Limitations of Dynamically Controlling the Listening Position In a 3-D Ambisonic Environment

Preprint 4460 (I3)

D. M. Burraston, M. P. Hollier, M. O. Hawksford
BT Laboratories, Ipswich, Great Britain
University of Essex, Colchester, Great Britain

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

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.

AN AUDIO ENGINEERING SOCIETY PREPRINT

Limitations of dynamically controlling the listening position in a 3D ambisonic environment.

D M Burraston (1), M P Hollier (1), M O Hawksford (2)
(1) BT Laboratories, Martlesham Heath, IPSWICH. IP5 7RE
(2) University of Essex, Wivenhoe Park, COLCHESTER. CO3 4SQ

Abstract

Telepresence is an important step in the evolution of communications: the facility to meet and interact in a shared virtual space. Work at BT Laboratories on spatial audio includes the use of 3D audio in interactive video environments (IVE's). A video image of the user provides location and gesture inputs to the system. This paper describes work to implement and assess the dynamic control of the spatial audio sweet-spot to match user location.

1.0 Introduction

The need to interact with remote environments has resulted in the development of telepresence and teleoperated systems. A sense of telepresence is achieved when a person receives sufficient, naturally displayed, information about a remote environment. Telepresence can enable an individual to feel and interact as if they were present in a remote environment. Visual, auditory and force sensations are all relevant to achieving perceptually compelling telepresence. The work described in this paper is part of BT Laboratories on-going research into this important emerging technology.

This paper examines the performance a 3D ambisonic sound system when the listener position is modified to accommodate movement within the reproduction loudspeaker array. For a description of ambisonics see Felgett [1], for the theory of periphonic reproduction see Gerzon[2], and Bamford [3] for wavefront mismatch errors.

It is expected that performance will degrade as the user moves away from the sweet-spot of an ambisonics system. However, if the position of the user is known it is possible to update a digital ambisonics decoder to form the sweet-spot at the users new position. The spatial audio performance can still be expected to degrade as the steered sweet-spot position becomes further away from the centre of the loudspeaker array, e.g. due to proximity to the loudspeakers, and the influence of loudspeaker array asymmetry.

In an experimental interactive video environment at BT Laboratories, user location is tracked using captured video data and used to update the sweet-spot position of the ambisonic system. In this way it is hoped that a useful working area can be created within a practical loudspeaker array for a dynamically controlled IVE spatial audio system. A subjective experiment to determine the size of the working area where the spatial audio remains perceptually compelling was conducted and the results are reported below.

2.0 Spatial audio environments

Most of the existing virtual environments under development at BT Laboratories include spatial audio. The audio is usually the voice of another user, but it may be sounds triggered by data or events. The environments are viewed using large monitors, video screens, video walls or large overhead projections inside domes. Interaction within these environments can take the form of a virtual/video conference, or an IVE where the user interacts with gesture and movement. In these environments (apart from users seated at personal monitors) the user requires the facility to move around, therefore spatial audio solutions must cater for this.

2.1 Spatial audio technology

Spatial audio is a vital component in many virtual environment and telepresence applications. In particular it provides a valuable navigation aid, alerting a user that either another person or a data object of interest is nearby; perhaps right behind them. In order to recreate perceptually compelling spatial audio a number or reproduction systems have been investigated:
(i) Transaural cancellation for single users located in front of a static monitor/display,
(ii) Ambisonics decoded to four loudspeakers around a single user, allowing greater freedom of movement, particularly head rotation,
(iii) Ambisonics decoded to multiple loudspeaker arrays for single users in large spaces, e.g. large immersive environments with interactive content, and
(iv) Panned decode of direct sound with ambisonic decode of reverberent sound for multiple users in large spaces.

Ambisonics decoded to multiple loudspeakers has been found to offer compelling spatial audio for a single stationary user. When there are multiple users in a space ambisonics is generally unsatisfactory due to the single sweet-spot. In order to provide compelling spatial audio for multiple users a hybrid solution with panning of direct sounds and ambisonically reproduced reverberation has been found to be successful

Since ambisonics is deemed to provide the most compelling spatial audio it would be the natural choice for a single user in an IVE, providing the user with excellent spatial audio while they are located at, or very near, the sweet-spot. A virtual sound source can be located anywhere around the user who can use natural small movements of the head in order to locate the sound. However, the performance of the system degrades noticeably once the user moves more than about 0.25m away from the sweet-spot in a loudspeaker array approximately located on the surface of a notional 4.5m diameter sphere. In an IVE location, information is available on the user location and it is logical to try and steer the ambisonic sweet-spot to the user location in order to maintain performance.

An IVE system at BT Laboratories was adapted to return location data provided from the video to the digital ambisonics decoder. The IVE system included; a Lake DSP Huron system and software, person finder software, a purpose built room with a large (4m by 3m) rear projection screen and network software to manage the transmission of geometry and control data. It is expected that the performance of the spatial audio reconstruction will degrade when the sweet-spot is generated away from the centre of the loudspeaker array for a number of reasons:
(i) nearfield effects, when in close proximity to loudspeakers
(ii) increased variation in level due to small movements when in close proximity to loudspeakers due to rising pressure gradient.
(iii) the array becomes increasingly asymmetric which can be expected to exaggerate the affect of any minor errors in the physical array or related decoder parameters.

In view of the above it was expected that there will be a usable region within the loudspeaker array in which the user location can be steered while maintaining spatial audio performance, and that beyond this region performance will degrade.

3.0 Subjective test on steerable spatial audio system

A subjective test was designed to provide evidence about the extent to which the ambisonic sweet spot could be moved. Identification of the direction of a series of stationary sounds is a critical test of a spatial audio system since it is subjectively more difficult than recognising the trajectory of a moving sound. Identification of the direction of a number of stationary sounds was therefore chosen to provide a critical metric for system performance. Initial informal tests were conducted which allowed the placement of the virtual source and user position at any desired location. These initial tests confirmed that the left-right symmetry of the speaker array allowing the test area to be reduced by half. The front and back of the room is not symmetrical and so the subjective tests included both these areas.

3.1 Experiment design

A forced-choice experiment using discrete sound sources and subject locations was chosen to determine the systems performance. It was decided that the system performance would be deemed unsatisfactory when the discrete locations of a set of virtual sound sources could no longer be reliably determined. The experiment was conducted in an acoustically treated room detailed in Fig 1. The experiment was conducted in conditions representative of actual use, which included air conditioning/computer equipment noise. The background noise level, measured at the centre of the speaker array, was 43.8dB(A).

The sound source chosen for the experiment was a speech. This contains suitable frequency content for spatialisation and is an important signal type for many applications under. A spectral plot of the speech signal is shown in Fig 2. The speech was taken from a corpus of level normalised speech data recorded at 16kHz, and contained two unrelated sentences : "I never knew he liked music" and "He rode down the country lane".

Six virtual sound source locations were chosen whose positions are set at the subject's ear height. The virtual sound source locations were relocated with the person when their location changes, i.e. the virtual sources do not remain at fixed. The six sound source locations chosen were : front left 45 degrees, left 90 degrees, rear left 135 degrees, front right 45 degrees, right 90 degrees, rear right 135 degrees. All the virtual sound sources were at a fixed distance of 1.5 meters in the horizontal plane. Six subject positions where chosen, based on the initial informal tests, and their relationship to the speakers is shown in Fig 3. The co-ordinates of the subject positions in metres with positive Y towards the screen were :

A	X= -1	Y= 1
B	X= -1	Y= 0
C	X= -1	Y= -1
D	X= 0	Y= 1
E	X= 0	Y= 0
F	X= 0	Y= -1

For each subject the combination of 6 sound source positions and 6 subject positions gave a total of 36 combinations. To test the influence of subjects getting used to the environment and test system, all conditions were presented twice. The order of presentation to the subject was selected using a Latin Square method, see Brownlee [4]. There were 12 subjects used in the experiment, six male and six female.

Subject responses were collected using a custom experiment control program, employing a visual representation of the experimental area. The subjects entered their choice of source direction, using a trackball, on a large graphical representation of the experimental set-up. These responses were stored automatically. Fig 4 shows a subject in the test area with the trackball. Fig 5 shows the screen image presented to the subject. This is a representation of a person, with 6 cubes around their head to signify the sound sources. At the start of the test the subject stands at one of 6 locations marked on the floor by A, B, C, D, E or F as directed the program. Once positioned on the appropriate spot, the subject used the trackball to select OK on the screen and was then presented with a short section of speech. The subject then selects one of the 6 green squares on the screen which they think best corresponds to the direction of the speech. The subject presses SUBMIT to confirm their choice. If, for any reason, they are unsure of the direction of the speech they press the UNSURE button. After pressing SUBMIT or UNSURE they are either prompted to move to a new location, in which case they move to the new location and then press OK, otherwise the next sound is presented.

3.2 Experiment results - overall performance

Analysis of variance on the results verified that the ease of correctly identifying the sound source is dependent on subject position. A visualisation of the responses for all subject positions is shown in figs 6 to 11. The sound source is represented by a sphere and the location of loudspeakers by cubes. Black lines represent correct responses, dark grey lines represent incorrect responses and light grey lines represent unsure responses. The length of the line represents the proportion of subjective responses in each category. A brief description of these results follows :

Position A
Front left 45 degrees mainly correct.
Left 90 degrees mainly incorrect.
Rear left 135 degrees just over half correct, the rest split between unsure and incorrect.

Front right 45 degrees virtually identical to left 90 degrees.
Right 90 degrees mainly correct.
Back right 135 degrees mainly incorrect, unsure being a high proportion of the remaining responses.

Position B
Front left 45 degrees just over half correct, the rest mainly incorrect.
Left 90 degrees mainly correct.
Rear left 135 degrees virtually identical to front left 45 degrees.

Front right 45 degrees unsure and high proportion incorrect.
Right 90 degrees virtually identical to left 90 degrees.
Rear right 135 degrees mixed result.

Position C
Front left 45 degrees just over half correct.
Left 90 degrees mainly incorrect.
Rear left 135 degrees all correct.

Front right 45 degrees mixed unsure and incorrect response.
Right 90 degrees mainly correct.
Rear right 135 degrees virtually identical to left 90 degrees.

Position D
Front left 45 degrees mainly correct.
Left 90 degrees just over half correct, remainder unsure.
Rear left 135 degrees almost half correct, remainder split between unsure or incorrect.

Front right 45 degrees just over half correct, remainder mainly unsure. Almost symmetrical along Y axis apart from the front right which shows a higher degree of uncertainty.
Right 90 degrees virtually identical to left 90 degrees.
Rear right 135 degrees identical to rear left 135 degrees.

Position E
The system sweet spot and should have had the best results. All positions were mainly correct except rear right 135 degrees, which was only correct for 13 out of the 24 responses.

Position F
Front left 45 degrees mainly correct, a quarter unsure.
Left 90 degrees mainly correct.
Rear left 135 degrees half correct, remainder mainly incorrect.

Front right 45 degrees mainly correct.

Right 90 degrees mainly correct.
Rear right 135 degrees almost half incorrect, remainder mainly correct.

3.3 Experiment results - individual locations

Due to the number of results, 864 responses in total, it was decided to use a custom built program to present the subject responses for individual locations. This program allowed the creation of animations for all the results in various configurations of subject, response and subject and sound position. A visualisation of the responses for 6 examples from the 36 possible sound sources is shown in Figs 12 to 17. In these diagrams, the incorrect location choices made by the subjects are also shown. Unsure responses are shown in light grey and pointing towards the intended sound source. A brief description of each representation follows:

1. **Position A Left 90 Degrees** : Almost all responses collapsed to the front left speaker.
2. **Position A Front Right 45 Degrees** : Almost all responses collapsed to front left speaker or were influenced on the left side.
3. **Position B Front Right 45 Degrees** : Just over half unsure, the rest composed of all the same incorrect responses and three correct responses.
4. **Position C Left 90 Degrees** : Almost all responses collapsed to the rear left speaker.
5. **Position E Front Left 45 Degrees** : Almost all responses correct at the sweet spot.
6. **Position F Rear Right 135 Degrees** : Incorrect responses from the right and front right. Errors were influenced on the right hand side.

More tests using the same method are planned in order to extend the data set and compare different batches of subjects responses.

4.0 Discussion and conclusions

Allowing the user to move while maintaining good spatial audio performance is necessary in some IVEs. Locating the spatial audio sweet-spot at the user position, which can be determined from a video image, will provide improved spatial audio performance. In order to assess the value of this development knowledge of the limits of the system is required in terms of listener position and its effect on sound source localisation performance.

Performance depends on subject position and sound source location. The results for positions on the centre line of the room were better, i.e. positions D, E, F, due to being further away from the speakers. These 3 positions have shown less subjective error than positions at A, B and C. Positions D and F fail most at the two rear positions. Informal listening at D and F, and questioning of some of the test subjects, indicate that in some cases multiple sound sources were heard. Positions A and C both collapsed to the speaker when failing at the positions mentioned in 3.3 above. Position B failed mainly on the front and rear right sound sources. Further informal listening at B resulted in multiple sound sources.

It may be possible to maintain a degraded working region at A, B and C. This would require source locations near the regions of correct responses only e.g. Position A at front left 45 degrees, back left 135 degrees and right 90 degrees, although this may be of limited use in an IVE/teleconference. If we are to allow the user complete freedom of movement within the space, the current working volume must be extended. Possible directions for research are :

1. Larger dimensions for loudspeaker array.
2. Turn-off/turn-down nearest loudspeaker.
3. Re-specify the loudspeaker decode algorithm to reduce the dependency on nearest loudspeaker.

Results at the sweet spot were not perfect. In a practical system we must accept that we will not get the perfect theoretical response. It may be possible to improve the response at the sweet spot and this will be investigated.

Testing subject performance to identify specific source directions is a demanding test. In practice, a good impression of spatial sound was maintained in most positions when a virtual sound is moving. This is due to the far more robust percept created by a moving source. However, positions A and C and their symmetrical equivalents still suffered a noticeable loss of performance for moving sources.

Further work will be pursued to extend these results and other experiments could take the form of :

1. Source locations which remain independent of subject location.
2. Source location at different distances and/or heights.
3. Moving sound sources
4. Different locations within working space e.g. closer to speaker boundaries and half metre distances.
5. Investigation into other subjective responses e.g. how "sharp" the directionality was, how well the distance was perceived.

A subjective test of the listening area of our ambisonics system has been conducted to examine the practical limitations of the system and results from initial tests have been presented. A useful working region for IVE/telepresence applications has now been identified and reasons for this discussed. These results will help us to assess the benefit of dynamically controlling the listener location. The temporal performance of this will be affected by the speed of our motion tracking software, the network link to the ambisonics system and the ability of the ambisonics system to respond in real time to these requests. This temporal performance will be the subject of a further subjective test.

5.0 Acknowledgements

The authors would like to thank their colleagues at BT, especially Barry Eggleton, and Rupert Voelcker for their help and support. Thanks are also due to Lake DSP for their contribution to the audio processing software for spatial audio decoding.

6.0 References

[1] Felgett P,"Ambisonics. Part One: General System Description", Studio Sound, 1:20-22,40. August 1975

[2] Gerzon M A,"Periphony: With-height sound reproduction", Journal of the Audio Engineering Society, Vol.21, Number 1, Jan./Feb. 1973.

[3] Bamford J S,"An Analysis of Ambisonic Sound Systems of First and Second Order", Msc. Thesis presented at the University of Waterloo, Ontario, Canada. 1995

[4] Brownlee, K.A. Industrial Experimentation, Chemical Publishing Co., Inc. 1953.

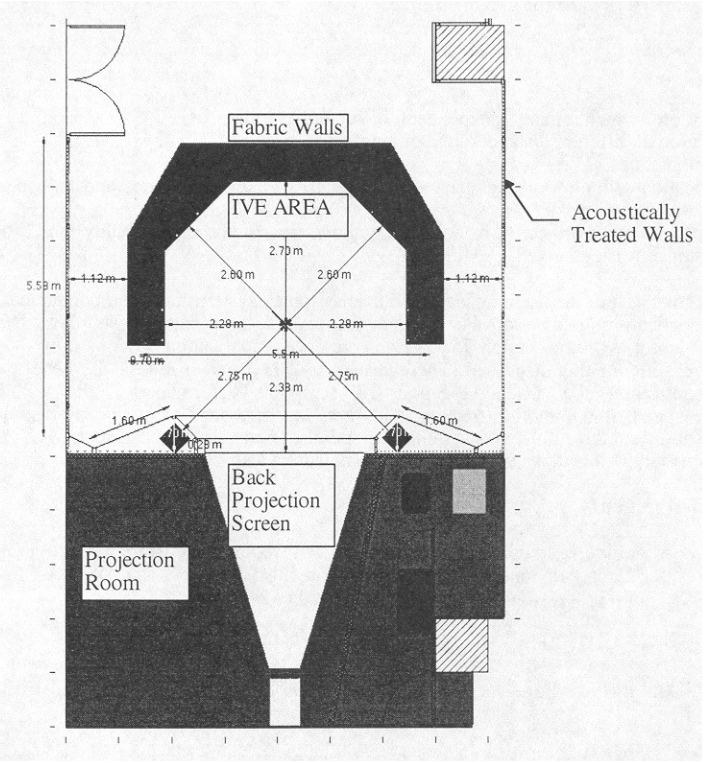


Fig 1. Detail of room layout.

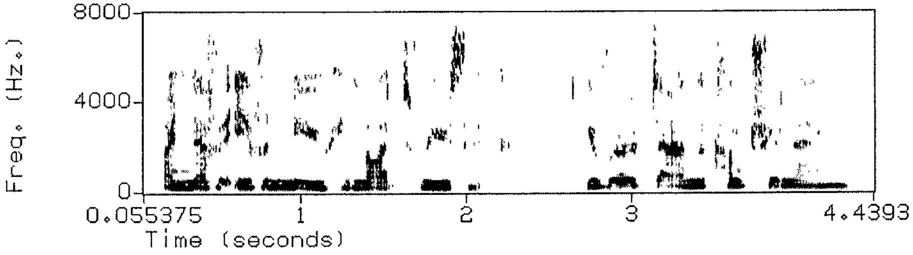


Fig 2. Spectral plot of test speech signal.

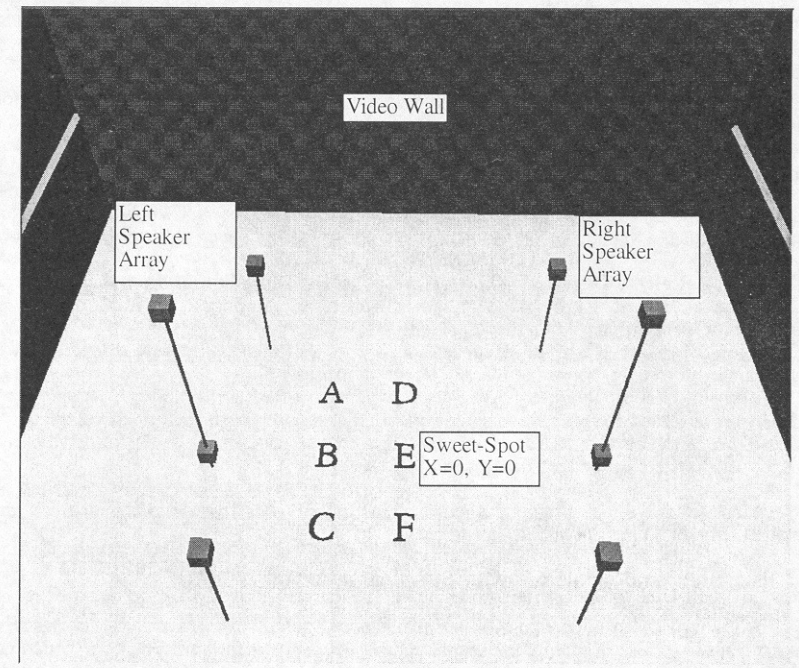


Fig 3. Listener positions and speaker locations

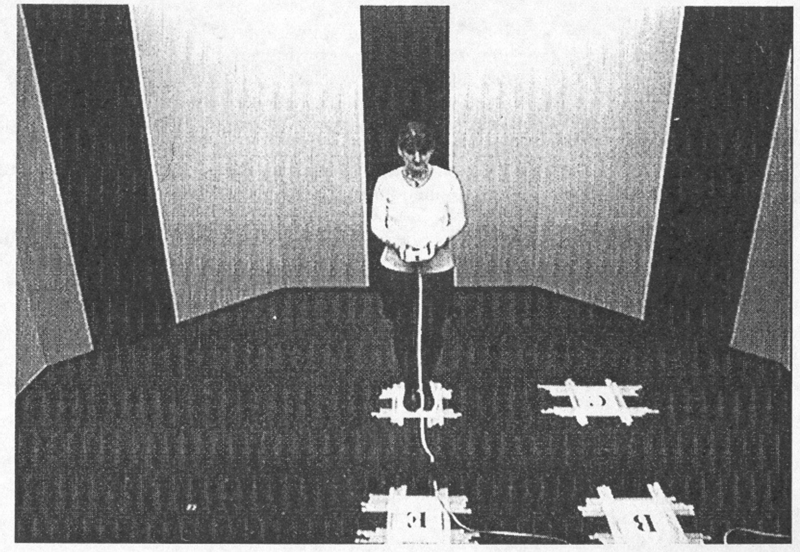
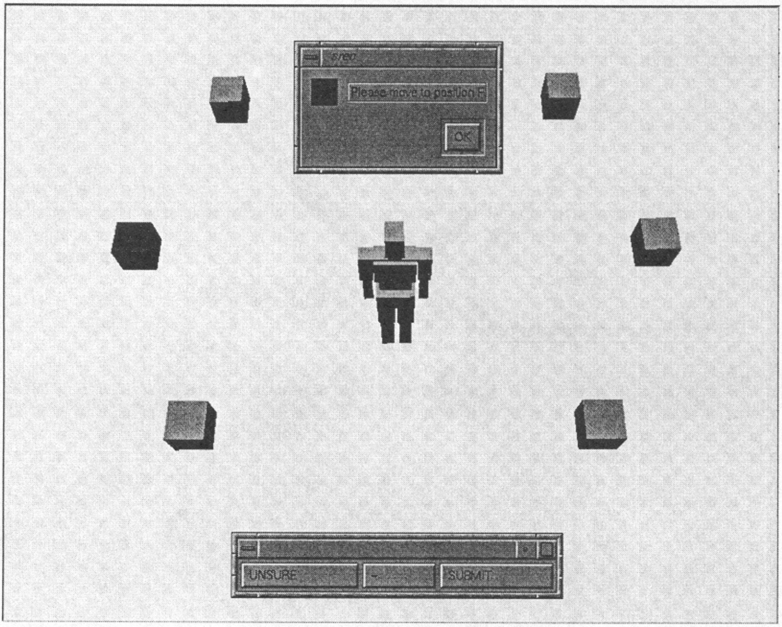


Fig 4. Subject in test area with trackball.



Please move to position F

Fig 5. Screen shot of subjective test program.

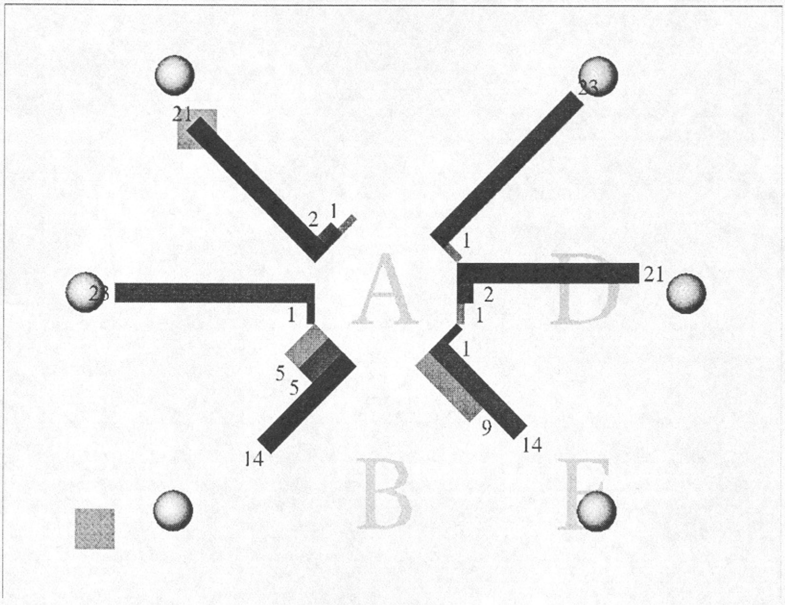


Figure 6. Response totals for Position A

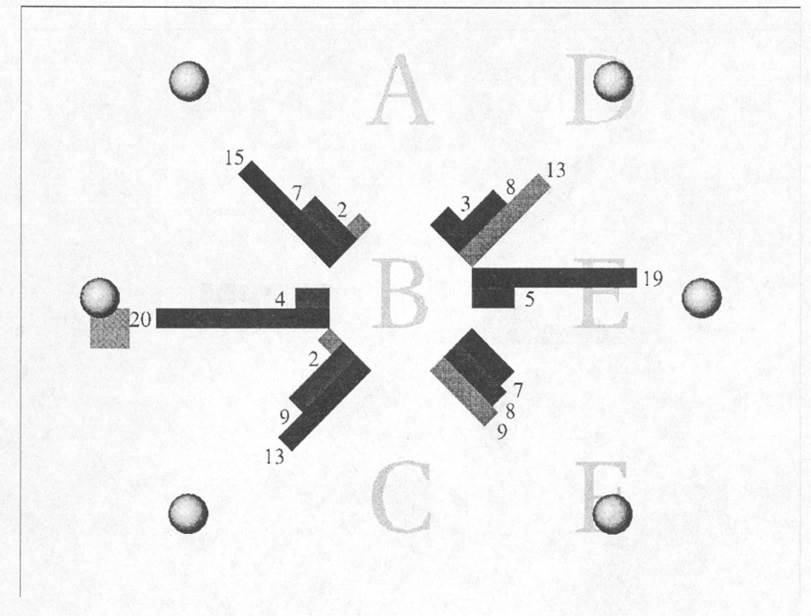


Figure 7. Response totals for Position B

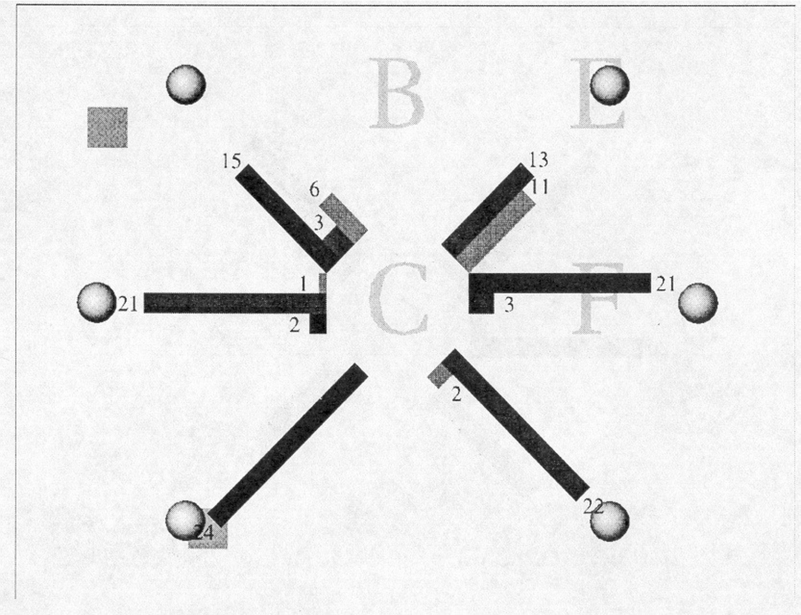


Figure 8. Response totals for Position C

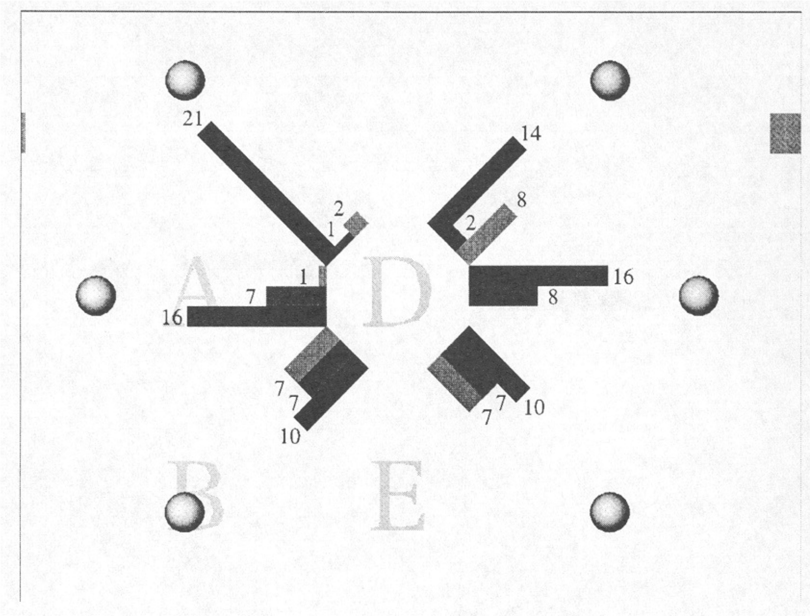


Figure 9. Response totals for Position D

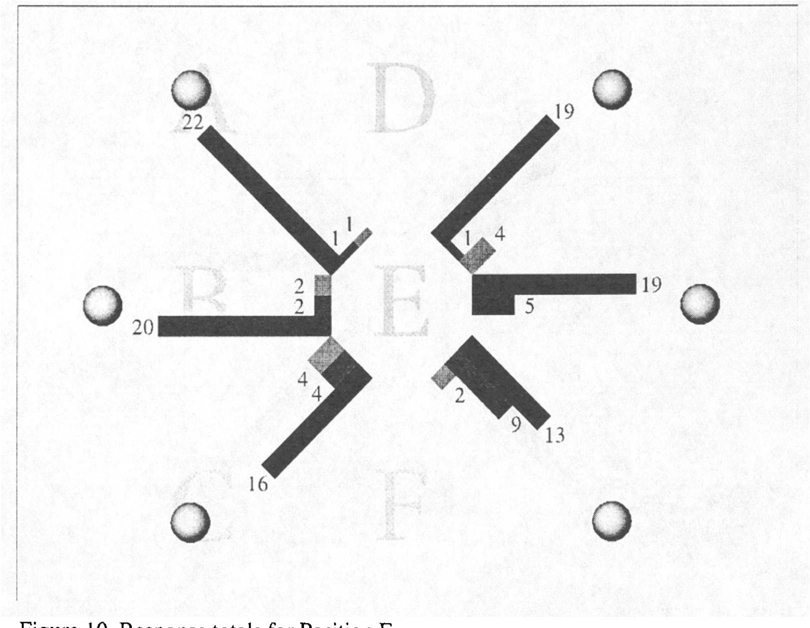


Figure 10. Response totals for Position E

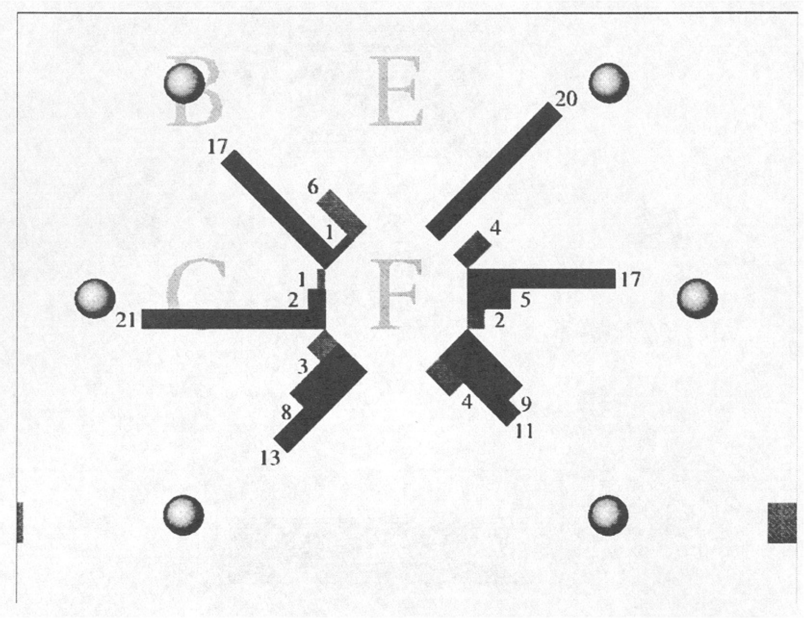


Figure 11. Response totals for Position F

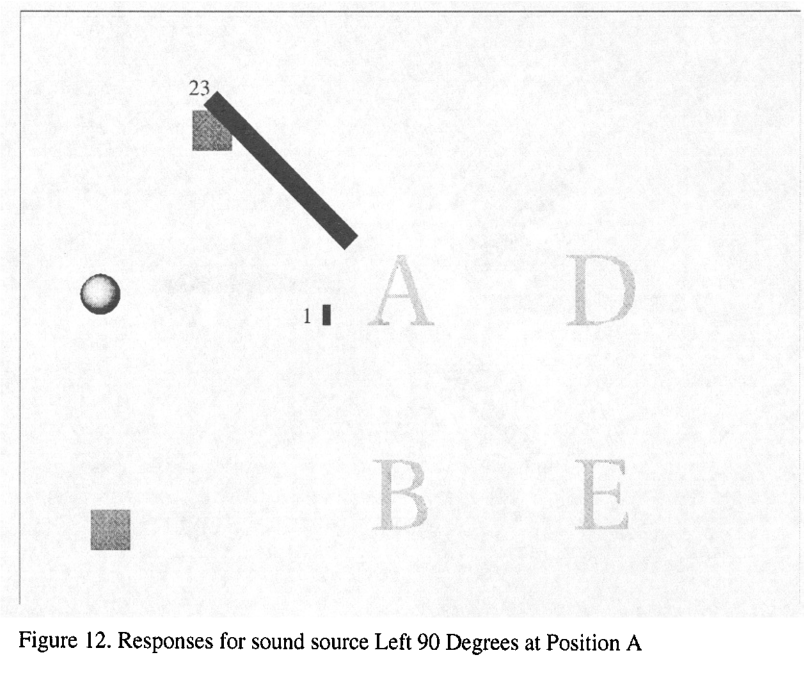


Figure 12. Responses for sound source Left 90 Degrees at Position A

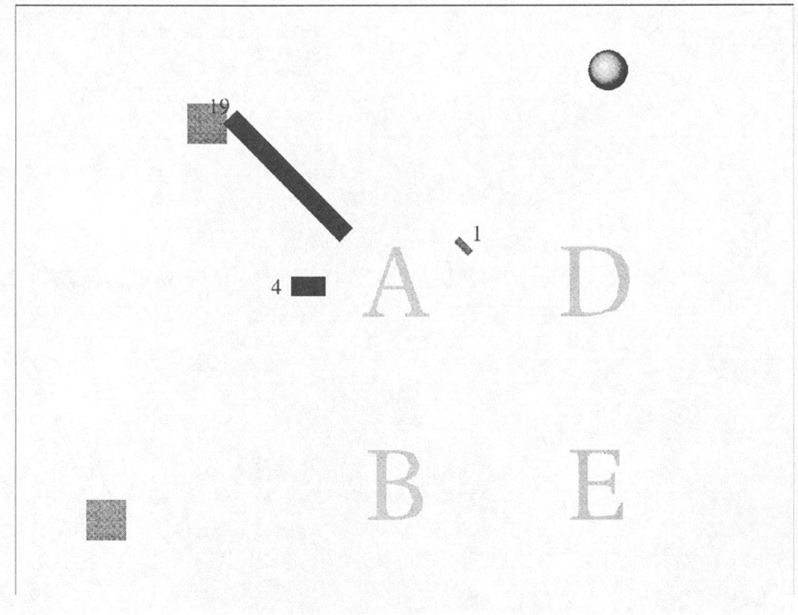


Figure 13. Responses for sound source Front Right 45 Degrees at Position A

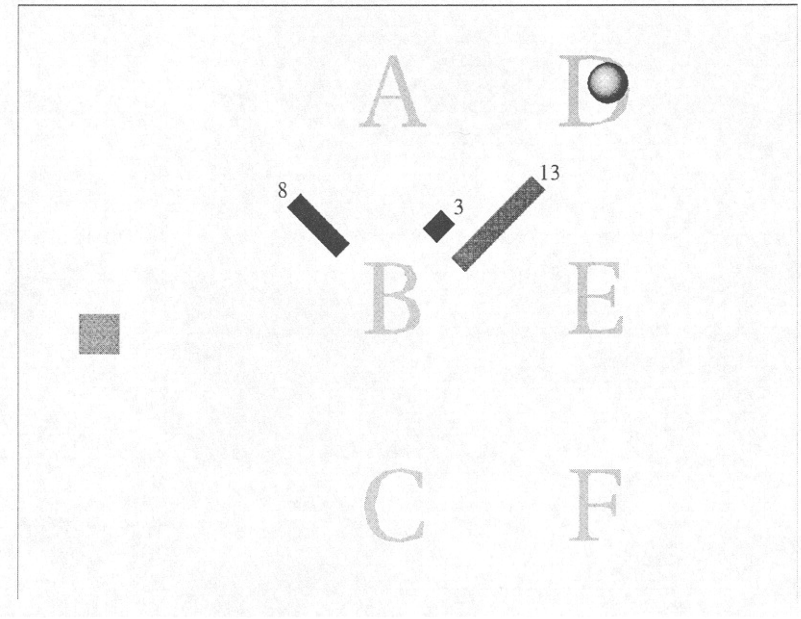


Figure 14. Responses for sound source Front Right 45 Degrees at Position B

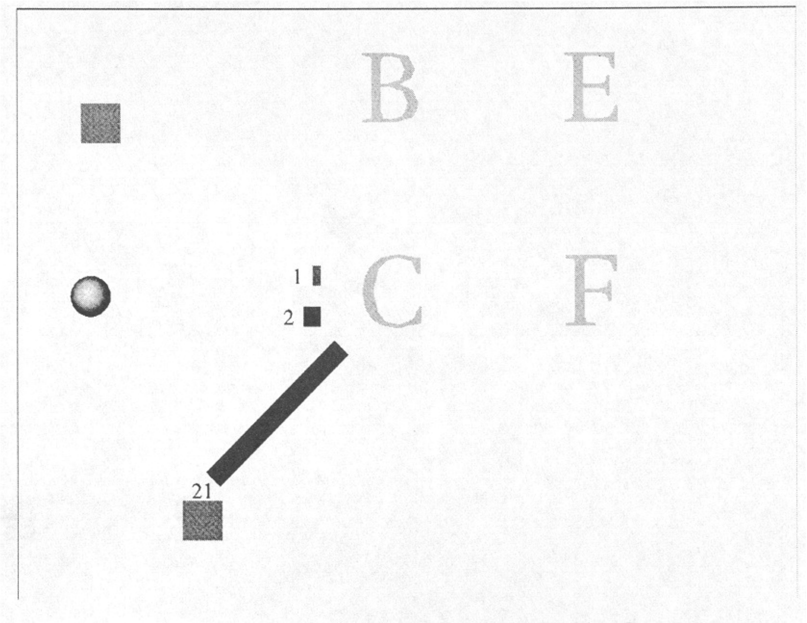


Figure 15. Responses for sound source Left 90 Degrees at Position C

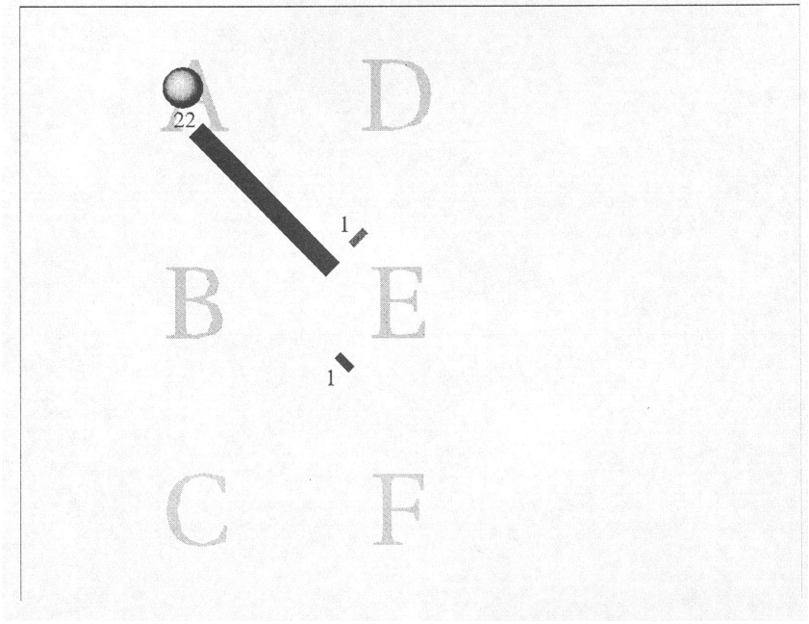


Figure 16. Responses for sound source Front Left 45 Degrees at Position E

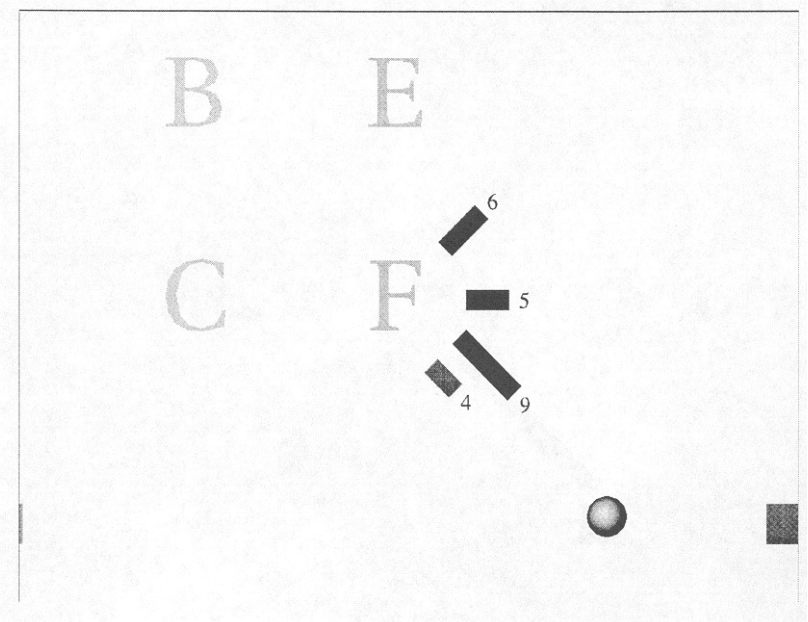


Figure 17. Responses for sound source Rear Right 135 Degrees at Position F