Virtual Acoustic Displays for Teleconferencing: Intelligibility Advantage for "Telephone-Grade" Audio*

DURAND R. BEGAULT

San Jose State University, Human Factors Research and Technology Division, NASA-Ames Research Center, Moffett Field, CA 94035 USA

Speech intelligibility was evaluated using a virtual acoustic (3-D audio) display using the method specified by ANSI (S3.2-1989). Ten subjects were evaluated with stimuli either unfiltered or low-pass filtered at 4-kHz. Results show that virtual acoustic techniques are advantageous for both full-bandwidth (44.1-kHz sampling rate) and low-bandwidth (8-kHz sampling rate) "telephone-grade" teleconferencing systems.

0 INTRODUCTION

In communication systems where more than one channel must be rendered intelligible, as used in police–fire–rescue, aeronautics, and space launch operations, it is common practice for multiple signals to be mixed to a single-channel headset, sometimes in conjunction with a telephone handset or loudspeakers. This is in spite of the fact that the advantage of spatial separation over headphones has been well known for some time [1]–[3]. The effect of including natural spatial cues—in particular, the head-related transfer function (HRTF)—was investigated in detail by Bronkhorst and Plomp [4], [5]. The use of a virtual acoustic (three-dimensional audio) display to accomplish this spatial separation for a stereo headphone user was previously investigated by the author; a hardware prototype is described in [6]. The technique is to separate multiple communication channels by placing each at a unique virtual audio position, thereby exploiting the well-known cocktail party effect [7], [8]. This approach allows an ordered, predictable approach to the distribution of sound sources compared to the use of headset–loudspeaker combinations, and increases the intelligibility of signals against noise relative to one-ear listening.

A preliminary experiment reported in [9] had shown a maximum intelligibility advantage of about 6 dB for communication call signs that were spatialized using three-dimensional audio techniques, against diotic speech babble. However, this study was conducted using a single male speaker as the signal source. A more rigorous approach specified in [10] was used in the current experiment, with three male speakers and two female speakers as sources for the speech stimuli, and a phonemically balanced word list in place of communication call signs.

1 HYPOTHESIS TESTED

The goal of the current experiment was to determine whether the low-pass characteristic of telephone-grade audio would significantly affect the binaural advantage expected with full-bandwidth stimuli. HRTF filtering, the main digital signal processing component of virtual audio, includes interaural time delays, interaural level differences, and high-frequency spectral shaping. Figs. 1 and 2 summarize the overall interaural delays and interaural level differences present in the set of binaural HRTF filters that were used to process the stimuli. Fig. 3 shows an example of the high-frequency spectral shaping present in the HRTF of a single ear that yields an elevation cue. By low-pass filtering the stimuli, most of the spectral shaping of the HRTF is effectively removed, thereby eliminating perceptual cues associated particularly with elevation perception and externalization [9], [11]. Data were gathered for non-low-pass-filtered stimuli for baseline comparison.

2 METHOD

The 50% intelligibility level for speech was evaluated for ten subjects using a virtual acoustic display that contained HRTFs of a single nonindividualized user, using a standard method for measuring the intelligibility of

* Presented at the 98th Convention of the Audio Engineering Society, Paris, France, 1995 February 25–28; revised 1999 September 18.

speech over communication systems [10]. A between-subjects experimental design was used. Five subjects were evaluated with stimuli low-pass filtered at 4 kHz, and an additional five subjects were evaluated with full-bandwidth stimuli (see Fig. 4). Eleven different spatialized azimuths were evaluated. The presentation of azimuths was randomized among subjects. In addition, thresholds were obtained for monotic (one-channel nonspatialized) playback, diotic (two-channel nonspatialized) playback, and for the stimuli in the absence of any noise masker.

All participants were screened prior to the experiment for normal hearing using standard audiometric techniques. Each subject wore Sennheiser HD-430 headphones within a soundproof booth, and gave responses using a computer terminal slaved to the experimental host computer. The host computer housed three-dimensional audio hardware (Crystal River Engineering Acoustetron) and custom experimental software. Diotic (two-channel monaural) white noise, filtered to approximate the spectral content of normal level speech, was used as the "speech-spectrum noise" masker. It was played at a constant level throughout each trial (52 dB, A-weighted); no amplitude modulation was used. The signal was spatialized (or monotic, or diotic) speech, consisting of a single word obtained from a modified rhyme test word list. This closed-set list (described fully in [10] and [12]) consists of 50 six-word sets of closely related American English monosyllabic words. These were mostly of the form consonant–vowel–consonant, with each set typically distinguished by a single phonemic element. Three male and two female speakers were used to prepare the signal stimuli. A digital recording was made for each word list entry for subsequent playback during the experiment.

The selected word, speaker, and spatial position were randomized for each trial. Subjects heard the diotic, speech-spectrum noise for 5 seconds. Within this period, after 2 seconds, the speech signal (duration around 1 second) was mixed in at a particular intensity level. They were then prompted to identify their best guess via the computer keyboard for the correct choice from six words seen on the computer screen. For example, a typical word list seen by the subject might be "cook shook look book took hook." There was no time limit for entering the correct answer.

A staircase algorithm was used to adjust the level of the signal within 0.5 dB of threshold for each experimental condition [13]. An "interleaved" staircase method was used, whereby the stimulus levels were adjusted both upward and downward (see details in Fig. 5). Within a single experimental block a subject heard randomly presented ascending and descending staircases for two different spatial positions that were randomly determined for each subject. Typically a subject would finish two to three blocks on a given day, each block requiring approximately 15 minutes. A separate block was given at the end of the experiment to assess the absolute threshold of nonspatialized speech in the absence of the noise masker. A practice block was used for training before the experiment began.

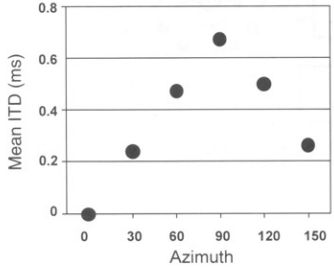

Fig. 1. Interaural delay present in HRTF set used in intelligibility experiment. Means of left and right symmetrical positions. 0^0 azimuth is directly in front of listener, with increasing azimuth angle moving toward rear of listener. Because the digital filters used were minimum-phase transformations of the original HRTF measurements, the interaural delay was constant across frequency (constant group delay) for a given HRTF.

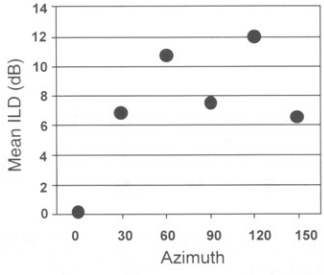

Fig. 2. Interaural level differences present in same HRTF set described in Fig. 1 (based on the rms level of a 1-s white noise burst). The means of left and right symmetrical positions are shown.

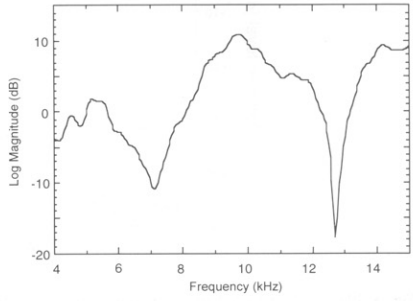

Fig. 3. High-frequency (>4-kHz) spectral information in HRTF at one ear, presumably used as a cue for source elevation, was eliminated in low-pass-filtered versions of stimuli used. Difference in HRTF magnitudes shown for sources elevated 54° above ear level and 36° below ear level relative to listener.

3 RESULTS

Fig. 6 shows the absolute threshold for the experimental conditions, with the mean of symmetrical azimuth positions given. The right part of the graph, labeled "dD" on the x axis, indicates the threshold for diotic (two-channel monaural) playback of the signal. The threshold for 50% intelligibility of the speech signal relative to the speech-noise masker ranges from +0 to +4 dB for the full-bandwidth stimuli and from +0 to +7.5 dB for the low-pass stimuli, as a function of azimuth.

Fig. 7 indicates the pattern of results in terms of the advantage of spatialized over diotic listening. These results were obtained by subtracting the mean threshold for the spatialized conditions from the diotic conditions. An analysis of variance indicated a significant difference in responses as a function of azimuth [$F(10, 80) = 4.82$, $p < 0.000$] but no significant difference between low-pass and full-bandwidth signal intelligibility [$F(1, 8) = 0.19$] or for the interaction between azimuth and bandwidth [$F(10, 80) = 0.67$]. Similar advantages are obtained by comparing spatialized stimuli to either left- or right-ear monotic listening.

The intelligibility advantage function matches the interaural time delay function shown in Fig. 1. Maximum intelligibility is at the azimuth position with maximal interaural delay (90°), as opposed to azimuths corres-

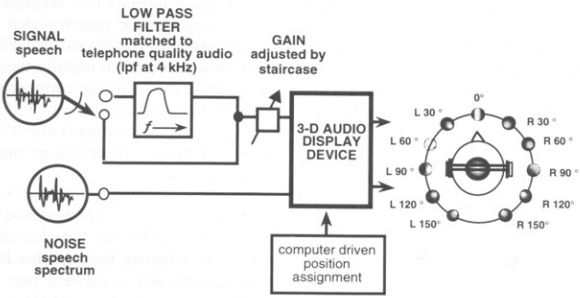

Fig. 4. Experiment playback configuration. At right, overhead view of spatialized positions used and diotic reference from listener's perspective.

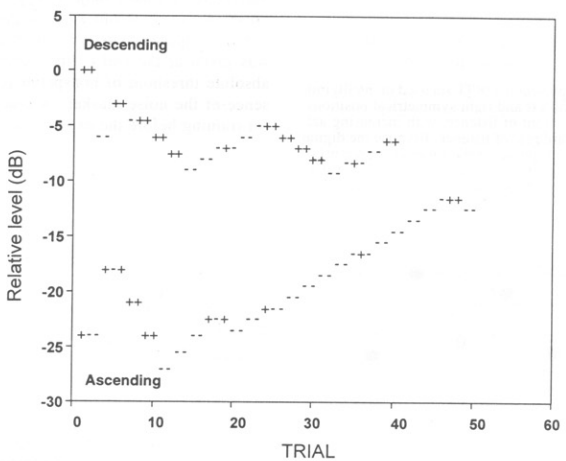

Fig. 5. Illustration of ascending and descending staircase adjustment of stimulus level for determining 50% intelligibility. + subject chose correct word from list of six words presented on computer screen; − incorrect choice. Step size (i.e., magnitude of intensity adjustment) is reduced or increased by 50% until minimum step size of 1 dB is attained: 6 dB—3 dB—1.5 dB—1 dB. Successive responses are evaluated to determine whether the intensity is adjusted upward or downward for the subsequent trial: the level is increased after (−, −), (−, +, −), or (+, −, −), and decreased after (+, +), (+, −, +), or (−, +, +). This "transformed up–down" technique is described in [13]. The threshold is defined as the mean of the first three reversals at the minimum step size for both ascending and descending staircases.

ponding to the maximum interaural level differences shown in Fig. 2. This result is not surprising, in that the bulk of the spectral energy content of speech lies in the frequency range where spatial location is cued by interaural time differences (below 1.5 kHz) as opposed to interaural level differences (above 1.5 kHz). Taking the mean of both conditions, the advantage ranges up to about 6 dB irrespective of the bandwidth following the general trend of the results found in [6] for call sign intelligibility (shown in Fig. 7 by +).

4 DISCUSSION

The results suggest that the interaural time delays resulting from virtual acoustic techniques (HRTF measurements) are advantageous for both low-sample-rate telephone-grade systems and full-bandwidth systems when more than one communication stream must be monitored simultaneously. The dB advantage over diotic listening using the virtual display is at least as good as for normal unaided hearing, as derived in [3]. On the other hand, the spectral modification caused by the HRTF is probably unimportant, particularly since it changes as a function of azimuth and elevation in frequencies higher than 4 kHz.

Overall these findings are important in that many communication systems transmit usable frequencies solely below 4 kHz, and that future computer-based telecommunication systems may economize resources by operating with less than full-bandwidth codecs. A device that utilized only the interaural delays derived from HRTFs might be suitable for the intelligibility of a single source against noise. It remains to be investigated what the implications might be for the advantage of HRTF spectral shaping versus a simple time-delay cue when *multiple* signal sources are involved at different azimuths, any of which could potentially be the desired signal. The latter reflects the state of affairs in a virtual acoustic

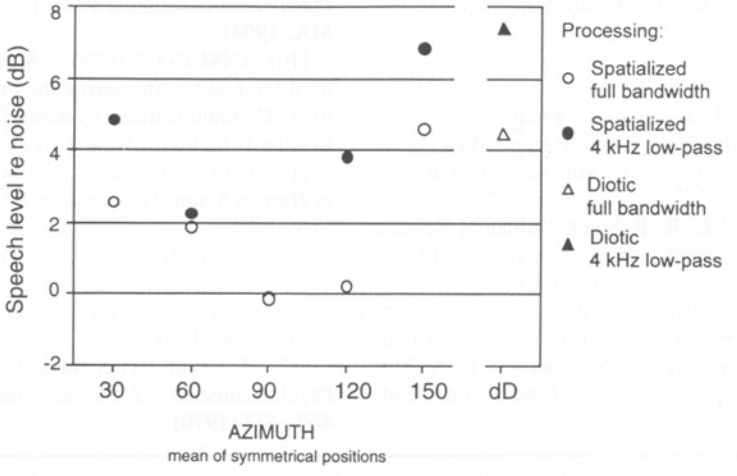

Fig. 6. Results for 50% intelligibility as a function of spatialized azimuths (mean of symmetrical left–right positions is shown) in terms of signal level in reference to noise level. Rightmost value labeled "dD" shows mean for diotic nonspatialized stimuli.

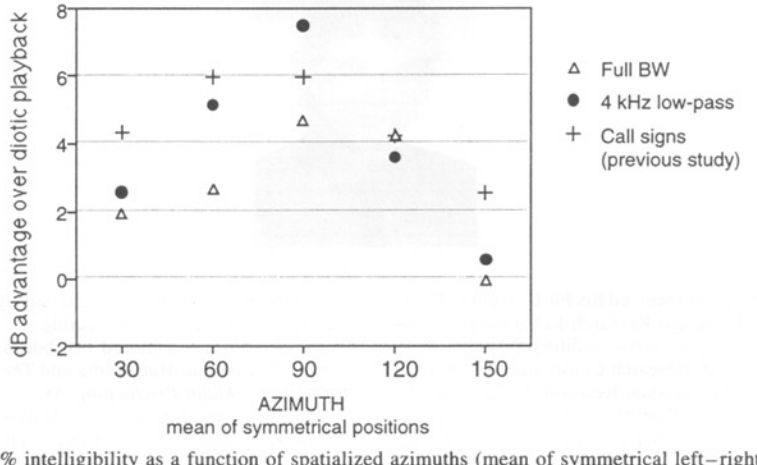

Fig. 7. Results for 50% intelligibility as a function of spatialized azimuths (mean of symmetrical left–right positions is shown) in terms of advantage of spatialized over diotic listening. △—full bandwidht advantage; ●—low-pass-filtered stimuli advantage. For comparison, results from [6] are also shown (+). Note that pattern of advantages follows general shape of curve for interaural delays shown in Fig. 1.

teleconference with many speaking voices. The results of [4] indicate that interaural level differences interfere with the unmasking achieved with interaural time differences, but that speech can potentially remain intelligible with as many as six interfering talkers, all speaking at the same level. Finally it is not known whether any additional intelligibility advantage may be gained when the subject's personal HRTFs or interaural time differences are used in the signal-processing portion of the three-dimensional audio communication display. Additional work in the application of three-dimensional audio technology to specific communication applications is currently underway at NASA Ames.

5 ACKNOWLEDGMENT

This work was supported by NASA Cooperative Agreement NCC 2-327. Special thanks to the personnel of the NASA Ames Spatial Auditory Display Laboratory: Elizabeth M. Wenzel, director; Joel Miller, programmer; and Rick Shrum, laboratory manager.

6 REFERENCES

[1] H. Levitt and L. R. Rabiner, "Predicting Binaural Gain in Intelligibility and Release from Masking of Speech," *J. Acoust. Soc. Am.*, vol. 42, pp. 820–829 (1967).

[2] H. Levitt and L. R. Rabiner, "Binaural Release from Masking for Speech and Gain in Intelligibility," *J. Acoust. Soc. Am.*, vol. 42, pp. 601–608 (1967).

[3] P. M. Zurek, "Binaural Advantages and Directional Effects in Speech Intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, G. A. Studebaker and I. Hochberg, Eds. (Allyn and Bacon, Needham Heights, MA, 1993).

[4] A. W. Bronkhorst and R. Plomp, "Effect of Multiple Speechlike Maskers on Binaural Speech Recognition in Normal and Impaired Hearing," *J. Acoust. Soc. Am.*, vol. 92, pp. 3132–3139 (1992).

[5] A. W. Bronkhorst and R. Plomp, "The Effect of Head-Induced Interaural Time and Level Differences on Speech Intelligibility in Noise," *J. Acoust. Soc. Am.*, vol. 83, pp. 1508–1516 (1988).

[6] D. R. Begault and T. Erbe, "Multichannel Spatial Auditory Display for Speech Communications," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 42, pp. 819–826 (1994 Oct.).

[7] E. C. Cherry, "Some Experiments on the Recognition of Speech with One and Two Ears," *J. Acoust. Soc. Am.*, vol. 25, pp. 975–979 (1953).

[8] E. C. Cherry and W. K. Taylor, "Some Further Experiments on the Recognition of Speech with One and with Two Ears," *J. Acoust. Soc. Am.*, vol. 26, pp. 549–554 (1954).

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

[10] ANSI (S3.2-1989), "American National Standard Method for Measuring the Intelligibility of Speech over Communication Systems," American National Standards Institute, New York, 1989.

[11] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1983).

[12] ISO 4870:1991(E), "Acoustics—The Construction and Calibration of Speech Intelligibility Tests," International Organization for Standardization, Geneva, Switzerland, 1991.

[13] H. Levitt, "Transformed Up–Down Methods in Psychoacoustics," *J. Acoust. Soc. Am.*, vol. 49, pp. 467–477 (1970).

THE AUTHOR

Durand R. Begault received his Ph.D. from U.C. San Diego (Computer Audio Research Laboratory). He has been a senior researcher at the Auditory Display Laboratory at NASA Ames Research Center since 1988, under the auspices of the National Research Council and San Jose-State University Foundation. His research specialties include acoustic displays for aviation, psychoacoustics of spatial hearing, digital signal processing, and applied acoustical engineering.

Dr. Begault has authored two books, *3-D Sound for Virtual Reality and Multimedia* and *The Sonic CD-ROM for Desktop Audio Production: An Electronic Guide to Producing Computer Audio for Multimedia*, both published by Academic Press Professional.