NASA Technical Memorandum 102826

Technical Aspects of a Demonstration Tape for Three-Dimensional Sound Displays

Durand R. Begault and Elizabeth M. Wenzel

October 1990

NASA
National Aeronautics and
Space Administration

NASA Technical Memorandum 102826

Technical Aspects of a Demonstration Tape for Three-Dimensional Sound Displays

Durand R. Begault and Elizabeth M. Wenzel, Ames Research Center, Moffett Field, California

October 1990

NASA
National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035-1000

SUMMARY

This document was developed to accompany an audio cassette that demonstrates work in three-dimensional auditory displays, developed at the Ames Research Center Aerospace Human Factors Division. It provides a text version of the audio material, and covers the theoretical and technical issues of spatial auditory displays in greater depth than on the cassette. This report also documents the technical procedures used in the production of the audio demonstrations, including methods for simulating rotorcraft radio communication, synthesizing auditory icons, and using the *Convolvotron*, a real-time spatialization device.

INTRODUCTION

This report documents the technical procedures used in the creation of a three-dimensional auditory display demonstration tape. The purpose of the tape and this report is to show some applications of the auditory research currently under way at the Aerospace Human Factors Research Division at Ames Research Center.[1]

The tape contains three separate demonstrations that illustrate the advantages of auditory displays using "headphone localization" as opposed to monaural and intensity stereo techniques. (Intensity stereo refers to interchannel level differences equal across all frequencies that are produced by an amplitude panpot on a mixing board.) A headphone localization technique is defined as a signal-processing method that allows a user to manipulate the apparent location of a sound to potentially any position in three-dimensional space.

The report is organized in three sections. First, a review of the theory and technique of headphone localization is given. Second, the software and hardware used for creating the tape is reviewed. Last, the three parts of the demonstration tape are described and the following subjects are emphasized: (1) purpose of the demo, (2) text used, and (3) technical means for producing the specific effect.

This work was performed while the first author held a National Research Council research associateship.

HEADPHONE LOCALIZATION THEORY AND TECHNIQUES

Head-Related Transfer Function Theory

The ultimate goal of three-dimensional auditory research is to have the ability to place sounds in an arbitrary position outside of the head for a listener who is wearing headphones. Thus, control of the listener's localization can be extended to anywhere in three-dimensional space. At one time, headphone localization was considered impossible; localization was thought to be possible only with "real world" (non-headphone) listening. Plenge's study of the perception of binaural ("dummy head") recordings showed that localization

[1] A copy of the demonstration tape is available directly from the authors at NASA Ames Research Center, MS 262-2, Moffett Field, CA 94035-1000. R-DAT or standard cassette dubs can be made if a high-quality blank tape is supplied with the request. Please indicate tape bias and desired noise-reduction format (Dolby B, Dolby C, or no noise reduction).

was indeed possible with headphone listening (ref. 1). The cues provided by the spectral modification of the outer ear were asserted to be the relevant factor. More recently, the work of Wightman and Kistler (refs. 2 and 3) has reasserted the notion that externalized localization of sounds in three-dimensional space is possible with headphone listening. They substituted digital signal-processing techniques for the dummy head recording process, a technique that according to Blauert was first developed in West Germany in the early 1970s (ref. 4).

The technique for implementing headphone localization involves the creation of a digital filter based on measurements of the head-related transfer function (HRTF). The HRTF can be thought of as a frequency-dependent amplitude and time delay that results from the resonances of the pinnae and the ear canal and the effects of head shadowing. These effects combine differentially across the frequency spectrum, as a function of sound source direction; hence, there is a transfer function imposed on an incoming signal that is unique for any given source position. In other words, the spectrum of the HRTF alters the spectrum of the input signal in a spatially dependent way.

The HRTF as a psychoacoustic cue for spatial hearing complements the "duplex theory," a long-held view that the principal cues for localization involved interaural level and time differences (ILD and ITD, respectively). The HRTF explains median plane perception of elevation and front–back positions, situations where ILD and ITD are close to 0 and/or below threshold. Researchers have also shown that overall localization acuity is diminished when pinnae cues are removed (ref. 5).

In applications contexts, implementation of audio signal processing based on the HRTF allows an operator a large degree of control over audio spatial manipulation, which was previously impossible with traditional stereo techniques, with both speakers and headphones. In the commercial audio industry, a large number of designs have been patented that are based on HRTF filtering, with the end purpose involving a relatively greater or lesser degree of control over spatial manipulation.[2] The availabilty and lower cost of special-purpose digital signal-processing chips, such as the Motorola 56001, have helped the implementation of HRTF-based spatial manipulation into practicable, realizable hardware designs.

HRTF Measurement and Filtering

The HRTF is measured by placing a probe microphone close to the eardrum of a subject, or at the entrance of the ear canal (refs. 2, 4, and 6). The goal is to obtain an impulse response for use in subsequent digital filtering algorithms, and a spectral measurement for analysis. In simplified terms, an impulse $x_{(n)}$ is sounded from a speaker at a carefully adjusted position in relation to a listener whose head is immobilized. The signal at the microphone $y_{(n)}$ is then recorded, and the procedure is repeated for the desired number of positions. The impulse response of $h_{(n)}$ (the HRTF for that position) is therefore obtained via the convolution of the speaker signal with its path of transmission to the microphone:

$$y_{(n)} = x_{(n)} * h_{(n)}$$

Figure 1 shows the magnitude of the HRTF for a single subject at one ear for different angles of incidence.

[2] For representative U.S. patents, see 4,118,599 (Iwahara et al.); 4,219,696 (Kogure et al.); 4,139,728 (Haramoto et al.); 4,731,848 (Kendall et al.); 4,817,149 (Myers); and 4,774,515 (Gehring).

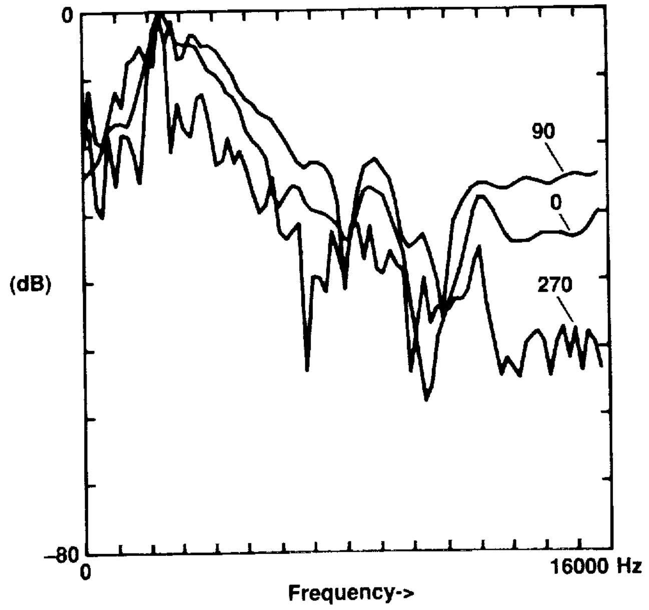

Figure 1. HRTF measurements: one subject, left ear, source at 0, 90 and 270° azimuth (adapted from ref. 2).

For implementation into spatial-manipulation hardware, the spectrum of the sound to be spatially processed must be multiplied by the spectra of two HRTF measurements, one for each ear. This is accomplished by *convolving* the input signal in the time domain with the impulse response of the HRTFs, using two finite impulse response (FIR) filters (ref. 7):

$$y_{(n)Left} = x_{(n)} * h_{(n)left}$$

$$y_{(n)Right} = x_{(n)} * h_{(n)right}$$

Figure 2 shows this method. Frequency responses of two filters for synthesizing a source directly opposite the right ear are shown. Incidentally, these particular frequency responses are not the HRTF of a single person, but rather are based on the *complex average* of 12 people, as calculated by Blauert (ref. 4). These measurements were then approximated by Begault using a filter design program (ref. 8). This technique can be referred to as direct HRTF filter design.

Perceptual Veridicality of Headphone Localization

The perceptual results of using headphone localization techniques are not completely predictable for the general population. Some people are unable to externalize HRTF-processed sound heard through headphones. The most common difficulty is in simulating sound sources on the median plane; sources synthesized to appear from the front of the listener usually sound as if they are inside the head. Often, a "bow tie" pattern is perceived when an HRTF-processed sound is intended to synthesize a circle with constant radius from the center of the head (fig. 3).

In theory, there ought to be improvements in localization performance when the filtering changes dynamically in relation to head movement—the means by which a person scans their auditory environment.

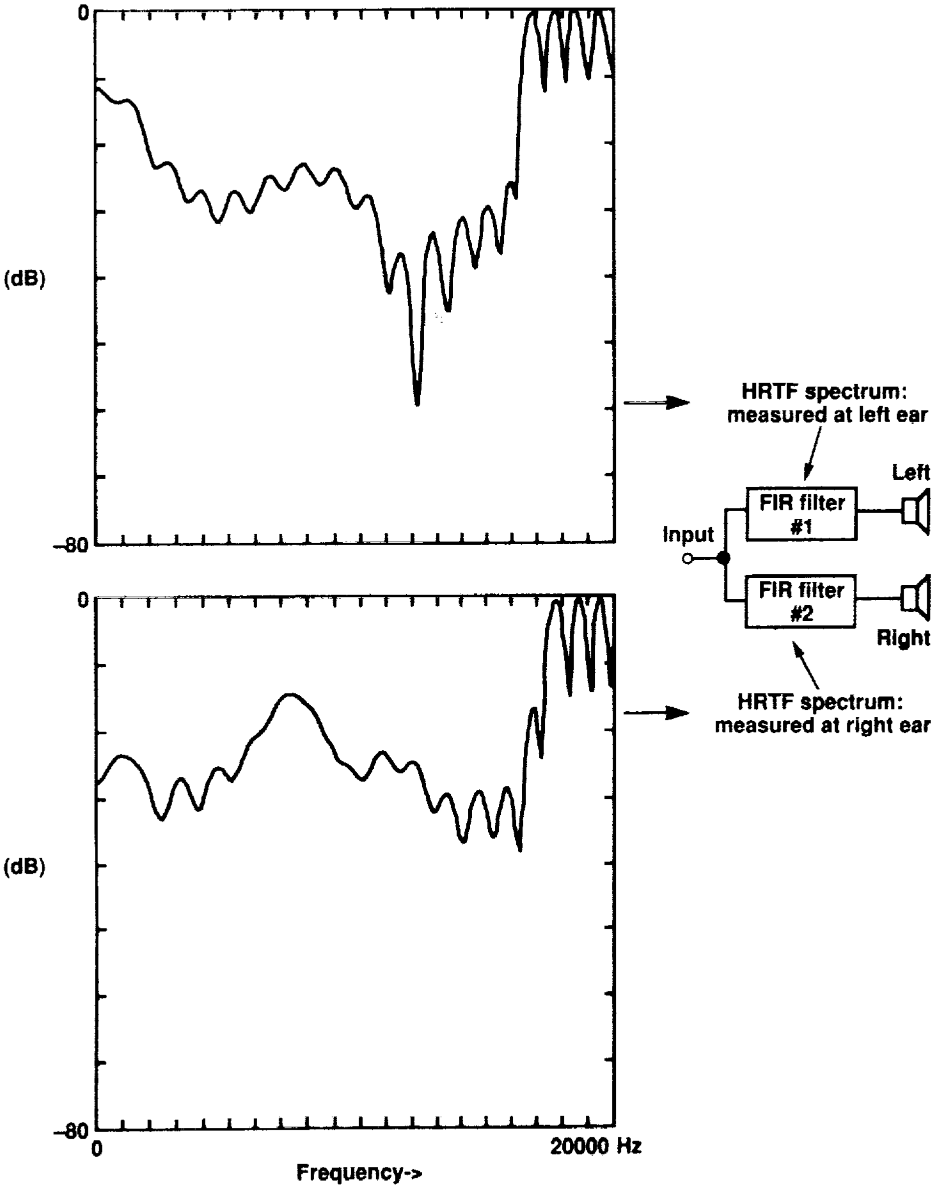

Figure 2. Headphone localization technique based on HRTF filtering. Direct HRTF filter design by Begault is shown, based on measurements obtained by Blauert.

This is an area of current psychoacoustic research at Ames and other institutions. Wenzel et al. have developed a hardware/software system known as the *Convolvotron* (ref. 9) that takes the output coordinates of a head tracker and assigns appropriate FIR filters to an input source. The magnetic head–tracking device is attached to a set of headphones that transmits three-dimensional coordinates of the listener's head position to a receiving device. The use of the head-tracking device is demonstrated on the third part of the tape. Begault is also researching the use of *synthetic reverberation* to capture features of the auditory environment to promote externalization and to possibly mitigate front-back reversals.

With continued research, the externalization and front-back reversal problems inherent in headphone localization should be alleviated. Until the problems are alleviated, it is impossible to map all dimensions

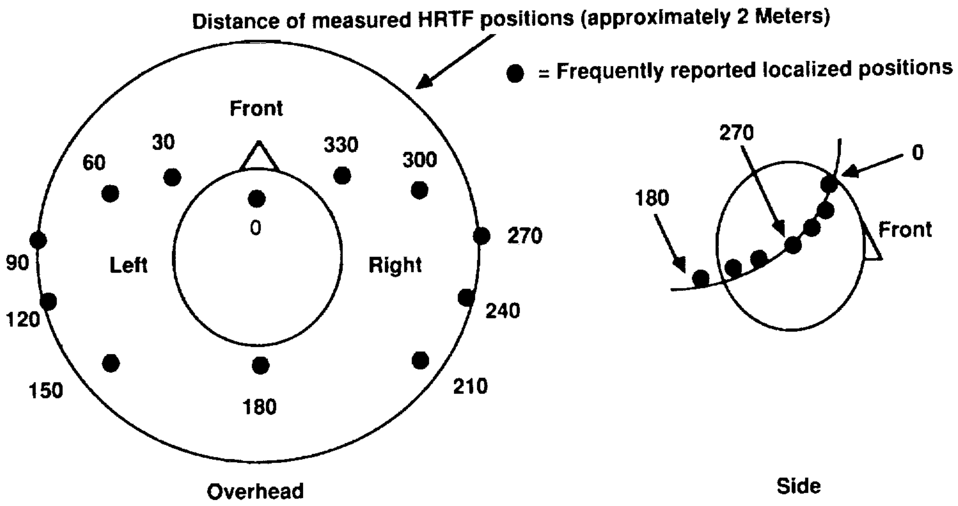
Distance of measured HRTF positions (approximately 2 Meters)
● = Frequently reported localized positions
Front
30
60
330
300
0
90
Left
Right
270
120
240
150
180
210
Overhead
270
0
180
Front
Side

Figure 3. Perceptual consequences in informal listening studies. The diagram shows positions frequently reported by subjects when presented with stimuli processed to sound equally distant from the center of the head. Note difficulty in externalizing sounds synthesized to the front and rear.

of actual auditory space on a virtual one with complete accuracy. However, it is possible to use *some* of the HRTF-synthesized positions, especially those to the sides. Their effect has been found, informally, by several authors, to be an improvement over traditional intensity or time difference stereo techniques for creating spatial auditory displays (refs. 4, 10, 11, and 12).

SOUND ASSEMBLY OF THE DEMONSTRATION TAPE

The master demonstration tape was produced by using a Panasonic SV-3500 R-DAT machine. Mixing, editing, and assembly of the spatially processed sound sources and the narration was accomplished using a Digidesign *Sound Tools* package: an *AD IN* (two-channel analog-digital converter), a *Sound Accelerator* card (two-channel digital-analog conversion, Motorola 56001 DSP chip, and host interface), and a *Sound Designer II* software package. An Apple Macintosh IIcx with 4 MB RAM was the host machine for the *Sound Designer* package.

The digital filtering (convolution) was accomplished two ways for the demonstration tape: (1) in real time, with the *Convolvotron,* and (2) in non-real time, using software signal processing. The third demonstration used two RDAT machines linked to the input and output, respectively, of the *Convolvotron,* which uses an IBM PC-AT host. This allowed the head-tracking device to control the virtual spatial position of the sound for the third demonstration.

Software-based, non-real-time convolution of sound sources inherently has a better signal-to-noise ratio than any real-time device. Two additional stages of analog processing are necessary with the *Convolvotron,* specifically, digital-to-analog conversion from the RDAT source and analog-to-digital conversion back to the Macintosh for mixing and editing. The fact that only static positions needed to be processed for the first and second demonstrations allowed work to be performed completely in the digital domain. On each of the three demonstrations on the tape, the narrator uses the phrase "processed with the Convolvotron," but, in actuality, software domain convolution was used for demonstrations 1 and 2. The phrase "processed using HRTF filtering" is used below in the written version of the narrator's script, when it was actually used. It must be emphasized that convolution performed in the software domain and in real time by the *Convolvotron* are equivalent.

Software signal processing was accomplished on a DEC Vax 11-780, using convolution software from the Computer Audio Research Laboratory (CARL) software distribution (ref. 13). Because the Macintosh was the platform for the Digidesign software for mixing and editing, compatibility of soundfiles between this system and the Vax was necessary. Custom software was developed to interchange Digidesign and CARL soundfile headers and to allow compatibility between the two different processors. Begault wrote the soundfile header software based on a program written by Phil Stone of Sterling Software Federal Systems, Palo Alto, CA.

The HRTFs used for this demonstration were obtained from measurements conducted by Wightman and Kistler at the University of Wisconsin, Madison (refs. 2 and 3). They are of a typical localizer. The impulse response of the HRTFs obtained for SDO were converted into 512 coefficient impulse responses readable by the CARL software. The original sampling rate of Wightman's measurements was 50 kHz, which posed a problem in that the soundfiles to be spatially processed were recorded at 48 kHz. The CARL software sample rate converter (*Convert*, written by Mark Dolson) was used instead of Digidesign's, because it allowed explicit indication of the number of coefficients used to indicate the converter's sync function.

The procedure for convolution, data transfer, and compatabilty of the spatially processed sound can be summarized as follows:

1. Record original sound source as a Digidesign (48-kHz sample rate, 16-bit word length, signed integer) soundfile, using *AD IN* and *Sound Designer.*

2. Edit the soundfile to the desired duration, and then normalize its gain, using *Sound Designer.*

3. Transfer file from the Macintosh to the Vax, via Ethernet connection.

4. Strip the Digidesign header from the soundfile and swap the bytes, using the custom software described previously.

5. Using the UNIX pipe facility with the CARL software, convert the soundfile from signed integer to 32-bit floating point (the format used by CARL software); convert the soundfile sample rate to 50-kHz; and then convolve the soundfile with the desired 50-kHz HRTF impulse response at the left ear for a given position.

6. The same procedures in the above item are repeated, but this time convolve the soundfile with the HRTF of the right ear for the same position.

7. The above two items are repeated for each desired spatially processed position—once for the left ear, once for the right ear.

8. Each separate left/right file is then sample-rate converted back to 48 kHz, converted from floating point back to signed integer, and then reformatted with the custom software with a header and byte swapped.

9. The file is transferred over the Ethernet connection back to the Macintosh. Left and right files are combined into a single stereo file using the Digidesign mix facility. (The header for a stereo Digidesign soundfile was, at the time, proprietary, making it impossible to do left/right soundfile combining at the level of the custom software.)

10. The spatially processed soundfiles are then mixed with other soundfiles using *Sound Designer*.

TAPE INTRODUCTION

Text

Text of the introductory narration of the tape follows:

This tape illustrates some of the capabilities and applications of the 3-D auditory display system being developed in the Aerospace Human Factors Research Division of NASA Ames Research Center. With this system, an arbitrary signal can be made to sound as if it is coming from any particular location in space even though the user is listening over headphones. The synthesis is achieved by digitally filtering the sounds with head-related transfer functions measured for a particular location in space. A special purpose signal processor known as the *Convolvotron* allows this spatialization to take place in real time. The project is a joint effort of Dr. Elizabeth Wenzel and Dr. Durand Begault of NASA-Ames, Dr. Frederic Wightman of the University of Wisconsin, Madison, and Mr. Scott Foster of Crystal River Engineering.

DEMONSTRATION 1

Purpose

The purpose of the first demonstration was to show the binaural advantage of increased intelligibilty of multiple streams of speech against noise. A rotorcraft pilot's aural environment was chosen as the scenario, because it is an excellent example of a high-workload environment that requires the monitoring of multiple communications channels. The high level of noise present from a helicopter's transmission engine also made the rotorcraft scenario a desirable one for showing the binaural versus the monaural advantage.

The advantage of binaural versus monaural listening is not only that localization cues such as ITD and ILD can be used, but also that binaural listening is superior to monaural listening for suppressing undesirable auditory input such as noise. The binaural advantage can be easily demonstrated by listening to a person speak in a noisy environment while you have one ear plugged. The noise seems to interfere less with speech when both ears are open. This is in stark contrast to the situation where a pilot must listen to many undifferentiated voices coming over a monotic headset.

Studies by Cherry et al. established the existence of what is commonly called "the cocktail party effect" and its relation to binaural hearing (refs. 14 and 15). The term comes from the observation that in a group of people who are all speaking simultaneously, it is still possible to understand a single stream of speech. This has led to many studies comparing the intelligibility of speech under binaural and monaural presentation. The difference between a masking level signal (noise or other voices) and the necessary level for intelligibility of the desired signal is termed the binaural intelligibility level difference, or BILD. Results from experimental data evaluating the BILD differ depending on the stimuli used and the criteria for evaluating intelligibility; generally, it ranges from 3 to 12 dB (ref. 4). Koening and Zurek also described the advantage of binaural over monaural hearing for squelching noise (refs. 16 and 17).

Work by Bronkhorst and Plomp measured the BILD as a function of the angle of incidence, using conditions that essentially compare the lateralization techniques discussed earlier (pure ITD or ILD) to headphone localization techniques (listening through the ears of a dummy head). Their results showed that the improvement using headphone localization techniques is around 2-4 dB (ref. 18). In informal studies conducted by Begault at Ames, the externalization provided by HRTFs with maximal ITDs (60, 90, 240, and 300° azimuth) was judged to be superior to traditional stereo techniques for selective attention to four voices. Because of binaural summation of loudness, a binaurally equipped pilot would also need less amplitude of the signal at each ear, resulting in an additional advantage in minimizing hearing fatigue.

Text

Demonstration 1

In the aviation environment, it is often the case that both air traffic controllers and the cockpit crew may need to monitor simultaneous channels of communication. The ability to separate multiple voices by location improves the intelligibility of the individual voices and makes it easier to attend to any one particular message. The sound can also provide information about the spatial location of the sources.

In this example, you will hear four different speakers embedded in a background of helicopter noise. First the voices and noise are mixed without spatialization, as would be the case in a typical communication context.

sound insert 1-a

Next the four voices have been processed by HRTF filtering so that they appear to come from four different locations around your head.

sound insert 1-b

Again, the non-spatially processed and spatially processed sounds are contrasted by alternation between the two cases.

sound insert 1-c

Technical Procedure

For this demonstration, it was necessary to create a realistic-sounding radio communication scenario for a helicopter pilot, and to obtain a sound source of a helicopter from the pilot's perspective. It is not unusual for a helicopter pilot to monitor four communications channels simultaneously; hence, it was decided to form a sound bed consisting of one ground control person and three other pilots, with a background of UH-1 (Huey) helicopter noise as the "masker."

The source for the four communication streams was a set of tapes previously recorded by actual pilots and other laboratory personnel for use in a rotorcraft applications experiment. These recordings were simulated in that they were made in normal office rooms at NASA by using radio communication equipment. This resulted in a realistic radio timbre as a result of nonlinearities of the microphone, but without cockpit environmental noise from engine, transmission, or wind. This was ideal for the demonstration tape in that the noise could be mixed independently of the spatially processed or monaural speech.

The source for the helicopter noise was a 3-sec recording made from within the cockpit of a UH-1 (Huey) helicopter. This sound was looped by using a digital sampling synthesizer (Yamaha TX16-W) and was then recorded for 45 sec as a Digidesign soundfile on the Macintosh. The helicopter noise was mixed in at the same level relative to both the monaural and spatially processed speech sounds in sound inserts *1-a* to *1-c*.

Figure 4 shows the mixing schemes for the vocal communication in sound inserts *1-a* to *1-c*. In sound insert *1-a*, the 4 × 1 mix was used—the signals were merely added to each other and were delivered diotically (identically to both ears). In sound inserts *1-b* and *1-c*, HRTF-spatially processed versions of each communication channel were formed with an 8 × 2 mix and delivered binaurally. Each voice stream was spatially processed to a unique position: 60, 150, 210, or 300° azimuth (all at 0° elevation). These positions were chosen as those most easily distinguishable from one another, from a set of 24 positions spatially processed at every 15° azimuth.

It should be noted that using purely dichotic helicopter noise in the presence of spatially processed communication channels could possibly enhance the signal-to-noise ratio somewhat more than that of the actual cockpit context, where the ambient sound would be partially correlated at the ears. However, because the purpose of the demonstration was simply to illustrate the effect of spatialization on intelligibility, and because binaurally recorded helicopter sounds were not available, we felt this was not a critical deviation from the actual situation.

DEMONSTRATION 2

Purpose

The second demonstration emphasized the application of spatial audio to nonspeech, auditory signals—*auditory icons*. The binaural advantage for discriminating between multiple sources of sound delivered over headphones was again the focus of the demonstration. In contrast to the first demonstration, the two auditory icons used here were nearly identical in timbre, and differed only in pitch (600 and 800 Hz center

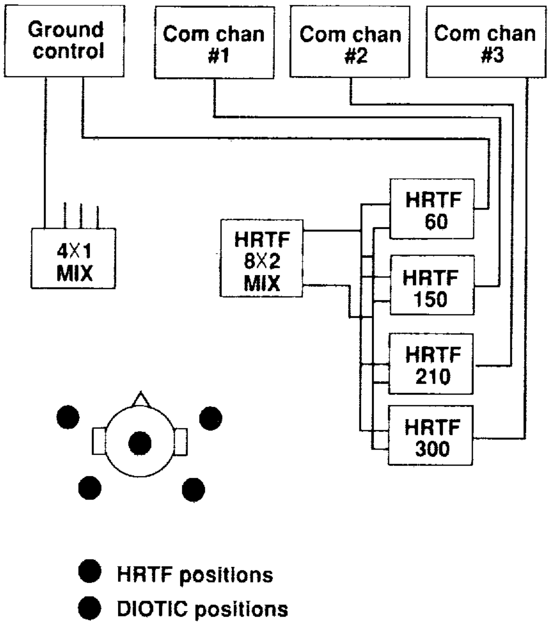

Figure 4. Mixing scheme for demonstration 1 (rotocraft example).

frequencies) and repetition rate, for reasons explained below. Unlike the first demonstration, no masking sound was mixed into the signal.

The goal of the demonstration was to show how spatialization facilitates the identification of a signal of interest from a signal that is relatively unimportant. We intended to suggest the auditory environment encountered by an operator of a sonar display in a submarine, an applications context that met this particular goal. The low and high frequencies of the auditory icons were intended to represent two different underwater vessels that the operator was monitoring. Although these stimuli do not correspond to actual sonar signals, this was unimportant for purposes of the demonstration.

In the demonstration, the higher-pitched signal maintains a constant repetition rate, while the lower-pitched signal gradually increases its repetition rate over the total duration, approximately 10 sec. This would correspond to a hypothetical situation faced by a sonar operator, who was identifying which vessel was moving closer and which one was stationary. An increase in the repetition rate of the icon would mean that the reflecting object was moving closer.

Figure 5 symbolically illustrates the two auditory icons and shows their temporal arrangement. The auditory icon itself is represented by the four arrows, which correspond to amplitude peaks. The upward arrow represents the higher frequency icon and the downward arrow represents the lower frequency icon. When the two signals are heard diotically, as in (c), it becomes difficult to segregate the two except on the basis of timbre. By applying HRTF spatialization to (a) and (b) and playing the sound binaurally, segregation is facilitated for the listener.

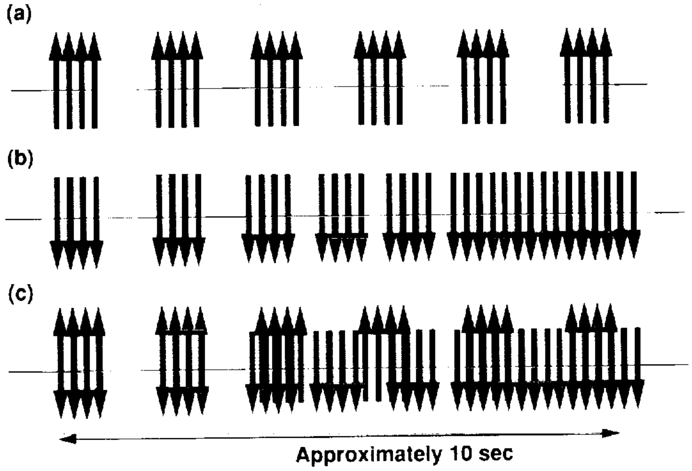

Figure 5. Temporal arrangement of sonar-like auditory icons. (a) Signal with higher frequency: rate constant; (b) signal with lower frequency: rate accelerates; (c) combination of (a) and (b), as heard diotically.

Text

Nonspeech audio cues, as in cockpit warnings and alarms and sonar displays, could also benefit from spatialization. For example, in an auditory sonar display, it might be appropriate for the operator to monitor an acoustic feature such as a change in the rate of returning echoes in order to detect the motion of one or more potential targets.

Here are two auditory icons which represent this type of situation. Note that in the first example, the rate of the higher-pitched sound is constant.

sound insert 2-a

In the second example, the rate of the lower-pitched sound increases.

sound insert 2-b

Here are the same two icons mixed together monaurally. It is difficult to detect which of the two has the faster rate.

sound insert 2-c

Next, the two icons have been spatially processed by HRTF filtering. Now it is much easier to detect which of the two has increased in rate.

sound insert 2-d

Technical Procedure

Using the same procedures as in the first demonstration, the two different auditory icons were presented in monaural and spatially processed stereo mixes, using 120 and 240° HRTF filters for the latter version. The auditory icons used in this demonstration were synthesized, rather than using actual sonar recordings. This allowed tailoring the sounds to the specific requirements of the demonstration, which were (1) wide harmonic bandwidth, but with two discernable pitches; (2) high signal-noise ratio of the icon itself; and (3) ease of manipulation of the repetition rate to create the two temporal patterns. Below, with reference to figure 6, we describe the steps used in creating the auditory icons.

1. A Digidesign program called *Softsynth* was used to generate the initial source signals A, B, C (Step 1 in fig. 6). This software synthesis program allows independent control over the amplitude and pitch envelopes of up to 32 harmonics, which can be sinusoids or harmonically rich waveforms in themselves. The three signals were a click and two types of impulsive "wood-block" timbres at two different center frequencies, which provided a fast rise-time of the amplitude envelope.

2. To design the icon so that it was harmonically more complex, the three *Softsynth* soundfiles were combined using *Sound Designer* mixing software (Step 2 in fig. 6). The duration of the waveform at this point is approximately 100 msec, with no discernible pitch.

3. Step 3 of figure 6 shows how the mixed waveform was overlapped with a time-delayed version of itself, again using the *Sound Designer* mixing facility. Two versions were created: a two-overlap version X and a three-overlap version Y. The overlapping of the amplitude envelopes is significant especially because the initial onset-release portion of the envelope is repeated. This gives the auditory system as much transient information in as short a time as possible, thereby facilitating localization. See the study by Hartmann on the effect of amplitude envelopes on localization (ref. 19). The time interval between the onset peaks of the envelope was adjusted to approximately 40 msec to prevent perceptual fusion ("smearing") of the transients. Each overlapped waveform was edited to have a total duration of 150 msec.

4. Step 4 of figure 6 shows how the two overlapped waveforms X and Y were arranged in time. A silence of 100 msec was inserted between each iteration of the overlapped signals (i.e., the black boxes of step 4). This resulted in an auditory icon with a total duration of 1 sec.

At this point, band-pass infinite impulse response (IIR) filtering was used to create two differently pitched versions of the icon, corresponding to the two different vessels detected by the radar. Center frequencies were set at 600 and 800 Hz, with a 75-Hz bandwidth and 20-dB relative gain in the passband. The filtering was also accomplished using the *Sound Designer* software.

With reference to the lower-pitched icon, figure 7 shows the time waveform of X, and figure 8 shows its amplitude spectrum over time. In figure 8, note that the icon has been designed so that energy is maintained over time in the center frequency region around 600 Hz, thus dominating as the perceived pitch. Note also the higher frequency energy extending up to 12.48 kHz that results from the two transients at 40-msec intervals. The transient envelope and the wide-band spectral spread of energy facilitate localization of the icon.

5. Step 5 of figure 6 shows the final temporal arrangement of the icons in the demonstration: 1 sec of icon, followed by 1 sec of silence. The overall pattern in the demonstration itself can be gleaned from figure 5. Hence, the repetition rate of the higher pitch icon is 0.5 Hz (once every 2 sec). To create the

effect of increasing repetition rate for the lower-pitched icon, the duration of the silence was progressively shortened throughout the course of the demonstration.

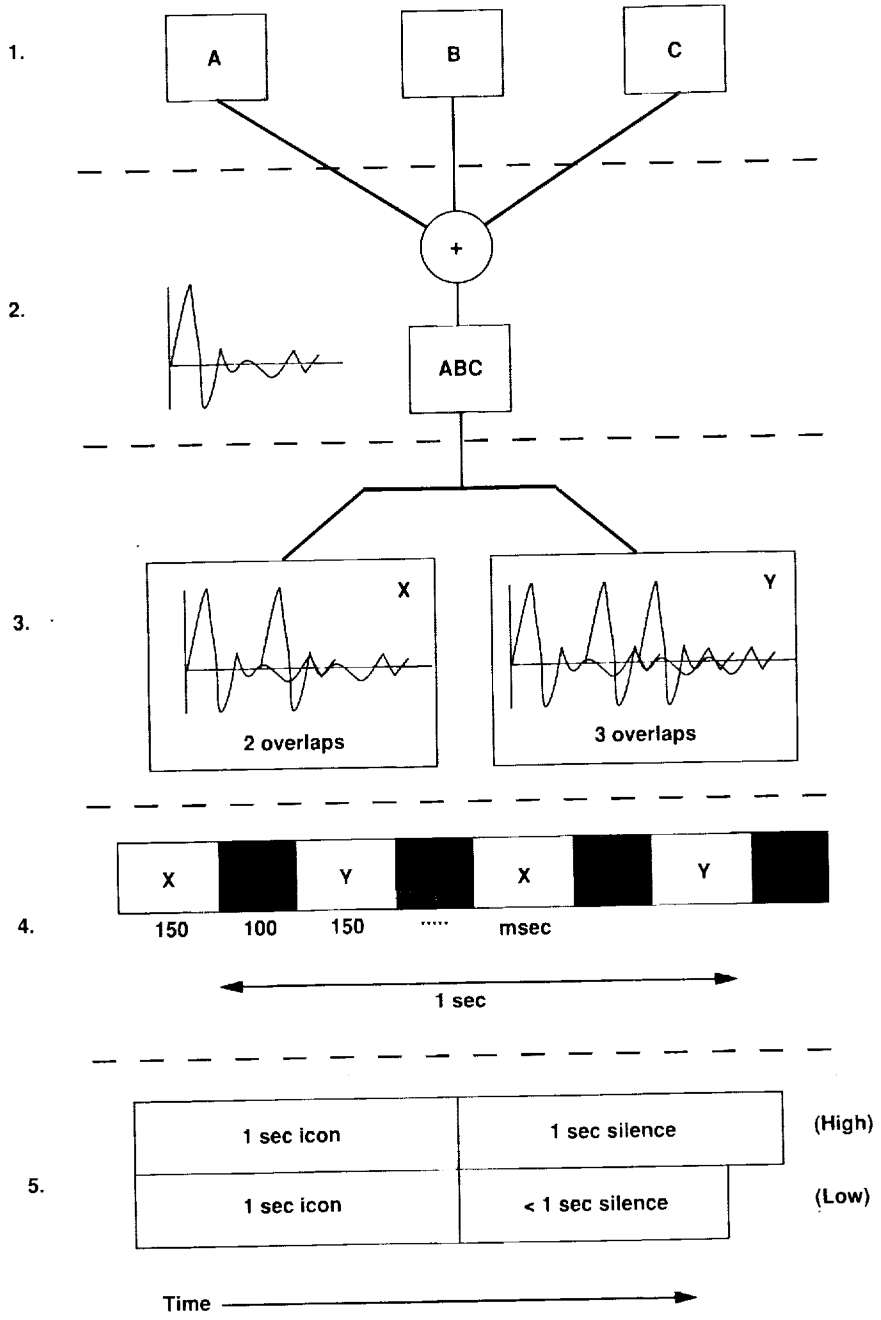

Figure 6. Steps taken in the formation of sonar-like auditory icon. See text for explanation.

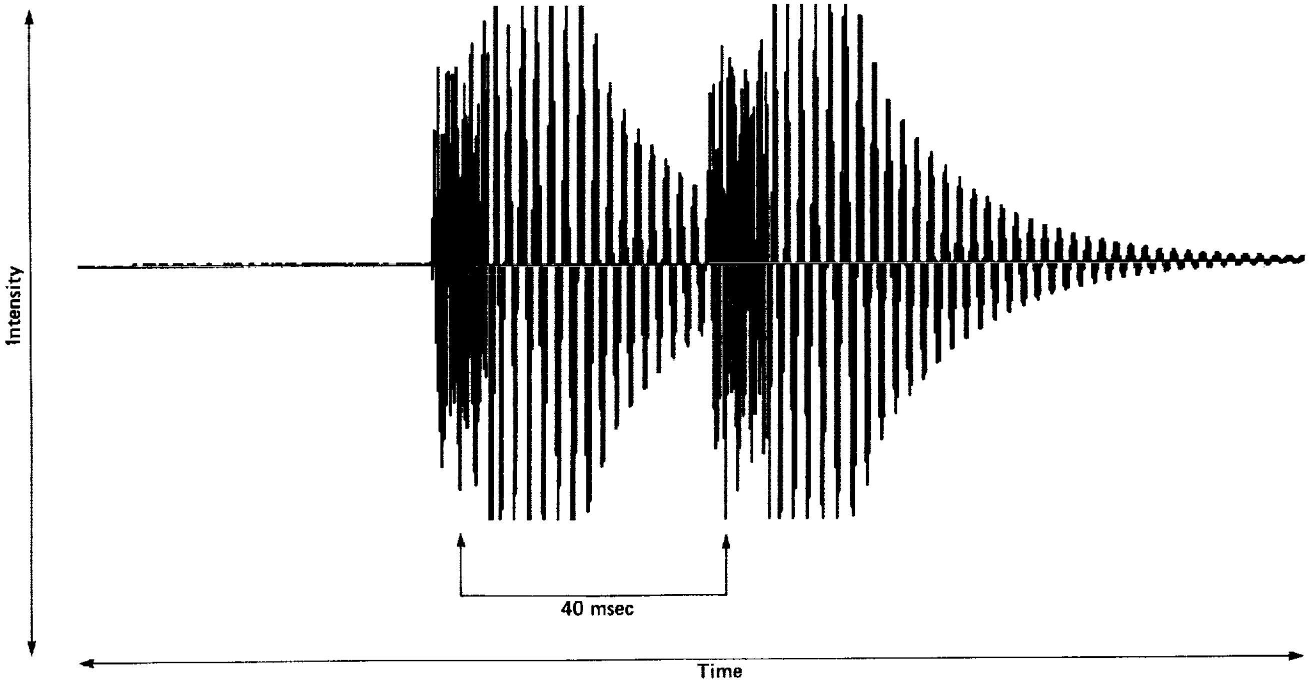

Figure 7. Time display of X waveform. Note the overlap at approximately 40 msec.

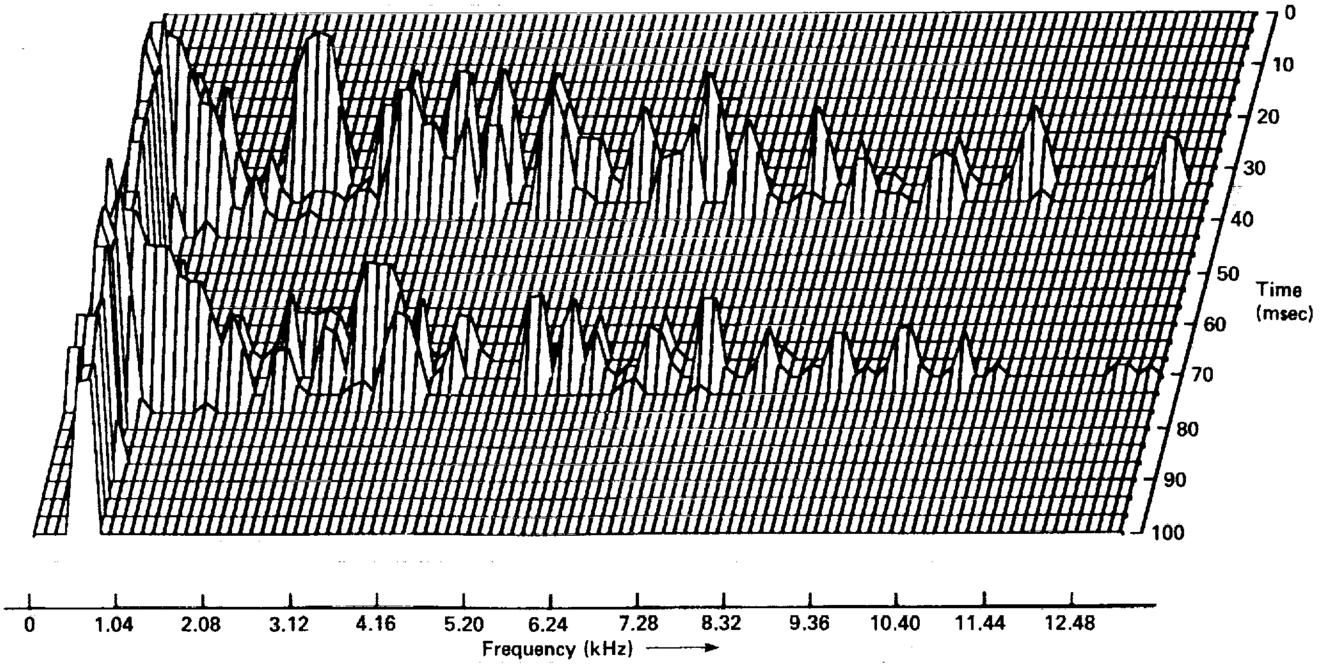

Figure 8. Spectrum of X waveform (lower pitch). Note that energy is maintained at the 600-Hz center frequency for the entire duration, and that the transients supply energy through 12 kHz at two 40-msec intervals.

DEMONSTRATION 3

Purpose

The purpose of the third demonstration was to focus on the real-time spatialization aspects of the *Convolvotron*. The demonstration consists of three sections:

1. comparison of monaural, stereo, and HRTF spatialization;
2. the *Convolvotron*: headtracking; and
3. the *Convolvotron*: moving and fixed source.

Although this demonstration was primarily designed to focus on the *Convolvotron*, the first part was designed to compare HRTF spatialization to normal stereo. A speech intellegibility paradigm was used to bring this into focus. The following definitions were paraphrased from a manual on English usage (ref. 20). They were read independently and mixed in monaural, intensity stereo, and HRTF spatial versions. The objective was to illustrate the difficulty of separating the words *sprain* and *strain* without spatialization.

Sprain describes the result of a momentary wrench or twist, especially of an ankle or a wrist.

Strain describes an exertion of a muscle too strong or too long for its capacity.

Perceptually, the monaural version is almost unintelligible, the stereo version allows a little more segregation of the two streams, and the HRTF spatially processed version allows the words *strain* and *sprain* to be clearly distinguished.

In the second and third parts of this last demonstration, two sound sources were spatially processed in real time with the *Convolvotron*: a speech recording and a simple noise burst. These sources were chosen for their contrast. In the second part, the sources remain "fixed" in virtual space at two positions; the spatial effect is caused by the operator of the *Convolvotron* moving its magnetic sensor by head movement. In the third part, the head remains fixed and the noise source moves via spatial trajectory software designed to simulate circular movement about the head, while the speech is fixed at 120° azimuth, 0° elevation.

Text

The final demonstration illustrates how the dynamic, real-time capabilities of the *Convolvotron* can enhance the simulation of a virtual acoustic space. First, the effects of increasing spatialization are heard by comparing monaural, normal stereo, and filtering by head-related transfer functions.

Here, two passages of speech by the same speaker are mixed monaurally.

sound insert 3-a

Next, a simple amplitude difference analogous to normal stereo is introduced.

sound insert 3-b

This time, the passages have been filtered by the *Convolvotron.* Note that it is now much easier to attend to each sentence.

sound insert 3-c

The dynamic capabilities of the *Convolvotron* allow a listener to move their head with respect to one or more virtual sound sources. This is accomplished in real time by providing information about a listener's head position with a magnetic head-tracking device. In the next example, imagine that you are wearing such a head tracker and that you are moving your head among two virtual sound sources.

sound insert 3-d

Finally, we synthesize a moving sound source. In this example, imagine that you are sitting still. The voice is now stationary while the noiseburst moves around your head.

sound insert 3-e

Technical Procedure

The procedures for creating these stimuli were essentially identical to those used for Demonstration 1. Sound insert *3-a* was simply a diotic mixture of the two speech passages, sound insert *3-b* had an overall interaural intensity difference of 9 dB between the two passages, and in sound insert *3-c* the passages were filtered by HRTFs corresponding to 120° azimuth, 0° elevation and 240° azimuth, 0° elevation.

Sound inserts *3-d* and *3-e* used the real-time filtering (convolution) capabilities of the Convolvotron system (see ref. 9) in lieu of the CARL software. Again, two RDAT machines were used to provide the speech inputs to the Convolvotron and record its output signal output after real-time spatialization. However, the 1-sec noiseburst of the final example was generated digitally in real time using an algorithm for synthesizing Gaussian noise. The noise was generated with a TMS 32020/C25 digital signal-processing board also resident in the IBM PC-AT that hosts the Convolvotron. The noise was converted to an analog signal via the digital-to-analog of the TMS board and routed to an analog-to-digital on the Convolvotron. This extra conversion was required because the Convolvotron does not have provisions for direct digital input.

The dynamic, interactive capability of the Convolvotron illustrated in sound insert *3-d* is aided by the use of a magnetic head-tracking device, the Polhemus 3Space Isotrack. This device can monitor the location and orientation of the listener's head in 6 degrees of freedom; absolute position in x, y, z coordinates and pitch, roll, and yaw in angular coordinates. Position and orientation are measured by means of the relative relationship between a source that generates a magnetic field in three orthogonal axes, and a sensor that monitors the strength of the field. In the Convolvotron system, the sensor is mounted at the top of the listener's headphones while the source is mounted in a stationary position. The coordinates are sent to the AT host via an RS-232 serial line at a nominal update rate of 60 Hz, with an angular resolution of 0.85° and a position resolution of 0.25 in. (sensor-to-source distance of 30 in.).

In practice, the spatial sampling (update) rate is at least 20 Hz; that is, the shortest interval at which a source location can be updated by the head tracker is about every 50 msec. An additional delay, about 30-40 msec, is introduced by the host computer and the Convolvotron which depends upon the number of sources (1 to 4) and the size of the HRTFs (128 to 512 coefficients). Thus, the total lag through the system is about 80 to 90 msec. This implies that relative angular source velocities of 360 deg/sec can be updated approximately every 29 to 32°, 180 deg/sec can be updated every 14 to 16°, and so on. However, when the system corrects for head position, as in sound insert *3-d*, or simulates a motion trajectory as in sound insert *3-e*, locations at greater resolution than the original empirical measurements are simulated by interpolation with linear weighting functions. That is, an "in-between" location is synthesized from a weighted combination of the four measured HRTFs that are closest to the desired location. The particular interpolation algorithm used effectively results in a time-varying filter which smoothly changes from one spatial location (HRTF) to the next at the sampling rate (50 kHz) of the system. Thus, audible artifacts due to abrupt changes in spatial location (switching between HRTFs) are avoided even though spatial sampling may occur at lower resolutions due to the 80- to 90-msec delay through the system.

As before, final mixing of the stimuli was accomplished using the Digidesign software.

CONCLUSION

The technical procedures used in the creation of a tape demonstrating the features and capabilities of a three-dimensional auditory display were reviewed in this report. The demonstrations illustrate the types of applications, such as communications in the cockpit and advanced sonar displays, which may benefit from this type of acoustic processing. They also illustrate the notion of a nonspeech acoustic stimulus, or auditory icon, and suggest the interactive capabilities of a real-time system—the Convolvotron—based on convolution with head-related transfer functions.

REFERENCES

1. Plenge, G.: On the Differences Between Localization and Lateralization. J. Acous. Soc. America, vol. 56, no. 3, 1974, pp. 944-951.

2. Wightman, Frederic L.; and Kistler, Doris J.: Headphone Simulation of Free-Field Listening. I: Stimulus Synthesis. J. Acous. Soc. America, vol. 85, no. 2, 1989, pp. 858-867.

3. Wightman, Frederic L.; and Kistler, Doris J.: Headphone Simulation of Free-Field Listening. II: Psychophysical Validation. J. Acous. Soc. America, vol. 85, no. 2, 1989, pp. 868-878.

4. Blauert, Jens: Spatial Hearing: The Psychophysics of Human Sound Localization. MIT Press, 1983.

5. Oldfield, Simon R.; and Parker, Simon P. A.: Acuity of Sound Localisation: A Topography of Sound Space: Pinna Cues Absent. Perception, vol. 13, 1984, pp. 600-617.

6. Mehrgardt, S.; and Mellert, V.: Transformation Characteristics of the External Human Ear. J. Acous. Soc. America, vol. 61, no. 6, 1977, pp. 1567-1576.

7. Oppenheim, A. V.; and Schafer, R. W.: Digital Signal Processing. Prentice-Hall Inc., 1975.

8. Begault, Durand R.: Control of Auditory Distance. Ph.D. Thesis, University of California, San Diego, 1987.

9. Wenzel, E. M.; Wightman, F. L.; and Foster, S. H.: A Virtual Display System for Conveying Three-Dimensional Acoustic Information. *In* Proceedings of the Human Factors Society 32nd Annual Meeting, Human Factors Society, 1988, pp. 86-90.

10. Griesinger, David: Equalization and Spatial Equalization of Dummy-head Recordings for Loudspeaker Reproduction. J. Audio Engineering Soc., vol. 37, no. 1-2, 1989, pp. 20-29.

11. Begault, Durand R.: Spatial Manipulation and Computers. Ex Tempore, vol. 4, no. 1, 1986, pp. 56-88.

12. Kendall, Gary S.; and Martens, William L.: Simulating the Cues of Spatial Hearing in Natural Environments. Technical Report, Northwestern University Computer Music, 1984.

13. Moore, F. Richard; Loy, D. Gareth; and Dolson, Mark: CARL Startup Kit. Center for Music Experiment, 1985.

14. Cherry, E. C.: Some Experiments on the Recognition of Speech with One and with Two Ears. J. Acous. Soc. America, vol. 25, 1953, pp. 975-979.

15. Cherry, E. C.; and Taylor, W. K.: Some Further Experiments on the Recognition of Speech with One and with Two Ears. J. Acous. Soc. America, vol. 26, 1954, pp. 549-554.

16. Koening, W.: Subjective Effects in Binaural Hearing. J. Acous. Soc. America, vol. 22, no. 1, 1950, pp. 61-62.

17. Zurek, P. M.: Measurements of Binaural Echo Suppression. J. Acous. Soc. America, vol. 66, no. 6, 1979, pp. 1750-1757.

18. Bronkhorst, A. W.; and Plomp, R.: The Effect of Head-Induced Interaural Time and Level Differences on Speech Intelligibility in Noise. J. Acous. Soc. America, vol. 83, no. 4, 1988, pp. 1508-1516.

19. Hartmann, W. M.: Localization of Sound in Rooms. III: Onset and Duration Effects. J. Acous. Soc. America, vol. 80, 1986, pp. 1695-1706.

20. Fowler, H. W.: Modern English Usage. Second ed. Oxford University Press, 1965.

NASA
National Aeronautics and Space Administration

Report Documentation Page

1. Report No.
NASA TM-102826

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Technical Aspects of a Demonstration Tape for Three-Dimensional Sound Displays

5. Report Date
October 1990

6. Performing Organization Code

7. Author(s)
Durand R. Begault and Elizabeth M. Wenzel

8. Performing Organization Report No.
A-90162

10. Work Unit No.
505-67-01

9. Performing Organization Name and Address
Ames Research Center
Moffett Field, CA 94035

11. Contract or Grant No.

13. Type of Report and Period Covered
Technical Memorandum

12. Sponsoring Agency Name and Address
National Aeronautics and Space Administration
Washington, DC 20546-0001

14. Sponsoring Agency Code

15. Supplementary Notes
Point of Contact: Durand R. Begault, Ames Research Center, MS 262-2, Moffett Field, California 94035-1000, (415) 604-3920 or FTS 464-3920

16. Abstract

This document was developed to accompany an audio cassette that demonstrates work in three-dimensional auditory displays, developed at the Ames Research Center Aerospace Human Factors Division. It provides a text version of the audio material, and covers the theoretical and technical issues of spatial auditory displays in greater depth than on the cassette. This report also documents the technical procedures used in the production of the audio demonstrations, including methods for simulating rotorcraft radio communication, synthesizing auditory icons, and using the *Convolvotron*, a real-time spatialization device.

17. Key Words (Suggested by Author(s))
Cockpit displays, Spatial sound, Audio, Speech communications, Auditory icons

18. Distribution Statement
Unclassified–Unlimited

Subject Category–53

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
22

22. Price
A02

For sale by the National Technical Information Service, Springfield, Virginia 22161