

**Objective Assessment of Phantom Images in a
3-Dimensional Sound Field Using a Virtual Listener**

Preprint 4462 (I5)

Bernd Theiss, Malcolm O. J. Hawksford
Audio Physic Gerhard GmbH, Brilon, Germany
University of Essex, Essex, Great Britain

**Presented at
the 102nd Convention
1997 March 22–25
Munich, Germany**



AES

This preprint has been reproduced from the author's advance manuscript, without editing, corrections or consideration by the Review Board. The AES takes no responsibility for the contents.

Additional preprints may be obtained by sending request and remittance to the Audio Engineering Society, 60 East 42nd St., New York, New York 10165-2520, USA.

All rights reserved. Reproduction of this preprint, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

AN AUDIO ENGINEERING SOCIETY PREPRINT

**Objective assessment of phantom images in a 3-dimensional
sound field using a virtual listener**

Bernd Theiß* and Prof. Malcolm O. J. Hawksford**

Abstract

A method of objective quality assessment of phantom images using a virtual listener is presented. A spherical array of microphones combined with binaural processing is used to measure attributes of a 3-dimensional sound field. The virtual listener is calibrated against human subjects in a formal experimental procedure to enable objective measurements of spatial auditory parameters to be made automatically.

* Centre for Audio Research Engineering, University of Essex,
Colchester, Essex CO4 3SQ, United Kingdom and Audio Physic
Gerhard GmbH, 59929 Brilon, Germany

** Centre for Audio Research Engineering, University of Essex,
Colchester, Essex CO4 3SQ, United Kingdom

0 Introduction

Numerous sound reproduction schemes have been developed which attempt to create auditory images which humans can localize from directions where no real sound source is present, these images are called phantom images. With the advent of higher capacity storage media (e.g. DVD) even more sophisticated techniques can be anticipated in the future (ref ARA document, see web), consequently it is desirable to identify objective methods for assessing the accuracy of reproduced images that can include both image coding and the capability of the equipment in a given environment.

This paper presents an objective measurement system designated a Virtual Listener (VL) for quantifying the accuracy of phantom images. The assessment of a sound reproduction system by formal subjective listening experiments is a difficult, time consuming and hence expensive task and this is a principle reason why so little information has been published about reproduced image quality. However, we maintain that by measuring the quality of phantom images with a system that embeds approximate physical characteristics of a human head and then analyses data according to the dominant human localization mechanisms, that useful performance indicators of the system under test (SUT) can be extracted ([1], [2]).

Our attempt to extract all the relevant information available to the human auditory system will be described in Section 1. However, the final arbiter about the quality of the SUT cannot be just a set of measurements per se but must be filtered to target the perceptual clues derived from typical human listeners. To obtain a close matching between the measured data and the perception made by humans under the same listening conditions, a calibration process is required that encompasses each variable critical to human localization, e.g. early reflections, stability of wavefront etc.

Comparative listening experiments are performed under controlled conditions, with only one variable changing at a time, so that changes in perception can be evaluated. Then, by systematic measurement under the same experimental conditions using the VL, correlations between listening experiments and measured data can be identified and used for the calibration. The listening experiments planned for the calibration process will be described in Section 2.

1 Virtual listener

We maintain that for an optimum design, a VL should closely resemble a human listener both in terms of physical appearance and signal processing and should account for the major cues used for sound localization. For example, these include: Interaural Time Difference (ITD), Interaural Amplitude Difference (IAD), Spectral Cues and changes in ear signals under head rotation.

1.1 Sound field capturing

A first-order approximation of the human head is a rigid sphere with a radius of ~90 mm with the ears lying on the horizontal great circle at $\pm 100^\circ$ from due front. These values are used often in the literature, see for instance: Cooper and Bauck [3] or Blauert (Chapter 2) [4]. An investigation of the influence of head shape, diameter and ear position can be found in Rasmussen and Juhl [5]. By comparing amplitude and phase relationships of those two stereophonic loudspeaker feeds, which are required to produce ear signals which are the same as those of a plane wave at the ear positions, they compared the rigid sphere postulated by Cooper and Bauck with an axi-symmetric body. This body was chosen to have a cross section around the horizontal (right to left) axis which resembled the horizontal section, passing through the ear positions, of the Bruel & Kjaer 4128 head-and-torso simulator. They concluded that for ear positions at $\pm 100^\circ$ from due front a sphere with a radius of 68.4 mm is more appropriate as an approximation of the human

head. The outline of the head approximation actually used for sound field capturing can be found in Figure 1. It contains a cylinder of 48 mm height within two half spheres of 69 mm radius, to account for heads being higher than wide. The exclusion of pinnae from this head model will change the ear signals presented to the analysing program above approximately 4 kHz ([6], [7]) in comparison to those available with a dummy head including pinnae. But there are still spectral cues available to the model because the microphones are, at least in the horizontal plane, not symmetrically located. Three pairs of microphones (Sennheiser KE 4-211-2) at $+90^\circ/-110^\circ$, $\pm 100^\circ$ and $+110^\circ/-90^\circ$ from due front are used for capturing the sound field. This enables the sound field to be measured with the head turned either 10° to the right or to the left from the reference position without physically moving the head. This ability is important as rotation of the head is the most often observed movement when humans are asked to localize a sound source [4]. Actual impulse response measurements are done with MLSSA [8] and the captured data are then exported to a numeric computation program for post analysis.

1.2 Ear signal processing

Once the impulse responses of the three pairs of microphones for a given (phantom) source have been captured they can be convolved with any desired input signal to derive the stimuli for which actual perception data exist. Ideally, these stimuli should be processed in a way similar to the processing which takes place in the auditory system of humans. The angle independent frequency shaping function of the middle ear can be omitted for this purpose, however the important impedance matching function should be noted. Within the inner ear, the cochlea is responsible for our ability to perform pitch discrimination. The cochlea contains three fluid-filled channels where two are separated by the basilar membrane which is formed in a spiral configuration. The cochlea can be modelled as a non-uniform transmission line, where a

frequency-position mapping takes place on the basilar membrane [9]. With this feature, the cochlea performs a running frequency filter function with a bandpass characteristic and a slightly frequency dependent bandwidth [10]. For the purpose of this work the cochlea is modelled with 24 third-octave spaced discrete bandpass filters (critical bands) described by the following impulse responses (from [11]):

$$h(t) = (f_0 t - 1)^2 e^{-(f_0 t - 1)} \sin(2\pi f_0 t) \quad f_0 t \geq 1 \quad [1]$$

$$h(t) = 0 \quad f_0 t < 1 \quad [2]$$

with

$$f_0 = 99.21 \text{ Hz} * 2^{\frac{n-1}{3}} \quad \text{for} \quad n = 1..24 \quad [3]$$

The motion of the basilar membrane is transformed into electrical signals (sequences of neural impulses) via hair cells. The conversion from motion to neural impulse rate is non-linear and it is also time dependent with a higher conversion rate during signal onsets, exhibiting an adaptation time and a refractory time after cessation of input stimulus. Nevertheless, for the levels of the stimuli used in the localization experiments (~ 60 dBa) the process can be modelled as half-wave rectification [9] with subsequent first-order low-pass filtering where $f_{3dB} = 800 \text{ Hz}$ [11]. We now have appropriate signals which would be available to the human brain for localization. Many models exist which attempt to describe how the brain processes these signals to extract information that includes localization, where for further information the reader is referred to the overview given in [12]. However, for the purpose of this work each auditory cue used for localization will be analysed independently, where a block diagram of the ear signal processing software is given in Figure 2.

1.2.1 Estimation of IAD

IADs are a consequence of the head forming an acoustical shadow for sound having wavelengths comparable to or smaller than the head [7]. A measure of the IAD can be calculated with:

$$IAD(t_2) = 10 \log \left(\int_{t_1}^{t_2} L^2(t) dt / \int_{t_1}^{t_2} R^2(t) dt \right) \quad [4]$$

A number of overlapping and uniformly spaced time intervals $t_2 - t_1$ have to be evaluated independently and weighted according to their importance for localization (e.g. concerning the precedence effect [13]) to estimate a final value. This can be compared against a map of free-field localization IADs, to make a direction estimation, where the map required for horizontal directional judgements is shown in Figure 3. It should be noted that there may be some ambiguity, one direction estimation will be found when only the frontal hemisphere should be investigated (e.g. intensity stereo), two possible direction estimations for a 360° investigation (e.g. horizontal Ambisonics) and a "cone of confusion" [14] for a full sphere investigation.

1.2.2 Estimation of ITD

ITDs play a dominant role in image localization because a sound source located to the left will produce a soundwave that reaches the left ear sooner than the right one, the amount of time delay is mainly dependent on the angle between the median plane through the listener and the sound source. ITDs are used up to 2 kHz while from above 150 Hz the ITDs of the two signals envelopes become progressively more important, representing the only ITD cue above 2 kHz.

The time difference between two similar signals L and R can be

calculated using cross correlation functions in the form of [11]:

$$\Psi_{L,R}(t, \tau) = \int_{-\infty}^t R(v) L(v-\tau) W(t-v) dv \quad (L \text{ leading } R) \quad [5]$$

or

$$\Psi_{R,L}(t, \tau) = \int_{-\infty}^t L(v) R(v-\tau) W(t-v) dv \quad (R \text{ leading } L) \quad [6]$$

$W(t-v)$ is a weighting function used to prevent an unreasonably high contribution of correlation products from past data. For the purpose of this model the cross correlation function is normalized, so it can become unity at maximum when both signals are equal except of a time delay. The time delay for the maximum of the cross correlation function is calculated for equally spaced time steps as long as the stimulus is present, and the values are stored together with the values of their magnitudes. The magnitude is supposed to be one measure of how sharp an image will be focused. Figures 4 to 6 show solutions of the cross correlation function for a single sound source under free-field condition stimulated with an impulse. Contributions of each time step can be weighted according to their importance on localization and then summed to a final time delay, where additional discussion follows in Section 2. This can be compared to a map of pre-learned time delays for free field localization of a single sound source, and an example map is shown in Figure 7. Again ambiguities similar to those mentioned with IADs exist.

1.2.3 Estimation of spectral cues

To incorporate spectral cues for localization a pattern recognition process is necessary. The pattern to be recognised can be created by constructing a map of the difference of the average power between both ears for each critical band using a

broadband excitation of the virtual listener. This has to be performed for a number of equally spaced directions. For the stimulus under investigation the same power differences can be calculated. The resulting set of 24 power differences is then compared to the entries of the map by a mean-square error algorithm, and the entry leading to the smallest error is the direction decision. This part of the model is so far incomplete and is pending further research.

1.2.4 Localization cues from head rotation

As well as for providing two additional sets of IAD, ITD and spectral cue information, head rotation has another important function in localization. For a sound source located left/front of the listener, turning the head to the left will lower IAD and ITD while if the source is left/back then IAD and ITD will increase. In this way head rotation can be used to resolve ambiguities in IADs and ITDs. It is further believed that the instability of the position of the phantom source under head rotation is a direct measure of its diffuseness.

1.3 Direction estimate processing

The final direction estimation requires initial identification of the quadrant of the phantom image, which for horizontal localization can be deduced by a majority decision over all changes in IADs and ITDs under head rotation. By this process most if not all directional ambiguities of single results can be resolved. The final azimuth can now be calculated as a weighted average taken over all estimates. Weighting should be applied according to the reliability the estimate has in relation to the other factors, for instance IADs have low reliability at low frequencies because of their low angle-dependency (see Figure 3). Finally, from the deviation of all the estimates derived from the final azimuth, a diffuseness parameter is calculated.

2 Formal experiments

It is the aim of the formal experiments to investigate those variables which are known to have an influence on localization accuracy and image sharpness. The general procedure of the localization experiment has already been described in an earlier paper [15] which presented a comparison between a reference source, either real or phantom, and the phantom source under evaluation. In this case the only variable significant to localization which is changed between reference and phantom is the variable under evaluation. The experiment in some cases requires an adjustment procedure, where the subject is asked to horizontally shift the phantom source to the position of the reference and also a comparison procedure, where the subject is asked to judge a number of acoustical properties of the phantom source in comparison to the reference. Important properties to be judged here are: direction (in some of the experiments) and diffuseness of image.

The experiments should be conducted in a room which is representative of a good listening room, as the aim is to use the VL to judge the quality of SUTs in the environment for which they are designed.

2.1 Stability of wavefront

An ideal sound reproduction system would be able to create a sound field which is defined for any point within a restricted volume. For $\lambda/8$ precision at 20 kHz 102 points, and hence 102 channels, per 1 cm^3 must be specified [16], which is too excessive for any practical application. A more relaxed demand of a sound reproduction system would require it to reconstruct a plane wave at least around the circumference of a human head for every direction. Even this system would need one loudspeaker every 10° for all components not assignable to a plane wave

being at least 20 dB below the plane wave component for frequencies up to 5 kHz, and even more loudspeakers above this frequency [17].

A plane wave radiating from the intended direction would then ensure that all the interaural intensity and phase informations from this direction are those of a distant sound source. To investigate the stability of wavefront a symmetrical three loudspeaker layout where the centre speaker is also the reference can be used. The sound radiation of the centre speaker can be assumed to produce a plane wave at the listener and with appropriate signals fed to the side loudspeakers, which still have to be calculated, a fixed departure from this can be simulated where this departure should be representative for some typical loudspeaker layouts. It is particularly important that the three loudspeakers are arranged in the room in a way to avoid early reflections which otherwise would adversely influence localization. This is possible with the layout shown in Figure 8, for which azimuth reflections are more than 5 ms delayed and thus have no influence on localization, see Blauert (chapter 4) [18]. However, the unavoidable floor and ceiling reflections which are in the sub 5 ms range are, according to experiments described in [19], uncritical. At least two angles of incident, the most uncritical and the most critical in this respect, should be investigated, centre front and 90° from centre front.

2.2 Early Reflections

In most cases early reflections are not perceived as discrete sources of sound, but merge with the original sound source when they arrive after delay times shorter than 5 ms. If they are delayed by more than 5 ms and emanate from lateral positions they can enhance the feeling of spaciousness and envelopment of the sound field [20], [21], but regardless of direction they will not influence the localization of the original sound

source. The strategy behind the suppression of localization information arriving later than 5 ms after the onset of a new sound sensation is straightforward: It excludes wrong localisation in reflective environments.

The median plane symmetric layout of Figure 9 allows an investigation into the influence of early reflections. For one pair of loudspeakers, in this case LA and RA, the walls leading to early reflections are covered with highly absorptive damping material, the mirror image walls for the other pair (LB and RB) are maintained as reflective. Now for LA and RA with the gains adjusted to produce a virtual image approximately at centre front, act as a reference against which the other pair, LB and RB is judged. The experiment has to be conducted in two stages for every boundary condition, where in the second stage signals and damping are changed between RB and LA and between LB and RA to cancel out residual asymmetry in the hearing capability of the subjects and for any calibration errors in adjusting the phantom image of the reference pair exactly to centre front. While the layout should be fixed in regard of subject-to-loudspeaker position, the whole experiment should be conducted at different places within the room, to allow for the influence of reflections from different directions to be investigated independently.

2.2.1 Early reflections < 2.5 ms

Early reflections occurring less than 2.5 ms after the original sound sensation are known to shift the image towards their direction and to blur the image. By the experiments described above, we try to identify the time weighting curves for IADs and ITDs which give correct predictions of the subject's directional responses together with predictions how the spreading of IADs and ITDs over time and critical band will lead to an increased diffuseness.

2.2.2 Early reflections < 5 ms

Early reflections occurring more than 2.5 ms but less than 5 ms after the original sound sensation are known to blur the image, although they keep the direction of the image constant [18]. However true this assertion, it might be speculated that a higher reliability rating is assigned to images which deliver the same cues beyond the 2.5 ms "border", so the influence of these reflections will be incorporated into the diffuseness rating.

2.3 Reverberation

The influence of reverberation on our ability to localize sounds is obvious to anybody who has ever listened carefully within a highly reflective environment, for instance a church. The influence of the energy of the reverberant sound field in comparison to the energy of the direct sound field can be evaluated approximately by using a monopole and a dipole loudspeaker to represent the different sound field components. The monopole loudspeaker is placed near the subject to ensure that its direct energy dominates its reverberant energy, while a dipole speaker is placed about 1.7 m behind the monopole speaker in such a way that the side lobe (i.e. a null) is facing the listener. By feeding either the monopole alone or the monopole with an amplitude reduced signal and the dipole with an amplitude reduced and time delayed signal, the direct-to-reverberation energy ratio can be changed and compared without a change in loudness or early reflection pattern. This experiment assumes that not just the reverberation time but the direct-to-reverberant energy ratio has a major impact on localization.

As the stimuli used for the experiments are repetitive, a way to incorporate the influence of reverberation into our model is to overlay and add the impulse response of the direct sound field with those elements of the impulse response where the first, second, third,.... etc. repetitions occur and then to analyse

the resulting impulse response.

3 Conclusions

A method has been presented which allows assessment spatial accuracy of sound reproduction systems by measurement. This method includes measuring the sound field with an approximation to a human head and then post-processing the measurements to extract those quantities which are available to the brain for localization purposes. An azimuth and a diffuseness estimate can be calculated from these quantities, making the device a Virtual Listener.

A number of independent and significant parameters have been introduced, which are known to have a direct influence on our perception of phantom sources. For each of these parameters, a listening experiment has been described which allows investigation and assessment without significant parameter interaction. By comparing the changes in output data of the localization model with the changes in perception of humans under the same conditions, the Virtual Listener can be calibrated. We believe that a way to physically measure and quantify the quality of phantom images is a worthwhile step towards a more objective and less time consuming evaluation of both present and future sound reproduction standards.

4 Acknowledgment

We would like to thank Joachim Gerhard, and Audio Physic Gerhard GmbH, for the financial support of this work.

5 References

- [1] C.J. Mac Cabe & D.J. Furlong, "Virtual Imaging Capabilities of Surround Sound Systems," J. Audio Eng. Soc., vol. 42, no. 1/2, pp. 38-49 (1994)
- [2] E.A. Macpherson, "A Computer Model of Binaural Localization for Stereo Image Measurement," J. Audio Eng. Soc., vol. 39, no. 9, pp. 604 - 622 (1991)
- [3] D.H. Cooper & J.L. Bauck, "On Acoustical Specification of Natural Stereo Imaging," Preprint 1616 of the 65th Audio Engineering Society Convention, Vienna (1980 Feb.)
- [4] J. Blauert, "Räumliches Hören," S. Hirzel Verlag, Stuttgart, (1974)
- [5] K.B. Rasmaussen & P.M. Juhl, "The Effect of Head Shape on Spectral Stereo Theory," J. Audio Eng. Soc., vol. 41 no. 3, pp. 135-142 (1993)
- [6] A.D. Musicant & R.A. Butler, "The Influence of Pinnae-Based Spectral Cues on Sound Localization," J. Acoust. Soc. Am., vol. 75, pp. 1195-1199, (1984 Apr.)
- [7] D.H. Cooper, "Problems with Shadowless Stereo Theory: Asymptotic Spectral Status," J. Audio Eng. Soc., vol. 35, no. 9, pp. 629-642 (1987)
- [8] D.D. Rife & J. Vanderkooy, "Transfer Function Measurement with Maximum-Length Sequences," J. Audio Eng. Soc., vol. 37, pp. 419-444 (1989 June)
- [9] M.R. Schroeder, "Models of Hearing," Proc. IEEE, vol. 63, pp. 1332-1350 (1975 Sept.)
- [10] Various Authors, "Die Musik in Geschichte und Gegenwart," vol. 3, pp. 1109-1111, second edition, (1995)
- [11] J. Blauert & W. Cobben, "Some considerations of binaural crosscorrelation analysis," Acoustica 39, pp. 96-104 (1978).

- [12] H.S. Colburn & N.I. Durlach, "Models of binaural interaction," in: E.C. Carterette and M.P. Friedman (Eds.), "Handbook of Perception," vol. 4, pp. 467-518, Academic Press, New York, (1978).
- [13] P.M. Zurek, "The Precedence Effect," in "Directional Hearing", Springer, New York, (1987)
- [14] S.A. Gelfand, "Hearing: An Introduction to Physiological and Psychological Acoustics," Marcel Dekker, New York, (1981)
- [15] B. Theiß & M. O. J. Hawksford, "Localization Experiments in Three-Dimensional Sound Reproduction," Preprint presented at the 100th AES convention, Copenhagen (1996 May)
- [16] A. Romano, "Three-Dimensional Image Reconstruction in Audio," J. Audio Eng. Soc., vol. 35, no. 10, pp. 749 - 759 (1987)
- [17] J.C. Bennett, K. Barker & F.O. Edeko, "A New Approach to the Assessment of Stereophonic Sound System Performance," J. Audio Eng. Soc., vol. 33, no. 5, p. 314 - 321 (1985)
- [18] J. Blauert, "Räumliches Hören - Nachschrift," S. Hirzel Verlag, (1985)
- [19] B. Rakerd & W.M. Hartmann, "Localization of Sound in Rooms, II: The Effect of a Single Reflecting Surface," J. Acoust. Soc. Am., vol. 78, pp. 524-533, (1985 Aug.)
- [20] Y. Ando, "Concert Hall Acoustics," Springer, New York, (1985)
- [21] D. Griesinger, "Spaciousness and Localization in Listening Rooms and Their Effects on Recording Technique," J. Audio Eng. Soc., vol 34, pp 255-267, (1986 April)

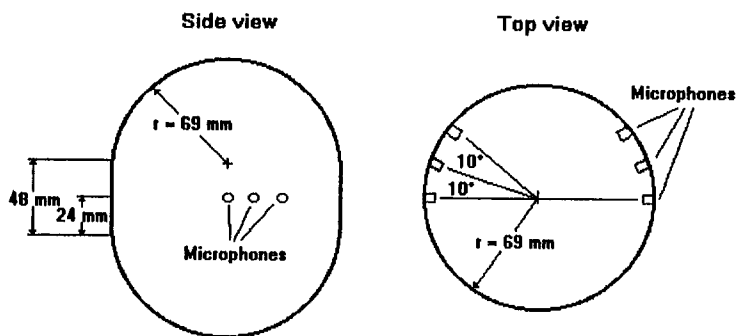


Figure 1 Approximation of the human head used for sound field measurements

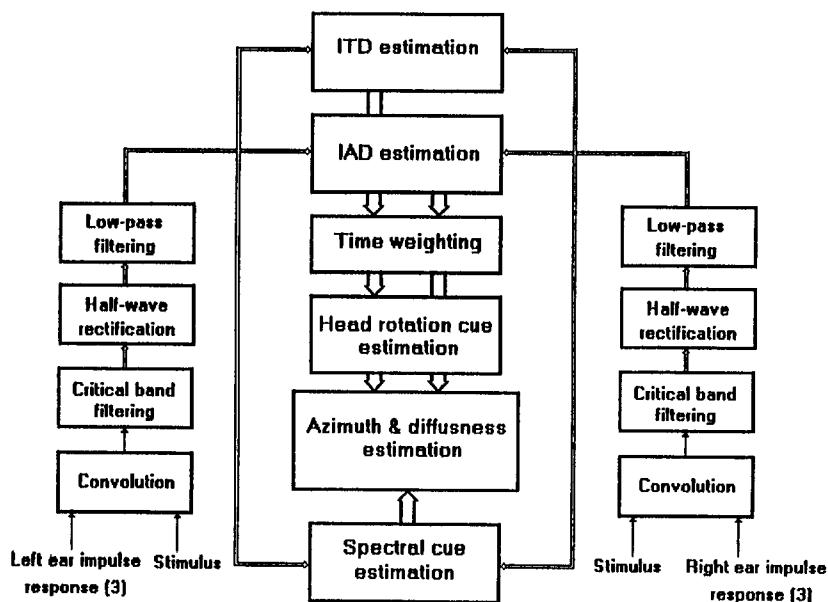


Figure 2 Block diagram of ear signal processing software

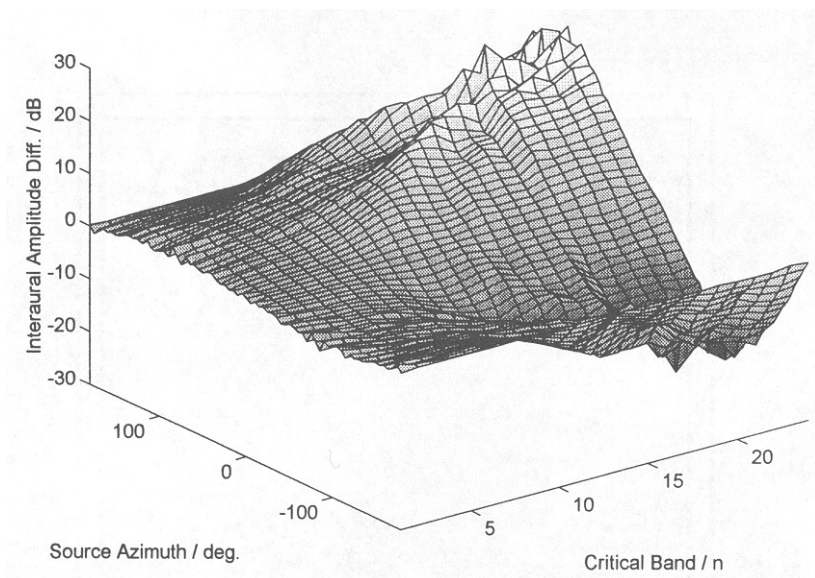


Figure 3 IAD versus angle of incident and number n of critical band

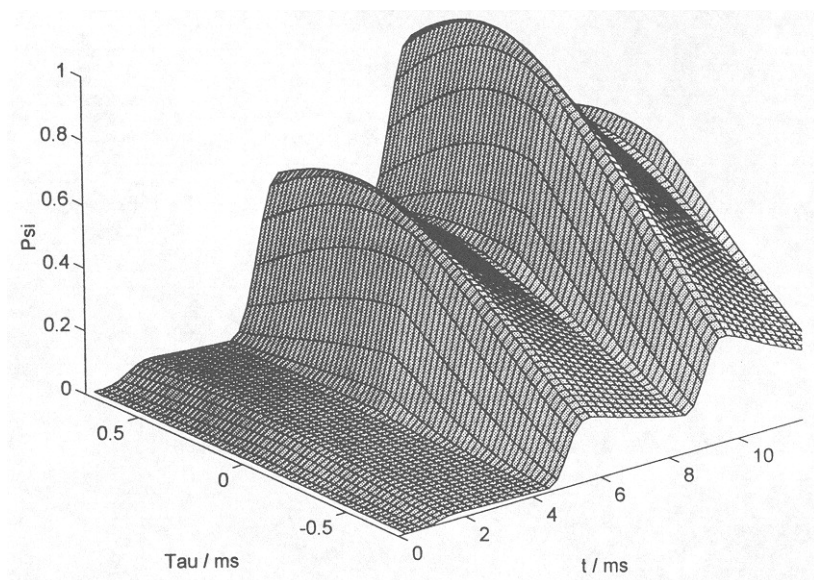


Figure 4 Solution Ψ (Ψ) of normalized cross correlation function versus time displacement τ (τ) and time t for critical band $n = 5$ (250 Hz), source at 30°

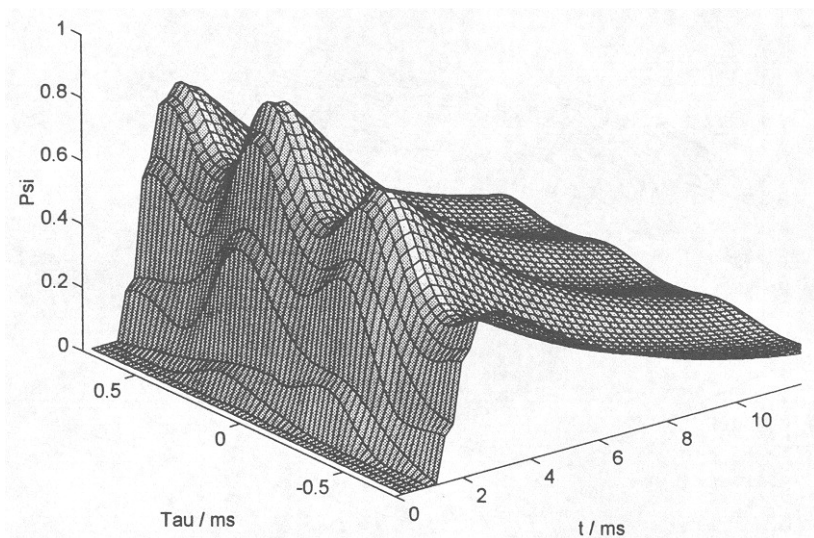


Figure 5 Solution Ψ (Psi) of normalized cross correlation function versus time displacement τ (Tau) and time t for critical band $n = 14$ (2 kHz), source at 30°

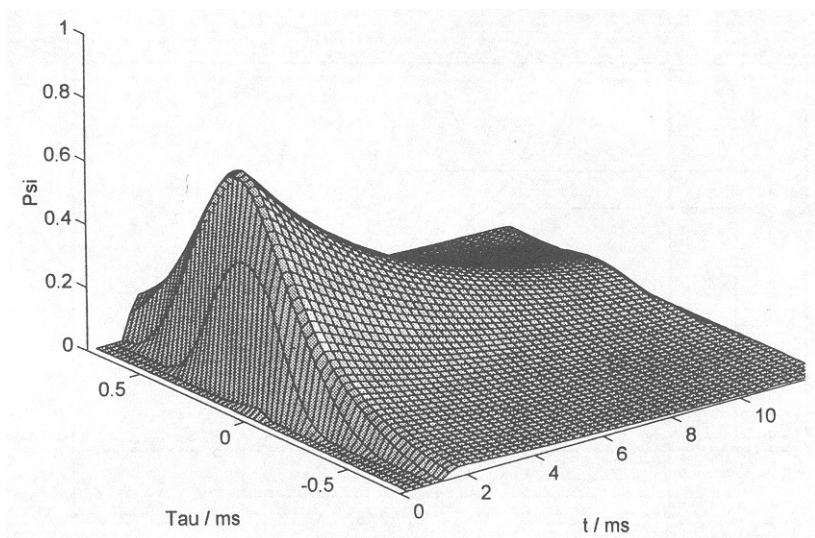


Figure 6 Solution Ψ (Psi) of normalized cross correlation function versus time displacement τ (Tau) and time t for critical band $n = 23$ (16 kHz), source at 30°

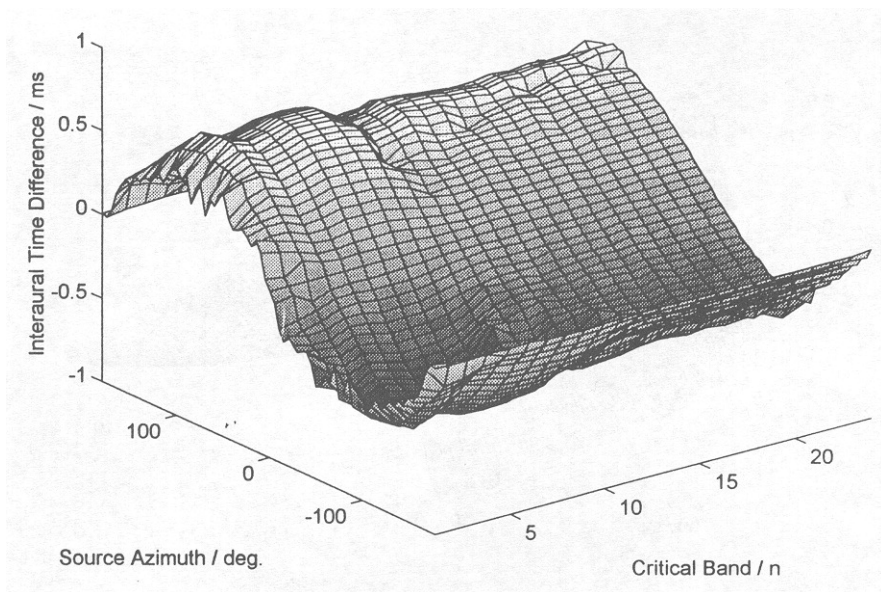


Figure 7 ITD versus angle of incident and number n of critical band

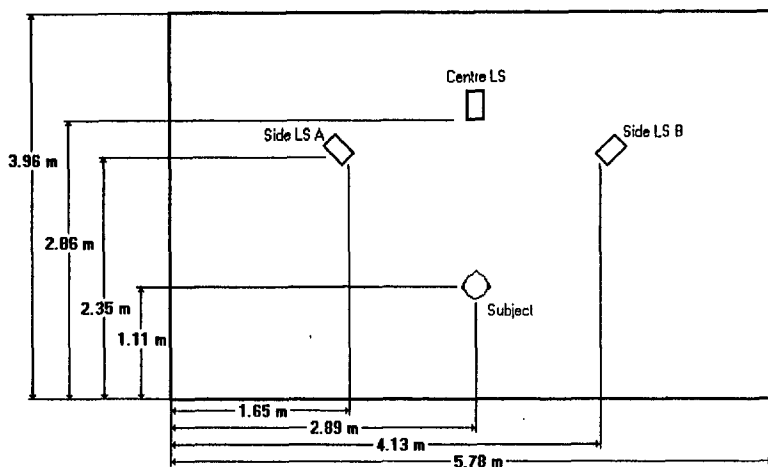


Figure 8 Loudspeaker layout for the investigation into localization of non-plane waves

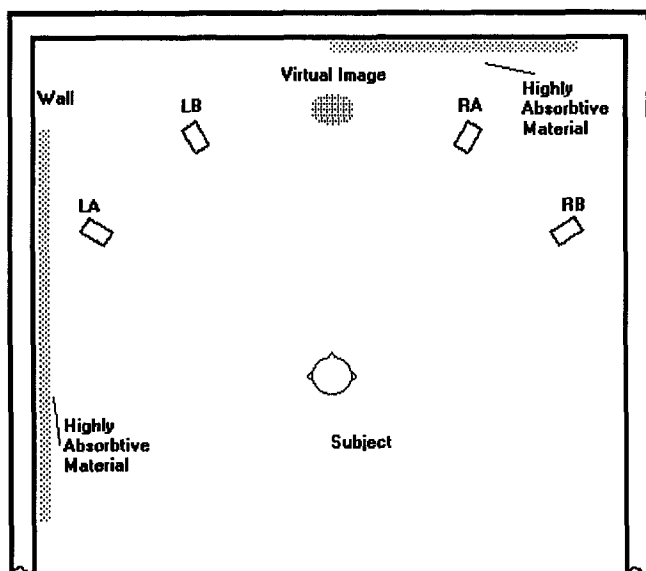


Figure 9 Sketch of loudspeaker layout for the investigation into the influence of early reflection on localization