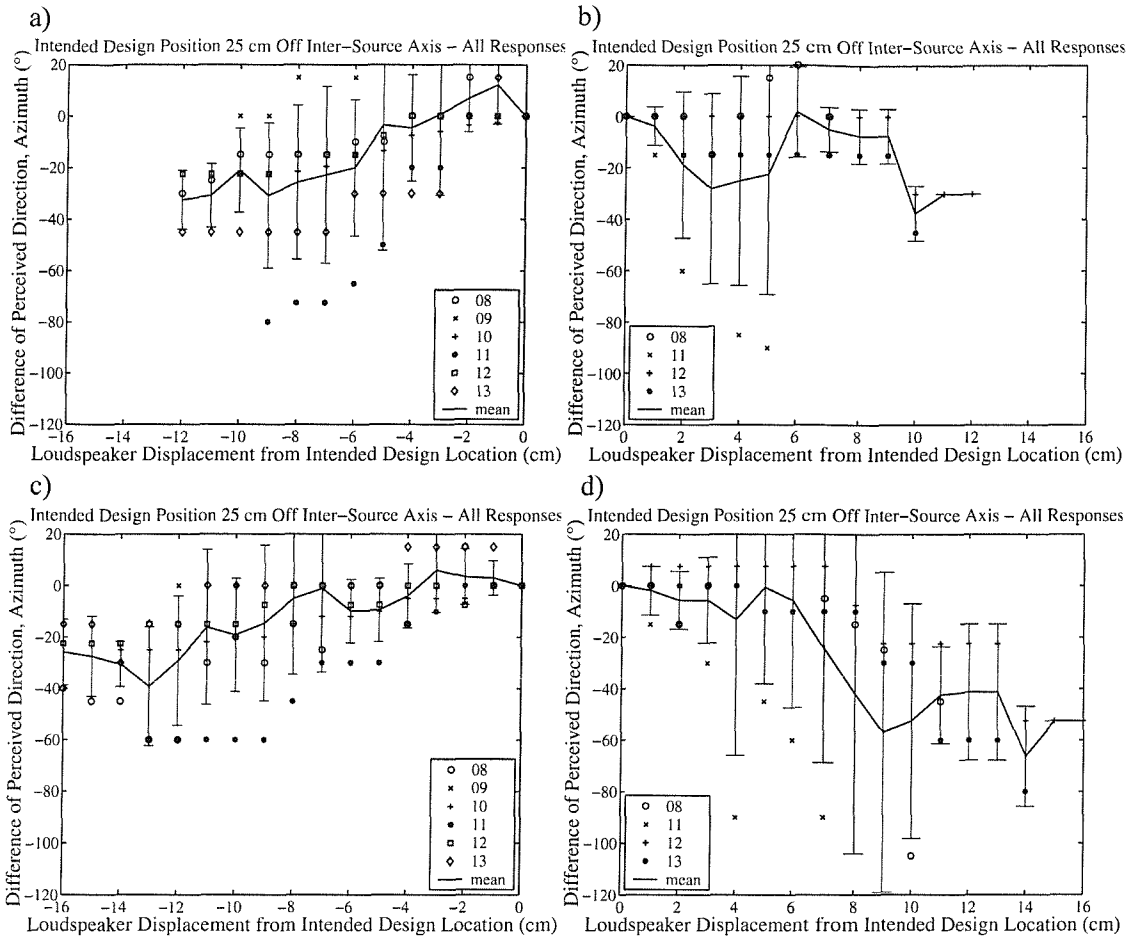


Fig. A.13 *Subjects' responses for the 25 cm off-axis intended filter design position with front-back confusions resolved during the static subjective experiment discussed in chapter 5. Different symbols denote different subjects. Responses are shown with a) loudspeakers moving away from intended filter design location, to the left, b) loudspeakers moving away from intended filter design location, to the right, c) loudspeakers moving toward the intended filter design location, from the left, and d) loudspeakers moving toward the intended filter design location, from the right.*

Appendix 2

Table B.1 *ANOVA table of the perceived angle difference for the different stimuli during all of the trials for the dynamic subjective experiment discussed in chapter 6*

	Sum of Squares	df	Mean Square	F	Sig.
Between Stimuli (Combined)	2148.945	1	2148.945	4.112	0.043
Within Stimuli	291636.039	558	522.645		
Total	293784.984	559			

Table B.2 *Mean and standard deviations of the perceived angle difference (°) for the different stimuli for the dynamic subjective experiment discussed in chapter 6*

Filter Update Movement Increment (cm)	Stimuli			
	Speech		White Noise	
	Mean (°)	Std. Deviation (°)	Mean (°)	Std. Deviation (°)
0	11.75	25.834	18.00	20.903
2	25.18	17.872	25.89	17.054
3	9.75	22.672	15.13	22.206
4	6.70	24.159	12.39	25.958
5	9.04	24.133	15.67	24.455
6	8.00	21.833	10.50	20.438
7	15.42	19.942	13.11	22.043
All	11.45	23.167	15.36	22.552

Table B.3 *Mean and standard deviations of the percentage of time the image is stable (%) for the different stimuli for the dynamic subjective experiment discussed in chapter 6*

Filter Update Movement Increment (cm)	Stimuli			
	Speech		White Noise	
	Mean (%)	Standard Deviation (%)	Mean (%)	Standard Deviation (%)
0	96.50	6.998	95.38	9.567
2	84.82	17.400	85.71	19.086
3	84.25	23.739	84.00	23.183
4	80.34	26.422	80.57	22.314
5	79.13	24.728	74.33	24.092
6	58.63	25.394	65.88	24.830
7	49.72	28.106	54.44	32.924
All	76.39	27.173	77.07	26.038

Table B.4 *ANOVA table of the percentage of time the image is stable (%) for the different stimuli during the stationary reference trials for the dynamic subjective experiment discussed in chapter 6*

	Sum of Squares	df	Mean Square	F	Sig.
Between Stimuli (Combined)	25.313	1	25.313	0.360	0.550
Within Stimuli	5479.375	78	70.248		
Total	5504.688	79			

Table B.5 *Mean and standard deviations of the perceived angle difference (°) for the different loudspeaker movement paths for the dynamic subjective experiment discussed in chapter 6*

Filter Update Movement Increment (cm)	Loudspeaker Movement Path							
	aa		*at*		*ta*		*tt*	
	Mean (°)	Std. Dev. (°)	Mean (°)	Std. Dev. (°)	Mean (°)	Std. Dev. (°)	Mean (°)	Std. Dev. (°)
2	26.43	11.837	18.93	19.921	29.64	14.205	27.14	21.458
3	13.00	20.092	10.25	20.357	15.25	30.585	11.25	18.272
4	5.45	27.599	9.55	17.856	15.68	28.673	7.50	25.390
5	12.50	22.989	11.92	21.170	15.00	29.223	10.00	24.698
6	2.50	19.297	5.00	14.956	21.50	22.775	8.00	22.384
7	3.06	20.661	14.17	13.422	23.72	22.996	16.11	21.390
All	9.83	22.379	11.21	18.357	19.27	26.074	12.33	22.972

Table B.6 *ANOVA table of the perceived angle difference for the different loudspeaker movement paths during all of the trials for the dynamic subjective experiment discussed in chapter 6*

	Sum of Squares	df	Mean Square	F	Sig.
Between Movement Types (Combined)	6543.642	4	1635.911	3.161	0.014
Within Movement Types	287241.342	555	517.552		
Total	293784.984	559			

Table B.7 *Mean and standard deviations of the percentage of time the image is stable (%) for the different loudspeaker movement paths for the dynamic experiment discussed in chapter 6*

Filter Update Movement Increment (cm)	Loudspeaker Movement Path							
	aa		*at*		*ta*		*tt*	
	Mean (%)	Std. Dev. (%)	Mean (%)	Std. Dev. (%)	Mean (%)	Std. Dev. (%)	Mean (%)	Std. Dev. (%)
2	73.57	16.919	90.00	13.587	90.36	19.262	87.14	18.472
3	70.50	27.621	89.00	19.440	93.00	15.927	84.00	23.709
4	74.09	28.395	82.05	22.817	90.68	12.564	75.00	27.903
5	68.46	29.992	73.46	24.810	88.65	13.532	76.35	22.959
6	49.25	22.727	63.25	23.802	80.25	19.227	56.25	25.177
7	38.06	27.714	56.39	25.077	70.00	30.438	43.89	30.320
All	62.67	29.226	75.29	25.032	85.75	19.975	70.42	28.694

Table B.8 *Mean and standard deviations of the perceived angle difference (°) for the different filter update movement increments for the dynamic subjective experiment discussed in chapter 6*

Filter Update Movement Increment (cm)	Mean (°)	Std. Deviation (°)
0	14.88	23.559
2	25.54	17.312
3	12.44	22.488
4	9.55	25.093
5	12.36	24.405
6	9.25	21.050
7	14.26	20.902
All	13.41	22.925

Table B.9 *ANOVA table of the perceived angle difference for the different filter update movement increments during all of the trials for the dynamic subjective experiment discussed in chapter 6*

	Sum of Squares	df	Mean Square	F	Sig.
Between Filter Update Movement Increments (Combined)	11347.977	6	1891.330	3.703	0.001
Within Filter Update Movement Increments	282437.007	553	510.736		
Total	293784.984	559			

Table B.10 *ANOVA table of the perceived angle difference for the different filter update movement increments excluding the 2 cm filter update movement increment trials for the dynamic subjective experiment discussed in chapter 6*

	Sum of Squares	df	Mean Square	F	Sig.
Between Filter Update Movement Increments (Combined)	2192.253	5	438.451	0.821	0.535
Within Filter Update Movement Increments	265953.078	498	534.042		
Total	268145.331	503			

Table B.11 *Mean and standard deviations of the percentage of time the image is stable (%) for the different filter update movement increments for the dynamic subjective experiment discussed in chapter 6*

Filter Update Movement Increment (cm)	Mean (%)	Standard Deviation (%)
0	95.94	8.347
2	85.27	18.101
3	84.13	23.314
4	80.45	24.314
5	76.73	24.413
6	62.25	25.219
7	52.08	30.486
All	76.73	26.590

Table B.12 *ANOVA table of the percentage of time the image is stable (%) for the different filter update movement increments during all of the trials for the dynamic subjective experiment discussed in chapter 6*

	Sum of Squares	df	Mean Square	F	Sig.
Between Filter Update Movement Increments (Combined)	99702.622	6	16617.104	31.096	0.000
Within Filter Update Movement Increments	295517.199	553	534.389		
Total	395219.821	559			

Appendix 3

template0.m

```
clear, close all, filt1=1; Tran=2; filder=0.04*[b zeros(1,16383)];

vfm('configformat'), vfm('preview', 1); vfm('show', 1);
a1=vfm('grab', 1);  T0=rgb2gray(Im2double(a1));
M=size(a1,1);  N=size(a1,2);  Tx=round(N/2)-50;  Ty=round(M/2)-20;
xmin=(N-Tx)/2;  xmax=xmin+Tx;  ymin=(M-Ty)/2;  ymax=ymin+Ty;
T=imcrop(T0,[xmin,ymin,xmax-xmin,ymax-ymin]);
x0=round((xmax+xmin)/2);  y0=round((ymax+ymin)/2);
```

trackme6.m

```
tmp=0; l=0; flag=0; clear tf FF tt u v
while tmp<0.0769,
   stat=hurontrn('trana',Tran,zeros(1,1));
   stat=hurontrn('gen', Tran, 1);
   tmp=hurontrn('capta', Tran, 1, 1)';
end

t0=clock;  tic; tt(1)=toc;  tf(1)=tt(1);
Bx1=x0;  By1=y0; FF=[x0 x0];
while flag == 0,
   l=l+1;
   u(l)=Bx1;   v(l)=By1;

   a2=vfm('grab', 1);  Im=rgb2gray(Im2double(a2));
   Dx=1;  Dy=1;

   while Dx>=1 & Dy>=1,
      [xa ya J1 J2]=explore(Dx, Dy, x0, y0, Bx1, By1, Im, T, xmin,
xmax, ymin, ymax);
      Bx2=Bx1+xa;  By2=By1-ya;  if (J2<J1)  des=0;  else  des=1;  end
      while J2<J1,
         tx=2*Bx2-Bx1;  ty=2*By2-By1;  Bx1=Bx2;  By1=By2;  J1=J2;
         [xa ya blah J2]=explore(Dx, Dy, x0, y0, tx, ty, Im, T, xmin,
xmax, ymin, ymax);
         Bx2=tx+xa;  By2=ty-ya;
      end
      if (des==1)  Dx=Dx-1;  Dy=Dy-1;  end
   end
   tt(l+1)=toc;   tf=[tf toc];   tempupdate,   tf=[tf toc];
   if toc>=secs, flag=1; end
end
u(l+1)=Bx1; v(l+1)=By1;
```

explore.m

```
function [xa, ya, J1, J2]=explore(Dx, Dy, x0, y0, Bx, By, Im, T,
xmin, xmax, ymin, ymax)
%EXPLORE Executes exploratory moves in the pattern search
optimisation procedure, in 2 dimensions.
%   [xa, ya, J1, J2]=EXPLORE(Dx, Dy, x0, y0, Bx, By, Im, T, xmin,
xmax, ymin, ymax) calculates the change
%   in the x-direction (xa) change in the y-direction (ya) and the
cost function at this temporary head
%   (J2), given the step size's in the x-direction (Dx) and y-
direction, the image (Im), the template (T).

J1=cost(Im(ymin-(By-y0):ymax-(By-y0), xmin+Bx-x0:xmax+Bx-x0),T);
J=cost(Im(ymin-(By-y0):ymax-(By-y0), xmin+Bx-x0+Dx:xmax+Bx-x0+Dx),T);

if (J<J1)   xa=Dx;
elseif (J1<J)   J=cost(Im(ymin-(By-y0):ymax-(By-y0), xmin+Bx-x0-
Dx:xmax+Bx-x0-Dx),T);
   if (J<J1)      xa=-Dx;   else      xa=0;      J=J1;   end
else   xa=0;  end

J2=cost(Im(ymin-(By-y0)-Dy:ymax-(By-y0)-Dy, xmin+Bx-x0+xa:xmax+Bx-
x0+xa),T);

if (J2<J)   ya=-Dy;
elseif (J<J2)   J2=cost(Im(ymin-(By-y0)+Dy:ymax-(By-y0)+Dy, xmin+Bx-
x0+xa:xmax+Bx-x0+xa),T);
   if (J2<J)      ya=Dy;   else      ya=0;      J2=J;   end
else   ya=0;  end
```

cost.m

```
function J=cost(IM, TM)
%COST Sum of squared difference between to matrices (images).
%   COST(IM, TM) Calculates the sum of squared difference Im - Tm
%   over all of the elements in the two matrices.

J=sum(sum((IM-TM).^2));
```

Table C.1 *Mean and standard deviations of the perceived angle difference (°) for the different stimuli for the dynamic subjective experiment with video tracking discussed in chapter 7*

Filter Update Movement Increment (cm)	Stimuli			
	Speech		White Noise	
	Mean (°)	Std. Deviation (°)	Mean (°)	Std. Deviation (°)
0	16.41	21.710	19.53	23.292
3	17.50	20.373	18.05	21.056
5	15.47	22.951	16.27	22.179
3\5	12.68	23.233	15.96	23.410
All	15.90	21.815	17.41	22.025

Table C.2 *ANOVA table of the perceived angle difference for the different stimuli during all of the trials for the dynamic subjective experiment with video tracking discussed in chapter 7*

	Sum of Squares	df	Mean Square	F	Sig.
Between Stimuli (Combined)	210.121	1	210.121	0.437	0.768
Within Stimuli	178250.87	371	480.461		
Total	178460.99	372			

Table C.3 *Mean and standard deviations of the percentage of time the image is stable (%) for the different stimuli for the dynamic subjective experiment with video tracking discussed in chapter 7*

Filter Update Movement Increment (cm)	Stimuli			
	Speech		White Noise	
	Mean (%)	Standard Deviation (%)	Mean (%)	Standard Deviation (%)
0	94.38	9.136	94.03	16.618
3	67.97	26.167	67.97	22.882
5	62.98	24.483	63.17	21.966
3/5	60.71	25.375	65.19	24.351
All	69.69	25.926	70.45	24.286

Table C.4 *ANOVA table of the percentage of time the image is stable (%) for the different stimuli during all trials for the dynamic subjective experiment with video tracking discussed in chapter 7*

	Sum of Squares	df	Mean Square	F	Sig.
Between Stimuli (Combined)	54.981	1	54.981	0.087	0.768
Within Stimuli	234222.34	371	631.327		
Total	234277.32	372			

Table C.5 *ANOVA table of the percentage of time the image is stable (%) for the different stimuli during the 3\5 cm filter update movement increment trials for the dynamic subjective experiment with video tracking discussed in chapter 7*

	Sum of Squares	df	Mean Square	F	Sig.
Between Stimuli (Combined)	270.340	1	270.340	0.436	0.512
Within Stimuli	32209.753	52	52		
Total	32480.93	53	53		

Table C.6 *Mean and standard deviations of the perceived angle difference (°) for the different loudspeaker movement paths for the dynamic subjective experiment with video tracking discussed in chapter 7*

Filter Update Movement Increment (cm)	Loudspeaker Movement Path							
	aa		*at*		*ta*		*tt*	
	Mean (°)	Std. Dev. (°)	Mean (°)	Std. Dev. (°)	Mean (°)	Std. Dev. (°)	Mean (°)	Std. Dev. (°)
3	13.13	24.388	20.63	19.787	20.47	16.818	16.88	20.897
5	19.53	19.525	9.38	24.846	20.97	22.928	13.75	21.440
3\5	16.67	22.744	N\A	N\A	N\A	N\A	11.85	23.743
All	16.43	22.202	15.00	22.991	20.71	19.896	14.29	21.815

Table C.7 *ANOVA table of the perceived angle difference for the different loudspeaker movement paths during all of the trials for the dynamic subjective experiment with video tracking discussed in chapter 7*

	Sum of Squares	df	Mean Square	F	Sig.
Between Movement Types (Combined)	1839.340	4	459.835	0.958	0.431
Within Movement Types	176621.65	368	479.950		
Total	178460.99	372			

Table C.8 *ANOVA table of the perceived angle difference for the different loudspeaker movement paths during the 5 cm filter update movement increment trials for the dynamic subjective experiment with video tracking discussed in chapter 7*

	Sum of Squares	df	Mean Square	F	Sig.
Between Movement Types (Combined)	2728.288	3	909.429	1.834	0.144
Within Movement Types	60976.436	123	495.743		
Total	63704.724	126			

Table C.9 *Mean and standard deviations of the percentage of time the image is stable (%) for the different loudspeaker movement paths for the dynamic experiment with video tracking discussed in chapter 7*

Filter Update Movement Increment (cm)	Loudspeaker Movement Path							
	aa		*at*		*ta*		*tt*	
	Mean (%)	Std. Dev. (%)	Mean (%)	Std. Dev. (%)	Mean (%)	Std. Dev. (%)	Mean (%)	Std. Dev. (%)
3	64.53	25.945	69.69	23.519	74.84	22.842	62.81	24.820
5	59.94	22.686	59.69	22.250	68.48	22.021	64.38	25.519
3\5	62.96	23.829	N\A	N\A	N\A	N\A	62.78	26.104
All	62.45	24.021	64.69	23.263	71.71	22.490	63.35	25.178

Table C.10 *Mean and standard deviations of the perceived angle difference (°) for the different filter update movement increments for the dynamic subjective experiment with video tracking discussed in chapter 7*

Filter Update Movement Increment (cm)	Mean (°)	Std. Deviation (°)
0	17.97	22.391
3	17.77	20.638
5	15.87	22.485
3\5	14.26	23.157
All	16.65	21.903

Table C.11 *ANOVA table of the perceived angle difference for the different filter update movement increments during all of the trials for the dynamic subjective experiment with video tracking discussed in chapter 7*

	Sum of Squares	df	Mean Square	F	Sig.
Between Filter Update Movement Increments (Combined)	659.530	3	219.843	0.456	0.713
Within Filter Update Movement Increments	177801.46	369	481.847		
Total	178460.99	372			

Table C.12 *Mean and standard deviations of the percentage of time the image is stable (%) for the different filter update movement increments for the dynamic subjective experiment with video tracking discussed in chapter 7*

Filter Update Movement Increment (cm)	Mean (%)	Standard Deviation (%)
0	94.20	13.304
3	67.97	24.482
5	63.08	23.176
3\5	62.87	24.755
All	70.07	25.095

Table C.13 *ANOVA table of the percentage of time the image is stable (%) for the different filter update movement increments during all of the trials for the dynamic subjective experiment with video tracking discussed in chapter 7*

	Sum of Squares	df	Mean Square	F	Sig.
Between Filter Update Movement Increments (Combined)	46845.785	3	15615.262	30.742	0.000
Within Filter Update Movement Increments	187431.54	369	507.945		
Total	234277.32	372			

REFERENCES

[1] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia* (Academic Press, London, 1994).

[2] J. Garas, "Adaptive 3D Sound Systems" (Ph.D. thesis, Eindhoven: Technische Universiteit Eindhoven, 1999).

[3] C. Kyriakakis, T. Holman, J. Lim, H. Hong, and H. Neven, "Signal processing, acoustics, and psychoacoustics for high quality desktop audio," Journal of Visual Communication and Image Representation, **9**, 51-61 (1998).

[4] W. G. Gardner, "3-D Audio Using Loudspeakers" (Ph.D. thesis, MIT Media Laboratory, Cambridge, MA, 1997).

[5] J. F. W. Rose, "A Visually Adaptive Virtual Sound Imaging System" (MSc. thesis, Institute of Sound and Vibration Research, Southampton, UK, 1999).

[6] W. M. Hartmann, "How we localize sound," Phys. Today **52**(11), 24-29 (1999).

[7] J. Blauert (**1997**), *Spatial Hearing: The Psychophysics of Human Sound Localization* (The MIT Press, Cambridge, MA).

[8] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," Annu. Rev. Psychol. **42**, 135-159 (1991).

[9] W. A. Yost, "Lateral position of sinusoids presented with interaural intensive and temporal differences," J. Acoust. Soc. Am. **70**, 397-409 (1981).

[10] A. W. Mills, "Lateralization of high-frequency tones," J. Acoust. Soc. Am. **32**, 132-134 (1960).

[11] G. B. Henning, "Detectability of interaural delay in high-frequency complex waveforms," J. Acoust. Soc. Am. **55**, 84-90 (1974).

[12] D. McFadden and E. G. Pasanen, "Lateralization at high frequencies based on interaural time differences," J. Acoust. Soc. Am. **59**, 634-639 (1976).

[13] P. J. Bloom and P. J. Jones, "Lateralization thresholds based on interaural time differences for middle and high-frequency three-tone harmonic complexes," Acoustica **39**, 283-291 (1978).

[14] P. J. Jones and R. P. Williams, "An experiment to determine whether the interaural time differences used in lateralizing middle and high frequency complex tomes is dependent in any way on fine structure information," Acoustica **47**, 164-169 (1981).

[15] 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).

[16] R. G. Klumpp and H. R. Eady, "Some measurements of interaural time difference thresholds," J. Acoust. Soc. Am. **28**, 859-860 (1956).

[17] L. Rayleigh, "On our perception of sound direction," Philos. Mag. **13**, 214-232 (1907).

[18] J. Hebrank and D. Wright "Spectral cues used in the localization of sound sources on the median plane," J. Acoust. Soc. Am. **56**, 1829-1834 (1974).

[19] W. R. Thurlow, and P. S. Runge, "Effect of induced head movements on localization of direction of sounds," J. Acoust. Soc. Am. **42**, 480-488, (1967).

[20] J. Sandvad, "Dynamic aspects of auditory virtual environments," 100th Audio Engineering Society Convention Preprint 4226 (M-1), 1996.

[21] J. A. Altman and I. V. Kalmykova "Role of the dog's auditory cortex in discrimination of sound signals simulating sound source movement," Hearing Research **24**, 243-253, (1986).

[22] D. R. Perrott and K. Saberi "Minimum audible angle thresholds for sources varying in both elevation and azimuth," J. Acoust. Soc. Am. **87**, 1728-1731, (1990).

[23] D. W. Chandler and D. W. Grantham "Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity," J. Acoust. Soc. Am. **91**, 1624-1636, (1992).

[24] D. W. Grantham, "Auditory motion perception: snapshots revisited," ch. 15 of *Binaural and Spatial Hearing in Real and Virtual Environments* edited by R. H. Gilkey and T. R. Anderson, (Lawrence Erlbaum Associates, Inc., 1997).

[25] D. R. Begault, "Auditory and non-auditory factors that potentially influence virtual acoustic imagery," AES 16th International conference, pp. 13-26

[26] O. Kirkeby, P. A. Nelson, and H. Hamada, "Stereo dipole," Patent Application, PCT/GB97/00415, 1997.

[27] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. I: stimulus synthesis," J. Acoust. Soc. Am. **85**, 858-867 (1989).

[28] P. A. Nelson and S. J. Elliott, *Active Control of Sound* (Academic Press, London, 1992), Chap. 1, p.26.

[29] O. Kirkeby, P. A. Nelson, and H. Hamada, "Local sound field reproduction using two closely spaced loudspeakers," J. Acoust. Soc. Am. **104**, 1973-1981 (1998).

[30] F. A. Firestone, "The phase difference and amplitude ratio at the ears due to a source of pure tone," J. Acoust. Soc. Am. **2**, 260-270 (1930).

[31] W. E. Feddersen, T. T. Sandel, D. C. Teas, and L. A. Jeffress, "Localization of high frequency tones," J. Acoust. Soc. Am. **29**, 988-991 (1957).

[32] M. D. Burkhard and R. M. Sachs, "Anthropometric manikin for acoustic research," J. Acoust. Soc. Am. **58**, 214-222 (1975).

[33] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," J. Acoust. Soc. Am. **97**, 3907-3908 (1995).

[34] W. G. Gardner and K. D. Martin, KEMAR HRTF Measurements, (MIT's Media Lab through http://sound.media.mit.edu/KEMAR.html, 1994).

[35] T. Takeuchi, "Systems for virtual acoustic imaging using the binaural principle" (PhD. thesis, Institute of Sound and Vibration Research, Southampton, UK, 2001).

[36] E. H. A. Langendijk and A. W. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," J. Acoust. Soc. Am. **107**, 528-537 (2000).

[37] J. C. Middlebrooks, "Narrow-band sound localization related to external ear acoustics," J. Acoust. Soc. Am. **92**, 2607-2624 (1992).

[38] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," J. Acoust. Soc. Am. **91**, 1637-1646 (1992).

[39] Z. Wu, F. H. Y. Chan, F. K. Lam, and J. C. K. Chan, "A time domain binaural model based on spatial feature extraction for the head-related transfer function," J. Acoust. Soc. Am. **102**, 2211-2218 (1997).

[40] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Analyzing head-related transfer function measurements using surface spherical harmoncis," J. Acoust. Soc. Am. **104**, 2400-2411 (1998).

[41] T. Nishino, S. Kajita, K. Takeda, and F. Itakura, "Interpolating head related transfer functions in the median plane," Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 167-170 (1999).

[42] D.A. Bies and C.H. Hansen, *Engineering Noise Control: Theory and Practice*, Unwin Hyman (1988).

[43] T. I. Laakso, V. Valimaki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay. Tools for fractional delay filter design," IEEE Signal Process. Mag. **13**, 30-60 (1996).

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

[45] T. Takeuchi and P. A. Nelson, "Robustness to head misalignment of virtual sound imaging systems," J. Acoust. Soc. Am. **109**, 958-971 (2001).

[46] D. B. Ward and G. W. Elko, "Effect of loudspeaker position on the robustness of acoustic crosstalk cancellation," IEEE Signal Process. Lett. **6**(5), 106-108 (1999).

[47] P. A. Nelson, "Active control for virtual acoustics," presented at Active 2002, University of Southampton, Southampton, UK.

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

[49] O. Kirkeby, P. A. Nelson, and H. Hamada, "The 'Stereo Dipole' – a virtual source imaging system using two closely spaced loudspeakers," J. Audio Eng. Soc. **46**, 387-395 (1998).

[50] T. Takeuchi, and P. A. Nelson, "Optimal source distribution for binaural synthsis over loudspeakers," Acoustics Research Letters Online, **2**, 7-12 (2000).

[51] T. Takeuchi and P. A. Nelson, "Optimal source distribution system for virtual acoustic imaging," presented at the 110th AES Convention 2001 May 12-15 Amsterdam, The Netherlands, preprint 5372

[52] T. Takeuchi and P. A. Nelson, "Subjective evaluation of the optimal source distribution system for virtual acoustic imaging," Proceedings of the 19th AES International Conference, Schloss Elmau, Germany (2001).

[53] P. A. Nelson, F. Orduna-Bustamante, and D. Engler, "Experiments on a system for the synthesis of virtual acoustic sources," J. Audio Eng. Soc. **44**, 990-1007 (1996).

[54] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast deconvolution of multichannel systems using regularization," IEEE Transactions on Speech and Audio Processing **6**, 189-194 (1998).

[55] J. C. R. Licklider, in *Information Theory: Third London Symposium*, edited by C Cherry (Butterworths Scientific Publications, London, 1956), p.253-268.

[56] B. McA. Sayers and E. C. Cherry, "Mechanism of binaural fucion in the hearing of speech," J. Acoust. Soc. Am. **29**, 973-987 (1957).

[57] J. C. Licklider, in *Psychology: A Study of a Science, Study I. Conceptual and Systematic, Volume 1. Sensory, Perceptual, and Physiological Formulations*, edited by S. Koch (McGraw-Hill Book Company, Inc., New York, 1959), Vol. **1**, p.41-144.

[58] B. McA. Sayers, "Acoustic-image lateralization judgments with binaural tones," J. Acoust. Soc. Am. **36**, 923-926 (1964).

[59] J. M. Goldberg and P. B. Brown, "Response of binaural neurons of dog superior ollivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization," J. Neurophysiol. **32**, 613-636 (1969).

[60] T. C. Yin and J. C. K. Chan, "Interaural time sensitivity in medial superior olive of cat," J. Neurophysiol. **64**, 465-488 (1990).

[61] R. Batra, S. Kuwada, and T. R. Stanford, "Temporal coding of envelopes and their interaural delays in the inferior colliculus of the unanesthetized rabbit," J. Neurophysiol. **61**, 257-268 (1989).

[62] G. Msoushegian, A. L. Rupert, and J. S. Gidda, "Functional characteristics of superior olivary neurons to binaural stimuli," J. Neurophysiol. **38**, 1037-1048 (1975).

[63] C. E. Carr and M. Konishi, "A circuit for detection of interaural time differences in the brain stem of the barn owl," J. Neuroscience. **10**, 3227-3246 (1990).

[64] B. Masterton, G. C. Thompson, J. K. Bechtold, and M. J. RoBards, "Neuroanatomical basis of binaural phase-difference analysis for sound localization: a comparative study," J. Comparative and Physiological Psychology **89**, 379-386 (1975).

[65] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals," J. Acoust. Soc. Am. **80**, 1608-1622 (1986).

[66] W. Gaik, "Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling," J. Acoust. Soc. Am. **94**, 98-110 (1993).

[67] R. A. Johnson and G. K. Bhattacharyya (**2001**), Statistics: principles and methods – 4th ed. (John Wiley & Sons, Inc., NY).

[68] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of Box Plots," American Statistician **32**, 12-15 (1978).

[69] P.A. Nelson, F. Orduna-Bustamante and H. Hamada, "Inverse filter design and equalisation zones in multi-channel sound reproduction", IEEE Transactions on Speech and Audio Processing **3**, 185-192 (1995).

[70] M. Subbarao, *Interpretation of Visual Motion: A Computational Study*, Pitman Publishing, London (1988).

[71] S. Basu, I. Essa, and A. Pentland, "Motion regularization for model-based head tracking," M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 362, Submitted (Jan 1996): 13th International Conference on Pattern Recognition (ICPR '96), Vienna, Austria, August 25-30 (1996).

[72] S. H. Or, W. S. Luk, K. H. Wong, and I. King, "An efficient iterative pose estimation algorithm," Image and Vision Computing, **16**, 353-362 (1998).

[73] C. Kyriakakis and T. Holman, "Video-based head tracking for improvements in multichannel loudspeaker audio," presented at the 105th Convention of the Audio Engineering Society, San Francisco, California (1998).

[74] W. K. P. Pratt, *Digital Image Processing*, John Wiley & Sons, Inc., United States of America, 651-653 (1991).

[75] R. Hooke and T. A. Jeeves, "'Direct search' solution of numerical and statistical problems," Journal of the Association of Computing Machines, **8**, 212-229 (1962).

[76] D.J. Wilde, *Optimum Seeking Methods*, Prentice-Hall, Englewood Cliffs, New Jersey (1964).

[77] F. Pezeshkpour, *Vision For Matlab*, School of Information Systems, University of East Anglia through http://www.sys.uea.ac.uk/~fuzz/vfm/default.html., (1998).