Spatial audio in graphical applications

J.D. Mulder, E.H. Dooijes

Computer Science/Department of Interactive Systems

Report CS-R9434 May 1994

CWI is the National Research Institute for Mathematics and Computer Science. CWI is part of the Stichting Mathematisch Centrum (SMC), the Dutch foundation for promotion of mathematics and computer science and their applications.
SMC is sponsored by the Netherlands Organization for Scientific Research (NWO). CWI is a member of ERCIM, the European Research Consortium for Informatics and Mathematics.

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

$\mathcal{S}$patial $\mathcal{A}$udio in $\mathcal{G}$raphical $\mathcal{A}$pplications

J.D. Mulder
CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands
mullie@cwi.nl

E.H. Dooijes
FWI, UvA
Kruislaan 403
1098 SJ Amsterdam
The Netherlands
edoh@fwi.uva.nl

Abstract

The use of audio in human-computer interfaces is gaining much attention nowadays. A neglected issue in this regard is the fact that human beings perceive sounds in a spatial context. Simulated spatial sound presented over headphones has many potential uses in a number of fields, such as virtual reality and visualization.

In this paper a simple experimental system for spatial sound simulation is presented, along with the results of several experiments that were conducted with this system. The system, called $\mathcal{SAGA}$ (Spatial Audio in Graphical Applications), was built to investigate the usability of spatial sound in three-dimensional graphical applications such as visualization.

Two primary uses of spatial sound in such applications are envisioned: (1) the active localization of virtual sound sources in a three-dimensional virtual environment in order to be able to move the point of view or a three-dimensional cursor towards this sound source, and (2) as an aid in the position determination of a cursor in a three-dimensional virtual environment. The results of the experiments conducted with the $\mathcal{SAGA}$ system show that the first use is very feasible, but the second will be more difficult to accomplish.

CR Subject Classification (1991): I.3.6: Methodology and techniques - *interaction techniques.*
Keywords & Phrases: Spatial audio, scientific visualization, localization

1 Introduction

One of the modes for human-computer interaction gaining much attention nowadays is the use of audio. Early usage of sound in computer interfaces was limited to simple beeps, for instance used as warning signs. With the increasing computational power available, and the development of fast dedicated signal processing hardware (digital signal processors), more complex audio came within reach. This led to an increasing interest in the use of computers as sound producing machines, particularly for the production of music. In addition, researchers realized the advantages and potentials of the use of both speech and non-speech audio in human-computer interfaces and applications, in the input as well as in the output direction. Particularly the use of the latter was introduced in a wide variety of applications. Some examples are in scientific visualization (to extract meaning from complex data [SC91]), in a human-computer interface for the blind [Edw89], or as an addition to existing interfaces, such as the "Sonic Finder", an interface that uses auditory icons [Gav89]. However, most of these applications were confined to monophonic sounds emitted by one simple speaker inside the computer.

Much research has been conducted on spatial sound (-perception) by physicists and psychologists. However, only a few researchers in computer science investigated the use of their results for computer interfaces and applications.

Report CS-R9434
ISSN 0169-118X
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Spatial sound offers some major advantages over the use of monophonic sound produced by one loudspeaker. First of all, auditory perception of human beings is omni-directional, as opposed to vision: we can hear sounds emitting from any direction, whereas we can only see what is in front of us. Second, spatial sound can also provide information about the distance of a sound source. These observations imply that the 'display area' of a computer would no longer have to be limited to the physical size of the screen used: objects, events, or whatever it is we want the user to be aware of, could be located anywhere in space, not just on the screen.

A feasible area for the use of spatial audio is scientific visualization. Some research has already been done in this area, primarily on the subject of mapping certain attributes of high-dimensional data to specific acoustic features. In [SBG90], for instance, several auditory data representation techniques are described, including the generation of stereophonic sound with apparent depth and sound that appears to emanate from a two-dimensional area. And in [WPFF90], an acoustic display system for visualization purposes is presented that is capable of generating localized acoustic cues linked with information or events in the (visual) display.

1.1 Aims of This Research Project

This research project was conducted to investigate usability of spatial sound simulation for three-dimensional graphical applications. The envisioned use of spatial audio in three-dimensional graphical applications was focussed on two related areas:

- The search of virtual sound sources in a three-dimensional virtual environment.
- Position estimation of a cursor in a three-dimensional virtual environment.

For instance, in a visualization system, spatial sound could be used as an aid in the search for an interesting point, object, or area in a visualized volumetric data set. This could be used to change the current point of view so as to focus on that particular area or object, but also to direct a three-dimensional cursor to the area, or to be able to select an interesting point or object with the cursor. Another example is the possible use in three-dimensional drawing applications. The spatial sound could be used as an aid in cursor positioning in the three-dimensional space.

To investigate the applicability of spatial sound in three-dimensional graphical applications, a simple spatial sound simulation system was built, called *SAGA* (Spatial Audio in Graphical Applications), and several experiments with this system were performed.

1.2 System and Experiments

The *SAGA* system is basically built up of two components: one for the graphics and one for the spatial sound. The graphics component of the system can visualize a virtual environment with arbitrary objects in real-time, using either mono-graphics or (time-multiplexed) stereo-graphics. The sound component of the system can simulate two spatial sound sources, static or moving and located anywhere in the environment. A simple model of the cues used by humans to localize sound sources is used to simulate spatial sound.

The 'point of hearing' can in principle be located at an arbitrary position in the environment, but in the experiments it was restricted to two types of location. The first is the same position as the viewpoint, so the user hears the sound as if he is part of the environment. The second is a position incident with a three-dimensional cursor which can be arbitrarily moved through the environment. In this case, the user hears the sound as if his head is positioned at the same location as the cursor, having the same orientation as the cursor. In other words: the user hears what the cursor hears.

One part of the experiments focussed on the active localization of sound sources, both by moving the cursor hearing what the cursor hears, as well as by moving the viewpoint hearing the sounds as they arrive at the viewpoint. The other part of the experiments concerned cursor position determination in three-dimensional space.

1.3 Outline

The outline of the remainder of this paper is as follows. Section 2 summarizes some important aspects of spatial sound perception by humans, along with some techniques on how these cues can be simulated over headphones. In Section 4, a description is given of the *SAGA* system developed during this project. Section 5 describes the conducted experiments, and discusses their results.

2 Spatial Sound

There are eight different cues that are of particular importance in sound localization by humans, which can be divided into two classes. The first class contains those cues that are related to the shape of a person's (outer) ears, head and shoulders. The four cues of this class are:

- Interaural time delay
- Head shadow
- Pinna response
- Shoulder echoes

Together, these cues define the so called *Head Related Transfer Functions* (HRTFs), where there is one function for every direction of a sound source.

The other class consists of four cues not (directly) related to the shape of a person's head, shoulders and ears:

- Head motion
- Vision
- Intensity
- Early echo response and reverberation

2.1 Interaural Time Delay

Interaural time delay, or interaural time difference, ITD, is the delay between a sound reaching the closer and farther ear. In pure tones (sinusoidal waves) the ITD manifests itself as a phase difference. Frequencies above approximately 1.6 kHz relate ambiguously to this phase difference. Indeed, it has been shown that the auditory system is insensitive to interaural time differences in pure tones of frequencies greater than approximately 1.5 kHz. However, in complex waveforms, the ITD manifests itself as an envelope delay. ITDs can be used in complex high-frequency sounds if these sounds have relatively slow modulations that the auditory system can lock onto over the duration of the signal [Hen74].

Although ITDs provide a primary cue for determining the position of a sound source, especially in the lateral planes, they correspond ambiguously with the direction of a source. Assuming a stationary, spherical model of the head, a given ITD value constrains the direction to positions located on a conical shell around the interaural axis (the *cone of confusion*).

Actual heads are not quite spherical and the ears are placed a little asymmetrically (each is about 7° in front of the idealized interaural axis), but interaural time differences computed from such a model are in fact a good approximation of the actual differences [Mil72].

Another form of interaural time differences is transient disparity. These are interaural differences in the time of onset, or time of arrival of the first wave of a tone pulse at the ears. They are not limited by the period of the tone, and they are not subject to the phase ambiguities of steady tones, but the onset information ends when the sound has reached both ears [Mil72].

2.2 Head Shadow

Another important cue is the effect of the head shadow, also known as the interaural intensity difference or IID. This effect is due to the fact that the sound has to travel around the head to reach the far ear. The head acts as a filter for the far ear, but its properties are not a simple function of either the direction or the frequency of the sound. Interaural intensity differences are negligible at very low frequencies (below approximately 1 kHz, because diffraction effects around the head only become significant above that frequency), but may be as great as 20 dB at high frequencies [Mil72].

The ITDs and IIDs, also called binaural differences, are primarily thought of as important cues for direction and although there seems to be evidence that binaural differences resulting from distance changes should be detectable by the ear, it is not known whether this binaural information is translated by the listener into distance judgments [Col63].

For a simple spherical head model, IID will have the same basic circular symmetry about the interaural axis as ITD (the cone of confusion) [SBDC76], which to a certain extend has also been shown from measurements [MMG89].

Modeling the interaural intensity differences is somewhat complicated. Assuming a simple relation between the loss of intensity and the distance a sound travels through the air, for instance the "$\frac{1}{r^2}$ loss", is not a very good model of the actual situation. The sound reaching the farther ear has to travel around, or even through the head, which will lead to different intensity losses. Another aspect is that the frequency spectrum of the sound is attenuated, as well as when it propagates through air (depending on a number of factors, such as the humidity and density) as when it propagates around and through the head.

2.3 Pinna Response

The outer ear, called pinna, acts as a filter to incoming sounds. The response of the filter is highly dependent on the direction and the frequency of the sound. The structure of the external ear interacts significantly with the incident sound when the wavelength is of the same order or shorter than the ear's dimension, that is, for frequencies from about 4–5 kHz up. The effect of the pinna is believed to be an important cue for determining both the elevation and the azimuth of a sound source, especially for sound localization in the median plane and for resolving the cone of confusion.

It is also suggested that pinna effects, both monaural as binaural, may yield information about distance as well, but a quantitative evaluation of this aspect of pinna responses has not yet been made, and no psychophysical study of these effects is available [Col63].

Although much research has been done on the subject of pinna response, a general model has not yet been developed, although there have been some attempts. For instance, the model used in [Wat78] assumed that the pinna essentially acts as a reflector which introduces delayed replications of the high frequency components of the incident sound.

There have been many measurements of the transfer functions of the outer ear, which indicated important aspects of the pinna response, such as certain notches and peaks in the frequency response when a sound source is moved up-down in the median plane [BB77]. Most of these measurements also include the effects of the head and shoulder characteristics since they are performed on subjects or artificial heads and not on an ear by itself.

2.4 Shoulder Echoes

Apart from the head and (outer) ears, the shoulders and the upper body also contribute to the spectral shaping of the sound. Not only do the shoulders cause echoes which reach the ear with a delay dependent on the elevation of the source, the effect on the spectrum of the sound, caused by the reflections, is direction dependent as well. It turns out that frequencies in the range of roughly 1–3 kHz do reflect from the shoulder [Gar73]. In case of a familiar sound, shoulder echoes may

provide both elevation and azimuth information, although previously mentioned cues are likely to be of greater importance [SBDC76].

When it comes to modeling the echoes from shoulders and upper torso, little research has been done, and no models have been established for these effects.

2.5 Head Motion

It is agreed that head movements play an important role in the localization of sound sources [TR67, TMR67, SWKE89, Mil72]. Head movements are especially useful for solving ambiguous directional information. Rotating the head around any axis generates specific changes in the perception of the cues related to the HRTFs, thereby limiting the possible positions of a sound source.

Translation of the head does not only contribute to the determination of the direction of a source but also serves as a distance cue. As a listener passes by a sound source, the direction of the source changes. For a given direction, the angular velocity of the source is inversely proportional to distance for a constant translational velocity.

Incorporating head motions in a spatial sound system can be achieved by the use of a head-tracking device. The sensed orientation can be used to select the proper HRTFs, whereas the sensed position can be used to adjust the distance cues such as intensity and reverberation.

2.6 Vision

The role of visual cues is extremely important in our perception of spatial sound. Our visual system is more spatially acute than the auditory system, especially in sensing the angle of elevation. But, as stated earlier, the visual spatial field is limited to the region in front of the user, whereas the auditory spatial field is unbounded and surrounds the listener. It is very likely that ambiguous direction perception of sound sources is resolved by visual cues (e.g. if we do not see a sound source in front of us, it must be behind).

Visual cues are also an important factor in distance judgments of a sound source. An example is the so called "proximity effect", being the tendency on the part of the observer to hear a sound as if it were coming from the closest of the rationally possible (visible) sources [Gar68].

We sometimes even rely so heavily on visual correlates, that we tend to ignore our auditory directional cues if they are conflicting with our visual cues. For instance, at a drive-in movie, one is strongly aware that the sound source and the visual image do not coincide. After some time, however, the discrepancy is no longer noticed, and one adapts to the displaced sound.

Incorporating visual cues into a spatial sound system is somewhat more complicated since visual display systems usually consist of an essentially two-dimensional screen. Actually, as mentioned in the introduction, this limitation is one of the main motives for the investigation and development of spatial sound systems in the first place.

2.7 Intensity

Apart from the interaural intensity differences that form a cue for the direction, intensity as such is a cue for the distance of a sound source. The intensity of a sound varies inversely with the square of the distance from its source, in other words, there is a 6 dB loss in sound pressure for each doubling of distance in free space. It is not surprising, therefore, to find that distance estimates increase systematically, when the intensity of the sound at the ear is decreased [Gar69].

From several experiments, it transpired that the intensity of sound primarily serves as cue to changes in distance [MK75], and that practice is necessary before a listener can use the intensity of an unfamiliar sound as a basis for distance judgments [Col62]. As an example, if speech is used as stimulus, distance judgments are better than when tones or white noise are used.

The intensity cue can be simulated by using natural damping and dispersion laws.

2.8 Early Echo Response and Reverberation

Early echo response is a term for the clear echoes we hear (but do not consciously perceive) in the first 50 to 100 ms after a sound starts. Early echo response and reverberation are caused by the room acoustics and are believed to be an important cue for distance perception.

Reverberation provides an absolute, and thus also relative, distance cue, with greater reverberation delays being associated with greater perceived distances [MK75]. As a sound moves away from the listener in a natural indoor setting, the proportion of sound energy directly reaching the observer's ears decreases, while the proportion reaching the observer's ears after reflection (and thus delayed) from surrounding surfaces increases. The delay of the reverberant sound relative to the direct sound may also change with the distance of the source but neither of these effects is an invariant transform of distance. The intensity and delay of the reflected sound depends upon the acoustic properties of the space; the effect on perceived distance varies with the listener's familiarity with the space [Mil72].

Echoes reflected from surfaces in the environment could provide additional information about the location of a sound source if the auditory system were equipped to use them and if the listener were familiar with the arrangements of the reflectors. In unfamiliar environments, however, such echoes would be confusing, and the auditory system seems designed to suppress awareness of them for direction judgments. The suppression of these echoes for localization is called the "precedence effect": the sound that reaches the ears first (by the most direct path) preempts the perception of direction [WNR49].

Incorporating room acoustics into the simulation of spatial sound greatly improves the overall spatial quality of the sound, but no method has yet been developed that is capable of real-time simulation of arbitrary rooms.

2.9 Other Possible Cues

Several other possible cues for auditory localization have been suggested and examined. For instance, the distinction between frequency spectra of stimuli at near and far distances, are thought of as a possible distance cue. Based upon physical analysis of acoustic stimuli it was suggested that a decrease in the high-frequency content of a sound would result in decreasing the judged distance if the sound were perceived to be close, and increasing the judged distance if the sound were perceived to be farther away. However, insufficient empirical data to support this theory was presented, and the significance of this cue could not be determined [Col63, Col68].

As mentioned before, familiar sounds are easier to localize than unfamiliar ones. In addition, the 'quality' of the sound is also of importance. An illustrative example in this case is the observation that the distance to a whispering talker is underestimated, whereas the distance to a shouting speaker is overestimated [Gar69].

Another possible cue could be the influence of bone-conduction. A sound can reach the inner ear via two paths: by the ear canal, ear drum, etc., and by bone conduction via the skull. It was suggested that the portion of sound reaching the inner ear by bone conduction varies with the angle of incidence and so contributes to the formation of the direction of the sound sensation. However, these and other forms of tactile stimuli, such as the ones perceived by the pinna or the skin of the neck, are regarded as less important.

3 Common Problems

From the previous sections, it has become clear that modeling each of the HRTF cues separately is difficult, and up till now no complete model has been developed that encounters the full scale of useful cues for spatial sound simulation. Therefore, the possibility of using measured HRTFs has been suggested and investigated [WK89a, WK89b, Wen92]. The main advantage of the use of measured HRTFs is that the filters obtained from the measurements enclose all HRTF cues. Disadvantages are, among others, that the use of the obtained filters is computationally expensive and that measured HRTFs tend to be more or less personal. In [Bur92] several suggestions were

made on how to reduce the computational demands and in [WWK91] it has been investigated whether an individual can use a set of HRTFs other than his own. The ideal set of HRTFs still has to be found, if it exists at all.

If normal sounds are heard over headphones (without any spatial simulation), they are perceived as being located inside the head, while adequately simulated spatial sounds are perceived as having a certain spatial position. This is referred to as the "out of the head sensation". It is believed that just about every cue mentioned in the previous section contributes to the out of the head sensation, although pinna response, head movements and reverberation seem to be of particular importance.

Another common problem is front-back reversals. Actually there are two classes of confusions. One class are the azimuth confusions consisting of front-to-back confusion: perception of a forward target in the rear hemisphere, and, vice versa, back-to-front confusion. The other class are the elevation confusions with up-to-down confusion: perception of a upper target in the lower hemisphere, and, vice versa, down-to-up confusion. Adding the cues of head motion and vision to the auditory display system is probably the best approach for eliminating these kind of confusions.

4 The *SAGA* System

The *SAGA* system (Spatial Audio in Graphical Applications) can simulate one or two static or moving sound sources which can be positioned anywhere in a virtual environment. This environment is displayed on the screen in either mono or stereo graphics.

In the current configuration of the system, there are two options for the location of the point of hearing. In the first, it is located at the same position as the viewpoint; the user can then manoeuvre the point of view through the environment by the use of two joysticks and hear the sounds as they arrive at the viewpoint, that is, as if he is part of the virtual environment. The other option is that the point of hearing is located at the same position as a three-dimensional cursor; the user looks from a static viewpoint into the environment and can move the three-dimensional cursor through the environment while hearing what the cursor hears, i.e. what he would hear if he had the position and orientation of the cursor.

Figure 1 depicts a block diagram of the *SAGA* system. It basically consists of three elements: the host computer, the digital signal processor, and the digital to analog conversion unit.

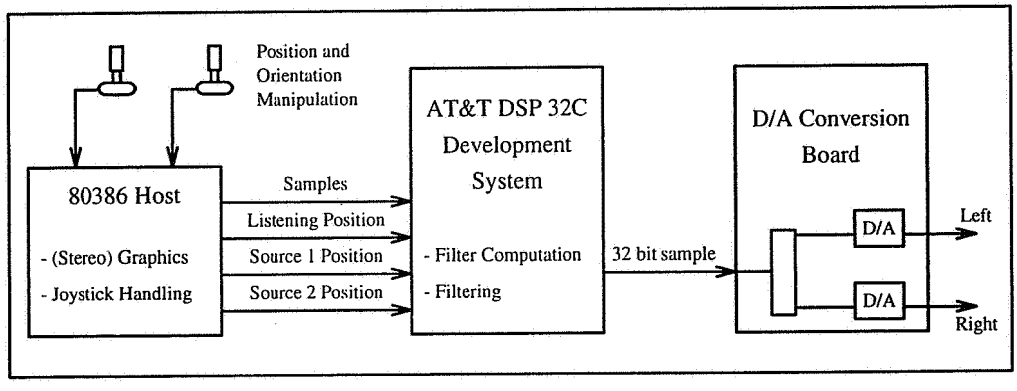

Figure 1: Block diagram of the *SAGA* system.

The host computer is an 80386, 33 MHz IBM-compatible PC. The main functions of this host are:

1. **Graphics:** It displays the virtual environment in either mono-graphics or stereo-graphics.
2. **Motion:** It handles the changes in the positions and orientations of the cursor and viewpoint (which are invoked by the user through the joysticks), as well as the (changing) positions of the sound sources.

3. **DSP interface:** It provides the digital signal processor with the sound samples for the sound sources, the position and orientation (-changes) of the 'listenpoint', and the positions (-changes) of the sound sources.

The digital signal processor board is an AT&T DSP32C development system which operates at 50 MHz. It was configured to produce appropriate sound samples at a rate of 24 kHz, which provides a good trade-off between sound quality and computational bandwidth. The main functions of the DSP are:

1. **Compute Filters:** From the information it receives from the host through direct memory access via its parallel port (the positions of the sound sources, the position and orientation of the 'listenpoint'), it computes the correct filters.

2. **Filtering:** It has to filter the sound signals with the appropriate spatial filters and direct the computed left and right samples, which are packed into one 32 bit sample, at a rate of 24 kHz to the digital to analog conversion board through its serial port.

The function of the digital to analog conversion board is simple: it has to split the serial 32 bit sample it receives from the DSP into two 16 bit samples, one for the right ear and one for the left ear, and convert them into analog output at a rate of 24 kHz. The board is built around an inexpensive dual 16 bit DAC (Philips TDA1543).

4.1 Spatial Sound Simulation

A model was developed for the spatial sound simulation. Because of the limitations inherent to the hardware, the model had to be simple and fast. As a consequence, the spatial sound cues are modeled in a rather simple or even crude manner to allow real-time computation on the DSP. For the same reason, it was decided to use pre-sampled sounds. The cues incorporated in the spatial sound model are:

- Interaural time difference. By computing the distances from a sound source to each ear the delay of the sound reaching each ear can be determined.

- Interaural intensity difference and intensity loss. As for intensity loss of a sound coming from the point source that can propagate through the air without having to travel around the head, a $\frac{1}{r^2}$ loss is assumed. If a sound has to travel around the head a greater loss is assumed. Both losses are overall intensity losses, i.e. they do not affect the frequency spectrum of the sound.

- Pinna effects for elevation. For the pinna effects a similar 'reflection model' is used as in [Wat78], although only the elevation reflections are simulated because the azimuth reflections were not specified for the full 180° in [Wat78] and incorporating both azimuth and elevation echoes would be computationally too expensive.

- Simple reverberation. As an additional distance cue and in an effort to enhance externalization, a simple reverberation unit is added to the filter. To simulate the attenuation of the high-frequency components of the sound by the air, a low-pass filter is connected in the loop of this reverberation unit. The reverberant sound does not cause any elevation echoes, nor is it affected by any time differences between the ears for there is no location from which it originates. Also, it is not affected by the distance between the source and the ears.

- Head motion. The virtual head, or 'listenpoint', can be placed anywhere in the virtual environment and freely moved around; both its position as well as its orientation can change.

- Vision. For the graphics, an experimental software development package called "REND386" is used. This package consists of several libraries containing routines for fast 3D polygon rendering. It supports perspective projections of three-dimensional scenes for either mono-vision or stereo-vision (time-multiplexed, mirrored, or head-mounted). It was developed by

Dave Stampe and Bernie Roehl of the Electrical and Computer Engineering Department of the University of Waterloo in Ontario, Canada. At the moment, both mono-graphics and time-multiplexed stereo-graphics (using a pair of SEGA LCD shutter-glasses) are used for the visual display of the *SAGA* system.

In the current configuration of the system, the user has two joysticks at his disposal to navigate the cursor or the viewpoint through the virtual environment; one for the translations and one for the rotations. When manipulating the viewpoint, the virtual head is located at the same position as the viewpoint with the same orientation: their coordinate systems are incident. The invoked translations and rotations are along and about the principal axes of the viewpoint (head) coordinate system. If the user is manipulating the cursor with the joysticks, the coordinate systems of the virtual head and the cursor are incident. The translations are performed in the directions of the principal axes of the viewpoint coordinate system. The rotations, however, are now performed about the principal axis of the cursor (the virtual head).

5 Experiments

The experiments can be divided into two categories according to their objectives. The first category concerned active sound source localization (described in Section 5.1). Here, the main objective was to investigate whether the used spatial sound cue models provided enough information for subjects to accurately find sound sources in a cubic room. The other series involved cursor position estimation (described in Section 5.3). It was investigated to what extent the subjects could estimate the position of an invisible cursor in a cubic room by using information provided by one or two sound sources.

The experiments were performed on four subjects, all undergraduate students at the faculty of Mathematics and Computer Science of the University of Amsterdam. None of the subjects had ever experienced any hearing problems nor had had any previous experience in sound localization tasks. The experiments were performed in the same order as described here and the total session lasted about two and a half hours per subject, depending on their performance.

In all experiments, the subjects were seated in front of the screen, the rotations joystick in their left hand, the translations joystick in their right hand, wearing the headphone, and, when using stereo-graphics, wearing the shutter-glasses. The virtual environment used in the experiments consisted of a cubic room of $5 \times 5 \times 5$ m. When sound sources were made visible, they were depicted as small cubes. The cursor was depicted as a three-dimensional arrow, with different colors for the top and bottom, so its current orientation could be determined simply by looking at it. As an example of a view, consider Figure 2. Here, the viewpoint is located outside the room and the cursor is positioned near the center of the room, while there are two sound sources visible.

Before a subject started the experiments, he was given a thirty minute training period to get used to the system, e.g. the graphics (stereo and mono) and the way of moving the viewpoint and cursor. This training session was performed using the same cubic room with two visible sound sources, each emitting a different sound: one calling the word "one" every other second, the other the word "two" every other one and a half seconds.

5.1 Session I: Sound Source Localization

For the first experiments, the space within the cubic room was rasterized with a resolution of 1.5 m and at all raster points, except the center of the room, (identical) cubes were shown. One of these cubes was the actual sound source, and the subjects had to move the viewpoint or cursor 'into' this sound source as quickly as possible, starting at the center of the room. When moving the cursor, the viewpoint was positioned at a fixed location outside the room, translated -6.5 m along the viewing axis away from the center of the room. From this position, all the sources could be seen simultaneously.

In the experiments, the source emitted a familiar sound (the word "here") every other second, and the subjects had to find the source under different circumstances:

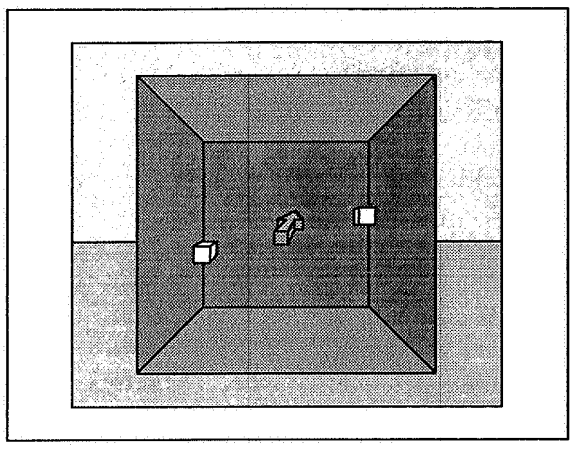

Figure 2: Example view of the graphics as used in the experiments. The viewpoint is looking into the room from the outside, so three walls, the ceiling, the floor, and the horizon can be seen. Two visible sound sources and the cursor are present in the room.

- Moving the viewpoint using stereo-vision and mono-vision.
- Moving the cursor using stereo-vision and mono-vision, hearing what the cursor heard.

In each case, four sound sources had to be found, each at some 'random' raster point in the room. To gain insight into whether it was an advantage that the sound source made a familiar sound, one experiment was repeated with an unfamiliar sound (random noise). In a second series of experiments the subject had to localize an invisible source in an empty room. The same variations as in the former series were applied. All the movements of the viewpoint or cursor (invoked by the subjects) were logged and these data were later used for analysis.

5.2 Results of Session I

Three important aspects of the localization tasks were extracted from the data: localization accuracy, localization speed, and how the sources were localized, i.e. which translations and rotations were performed to reach the sources' positions.

To get an impression of what the paths taken by the subjects to the sound sources look like, Figures 3 and 4 each show a path of a subject as performed in one of the localization experiments, reconstructed from a log-file. No performed rotations are depicted in the figures, and movements back and forth cannot always be seen, because all timing information is lost. Figure 3 depicts a short path, i.e. the source was found fast and accurately. It was performed by a subject in the configuration of moving the cursor, mono-vision, invisible sources and familiar sound. Figure 4 depicts a very long path, i.e. the subject had some trouble finding the source. It was performed by an other subject in the configuration of moving the cursor, stereo-vision, visible sources, and familiar sound.

5.2.1 Localization Accuracy

In each configuration, practically all sources could be localized. In the experiments with visualized sources, only once an incorrect source was pointed out by a subject. In the case of invisible sources, subjects had difficulty reaching the exact position of the source. A source was localized exactly if it was positioned within the 'virtual head', that is, if the distance from the center of the head to the source position was less than the radius of the head (8.75 cm). From the obtained results it seems that adding the cue of depth to the visual perception by means of stereo-vision does

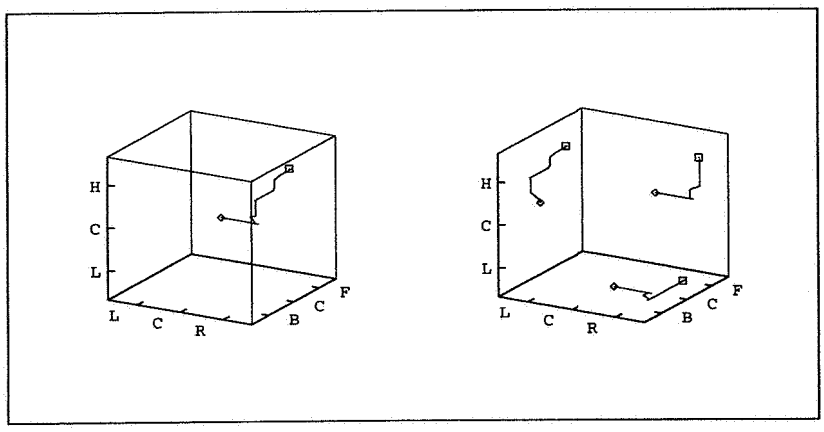

Figure 3: Example of a good localization path. ◇ is the start position (Center-Center-Center), □ is the source position (Right-High-Front). Left is the actual path inside the room, right are the projections of the path on three walls. This path was performed by a subject in the configuration of moving the cursor, disabled stereo-vision, invisible sources and familiar sound.

not enhance the accuracy of sound localization with the cursor, but it does increase localization accuracy in the case of viewpoint localization.

In the case of mono-vision, the localizations with the cursor were performed more accurately than with the viewpoint, whereas with stereo-vision there was no (big) difference. So, apparently, the indirect connection between the sound and the subject's head (the subject heard what the cursor heard) did not decrease localization accuracy. Perhaps the almost complete absence of visual cues which would contribute to the sense of movement in case of localization with the viewpoint (particularly in the case of mono-vision), and, on the other hand, the ability of seeing the induced movements of the cursor, did contribute to the localization performance for the cursor tasks; subjects probably had a better sense of movement and could relate changes in position with changes in perceived sound more accurately.

Figure 5 depicts the final positions of the sources that were not exactly localized in the localization tasks with the cursor, projected in the vertical and horizontal lateral plane of the virtual head (the final positions of the sources in case of localization with the viewpoint show a similar result). The sources are mostly located at or near the median plane. From this observation it can be concluded that the cues for median plane localization in the present spatial sound model are of limited use, although it must be stated that median plane localization is more difficult than lateral plane localization in real life as well.

5.2.2 Localization Speed

Since the rate of position and orientation updates depends on the graphical configuration, comparisons among localization speeds acquired under different circumstances cannot be done by simply comparing the time it took to find the sources' locations. A better criterion is the total distance covered by the viewpoint or cursor to reach the sources.

The distances covered in all localization tasks showed a great variety among the subjects as well as among the different source positions. There seemed to be no direct relation between the spatial dimensions of the sources' positions and the covered distances. This could indicate that the 'static cues' for localization did not provide sufficient information for the subjects to reach a source's position with minimal movements. Since the movements did result in a change in the perceived sound, and since the average covered distances were relatively large compared to the actual distances from the start position to the sources, it seems that the *changes* in the perceived sound were of particular importance for the localization.

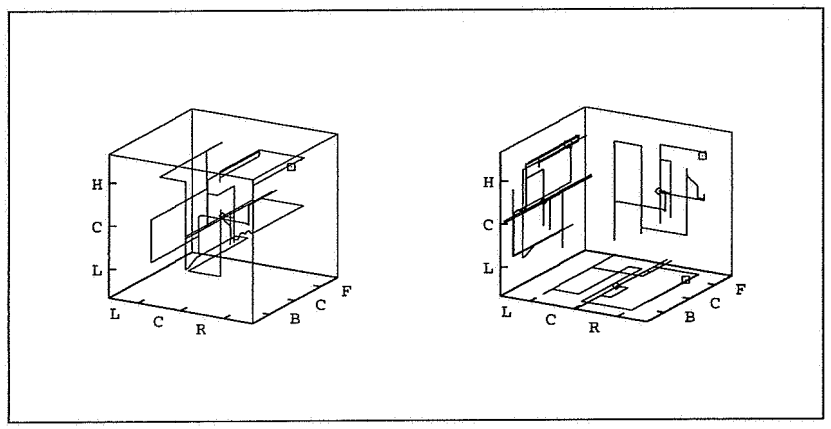

Figure 4: Example of a poor localization path. ◇ is the start position (Center-Center-Center), □ is the source position (Right-High-Front). Left is the actual path inside the room, right are the projections of the path on three walls. This path was performed by a subject in the configuration of moving the cursor, enabled stereo-vision, visible sources and familiar sound.

Comparisons among the averages over all subjects for the different configurations yielded some rather strange results. For viewpoint localization, average distances were larger in the case of non-visualized sources than with visualized sources. This does not come as much of a surprise since it could be expected that the addition of visual cues would increase localization performance. However, in the case of localization with the cursor, there appeared to be no significant difference in the average distances between visualized and non-visualized sources. Perhaps visualizing the sources tempted the subjects to pay less attention to the audible cues and use a 'trial and error' method on the visible sources to localize the source that actually emitted the sound, herewith making superfluous movements. Such a trial and error method is less feasible in the case of viewpoint localization because when the viewpoint was located inside the room, only a few sources could be seen at once, whereas the static viewpoint used for localization with the cursor overlooks the entire room, so each source was visible.

The use of either stereo-vision or mono-vision does not seem to have been of great influence in the case of localization with the cursor. For localization with the viewpoint, however, there was a larger difference between mono-vision and stereo-vision results in the case of invisible sources. Again, the absence of visible objects seems to have degraded the visual spatial perception of the scene, and therefore stereo-vision became more important for the (visual) spatial orientation.

It was remarkable to see how little difference there was between viewpoint and localization with the cursor. Apparently, both approaches can be used for sound localization under these circumstances, where each approach has its advantages and disadvantages, depending on the configuration. Cursor localization seemed to benefit from the fact that the subject could directly see the entire environment and the result of the invoked movements, whereas with localization with the viewpoint only a small part of the environment was visible at once and therefore sometimes a good sense of motion was lacking, especially if there were no sources visualized. Viewpoint localization, however, did not suffer from 'indirect hearing', as was the case with localization with the cursor. Subjects heared what the viewpoint heared and did not have to displace themselves mentally to the position and orientation of the cursor.

One experiment was conducted to see if the usage of an unfamiliar sound (noise instead of the word "here") would influence the localization performance in case of localization with the cursor. In Section 2 it was stated that the static localization of unfamiliar sounds is more difficult than if familiar sounds are used. However, under the circumstances used here, particularly the active manner in which sounds were to be localized, there was no significant difference.

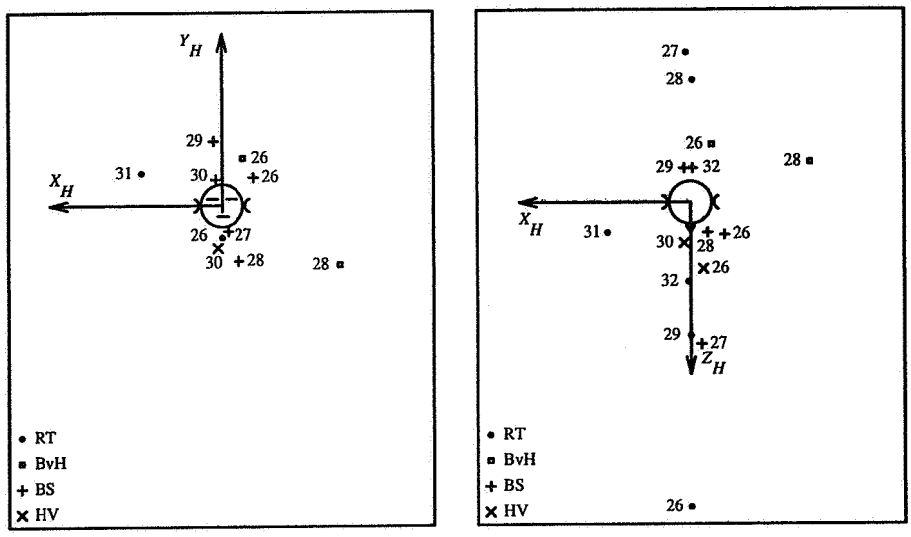

Figure 5: Final positions of inexactly localized sources in localization with the cursor. Depicted are the projections of the sources' positions at the vertical and horizontal lateral plane (on the left respectively right). Positions that would be projected inside the head in one of the planes are omitted in that plane.

5.2.3 Localization Movements

Another interesting aspect is to investigate which movements the subjects used in the localization tasks. The results clearly showed that the front-back movement was the most used *translation* for (practically) all subjects in each configuration. Observations made during the experiments and a closer study of the log-files lead to the following explanation. First of all, the left-right localization seemed to be the easiest part of the localization tasks, so less movements along the inter-aural axis were necessary. This assumption is supported by the results of the analysis on localization accuracy, discussed in the previous section. Secondly, translations along the vertical axis were invoked by pressing the joystick's buttons, a method that gave the subject less control over their movements compared to left-right and front-back movements because of the small (fixed) step size and therefore relatively slow motions. Thirdly, it was observed that the subjects tended to move back and forth a lot if they thought they were close to a source's position. And finally, perhaps moving back and forth is just a more 'natural' translation compared to left-right and up-down; human beings tend to move forward in everyday life, not left, right, up, or down.

The same analysis was done for the cursor translations. The averages showed the same tendencies as with localization with the viewpoint, so the same conclusions can be drawn here.

The total amounts of the different *rotations* (tilt, pan, and roll) about the axes of the head coordinate system were examined to extract the subjects' mostly used 'head rotations' in the localization tasks with the viewpoint.

The differences in the amount of rotations between the subjects was significant. There seemed to be no relation between the amount of rotations and the covered distance to localize a source in localization with the viewpoint. The rotations analysis did show, however, that of all possible rotations, the rotation about the vertical axis was performed most. Apparently, turning left-to-right provided the most useful changes in the perceived sound. It was observed during the experiments, that some subjects tended to face the source's (apparent) lateral position, followed by a translation in the direction of the sound source. This would also explain the preference of the subjects to use front-back translation rather than other translations.

Although there were quite a few differences between the various configurations (stereo-vision or mono-vision, with or without visualized sources), the changes in the quantities did not show

any correlation with the changing circumstances. Even for individual subjects, the differences in the amounts of rotations for the different configurations were not consistent.

The rotations analysis in case of localization tasks with the cursor showed the same tendencies as in case of localization with the viewpoint. However, it was observed during the experiments, that subjects tended to align the cursor parallel to one of the axes of the viewpoint coordinate system (which had the same orientation as the world coordinate system), and often even brought the cursor back to its original orientation, so it had the same orientation as the viewpoint again. Apparently, it was difficult for the subjects to displace themselves mentally into a cursor with some other orientation.

From these observations, it can probably be concluded that the method used to rotate the cursor does not provide an easy, intuitive relation between these rotations and the (changes in the) perceived sound; most likely, a more direct relation between the subject's own head and the cursor is needed. This can perhaps be accomplished by the use of a head-tracking device: the rotations of a subject's head would result in exactly the same rotations of the cursor, while translations could still be performed with some 'indirect' input device, such as a joystick.

5.3 Session II: Cursor Position Estimation

All experiments in this session were performed with a static viewpoint, stereo-vision enabled, and only the use of the translation joystick to move the cursor, where the cursor had the same orientation as the viewpoint.

In the first series, two visible sound sources emitting (different) familiar sounds were used and the subject heard what the cursor heard. To begin with, the subjects got a few minutes to move around the room, getting used to the stimuli and the way the sound was perceived by the cursor at different positions. Then, the cursor had to be returned to approximately its original position (the center of the room), whereupon the cursor was made invisible and 'randomly' located somewhere within the room (still invisible). Now, the subject had to guess the position of the cursor in the room and indicate it on a drawing of the room, i.e. a ground-plan and cross-section. To facilitate things slightly, the number of positions where the cursor could be located was limited to the 27 raster points inside the room.

The used sounds and their sources' positions were "front" and "back" with locations diagonal (Left-Bottom-Back and Right-Top-Front), and "high" and "low" located in front of the back wall with a 3 m vertical position difference (Center-High-Back and Center-Low-Back). For both set-ups, four cursor positions had to be estimated.

In the second series of experiments in this session, the *cursor* was emitting sound (again the word "here") and heard by the user from the viewpoint. After a few minutes of free cursor manipulation, the cursor was again placed at the center of the room, made invisible, and 'randomly' located at four raster positions within the room. Again, the subject had to guess the position of the cursor. Next an extra (visible) sound source was positioned at the center of the room emitting the word "center" and the same procedure was followed. This was done in order to see if this (static) sound source could be useful as a reference position.

5.4 Results of Session II

According to the subjects, the experiments of Session II were more difficult than those of Session I. In the case of two sound sources placed diagonally and vertically, left-right judgments of the cursor's position were most accurate. Low-high and front-back judgments were less accurate, where it was remarkable that in case of diagonal source positions, low-high judgments were considerably worse than for the vertical positioned sources. An explanation for this might be that the subjects realized that because in the latter case the sources should have been of 'equal loudness' for any distance, a difference in loudness was used for the judgment of height, while for the diagonal sources, such a relation between loudness and height was less well defined (if one sound was perceived louder than the other, it could also have been caused by a left-right or front-back displacement).

The results of the final two cursor position estimation tasks showed a poor overall performance. Indeed, some of the results would have been better if the subjects would have taken a blind guess (or better, a 'deaf guess'). It must be concluded that the current spatial sound model does not provide sufficient cues for judgments in the front-back and low-high dimensions under these circumstances.

There was, however, a noticeable difference between the two configurations, particularly in the left-right judgments: In the case where the reference sound source was present, left-right judgments were far more accurate than without that source. Apparently, subjects were unable to estimate the cursor's position with only the cursor as a sound source, even though there was a difference in the sound at different (left-right) positions. Adding a static sound source enhanced the perception of the differences, and, being aware of the position of this reference sound source, these differences could be used for a (left-right) position discrimination.

5.5 Some More Observations

During the experiments several observations were made which were not discussed in the previous sections, but are interesting enough to mention. First, a few words about externalization.

Unfortunately, it was reported by the subjects that most of the time, the sounds were perceived inside the head. At best, the subjects sometimes heard the sources at some position outside, but still very close to the head, mostly somewhere on their foreheads. So, the simple spatial sound model used in the *SAGA* system cannot establish the "out of the head sensation", even though it is complemented with (indirect) head movements, reverberation, and distance dependent intensity. Other research, however, has indicated the possibility of externalized sounds, even with fairly simple models. For instance, the system used for experiments on active sound source localization described in [LHC90] used a fairly simple model to simulate spatial sound, but it was reported that most of the subjects did indeed perceive the sound as out of the head. One of the major differences between the two systems is the way head motions are implemented. Whereas the *SAGA* system only provides indirect head motions, to be invoked by the subject through the use of the joysticks, the system described in [LHC90] made use of a head-tracking device to incorporate head rotations and translations.

As mentioned before, subjects had difficulty placing themselves mentally into the cursor if the cursor had some orientation other than incident with the subject's head (i.e. the viewpoint). It was observed that the subjects sometimes could not clearly determine what orientation the cursor had. Again, this is an indication that the used method for cursor rotations is far from ideal if the perceived sound is related to the cursor's orientation.

Another interesting observation is that all subjects now and then moved their own head in an attempt to get a more accurate estimation of where a sound source was located. Whereas in free field localization such movements do result in a better sense of the sound's direction, in this case it (obviously) had no effect, and most subjects clearly showed some sign of disappointment or frustration as they realized this.

Considering these observations, and the analysis of the experiments, we suggest that adding a head tracking device to the system used here could be an important improvement. In case of cursor localization, the cursor's orientation could be directly coupled to the orientation of the subject's head. With localization with the viewpoint, there would still be the need for some other interfacing device to enable larger rotations (looking around), but small movements of the head, made by the user to get a better impression of the sound's direction, could be incorporated by the use of such a device. In both cases, the result would probably be an enhancement of the out of the head sensation, as well as an improvement in localization speed and accuracy for any user of the system.

6 Conclusion

In this paper we addressed the question whether (simple) spatial sound systems could be useful for three-dimensional graphical applications. This was narrowed down to two issues: the active (indirect) localization of virtual sound sources and the position estimation of a three-dimensional cursor with the aid of spatial sound.

The results obtained from the experiments performed with the *SAGA* system clearly showed that spatial sound systems (even if they are as simple as the *SAGA* system) can be used as an aid in finding interesting objects, points, or areas in visualization applications. Comparing localization performance between 'viewpoint listening' and 'cursor listening' showed that the localization with the cursor was no more difficult than localization with the viewpoint.

The use of spatial sound for cursor position determination in three-dimensional environments is less feasible with the *SAGA* system. The results obtained from the experiments on this issue are not very encouraging. As it is concluded from the localization experiments that subjects primarily used the changes in the perceived sounds, better results in cursor position determination with the *SAGA* system can probably be obtained if some form of motion is incorporated, for instance by the use of moving sound sources or by allowing the user to rotate the cursor.

References

[BB77] R.A. Butler and K. Belendiuk. Spectral cues utilized in the localization of sound in the median sagittal plane. *Journal of the Acoustical Society of America*, 61:1264–1269, 1977.

[Bur92] D.A. Burgess. Techniques for low cost spatial audio. In *Proceedings of UIST '92*, pages 53–59, November 1992.

[Col62] P.D. Coleman. Failure to localize the source distance of an unfamiliar sound. *Journal of the Acoustical Society of America*, 34(3):345–346, March 1962.

[Col63] P.D. Coleman. An analysis of cues to auditory depth perception in free space. *Psychological Bulletin*, 60:302–315, 1963.

[Col68] P.D. Coleman. Dual role of frequency spectrum in determination of auditory distance. *Journal of the Acoustical Society of America*, 44(2):631–632, 1968.

[Edw89] A.D.N. Edwards. Soundtrack: An auditory interface for blind users. *Human Computer Interaction*, 4(1):45–66, 1989.

[Gar68] M.B. Gardner. Proximity image effect in sound localization. *Journal of the Acoustical Society of America*, 43:163, 1968.

[Gar69] M.B. Gardner. Distance estimation of 0° or apparent 0° oriented speech signals in anechoic space. *Journal of the Acoustical Society of America*, 45(1):47–53, 1969.

[Gar73] M.B. Gardner. Some monaural and binaural facets of median plane localization. *Journal of the Acoustical Society of America*, 54(6):1489–1495, 1973.

[Gav89] W.W. Gaver. The sonicfinder: an interface that uses auditory icons. *Human-Computer Interaction*, 4(1):67–94, 1989.

[Hen74] G.B. Henning. Detectability of interaural delay in high-frequency complex waveforms. *Journal of the Acoustical Society of America*, 55:84–90, 1974.

[LHC90] J.M. Loomis, C. Herbert, and J.G. Cicinelli. Active localization of virtual sounds. *Journal of the Acoustical Society of America*, 88(4):1757–1764, October 1990.

[Mil72] A.W. Mills. Auditory localization. In J.V. Tobias, editor, *Foundations of Modern Auditory Theory, Volume II*, chapter 8, pages 303–348. Academic Press, New York, 1972.

[MK75] D.H. Mershon and L.E. King. Intensity and reverberation as factors in the auditory perception of egocentric distance. *Perception and Psychophysics*, 18:409–415, 1975.

[MMG89] J.C. Middlebrooks, J.C. Makous, and D.M. Green. Directional sensitivity of sound-pressure levels in the human ear canal. *Journal of the Acoustical Society of America*, 86(1):89–108, July 1989.

[SBDC76] C.L. Searle, L.D. Braida, M.F. Davis, and H.S. Colburn. Model for auditory localization. *Journal of the Acoustical Society of America*, 60(5):1164–1175, November 1976.

[SBG90] S. Smith, R.D. Bergeron, and G.G. Grinstein. Stereophonic and surface sound generation for exploratory data analysis. In *Proceedings of the CHI '90, ACM Conference on Human Factors in Computing Systems*, pages 123–132, April 1990.

[SC91] C. Scaletti and A.B. Craig. Using sound to extract meaning from complex data. In *Proceedings SPIE, 1459*, pages 207–219, 1991.

[SWKE89] R.D. Sorkin, F.L. Wightman, D.J. Kistler, and G.C. Elvers. An exploratory study of the use of movement-correlated cues in an auditory heads-up display. *Human Factors*, 31:161–166, 1989.

[TMR67] W.R. Thurlow, J.W. Mangels, and P.S. Runge. Head movements during sound localization. *Journal of the Acoustical Society of America*, 42(2):489–493, 1967.

[TR67] W.R. Thurlow and P.S. Runge. Effect of induced head movements on the localization of direction of sounds. *Journal of the Acoustical Society of America*, 42(2):480–488, 1967.

[Wat78] A.J. Watkins. Psychoacoustical aspects of synthesized vertical locale cues. *Journal of the Acoustical Society of America*, 63:1152–1165, 1978.

[Wen92] E.M. Wenzel. Three-dimensional virtual acoustic displays. In M.M Blattner and R.B. Dannenberg, editors, *MultiMedia Interface Design*, chapter 15, pages 257–288. ACM Press, New York, 1992. Condensed version of "Localization in Virtual Acoustic Displays", PRESENCE: Teleoperators and Virtual Environments, Vol. 1 No. 1, March 1992.

[WK89a] F.L. Wightman and D.J. Kistler. Headphone simulation of free-field listening. I: Stimulus synthesis. *Journal of the Acoustical Society of America*, 85(2):858–867, 1989.

[WK89b] F.L. Wightman and D.J. Kistler. Headphone simulation of free-field listening. II: Psychophysical validation. *Journal of the Acoustical Society of America*, 85(2):868–878, 1989.

[WNR49] H. Wallach, E.B. Newman, and M.R. Rosenzweig. The precedence effect in sound localization. *American Journal of Psychology*, 62:315–336, 1949.

[WPFF90] E.M. Wenzel, P.K.Stone, S.S. Fisher, and S.H. Foster. A system for three-dimensional acoustic 'visualization' in a virtual environment workstation. In *Proceedings of the IEEE Visualization '90 Conference*, pages 329–337, October 1990.

[WWK91] E.M. Wenzel, F.L. Wightman, and D.J. Kistler. Localization with non-individualised virtual acoustic display cues. In *Proceedings of the CHI '91, ACM Conference on Computer-Human Interaction*, pages 351–359, April 1991.