Challenges to the Successful Implementation of 3-D Sound*

DURAND R. BEGAULT, *AES Member*

Aerospace Human Factors Division, NASA-Ames Research Center, Moffett Field, CA 94035, USA

The major challenges for the successful implementation of 3-D audio systems involve minimizing reversals, intracranially heard sound, and localization error for listeners. Designers of 3-D audio systems are faced with additional challenges in data reduction and low-frequency response characteristics. The relationship of the head-related transfer function (HRTF) to these challenges is shown, along with some preliminary psychoacoustic results gathered at NASA–Ames.

0 INTRODUCTION

"3-D audio technology" is a generic term for the promised result of a host of new systems that have only recently made the transition from the laboratory to the commercial audio world. A number of other terms have been used both commercially and technically to describe the technique (such as dummy-head synthesis and spatial sound processing), but all are related in their promise of a "psychoacoustically enhanced" auditory display. Much in the same way that stereophonic and quadrophonic signal-processing devices were introduced as improvements over their predecessors, 3-D audio technology could be considered as the latest innovation for both mixing consoles and reverberation devices.

The heart of 3-D audio technology involves digital filtering according to the head-related transfer function (HRTF). The spectral modification imposed by the HRTF on an incoming sound has been established in the psychoacoustic literature as an important cue for auditory spatial perception, along with interaural level and amplitude differences [1]–[4]. The HRTF imposes a unique frequency response for a given sound-source position outside of the head, which can be measured by recording the impulse response in or at the ear canal and then examining its frequency response using fast Fourier Transform techniques. The binaural impulse response can also be directly implemented into a pair of digital filters for use in a 3-D audio system, using convolution techniques [5]–[7].

Fig. 1 shows the difference between HRTF measurements at the left ear for two different persons, for a source at 0° azimuth, 0° elevation. (In this engineering report azimuth increases to the left or right on the horizontal plane, with 0° directly in front of the listener and 180° directly behind the listener; for elevation, 0° is at ear level, increasing upward to +90° and decreasing downward to −90°). Note that the HRTF indicated by the solid line contains a 10-dB peak between 8 and 9 kHz and an even more pronounced notch around 10.5 kHz that is not present in the other person's HRTF.

In the commercial world it is often assumed that HRTF processing of audio is tantamount to having control over a listener's perception of sound in three-

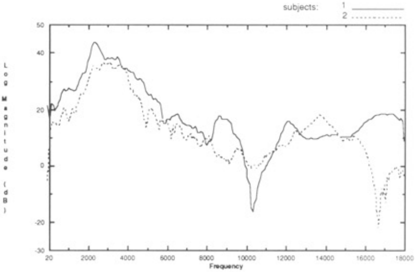


Fig. 1. HRTF spectra for two different persons: left ear, source at 0° azimuth, 0° elevation. Although the overall shapes of the spectra are similar between 20 Hz and 8 kHz, note that the HRTF for subject 1 contains peaks and troughs between 8 and 12 kHz that are not found in the HRTF for subject 2.

* Presented at the 89th Convention of the Audio Engineering Society, Los Angeles, CA, 1990 September 21–25; revised 1991 February 5 and August 23.

dimensional space. But the veracity of this assumption is offset by the fact that there is often a mismatch between operator specification and listener perception. Another problem lies in the fact that HRTF measurements are both difficult to make accurately and costly in terms of memory requirements. Three major challenges must be overcome in order to surmount these problems:

1) Eliminating front–back reversals and intracranially heard sound, and minimizing localization error

2) Reducing the amount of data necessary to represent the most perceptually salient features of HRTF measurements

3) Resolving conflicts between desired frequency and phase response characteristics and measured HRTFs.

In this study these challenges are illuminated along with some preliminary psychoacoustic results obtained at the Auditory Lab at NASA–Ames Research Center.

1 POSSIBLE CAUSES OF MISMATCH BETWEEN SPECIFICATION AND PERCEPTION

Fig. 2 shows a source–receiver communication-chain model to illustrate potential sources of localization error that can occur when using a 3-D audio system. (The description is restricted to headphone playback for simplicity.) The recording engineer, the 3-D audio system, and the playback chain through the headphones used by the listener can be thought of collectively as the "source" for a sound placed in 3-D space. The "receiver" refers collectively to a listener's perceptual and cognitive abilities to identify the location of an actual or virtual sound source under various conditions. "Error" refers to the difference between a recording engineer's specification for a sound to be heard at position X and the listener's perception of it at position Y.

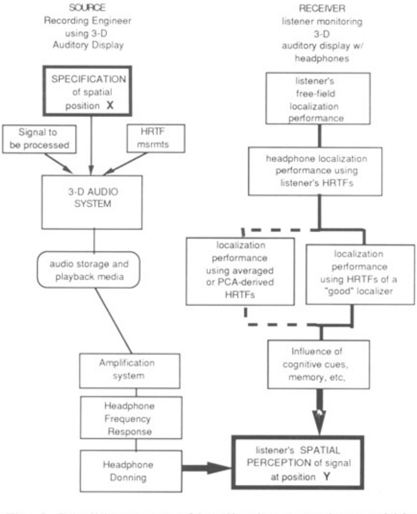

SOURCE
Recording Engineer using 3-D Auditory Display
RECEIVER
listener monitoring 3-D auditory display w/ headphones
SPECIFICATION of spatial position X
Signal to be processed
HRTF msrmts
3-D AUDIO SYSTEM
audio storage and playback media
Amplification system
Headphone Frequency Response
Headphone Donning
listener's free-field localization performance
headphone localization performance using listener's HRTFs
localization performance using averaged or PCA-derived HRTFs
localization performance using HRTFs of a "good" localizer
Influence of cognitive cues, memory, etc.
listener's SPATIAL PERCEPTION of signal at position Y

Fig. 2. Possible sources of localization error shown within source–receiver model.

There are three input components that are key to the overall success of any 3-D audio system. First, and perhaps most important, is the particular set of HRTFs that are used in the system. Practical experience has shown that some HRTF sets simply "work" better than others for particular individuals, especially in terms of externalization and mitigation of reversals. Different HRTF sets can also have large overall timbral differences, depending on the particular person measured and the measurement technique itself. Second, the differential character of various sounds that are input into a 3-D audio system will affect performance. For example, broad-band, impulsive sounds will be easier to localize to a specific position than low-frequency sounds with slow amplitude envelopes. Finally, the spatial resolution demanded by the recording engineer will determine the criterion for evaluating the overall performance of a system. This manifests itself in terms of the user's specification; for example, specifying that a listener should hear an externalized source behind him or her toward the left is easier to attain than a specification that a sound be heard at left 140° azimuth, up 20° elevation, and at 3 m distance.

The nonlinearities in amplification, headphone frequency response, and donning of headphones by the listener are additional sources of error in any audio reproduction system. Interchannel frequency and phase response differences can be particularly problematic for conveying 3-D audio imagery, since a binaural HRTF contains perceptually significant interaural intensity and time differences.

Even if the error due to the source has been minimized, error will still be present due to the varied localization accuracy of a listener. A listener's localization accuracy under free-field conditions will contain error. The listener's headphone localization accuracy is usually worse, and will vary according to the particular HRTFs used. Studies by Wightman and Kistler have shown that a listener's headphone localization performance using the listener's own HRTFs can come close to this listener's performance under free-field conditions [7]. Unfortunately it is not practical for each user to install his or her own HRTF measurements into a 3-D audio system. Hence, a goal of many researchers is to design a general set of HRTFs for the overall population. Since most research has shown that performance is somewhat poorer when listening to sound processed with HRTFs other than one's own [8], [9], this remains a formidable goal.

Two approaches can be used in developing a set of general HRTFs. One involves synthesizing HRTFs via averaging, structural modeling, or through principal-component analysis. The other approach is to use the actual HRTFs of a "good localizer": a subject whose free-field localization performance is better than average, and whose headphone localization performance closely matches their subject's free-field localization performance. One such set of measurements (from subject SDO in [10]) has been used as the non-listener-specific HRTFs in several studies measuring localization error of untrained subjects under various conditions [8], [9], [11].

The final source of receiver error illustrated in Fig. 2 is the influence of cognitive and visual cues. Auditory localization judgments are highly malleable as a function of expectation or memory. For instance, how often do you think that it must be your telephone that's ringing in the office? Visually acquired stimuli can also modify auditory localization, as in the cinema. Cognitive cues have been used to advantage by manufacturers of 3-D audio systems in their demonstration tapes. For example, one particular company argued that its system avoided the commonly reported problem of front-to-back reversals (e.g., hearing a 0° sound source at 180°). When, on the company's demonstration tape, a person lights a cigarette and drinks a glass of water, it is probably difficult to imagine the virtual source to the rear simply because we know our mouth is positioned at the front of the head.

2 MEASUREMENTS OF LOCALIZATION ERROR: SOME PRELIMINARY RESULTS FOR SPEECH STIMULI

Inaccuracies in localization judgments can be attributed to two sources of perceptual error. The first source, localization error, refers to the variability between a target position and its judged location, measured in degrees azimuth or elevation. The second source is reversal error, a situation where a stimulus at a given position is heard at its "mirror" position; for instance, making a judgment of right 30° azimuth for a right 150° target. Reversal errors are ascribed to static sound sources placed at target positions where overall interaural time and level difference cues are constant. In theory, this problem is diminished when the head or sound source is allowed to move [12]. Fig. 3 shows how the spectral cue provided by the HRTF differentiates front and back mirror positions in a static source–listener context.

Several investigations into subjective localization error using a 3-D auditory display with non-listener-specific HRTFs have been conducted or are currently under way at NASA–Ames Research Center. Some of the main results of a study are given in this section, where azimuth, elevation, and distance judgments were gathered from 11 subjects, focusing on judgments for 0 and 180° azimuth targets [8]. Speech stimuli were used because of their applicability to workstations, teleconferencing, and communications systems in general.

Stimuli were generated by digitally filtering a set of 45 one- or two-syllable words, each representing a particular international phonetic alphabet phoneme, with HRTFs at target positions of 0 and 180°, and left and right 30, 60, 90, 120, and 150° azimuth, all at 0° 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 and Kistler [10] and also used in [7]–[9], [11]. The spectrum of the headphones used (Sennheiser HD-430) was divided out of the HRTFs.

During a trial, subjects heard a given speech segment repeated 5 times and then called out estimates of the azimuth (0 to 180° left or right; 0° in front), elevation (0 to 90° up or down; 0° at ear level), and distance (0 in at the center, 4 in at the edge of the head), which were recorded by the experimenter. Over the course of 2 to 3 days, each subject listened to 15 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. For distance, subjects were instructed to use 0 in if the sound was directly in the middle of their head, >0 and <4 in for anywhere inside the head, 4 in for a sound at the edge of the head, and >4 in for externalized sounds.

Figs. 4–6 summarize the mean azimuth error resulting from both localization error and reversals for all subjects. The data are presented here in terms of tolerances: the percentage of judgments for the given position within 0–10°, 11–30°, and >30°. The percentages for 0 and 180° azimuth targets are based on 150 judgments from each subject; the other azimuths were based on 15 judgments from each subject.

In general, the error was highest with the front azimuth positions (0°, left and right 30°) and lowest at the sides (left and right 90°). It must be noted that the intersubject variability in estimating azimuth was quite high; for instance, the standard deviations for the 0–10° tolerances were 35 and 38% at 0 and 180°. Fig. 7 gives an example; it shows polar plots of the azimuth and distance judgments of two different subjects for

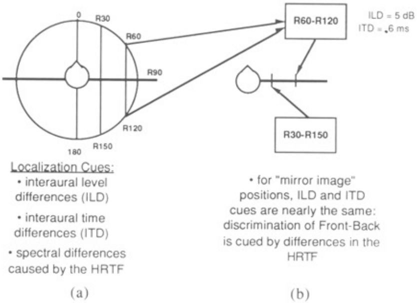

0
R30
R60
R90
R120
R150
180
R60-R120
ILD = 5 dB
ITD = .6 ms
R30-R150
Localization Cues:
• interaural level differences (ILD)
• interaural time differences (ITD)
• spectral differences caused by the HRTF
• for "mirror image" positions, ILD and ITD cues are nearly the same: discrimination of Front-Back is cued by differences in the HRTF
(a)
(b)

Fig. 3. Depiction of role of HRTF as a cue for mirror-image positions with similar interaural time and level differences. (a) Overhead view of listener with positions "specified" by user of 3-D audio system. (b) Possible perceptual result involving reversal errors.

the 0° azimuth HRTF stimuli. The judgments shown in Fig. 7(a) are for a subject who heard everything to the rear; Fig. 7(b) shows a person with highly variable judgments.

How much of the error shown in Figs. 4–6 can be ascribed to reversals? The results echo the problem commonly reported with dummy-head recordings: about 30% of all judgments were reversed, with the ratio of front–back versus back–front reversals about 4:1. Hence a large proportion of the judgments in the >30° tolerance categories shown in Figs. 4–6 is a result of reversals.

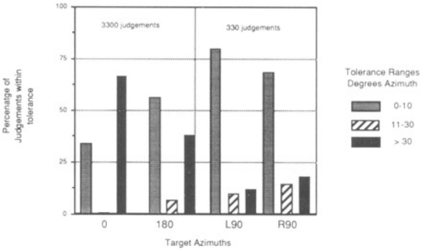


Fig. 4. Errors in headphone localization for 0 and 180°, left and right 90° target azimuths at 0° elevation. Speech stimuli processed with non-listener-specific HRTFs; inexperienced subjects.

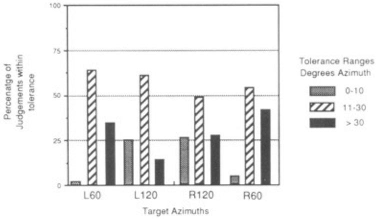


Fig. 5. Errors in headphone localization for left and right 60 and 120° target azimuths at 0° elevation. Speech processed with non-listener-specific HRTFs; inexperienced subjects.

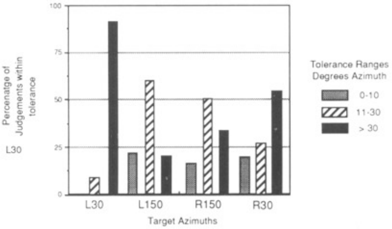
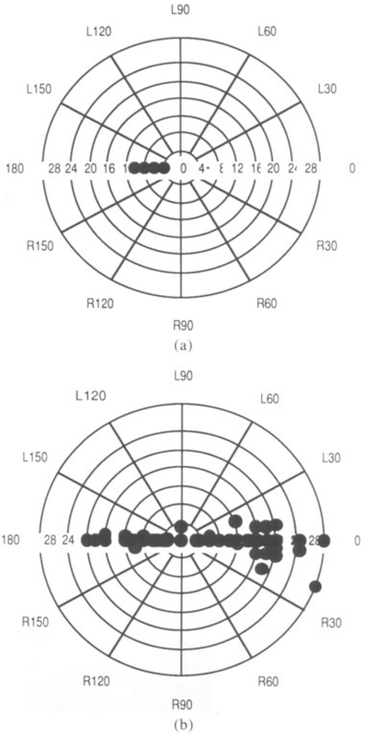


Fig. 6. Errors in headphone localization for left and right 30 and 150° target azimuths at 0° elevation. Speech processed with non-listener-specific HRTFs; inexperienced subjects.

Elevation and distance judgments are similar for both 0 and 180° HRTF filtered speech. Fig. 8 shows means and standard deviations for nine of the eleven subjects, collectively and individually. The means for elevation judgments, both across subjects and within each subject, were above ear level (>0°) for both 0 and 180° azimuth, with the exception of one subject (s9), whose mean was at ear level for the 180° azimuth target. The mean elevation for all subjects is slightly higher for 0° than for 180° (51.4 versus 45.1), with large standard deviations in both cases. An overall tendency for subjects to elevate their elevation judgments above eye level was observed for all azimuth targets, paralleling the results found in another speech study by the author [11].

Although the mean value for distance judgments for all subjects was externalized (that is, greater than 4 in) for both 0 and 180° positions, the standard deviations and individual means indicate that all subjects heard a



(a)

(b)

Fig. 7. Azimuth and distance judgments for 2 subjects: 0° azimuth, 0° elevation HRTF-processed speech. Overhead polar plot. Each circle is 4 in; points within center ring represent unexternalized judgments.

proportion of the sounds intracranially. In particular, subject s4 heard all sounds inside the head. The means and the standard deviations of the percentage of all judgments heard intracranially were very similar for the two positions: mean 38%, standard deviation 31% and mean 36%, standard deviation 31% for 0 and 180°, respectively. A possible reason for this effect is the absence of environmental cues such as reverberation in the HRTF measurements. A recently finished study with five subjects demonstrated that adding HRTF-processed room reflection information to the anechoic HRTFs minimizes intracranially heard sound [11].

3 LOW-FREQUENCY RESPONSE AND DATA REDUCTION OF HRTF MEASUREMENTS

HRTF measurements have been discussed as an important aspect for evaluating the performance of 3-D audio systems from a localization standpoint. In an applications context, there are additional challenges posed by the HRTF measurements in terms of their impulse response duration and low-frequency response.

Measurement of the HRTF at or in the ear canal requires the use of miniature probe microphones and other apparatus that collectively pose a great challenge for obtaining linear frequency responses and low noise. Although the system used for the HRTF measurements of subject SDO was in many respects state of the art, it did not allow for the measurement of frequencies below 200 Hz [10]. Also, the impulses were transduced over inexpensive miniature loudspeakers (Realistic Minimus-7) so that a point source could be approximated; however, these loudspeakers are inefficient in their overall low-frequency response. The relatively diminished low-frequency response in the magnitude of these measured HRTFs becomes particularly evident when filtering music or other types of audio.

Fig. 9 shows an example of this problem with a closer look at the same HRTF response shown by the solid line in Fig. 1 (0° azimuth and elevation). The peak amplitude of the filter occurs at approximately 2350 Hz. Between 2000 and 4000 Hz the level is on average approximately 6 dB lower than this. But between 20 and 500 Hz, the amplitude is on average approximately 25 dB lower than at the peak value, and 19 dB lower than the average value between 2000 and 4000 Hz. Processing a chromatic scale played across the entire keyboard of a piano with this filter in a 3-D audio system would probably have unacceptable results in terms of realistic dynamics.

The durations of most HRTF impulse responses translate into FIR filter coefficient arrays that are expensive to implement on currently available hardware for real-time applications. The HRTFs used in the experiments described had an impulse response length of 512 coefficients, or 0.01024 s at a 50-kHz sampling rate. For real-time applications it would be desirable to reduce the number of coefficients to accommodate inexpensive DSP chips that accommodate FIR filtering.

Several approaches can be taken to reduce the number of coefficients of a given HRTF measurement to obtain a desired magnitude response, including windowing techniques. Windowing necessarily reduces the amount of detail present in the original HRTF; it remains an open question as to how much windowing is psychoacoustically transparent. In my own work I have used

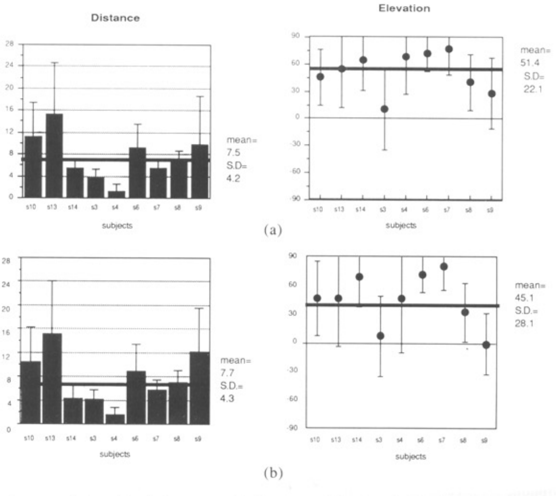
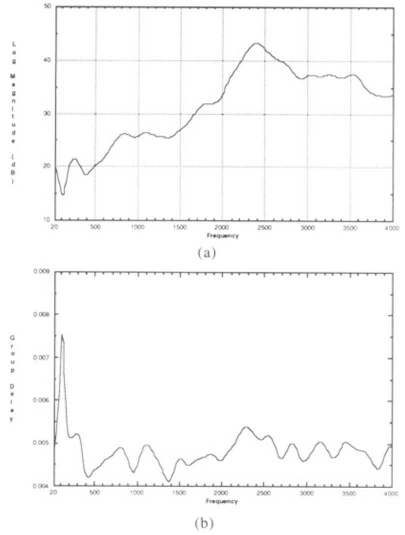


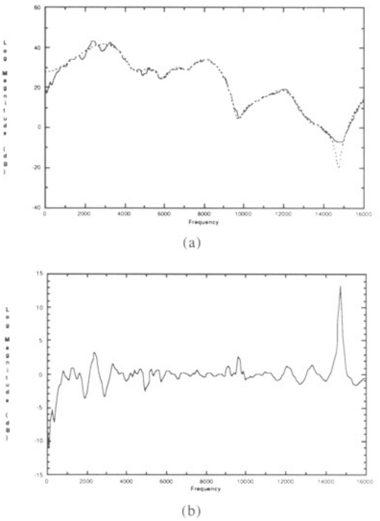
Fig. 8. Summary of distance and elevation judgments. (a) Targets at 0° azimuth, 0° elevation. (b) Targets at 180° azimuth, 0° elevation.



(a)

(b)

Fig. 9. (a) 20-Hz to 4-kHz frequency response and (b) group delay of HRTF shown in Fig. 1 (solid line). Group-delay vertical axis indicates seconds. Maximum peak amplitude of response from 20 Hz to 16 kHz is at approximately 2350 Hz; average amplitude between 2 and 4 kHz is approximately 6 dB down from this peak; average amplitude between 20 and 500 Hz is approximately 25 dB down.



(a)

(b)

Fig. 10. (a) Original 512-coefficient HRTF for 90° azimuth, 0° elevation (solid line) and filter-design approximation (dashed line). (b) Magnitude difference between both filters.

Parks–McClellan and other filter design algorithms [13] to create approximations of HRTF magnitude responses published by both Blauert, and Wightman and Kistler [1], [10]. The group delay is implemented with some filter designs as a single value across all frequencies, and in others as an approximation of the original group-delay response. The low-frequency problem mentioned can be improved by changing the measured decibel response in the initial analysis and "directly" specifying a desired response.

Fig. 10(a) shows the decibel response of a measured HRTF with 512 coefficients and of a filter with 80 coefficients designed to approximate it; the difference between the two responses is shown in Fig. 10(b). The questions of how important magnitude- and frequency-dependent group-delay approximations are for creating an effective 3-D auditory display is a question currently under study here at NASA–Ames. Ultimately, psychoacoustic investigations will need to be conducted that compare real and synthetic HRTFs to determine their effect on both localization and timbre.

4 SUMMARY

Through an interdisciplinary approach to the design and evaluation of 3-D audio systems it may be possible in the near future to confront the challenges described in this study successfully. The need for additional work in the psychoacoustic domain is apparent; development of products always outpaces research. Already, it is possible to perform creative audio processing with HRTF filtering techniques that is not possible with other technologies. Ultimately it will be the specificity of the demands of recording engineers and listeners that will determine the rigor and quality involved in the design of 3-D audio systems.

5 REFERENCES

[1] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (M.I.T. Press, Cambridge, MA, 1983).

[2] L. A. Abbagnaro, B. B. Bauer, and E. L. Torick, "Measurement of Diffraction and Interaural Delay of a Progressive Sound Wave Caused by the Human Head," *J. Acoust. Soc. Am.*, vol. 58, p. 693 (1975).

[3] D. H. Cooper, "Calculator Program for Head-Related Transfer Function," *J. Audio Eng. Soc.*, vol. 30, pp. 34–38 (1982 Jan./Feb.).

[4] D. H. Cooper, "Phase Reference in HRTF Calculation," *J. Audio Eng. Soc.* (*Letters to the Editor*), vol. 31, p. 760 (1983 Oct.).

[5] D. R. Begault, "Control of Auditory Distance," dissertation, University of California, San Diego (1987).

[6] E. M. Wenzel, F. L. Wightman, and S. H. Foster, "A Virtual Display System for Conveying Three-Dimensional Acoustic Information," in *Proc. Human*

Factors Society 32nd Annual Meeting (1988), pp. 86–90.

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

[8] D. R. Begault and E. M. Wenzel, "Headphone Localization of Speech Stimuli," in *Proc. Human Factors Society 35th Annual Meeting* (1991), pp. 82–86.

[9] E. M. Wenzel, F. L. Wightman, and D. J. Kistler, "Localization of Nonindividualized Virtual Acoustic Display Cues," *Proc. CHI 91 ACM Conf. on Computer–Human Interaction* (1991), pp. 351–359.

[10] F. L. Wightman and D. J. Kistler, "Headphone Simulation of Free-Field Listening. I: Stimulus Synthesis," *J. Acoust. Soc. Am.*, vol. 85, pp. 858–867 (1989).

[11] D. R. Begault, "Perceptual Effects of Synthetic Reverberation on 3-D Audio Systems," presented at the 91st Convention of the Audio Engineering Society, New York, 1991 Oct. 4–8, preprint 3212.

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

[13] J. H. McClellan, T. W. Parks, and L. R. Rabiner, "FIR Linear Phase Filter Design Program," in *Programs for Digital Signal Processing* (IEEE Press, New York, 1979).

THE AUTHOR

Durand R. Begault received the Ph.D. degree in computer music from the University of California, San Diego, in 1987, where he studied with F. Richard Moore. He is a visiting researcher with the Aerospace Human Factors Research Division of NASA–Ames Research Center under the auspices of the National Research Council. His work has included basic research into sound localization and applications research involving advanced audio displays in commercial airline cockpits.