HEADPHONE LOCALIZATION OF SPEECH STIMULI

Durand R. Begault
Elizabeth M. Wenzel
Aerospace Human Factors Research Division
NASA-Ames Research Center
Moffett Field, California 94035

Recently, three-dimensional acoustic display systems have 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 outer ears. Here, 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 judgements toward either the median or the lateral-vertical planes, and estimates were almost always elevated. Individual differences were pronounced for the distance judgements; 15 to 46% of stimuli were heard inside the head with the shortest estimates near the median plane. The results infer that most listeners can obtain useful azimuth information from speech stimuli filtered by non-individualized 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 a truly three-dimensional, interactive display called the virtual interface (e.g., Fisher, et al., 1988). In general, virtual displays have emphasized visual information, although many investigators have proposed the binaural auditory system as an important alternative channel of information (e.g., Patterson, 1982). Potentially useful applications of a three-dimensional auditory display include tasks involving spatial situational awareness, as in air traffic control displays and traffic collision avoidance systems (Begault and Wenzel 1990).

Current approaches to 3-D auditory displays are based on the use of digital filters that capture the spectral and phase characteristics of the Head-Related Transfer Function (HRTF); the listener-specific, direction-dependent effects imposed on an incoming signal by the outer ears (e.g., Begault & Wenzel, 1990; Wenzel, et. al, 1988). Because it may not always be possible to individually-tailor a set of HRTFs to a particular user, subjective localization performance with non-individualized HRTFs becomes a critical issue for applied research. An additional consideration is the fact that many 3-D auditory display systems will require speech input, which could degrade localization performance compared to the broadband noise stimuli used in most previous studies. Finally, the performance of untrained subjects is an important baseline measure reflecting how readily the general population could use a 3-D auditory display system.

We examined the headphone localization error of untrained subjects, listening to speech stimuli processed with non-individualized HRTFs, for a limited set of virtual auditory targets. 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). We were also interested in: (1) how comparable is azimuth blur of speech to azimuth blur of broadband noise, as measured in a previous study (Wenzel et al. 1991) that used the same non-individualized HRTFs; (2) how precise are elevation judgements, and how consistent are they, given the same target elevation but different target azimuths; (3) do non-individualized HRTFs allow externalization; and (4) how consistent are absolute distance judgements of speech at various target azimuths.

METHOD

Eleven adults served as paid volunteers in the study (ages 19 - 47, 8 males, 3 females). All reported no known hearing loss or history of hearing problems, and none had previously participated in a headphone localization experiment.

Stimuli were generated by digitally filtering a set of 45 one or two-syllable words, each representing a particular phoneme, with HRTFs at target positions of 0, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300 and 330 degrees azimuth, all at 0 degrees elevation. For each trial, the particular combination of speech segment and target location was chosen randomly. All stimuli were presented via headphones. The HRTFs (measured under anechoic conditions) were derived from a representative subject in the study by Wightman & Kistler (1989a) and also used in the study by Wenzel et al. (1991).

During a trial, subjects heard a given speech segment repeated 5 times and then called out estimates of the azimuth (0 to 180 degrees left or right; 0 in front), elevation (0 to 90 degrees up or down; 0 at ear level), and distance (0 inches at the center, 4 inches at the edge of the head) which were recorded by the experimenter. Over the course of 2 to 3 days, each subject listened to 16 blocks of 30 stimuli containing a randomized ordering of the azimuth positions; targets at 0 and 180 were heard 150 times and all other locations were repeated 15 times.

Meta Platforms, Inc. - Ex. 1047
Meta Platforms, Inc. v. Eight kHz, LLC
IPR2023-01003, Page 1 of 5

RESULTS

Reversals

The presence of front-back "reversals" has been observed in nearly all studies of localization. These are responses which 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 is also heard. In the literature, reversals have generally been resolved when computing descriptive statistics (i.e., the responses are coded as if they 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 (e.g., Oldfield and Parker 1984; Wightman and Kistler 1989b).

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

The percentage of reversed judgements, averaged across all subjects, is shown in Figure 1 for the total number of judgements obtained at each target. The ratio of front to-back vs. back-to-front reversals found here (approximately 4:1) is comparable to that reported by Wenzel et al. (1991), but higher than the ratio of about 2:1 reported by Wightman & Kistler (1989b).

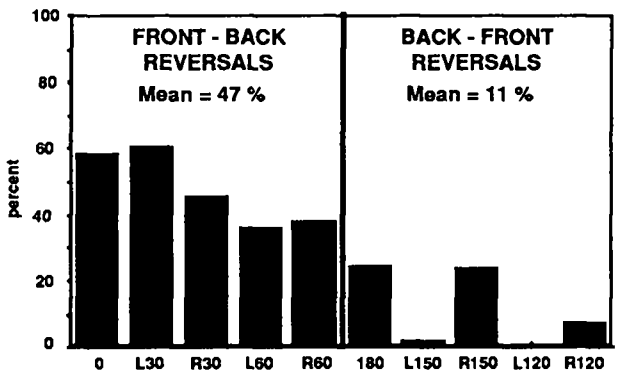

FIGURE 1: Total percentage of reversed judgements for each target position. 1650 judgements for 0, 180 degree targets; 165 judgements for all other targets; 11 subjects

Localization Error

Two measures of azimuth and elevation judgements, analogous to means and variances, were computed using spherical statistics (Fisher et al. 1987); the judgement centroid, the "mean direction" of a set of judgement vectors for a given target, and inverse kappa (K-1), a measure of dispersion around the centroid on the surface of a sphere. The average angle of error, the mean of the unsigned magnitude of judgement errors, was also computed.

Table I shows the reversal-corrected judgement 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. The mean value of K-1 from Table I is .102 and the mean average error angle for azimuth and elevation is 28 degrees. These values are somewhat larger than the corresponding values computed by Wightman & Kistler (1989b) for subject SDO listening with her own HRTFs; K-1 was .06 and the mean error angle was 20.5 degrees, for a range of different azimuths at middle elevations (0 and 18 degrees- up).

Target Azimuth	Azimuth Centroid	Elevation Centroid	Inverse kappa	Average error angle
L0	L5	U19	0.122	24.6
L30	L52	U38	0.126	47
L60	L79	U16	0.067	29.6
L90	L97	U13	0.053	17.6
L120	L111	U8	0.092	22.2
L150	L142	D2	0.131	25.3
180	L172	U9	0.127	21.7
R150	R143	U18	0.093	28.1
R120	R115	U16	0.102	27.5
R90	R102	U18	0.088	25.7
R60	R65	U25	0.094	31.5
R30	R39	U27	0.134	35
Mean			0.102	27.9

TABLE I: Localization data collapsed across all subjects for each target

Figure 2 shows the mean value for the azimuth centroids of all subjects, compared to the data for untrained subjects from the study by Wenzel et al. (1991). The latter study presented noise stimuli from a range of different elevations, but used the same HRTFs as here. The centroids for both studies, based on resolved judgements, are close to their intended target positions when viewed in this way. This suggests that the broadband characteristic of noise, compared to the relatively narrowband spectral content of speech, is not critical for equivalent azimuth performance in a 3-D headphone display system.

Meta Platforms, Inc. - Ex. 1047
Meta Platforms, Inc. v. Eight kHz, LLC
IPR2023-01003, Page 2 of 5

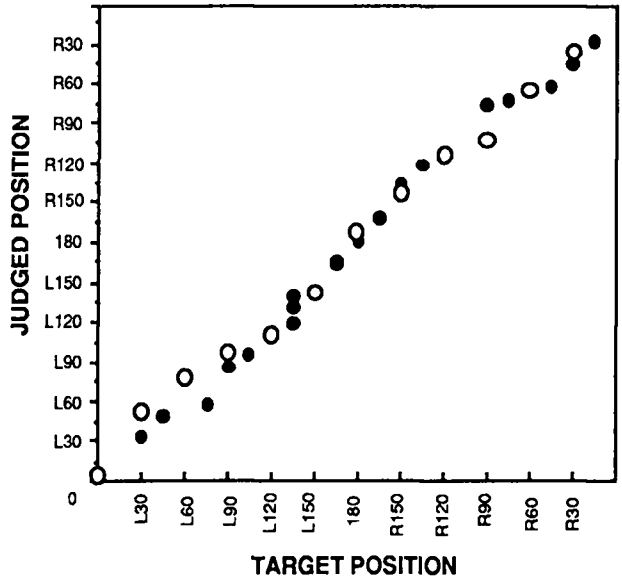

FIGURE 2: Centroids of azimuth judgements, based on reversal corrected judgements. Open circles: this study (speech). Closed circles: study by Wenzel et al. 1991 (noise).

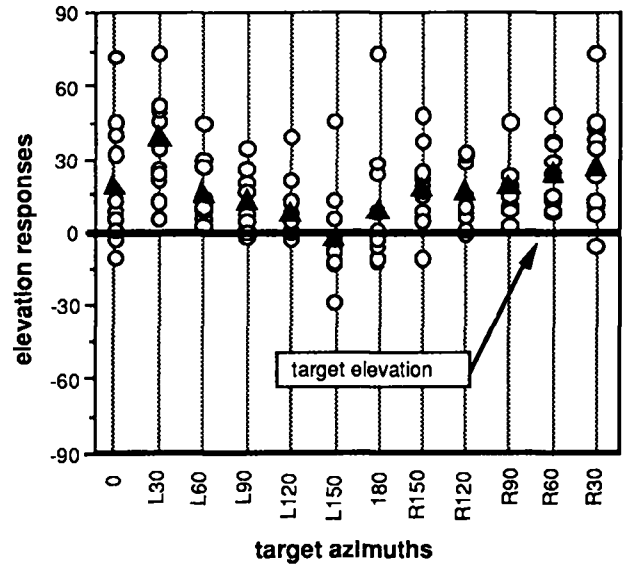

FIGURE 3: Centroids of elevation judgements, for each target azimuth. Open circles: individual judgements. Triangles: mean centroids collapsed across subjects. 1650 judgements for 0, 180 degree target; 165 judgements for other targets; 11 subjects

In looking at individuals' behavior, we observed three patterns of performance. The first pattern, seen with five subjects, was for resolved judgement centroids to be strongly correlated with the target positions. In the second pattern, another five subjects displayed a response bias in which their judgements tended to clump or "pull" toward the vertical-lateral plane passing through left and right ears . This pattern was also observed in the extensive free-field study conducted by Oldfield & Parker (1984) which they characterized as "defaults to 90". Finally, one subject "pulled" their judgements towards the vertical-median plane; that is, this listener had a response bias toward the front and rear positions.

Elevation

For most subjects, the HRTF-processed speech was perceived at a higher elevation than the target of 0 degrees (eye level); the mean value across azimuths is up-17 degrees (S.D., 10 degrees). Figure 3 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 degrees except for the left 150 target, and that the range of individual centroids is quite large (down 2 degrees to up 40 degrees).

The large dispersion of judgements is surprising in light of the fact that the target elevation of the stimuli never varied. This poor elevation performance is possibly due 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 little spectral energy (Blauert 1969; Rodgers 1981). However, the lack of a spectral cue to elevation does not explain the overall bias found here towards upward judgements. An elevation bias was not observed in the comparable study by Wenzel et al. (1991).

Distance and Externalization

Subjective impression of the distance of speech is particularly interesting since it is often supposed that distance perception is largely mediated by a listener's familiarity with the stimulus (Coleman 1962). The playback level of 53 dB SPL used in this experiment corresponds to a person speaking at a meter distant from the listener . In all cases the distance was underestimated (mean = 6.6 - 8.7 inches, S.D. 4.1 - 5.6) , as has been reported in distance studies with actual sound sources (Holt and Thurlow, 1969). Although the mean value for distance judgements for all subjects was externalized (i.e., greater than 4 inches) for all target positions , the standard deviations indicate that a large proportion of the sounds were heard inside the head.

Figure 4 shows the combined percentage of total judgements for targets reported to be less than four inches in distance (intracranial) and exactly at four inches distance (verged-cranial). Note that these combined percentages are always less than the percentage of externalized judgements. In particular, Figure 4 shows that about a third of the judgements for 0 and 180 degrees 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 the 0 and 180-degree positions; the range of responses heard verged or intracranially ranged from 15 - 46%.

Meta Platforms, Inc. - Ex. 1047
Meta Platforms, Inc. v. Eight kHz, LLC
IPR2023-01003, Page 3 of 5

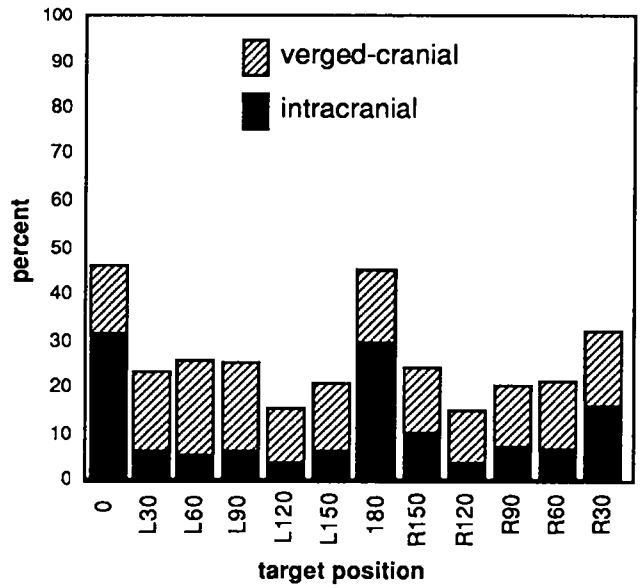

FIGURE 4: Percentage of judgements for each target position heard inside the head (intracranially: less than 4 inches) or on the edge of the head (verged cranially; at exactly 4 inches)

These results contrast strongly with those obtained by Plenge (1974), who used an artificial head in a reverberant environment and found that the vast majority of judgements were externalized. One possible reason for the relative lack of externalization here is the absence of environmental cues such as reverberation in the original HRTF measurements. We are currently examining whether adding modeled room reflections to speech stimuli increases externalization; pilot data from five subjects and data from a related study by Sakamoto et al. (1976) suggest that this is the case. Another factor may be the static nature of the stimuli, due to the lack of head movement (Wallach 1940).

DISCUSSION

Overall, the localization performance of untrained subjects was worse than data obtained for experienced subjects (Wightman & Kistler, 1989), paralleling the observations of Wenzel et al. (1991). Further, performance variability measures and rates of azimuth confusions for our inexperienced listeners with speech stimuli were generally comparable to localization of broadband noise by the untrained subjects of Wenzel et al. (1991). A primary difference between the studies was the fact that performance could be grouped according to response biases peculiar to individual subjects, for example, the tendency to clump responses near the vertical lateral or median planes. Another tendency was to elevate judgements, particularly for positions at the front. Finally, in a measure that was not reported in the studies cited above, we observed that stimuli failed to externalize for 15 to 46% of the judgements, with the smallest distances occurring near the median plane.

These data suggest that non-individualized HRTFs may have limited effectiveness for untrained subjects listening to speech in a 3-D auditory display system. It is possible, as Asano, et. al. (1990) have claimed, that errors diminish as subjects adapt to the unusual listening conditions provided by static anechoic sources. The difference in the performance of subjects in this study and the more experienced subjects of Wightman & Kistler (1989b) tend to support this view. Thus, it may be that some form of adaptation or training will be required to take full advantage of a virtual acoustic display.

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

Begault, D. R. (1990, September). Challenges to the successful implementation of 3-D sound. (Preprint 2948). Paper presented to the Audio Engineering Society 89th Convention, Los Angeles, California.

Begault, D. R., and Wenzel, E. M. (1990). Techniques and applications for binaural sound manipulation in human-machine interfaces (Tech. Memorandum 102279). Moffett Field: NASA-Ames Research Center.

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.

Coleman, P. D. (1962). Failure to localize the source distance of an unfamiliar sound. Journal of the Acoustical Society of America, 34(3), 345-346.

Fisher, N. I., Lewis, T., and Embleton, B. J. (1987). Statistical Analysis of Spherical Data. Cambridge, United Kingdom: Cambridge University Press.

Fisher, S.S., Wenzel, E.M., Coler, C. and McGreevy, (1988) M.W. Virtual interface environment workstations. Proceedings of the Human Factors Society 32nd Annual Meeting, 91-95.

Holt, R. E., and Thurlow, W. R. (1969). Subject orientation and judgement of distance of a sound source. Journal of the Acoustical Society of America, 46(6), 1584-5.

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

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 (Paper No. 82017) London, U.K.: Civil Aviation Authority.

Meta Platforms, Inc. - Ex. 1047
Meta Platforms, Inc. v. Eight kHz, LLC
IPR2023-01003, Page 4 of 5

Plenge, G. (1974). On the difference between localization and lateralization. Journal of the Acoustical Society of America, 56(3), 944-951.

Rodgers, C. A. (1981). Pinna transformations and sound reproduction. Journal of the Audio Engineering Society, 29(4), 226-234.

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

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

Wenzel, E. M., Wightman, F. L., and Kistler, D. J. (1991). Localization of non-individualized virtual acoustic display cues (in press). Proceedings of CHI'91 ACM Conference on Human Factors in Computing Systems New York: Association for Computing Machinery

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(2), 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(2), 868-878.

Meta Platforms, Inc. - Ex. 1047
Meta Platforms, Inc. v. Eight kHz, LLC
IPR2023-01003, Page 5 of 5