Intelligibility of Speech in a Virtual 3-D Environment

Justin A. MacDonald and **J. D. Balakrishnan,** Purdue University, West Lafayette, Indiana, and **Michael D. Orosz** and **Walter J. Karplus,** University of California, Los Angeles, Los Angeles, California

In a simulated air traffic control task, improvement in the detection of auditory warnings when using virtual 3-D audio depended on the spatial configuration of the sounds. Performance improved substantially when two of four sources were placed to the left and the remaining two were placed to the right of the participant. Surprisingly, little or no benefits were observed for configurations involving the elevation or transverse (front/back) dimensions of virtual space, suggesting that position on the interaural (left/right) axis is the crucial factor to consider in auditory display design. The relative importance of interaural spacing effects was corroborated in a second, free-field (real space) experiment. Two additional experiments showed that (a) positioning signals to the side of the listener is superior to placing them in front even when two sounds are presented in the same location, and (b) the optimal distance on the interaural axis varies with the amplitude of the sounds. These results are well predicted by the behavior of an ideal observer under the different display conditions. This suggests that guidelines for auditory display design that allow for effective perception of speech information can be developed from an analysis of the physical sound patterns.

INTRODUCTION

The number of simultaneously active sound feeds often limits the ability of human operators to interpret and respond to auditory messages in electronic communication systems (e.g., telecommunications). Cockpits and air traffic control rooms, military and police radio communications, and teleconferencing are examples of common situations in which many different sounds might become important at one time or another and the listener must be able to selectively attend to one or more of them while entirely or partially tuning out some or all of the others. Controlling volume levels of different channels in the system, following or enforcing communication protocols, and directing attention efficiently to a given sound ensemble can be especially difficult during the chaotic interchanges typical of a mission-critical or crisis situation.

Recently, the three-dimensional (3-D) auditory displays of virtual reality have been advocated as a means of facilitating this multichannel listening process (e.g., Begault & Wenzel, 1993; Burdea, Richard, & Coiffet, 1996; Doll & Hanna, 1995; Doll, Hanna, & Russotti, 1992; King & Oldfield, 1997; Noro, Kawai, & Takao, 1996; Ricard & Meirs, 1994). Instead of piping all sound channels directly into speakers or earphones at equal volumes, a specialized sound card is used to create the illusion that the different sounds in the system originate from different locations in the space surrounding the listener. The 3-D effect is achieved by first introducing a time delay to simulate the interaural arrival time differences associated with the different distances of a sound from the left and right ear. The two ear channels are then subjected to a series of preprogrammable filters (the head-related transfer function, or HRTF) to simulate the effects of the head, pinnae, and torso on a

Address correspondence to Justin A. MacDonald, 1364 Psychological Sciences, Purdue University, West Lafayette, IN 47907; macdonja@purdue.edu. **HUMAN FACTORS,** Vol. 44, No. 2, Summer 2002, pp. 272–286.

waveform (e.g., Begault, 1994; Begault & Wenzel, 1993; Wenzel, Arruda, Kistler, & Wightman, 1993). Differences in the sound pressure level and arrival times of the waveform at the two ears provide lateral direction cues, and effects of anatomical structure (e.g., pinnae shape) provide information about elevation and position on the transverse (front/back) axis. The effects of head shadow on the intensity of stimuli at the ears become more pronounced in higher-frequency regions, whereas the interaural time difference (ITD) is more noticeable at lower frequencies.

Separating sounds in both real and virtual spaces has been shown to increase the intelligibility of speech paired with a noise masker (e.g., Doll & Hanna, 1995; Ricard & Meirs, 1994; Saberi, Dostal, Sadralodabai, Bull, & Perrott, 1991) and with interfering speech (e.g., Dirks & Wilson, 1969; Ericson & McKinley, 1997; Yost, Dye, & Sheft, 1996). As one might expect, the most effective 3-D simulation is usually obtained using HRTFs that take into account the unique anatomy of the listener, including the size of the listener's head and the shape of the pinnae. In many applied situations, these individualized filters would be impractical. Fortunately, however, nearly equivalent results can be achieved with a single nonindividualized set of filters (Wenzel et al., 1993), which can be derived from a model of the human head such as the Knowles Electronics Mannequin for Acoustics Research (Knowles Electronics, Inc., Itasca, Illinois) or from a listener who is particularly good at localizing sounds (Begault & Wenzel, 1993). Begault and Wenzel showed that the performance of poor sound localizers can be increased by replacing their HRTFs with those of an exceptionally good localizer.

The main disadvantages of nonindividualized HRTFs are increased frequency of front/back reversals and more difficulty in simulating elevation (Wightman & Kistler, 1989). The same kinds of benefits for 3-D sound displays over monophonic listening, however, have been observed using both individualized and nonindividualized methods. It seems likely, therefore, that any general design issues discovered using nonindividualized HRTFs would apply just as well to individualized systems. In the present study we used nonindividualized HRTFs to examine some potential general rules relating spatial layout of a sound display to speech intelligibility. Although the interaural intensity difference (IID) is likely to be less of a factor at the frequency ranges common to speech, some effect still exists, and any observed effect of location on speech intelligibility would presumably be a result of a combination of the IID and ITD.

Performance Effects of 3-D Sound Simulation

Because the intelligibility of a signal depends on its amplitude at the two ears and this incident amplitude decreases with the distance of the source from the listener, investigators typically compare 3-D with traditional display conditions only when the distances of the signals from the center of the head are held constant. Coordinates of the sounds are usually also restricted to points on the horizontal plane passing through the two ears. When one of the two sources is placed on the interaural axis (right or left of the head), thresholds are lower when the angle between the two sources is at least 90° (Dirks & Wilson, 1969; Ericson & McKinley, 1997; Tonning, 1971). However, when either the signal or the masker is placed in front of the listener (0° azimuth), detection thresholds decrease until the angle between the signal and masker is roughly 90° and then appear to increase again from 90° to 180° (Bronkhorst & Plomp, 1988, 1990, 1992; Decroix & Dehaussy, 1964; Duquesnoy, 1983; Plomp & Mimpen, 1981; Saberi et al., 1991; but see Ricard & Meirs, 1994, for an exception and noticeable individual differences). Thus increased spacing may be expected to improve performance in some conditions, but maximizing the spacing between two sources apparently is not the optimal general rule for sound placement.

Even if more were known about the effects of simulating space between a signal and a noise source, it is not obvious that this information could be used to predict behavior when there are more than two talkers, as is typical in air traffic control and many other workplace environments. Two recent studies (Ericson & McKinley, 1997; Yost et al., 1996) have attempted to address this issue by measuring the intelligibility

of three or four speech sources as a function of their separation in virtual 3-D space. Although both studies show that increasing separation along the azimuthal plane leads to increased intelligibility, these results are relatively limited in their application because they tested only a few spatial configurations.

The first experiment in the current study remedied this problem by investigating the effects of a wide variety of spatial configurations on the intelligibility of four speech sources. Participants performed a simulated air traffic control task with four speech channels and a visual graphical display of aircraft moving through airspace. Experiment 2 compared the effects of spacing on the transverse (front/back) axis versus that on the interaural axis in free-field (real space) listening in order to show that the main conclusions of Experiment 1 should not be attributed to limitations in the 3-D simulation techniques. The last two experiments tested the hypothesis that performance levels for different lateral positions can be predicted from an analysis of the quantifiable information content of the physical stimulus patterns.

EXPERIMENT 1

Experiment 1 was designed to reproduce some basic elements of flight test control operations at the National Aeronautics and Space Administration (NASA) Dryden Flight Research Center. The NASA Dryden facility is responsible for carrying out most in-flight test simulations of experimental aircraft designs currently supported by NASA. In a typical test the pilot has a schedule of maneuvers to perform, and several flight test engineers simultaneously monitor a number of communications and data streams, which may arrive over audio channels or appear on video monitors. Critical decisions during the flight are made by the chief engineer (NASA-1), who is the only ground personnel member permitted to communicate directly with the pilot. NASA-1's communication system has four audio channels: radioed signals from the test pilot, radioed signals from the pilot of a chase plane, intercom communications from other test engineers in the control room, and radio communications from the air traffic control tower and other downrange radar positions. Based on visual graphic displays, lookup tables, and communications from the pilots, NASA-1 must identify stages and key events occurring during the flight test and decide as quickly as possible whether to continue, abort the mission, or instruct the pilot to eject from the aircraft.

Visual components of the flight control task were simulated in the experiment by presenting a two-dimensional graphical image of the NASA Dryden Flight Research Center airspace and several moving aircraft that the participant could redirect. A monochrome snapshot of this visual graphic is shown in Figure 1. Four audio channels containing continuous aviation-related dialogues were played continuously under 1 of 19 display configuration conditions, including several conditions in which one of the four sound sources was continuously moving in a horizontal or vertical direction, as they would in free-field listening if the participant could move his or her head. Except for the *center condition,* in which all sounds were presented at equal volumes to both ears, each configuration was run with and without a visual graphic insert on the screen showing the spatial layout of the sounds. Participants were required to detect the utterance of a warning and also to identify the talker as quickly as possible while at the same time ensuring that aircraft appearing on the graphic display avoided the regions designated as restricted airspaces.

Method

Participants. A group of 30 volunteers consisting of test pilots, flight test air controllers, and test engineers from the NASA Dryden Flight Research Center at Edwards Air Force Base, as well as several graduate students from UCLA, participated in three separate 1-h sessions. All participants reported normal hearing and normal or corrected-to-normal vision. Pilots often have some degree of hearing loss, and it is possible that they were unaware of some deficiencies. However, previous studies have shown that a bilateral hearing loss decreases overall localization performance but preserves the trends exhibited by normal listeners (Hay, 1996).

Apparatus. Visual graphics were presented on an 11-inch (28-cm) active matrix color screen controlled by a laptop computer running at

640 × 480 resolution. The audio subsystem was run on a separate computer system. Communications between the two subsystems was via a custom-designed serial cable. A Crystal River Engineering Alphatron sound card (Crystal River Engineering, Fremont, California) simulated 3-D auditory positions over a pair of AKG Acoustics K-240M stereo headphones (AKG Acoustics, Vienna, Austria). The HRTFs used in the Alphatron were measured from an adult head and are available through Chapin (2001).

Stimuli. The auditory stimuli were generated from monaural sound files recorded in 16-bit format with a 22-kHz sampling rate. Each file consisted of prerecorded speech segments commonly heard in air traffic control environments. Each segment was uttered by four speakers, one female and three male. The test bed image shown in Figure 1 was presented on the laptop LCD screen together with menu options to report a warning and to identify the warning speaker. In the *graphic condition,* the simulated positions of the four speakers were represented as rectangular graphic icons. These icons were positioned in the graphic display at locations that corresponded to the horizontal position of each speech source relative to the location of the test participant (see graphic icons labeled "Mike," "Nina," etc., in Figure 1).

Procedure. At the beginning of each session participants received instructions about the concurrent visual and auditory tasks and the operation of the mouse. Recorded experimental trials began after a short practice session. Participants were asked to monitor and respond to both routine (movement of aircraft icons on the visual graphic) and nonroutine events (warning utterances). The four graphic aircraft icons moved continuously toward the center of the display, and the participants attempted to prevent any single aircraft from reaching the display center and also to prevent two or more aircraft from occupying the same restricted airspace. When an aircraft entered a restricted region, the color of the region's border changed from green to red and a red crosshatch fill pattern covered the interior. If another aircraft entered the region while it was already occupied, the red-crosshatched region began to blink

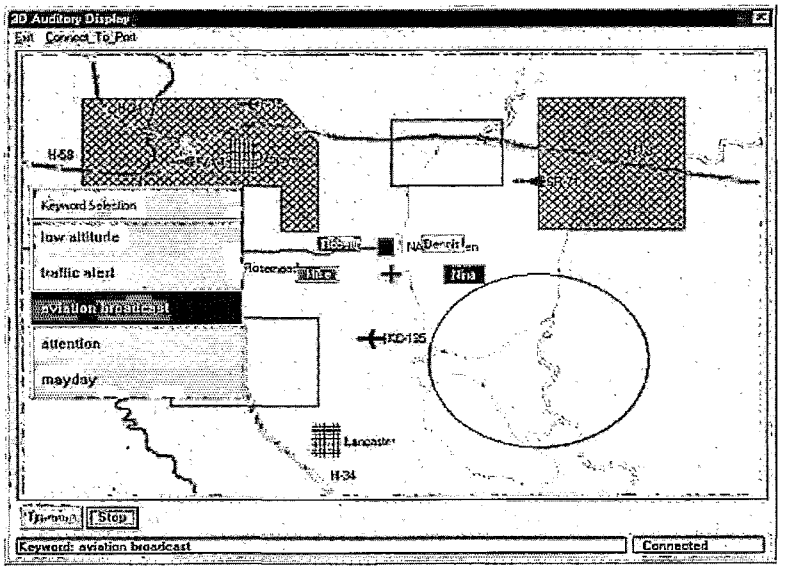


Figure 1. An image captured from the graphic display interface used in Experiment 1 to simulate visual and auditory components of air traffic control. The solid rectangles are restricted airspaces that should contain no more than one aircraft icon at any given time. Participants are also required to keep the aircraft from reaching the center of the airfield ("NASA-Dryden") while monitoring four speech signals for warnings. The computer mouse is used to open a menu box and to highlight one of the five warning sounds and then to choose one of four speakers.

on and off, indicating a violation of airspace. To avoid airspace violations, participants repositioned aircraft by using the mouse to select and drag them to a safe location. Once an aircraft icon was repositioned, the program recalculated a movement vector to the center of the graphic display space.

Speech trains from each source (three males, one female) consisted of 15 prerecorded speech segments commonly heard in air traffic control environments. Five of these segments contained a warning ("low altitude alert," "traffic alert," "aviation broadcast," "attention all aircraft," or "mayday, mayday, mayday"). Speech segments for the four speakers were selected at random from the 15 samples with the constraint that only one sample contained a warning. Sequences were also selected so that each speaker uttered each of the five warnings only once during a test (i.e., warnings were never repeated by the same speaker during a test trial). The experimental trials could therefore be divided into sequences of 20 warning presentations. Participants were instructed to select the detected warning from a list of warnings (maintained in the graphic display space) and then identify the source (speaker) of the warning.

Participants performed the task in each of 19 spatial display configurations. Figure 2 summarizes the different spatial conditions tested (the center condition is not shown). In addition to differences in spacing, 10 of the 11 conditions included in Configuration 4 incorporated a single moving speech source. This moving speech source maintained the same distance from the center of the head while subtending angles ranging from 15° to 360° in the horizontal or vertical plane (see Figure 2). Configurations 7 and 8 "tilted" the sounds in the front and back locations so that one source was located 45° above the azimuthal plane and the other was located 45° below the azimuthal plane. Except for a single condition in which all sounds were located at the center of head, all test conditions were repeated twice, once with a graphic representation of the speech sources and once without it. The resulting 37 conditions (18 with a graphic, 19 without a graphic) were divided into three blocks, and the presentation order of these blocks was counterbalanced across participants. All sounds (with the exception of those in the center configuration) were simulated to appear 10 m from the center of the head. The standard formula for the attenuation of sound over distance was used to simulate distance cues.

Data analysis. Performance measures were the number of warnings correctly detected in a given listening condition, f_d ($f_d \leq 20$, given that there were 20 warning presentations), and the mean response time (RT; i.e., the mean of the f_d detection times). False alarms were rare (475 occurrences in total, compared with 22 195 true warning trials) and were therefore ignored. Response time was also undefined for unreported warnings (misses). The 37 listening conditions were treated first as a single independent variable and were then divided into subgroups to test for effects of number of sound locations, motion, tilt, and the presence/absence of the configuration graphic.

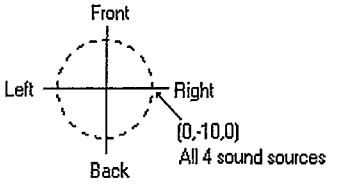
Results and Discussion

Warning detection rates and mean RTs for each display configuration are given in Table 1. RTs showed a strong positive correlation with detection rates and were therefore not included in the analysis. A one-way repeated-measures analysis of variance (ANOVA) revealed a significant effect of configuration on detection rate, $F(36, 1044) = 3.082$, $p < .001$. No significant differences were observed between conditions with and without a graphic icon ($p > .5$). Tilting and motion in the display (Configurations 7 and 8 in Figure 2) also had no significant effect on performance, $F(1, 29) = 0.534$, $p > .5$ for tilted versus horizontal; $F(1, 29) = 3.611$, $p > .05$ for moving versus static displays. The main effect of configuration was apparently predominantly attributable to differences between conditions in which the four sounds were separated into at least two static locations versus those with only one location, $F(1, 29) = 77.428$, $p < .001$. In fact, performance in the spatialized displays was higher even than that in the center condition, which was considerably louder because of the different simulated distance (zero) from the head (see Table 1).

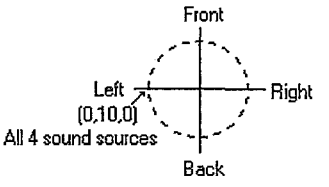
It is conceivable that asymmetrical hearing loss in our participants resulted in reduced spatialization ability, thereby decreasing performance in all spatialized configurations. Even

if this were the case, a clear increase in performance was observed when sounds were separated into at least two locations. This result indicates that even if our participants' spatialization ability was compromised by asymmetrical hearing loss, separating the sounds in space still led to a highly significant increase in identification performance.

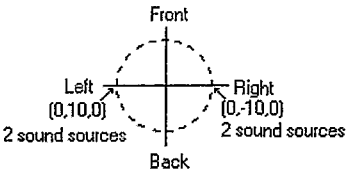
Assuming that the ability to segment the sound stream is somehow facilitated by selective attention to positions in perceptual space, performance levels should have been highest when the four sounds were placed in four different locations. Instead, they were higher in the two-location conditions, a difference that was marginally significant, $F(1, 29) = 3.822$,



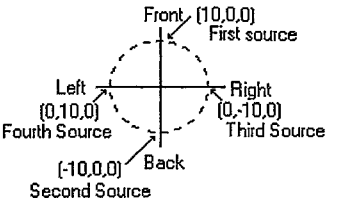
Configuration 1



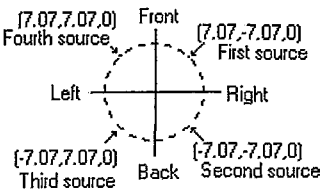
Configuration 2



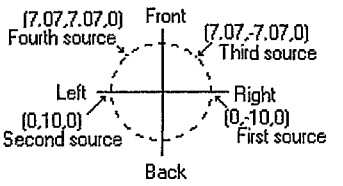
Configuration 3



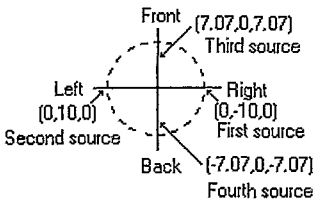
Configuration 4



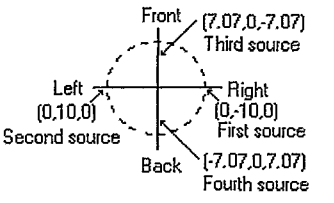
Configuration 5



Configuration 6



Configuration 7



Configuration 8

Figure 2. Location configurations used in Experiment 1. The configuration for center of the head is not shown. Each diagram includes the *x*, *y*, and *z* coordinates in meters for each of the four sound sources. Configuration 4 represents a combination of 11 conditions, 10 with different types of motion and 1 stationary condition.

TABLE 1: Proportion of Warnings Detected and Mean Time to Respond in Experiment 1 When a Visual Graphic of the Sound Configuration Was Present or Absent

	Graphic		No Graphic	
Configuration	Proportion Correct	Mean RT (s)	Proportion Correct	Mean RT (s)
Center	N/A	N/A	.752	4.891
1	.712	5.469	.703	6.027
2	.688	5.668	.692	5.643
3	.798	4.464	.827	4.314
4 (Stationary)	.797	4.582	.787	4.625
4 (Motion)	.784	4.710	.799	4.694
5	.813	4.412	.803	4.553
6	.840	4.308	.808	4.482
7	.773	4.806	.770	4.772
8	.775	4.858	.788	4.877

Note: Scores in the motion conditions of Configuration 4 are combined over 10 movement patterns, including horizontal motion within 20° in front, back, or both front and back sources; vertical motion within 20° in front, back, or both sources; and complete rotation (360°) of sounds in front and back of the listener. Except for the center condition (all sounds inside the head), the four sounds were simulated 10 m from the center of the head.

$p = .06$. Most previous studies have compared configuration conditions with only two sounds, so it is possible that the advantages of spatialized displays simply do not extend beyond two simulated locations. However, these earlier studies have also compared spatialized with nonspatialized displays by moving one of the two sources around the azimuth (so that distance from the listener is constant). As a result, the degree of separation on the transverse axis was always correlated with separation on the interaural axis. It is possible, therefore, that our four-location conditions were no different from the two-location condition because spatialization on the transverse and elevation axes is generally ineffective (i.e., the segmentation by spatial attention hypothesis doesn't hold for these axes). It is also important to consider the possibility that our simulation of sounds in front and back or above and below the listener was simply not good enough to facilitate distributed spatial attention.

The next two experiments were designed to address these issues. In Experiment 2 we used pairs of speech sounds presented in real 3-D space to compare the effects of spacing on the transverse axis with that of spacing on the interaural axis under conditions in which the physical spacing effects were unambiguous. Experiment 3 used the same speech sounds presented in simulated positions around the listener to show that even when sounds emanate from a single location, their position on the interaural axis is still the crucial factor determining their intelligibility.

EXPERIMENT 2

To determine whether the results of Experiment 1 could be attributed to limitations of the 3-D simulation or whether the benefits of 3-D audio displays are predominantly or even entirely attributable to the spacing of the sounds on the interaural axis, participants in Experiment 2 were asked to identify two simultaneous speech sounds (letters of the English alphabet) that emanated from external speakers positioned either in the same location (both to the left, both to the right, both to the front, or both in back) or in two different locations (one on the left and one on the right, or one in front and one in back). From the results of Experiment 1, we would expect the participants to perform better when the sounds are lateralized to the left and right, compared with when they are both on the left or both on the right, but equally well in all conditions involving spacing only on the transverse axis.

Method

Participants. Six participants each completed four 1-h sessions and were compensated at the

rate of $6.50 per session. All participants reported normal hearing and normal or corrected-to-normal vision.

Apparatus. Two personal computers presented stimuli over two pairs of Altec Lansing (Altec Lansing Technologies, Inc., Milford, Pennsylvania) ACS340 speakers, one pair of speakers connected to each computer. The experimental control graphics were displayed on a 17-inch (43 cm) color monitor (1024 × 768 resolution), and responses were given on a standard 104-key keyboard.

Stimuli. The speech signals were vocalizations of the 26 letters of the alphabet produced by a single male speaker and recorded in a monaural 16-bit format with a 22-kHz sampling rate. The length of the files ranged from 450 to 700 ms.

Procedure. The experiment was divided into four blocks, one for each session. In each block the computer speakers were situated either on the transverse (two sessions) or interaural axis (two sessions). The axis location of the speakers was alternated from block to block, and the ordering of the blocks was counterbalanced across participants.

The four speakers were positioned in pairs facing toward the participant's head, either directly in front and directly behind or to the left and right, at approximately 38 cm from the center of the participant's head. Separation between the centers of adjacent speakers in a pair was approximately 8 cm. Each trial consisted of a randomly selected pair of two letter sounds from the English alphabet presented simultaneously. The letter sounds were always presented from only two of the four speakers. In the *separate condition,* one speaker from each pair was used to present the sounds so that one sound emanated from the left (or front) of the participant and the other from the right (or behind). In the *together condition,* both letter sounds were presented using speakers from the same pair (one letter from each speaker in the pair) and hence originated from virtually the same physical location. The result was a 2 × 2 design, with the two variables being location axis (interaural or transverse) and proximity (together or separate). In all four conditions the letter sounds were presented at a mean sound level of 54.2 dB(A).

Participants pressed the Enter key to begin each trial and indicated which letters were presented by pressing the appropriate keys on the keyboard. Nonletter key presses were not accepted. Participants were asked to respond as quickly as possible while maintaining a high level of accuracy. Collapsed across participants, a minimum of 4489 trials were completed in each condition.

Results and Discussion

Pooled across participants, mean proportion correct scores for the four conditions are shown in Figure 3. A repeated-measures ANOVA revealed a significant effect of the axis of presentation, $F(1, 5) = 29.509$, $p < .01$, as well as proximity, $F(1, 5) = 11.983$, $p < .05$. The Axis × Proximity interaction was also significant, $F(1, 5) = 25.774$, $p < .01$. Overall performance was substantially lower in the transverse axis condition than in the interaural condition, suggesting that when distance from the listener is constant, positioning sounds to the left or right is superior to positioning them in front or back of the listener, even if the 3-D simulation on the transverse axis is ideal. Separating the sounds on the transverse axis also had virtually no effect on performance, whereas separating them along the interaural axis increased the identification rate. This result is consistent with the results of Experiment 1, indicating that configurations that rely on separation along the interaural axis lead to better identification rates than do those with separation along the transverse axis (e.g., Configuration 6 vs. Configuration 4 in Figure 2).

The importance of the interaural dimension and the risks associated with spacing on the transverse axis apparently are not limited to artificial sound displays. The next two experiments examined why position on the interaural axis should be crucial and why positioning sound sources in the most natural location – in front of the listener – would be disadvantageous even when multiple sound sources are presented in the same location.

EXPERIMENT 3

When a speech sound is placed in front of the listener and a noise source is shifted from

the center to the left or right (i.e., toward one ear and away from the other), the signal-to-noise ratio will increase in one ear and decrease in the other. Doll and Hanna (1995) pointed out that this physical trade-off effect could be sufficient in itself to predict effects of source position on intelligibility if the listener bases his or her judgment entirely on the inputs from one of the two ears (i.e., if detection is monaural). However, Dirks and Wilson (1969) reported results suggesting that both ears are important in the detection of speech, as the obstruction of the ear either closest to or farthest from the speech signal resulted in impaired performance. More recent work in neural network modeling has suggested that the use of a single ear to localize sounds is not enough to achieve a level of performance comparable to that of using two ears (Janko, Anderson, & Gilkey, 1997).

Assuming that observers will combine information from the two ears, it can be shown that detection rates should still be expected to increase when the noise source is moved away from the signal and toward one of the two ears under some minimal assumptions about the nature of the ambient noise. Specifically, using the *ideal observer* as a measure of task difficulty and adopting the reasonable assumption that some internal noise exists in the auditory perceptual system, detection rates should also increase if the signal and/or noise sources are shifted from the front or back of the listener toward one of the two ears (see MacDonald, 1999).

To test this fundamental prediction about the role of information content in spatial displays, speech sounds and noise in Experiment 3 were both presented at the same simulated spatial location, and speech recognition performance was compared as the pair was rotated on the azimuth. If the effect of source spacing is related to differences in information content of the waveforms after they are transformed to create the spatial effect, rather than to any higher-level cognitive processes associated with more realistic listening environments, speech intelligibility should be highest when the sounds are located to the left or right of the listener and should decrease as they are moved toward the front of the head (holding distance constant). Any effects of position in this case could not be attributed to effects such as facilitated attention to different sounds in different spatial locations because both sounds are presented in a single location.

Method

Participants. Twenty-six participants from an introductory psychology course at Purdue University each completed a 1-h session in partial fulfillment of a course requirement. All participants reported normal hearing and normal or corrected-to-normal vision.

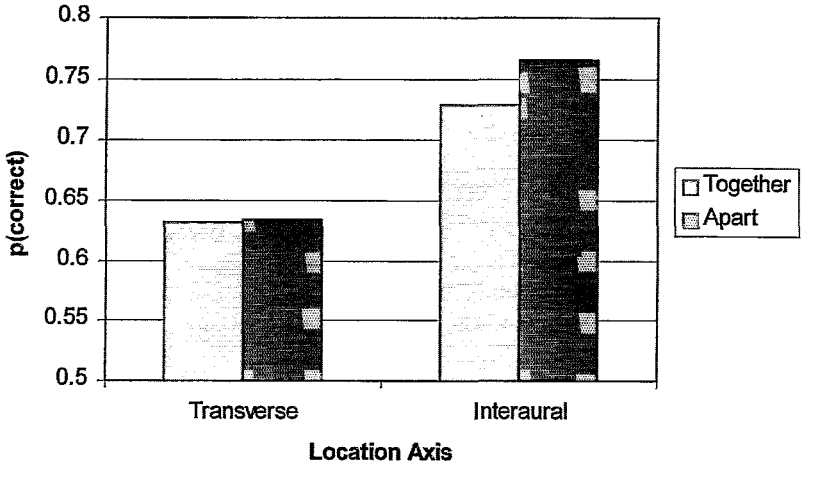


Figure 3. Mean proportion correct for each condition in Experiment 2. Participants were asked to correctly identify two letters during each trial.

Apparatus. A personal computer equipped with a Crystal River Engineering Alphatron sound card simulated 3-D auditory positions over a pair of AKG Acoustics K-240M stereo headphones. Visual graphics were displayed on a 13-inch (33 cm) color monitor (640 × 480 resolution), and responses were given on a standard 104-key keyboard.

Stimuli. The speech signals were the same as those used in Experiment 2. The 26 distractor stimuli were monaural 16-bit recordings of white noise recorded at a 22-kHz sampling rate for 750 ms.

Procedure. Each trial consisted of a single letter sound paired with a white noise source. Both the letters and noise files were chosen randomly from the set. Because the 26 white noise stimuli were longer in duration than any of the letter waveforms, the presentation of the letter waveform was always completed prior to termination of the white noise.

The letter and noise stimuli were presented in the same simulated location, which was chosen at random from nine positions on the ear-level horizontal plane: 0° azimuth (front), 45° azimuth (front right), 90° azimuth (right), 135° azimuth (back right), 180° azimuth (back), 225° azimuth (back left), 270° azimuth (left), 315° azimuth (front left), and center (inside the head). Except for the last condition, all locations were 38 cm from the center of the listener's head. Stimuli in the center location were presented at a mean level of 60 dB(A) at the output of the headphones, whereas all other stimuli were presented at a mean level of 53.2 dB(A) at the output of the headphones.

At the start of each trial, a visual graphic depicting the location of the upcoming sound pair was displayed on the computer monitor. Participants then pressed a "start" key, and the sound pair was presented. A response query was presented on the screen 1 s after the sound presentation. Participants indicated which letter was presented by pressing the appropriate key on the keyboard. Nonletter key presses were not accepted. Participants were asked to respond as quickly as possible while maintaining a high level of accuracy. After entering a valid response and receiving feedback about the correct response, participants pressed a key to proceed to the next trial. Collapsed across participants, a minimum of 1972 trials were completed in each location.

Results and Discussion

Pooled across all participants, proportion correct and mean RT for the nine locations are shown in Table 2. RTs showed a strong positive correlation with proportion correct and were therefore not included in the analysis. Identification rates were highest when the speech sounds were presented to the immediate right or left of the participant and lowest when they were presented in the front or back. Identification performance was highest in the center location, presumably a result of the relatively high sound level of the stimuli presented in this location (the mean level of the center-location stimuli was approximately 7 dB higher than that in the other locations). A repeated-measures ANOVA revealed a significant effect of location on identification rate, $F(8, 200) = 13.326$, $p < .001$, and a Scheffé post hoc test comparing proportion correct results by axis (transverse vs. interaural) was significant, $FS = 67.39$, $p < .01$. In fact, identification rates in the front and back conditions were significantly lower than that in four of the six lateral conditions ($HSD = 0.0544$, $p < .05$). The most "natural" position of a sound source – directly in front of the listener – was clearly the least effective with respect to intelligibility.

Effects of asymmetric hearing loss. Because the perception of 3-D sounds depends on input from both ears, symmetric hearing ability is an

TABLE 2: Proportion of Correct Identification Responses and Mean RT by Location in Experiment 3

Location	Proportion Correct	Mean RT (s)
Center	.440	1.281
Left	.371	1.385
Right	.379	1.276
Front left	.365	1.293
Front right	.356	1.341
Back left	.341	1.346
Back right	.338	1.459
Front	.294	1.445
Back	.296	1.483

important presupposition of our analysis. If a participant has experienced hearing loss in the left ear, for example, sounds simulated to the left of the head would be harder to perceive than sounds located to the right. An apparent advantage of lateralized over nonlateralized sounds could therefore be attributed to higher performance when the sounds are presented to the nondeficient ear that more than compensates for poorer performance when sounds are presented to the deficient ear. For example, if a pair of lateralized sounds is presented to the participant, the performance increase attributable to locating one of the sounds adjacent to the good ear could be greater in magnitude than the decrease in performance resulting from locating the other sound adjacent to the bad ear. To rule out this possibility, we compared performance in the left and right conditions for each participant to identify which ear might be deficient (i.e., the smaller of the two performance levels). Proportion correct in the lower of the two lateral conditions was .347, compared with .282 in the front condition ($HSD = 0.0544$, $p < .05$, with approximately 2000 trials in each condition). Thus it would be difficult to attribute the effects of location to hearing loss.

Effects of spatial location on information content. To see whether the relationship between position and recognition rates could be explained by differences in the information content of a signal arriving at the ears as a function of its position in space, we simulated the behavior of an optimal decision maker for each sound pattern and source location in Experiment 3. If the behavior of an ideal observer under the conditions of the experiment is similar to the behavior of human participants, it is reasonable to suppose that computational aspects of the task, rather than any unique properties of human auditory perception, will be the main factors to consider in auditory display design. For each of the nine location conditions, recordings were made of the letter sound output by the sound card at each ear, creating a set of 26 pairs of signal templates that incorporated all aspects of the spatial filter. Twenty-six white noise files were also recorded for each location at each ear to simulate the white noise background added in the experiment and the effects of filtering on these external stimuli.

An experimental trial was simulated by selecting a letter sound at random from within a given location set. A white noise pattern for the given location was then selected and added to the left ear template for the chosen letter sound. This process was repeated to simulate the propagation of the letter and noise sounds to the right ear, and the input to the decision maker was therefore two noise-added sound files, one for each ear. Finally, internal noise was simulated by adding another white noise pattern to each of these sound files.

The ideal observer's identification response was defined by computing the most probable letter stimulus given the input (i.e., the decision rule maximized the probability of a correct identification response on each trial and, hence, the proportion of correct responses across trials; see, e.g., Whalen, 1971). This was accomplished by computing the sum of squared errors between the noisy input stimulus and each of the 26 noise-free letter stimuli. For example, the squared error was calculated between the first sample of the noisy input and the first sample of the letter *A* stimulus. This process continued by computing squared errors for all subsequent pairs of samples, which were then summed to provide a measurement of the similarity between the noisy input and the letter *A* stimulus. The letter stimulus that most closely resembled the noisy input (i.e., the stimulus that resulted in the smallest sum of squared errors) was output as the most probable letter stimulus. The simulation program recorded whether this prediction was correct or incorrect and then proceeded to the next trial.

Results obtained from 90 000 simulation trials (10 000 in each location) are shown in Table 3. Because the level of internal noise in the simulation was chosen arbitrarily, the exact numerical values are not comparable to those from the experiment. The point of the analysis is to illustrate how the nine location conditions would be ordered on the basis of information content alone. The last two columns in the table compare the ranks of the configurations in the experiment with those of the simulation. Except for conditions that were not significantly different in the experiment or in the simulation results (e.g., left vs. right and front vs. back), the ideal observer predicts the order perfectly.

TABLE 3: Proportion of Correct Identification Responses by Location for the Ideal Observer and Comparison of the Rank Order with That of Human Performance in Experiment 3

Location	Proportion Correct	Rank Order	Rank in Experiment 3
Center	.907	1	1
Left	.524	2	3
Right	.521	3	2
Front left	.491	5	4
Front right	.497	4	5
Back left	.486	7	6
Back right	.488	6	7
Front	.475	8	9
Back	.466	9	8

EXPERIMENT 4

The results of the previous experiments suggest that the three dimensions of virtual auditory space should not be treated equally in audio interface design. Positioning along the interaural axis appears to be crucial, whereas placement on the other two axes may be relatively inconsequential, even when several sound channels exist in the display. Taking this thesis for granted, Experiment 4 examined the effects of two other design issues: the sound level and the distance of the sources from the head. If the optimal configuration is determined mostly by effects of position on information content, performance should increase and then decrease as a sound is displaced from the center to either side of the listener.

Method

Participants. Twenty-five participants from an introductory psychology course at Purdue University each completed a 1-h session in partial fulfillment of a course requirement.

Apparatus. Stimuli were generated using the same equipment as in Experiment 3.

Stimuli. The sound patterns were the same as those used in Experiment 3, with the exception that amplification was decreased in the low-amplitude conditions. The letter and white noise stimuli were not amplified in the high-amplitude conditions and were attenuated by 5 dB relative to their original level in the low-amplitude conditions.

Procedure. The procedure was the same as in Experiment 3 with the exception that stimuli were randomly sampled from a set of 14 location conditions: 10 simulated locations on a line parallel to the interaural axis and 20 cm in front of the head (200, 100, 50, 30, and 15 cm to the left and right of the midline) and 4 positions on the midline (20, 10, 5, and 0 cm from the center of the head). Amplitude was also randomly selected from two values on each trial. This resulted in approximately 600 trials for each cell. As in Experiment 3, all sounds were simulated at ear level.

Results and Discussion

Proportion correct and mean RT by location and amplitude are listed in Table 4. Reaction times showed a strong positive correlation with proportion correct and were therefore not included in the analysis. A repeated-measures ANOVA revealed a main effect of location on identification rate, $F(13, 312) = 30.1$, $p < .001$. The main effect of amplitude on identification rate was also significant, $F(1, 24) = 61.562$, $p < .001$. Identification rates increased with increasing amplitude. The Location × Amplitude interaction was not significant, $F(13, 312) = 1.265$, $p = .233$. Somewhat interestingly, however, in the low-amplitude condition identification rates appeared to increase and then decrease as the sounds moved away from the listener, whereas in the high-amplitude condition the scores simply decreased with distance from the head.

It is not surprising to find that identification rates begin to decrease at some point as speech sounds are located farther from the head (and hence the sound level decreases). More important is the dependence of this function on the amplitude of the source. When amplitude was high, the optimal position appeared to have moved closer to the head, despite the fact that overall performance levels were clearly higher. Intuitively, it would seem that placing sounds closer to the head should be more important when sound levels are relatively low.

To see whether this counterintuitive result could be attributed to changes in information content under different display configurations,

TABLE 4: Proportion of Correct Letter Identification Responses and Mean RT by Location and Amplitude in Experiment 4

	Proportion Correct		Mean RT	
Location	Low Amplitude	High Amplitude	Low Amplitude	High Amplitude
200, 20	.085	.174	1.434	1.479
100, 20	.133	.198	1.259	1.133
50, 20	.198	.221	1.225	1.363
30, 20	.255	.307	1.111	1.148
15, 20	.287	.311	1.188	1.142
0, 20	.249	.349	1.072	1.044
–15, 20	.289	.331	1.120	1.064
–30, 20	.247	.300	1.165	1.100
–50, 20	.191	.285	1.184	1.218
–100, 20	.147	.205	1.344	1.188
–200, 20	.088	.172	1.451	1.319
0, 10	.294	.360	1.110	1.074
0, 5	.336	.358	1.092	1.102
0, 0	.369	.377	1.069	1.210

the ideal observer was simulated once more, this time using the two source amplitudes and six of the location parameters from the experiment. Results of 120 000 simulation trials (10 000 in each of the 12 simulated conditions) are given in Table 5, and the rank ordering of the conditions is compared with that of human data in Table 6. With one exception, the model exhibits the correct pattern of effects under both levels of source amplitude, including the shift in the optimal location of the source toward the head as amplitude is increased. As in Experiment 3, the major factor in display design seems to be the computational demands related to different spatial filtering patterns rather than the perceived locations of the sources.

TABLE 5: Proportion of Correct Letter Identification Responses by Location and Amplitude for the Ideal Observer in Experiment 4

	Proportion Correct	
Location	Low Amplitude	High Amplitude
100, 20	.431	.446
30, 20	.447	.512
15, 20	.474	.527
0, 20	.470	.543
0, 10	.518	.634
0, 0	.781	.907

SUMMARY AND CONCLUSIONS

The natural listening environment may be superior to the traditional, inside-the-head format of many telecommunication systems, but it does not appear to be an optimal model for simulated 3-D sound displays. Reviewing results from both free-field and virtual listening studies, Yost (1997) concluded that spatial cues contribute relatively little to speech perception when there are multiple speech sources, even though it is easier to segment the sound stream when the sources are spatially distributed.

Results of the present experiments suggest that aspects of the sound patterns that are sometimes correlated with spatial location are more important factors for speech recognition than is spatialization per se. In Experiment 1 we found no appreciable increase in verbal warning identification rates when four sounds were placed in four different locations, as compared with placing them in pairs to the left and right of the listener. This finding suggests that position on the interaural axis is the key factor to consider, rather than realism or maximum spacing. In Experiment 3, placing a single sound in its most natural location – directly in front of the listener – was clearly suboptimal compared with every other position tested except the one directly behind the head. Apparently communication

TABLE 6: Source Location Coordinates Ranked by Identification Performance of the Ideal Observer and Humans in the Two Amplitude Conditions in Experiment 4

Low Amplitude		High Amplitude	
Ideal Observer	Humans	Ideal Observer	Humans
100, 20	100, 20	100, 20	100, 20
30, 20	0, 20	30, 20	30, 20
0, 20	30, 20	15, 20	15, 20
15, 20	15, 20	0, 20	0, 20
0, 10	0, 10	0, 10	0, 10
0, 0	0, 0	0, 0	0, 0

Note: The center of the head is 0, 0. Discrepancies between the ideal observer and humans occur only for two low-amplitude conditions (30, 20 vs. 0, 20), which were not significantly different in the experiment or in the simulations ($HSD = 0.0274$, $p > .05$).

in the real world would be improved if one could see the speaker in front of one but listen to him or her from the side.

When distance and amplitude are controlled, varying the position of a source on the interaural axis has relatively substantial effects on the timing and amplitude levels of the sounds arriving at the ears, which appears to change the computational demands of the identification problem. Masking effects of a noise source should be expected to diminish when the source is moved away from the signal, in part because source spacing increases the maximum signal-to-noise ratio in one of the two ears (Doll & Hanna, 1995). Developing this line of reasoning further, we found that the ideal observer predicts the human data well, including the effect of amplitude on the optimal location of a source on the interaural axis. Generating predictions from this model is somewhat difficult because the only way one can estimate its performance is to run large and computationally intensive simulations. In principle, however, the approach could be used to estimate the optimal location of an arbitrary class of sound stimuli, even without an accurate model of the perceptual systems or the higher-level cognition involved in speech recognition.

ACKNOWLEDGMENTS

Parts of this work were supported by NASA Ames Human Computing Center Grant NAS7-1407 (MacDonald and Balakrishnan) and by NASA Dryden Flight Research Center Grant NCC2-374 (Orosz and Karplus).

REFERENCES

Begault, D. R. (1994). *3-D sound for virtual reality and multimedia.* Boston: AP Professional.

Begault, D. R., & Wenzel, E. M. (1993). Headphone localization of speech. *Human Factors, 35,* 361–376.

Bronkhorst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *Journal of the Acoustical Society of America, 83,* 1508–1516.

Bronkhorst, A. W., & Plomp, R. (1990). A clinical test for the assessment of binaural speech perception in noise. *Audiology, 29,* 275–285.

Bronkhorst, A. W., & Plomp, R. (1992). Effect of multiple speech like maskers on binaural speech recognition in noise. *Journal of the Acoustical Society of America, 92,* 3132–3139.

Burdea, G., Richard, P., & Coiffet, P. (1996). Multimodal virtual reality: Input-output devices, system integration, and human factors. *International Journal of Human-Computer Interaction, 8,* 5–24.

Chapin, W. (2001). *CRE solutions* [On-line]. Available: http://ausim3d.com/products/crestuff.html.

Decroix, G., & Dehaussy, J. (1964). Binaural hearing and intelligibility. *Journal of Auditory Research, 4,* 115–134.

Dirks, D. D., & Wilson, R. H. (1969). The effect of spatially separated sound sources on speech intelligibility. *Journal of Speech and Hearing Research, 12,* 5–38.

Doll, T. J., & Hanna, T. E. (1995). Spatial and spectral release from masking in three-dimensional auditory displays. *Human Factors, 37,* 341–355.

Doll, T. J., Hanna, T. E., & Russotti, J. S. (1992). Masking in three-dimensional auditory displays. *Human Factors, 34,* 255–265.

Duquesnoy, A. J. (1983). Effect of a single interfering noise or speech source on the binaural sentence intelligibility of aged persons. *Journal of the Acoustical Society of America, 74,* 739–743.

Ericson, M. A., & McKinley, R. L. (1997). The intelligibility of multiple talkers separated spatially in noise. In R. H. Gilkey & T. R. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments* (pp. 701–724). Mahwah, NJ: Erlbaum.

Hay, V. H. (1996). Sound localization: The interaction of aging, hearing loss, and hearing protection. *Scandinavian Audiology, 25,* 3–12.

Janko, J. A., Anderson, T. R., & Gilkey, R. H. (1997). Using neural networks to evaluate the viability of monaural and interaural cues for sound localization. In R. H. Gilkey & T. R. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments* (pp. 557–570). Mahwah, NJ: Erlbaum.

King, R. B., & Oldfield, S. R. (1997). The impact of signal bandwidth on auditory localization: Implications for the design of three-dimensional auditory displays. *Human Factors, 39,* 287–295.

MacDonald, J. A. (1999). *The optimal decision maker — A tool for the analysis of speech recognition data.* Unpublished master's thesis, Purdue University, West Lafayette, IN.

Noro, K., Kawai, T., & Takao, H. (1996). The development of a dummy head for 3-D audiovisual recording for transmitting telepresence. *Ergonomics, 39,* 1381–1389.

Plomp, R., & Mimpen, A. M. (1981). Effect of the orientation of the speaker's head and the azimuth of a noise source on the speech reception threshold for sentences. *Acustica, 48,* 325–328.

Ricard, G. L., & Meirs, S. L. (1994). Intelligibility and localization of speech from virtual directions. *Human Factors, 36,* 120–128.

Saberi, K., Dostal, L., Sadralodabai, T., Bull, V., & Perrott, D. (1991). Free-field release from masking. *Journal of the Acoustical Society of America, 90,* 1355–1370.

Tonning, F. M. (1971). Directional audiometry: II. The influence of azimuth on the perception of speech. *Acta Otolaryngologica, 72,* 352–357.

Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America, 94,* 111–123.

Whalen, A. D. (1971). *Detection of signals in noise.* New York: Academic.

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

Yost, W. A. (1997). The cocktail party problem: Forty years later. In R. H. Gilkey & T. R. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments* (pp. 329–347). Mahwah, NJ: Erlbaum.

Yost, W. A., Dye, R. H., & Sheft, S. (1996). A simulated "cocktail party" with up to three sound sources. *Perception and Psychophysics, 58,* 1026–1036.

Justin A. MacDonald is a doctoral candidate at Purdue University, where he obtained his M.S. in quantitative-mathematical psychology in 1999.

J. D. Balakrishnan is an associate professor of quantitative-mathematical psychology at Purdue University. He obtained his Ph.D. in cognitive-mathematical psychology from the University of California, Santa Barbara, in 1991.

Michael D. Orosz is an assistant researcher at the University of California, Los Angeles, where he obtained his Ph.D. in computer science in 1999.

Walter J. Karplus was professor emeritus in computer science at the University of California, Los Angeles. He passed away on November 11, 2001. He obtained his Ph.D. from the University of California, Los Angeles, in 1955.

Date received: September 7, 2000
Date accepted: August 17, 2001