# Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source*

**DURAND R. BEGAULT,** *AES Member,* **AND ELIZABETH M. WENZEL,** *AES Member*

*Human Factors Research and Technology Division, NASA Ames Research Center, Moffett Field, CA 94035, USA*

**AND**

**MARK R. ANDERSON**

*Raytheon Technical Services Company, Moffett Field, CA 94035, USA*

A study of sound localization performance was conducted using headphone-delivered virtual speech stimuli, rendered via HRTF-based acoustic auralization software and hardware, and blocked-meatus HRTF measurements. The independent variables were chosen to evaluate commonly held assumptions in the literature regarding improved localization: inclusion of head tracking, individualized HRTFs, and early and diffuse reflections. Significant effects were found for azimuth and elevation error, reversal rates, and externalization.

## 0 INTRODUCTION

A review of the perceptual literature relevant to virtual acoustic displays ("3-D sound") suggests that both localization performance and perceived realism are optimal when the cues used in everyday spatial hearing are reproduced as faithfully as possible. Currently the technological approaches typically used for causing more "naturalistic" spatial hearing experiences in virtual acoustic displays are:

1) Means for allowing a virtual source to remain in a constant position relative to the orientation of the listener, by the use of head-tracked virtual stimuli [1]–[3]

2) Use of individualized (as opposed to nonindividualized or "generic") head-related transfer functions (HRTFs) for determining digital filters, so that the processing of sound sources corresponds to the listener's own body, head, and pinnae [4]–[6]

3) Synthesis of a realistic reverberant context for sound sources, via "auralization" techniques that use HRTF-filtered early reflections and similarly realistic representations of the diffuse reverberant sound field [7]–[9].

It is surprising that the relative advantage of these methods has never been compared within a single experimental paradigm, particularly for the purpose of minimizing a listener's "localization error" when using an acoustic display. Unfortunately one cannot usefully compare previous studies that investigate these techniques in isolation because experimental designs and methodologies differ. In addition, the results of localization studies using artificial stimuli such as noise or clicks cannot be compared to those studies that use "real-world" stimuli such as speech.[1] For example, a subject cannot use a cognitive reference to previous experience in order to determine the distance of a noise burst.

The current study determines, in a directly comparable manner, the contribution of head tracking, reverberation, and individualized-HRTF cues to the reduction in local-

* Presented at the 108th Convention of the Audio Engineering Society, Paris, France, 2000 February 19–22; revised 2001 August 29.

[1] Noise stimuli can simulate spatial cues in higher frequency regions of the audible range that might otherwise not be observable. By contrast, the long-term average level of speech stimuli has maximum energy in the 500-Hz octave band, with a 6-dB amplitude rolloff per octave frequency above and below this band.

ization errors within an auditory display. This study also gathered overall judgments for perceived realism of the stimuli. In this experiment all listeners are exposed to all conditions and all combinations of these independent variables. Although these cues are available simultaneously under normal listening conditions, it is possible within a simulation context to isolate each of these cues. Speech stimuli were studied because of their use in applications such as virtual audio teleconferencing.

Fig. 1 overviews a taxonomy of sound localization errors that are analyzed in the current experiment and that are characteristically analyzed in the literature. *Localization* error refers to the deviation of the reported position of a sound stimulus from a measured or synthesized "target" location. In this study, azimuth errors (deviations along the horizontal plane) and elevation errors (deviations from eye-level elevation) are evaluated separately. An *externalization* error refers to a judgment of the distance of a sound stimulus as within or at the edge of the head. This is sometimes termed "inside-the-head localization," "lateralization," or "intracranial localization" [10], [11]. The goal of virtual acoustic synthesis is usually to produce sounds that seem externalized, that is, outside the listener's body. A *reversal* error (sometimes termed front–back or back–front "confusion") refers to the judgment of a sound stimulus as located on the opposite side of the interaural axis than the target position. Front–back reversals are particularly endemic with three-dimensional sound reproduction over headphones [12].

Virtual stimuli processed using nonindividualized (generic) HRTFs as opposed to individualized HRTFs have been cited in the literature as degrading localization accuracy, decreasing externalization, and increasing reversal errors [1], [6], [13]–[16]. However, these conclusions are typically based on full-spectrum noise stimuli. Results

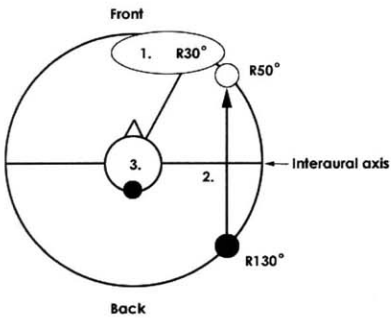


Fig. 1. Typical methods of categorizing localization errors in virtual acoustic headphone studies include: 1) description of deviation of a judged position from a target position in terms of azimuth and elevation (several presentations of a target stimulus at right 30° might be heard at azimuths within region shown); 2) reversal rates, that is, rate at which stimuli are identified on incorrect side of interaural axis (shown is a back–front reversal, with a target at 130° judged to be at 50°); 3) distance errors, categorized in terms of whether or not sound is heard outside or inside head (shown is an unexternalized target heard at the edge of the head; alternatively, quantitative rather than ordinal judgments of distance can be elicited).

that are to some extent analogous to a comparison of noise and speech stimuli were obtained by Bronkhorst, who compared 7-kHz low-pass-filtered harmonic stimuli with a fundamental frequency of 250 Hz to an unfiltered version extending to 15 kHz (the one-third-octave power spectrum was flat) [16]. Both real and virtual source localization (head-tracked virtual stimuli using individualized HRTFs) was evaluated. Localization of the low-pass virtual and real sources was nearly identical, but localization was significantly better for real sources with the unfiltered stimuli. This was suggested to be a result of inaccuracies in the simulation of pinna spectral cues above 7 kHz for the virtual stimuli.

Reversals of stimuli across the interaural axis have been cited in the literature as diminishing when head movements correspond to realistic changes in stimuli position. This has been hypothesized as resulting from the ability to track the size and direction of interaural cues over time [17], [18]; see also [10]. For virtual acoustic presentation of broad-band Gaussian noise stimuli, one study found about a 7:1 decrease in front–back reversals (42%–7%) and a 2:1 decrease in back–front reversals (13%–7%) when head motion cues were supplied within a virtual simulation [19]. Another study found that either head movement or source movement under the listener's control reduced front–back errors significantly, but such errors were not reduced with source movement alone [20]. Although it has been proposed informally that reverberant cues in the form of unique patterns of early reflections may help front–back discrimination based on the familiarity of the effect on timbre cues, this has not yet been verified experimentally [21].

Bronkhorst found no significant effect of using individualized HRTFs on reversals, in contrast to a noise-stimulus experiment by Wenzel et al., which indicated that individualized HRTFs mitigated reversal "confusions" [4], [16], [22]. Møller et al. conducted a source-identification experiment using speech stimuli, which suggested that nonindividualized HRTFs resulted in an increased number of reversals, but had no effect on externalization. However, all the conditions in that study were made under reverberant conditions and without head tracking [6].

The literature indicates that reverberation, even in the form of a few "early reflections" or attenuated, delayed copies of the direct sound, is sufficient to produce image externalization [8], [23]. Like front–back reversals, externalization has been recognized as a problem for headphone reproduction for some time. A typical paradigm in externalization studies is to have subjects compare anechoic and reverberant stimuli. Durlach and Colburn have stated that the externalization of a sound source is difficult to predict with precision, but "… it increases as the stimulation approximates more closely a stimulation that is natural" [24]. The means of simulating externalized stimuli successfully is typically attributed to the use of personal binaural HRTFs, head movement, and/or reverberation. Another approach has been from the perspective of simulating a loudspeaker listening experience over headphones, although there are few experimental data to support the engineering concept described in [10], [25].

## 1 EXPERIMENTAL DESIGN

The experimental design shown in Table 1 was developed to evaluate all combinations of variables to be studied. "Anechoic," "early reflection," and "full auralization" refer to the level of diffuse-field simulation derived from a room modeling program. Anechoic stimuli contained only a model of the direct path of an HRTF-filtered signal. Early-reflection stimuli added HRTF-filtered early reflections derived from the room model to the direct path, based on the first 80 ms of the impulse response. Full-auralization stimuli included the entire diffuse sound field (the same anechoic and early reflection stimuli, plus additional simulation of the late reverberant sound field from 80 ms to 2.2 s). Individualized HRTFs were measured for each subject, whereas generic HRTFs were derived from a dummy head developed at the Institut für Technische Akustik at the University of Aachen (supplied as part of the room modeling software). "Tracking on–off" refers to whether or not the direct-sound HRTFs and (when applicable) the early-reflection HRTFs were updated in real time in response to head movement.

Each subject was run under each of the treatment combinations shown in Table 1, resulting in twelve conditions. All positions simulated were at eye level (0° elevation). Only six azimuth positions were used: left and right 45°; left and right 135°; and 0 and 180° (0° is referenced to directly ahead of the listener). These positions represent pairs that lie on two different "cones of confusion" and a pair directly on the median plane [26]. The elements of each pair are similar in that the interaural time difference (ITD) is nearly the same for left 45 and 135°, 0 and 180°, and right 45 and 135° for frequencies below 1.5 kHz.

Each azimuth position was evaluated five times in each block, resulting in 30 trials per block. The order of blocks, azimuth positions, and speech stimuli was randomized, whereas the particular combination of experimental treatments was held constant within each block. Each trial took approximately 10 s to complete.

## 2 SUBJECTS

Nine paid participants (four female, five male, age range 18 – 40, hearing ability ≤15 dB HL) inexperienced with virtual acoustic experiments participated. They were given no details about the experiment. Prior to the experiment, subjects were instructed how to make azimuth, elevation, distance, and realism judgments on the interactive, self-paced software interface to be described. For distance judgments they were instructed to pay attention to whether or not the sound seemed inside or outside the head, and that the maximum distance judgment possible on the graphic corresponded to "more than 4 inches outside the edge of the head." Instructions were given on the computer screen before each experimental trial that utilized head tracking requesting that subjects move their heads; they were not trained to move their heads in any particular manner. Subjects were encouraged to stop when fatigued, and were given breaks between experimental blocks at their request. Three to four days were typically required to complete the entire experiment.

## 3 STIMULI

### 3.1 Experimental Hardware and Software

Brief (3-s) segments of speech stimuli were used in the current experiment, both because reverberation was to be evaluated and because of the relevance of the spoken word to three-dimensional audio applications such as teleconferencing, virtual home theater, and so on. The tradeoff in using speech for evaluating localization accuracy is that a long-term average speech spectrum does not contain a significant level of acoustical power at those frequencies where the HRTF yields elevation cues that can be used by a listener. For that reason, only eye-level elevations were evaluated in the current study, as in a previous speech localization study we had conducted [27].

Stimuli were presented to subjects over stereophonic headphones (Sennheiser HD-430) in a double-walled soundproof booth (Industrial Acoustics Company), at a level of approximately 60 dB (A weighted). This level corresponds to normal speech at a distance of 1 m. The A-weighted background noise level in the booth was 19 dB. A head-tracking device (Polhemus FastTrak) was attached at all times to the subjects' headphones. The head tracker was interfaced via a TCP/IP socket connection with the simulation software and hardware (Lake Technology Headscape and Vrack software, CP-4 hardware) updating the stimuli in response to head movements at an update interval of 33 Hz (30 ms). The end-to-end latency throughout the entire hardware system averaged 45.3 ms (SD = 13.1 ms), as measured via a swing-arm apparatus (described in [28], [29]).

For simulating different azimuths, the virtual simulation reoriented the receiver position about the listener's

Table 1. Experimental conditions.

| Reverberation type | Anechoic | | | | Early reflections | | | | Full auralization | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HRTF used | Individual | | Generic | | Individual | | Generic | | Individual | | Generic | |
| Head tracking | On | Off | On | Off | On | Off | On | Off | On | Off | On | Off |
| Variables | | | | | | | | | | | | |
| Externalization error | • | • | • | • | • | • | • | • | • | • | • | • |
| Reversal error | • | • | • | • | • | • | • | • | • | • | • | • |
| Azimuth error | • | • | • | • | • | • | • | • | • | • | • | • |
| Elevation error | • | • | • | • | • | • | • | • | • | • | • | • |
| Realism rating | • | • | • | • | • | • | • | • | • | • | • | • |

central axis to face different directions within the room, rather than moving the virtual sound source about the listener. The hardware did not resynthesize the stimuli in response to subject tilt or roll, only to yaw (azimuth). Nevertheless, the simulation always included a full three-dimensional rendering of the reverberant field.

Custom software drove the experimental trials and data collection, and generated playback of speech recordings from a digital sampler (randomized 3-s anechoic female and male voice segments taken from the psychoacoustic test CD "Music for Archimedes" [30]). The software also coordinated the Lake Vrack program, the sound generator, the head-tracking device, and the stimulus generation.

Subjects indicated their responses via computer mouse using an interactive graphic, as shown in Fig. 2. The details of the interface were described previously [31]. Subjects were required to indicate first the azimuth and the distance on the left panel, and then the elevation on the right panel of the display, and then finally to adjust the slider at the bottom to indicate "perceived realism." Subjects were forced to make all judgments before they could proceed to the next trial.

The distance of the outer circle from the center of the head was twice the radius from the center to the edge of the head in order to prevent biasing the externalization judgments. For example, a very large head relative to the available area on the display would probably yield more inside-the-head localization judgments. The perceived realism was indicated on a continuous slider bar with labels "bad," "poor," "fair," "good," and "excellent" present at locations corresponding to a 0 – 4 rating scale. No special instructions were given on how to interpret "realism" or what to listen for specifically.

### 3.2 HRTF Measurement

Before starting the experiment, subjects had their HRTFs measured using a modified Crystal River Engineering Snapshot system within the same soundproof booth used for the presentation of the stimuli. The Snapshot system is based a blocked-meatus measurement technique (see, for example, [32]) and has been used in our previous experiments. The HRTF was measured iteratively via a single loudspeaker, with the subject moving on a rotating chair for each measurement; the loudspeaker was moved vertically along a pole to obtain a set of azimuth measurements at a specific elevation. Wall reflections are eliminated analytically from the measurement, and the frequency response is made flat below 400 Hz to compensate for the nonlinear response of the loudspeaker and the listening booth.

A software script (Matlab from The MathWorks, Inc.) guided the experimenter in measuring the HRTF map at 30° azimuth increments, at six elevations: −36, −18, 0, 18, 36, and 54° relative to eye level. A head-tracking device was monitored by the experimenter to position the subject for each measurement. Golay code pairs were used to obtain the raw impulse response. Subsequently a diffuse-field equalization that compensated for the headphone, microphone, and loudspeaker transfer functions was applied. The measurements are eventually processed into a HRTF "map," consisting of an array of 128-point minimum-phase impulse responses at a sample rate of 44.1 kHz.

### 3.3 Reverberation Simulation

A room prediction–auralization software package (CATT-Acoustic) was used to generate both an individualized and a generic binaural version of an existing multipurpose performance space. This space has a volume of about 1000 $m^3$ (8.9 m width, 14.3 m length, 7.9 m height) and is essentially a rectangular box with gypsum board mounted on stud walls, a cement ceiling, and a wooden parquet floor with seating. Velour curtains are on the two opposite long sides of the hall. The modeled source and receiver were positioned asymmetrically in the room 5.3 m apart, 1 m off the centerline of the room, and with the source 4.7 m off the back wall. The source directivity was specified in octave bands based on a model of the human

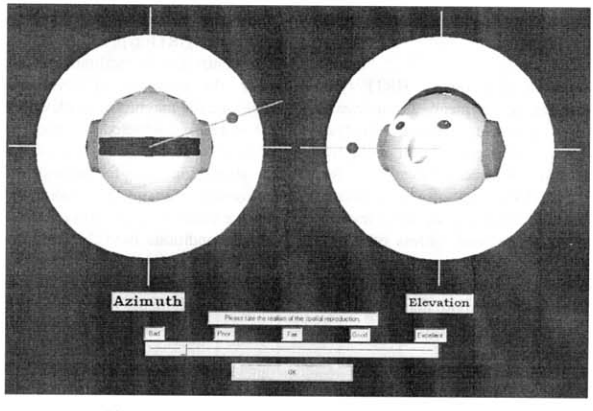


Fig. 2. Subjective response screen graphic used in experiment. Azimuth and distance judgments are made via a computer mouse on left view; elevation on right view; perceived realism on slider.

voice derived from [33]. 10 088 rays were calculated using a hybrid cone-tracing technique (see [34], [35]).

The acoustical characteristics of the modeled room in the experiment were initially compared to measurements made in the real room to verify the match between significant acoustical parameters such as octave-band reverberation times and early-reflection arrivals (see [36]). The agreement between the modeled and real room was within 0.2 s in each octave band from 125 Hz to 4 kHz, with a midband reverberation time of 0.9 s. However, the existing space was considered overly dry (nonreverberant) for purposes of the current experiment; the reverberation for the full auralization was meant to be obvious to the nonexpert subjects. Consequently the absorptive materials in the virtual room were adjusted to achieve a midband reverberation time of 1.5 s by altering the absorption coefficients used in the room model. (The velour curtains were "opened" to expose the gypsum board panels on the sidewalls and audience absorption was removed.) The resulting reverberation times are shown in Table 2. The early reflections out to 80 ms contained no noticeable echoes, and were uniformly dense from all directions after about 12 ms.

A customized software script (Mathwork's Matlab) took the complete measured HRTF map made by Snapshot and produced an interpolated, upsampled (44.1 – 48-kHz) version for implementation into the room modeling program. The room modeling program then generated 256 binaural impulse responses of the first 80 ms of the room response, representing 1.5° increments of listener motion relative to the source. In addition, two binaural impulse responses were generated representing the diffuse field from 80 ms to 2 s. The early-reflection binaural impulse responses were updated in real time for both the anechoic and the early-reflection conditions by the Lake hardware/software for real-time, low-latency convolution. The anechoic condition was achieved by using only the part of the impulse response representing the direct arrival. The full-auralization condition included the late diffuse reverberation response (80 ms to 2.2 s), which did not vary with head motion.

## 4 RESULTS

A 2 × 2 × 3 (head tracking: on/off; HRTF type: generic/individual; reverberation treatment: anechoic/early reflections/diffuse) univariate repeated-measures analysis of variance (ANOVA) was conducted for each of the five dependent measures: azimuth error, elevation error, front–back reversal rate, externalization, and subjective rating of sound realism (see Table 1). An alpha level of 0.05 was used for all statistical tests, unless otherwise indicated. The Geisser–Greenhouse correction was used to adjust for assumed violations of sphericity when testing repeated-measures effects involving more than two levels.

### 4.1 Azimuth Error

Azimuth error was defined as the unsigned deviation, in degrees, of each azimuth judgment from the target azimuth location, corrected for frontal plane and median plane reversals. The azimuth error for each independent variable tested was calculated as the mean of the unsigned error across all of the target positions that were evaluated. The main effects for head tracking and HRTF type were nonsignificant, but a significant main effect was found for reverberation type, $F(2, 16) = 5.39$, $p = 0.016$; see Fig. 3. The analysis also revealed a significant two-way interaction between head tracking and HRTF for azimuth error, $F(1, 8) = 20.27$, $p = 0.002$; see Fig. 4. For generic HRTF conditions, head-tracked stimuli resulted in smaller azimuth errors (mean = 16.9, SD = 7.8) than stimuli without head tracking (mean = 21.7, SD = 7.8).

Fig. 5 shows the mean values for unsigned azimuth error for individual subjects in comparison to the overall mean, grouped by experimental condition. The relatively larger distribution of means between the anechoic condition and the reverberant conditions is visible. Also notable are the differences between subjects as a function of the manipulated variable. Subject LH (indicated by filled circles, dashed line) had an azimuth error consistently higher for anechoic (30 – 50°) versus reverberant conditions (15 – 19°). In other words, LH had a 2:1 improvement in azimuth performance with the presence of reverberation. By contrast, azimuth performance for subjects TZ (unfilled circles, dashed line) and EL (filled diamond) improved by nearly a 2:1 ratio when head tracking was present but appeared largely unaffected by the presence of reverberation.

### 4.2 Elevation Error

Elevation error was defined as the unsigned deviation, in degrees, of an elevation judgment from an eye-level target (0° elevation). A significant main effect was found for reverberation, $F(1.20, 9.57) = 5.15$, $p = 0.043$, applying the Geisser–Greenhouse correction; see Fig. 6. No significant main effects or interactions were found for head tracking or HRTF type.

In contrast to its facilitating effect in reducing azimuth error, the presence of reverberation in the stimuli increased the magnitude of elevation errors (mean = 28.7, SD = 17.8), compared to anechoic conditions (mean = 17.6, SD = 14.6). Fig. 7 shows the mean values for signed elevation judgments for individual subjects compared to the overall mean, grouped by experimental condition. The relative increase in elevation between anechoic and reverberant conditions is visible. Also notable are the differences between subjects with regard to an overall bias in all of their elevation judgments. For example, subject KP's elevation judgments (filled triangles, dashed lines) were

Table 2. Adjusted reverberation times used in room model.

| Octave-band center frequency (Hz) | 125 Hz | 250 Hz | 500 Hz | 1 kHz | 2 kHz | 4 kHz |
|---|---|---|---|---|---|---|
| Reverberation time *T-30* (s) | 2.2 | 1.9 | 1.5 | 1.5 | 1.5 | 1.4 |

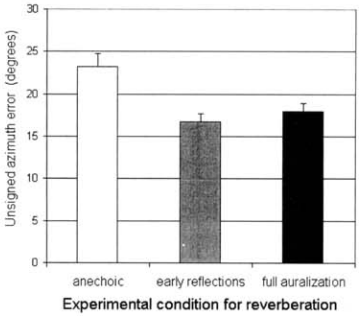


Fig. 3. Unsigned azimuth error—significant main effect for reverberation (mean values and standard error bars).

consistently higher (mean 42°) than those of subject BP (mean 0.3°, unfilled triangles, dashed line).

### 4.3 Reversal Rates

The ANOVA indicated that head tracking reduced reversals significantly (front–back and back–front confusion rates) compared to stimuli without head tracking, $F(1, 8) = 31.14$, $p = 0.001$; see Fig. 8. The overall mean reversal rate for head-tracked conditions was 28% (SD = 25%) compared to the 59% (SD = 12%) reversal rate for non-head-tracked conditions. No other treatments or interactions yielded significant effects on reversal rates.

Fig. 9 shows percentages of reversed judgments for individual subjects, categorized as to the percentage of frontal stimuli reversed to the rear (front–back reversals), to the front (back–front reversals), and for all stimuli

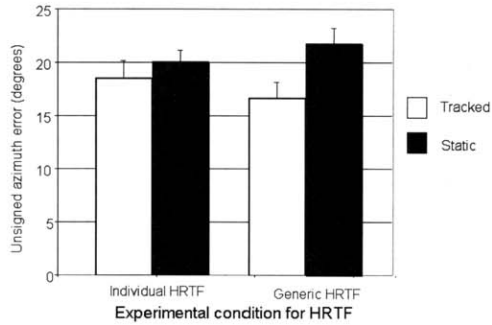


Fig. 4. Unsigned azimuth error—significant interaction for head tracking and HRTF-type reverberation (mean values and standard error bars).

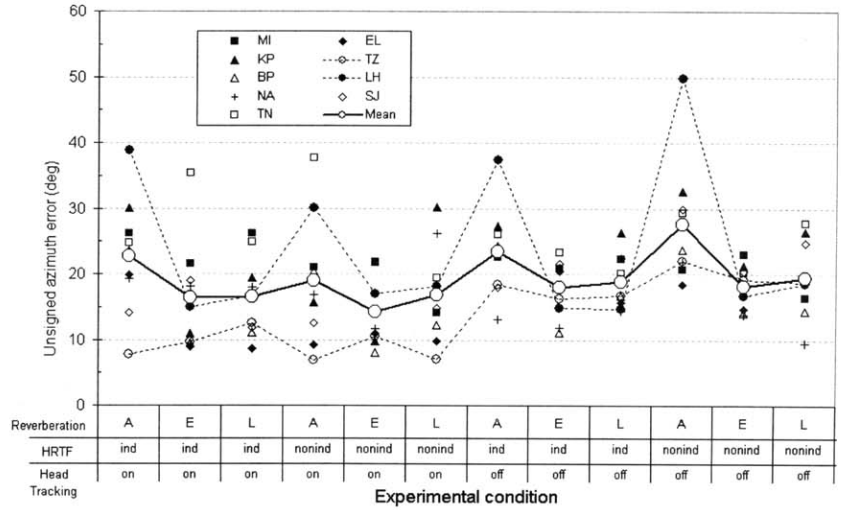


Fig. 5. Unsigned azimuth error—values for individual subjects, compared to mean values across subjects (unfilled circles connected by solid line). A—anechoic; E—early reflections; L—early and late reverberation (full auralization).

(mean of the front–back and back–front reversal rates). Across all conditions, four of the subjects (EL, MI, LH, BP) made mostly front–back reversed judgments. Two subjects' (NA and KP) reversals were mostly back–front. The remaining three subjects (TZ, SJ, and TN) had no specific bias for front–back versus back–front reversals.

### 4.4 Externalization

A sound is externalized if its location is judged to be outside the head. For this analysis, the cutoff point for treating a judgment as "externalized" was set to >5 inches in order to yield a conservative estimate that eliminated judgments perceived at the edge of the head ("verged cranial"; see [11]). Note that the unit of an inch is relative to the graphic shown in Fig. 1, and does not represent a literal judgment in inches by the listener. The edge of the head is set at 4 inches, and the maximum distance judgment possible by the subject was 8 inches.

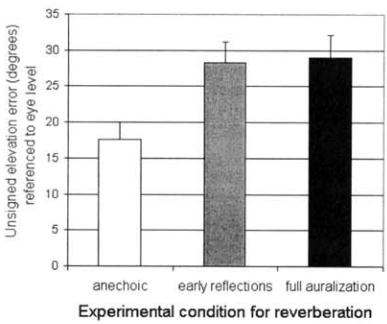


Fig. 6. Elevation—main effect for reverberation (mean values and standard error bars).

A significant main effect of reverberation was found for the proportion of externalized distance judgments, $F(1.43, 11.43) = 13.43$, $p = 0.002$ (Geisser–Greenhouse correction); see Fig. 10. Subjects externalized judgments at a mean rate of 79% (SD = 23%) under the combined reverberant conditions, compared to 40% (SD = 29%) under the anechoic condition. No other treatments or interactions yielded significant effects for externalized stimuli.

### 4.5 Perceived Realism

Listeners were asked to rate the perceived realism of each stimulus on a continuous scale, which was subsequently encoded from 0 (least realistic) to 4 (most realistic). No significant main effects or interactions were found. Realism ratings for each of the 12 conditions, averaged over all nine participants, varied only from 2.42 and 2.97, with an overall mean realism rating of 2.71 (SD = 0.55) on the 0 – 4 scale. This lack of variability suggests that the participants did not differentiate among conditions based on perceived realism, or that they did not have a common understanding of what "realism" meant.

## 5 DISCUSSION

It is possible to evaluate these results in light of system requirements for improved virtual acoustic simulation. Overall, these results would seem to indicate that stimuli which include reverberation will yield lower azimuth errors and higher externalization rates (here by a ratio of about 2:1), but at the sacrifice of elevation accuracy. The

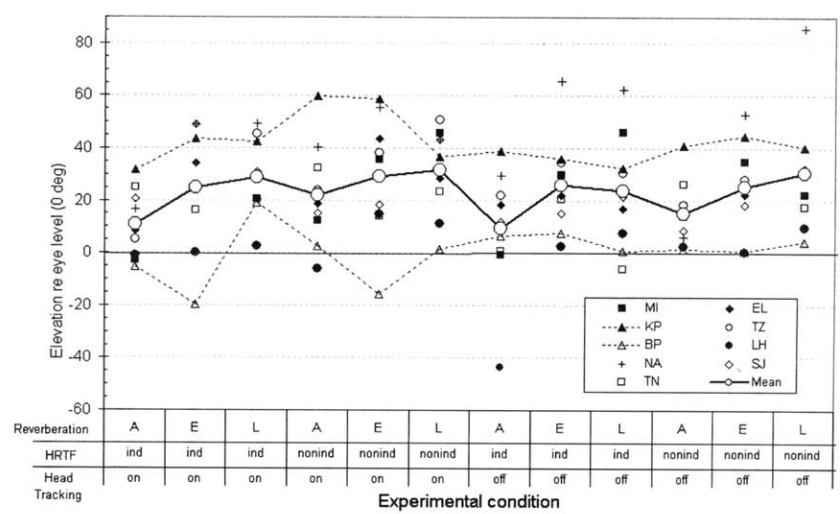


Fig. 7. Elevation error—values for individual subjects, compared to mean values across subjects (unfilled circles connected by solid line). A—anechoic; E—early reflections; L—early and late reverberation (full auralization).

inclusion of head tracking will reduce reversal rates significantly, also by a ratio of about 2:1, but does not improve localization accuracy or externalization. Except for the interaction of head tracking and HRTFs for azimuth error, there is no clear advantage to including individualized HRTFs for improving localization accuracy, externalization, or reversal rates within a virtual acoustic display of speech.

The fact that individualized HRTFs did not increase azimuth accuracy significantly is perhaps explained by the fact that most of the spectral energy of speech is in a frequency region where ITD cues are more significant than spectral cues. Møller et al. also found that individualized HRTFs gave no advantage for localization accuracy for speech [6]. One might expect a strong correlation between azimuth error and the magnitude of the ITD difference between individual and generic HRTFs for each subject. In other words, a predictor of localization errors might be the degree to which a subject's head size matches a nonindividualized head. The average of each subject's ITD at left and right 90° is indicative of the subject's relative head size; the range was 0.54 – 0.69 ms for the measured subjects, and 0.65 ms for the generic HRTF. The ITD difference for each subject relative to the generic HRTF was compared to the net increase in azimuth error when listening under the generic HRTF condition. This analysis showed no correlation between azimuth error and head size difference. This may be due to the limited number of target positions that were evaluated in the current study.

It has been suggested previously that reversals are mitigated by individualized HRTFs [4]–[6], [15]. For instance, for noise stimuli in a nondynamic simulation, an almost 3:1 decrease in reversals (31% to 11%) is found when comparing data for nonindividualized HRTFs used in [4], versus the individualized HRTFs used in [22]. This is not the case for the speech stimuli used here. In fact, for non-head-tracked speech stimuli, the mean reversal rate found here (59%) was much higher than that found previously (37%) [27].

Previous studies have reported on the problem of inside-the-head localization and the externalization advantage of using reverberation [8], [23], [37]. The current study indicates that the presence of early reflections out to 80 ms beyond the direct sound is sufficient to provide externalization; a full auralization of late reflections out to 1.5 s is not necessary. The early reflections cause a lowering of the interaural cross correlation of the binaural signal over time as the speech is voiced, relative to the ane-

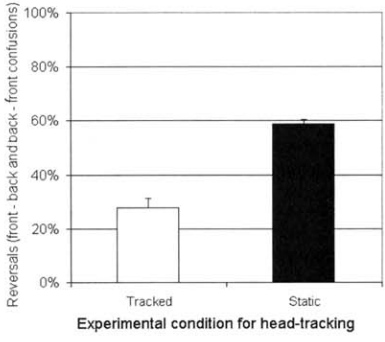


Fig. 8. Reversals—main effect for head tracking (mean values and standard error bars).

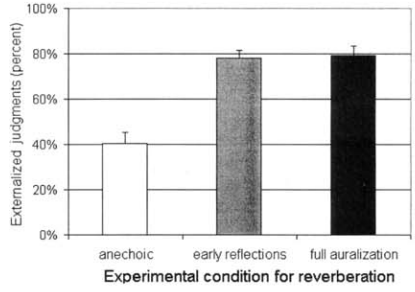


Fig. 10. Percentage of externalized judgments—main effect for reverberation (mean values and standard error bars).

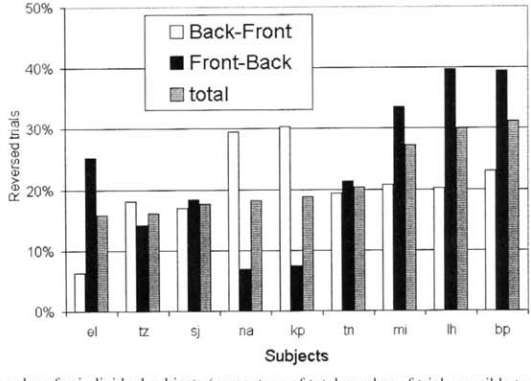


Fig. 9. Reversals—mean values for individual subjects (percentage of total number of trials possible to reverse in each direction).

choic stimuli [10]. It is possible that this increased differentiation of the binaural signal over time is responsible for the externalization effect, as opposed to the cognitive recognition of a room.

It was mildly surprising that the presence of reverberation caused the accuracy of azimuth estimation to improve by approximately 5°. Typically reverberation introduces a smearing effect in the form of image broadening that would make the loci of the localized image less precise. On the other hand, nonexternalized responses are not actually "localized" in the normal sense. It is possible that sounds heard within the head are less precisely localized, and that localization resolution improves beyond a certain distance. There could also have been a response bias; it may have been easier on the graphic response screen to represent a perceived azimuth more accurately when the distance judgment had increased.

Figs. 11 and 12 show the relationship between perceived externalization and azimuth and elevation judgment error, respectively. Open circles represent the anechoic simulation, and filled circles indicate both reverberation conditions used in this study. Although Fig. 11 indicates a mild azimuth error decrease with externalization, Fig. 12 shows a stronger increase in elevation error with externalization. Elevation biases were observed previously for virtual speech stimuli using nonindividualized HRTFs [21], [27]. Particularly for sound sources at 0° azimuth and elevation, virtual acoustic stimuli (as well as dummy-head recordings) are frequently perceived as elevated, within or at the edge of the head, when heard through headphones. Fig. 7 indicates that this upward bias is present in the current data. There is no explanation for this phenomenon at this point.

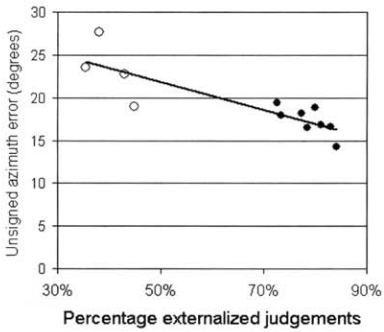


Fig. 11. Azimuth error as a function of externalization—mean values across subjects. ○ —anechoic stimuli; ● —reverberant stimuli (combined data for early reflections and full auralization); —— —linear curve fit to data.

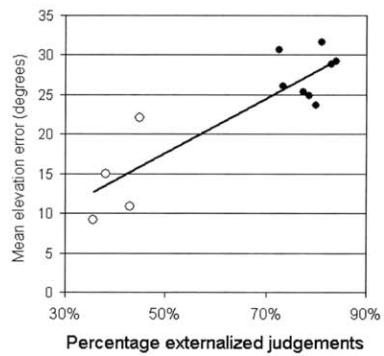


Fig. 12. Elevation error as a function of externalization—mean values across subjects. ○ —anechoic stimuli; ● —reverberant stimuli (combined data for early reflections and full auralization); —— —linear curve fit to data.

Head tracking did not increase externalization rates significantly, nor did it yield more accurate judgments of azimuth. Head tracking primarily served to eliminate reversals, a phenomenon explained by the differential integration of interaural cues over time as the cone of confusion is resolved by head motion [10]. These findings contrast previous results indicating that head movements enhance source externalization [23], [38], [39] and, to some extent, localization accuracy [40].

Fig. 13 shows a rank ordering of subjects by average azimuth error across all head-tracked conditions tested, along with the mean and the standard deviation of head movement (yaw). The Pearson product moment correlation coefficient $r$ between the mean value for yaw and azimuth error is –0.75, and between the standard deviation of the yaw and azimuth error it is –0.83. Although there is an overall trend, the data do not clearly suggest that each subject's localization accuracy is tied to the magnitude of average head movement, nor can it be implied that reducing or increasing head movement for any particular subject would create a corresponding change in their localization accuracy. Nevertheless, as seen in Fig. 13, it is interesting that the best localizer (TN) utilized the greatest

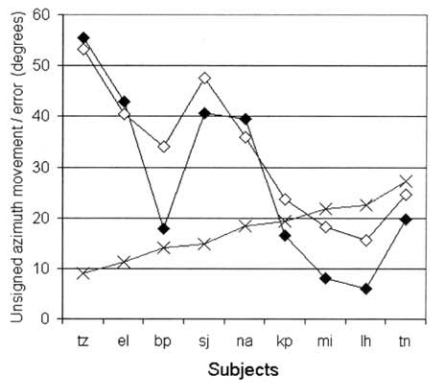


Fig. 13. Rank ordering of subjects by average azimuth error under head-tracked conditions. ×—mean azimuth error for each subject; ◆ —mean unsigned head movement (yaw) under head-tracked conditions; ◇ —standard deviation for unsigned head movement.

amount of head movement, whereas the worst three localizers' average head movements were less than their average azimuth errors. The fact that the stimuli were only 3 s long perhaps limited the ability of the subjects to "take advantage" of cues derived from head movement.

The insignificance of the ratings given to perceived realism as a function of experimental condition could have been affected by many factors. No instructions were given, perhaps causing subjects to utilize different criteria that all resulted in a relatively "neutral" judgment. It was surprising that full auralization with head tracking was not judged as significantly more realistic than anechoic conditions. It may simply be that realism is difficult to associate with 3-s segment of speech in a laboratory simulation.

As a final note it should be emphasized that the current experimental design was for the examination of the localization "error," that is, the divergence between reported auditory image positions and intended sound source positions. This approach is useful from the standpoint of audio or human factors engineering, but it does not address sound localization in terms of either an individual listener's localization bias or with respect to the frequent lack of correspondence between sound source events and auditory images [10], [41]. To examine the effect of the experimental variables on individual localization of auditory images, a separate analysis was conducted by referencing the data to the subject's personal localization bias. Ideally, this bias would be determined from the localization of actual sound sources within an anechoic chamber, but this was not possible in the current experiment. As an alternative, localization judgments made under the virtual source condition most comparable to anechoic localization of actual sources was used as a reference. This criterion was best met by the head-tracked individualized-HRTF anechoic condition.

Fig. 14 shows, for this reference condition, the means of the five judgments made for each target azimuth. These data were used to make a signed correction to each subject's azimuth judgments made under the other experimental conditions. The unsigned magnitude of the error judgment was then calculated across all azimuths for each condition, as described in Section 4.1. The subsequent ANOVA showed a significant two-way interaction between reverberation type and head tracking, $F(2, 16) = 7.88$, $p = 0.004$ (see Fig. 15). This contrasts with the analysis made in Section 4.1 for localization errors, where there was a significant effect of reverberation and a significant two-way interaction between head tracking and HRTF type (Figs. 3 and 4). Overall, the effect of correcting the data to the reference condition was small, with

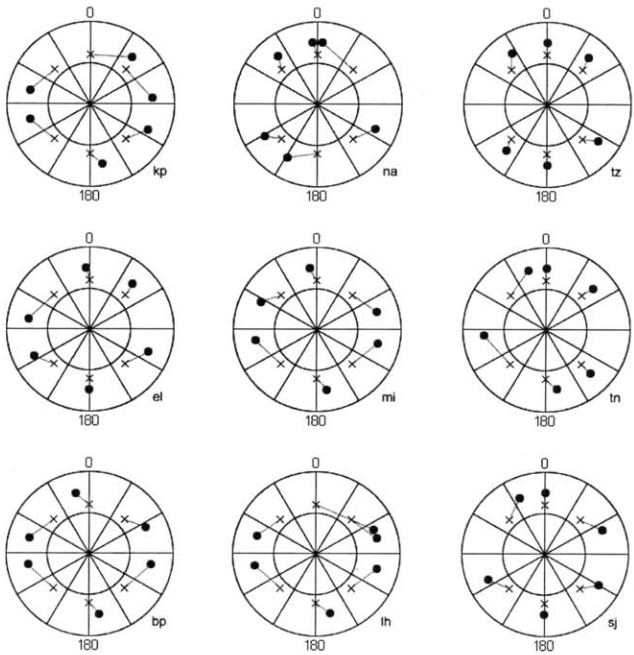


Fig. 14. Azimuth judgments for each subject under head-tracked individualized-HRTF anechoic conditions. Each filled circle ● represents average of five judgments used to correct bias in each subject's data set for ANOVA for localization of auditory images. Original target position (×) is connected by a line. Distance judgments are not shown

regard to both mean values and standard deviations.

These results tentatively suggest that the experimental variables work differently when localization judgments are adjusted for individual biases. A definitive determination could be made in an experiment similar to the current one but using a reference condition involving the localization of actual sound sources. A larger number of judgments for each target azimuth would also be necessary to reduce variability, since the bias adjustments are performed individually for each target azimuth. For example, the notably high offset from the 0° azimuth target for subjects KP and LH seen in Fig. 14 resulted from the fact that the averaged value included positions that were judged near the center of the head at lateral positions (near 90°). A larger number of judgments would minimize the impact of these types of data outliers.

## 6 SUMMARY

A review of the research and product literature related to the perception of headphone-delivered three-dimensional sound suggests that optimal localization performance results from a combination of the following factors: 1) head-tracked virtual stimuli, 2) synthesis of a virtual room, and 3) use of individualized as opposed to generic HRTFs (the primary cue to auditory localization). However, these assumptions were never previously evaluated within a single study that directly compared these factors in all combinations. It was previously not known if all of these factors contributed equally to the accuracy and overall quality of auditory localization in a virtual acoustic display, or if instead these factors contributed only to specific aspects of localization.

An experiment was run using nine subjects to evaluate auditory localization, externalization of sound images, and perceived realism. Speech stimuli were used due to their relevance to three-dimensional audio applications such as teleconferencing and multiple-channel voice communications. Three levels of reverberation were used: anechoic (no reverberation), early reflections (first 80 ms of room reverberation), and full reverberation simulation. Two types of HRTFs were used: individualized HRTFs were measured for each subject, whereas generic HRTFs were from a dummy head. Finally, head tracking was either used to update the position of the stimuli in real time in response to head movement, or disregarded.

For azimuth errors (corrected for reversals) and for elevation errors (deviation from the target at eye level), a significant main effect was found only for reverberation. No significant main effects were found for head tracking or HRTF type. For azimuth errors an interaction was found for the head tracking and HRTF used, but the net effect on the localization error is minimal (about 5°). Head tracking reduced reversals significantly (front–back and back–front confusions of the location of the stimuli across the interaural axis), from a rate of 59% to 28%. Neither reverberation nor HRTF type yielded significant effects on the reversal rates.

Finally a significant main effect of reverberation was found for externalization. A mean value of 79% of the stimuli were heard outside the head under the reverberant conditions, compared to 40% for the anechoic condition. A post-hoc test indicated no significant difference between the early-reflection and full-reverberation conditions, meaning that externalized stimuli can be simulated using a minimal representation of a reverberant acoustic field. No other treatments or interactions yielded significant effects for externalized stimuli.

These results contradict some commonly head assumptions regarding the efficacy of head tracking and individualized HRTFs for virtual acoustic displays, and are therefore applicable to establishing the design criteria used for three-dimensional audio displays of speech for human–machine interfaces (including virtual reality), teleconferencing, multimedia, games, and other applications. It must be emphasized that these data apply only to speech stimuli, and that experimental results may differ when broadband stimuli such as noise or clicks are used. For speech applications, future technology engineering efforts will benefit from the current experiment's identification of perceptually relevant factors in the design of virtual acoustic displays.

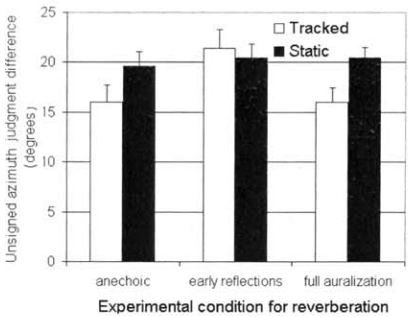


Fig. 15. Unsigned azimuth judgment referenced to head-tracked individualized-HRTF anechoic conditions for each subject—significant interaction for reverberation and HRTF type (mean values and standard error bars).

## 7 ACKNOWLEDGMENT

The authors wish to thank Bryan McClain, Joel Miller, and Kevin Jordan for their assistance in the completion of this work. Special thanks are due to Alexandra Lee for her work in data gathering and analysis, and for her assistance on an earlier version of this paper. This paper was partially supported by NASA–San Jose State University cooperative research grant NCC 2-1095.

## 8 REFERENCES

[1] F. L. Wightman and D. J. Kistler, "Factors Affecting the Relative Salience of Sound Localization Cues," in *Binaural and Spatial Hearing in Real and Virtual Environments*, R. Gilkey and T. Anderson, Eds. (Lawrence Erlbaum, Mahwah, NJ, 1997).

[2] F. L. Wightman and D. J. Kistler, "The Importance

of Head Movements for Localizing Virtual Auditory Display Objects," in *Proc. 1994 Conf. on Auditory Displays* (1995).

[3] E. M. Wenzel and D. R. Begault, "Are Individualized Head-Related Transfer Functions Required for Auditory Information Displays?," in *Proc. 137th Mtg. of the Acoustical Society of America, 2nd Conv. of the European Acoustics Association (Forum Acusticum 99)* and *25th German DAGA Conf.* (1999), p. 105.

[4] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization Using Nonindividualized Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 94, pp. 111–123 (1993).

[5] F. L. Wightman, D. J. Kistler, and M. E. Perkins, "A New Approach to the Study of Human Sound Localization," in *Directional Hearing*, W. Yost and G. Gourevitch, Eds. (Springer, New York, 1987).

[6] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural Technique: Do We Need Individual Recordings?," *J. Audio Eng. Soc.*, vol. 44, pp. 451–469 (1996 June).

[7] H. Lehnert and J. Blauert, "Principals of Binaural Room Simulation," *Appl. Acoust.*, vol. 36, pp. 259–291 (1992).

[8] D. R. Begault, "Perceptual Effects of Synthetic Reverberation on Three-Dimensional Audio Systems," *J. Audio Eng. Soc.*, vol. 40, pp. 895–904 (1992 Nov.).

[9] M. Kleiner, B. I. Dalenbäck, and P. Svensson, "Auralization—An Overview," *J. Audio Eng. Soc.*, vol. 41, pp. 861–875 (1993 Nov.).

[10] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, rev. ed. (MIT Press, Cambridge, MA, 1997).

[11] G. Plenge, "On the Difference between Localization and Lateralization," *J. Acoust. Soc. Am.*, vol. 56, pp. 944–951 (1974).

[12] D. R. Begault, *3-D Sound for Virtual Reality and Multimedia* (Academic Press Professional, Cambridge, MA, 1994).

[13] H. Fisher and S. J. Freedman, "The Role of the Pinnae in Auditory Localization," *J. Audit. Res*, vol. 8, pp. 15–26 (1968).

[14] E. M. Wenzel and S. H. Foster, "Perceptual Consequences of Interpolating Head-Related Transfer Functions during Spatial Synthesis," in *Proc. ASSP (IEEE) Workshop on Applications of Signal Processing to Audio and Acoustics* (1993).

[15] S. Weinrich, "The Problem of Front–Back Localization in Binaural Hearing," in *10th Danavox Symp. on Binaural Effects in Normal and Impaired Hearing*, vol. 11, suppl. 15 (1982).

[16] A. W. Bronkhorst, "Localization of Real and Virtual Sound Sources," *J. Acoust. Soc. Am.*, vol. 98, pp. 2542–2553 (1995).

[17] A. Wallach, "On Sound Localization," *J. Acoust. Soc. Am.*, vol. 10, pp. 270–274 (1939).

[18] H. Wallach, "The Role of Head Movements and Vestibular and Visual Cues in Sound Localization," *J. Exp. Psychol.*, vol. 27, pp. 339–368 (1940).

[19] E. M. Wenzel, "The Relative Contribution of Interaural Time and Magnitude Cues to Dynamic Sound Localization," in *Proc. 1995 ASSP (IEEE) Workshop on Applications of Signal Processing to Audio and Acoustics*, 95TH8144 (1995).

[20] F. L. Wightman and D. J. Kistler, "Resolution of Front–Back Ambiguity in Spatial Hearing by Listener and Source Movement," *J. Acoust. Soc. Am.*, vol. 105, pp. 2841–2853 (1999).

[21] D. R. Begault, "Perceptual Similarity of Measured and Synthetic HRTF Filtered Speech Stimuli," *J. Acoust. Soc. Am.*, vol. 92, p. 2334 (1992).

[22] F. L. Wightman and D. J. Kistler, "Headphone Simulation of Free-Field Listening. II: Psychophysical Validation," *J. Acoust. Soc. Am.*, vol. 85, pp. 868–878 (1989).

[23] N. I. Durlach, A. Rigopulos, X. D. Pang, W. S. Woods, A. Kulkkarni, H. S. Colburn, and E. Wenzel, "On the Externalization of Auditory Images," *Presence: Teleoperators and Virtual Environm.*, vol. 1 (1992).

[24] N. I. Durlach and H. S. Colburn, "Binural Phenomena," in *Handbook of Perception*, vol. 4: *Hearing*, E. C. Carterette and M. P. Friedman, Eds. (Academic Press, New York, 1978).

[25] B. B. Bauer, "Stereophonic Earphones and Binaural Loudspeakers," *J. Audio Eng. Soc.*, vol. 9, pp. 148–151 (1961).

[26] W. Mills, "Auditory Localization," in *Foundations of Modern Auditory Theory*, J. V. Tobias, Ed. (Academic Press, New York, 1972).

[27] D. R. Begault and E. M. Wenzel, "Headphone Localization of Speech," *Human Factors*, vol. 35, pp. 361–376 (1993).

[28] E. M. Wenzel, "The Impact of System Latency on Dynamic Performance in Virtual Acoustic Environments," in *Proc. 15th Int. Congr. on Acoustics* and *135th Mtg. of the Acoustical Society of America* (1998).

[29] B. D. Adelstein, E. R. Johnston, and S. R. Ellis, "Dynamic Response of Electromagnetic Spatial Displacement Trackers," *Presence: Teleoperators and Virtual Environm.*, vol. 5, pp. 302–318 (1996).

[30] Bang and Olufsen, *Music from Archimedes*, Compact Disc B&O 101 (1992).

[31] E. M. Wenzel, "Effect of Increasing System Latency on Localization of Virtual Sources," in *Proc. AES 16th Int. Conf. on Spatial Sound Reproduction* (1999).

[32] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-Related Transfer Functions of Human Subjects," *J. Audio Eng. Soc.*, vol. 43, pp. 300–321 (1995 May).

[33] A. H. Marshall and J. Meyer, "Directivity and Auditory Impression of Singers," *Acustica*, vol. 58, p. 130 (1985).

[34] B. I. Dalenbäck, "Verification of Prediction Based on Randomized Tail-Corrected Cone-Tracing and Array Modeling," in *Proc. 137th Mtg. of the Acoustical Society of America, 2nd Conv. of the European Acoustics Association (Forum Acusticum 99)*, and *25th German DAGA Conf.* (1999).

[35] B. I. Dalenbäck, "A New Model for Room Acoustic Prediction and Absorption," dissertation,

Chalmers University of Technology, Gothenburg, Sweden (1995).

[36] D. R. Begault and J. S. Abel, "Studying Room Acoustics Using a Monopole–Dipole Microphone Array," in *Proc. 16th Int. Congr. on Acoustics* and *135th Mtg. of the Acoustical Society of America* (Seattle, WA, 1998).

[37] F. E. Toole, "In-Head Localization of Acoustic Images," *J. Acoust. Soc. Am.*, vol. 48, pp. 943–949 (1970).

[38] E. M. Wenzel, "What Perception Implies about Implementation of Interactive Virtual Acoustic Environments," presented at the 101st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 44, p. 1165 (1996 Dec.), preprint 4353.

[39] J. M. Loomis, C. Hebert, and J. G. Cincinelli, "Active Localization of Virtual Sounds," *J. Acoust. Soc. Am.*, vol. 88, pp. 757–764 (1990).

[40] P. Mackensen, M. Fruhmann, M. Thanner, G. Theile, U. Horbach, and A. Karamustafaoglu, "Head-Tracker-Based Auralization Systems: Additional Consideration of Vertical Head Movements," presented at the 108th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 48, p. 359 (2000 Apr.), preprint 5135.

[41] W. L. Martens, "Uses and Misuses of Psychophysical Methods in the Evaluation of Spatial Sound Reproduction," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 49, p. 546 (2001 June), preprint 5403.

## THE AUTHORS

D. R. Begault

E. M. Wenzel

M. R. Anderson

Durand R. Begault received the Ph.D. degree in 1987 from the University of California San Diego, where he worked at the Computer Audio Research Laboratory. He has been associated with the Spatial Auditory Display Laboratory at NASA Ames Research Center since 1988. He authored the book *3-D Sound for Virtual Reality and Multimedia* in 1984, has published many articles on three-dimensional audio research and development, and holds two U.S. patents. Dr. Begault is vice-chair of the Audio Engineering Society Technical Committee on Perception and Subjective Evaluation of Audio Signals, and a member of the *Journal's* review board.

●

Elizabeth M. (Beth) Wenzel received the Ph.D. degree in cognitive psychology with an emphasis in psychoacoustics from the University of California, Berkeley, in 1984. From 1985 to 1986 she was a National Research Council postdoctoral research associate at NASA-Ames Research Center working on the auditory display of information for aviation systems. Since 1986 she has been director of the Spatial Auditory Displays Laboratory in the Human Factors Research and Technology Division at NASA-Ames, directing development of real-time display technology and conducting basic and applied research in auditory perception and localization in three-dimensional virtual acoustic displays. Dr. Wenzel is a senior editor of the journal *Presence* and has published a number of articles and spoken at many conferences on the topic of virtual acoustic environments.

●

Mark R. Anderson studied aeronautical engineering at the California Polytechnic University, San Luis Obispo, and at Northrop University, Inglewood, California, and holds a B.S. in computer science from San Jose State University. Mr. Anderson has been a senior programmer at the Spatial Auditory Display Laboratory at NASA-Ames Research Center since 1996, under contract from Raytheon Technical Services Company.