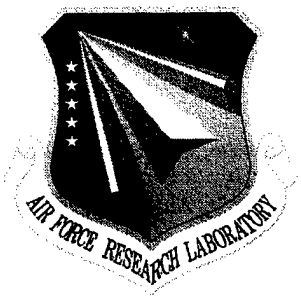
AFRL-IF-RS-TR-1999-148
Final Technical Report
July 1999



COMINT AUDIO INTERFACE

SRI International

David M. Morgan

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-1999-148 has been reviewed and is approved for publication.

APPROVED: Sharon M. Walter

SHARON M. WALTER
Project Engineer

FOR THE DIRECTOR: John V. McNamara

JOHN V. MCNAMARA, Technical Advisor
Info & Intel Exploitation Division
Information Directorate

If your address has changed or if you wish to be removed from the Air Force Research Laboratory Rome Research Site mailing list, or if the addressee is no longer employed by your organization, please notify AFRL/IFEC, 32 Brooks Road, Rome, NY 13441-4114. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 1999	Final Jun 98 - Jan 99

4. TITLE AND SUBTITLE

COMINT AUDIO INTERFACE

5. FUNDING NUMBERS

C - F30602-94-D-0055/07
PE - 35885G
PR - 1039
TA - QK
WU - 07

6. AUTHOR(S)

David Morgan

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

SRI International
333 Ravenswood Avenue
Menlo Park CA 94025

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/IFEC
32 Brooks Road
Rome NY 13441-4114

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AFRL-IF-RS-TR-1999-148

11. SUPPLEMENTARY NOTES

AFRL Project Engineer: Sharon M. Walter/IFEC/(315) 330-7890

12a. DISTRIBUTION AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This document represents the results of the COMINT Audio Interface study. The objective was to investigate 3-dimensional audio technology for application in the military linquist's work environment. Demonstrations conducted under this effort concluded that 3D audio localization techniques on their own have not been developed to the point where they achieve the fidelity necessary for the military work environment. Recommended areas for additional research in human audition, acoustic environment simulators, and invidualized Head Related Transfer Function filters were defined.

14. SUBJECT TERMS

3-Dimensional Audio, spatial audio, localization cues, head tracking

15. NUMBER OF PAGES

24

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

CONTENTS

I. Introduction

The objective of the COMINT Audio Interface project was to investigate three-dimensional (3D) audio technology for future application in the military linguist's work environment. Sound (or audio) localization in three-dimensional space, known in the multimedia community as spatial audio, is sound presented over headphones that has been processed to give the listener the impression of originating from a point in space a programmed direction and distance from the listener.

Audio localization may assist linguists in monitoring simultaneous, multiple channels of communications. It may further enhance individual channel intelligibility and improve the linguist's overall situational awareness.

Demonstrations conducted during this project led to the conclusion that 3D audio localization techniques have not been developed to the point of fidelity necessary for the military linguist's environment. Technical development has primarily targeted virtual environment applications for personal computer and game machine markets, combining spatial audio with visual cues and the simulation of motion. Improved simulation of the acoustic environment, including reverberation effects from simulated walls, may provide the needed additional 3D audio fidelity. Discussions with a 3D audio technologist indicated that further research is developing technology that will more closely match the military requirement for 3D audio without the need for visual or motion cues. That technology may be available for evaluation late in 1999.

Section II of this report provides background on human auditory processing. Section III describes the operational setting and tasking of the military linguist who is monitoring simultaneous, multiple communication channels. Section IV describes the activities conducted for the effort. Section V includes the author's conclusions, and Section VI contains the author's recommendations for follow-on activities.

II. Human Auditory Processing

In the 1950's the term "cocktail party effect" was coined to describe the ability to determine the sources of sounds in a multi-source acoustic environment. It is the human ability to spatially locate sounds that plays a major role in that effect. The problem is also defined as "one world, two ears". That is, we have but one acoustic stimulus (the complex sound field) and we use our two ears to sort out the sound sources that constitute that acoustic stimulus. The study of how we perceive sound is called psychoacoustics.

There are five basic qualities of any sound: pitch (frequency), loudness (amplitude), direction (left/right, front/back, high/low), distance (the listener's distance from the sound source), and timbre (complexity of a sound changing through time). The characteristics of these four as they relate to this study are as follows.

- Pitch - The range of human hearing is approximately 20-20,000 vibrations per second (Hertz).
- Loudness - Sound loses energy as it travels away from its source. Low frequencies (20-100Hz) travel further (with the same amount of initial energy) than higher frequencies.
- Direction - A sound coming from your left has less energy (amplitude) by the time it reaches your right ear.
- Distance - This gives the listener some spatial sense. It helps the listener visualize the listening environment. Primary components are reverberation time and timbre.
- Timbre - the harmonics of the multiple frequency components of a single sound through time.

Each of these sound qualities can be impacted by the environment of the listener and by his/her mental interpretation of sounds. The following are offered as examples:

- A listener's head creates an acoustic shadow which blocks frequencies above 1KHz, further attenuating these frequencies as they travel across the head. This impacts loudness and establishes how we determine direction.
- Head position in relation to sound source (phase angle) affects low frequency perception more than loudness does.
- Our minds filter out sounds deemed unimportant (noise, clicks, etc.).
- We favor expectation over surprise in tracking sounds.
- A human listener can distinguish (subconsciously) a 1/10,000th second difference between the same sound entering one, then the other, ear.

A further complication derives from the way in which humans organize simultaneous auditory events. We use heuristics to segregate sounds into streams representing an auditory event. If, for example, several components continue and a new sound is added, then the new component probably belongs to that original group or event. The main factors affecting how we group sounds include:

- Harmonics - We tend to group components that are harmonics of the same fundamental (such as a chord).

- Patterns - Components of natural sounds may start together and follow a similar pattern. We also tend to group sounds that have equal (or near equal) onset/offset times, or that are subject to the same amplitude or frequency modulations.
- Location of source - Sounds that can be interpreted as coming from the same spatial location tend to group.

III. Linguist Operator Environment and Tasking*

Having a clear picture of what is going on in the signal environment may require an operator to maintain an ongoing awareness of several activities or events. That is not an easy task in a speech environment characterized by poor quality signals, and speech segments averaging only two to four seconds in duration. Often an operator may have to process one or more transmissions on an individual channel to glean enough information to make any reasonable decision on identification or the importance of some communication.

Operators must monitor multiple channels of communications. The ability to divide attention among multiple speech inputs is a skill perfected by very few operators. Complicating the task is the physical limitation of "left ear, right ear, both ears" selections of the headphones in today's platforms.

The linguist/operator's environment is becoming more complex. New technical capabilities are making vast amounts of supporting information available to each operator's position. Communicating that information to the operator has created an almost unmanageable hands-busy, eyes-busy environment.

In the targeted operational environment, operators search for certain essential elements of information (EEI) in audio communications. The operator performs on-line, live 'gisting', detecting the specific terms and phrases that are of interest, in tactical communications. Tasking consists of searching for, recording, and monitoring a number of known channels, and possibly searching for potential new channels of interest. In addition, there is a site or platform intercom channel that the operator will want to monitor to remain cognizant of other operators' activities. There are occasions when an operator will have more than one channel of interest active at the same time. When this occurs, the operator will have difficulty filtering the multiple communications accurately while trying to maintain some level of awareness of the activities that are occurring on each channel.

The operator's input comes through stereo headphones. This provides the operator with options of using:

- One ear for intercom and one for channel monitoring, or
- One ear for each of two monitored channels, or
- One ear for each of two monitored channels and a third channel balanced between both ears.

None of these options are optimal. The second and third options offer the ability to focus on a limited number of multiple incoming active channels, but they exclude intercom availability unless it is fed through with another channel. Potentially valuable intercom information on other activities may be important in helping the operator recognize an intelligence event.

---

* The Principal Investigator for this effort was a military linguist/operator for twelve years. His knowledge of that environment was used as a baseline against which the initial investigations into human psychophysics and the demonstrations of the technology were directed.

IV. Three-Dimensional Audio Evaluation Activities

The AFRL Green Flag database of audio communications collected during a military exercise was used in this evaluation of 3D audio systems. The database consists of nearly three thousand recorded transmissions of airfield area activities including taxi, ready chatter, takeoffs, approaches, landings, runway assignments, and occasional communications checks. The terminology appropriate for essential elements of information (EEI) in this data set included call-signs, location place names, and the numbers and types of aircraft in formations. The only additional information, which might be used, was the occasional order to change communications channels. Separate files, each containing a set of transmissions representing a channel of activity, were prepared from the Green Flag database.

There are three basic characteristics of this database that made it less than ideal for experimentation on this effort. First, only one controller call-sign exists in all conversations. In the real-world environment each monitored channel has a different controller call-sign associated with it. Having the same call-sign show up in conversations coming from multiple, separate locations in our experiments might have created confusion for the operator. Second, due to the fact that these recordings were taken over a period of multiple days of military exercise activity, a number of different speakers (on-duty controllers) identified themselves by the controller call-sign. Therefore, there is only one controller call-sign but multiple speakers identify themselves by that call-sign. Again, multiple speakers identifying themselves by the same name creates a potential for confusion, not enhanced intelligibility. Finally, no timing information was provided with the files. Time delays between transmissions (based on the type of activity being conducted) had to be created.

Proposed procedures for this effort included designing and conducting experiments to produce data on operator performance using spatial audio technology in the course of performing simulations of their normal tasks. The range of experiment complexity proposed was from simple experiments, designed to address the general operating characteristics of the technology, to operations-oriented experiments designed around sets of communications data as they might appear in an operational work environment. Real operator tasks were to be modeled. The primary objective in each experiment was to capture measurable performance data. In the end, however, technical demonstrations were conducted instead of controlled experiments. These demonstrations were designed to show the capability of the 3D audio system to produce effects thought to be most beneficial in the target application. Those effects are:

- identification of at least four sound source locations at any operator selected positions in space around the listener
- a high level of localization fidelity to potentially enhance an operator's ability to extract information from multiple conversations

The demonstrations were not suitable for producing the controlled data product of an experiment, but they did provide observers with a good picture of the current capabilities and exposed some basic weaknesses of audio localization technology.

During the project kick-off meeting it had been suggested that hearing test data might be appropriate for the participants to aid in baselining every possible aspect of the

operators' performance in the tests. A brief Internet search for information on audiometric testing for this effort, however, found that audiometric testing considers normal hearing to be within a 15-20 decibel range for only a limited set of frequencies and has a resolution accuracy of no better than 5 decibels. Localization of complex signals such as speech is based on information integrated across the frequency spectrum, making it unlikely that sensitivity at single frequencies is an accurate predictor of localization performance. Based on this information, individual hearing tests for evaluation participants were not recommended.

Initially, the capabilities of a government-furnished audio spatialization unit were evaluated. At a minimum the operator task requires four simultaneous inputs. As the government-furnished unit permitted only two simultaneous inputs, it was found to be unsatisfactory for this effort.

A search for commercially available technology led to systems developed by Crystal River Engineering (CRE). Internet Web pages for multimedia laboratories at both Georgia Technical Institute, and Massachusetts Institute of Technology stated that the only realistic 3D audio units available were those of CRE. CRE's first stand-alone 3D audio system, developed for NASA, was called the CONVOLVOTRON. The unit is no longer produced but SRI's Virtual Reality Laboratory provided one for use in some limited technical demonstrations. Before it became known that a CONVOLVOTRON could be made available by SRI's Virtual Reality Laboratory, marketing literature, on-line information, and discussions with company marketing personnel described the Acoustetron II by CRE as a system that might satisfy this investigation's requirements. Evaluations of both the CONVOLVOTRON and the Acoustetron II were performed.

A. Acoustetron II Evaluation

The Acoustetron II was advertised as being controllable with SUN, SGI, or PC platforms. However, the PC version worked in a limited demonstration mode only. We would not be able to control any element of the system's operations to use our own data, or to set up our own spatial configurations. Communications with the developers confirmed that the original version of the unit worked with a PC, but revealed that later versions did not have the required updated software.

Additional communication with CRE technical personnel identified the need to use the Acoustic Room Simulation (ARS) package to produce echo/reflection effects. CRE representatives also noted that sound source movement would enhance the listener's perception of localization. Since motion was not our one of our objectives, the sound sources were given the slightest movement selectable, which we described as "dithering." The ARS permitted some control over the size of the virtual space that the listener and the sound source(s) were in, and simulated reflective wall surfaces made of various materials. Without using ARS, no localization effect was achievable with the system.

Final demonstrations of the Acoustetron II were held at SRI in State College, Pennsylvania. Participants included E.J. Cupples, the AFRL/IFEC Speech Program Manager; Sharon Walter, the Laboratory Project Manager for this effort; David M. Morgan, the SRI Principal Investigator, and James Grimplin, the SRI Engineer responsible for the Acoustctron II system set-up. Demonstrations used an ARS acoustically-simulated environment to produce the following auditory events:

- a single spatialized sound, listener placed center, source dithered,
- two spatialized sounds, listener placed center, source dithered,
- one sound from two locations less than 90 degrees apart, listener center,
- one sound source, listener off center in room simulation, source dithered.

The Acoustetron II system used in our evaluation consisted of:

- a 486DX2 based host system,
- 320Mbyte of wave file storage,
- 8Mbyte wave file playback memory,
- an Acoustic Room Simulator (ARS) package,
- 3 Head Related Transfer Function (HRTF) filters,
- 4 Motorola 56001 DSPs (80 million operations per second (MIPS)),
- and an ethernet connection.

Small sets of transmissions from the Green Flag files were used for this demonstration. Timing information was limited to a standardized delay between all transmissions. Start time was slightly staggered for each set, and the duration of each transmission supplied the other timing differences. Each small set of transmissions was used to represent a different channel of activity.

Each participant listened to every variation of the simulated environment. Some demonstrations were listened to multiple times with discussion periods between each session.

Throughout the evaluation at SRI the unit seemed to provide some level of separation, but only with moving sound sources. The actual sense of separation differed from participant to participant. In addition, the sense of distance from the listener was never achieved for any of the participants. Further contact with CRE technical personnel indicated that this implementation of the technology did not have the processing power to perform to the fidelity required for the military linguists' application. In addition, it was later learned that both the Acoustic Room Simulator (ARS) package and the HRTF filter set for this unit were much reduced in capability from previous versions.

The following are comments as noted by the SRI staff during the initial set-up of the system at SRI, and made by the participants during the observed demonstration of the Acoustetron II:

- System control is awkward with a "programming-like" interface requiring a lot of time to make adjustments.
- Stationary sound placement above 60 degrees or below 40 degrees of center (the listener) is not possible with this system. These areas are known as "trouble cones" and are apparently a known problem in 3D audio simulation.
- Spatialization requires use of the Acoustic Room Simulator for echo and reverberation effects. Without this there cannot be any localization information for the listener.
- Motion is apparently another requirement for high fidelity localization with this implementation of the technology. Either the sound source or the listener must

show movement. Since the contract did not support purchasing a headtracker, the sound source must have motion.

- Placement of a sound source directly in front of a listener eliminates localization information. This set-up apparently leaves the model using the same exact distance to the listener's ears, eliminating HRTF.
- The set of HRTF filters supplied with this system consists of only three human model variations. These may be inadequate for general population use.
- Sense of distance from the listener is poor. Any level of perceived sound source separation occurs more within the listener's head than outside the head.
- At distances of ten feet or more, the speech sounded muffled. This was alleviated by setting the doppler shift to zero.
- Distancing the speech lowers gain but seems to have little effect of perceived separation of sources.
- Some listeners experienced slightly more separation than did others. However, the amount of that effect appeared to be based on the listener concentrating on localizing the sound sources rather than processing the speech contained in them.

B. Convolvotron Evaluation

Test data for the CONVOLVOTRON demonstrations consisted of digital recordings of four different speakers reading various texts. The data consisted of quite clean speech of various signal strengths. Engineering staff at SRI's Virtual Reality Laboratory in Menlo Park performed preparatory activities for the Convolvotron evaluation. Participating in the final demonstration of the CONVOLVOTRON were: Nat Bletter (SRI Menlo Park), James Grimplin and David Morgan (both of SRI State College), and Sharon Walter (AFRL/IFEC).

Demonstrations were designed to produce the following auditory events:

- a moving sound (speech)
- a stationary sound placed at various locations around the listener
- multiple sound sources (up to four) placed around the speaker

The CONVOLVOTRON demonstration system set up in the Menlo Park Virtual Reality Laboratory included:

- PC hosted software for card control
- client/server software
- a two card set with 128 parallel 16x16 multipliers (320MIPS)
- 4 16-bit analog/digital converters
- TMS320C25 and OMS A100 DSPs
- 74 HRTF filters
- 2 acoustic models
- a Polhemus ISOTRAK II headtracker (not part of the CONVOLVOTRON)

The following are comments collected from the three observers during the first day of demonstrations and discussions.

- The unit represents old technology in the 3D Audio field.
- The demonstrations provided an improved sense of spatialized audio over the Acoustetron II demonstrations, but still fell far short of meeting AFRL/IFEC requirements.
- Up, down, and forward sound placement is inadequate.
- Headtracking helped with localization, of course only when the listener's head was in motion.
- According to the SRI system operator (who is experienced with this version of the technology), a sound source cannot be placed directly in front of a listener without a visual cue. This was experienced when he created a moving sound source, which was supposed to circle the listener on a horizontal plane. The source seemed to rise up beyond the listener's field of view as it was passing in front of the listener. According to the system operator, the mind performs an automatic transferal of the sound source to another position out of the field of sight of the listener because it has no visual confirmation of the sound source. With a stationary sound source the mind actually reverses the location to one behind the listener.
- Sense of distance from the listener was still not apparent to all participants, even with head movement.
- Acoustic environment models may be inadequate to provide the desired fidelity without being supported by motion and/or visual cues.
- The distinctly different sound sources may have added to the ability of the system to seem to produce a better level of sound separation than the Acoustetron II.

C. Further Discussion

The following comments are based on discussions conducted on 17 December, 1998 with William Chapin, formerly of CRE and Aureal Semiconductor, Inc., now Engineering Director of AuSIM Engineering Solutions. Most of the comments are Mr. Chapin's responses to questions from others present (same participants, listed above, from the CONVOLVOTRON evaluation).

- Because it was developed later, the Acoustetron II is better technology than the CONVOLVOTRON.
- The ARS package on the Acoustetron II is poor. In earlier versions of the ARS it was possible to simulate echoes from surrounding walls of a virtual room. The Acoustetron II version of the ARS supplies directional radiators, which produce an effect that is not as complete an environment model as the older ARS version produced. Mr. Chapin believes that ARS should be all or nothing. Either do a full room simulation or do nothing at all.
- The Acoustetron II has only one-tenth the processing power of the CONVOLVOTRON, and it uses fewer, shorter HRTFs.

- Generalized HRTFs require the listener to train for that set of measurements/parameters each time the system is used. That may be from four to ten minutes each training session. A new training session may be needed each time there is a significant silent passage in a sound source.
- The listeners in AFRL's target environment would experience some localization confusion until they became accustomed to the generalized HRTF. That could occur each time they left the environment and returned to their workstation. Also, they could become confused when they left the artificial HRTF environment until they became used to their own localization parameters again.
- Individualized HRTFs offer about 95% success rate for sensing the localization effect versus about 70% using generalized models.
- With Aureal's method, an individualized HRTF can be made in about 20 minutes.
- Head motion reinforces locations of static audio cues.
- Head tracking (minimally one degree of freedom) is highly advised. Without it the listener can lose the localization effect when he moves his head.
- For the most accurate localization of sound place the sources along the azimuth because HRTFs for up/down locations are more difficult to measure properly.
- Front-to-back reversal effects (the field of view problem) can be resolved with approximately a five-degree head movement by the listener.
- Aureal now has a new 3D audio product (A3D) selling for $99.00.

Mr. Chapin is currently under contract for the Navy to develop 3D audio for Navy communications officers. That environment was described as more of an open network where a number of communications are present at any one time. The communications officers want to have the ability to place various speakers on the network in locations so that they may better focus on individual activities and prioritize the communications. The Navy wants individualized HRTFs for the officers (approximately 2500 users). Mr. Chapin's first scheduled delivery is for late spring 1999. System specifications will be available on the AuSIM web site (http://www.ausim3d.com) in late January 1999. The base system will provide four sources to up to eight listeners. It will be scalable (more hardware more performance).

V. Conclusions

This section contains the author's opinions on human factors aspects of 3D audio technology, the required data and environment needed for testing and evaluation, and statements on the current state of 3D audio technology based on the limited investigation conducted.

A. The Human Factor

The complexity of human factors associated with realizing spatialized audio makes a case for individualized HRTF over generalized filters. The discovery of just how complex those factors could be was not fully realized until late in the investigation. Revisiting the evaluation of the Acoustetron II as part of analyzing the results made it clear that the observers' wide variations in the perceived performance of the unit were due in part to the degree of match/mismatch between the unit's HRTF filters and the particular listener. Some listeners could perceive some level of localization, while others could not for the same demonstration.

The human listener can be trained for a generalized HRTF. The closer the match to each listener's own precise measurements, the less training time required to adopt the new filter as their own. Unfortunately, listeners will require the same HRTF "reprogramming" each time they are removed from that HRTF environment and come back again. During the reprogramming time, the operator's capabilities for 3D processing will be at a lower level.

There is also a strong argument for using head tracking as part of the overall implementation of 3D audio in the linguist's environment. If the spatialized environment moves with the listener's head, the localization effect is lost. Each sound source should stay in its assigned place regardless of the position of the listener's head. This emulates real world audio localization by humans.

There is a significant amount of information available on the ability of human listeners to perform multiple sound source location. However, there are not vast resources providing information on how well individuals can monitor multiple conversations and retain any information from all of the sources. [J. Blauert, 1996], in Chapter 2, "Relation to Other Psychological Mechanisms: Attention," the researchers seemed to prove that there are functionally separate sets of verbal analyzers associated with different streams of input speech. This was done using the same speech in all streams with varied onset times. Then they provided two streams of dissimilar speech, one to each ear. The listener was asked to repeat speech heard from one ear. The listener was unable to report what speech was presented to the other ear. The listener was then "conditioned" by mild electric shock to the presence of certain words in the non-attended channel of input speech. The conditioning worked, and the listener could report the occurrences of those words in the non-attended channel.

Humans have a limited set of cognitive resources used for speech. Those resources are used for both the production and the processing of speech. In the experiments by Blauert, the listener was also a speaker, repeating the speech heard in the target channel. In effect, the experiment caused the participant to use additional

cognitive resources to produce speech while monitoring (processing) speech. The value of Blauert's experiments is that they proved that human attention capabilities may be expanded, or at least focused, through conditioning, not that it is a natural human trait.

There is little question that the function of multiple channel monitoring can be performed. There are still questions as to whether all listeners can be equally capable of multiple channel monitoring, and how long it will take to train ("condition") a listener to attain this skill.

B. Test Data and Environment

Testing technology in the exact environment into which it might be implemented is extremely valuable. However, that is not always an option, nor is it a must for success. After conducting the demonstrations with Green Flag data, and later with four totally different channels of clean, recorded speech, it seemed that the clean, easy to process speech allowed the participants to concentrate on the effects of the technology more than trying to discern the context of unfamiliar, poor quality speech. The clean speech approach may be best for initial evaluation of the technology if a true operational environment is not possible.

The Green Flag data is not satisfactory for this type of testing. Many characteristics of the data set can create conflicting cues for a human listener. The data in each channel should consist of completely separate conversations, by separate speakers, using different call-signs or other forms of identification.

C. State of the Technology

The technology evaluated for this investigation uses multiple techniques to provide the effect of spatial audio. If these various techniques are not used in combination, the fidelity of the 3D audio effect diminishes tremendously. Unfortunately, today's (and the near term) operational environment is not in a position to provide visual cues, and motion cues do not seem appropriate for the targeted application.

The ability to place sound sources in space without visual cues, and without moving the listener's head or the sound source, is necessary for AFRL's application. A similar effect can be realized today in high-end home audio systems through two speakers. This is called imaging, or stage presence, and allows the listener to locate a singer, or individual instrument in a musical group, within the listening room. There are no visual cues. Motion is not required. The listener must often sit in one particular area of the room to realize this effect. A different room will have a different "sweet spot" based on the size and materials of the room. This may be a clue to the problem of 3D technology today. The Acoustic Room Simulator packages are not providing adequate models (simulations) of true room effects. Rather, they are assisting 3D with visual- and motion-generated cues.

Ongoing development in this technical area seems geared toward the game machine and PC markets. Lowered cost and the ability to use multiple, combined techniques to provide the system user with the experience of 3-dimensional audio within a more complete virtual environment is driving this.

VI. Recommendations

In the following paragraphs, the Principal Investigator suggests a number of follow-on actions based on the results of this investigation. These suggestions include enhancing awareness of the target operating environment for the technology, defining the desired operating characteristics of a 3D audio system, and pursuing further technical research and development in recommended areas.

A. Application Development

The user environment should be documented and the technology need precisely defined. It would not be prudent to further the development of any areas of this technology without first knowing the specifics of the need. This may be conducted in two ways. First, as recommended in Section C below, the work done for the Navy for its communications officers may be reviewed to determine their specific user requirements and characteristics of their operational environment. If the fit is close enough to the Air Force linguists' operating environment, use those requirements and environment characteristics to conduct further experiments. Or, the Air Force COMINT operator environment may be analyzed for potential applications of 3D audio. A clear understanding of the limitations of human capabilities in a 3D environment, as referenced in Section C below, is a necessary precursor to either of these options.

B. Desired Technical Capabilities

AFRL/IFEC should try to determine some basic operating characteristics the technology should have. Without conducting an extensive study on the technology, suggested desirable operating characteristics for a 3D audio system implemented in the COMINT environment include:

- Provide operator-controlled placement of sound sources. The range of optimal or comfortable placement may differ from operator to operator. Also, an operator may want to group sound sources in certain regions based on their functions, real-world locations, or interest potential. This could be a dynamically changing environment requiring frequent changes of sound source locations, elimination of some sources, and/or additions of new sources as a mission progresses.
- Operators should be able to save their favorite sound source placements as defaults for various multi-channel scenarios.
- Provide a user-friendly interface for the operator to support all of the operations of sound placement.
- There must be a tie between the sound sources and individual gain control for the channels so that an operator can adjust gain levels for each channel independently if needed.

- High fidelity in the localization cues of sound sources is a must. Operators may want to maintain multiple target sources in one quadrant of the virtual space around them. Highly accurate cues would support this.
- High quality (one-degree-of-freedom) head tracking
- On-line individualized HRTF training

C. Technical Areas for Further Investigation

Acoustic Room Simulators - This seems to be a prime area for further investigation. It is recommended that AFRL/IFEC follow up on the comments made by William Chapin regarding the availability of better packages than the one used with the Acoustetron II. If they do not prove adequate, then the idea of further development of an existing package, or the development of a custom package may be appropriate.

Individualized Head-Related Transfer Functions (HRTFs) - Aureal seems to have a lock on the market for developing individualized HRTFs. The Principal Investigator on this effort believes that individualized filters are the best way to avoid potential operational problems when training and retraining with generalized HRTFs. One method of investigation would be to obtain the system being developed for the Navy this year by AuSIM, and have Aureal build individualized HRTFs for a small set of investigators to evaluate that system for the AFRL-targeted application. The system may not be the exact answer to the Air Force application, but it could answer some questions on Individualized HRTFs and the use of acoustic environment simulations.

Human psychophysics - There is still much that needs to be understood about human reaction and performance in a 3D environment. This effort was not intended to be an in-depth study of the human element in all of this, but the brief investigation that was performed identified a number of questions which do need answers. A prime example is: determining the human capacity for multiple conversation retention. If human performance is limited, the applicability of 3D audio in the operator's environment will be also. Some answers may be available from other laboratories researching human performance in multiple audio input environments, virtual reality environments, and effects of various physiological conditions on human performance in these environments. This information should be gathered and analyzed to complete a human factors profile for the operating environment of the linguist.

References:

Blauert, Jens, "Spatial Hearing: The Psychophysics of Human Sound Location (Revised Edition)," MIT Press, 1996.

Morgan, David M., "Concept of Operations for Speech Processing Technologies," HRB Systems, Inc. Unpublished Report, January 20, 1995.

Yost, W.A., A.N. Popper and R.R. Fay (eds.), "Springer Handbook of Auditory Research, Volume 3: Human Psychophysics," Springer-Verlag, New York, 1993.

MISSION OF AFRL/INFORMATION DIRECTORATE (IF)

The advancement and application of information systems science and technology for aerospace command and control and its transition to air, space, and ground systems to meet customer needs in the areas of Global Awareness, Dynamic Planning and Execution, and Global Information Exchange is the focus of this AFRL organization. The directorate's areas of investigation include a broad spectrum of information and fusion, communication, collaborative environment and modeling and simulation, defensive information warfare, and intelligent information systems technologies.