HUMAN FACTORS, 1993, 35(2), 361–376

Headphone Localization of Speech

DURAND R. BEGAULT[1] *and* ELIZABETH M. WENZEL, *NASA Ames Research Center, Moffett Field, California*

Three-dimensional acoustic display systems have recently been developed that synthesize virtual sound sources over headphones based on filtering by head-related transfer functions (HRTFs), the direction-dependent spectral changes caused primarily by the pinnae. In this study 11 inexperienced subjects judged the apparent spatial location of headphone-presented speech stimuli filtered with non-individualized HRTFs. About half of the subjects "pulled" their judgments toward either the median or the lateral-vertical planes, and estimates were almost always elevated. Individual differences were pronounced for the distance judgments; 15% to 46% of stimuli were heard inside the head, with the shortest estimates near the median plane. The results suggest that most listeners can obtain useful azimuth information from speech stimuli filtered by nonindividualized HRTFs. Measurements of localization error and reversal rates are comparable with a previous study that used broadband noise stimuli.

INTRODUCTION

Recently a considerable amount of attention has been focused on the development of a three-dimensional (3D) interactive display called the *virtual interface* (e.g., Fisher, Wenzel, Coler, and McGreevy, 1988). As with most research in information displays, that on virtual displays has generally emphasized visual information. Many investigators, however, have pointed out the importance of the auditory system as an alternative or supplementary information channel (e.g., Patterson, 1982). Primary advantages of the binaural auditory system include the ability to monitor and identify sources of information from all possible locations, improved intelligibility of sources in noise, enhanced segregation of multiple sources of speech (Cherry, 1953), and, in conjunction with other modalities, the reinforcement of information combined with a greater sense of presence or realism (Warren, Welch, and McCarthy, 1981). These features could be critical for applications such as air traffic control displays, applications involving cockpit communications (Begault and Wenzel, 1990, 1992; Calhoun, Janson, and Valencia, 1988), and telepresence applications such as teleconferencing, shared electronic workspaces (Ludwig, Pincever, and Cohen, 1990), and telerobotic control in hazardous situations (Wenzel, Stone, Fisher, and Foster, 1990).

Current approaches to developing 3D auditory displays have been based on the use of digital filters that capture the magnitude and phase characteristics of the head-related transfer function (HRTF); the listener-specific, direction-dependent acoustic effects

[1] Requests for reprints should be sent to Durand R. Begault, MS 262-2, NASA-Ames Research Center, Moffett Field, CA 94035.

imposed on an incoming signal by the pinnae (e.g., Begault, 1987; Kendall and Martens, 1984; McKinley and Ericson, 1988; Wenzel, Wightman, and Foster, 1988). The HRTF, measured near the tympanic membrane of the listener, demonstrates marked changes as a function of different source positions that accompany overall interaural level and time differences. Measured HRTFs also tend to vary considerably between subjects, probably because of differences in individual pinnae formation. Its complex spectrum provides a principal cue for localization, particularly for sources on the median plane (Blauert, 1983; Searle, Braida, Cuddy, and Davis, 1976). The use of HRTF filtering is also considered to be a primary determinant for externalizing headphone-delivered sound (Plenge, 1974; Wightman, Kistler, and Perkins, 1987).

To date relatively few experiments aimed at the perceptual validation of this synthesis technique have been done. Wightman and Kistler (1989b) conducted one groundwork study examining experienced listeners' performance under both free-field and headphone conditions with the subject's own HRTFs used to synthesize the stimuli. In general, Wightman and Kistler reported that localization accuracy for the free-field and headphone stimuli was comparable. With 3D auditory displays, however, it may not always be possible to tailor a set of HRTFs to a particular user; therefore, subjective localization performance with nonindividualized HRTFs becomes a critical issue for applied research (Begault, 1991).

An additional consideration is the fact that many 3D auditory display systems will require speech input, which could degrade localization performance compared with the broadband noise stimuli used in most previous studies. Finally, the performance of inexperienced subjects is important as a baseline measure of how readily the general population could use a 3D auditory display system.

The preliminary results of Butler and Belendiuk (1977) and of Wenzel, Wightman, Kistler, and Foster (1988) suggest the feasibility of using nonindividualized transfer functions to synthesize 3D auditory display cues. One approach suggested by Wenzel, Wightman, Kistler, and Foster (1988) was the use of HRTFs that are derived from a subject whose localization ability is relatively accurate and whose free-field and individualized HRTF headphone performance responses are closely matched. They proposed that the cues present in the HRTFs of a good localizer may work for another person in spite of the range of individual differences in HRTFs. This approach was suggested by the fact that a good localizer listening through the pinnae of a "bad localizer" exhibited degraded localization performance. The alternative approach of using HRTFs based on simple averages from several subjects has been discouraged because of the possibility of eliminating distinctive spectral features (Blauert, 1983). Using broadband noise stimuli, Wenzel, Arruda, Kistler, and Wightman (in press) completed an extensive study of virtual sources synthesized from the same nonindividualized, "good localizer" HRTFs used here. The results showed that localization of both free-field and virtual sources was accurate for 12 of the 16 subjects tested. Wenzel et al. (in press) concluded that most listeners can obtain useful directional information for azimuth using nonindividualized HRTFs.

The present study examined the headphone localization error of untrained subjects listening to speech stimuli processed with nonindividualized HRTFs. Performance was evaluated for a limited set of virtual auditory targets (12 different azimuths, all at ear level). The HRTFs used were derived from a representative subject (SDO) who had shown good localization performance in the experiment by Wightman and Kistler (1989b) under both headphone and free-field conditions. The perceptual deviation from the intended target was measured in terms of localization error (absolute errors in estimating azimuth and elevation), reversals (azimuth errors between the front and rear hemispheres), and distance errors (in particular, hearing the sound intracranially). These errors are illustrated in Figure 1.

The headphone-localization studies of Wightman and Kistler (1989b) and Wenzel et al. (in press) reported no data on absolute distance judgments, though Plenge (1974) and Laws (1973) reported such data in comparable experiments. The questions addressed by this study were (1) the comparability of overall azimuth error for speech to that of broadband noise, as measured in the study by Wenzel et al. (in press) that used the same nonindividualized HRTFs as used here; (2) the precision and consistency of elevation judgments for targets at the same elevation but at different azimuths; (3) whether or not nonindividualized HRTFs allow externalization; and (4) the consistency of absolute distance judgments of speech at various target azimuths.

METHOD

Subjects

Eleven adults served as paid volunteers in the study (ages 19–42; 8 males, 3 females). Although we did not conduct audiometric evaluations, we screened subjects orally with questions directed toward the following issues: noticeable overall hearing loss, noticeable differential hearing loss, recent exposure to loud noises (e.g., amplified music, motorcycle), work noise environment, and medical history. The use of oral reporting is not unusual in localization studies; other recent localization studies that have used oral screening methods without audiometric screening include Noble (1987), Asano, Suzuki, and Sone (1990), and Perrott, Sadralodabai, Saberi, and Strybel (1991).

Our rationale for not using audiometric screening procedures is as follows. Typically audiometric screening considers "normal hearing" to be within 15–20 dB HL for a limited set of frequencies, with a resolution accuracy of no better than 5 dB HL. Localization of complex signals such as speech is based on information integrated across the frequency spectrum, making it unlikely that

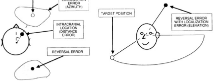
TARGET POSITION
REGION OF LOCALIZATION ERROR (AZIMUTH)
INTRACRANIAL LOCATION (DISTANCE ERROR)
REVERSAL ERROR
TARGET POSITION
REVERSAL ERROR WITH LOCALIZATION ERROR (ELEVATION)

Figure 1. *Types of localization error reported in the literature and discussed in the current experiment. Target position is the position of the sound source where the HRTF was measured. Left side: overhead view. Right side: perspective view.*

sensitivity at single tone frequencies is an accurate predictor of localization performance.

The lack of a clear relationship between audiometric screening and localization has also been observed empirically. For example, several studies have shown that audiometrically normal individuals may have deficits in the discrimination of interaural differences and binaural detection (e.g., Koehnke, Colburn, and Durlach, 1986; see also the review by Colburn and Trahiotis, 1991). Conversely, Gabriel, Koehnke, and Colburn (1991) underscored this result in a study of individuals who had audiometric loss: no apparent relation between audiometric measurements and binaural performance was found. Blauert (1983) cited several studies to support the point that "symmetrical hearing loss of peripheral origin of as much as 30–40 dB has almost no noticeable effect on localization blur. In particular, age-related hearing loss hardly detracts from spatial hearing" (p. 49). Blauert cited other literature to the effect that although asymmetrical hearing loss will alter localization, localization blur decreases with experience and becomes more normal, and "with time the direction of the auditory event coincides better and better with that of the sound source" (Blauert, 1983, p. 50). Thus, although temporary bilateral threshold shifts might conceivably result in poor binaural performance, it is unlikely that long-term hearing losses detected by a single audiometric test at the start of an experiment would have a predictable impact on localization ability. However, the possibility of temporary threshold shifts attributable to exposure to loud noises was the rationale for the verbal screening procedure.

Stimuli

Stimuli were generated from a set of 45 one- or two-syllable words, each representing a particular phoneme from an International Phonetic Alphabet list (Table 1), with durations ranging from 0.7 to 1.3 s. The speech was recorded digitally in a soundproof booth by a male speaker using an AKG microphone (451-EB), a Symetrix preamplifier (SX-202), and a Panasonic DAT recorder (SV-3500). The speech segments were then transferred to an Apple computer (Macintosh IIcx), edited with Digidesign Sound Tools hardware and software, and then digitally transferred to a Masscomp computer (MC-5500 DP) for non-real-time signal processing and real-time playback to subjects. The average spectrum of the speech segments is shown in Figure 2.

TABLE 1

Words Used for Stimuli, with International Phonetic Alphabet Symbol

Word	*IPA Symbol*	*Word*	*IPA Symbol*
pat	æ	toe	o
pay	e	caught	ɔ
care	Ɛr	noise	ɔI
father		took	U
bib	b	boot	u
church	tʃ	out	aU
deed	d	pop	p
pet	Ɛ	roar	r
bee	i	sauce	s
fife	f	ship	ʃ
gag	g	tight	t
hat	h	thin	θ
which	hw	this	∂
pit	I	cut	V
pie	aI	urge	3r
pier	Ir	valve	v
judge	d_3	with	w
kick	k	yes	j
lid	l	zebra	z
mum	m	vision	ʒ
no	n	about	∂
thing	η	butter	∂r
pot			

Stimuli for a given subject and a given set of trials were precomputed and normalized to a fixed root mean square (RMS) value on the Masscomp. Each of the 45 speech segments was digitally processed so that it would simulate a particular free-field location: target positions at 0 (front), 180, and left and right 30, 60, 90, 120, and 150 deg

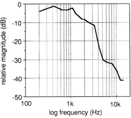

Figure 2. *Average spectrum of speech stimuli used in experiment prior to HRTF processing. This graph was obtained using a 256-point fast Fourier transform.*

azimuth, all at 0 deg elevation (ear level). As noted, the processing was based on the HRTFs of subject SDO, who was measured by Wightman and Kistler (1989a) and whose HRTFs were used by Wenzel et al. (in press). In order to compare results with those of previous studies using static sources, the system used here did not compensate for head position via a tracking device (such as described in Wenzel, Wightman, and Foster, 1988).

Each speech segment was processed independently for the left- and right-ear stimuli by cascading through two finite impulse response filter sections. The first filter section consisted of SDO's left (or right) ear HRTF for a given source position and the inverse of SDO's headphone-to-ear-canal transfer function for the same ear. The inverse transfer function was required to remove the spectral characteristics imposed by the headphones (Sennheiser HD 430) when worn by SDO (see Wightman and Kistler, 1989a). This same type of headphone was used for subject playback. The second filter section was a zero-phase bandpass filter (200 Hz–14 kHz) used to remove processing artifacts at low and high frequencies. In generating the stimuli for each experimental trial, the particular combination of speech segment and target location was chosen randomly. The signals were played back via Masscomp-controlled, 16-bit D/A converters at a rate of 50 kHz. The RMS level of the filtered, normalized stimuli was about 70 dB SPL.

Procedure

Subjects were blindfolded, and testing was conducted in a double-walled sound isolation chamber. Subjects sat relatively motionless at a table with their heads unrestrained and were instructed to give oral responses to the microphone located directly in front of them.

During each trial subjects heard five repetitions of a given speech segment and then called out estimates of the apparent azimuth, elevation, and distance of the virtual sound source using a modified spherical coordinate system. That is, azimuth was defined as 0 to 180 deg left or right (where 0 deg is directly in front) and elevation was defined as 0 to 90 deg up or down (where 0 degrees is at ear level). For distance, subjects were instructed to call out "0 inches" if the sound was directly at the center of their head, between 0 and 4 inches for positions inside the head, exactly 4 inches for a verged-cranial impression (at the edge of the head), and greater than 4 inches for externalized sounds. For example, a sound that seemed outside the head, slightly elevated, and to the right of the median plane might be reported as "right 30 degrees, up 15 degrees, and 30 inches." Subjects' estimates were recorded by an experimenter located outside the testing booth during an unfixed response interval. No feedback was given, but subjects were allowed to request that a particular trial be repeated.

Prior to the experimental runs, a 15-min training session was conducted that included an oral explanation of the response coordinates and a practice block of trials.

Subjects appeared to learn the task easily, and they quickly produced stable judgments. However, the practice block was not used in subsequent data analyses. To avoid errors attributable to headphone misplacement, subjects were asked to center a diotic 440-Hz sine wave (70 dB SPL) by adjusting the headphones at the beginning of each block.

Over the course of two to three days, each subject listened to 15 experimental blocks of 30 stimuli containing a different randomized ordering of the 12 azimuth positions; targets at 0 and 180 deg were heard a total of 150 times, and all other locations were repeated 15 times. Within each block, 10 of the stimuli were at 0 deg, 10 were at 180 deg, and the remaining 10 occurred at each of the following target azimuths: left 30, 60, 90, 120, and 150, and right 30, 60, 90, 120, and 150 deg. Subjects were given breaks at least every two or three blocks, and the total duration for a single day never exceeded 2½ h.

RESULTS

Reversals

Front-back "reversals" have been observed in nearly all studies of sound localization, for both real and virtual sources. These are responses that indicate that a source in the front hemisphere, usually near the median plane, was perceived to be in the rear hemisphere. Occasionally the reverse situation also occurs. In the literature reversals have generally been resolved when computing descriptive statistics (i.e., the responses are coded as if the subject had indicated the correct hemisphere), and then the number of reversals is reported as a separate statistic. The argument for treating reversals as a separate issue is based on the premise that localization blur would be unfairly inflated if reversals are left "uncorrected" in reporting the results of an experiment (Oldfield and Parker, 1984; Stevens and Newman 1936; Wightman and Kistler, 1989b).

The algorithm for resolving reversals used here tests whether the angle between the target and judged location is made smaller by reflecting the judgment about the vertical plane passing through the subject's ears. If the test proves true, the judgment is coded in reflected form and the percentage of reversals is increased. Note that there can be no reversed judgment for the 90-deg target because the target lies directly on this plane.

The percentage of reversed judgments, averaged across all subjects, is shown in Figure 3 for the total number of judgments obtained at each target. The mean value of the percentages of reversed judgments at each target is 29%. The percentage of reversals from back to front is significantly less than the percentage from front to back (11% vs. 47%, $\chi^2 = 597$, $p < 0.0001$), a phenomenon that has been observed informally for many years with recordings made in the "ear canals" of binaural artificial heads (Hudde and Schröter, 1981).

Wightman and Kistler (1989b) reported reversal rates for eight subjects listening to noise stimuli at a number of azimuth positions, with the data at their "middle elevations" closest to the locations tested here (0 to 18 deg up). In their study subjects in the headphone condition listened to stimuli processed by their own HRTFs. When results for the middle elevations are averaged across subjects, the reversal rate is about 6%, much lower than the mean value of 29% found here for reversals. The front-to-back reversal rate for the 0-deg target obtained in this study (58%) is also much higher than the percentage reported by Laws (1973), who found a 35% reversal rate using averaged HRTFs. However, the overall percentages are comparable to those of Wenzel et al. (in press), who found a reversal rate of 31% (25% front to back, 6% back to front) for untrained listeners

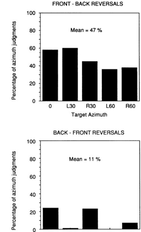

Figure 3. *Total percentage of reversed judgments by target (1650 judgments for 0- and 180-deg positions; 165 judgments for other positions; 11 subjects). Note that the following pairs of positions have roughly the same interaural differences: (0, 180); (30, 150); (60, 120). The overall mean of reversals = 29%.*

hearing broadband noise processed with the same HRTFs as in this experiment but over a larger range of target elevations and azimuths. The ratio of front-to-back versus back-to-front reversals found here (approximately 4:1) is comparable to that reported by Wenzel et al. (in press) but higher than the ratio of about 2:1 reported by Wightman and Kistler (1989b).

Localization Error

In order to analyze localization error, target and judged positions described as points in three-dimensional space on the surface of a sphere are compared. This sphere is of unit distance because the actual target distance remained constant in this experiment. As a result, standard mean and variance statistics are potentially misleading. For example, an azimuth error of 15 deg on the horizontal plane is much larger in terms of absolute distance than a 15-deg error at an elevation of 54 deg. Thus spherical statistical techniques are used to characterize the data (Fisher, Lewis, and Embleton, 1987); these techniques were first applied in localization studies by Wightman and Kistler (1989b). Such issues are somewhat less relevant to the present study, in which all targets occurred at an elevation of 0 deg, although subjects' responses included estimates of both azimuth and elevation. However, the ability to compare reversal-corrected judgments to the data of Wightman and Kistler (1989b) and Wenzel et al. (in press) was desirable.

The following descriptive spherical statistics were used here: average angle of error, judgment centroid, and inverse kappa (K^{-1}). The average angle of error is the mean of the unsigned angles between each judgment vector and the vector from the origin to the target position. The judgment centroid can be thought of as the "average direction" of a set of judgments from the origin (i.e., the center of the subject's head). It is defined as a unit-length vector with the same direction as the resultant, the vector sum of all the unit-length judgment vectors. The length of the resultant vector is determined by the dispersion of the judgments; judgments concentrated around the centroid are reflected in a long resultant, whereas scattered judgments produce a short resultant. K, the commonly used index of dispersion, is estimated from the length of the resultant. Generally the parameter K^{-1} is reported because the inverse value varies with dispersion in the same manner as a variance estimate; larger values of K^{-1} reflect larger deviations of the judgment vectors from the target. For example, when the number of trials is 150 (i.e., for the 0- and 180-deg targets), a K^{-1} of 0.01 corresponds to a 95% confidence angle of 1.2 deg, whereas a K^{-1} of 0.18 corresponds to a confidence angle of 5.4 deg, with respect to the centroid estimated for a particular target location. When the number of trials is 15 (i.e., all other targets), the confidence angles becomes 3.7 and 17.1 deg for K^{-1} values of 0.01 and 0.18, respectively. See Wightman and Kistler (1989b) for further details on spherical statistics applied to localization data.

Table 1 shows the reversal-corrected judgment centroids, average error angles, and K^{-1} values for each target position. The average error angle and K^{-1} values shown are based on the mean of each individual subject's mean values. In examining these data, it should be remembered that 150 judgments contributed to the 0- and 180-deg means and that 15 judgments contributed to the means of the other 10 positions. The mean value of K^{-1} from Table 2 is 0.102, and the mean average error angle is 28 deg. These values are somewhat larger than the corresponding values computed by Wightman and Kistler (1989b) for subject SDO listening with her own HRTFs; the mean value for K^{-1} was 0.06, and the mean error angle was 20.5 deg, for a range of different azimuths at middle elevations (0 and 18 deg up).

Azimuth Estimation Error

Figure 4 shows the mean value for the azimuth centroids of all subjects compared with the data for untrained subjects from the study by Wenzel et al. (in press). Again, the latter study presented noise stimuli from a range of different elevations but used the same HRTFs as in this study. The centroids for both studies, based on resolved judg-

TABLE 2

Means of Reversal-Corrected Judgment Centroids, Average Error Angles, and Inverse Kappa values Collapsed across All Subjects, Shown for Each Target Position

Target Azimuth	*Azimuth Centroid*	*Elevation Centroid*	*Inverse Kappa*	*Average Error Angle*	*Percentage Reversals*	*Mean Distance*
L0	L5	U19	0.122	24.6	58	6.6
L30	L52	U38	0.126	47.0	60	8.4
L60	L79	U16	0.067	29.6	36	7.5
L90	L97	U13	0.053	17.6	n/a	8.4
L120	L111	U8	0.092	22.2	0	8.7
L150	L142	D2	0.131	25.3	1	8.5
180	L172	U9	0.127	21.7	24	6.7
R150	R143	U18	0.093	28.1	23	8.1
R120	R115	U16	0.102	27.5	7	8.5
R90	R102	U18	0.088	25.7	n/a	8.0
R60	R65	U25	0.094	31.5	38	7.5
R30	R39	U27	0.134	35.0	45	7.5
Mean			0.102	27.9	29.2	7.8

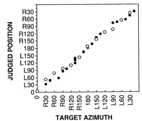

Figure 4. *Centroids of azimuth judgments, based on reversal-corrected judgments by 11 subjects. Open circles: this study (speech). Closed circles: study by Wenzel et al. (1991 [noise stimuli]; data collapsed across different elevations).*

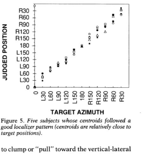

Figure 5. *Five subjects whose centroids followed a good localizer pattern (centroids are relatively close to target positions).*

ments, are close to their intended target positions when viewed in this way (for this study, $r^2 = 0.992$ [$y = 11.885 + 0.935x$]). This suggests that the broader spectral characteristics of white noise, compared with the spectral content of speech (see Figure 2), is not critical for equivalent azimuth performance in a 3D headphone display system. This is not surprising in light of previous data that indicate that lower frequencies are dominant in lateralization tasks based on interaural differences of time, a primary cue for azimuth (e.g., Bilsen and Raatgever, 1973; McFadden and Pasanen, 1976).

In looking at individuals' behavior, three patterns of performance were observed. The first pattern is seen in Figure 5 for the five subjects who showed "good" localization. That is, resolved judgment centroids were strongly correlated with the target positions, corresponding closely to *ideal performance:* a linear fit with a slope of +1.0. In the second pattern another 5 subjects displayed a response bias in which their judgments tended to clump or "pull" toward the vertical-lateral plane passing through the left and right ears (Figure 6). This pattern was also observed in the extensive free-field study conducted by Oldfield and Parker (1984), which they characterized as "defaults to 90." Finally, as shown in Figure 7, one subject's judgments were "pulled" toward the vertical-median plane; that is, this listener had a response bias toward the front and rear positions.

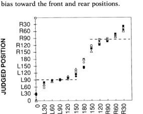

Figure 6. *Five subjects whose centroids followed a pattern of pulling toward the vertical-lateral plane.*

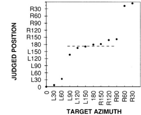

Figure 7. *The single subject whose centroids followed a pattern of pulling toward the vertical median plane.*

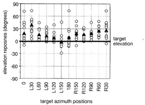

Figure 8. *Judgment centroids for elevation at each target azimuth for each subject (circles) and collapsed across all subjects (triangles). Total of 1650 judgments for 0- and 180-deg target azimuths; 165 judgments for other targets.*

These individual differences in localization performance tend to support Butler and Belendiuk's (1977) observations that there are "good" and "bad" localizers. However, it is also possible that the patterns shown in Figures 6 and 7 reflect a response bias attributable to the relatively limited number of positions tested and the high proportion of 0- and 180-deg targets. It is interesting that of the subjects who showed biases, most were biased toward the side positions. Perhaps in this case the subjects' strategy was to place the most infrequent positions as far away as possible from the most frequent stimuli in a kind of response-contrast effect.

Elevation Estimation Error

For most subjects the HRTF-processed speech was perceived to be at an elevation higher than the target of 0 deg (eye level); the mean value across azimuths was up 17 deg (*SD*, 10 deg). Figure 8 contrasts the elevation centroids for each subject with means collapsed across all subjects for each azimuth examined. Note that the means are all above the target elevation of 0 deg, except for the left 150 target, and that the range of individual centroids is quite large (down 2 deg to up 40 deg).

The large dispersion of judgments is perhaps surprising in light of the fact that the target elevation of the stimuli never varied. This poor elevation performance is possibly attributable to the use of speech stimuli; some researchers have contended that the spectral cue in the HRTF for elevation is in the region above 7 kHz, where speech has relatively less spectral energy (see Table 1; Blauert, 1969; Rodgers, 1981). However, the lack of a spectral cue to elevation does not explain the overall elevation bias in listening to speech, given the predominance of upward judgments. An elevation bias was not observed in the comparable study by Wenzel et al. (in press).

Distance and Externalization

Externalization of HRTF-processed sound is of interest because this is a perceptual characteristic frequently absent from dichotic signals produced with interaural level or time differences. Subjective impression of the distance of speech is particularly interesting because it is often supposed that distance perception is largely mediated by a listener's familiarity with the stimulus (Coleman, 1962; McGregor, Horn, and Todd, 1985). Here the level of 70 dB SPL corresponds to the average RMS speech level of a person at 1 m from the listener (Kryter, 1972). The actual distance of the sound source to the listener was 1.38 m (54.33 inches) when the HRTFs were originally measured.

The judgments for distance can be analyzed either categorically in terms of whether the sound source is externalized or heard inside the head, or based on the continuous response values, which presumably reflect a monotonic, ordinal relationship between the subjective impression and the distance of the sound source. Here distance was consistently underestimated, as has been reported in distance studies with actual sound sources (Holt and Thurlow, 1969) and with headphone-delivered sources processed by HRTFs (Begault, 1987). Figure 9 shows the means and standard deviations of distance judgments for each target position collapsed across the individual means of the 11 subjects. Although the mean value for distance judgments for all subjects was externalized (i.e., greater than 4 in) for all target positions (Figure 9), the standard deviations indicate that a large proportion of the sounds were heard inside the head.

Figure 11 shows the combined percentage of total judgments for targets reported to be less than 4 in in distance (intracranial) and those exactly at 4 in distance (verged-cranial). Note that these combined percentages are always less than the percentage of externalized judgments. In particular, Figure 10 shows that about one third of the judgments for 0 and 180 deg were heard intracranially, with a total of 50% not externalized for these positions, matching the 50% level reported by Laws (1973; averaged HRTFs). Other positions were externalized more often than were the 0- and 180-deg positions; the range of responses heard verged or intracranially ranged from 15% to 46%.

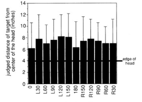

Figure 9. *Means and standard deviations of distance judgments collapsed across all subjects. Total of 1650 judgments for 0- and 180-degree targets; 165 judgments for other targets. Distances beyond 4 inches are externalized.*

These results contrast strongly with those obtained by Plenge (1974), who used an

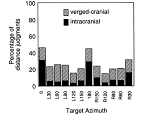

Figure 10. *Percentage of all distance judgments collapsed across subjects for each target position judged as inside the head (intracranial: between 0 and 4 inches) or on the edge of the head (verged-cranial: subjects reporting 4 inches distance). Note that the percentage of intracranial judgments exceeds verged cranial judgments only with the 0- and 180-deg target positions.*

artificial head in a reverberant environment and found that the majority of judgments were externalized, with almost none placed at the verged-cranial position. One possible reason for the relative lack of externalization in the present study is the absence of environmental cues such as reverberation in the original HRTF measurements. A solution is to add synthetic room reflections to the stimuli; data obtained by Begault (1992) using the same speech stimuli as in this study, as well as data from a related study by Sakamoto, Gotoh, and Kimura (1976), suggest that this is indeed the case.

Figure 11 shows polar plots of individual differences, combining reversal-corrected azimuth centroids and distance judgments. The perspective is from overhead, with the listener facing right. The top-left plot of Figure 11 shows what an ideal response set might look like: equidistant estimates corresponding exactly to the target locations. The vectors from the center represent the target positions, and the black dots are the location of the response centroids. The distance of the black dot from the center of the head is in inches; the inner circle represents the edge of the head at a radius of 4 inches. Several observations are recapitulated in these plots: the tendency of some subjects to collapse their azimuth judgments toward the median or lateral-vertical planes, the range of individual differences in the degree of perceived externalization, and the tendency for locations at the median plane to produce the shortest estimates of distance.

DISCUSSION

The motivation for this study was to evaluate the headphone localization performance of inexperienced subjects listening to HRTF-filtered speech stimuli and to compare results with a similar study by Wenzel et al. (in press) in which noise stimuli were used. The localization performance of inexperienced subjects in this study, particularly as evidenced by measures of dispersion and rates of azimuth reversals, was somewhat worse than that obtained with experienced subjects by Wightman and Kistler (1989b). However, the data reported here parallel the results obtained by Wenzel et al. (in press) for inexperienced subjects listening to noise stimuli.

Wightman and Kistler's (1989b) data suggest that reversal rates and azimuth error angles are generally lower when subjects listen to stimuli synthesized with their own HRTFs. The ability of listeners to adapt to the unique spectral cues provided by a particular set of HRTFs is probably a factor in the discrepancies between studies using experienced subjects and those using inexperienced subjects. For example, Asano et al. (1990) claimed that reversal errors diminish as subjects adapt to the unfamiliar and ambiguous cues for localization present in static, anechoic stimuli. Further, the existence of free-field reversals in both the Wightman and Kistler (1989b) and Wenzel et al. (in press) studies indicates that these reversals are not strictly the result of the simulation. The evidence of individual differences found in the present study may suggest that some listeners were able to adapt more easily to the spectral cues of the nonindividualized HRTFs than others; or, put another way, perhaps their ears were more similar to SDOs.

Perhaps the most striking feature of the data was the appearance of three distinct patterns of azimuth judgment behavior. About half of the subjects' resolved judgments of location closely matched the target positions, about half pulled toward the vertical-lateral plane, and a single subject clumped responses near the median plane. In spite of these individual differences, there were some common behavioral trends. For example, localization error was not equal for all azimuth targets: absolute accuracy tended to be better at the rear than at the front. The mean value

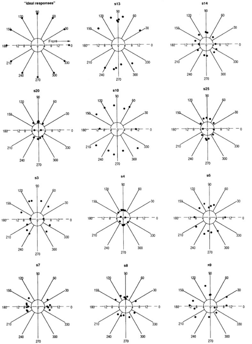

Figure 11. *Azimuth centroids and average distances for individual subjects. Dots represent the average distance reported by the subject for each target position. Upper left shows an "ideal response." Inner circle = 4-inch radius (representing the head). Overhead view of subject oriented facing right.*

of K^{-1} for left and right 30 and 60 deg is 0.105, whereas the mean value for left and right 120 and 150 deg is 0.085. This observation parallels Wightman and Kistler (1989b), who found that localization performance is affected by the azimuth of the source. Another trend was the tendency for subjects to elevate their judgments, particularly for positions at the front, even though the targets were all at 0 deg elevation. Finally, distance estimates were almost always smaller for targets at the median plane (0 and 180 deg) compared with the other targets, even though subjects often varied widely in their rates of perceived intracranial and verged-cranial targets (15%–46%, as a function of target position).

Although the reason for azimuth reversals is not completely understood, they are probably attributable in large part to the static nature of the stimulus and the ambiguity resulting from the so-called cone of confusion (Mills, 1972). Assuming a stationary, spherical model of the head, a given interaural time difference correlates ambiguously with several sound source locations to the front and to the rear. Several stimulus characteristics may help to minimize these errors, such as the addition of visual or dynamic cues correlated with head motion (Wallach, 1940). Other methods reported in the literature include altering the spectrum of averaged HRTFs to mimic loudspeaker transfer functions (Laws, 1973) and manipulating the subject's own HRTFs in the 0.5–7.0 kHz region (Weinrich, 1982).

From an applied standpoint, the data suggest that most listeners can obtain useful directional information from speech stimuli in an auditory display without requiring the use of individually tailored HRTFs, particularly for the dimension of azimuth. However, the data suggest that some azimuth targets may have a smaller range of error and a greater chance of reversal than do others. Similarly, although the stimuli were confined to a target elevation of 0 deg, the predominantly elevated judgments suggest that adequate synthesis of elevation cues is difficult. The high percentage of intracranial or verged-cranial localized stimuli was surprising, although the inclusion of synthetic reverberation with the same speech stimuli used here has been shown to mitigate this problem (Begault, 1992).

Overall, there is great potential for the use of HRTF-filtered speech in auditory displays. However, substantial questions remain regarding how to synthesize new HRTFs or modify existing ones so that they can be used within an applied context by inexperienced listeners. Designers and users will need to be keenly aware of both the possibilities and limitations in current implementations of these systems.

ACKNOWLEDGMENTS

This work was supported by NASA Ames Research Center, the National Research Council, and the Federal Aviation Authority. The authors gratefully acknowledge the invaluable assistance of Phil Stone and Rick Shrum during the technical phases of the study.

REFERENCES

Asano, F., Suzuki, Y., and Sone, T. (1990). Role of spectral cues in median plane localization. *Journal of the Acoustical Society of America, 88,* 159–168.

Begault, D. R. (1987). *Control of auditory distance.* Unpublished doctoral dissertation, University of California, San Diego.

Begault, D. R. (1991). Challenges to the successful implementation of 3-D sound. *Journal of the Audio Engineering Society, 39,* 864–870.

Begault, D. R. (1992). Perceptual effects of synthetic reverberation on three-dimensional audio systems. *Journal of the Audio Engineering Society, 40,* 895–904.

Begault, D. R., and Wenzel, E. M. (1990). *Technical aspects of a demonstration tape for three-dimensional auditory displays* (Tech. Memorandum 102286). Moffett Field, CA: NASA Ames Research Center.

Begault, D. R., and Wenzel, E. M. (1992). Techniques and applications for binaural sound manipulation in human-machine interfaces. *International Journal of Aviation Psychology, 2,* 1–22.

Bilsen, F. A., and Raatgever, J. (1973). Spectral dominance in binaural lateralization. *Acustica, 28,* 131–132.

Blauert, J. (1969). Sound localization in the median plane. *Acustica, 22,* 205–213.

Blauert, J. (1983). *Spatial hearing: The psychophysics of human sound localization.* Cambridge, MA: MIT Press.

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

Calhoun, G. L., Janson, W. P., and Valencia, G. (1988). Effectiveness of three-dimensional auditory directional cues. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 68–72). Santa Monica, CA: Human Factors and Ergonomics Society.

Cherry, E. C. (1953). Some experiments on the recognition of speech with one and two ears. *Journal of the Acoustical Society of America, 25*, 975–979.

Colburn, H. S., and Trahiotis, C. (1991). Effects of noise on binaural hearing. In A. Dancer, D. Henderson, R. Salvi, and R. Hamernik (Eds.), *Noise-induced hearing loss*. St. Louis, MO: Mosby.

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

Fisher, N. I., Lewis, T., and Embleton, B. J. (1987). *Statistical analysis of spherical data*. Cambridge, England: Cambridge University Press.

Fisher, S. S., Wenzel, E. M., Coler, C., and McGreevy, M. W. (1988). Virtual interface environment workstations. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 91–95). Santa Monica, CA: Human Factors and Ergonomics Society.

Gabriel, K. J., Koehnke, J., and Colburn, H. S. (1991). Frequency dependence of binaural performance in listeners with impaired binaural hearing. *Journal of the Acoustical Society of America, 91*, 336–347.

Holt, R. E., and Thurlow, W. R. (1969). Subject orientation and judgment of distance of a sound source. *Journal of the Acoustical Society of America, 46*, 1584–1585.

Hudde, H., and Schroter, J. (1981). Verbesserungen am Neumann-Kunstkopfsystem [Improvements in the Neumann Artificial Head System.] In *Rundfunktechnische Mitteilungen* (Report C-21-738-F, pp. 5–10). Berlin: Georg Neumann.

Kendall, G. S., and Martens, W. L. (1984). Simulating the cues of spatial hearing in natural environments. In *Proceedings of the 1984 International Computer Music Conference* (pp. 111–125). San Francisco: Computer Music Association.

Koehnke, J., Colburn, H. S., and Durlach, N. I. (1986). Performance in several binaural interaction experiments. *Journal of the Acoustical Society of America, 79*, 1558–1563.

Kryter, K. D. (1972). Speech communication. In H. P. Van Cott and R. G. Kinkade (Eds.), *Human engineering guide to equipment design* (pp. 161–226). Washington, DC: U.S. Government Printing Office.

Laws, P. (1973). Entfernungshören und das Problem der Im-Kopf-Lokalisiertheit von Hörereignissen [Auditory distance perception and the problem of "in-head localization" of sound images]. *Acustica, 29*, 243–259.

Ludwig, L., Pincever, N., and Cohen, M. (1990). Extending the notion of a window system to audio. *Computer, 23*(8), 66–72.

McFadden, D., and Pasanen, E. (1976). Lateralization of high frequencies based on interaural time differences. *Journal of the Acoustical Society of America, 59*, 634–639.

McGregor, P., Horn, A. G., and Todd, M. A. (1985). Are familiar sounds ranged more accurately? *Perceptual and Motor Skills, 61*, 1082.

McKinley, R. L., and Ericson, M. A. (1988). Digital synthesis of binaural auditory localization azimuth cues using headphones. *Journal of the Acoustical Society of America, 83*, S18.

Mills, W. (1972). Auditory localization. In J. V. Tobias (Ed.), *Foundations of modern auditory theory* (pp. 301–345). New York: Academic.

Noble, W. (1987). Auditory localization in the vertical plane: Accuracy and constraint on bodily movement. *Journal of the Acoustical Society of America, 82*, 1631–1636.

Oldfield, S. R., and Parker, S. P. A. (1984). Acuity of sound localization: A topography of sound space: I. Normal hearing conditions. *Perception, 13*, 581–600.

Patterson, R. R. (1982). *Guidelines for auditory warning systems on civil aircraft* (Tech. Paper 82017). London, England: Civil Aviation Authority.

Perrott, D. R., Sadralodabai, T., Saberi, K., and Strybel, T. Z. (1991). Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target. *Human Factors, 33*, 389–400.

Plenge, G. (1974). On the difference between localization and lateralization. *Journal of the Acoustical Society of America, 56*, 944–951.

Rodgers, C. A. (1981). Pinna transformations and sound reproduction. *Journal of the Audio Engineering Society, 29*, 226–234.

Sakamoto, N., Gotoh, T., and Kimura, Y. (1976). On "out-of-head localization" in headphone listening. *Journal of the Audio Engineering Society, 24*, 710–716.

Searle, C. L., Braida, L. D., Cuddy, D. R., and Davis, M. F. (1976). Model for auditory localization. *Journal of the Acoustical Society of America, 60*, 1164–1175.

Stevens, S. S., and Newman, E. B. (1936). The localization of actual sources of sound. *American Journal of Psychology, 48*, 297–306.

Wallach, H. (1940). The role of head movements and vestibular cues in sound localization. *Journal of Experimental Psychology, 27*, 339–368.

Warren, D. H., Welch, R. B., and McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception and Psychophysics, 30*, 557–564.

Weinrich, S. (1982). The problem of front-back localization in binaural hearing. *Scandinavian Audiology, 11* (Suppl. 15), 135–145.

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (in press). Localization using non-individualized head-related transfer functions. *Journal of the Acoustical Society of America.*

Wenzel, E. M., Stone, P. K., Fisher, S. S., and Foster, S. H. (1990). A system for three-dimensional acoustic "visualization" in a virtual environment workstation. In *Proceedings of the IEEE Visualization '90 Conference* (pp. 329–337). Los Alamitos, California: IEEE Computer Society Press.

Wenzel, E. M., Wightman, F. L., and Foster, S. H. (1988). A virtual display system for conveying three-dimensional acoustic information. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 86–90). Santa Monica, CA: Human Factors and Ergonomics Society.

Wenzel, E. M., Wightman, F. L., Kistler, D. J., and Foster, S. H. (1988). Acoustic origins of individual differences in sound localization behavior. *Journal of the Acoustical Society of America, 84,* S79.

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

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

Wightman, F. L., Kistler, D. J., and Perkins, M. E. (1987). A new approach to the study of human sound localization. In W. Yost and G. Gourevitch (Eds.), *Directional hearing* (pp. 26–48). New York: Springer-Verlag.