VIRTUAL ACOUSTICS: EVALUATION OF PSYCHOACOUSTIC PARAMETERS FOR AURALIZATION

Durand R. Begault, Ph.D.
San José State University
Flight Management and Human Factors Research Division
NASA Ames Research Center
Mail Stop 262-2
Moffett Field, California 94035-1000
Tel: (415) 604 3920 Fax: (415) 604 3323
Email: db@eos.arc.nasa.gov

ABSTRACT

Many of the current virtual acoustic displays for teleconferencing and virtual reality are limited to a very simple or non-existent rendering of reverberation–a fundamental part of the acoustic environmental context that is encountered in day-to-day hearing. Research shows that environmental cues dramatically improve perceptual performance within virtual acoustic displays, and that is possible to manipulate signal processing parameters to effectively reproduce aspects of virtual acoustic rooms in response to head movement. However, the computational resources for rendering a complete diffuse sound field in real time remain formidable. Our efforts at NASA Ames have been focused on several perceptually-based software/hardware strategies that include determination of what aspects of a calculated diffuse field are below threshold.

INTRODUCTION

An important milestone for the future of 3-D audio systems and for the rendering of virtual acoustic spaces was the recognition of what is now called *auralization* (Kleiner, Svensson & Dalenbäck, 1990). Auralization has been defined as "the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in a modeled space" (Kleiner, Dalenbäck & Svensson, 1993). The binaural listening experience is based on the application of Head Related Transfer Functions (HRTFs) to not only the direct sound ("3-D sound"), but to reflected sound as well, a technique pioneered in the early 1980s by the group at Northwestern Computer Music (Kendall & Martens, 1984).

The development of auralization systems has primarily been intended for the acoustical consultant, who, in combination with acoustical prediction software, can allow themselves or their clients to listen to the effect of an acoustical treatment to a building before it is built. This application has had a long history (see, e.g., Horrall, 1970; Kuttruff, 1993). Another application is the creation of a virtual acoustic sound field that can be tied in with direct sound synthesis for applications such as virtual reality (e.g., Takala et al., 1996). Yet another perhaps more futuristic application is the 'virtual psychoacoustic laboratory' for reverberation studies. A multiple loudspeaker system within an anechoic chamber could potentially be replaced by a virtual acoustic system, given the assurance that the signals produced were acoustically equivalent.

Unlike the virtual psychoacoustic laboratory just mentioned, virtual reality and acoustical consulting applications can tolerate greater differences between measured and synthesized signals. What matters is perceptual, not acoustical, equivalence between the signals. But the problem common to all of these applications is that the multiple source generation necessary for simulating a diffuse sound field can easily overwhelm computational resources. For instance, a simple image model of a rectangular room with early reflections up to fifth order results in over 200 valid sources (Jan & Flanagan, 1995) (with very unrealistic sounding reverberation!). Using the computational savings of overlap-add FFT techniques, 2 channels of a 3 sec impulse response still require around 22 million floating point operations (Lehnert & Blauert, 1992).

The problem becomes more significant when head-tracked 3-D audio systems as described in Wenzel (1992; 1996) are considered. Head tracking allows one to navigate through a space, turn the head towards and way from a sound source, and

enhances the sensation of immersion in the virtual environment (Foster, Wenzel & Taylor, 1991). Unfortunately, the computation problem for virtual diffuse fields then becomes worse. Four times as many coefficients as used in a static simulator are necessary to enable real-time interpolation.

With both head-tracking and rendering of diffuse sound fields, the perceptual performance in all types of 3-D sound applications is greatly improved (Begault, 1994). The overall goal is therefore accurate representation of the acoustical environment while reducing computation necessary to bring about such an effect. The following is a review of efforts made for reducing the computational complexity necessary for auralization, including ongoing work at NASA Ames.

COMPUTATION REDUCTION

Hardware and Software

Significant progress has been made during the last decade for decreasing the necessary processing time for auralization, thanks to new generations of software and hardware. For example, software acoustic prediction schemes have been proposed and implemented that simplify ray tracing into octave bands, and then use time integration over a constant such as 1 msec for late reflections (Dalenbäck, 1995; Kuttruff, 1993). In terms of hardware, large-scale convolution devices, such as the *Huron* and *HeadScape* systems manufactured by Lake DSP, allow head-tracked early reflections out to around 4000 taps (180 msec at 22.05 kHz sample rate) and overall reverberation out to 262144 taps, although the late reverberation must still be in essence "fixed" in time (see Reilly & McGrath, 1995; Reilly, McGrath & Dalenbäck, 1995). The ability to update finite impulse response (FIR) filters with low time latency is a feature of several newer approaches (Gardner, 1994; Single & McGrath, 1989).

Perceptual modeling

A different approach is to drive hardware requirements according to perceptual parameters. This strategy is seen in IRCAM's *Le Spatialisateur*, where perceptual parameters can be adjusted directly (Jot, 1996; Jot, Larcher & Warusfel, 1995). The system design is driven by the results of a considerable body of perceptual research into large room acoustics (concert halls, auditoria, etc.) (Jullien, et al., 1992). The DSP approach involves multichannel feedback delay networks, as opposed to an FIR representation of the impulse response. This results in a considerable savings in computation time, but does not allow direct implementation of a measured impulse response or one resulting from CAD-based acoustic prediction software.

An alternative perceptual approach involves starting with a physical acoustic prediction, but then restructuring either the available resources for computational rendering or the impulse response itself.

Lehnert and Blauert point out that "The perception of the auditory environment, like most other perception processes, includes a tremendous amount of information reduction from the physical to the perceptual domain, which is not well understood yet" (Lehnert & Blauert, 1992). They proposed a solution to computational complexity whereby a software "priority manager" is used to allocate DSP resources according to the most perceptually-relevant factor at a given time. For instance, head movement with a static source would require HRTF updates, but source movement would require additional real-time processing of reflection absorption coefficients.

The work at Ames has been focused on a similar strategy. We have proposed a pre-processing of reflections based on a data base of acoustic threshold data (Begault, 1992a). A sufficiently complete perceptual data base could be used to simplify the results of a room prediction model. In many cases, either the reflection itself can be eliminated, or the HRTF used can be simplified for computational efficiency. Another function might be to exaggerate spatial cues to take advantage of temporal discrimination thresholds, although we have yet to explore this method formally. Overall, the concept is to "prune" the reflectogram of information that otherwise may be perceptually irrelevant.

Matching spatial resolution to performance

Most research indicates that localization performance is best when using individualized HRTFs within a 3-D audio system. For a long time, the process for obtaining a set of HRTF measurements was cumbersome, requiring a specialized laboratory, anechoic chamber and in some

cases ear molds for measurements made near the ear canal (e.g., Wightman & Kistler, 1989).

Recently, we have been using a blocked meatus technique similar to that used at Aalborg University (Hammershøi, et al. 1992; Møller, 1992). Specifically, this is the *Snapshot* system developed by Jonathan Abel and Crystal River Engineering, which enables measurements to be made in a reflective room. The resulting HRTFs are minimum-phase, using a constant interaural delay across frequency. Several researchers have found that minimum-phase models of HRTFs are perceptually valid (Kistler & Wightman, 1992; Kulkarni, Isabelle & Colburn, 1995). The ability to separate HRTF magnitude and time delay processing greatly enhances the computational performance of an auralization system.

Although individualized HRTFs offer superior performance, the use of non-individualized HRTFs will be a necessary requirement for most systems. For instance, the distribution of a virtual acoustic environment over the internet cannot predict the outer ear characteristics or head size of the individual user. Once the requirement for individualized HRTFs is relaxed, greater simplification of the magnitude spectra is possible. Our data suggest that, at least for speech stimuli, any negative performance effects caused by approximation of the HRTF via a constant delay or simplified magnitude function is overwhelmed by perceptual inaccuracy.

Figure 1 shows two examples from a set of 13 subjects (Begault, 1992c). The speech stimuli used in this experiment were processed with 0 degree elevation HRTFs at 30 degree increments corresponding to clock positions about the head. The immediately striking aspect for these two particular individuals is the difference in distance and elevation judgments. Most all of subject A's judgments are elevated and heard inside or at the edge of the head. Subject B's judgments are mostly externalized and heard above or below the target elevation with a blur of ±45°. Figure 1 also indicates that while very few back-to-front reversals occurred, around 70% of frontal judgments were heard to the rear. Other data indicate that these reversals dissipate to much lower levels when head-motion cues are available (Wenzel, 1996).

Across all 13 subjects, the data show little difference in performance using actual, compared to synthetic, HRTFs; the average error for azimuth judgments (across azimuth only) is about 7°. For similar stimuli that has been reverberated, azimuth resolution is coarser, reaching about 15 degrees on average (Begault, 1992b). This could probably be due to the fact that the extent of the auditory image is increased by the presence of decorrelated reflected energy. On the other hand, reverberation vastly improves the externalization of virtual sound sources; in one study using similar stimuli, the rate of unexternalized stimuli dropped from 25% to 3% (Begault, 1992b). The azimuth resolution for reflections themselves is coarser still, due to summing localization.

An approach proposed by several authors for simplifying the spatialization of late reflections is to approximate the directions of arrival to 5 or 6 directions around the head (Dalenbäck, 1995; Gierlich, 1992; Wagener, 1971). The spectral features of the HRTF can also be simplified or possibly eliminated for late reverberation (> 100 msec), since it contains little energy at higher frequencies.

Taken together, these data suggest that the spatial resolution of an auralization system can be somewhat coarse. At the same time, a complimentary situation exists: head-tracking is important for minimizing reversals, while the presence of a diffuse field insures externalization. Interpolation for a head-tracked system might also be possible on a relatively coarser grid of stored HRTF measurements than is used for the direct sound.

Threshold studies

The absolute threshold for an early reflection is a function of angle of incidence, time of arrival and sound source type (Bech, 1995a, 1995b; Ebata, Sone & Nimura, 1968; Olive & Toole, 1989; Zurek, 1979). For patterns of reflections, forward and backward masking are also relevant (see Blauert, 1983). Kuttruff determined the echo threshold level for speech with a level of 70 dB, with frontal incidence for both direct and the reflected sound, as

$$(1)\ \Delta L \approx -0.6t - 8 \text{ dB}$$

where ΔL is the level of the reflected sound, and t is the time delay in msec (Kuttruff, 1991). Absolute thresholds for a single reflection as a function of lateral angles between 30° and 65° were summarized by Olive and Toole for a variety of test signals. For time delays of

2—25 msec, the absolute threshold is between –25 and –15 dB, with outliers at –10 dB (musical stimuli) and –45 dB (pulsed click stimuli).

Bech examined early reflection thresholds in the context of a loudspeaker-based simulation system, using 17 early reflections and a simulated diffuse field; a criteria of including only those reflections with intensity greater than –20 dB was used (Bech, 1995a, 1995b). The conclusion was that sensitivity may be increased anywhere from 2 to 5 dB when comparing early reflections without the presence of a dense reflection field ("late" reverberation) (Bech, 1995b).

Experimental Set-up

Our own work in reflection thresholds has used the virtual acoustic system described in Figure 2. Stimuli consist of 3-4 sec of spatially-processed speech from a single, randomly chosen anechoic speech sound file (EBU, 1988).

For simulations of a complete diffuse field with early and late reflections, we have used a commercially available system (CATT Acoustic) to generate binaural impulse responses based on a spherical model (Dalenbäck, 1995). The resulting impulse response is convolved with test material and then stored in a high-quality stereo sampler. The direct sound and additional early reflections, including those experimentally varied in level, are time-delayed by digital delays, and then amplitude-scaled and spatialized by an *Acoustetron* (Crystal River Engineering). This set-up allows individualized or non-individualized HRTF filtering to be applied, as well as head-tracking. The entire system is controlled by MIDI from the experiment platform.

Absolute thresholds are determined at either a 70.7% or 50% level within a tolerance of 1 dB using a staircase algorithm (Levitt, 1970). The threshold is defined for each subject as the mean of 5 staircase direction reversals at the 1 dB level. No special training is given subjects to adopt a particular criteria; their task is to listen in terms for any detectable difference. A three-alternative forced-choice paradigm is used, where the subject identifies which of three stimuli heard in succession is different from the other two. Two of these are reference stimuli and one is the probe stimulus, with presentation randomized. The reference stimulus is the same as the probe, except that the reflection(s) being investigated are absent.

In one study using individualized HRTFs, we found significant enough differences in absolute thresholds between individuals to recommend that the threshold for inclusion of an early reflection be based on a dB value lower than the mean; e.g., the lower boundary of the first standard deviation (Begault, 1996). The threshold determined in this way turned out to be –23 dB for reflections at 5–15 msec at "narrow" and "wide" angles of incidence (±30° and ±90°, respectively). The data from a study currently underway finds similar thresholds for the overall level of a diffuse field with an rt60 of 500 msec. Overall, thresholds decrease with increasing angle of incidence, which suggests that the interaural time delay is primarily responsible for unmasking. Figure 3 shows some of the obtained results. The right plot shows a greater sensitivity for a floor reflection (0° azimuth, –36° elevation) than predicted by equation (1): –14 versus –9 dB for a 3 msec delay.

Other types of threshold studies have yet to be fully explored. For instance, the author found in one preliminary study that increasing the *number* of spatialized early reflections influences the perceived envelopment and distance of a sound source (Begault, 1987). In another study, listeners could not discriminate between different spatial incidence patterns of six HRTF-filtered "virtual early reflections" (Begault, 1992a). Subjects auralized the convolution of test material under three configurations: The first was facing the sound source, as derived from a room model; the second, a version with the listener turned 180 degrees; and the third, with a *random* spatial distribution of reflections. (The same timings and amplitudes were used for the early reflections in each case; only the directional information contained in the HRTFs was varied). It was extremely difficult for anyone to discriminate *any* difference between the three examples, suggesting that the directional properties of the reflections in the particular configuration used were below threshold. A similar study reported that "incorrect" patterns of reflections do not affect overall localization performance, compared to a predicted pattern (Zahorik, Kistler & Wightman, 1994).

CONCLUSION

The application of an auralization system to a given problem will result in

accurate solutions only to the degree that both acoustical behavior and human perception are modeled accurately. On the other hand, some improvement in computational efficiency may be possible if auralization systems are matched to human performance. Concurrent improvements in hardware speed and understanding of perceptual parameters might eventually allow fully immersive simulations of auditory environments in the near future.

REFERENCES

Bech, S. (1995a). Perception of reproduced sound: Audibility of individual reflections in a complete sound field, II. Audio Engineering Society 99th Convention (Preprint 4093).

Bech, S. (1995b). Timbral aspects of reproduced small rooms. I. Journal of the Acoustical Society of America, 97, 1717-1726.

Begault, D. R. (1987). Control of auditory distance. Ph.D. Dissertation, University of California San Diego.

Begault, D. R. (1992a). Binaural Auralization and Perceptual Veridicality. Audio Engineering Society 93rd Convention, San Francisco (Preprint 3421)..

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

Begault, D. R. (1992c). Perceptual similarity of measured and synthetic HRTF filtered speech stimuli. Journal of the Acoustical Society of America, 92, 2334.

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

Begault, D. R. (1996). Audible and inaudible early reflections: thresholds for auralization system design. Audio Engineering Society 100th Convention, Copenhagen (Preprint 4244).

Blauert, J. (1983). Spatial hearing (J. Allen, Trans.). Cambridge: MIT Press.

Dalenbäck, B.-I. (1995). A new model for room acoustic prediction and absorption. Ph.D. Thesis, Chalmers University of Technology.

Ebata, M., Sone, T., & Nimura, T. (1968). On the perception of direction of echo. Journal of the Acoustical Society of America, 44(2), 542-547.

EBU (1988). Sound quality assessment material recordings for subjective tests (EBU SQAM). Hanover: Polygram.

Foster, S. H., Wenzel, E. M., & Taylor, R. (1991). Real-time synthesis of complex acoustic environments. 1991 IEEE ASSP Workshop on applications of signal processing to audio and acoustics, New Platz, NY.

Gardner, W. G. (1994). Efficient convolution without input/output delay. Audio Engineering Society 97th Convention, San Francisco (Preprint 3897).

Gierlich, H. W. (1992). The Application of Binaural Technology. Journal of the Audio Engineering Society, 36, 219-243.

Hammershøi, D., Møller, H., Sørensen, M., & Larsen, K. (1992). Head-Related Transfer Functions: Measurements on 24 Subjects. Audio Engineering Society 92nd Convention, Vienna (Preprint 3289).

Horrall, T. R. (1970). Auditorium acoustics simulator: form and uses. Audio Engineering Society 39th Convention, New York (Preprint 761).

Jan, E., & Flanagan, J. (1995). Image model for computer simulation of sound wave behavior in an enclosure. 1995 IEEE ASSP Workshop on applications of signal processing to audio and acoustics, New Platz, NY.

Jot, J. M. (1996). Synthesizing three-dimensional sound scenes in audio or multimedia production and interactive human-computer interfaces. Interface to Real and Virtual Worlds 5th International Conference, Montpellier, FR.

Jot, J. M., Larcher, V., & Warusfel, O. (1995). Digital signal processing issues in the context of binaural and transaural stereophony. Audio Engineering Society 98th Convention, Paris (Preprint 3980).

Jullien, J.-P., Kahle, E., Winsberg, S., & Warusfel, O. (1992). Some results on the objective and perceptual characterization of room acoustical quality in both laboratory and real environments. Proceedings of the Institute of Acoustics, Birmingham (Vol. XIV, no. 2).

Kendall, G. S., & Martens, W. L. (1984). Simulating the cues of spatial hearing in natural environments. Proceedings of the 1984 International Computer Music Conference, Paris.

Kistler, D. J., and WIghtman, F. L. (1992). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. Journal of the Acoustical Society of America, 91, 1637-1647.

Kleiner, M., Dalenbäck, B.-I., & Svensson, P. (1993). Auralization-an overview. Journal of the Audio Engineering Society, 41, 861-875.

Kleiner, M., Svensson, P., & Dalenbäck, B.-I. (1990). Auralization: Experiments in Acoustical CAD. 89th Convention of the Audio Engineering Society, Los Angeles, California (Preprint # 2990).

Kulkarni, A., Isabelle, S. K., & Colburn, H. S. (1995). On the minimum-phase approximation of head-related transfer functions. 1995 IEEE ASSP Workshop on applications of signal processing to audio and acoustics, New Platz, NY.

Kuttruff, H. (1991). Room Acoustics. (3rd ed.). Essex, UK: Elsevier Science Publishers.

Kuttruff, H. (1993). Auralization of Impulse Responses Modeled on the Basis of Ray-Tracing Results. Audio Engineering Society 91st Convention, New York (Preprint 3122).

Lehnert, H., & Blauert, J. (1992). Principals of Binaural Room Simulation. Applied Acoustics, 36, 259-91.

Levitt, H. (1970). Transformed up-down methods in psychoacoustics. Journal of the Acoustical Society of America, 49, 467-477.

Møller, H. (1992). Fundamentals of Binaural Technology. Applied Acoustics, 36, 171-218.

Olive, S. E., & Toole, F. E. (1989). The detection of reflections in typical rooms. Journal of the Audio Engineering Society, 37, 539-553.

Reilly, A., & McGrath, D. (1995). Convolution processing for realistic reverberation. AES 98th Convention, Paris (preprint 3977).

Reilly, A., McGrath, D., & Dalenbäck, B.-I. (1995). Using auralization for creating animated 3-D sound fields across multiple speakers. AES 99th Convention, New York (preprint 4127).

Single, P., & McGrath, D. (1989). Implementation of a 32768-tap FIR filter using real time fast convolution. Audio Engineering Society 87th Convention, New York (Preprint 2830).

Takala, T., Hanninen, R., Valimaki, V., Savioja, L., Huopaniemi, J., Huotilainen, T., & Karjalainen, M. (1996). An integrated system for virtual audio reality. Audio Engineering Society 100th Convention, Copenhagen (Preprint 4229).

Wagener, B. (1971). Räumlische Verteilungen der Hörrichtungen in synthetisher Schallfeldern. Acustica, 25, 203-19.

Wenzel, E. M. (1992). Localization in virtual acoustic displays. Presence: Teleoperators and Virtual Environments, 1, 80-107.

Wenzel, E. M. (1996). What perception implies about implementation of interactive virtual acoustic environments. Audio Engineering Society 101st Convention, Los Angeles (Preprint 4353).

Wightman, F. L., & Kistler, D. J. (1989). Headphone simulation of free-field listening. I: Stimulus synthesis. Journal of the Acoustical Society of America, 85(2), 858-867.

Zahorik, P., Kistler, D., & Wightman, F. (1994). Sound localization in varying virtual acoustic environments. Proceedings of the Second International Conference on Auditory Display, ICAD '94, Santa Fe, NM.

Zurek, P. M. (1979). Measurements of binaural echo suppression. Journal of the Acoustical Society of America, 66(6), 1750-1757.

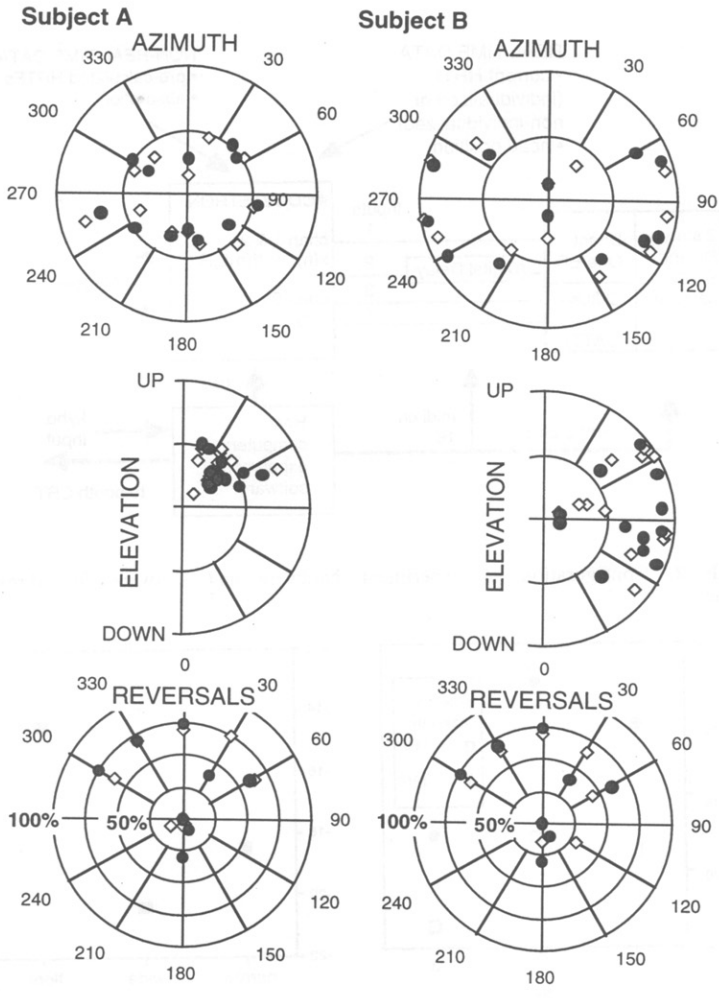

FIGURE 1. Localization data for two subjects, HRTF-filtered speech stimuli (from Begault, 1992). Top: azimuth vs. distance; Middle: elevation vs. distance; Bottom: target azimuth vs. percentage of reversals. Filled circles- HRTF measurements (512 coefficients); open diamonds- synthetic HRTFs, linear phase constant ITD (65 coefficient); 260 judgements per condition per azimuth. For azimuth and elevation plots, the inner ring indicates the edge of the head; the outer ring, complete externalization of all stimuli at that azimuth. For the reversals plot, each ring indicates increments of 25% of total number of stimuli that were reversed.

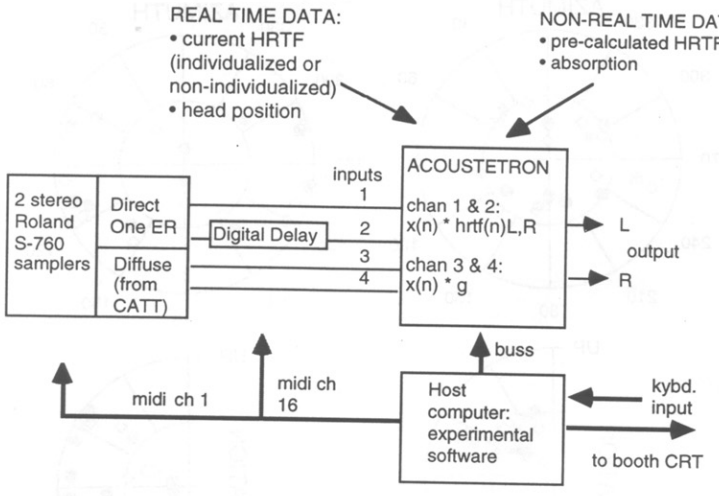

FIGURE 2. Configuration of experiment hardware and software for threshold experiments.

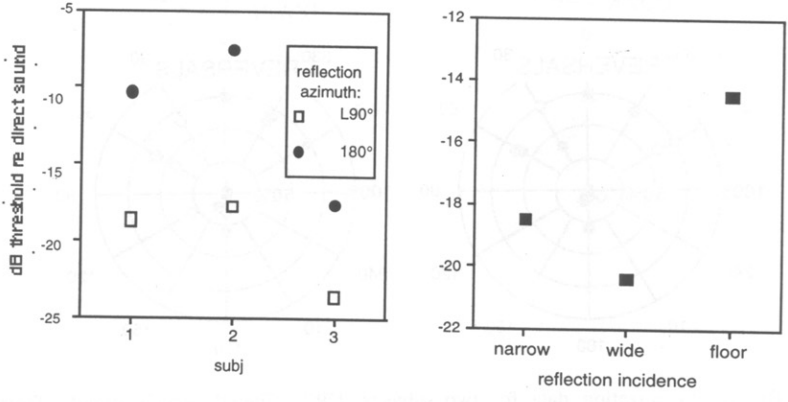

FIGURE 3. Left: illustration of threshold for a single reflection at two different azimuths, three different subjects. Note that subject 3's highest threshold is about the same level as the lowest threshold for subject 1 and 2. The reflection delay is 18 msec. Right: results for "narrow" (30° azimuth) and "wide" (90° azimuth) early reflections. For comparison the threshold is given for the a "floor" reflection at 0° azimuth, –36° elevation.