

MONITORING THE SIMULTANEOUS PRESENTATION OF SPATIALIZED SPEECH SIGNALS IN A VIRTUAL ACOUSTIC ENVIRONMENT

W. Todd Nelson, Ph.D.
Engineering Research Psychologist
Crew System Interface Division
Air Force Research Laboratory

Robert S. Bolia
Research Scientist
Veridian

Mark A. Ericson
Electrical Engineer
&
Richard L. McKinley
Biomedical Engineer
Crew Survivability and Logistics Division
Air Force Research Laboratory

Abstract

The effect of spatial auditory information on a listener's ability to detect, identify, and monitor multiple simultaneous speech signals was evaluated using virtual audio technology. Factorial combinations of three variables - the number of localized speech signals, the location of the speech signals around the horizontal plane, and the sex of the talker - were employed using a within-subjects design. Participants were required to detect the presentation of a critical speech signal among a background of non-speech events. Results indicated that the spatialization of simultaneous speech signals (1) increased the percentage of correctly detected and identified critical speech signals and (2) did not affect the response times of correctly detected signals. Implications for the use of this technology as an applied interface are discussed.

Introduction

The advanced capabilities of modern fighter aircraft, in conjunction with the burgeoning complexity and lethality of future air combat environments, are likely to place significant limits on the abilities of pilots and crew members to perceive effectively the critical information presented on aircraft displays. In addition, as noted by several researchers (Brickman, Hettinger, Haas, & Dennis, 1998;

Furness, 1986; Haas & Hettinger, 1993), continued advances in weapons technology and aircraft performance are likely to be associated with concomitant increases in the perceptual, perceptual-motor, and cognitive demands placed upon pilots. Accordingly, there is a compelling need to develop interfaces that will be able to compensate for these effects, thereby enabling pilots and crew members to operate effectively in these challenging environments.

One potential way to offset the problems caused by these factors would be to exploit the human operator's ability to perceive and process spatial auditory information. Along this line, numerous researchers (Begault, 1993; Begault & Pittman, 1996; Bronkhorst, Veltman, & van Breda, 1996; Doll, 1986; McKinley, Ericson, & D'Angelo, 1994; Nelson, Hettinger, Cunningham, Brickman, Haas, & McKinley, in press; Perrott, Cisneros, McKinley, & D'Angelo, 1996) have demonstrated that spatialized auditory displays increase performance efficiency for a variety of tasks that are relevant to airborne applications, including target detection and identification, navigation, and collision avoidance.

For example, Bronkhorst, Veltman, and van Breda (1996) recently demonstrated that 3-dimensional auditory displays may be effective in providing pilots with directional information about the location of targets. In brief, participants (Royal Netherlands Air Force pilots) used a simulated F-16 aircraft to follow an F-18 target aircraft that would suddenly disappear and then reappear at an unknown location.

Participants' main task was to locate and follow the target aircraft as quickly as possible under one of four display conditions; (1) no display (2), 3-D auditory display, (3) visual display (bird's-eye-view radar), and (4) 3-D auditory and visual display. Results indicated that mean search time was significantly shorter when the 3-D auditory display and visual display were combined as compared to the visual-only, 3-D auditory-only and the no display conditions. As Bronkhorst and his colleagues (1996) pointed out, such an outcome is consistent with the idea that multi-sensory presentation of spatial information can serve to enhance performance efficiency.

Displays that provide spatialized auditory information may also afford more efficient segregation, monitoring, and attentional shifts among speech signals that are presented simultaneously. This notion is based, in part, upon the recognition that the spatial separation of acoustic signals improves the intelligibility of signals in noise and assists in the segregation of multiple sound streams, the so-called "cocktail party effect" (Bregman, 1990; Bronkhorst & Plomp, 1988; Cherry, 1953; Yost, Dye, & Sheft, 1996). As noted by Yost and his colleagues (1996), spatial hearing plays an important role in tasks that characterize the "cocktail party effect," especially when more than two speech signals are presented simultaneously.

Empirical evidence along these lines has been provided by Ricard and Meirs (1994). Their experiment was designed to determine what impact, if any, localized speech might have on the intelligibility and localization of speech signals. In regard to the latter, their data suggested that the accuracy of localizing speech stimuli was comparable to that of non-speech stimuli presented via headphones using non-individualized head-related transfer functions (HRTFs). Similar results have been reported by Begault and Wenzel (1993). In addition, Ricard and Meirs showed that when localized single-word stimuli were presented in the presence of a masking white noise, intelligibility increased by an average of 4 to 5 dB relative to that of non-localized speech stimuli.

Recently, Ericson and McKinley (1997) investigated the effects of wide band noise and the separation of multiple talkers on the intelligibility of a call sign phrase. Wide band noise consisted of correlated diotic pink noise, uncorrelated pink noise, and ambient pink noise. The diotic and dichotic pink noise were mixed with speech signals and presented to the listeners over headphones, while the ambient

pink noise was played over loudspeakers in a reverberant chamber. Sex of talker was an experimental variable and competing talkers consisted of either one or three same or mixed sex talkers. Ericson and McKinley (1997) found that angular separation greater than or equal to $\pm 45^\circ$ provided the highest levels of intelligibility. In addition, diotic noise was associated with the greatest degradations in speech intelligibility, followed by ambient noise and dichotic noise, respectively. Finally, same female, same male, and mixed sex were found to be the most to least degrading on speech intelligibility.

Given the evidence just described, it is reasonable to expect that spatially-separated speech may enhance the effectiveness of speech communications in noisy or competing message environments. - i.e., aircraft operations such as cockpit communications, air traffic control, and AWACS applications (Ericson & McKinley, 1997, Wenzel; 1992) Moreover, displays that present *spatialized* speech may potentially benefit airborne applications by: 1) increasing performance efficiency; 2) lowering operator workload; and 3) enhancing situation awareness.

The primary objective of this research project was to assess the effects of 3-D auditory information on an operator's ability to detect, identify, and monitor the presentation of a critical call sign phrase among multiple simultaneous speech signals. Toward that end, the research presented hereinafter describes the results of an empirical investigation aimed at assessing the effects of numerous factors that are believed to influence a listener's ability to effectively perceive spatialized speech signals - specifically, the number of simultaneous signals, the location and spatial separation of the speech signals, and the sex of the talker. To date, investigations of this sort have been extremely sparse (Koehnke, Besing, Abouchacra, & Tran, 1998; Ricard & Meirs, 1994; Yost et al., 1996); hence, it is anticipated that research of this sort will be useful to researchers and interface designers at both the basic and applied levels.

Method

Participants

Four males and four females, naïve to the purposes of the experiment, served as paid participants. Their ages ranged from 19 to 47 years with a mean of 29 years. All participants had normal hearing and normal localization acuity.

Experimental Design

Five spatialization conditions (front right quadrant (RQ), front hemifield (FH), right hemifield (RH), full 360° (F), and a non-spatialized control (C)) were combined factorially with eight talker conditions (1, 2, 3, 4, 5, 6, 7, or 8 talkers) and the sex of the critical speech signal (male and female) to provide 80 experimental conditions. Participants completed all combinations of the experimental conditions with the constraint that each of the five spatialization conditions was performed during a separate experimental session.

Apparatus

Facility. The experiment was conducted at the Air Force Research Laboratory's Auditory Localization Facility (ALF) - a geodesic sphere of radius 2.3 m housed within a cubic anechoic chamber of volume (6.7 m)³ - at Wright-Patterson Air Force Base, Ohio.

Recording and Playback. Four males and four females served as talkers. All of the talkers were recorded saying each of the 256 possible phrases in the modified Coordinated Response Measure (CRM; Moore, 1981). Speech signals were recorded in a quiet room using a Tucker-Davis DD1 combined analog-to-digital/digital-to-analog converter at a sampling rate of 40 kHz. Each of the 2048 signals was then high-pass filtered at 100 Hz, low-pass filtered at 8 kHz, and scaled such that all of the phrases had the same average power. Additionally, all silence was removed from the beginning of each waveform for purposes of synchronization. Simultaneous playback of up to eight phrases was achieved using the Tucker-Davis DA3-8, an eight-channel digital-to-analog converter.

Virtual Auditory Display. Spatialization of the signals was achieved using two of the Air Force Research Laboratory's four-channel 3-D Auditory Display Generators (3-D ADG) coupled to the same Polhemus 3Space position tracker. The 3-D ADG uses the techniques of digital signal processing to encode naturally occurring spatial information in an audio signal and present the resulting "spatialized" image over stereo headphones (Sennheiser HD-560).

This encoding is accomplished via digitally filtering a sound source by means of an FIR filter created from measurements of a human's head-related transfer functions (HRTFs), which represent the modification of a sound source by

the person's head, torso, and pinnae (see Wightman and Kistler, 1997, for a review of HRTFs). A block diagram of the experimental apparatus is depicted in Fig. 1.

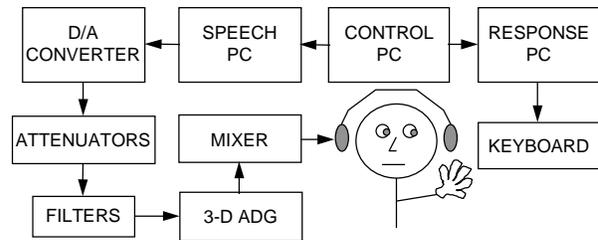
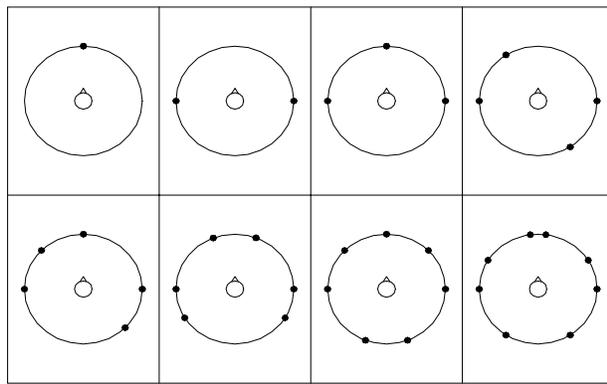


Fig. 1 A block diagram of the experimental apparatus.

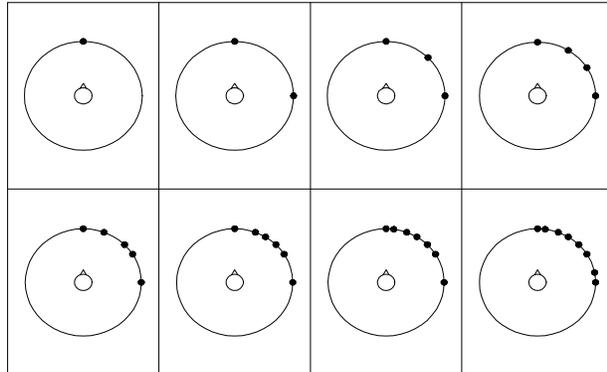
Location of the Speech Signals. In order to be able to compare the data with that obtained in a previous study (Nelson, Bolia, Ericson, & McKinley, 1998), the spatial locations of the potential targets and distractors were constrained in such a way that they mapped onto the location of loudspeakers in the ALF. The algorithm for selection of spatial locations was simple. If, on a given trial, there was only a single talker, the signal always emanated from directly in front of the listener. If there were two or more talkers, the signals were positioned such that the average difference in source-midline distance (SMD) was a maximum for the configuration. If, as might occur in the right hemifield and full 360° conditions, two potential locations had the same average SMD difference, the location was chosen which maximized angular separation. Figures 2a-d show the eight possible target+distractor configurations for each of the spatialization conditions.

Procedure

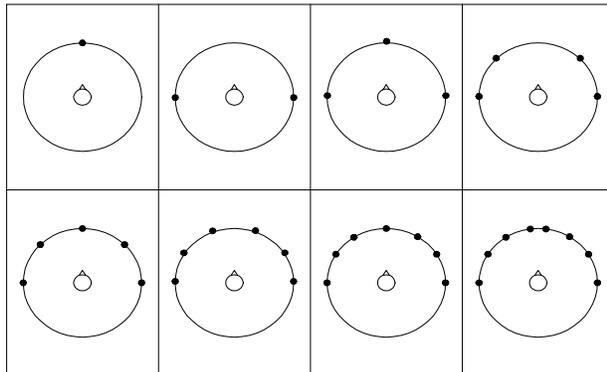
For each trial, between one and eight speech signals were selected from a set of phrases from a modified version of the CRM (Moore, 1981). Each phrase consisted of a call sign (Baron, Ringo, Laker, Charlie, Hopper, Arrow, ...), a color (Red, White, Green, Blue), and a number (1, 2, 3, 4, 5, 6, 7, 8), embedded within a carrier phrase. Phrases were selected at random with the constraints that 1) the target phrase always contained the call sign "Baron," and 2) within a given trial on which a critical signal was present, neither talkers nor call signs were repeated. Hence, in the 8-talker condition, a listener would hear eight different talkers, each uttering a different call sign.



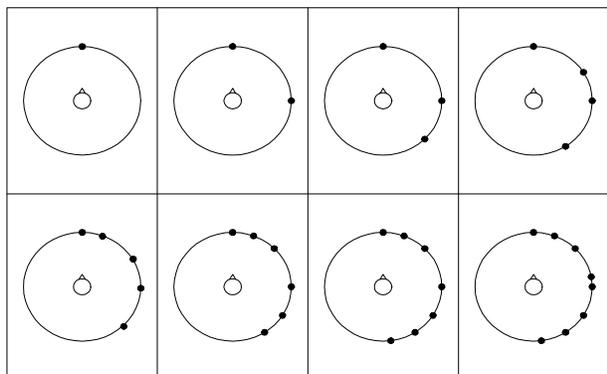
(a) full 360° (F) condition



(b) front right quadrant (RQ) condition



(c) front hemifield (FH) condition



(d) right hemifield (RH) condition

Fig. 2a-d Top down view of the eight possible target and distractor locations for each of the spatialization conditions: a) full 360° (F), b) front right quadrant (RQ), c) front hemifield (FH), and d) right hemifield (RH).

During each trial, participants monitored the simultaneous presentation of multiple spatialized speech signals. Their task was to listen for the occurrence of a critical call sign (“Baron”) and to identify the color-number combination that appeared to emanate from the same spatial location as the critical call sign. This was done by pressing the key on the response device that was of the appropriate color and marked by the appropriate number.

Thus, the appropriate response to “Ready Baron Go To Red Six Now” would have been to press the red key labeled with a number six. If the critical call sign were not present, the listener was required to press the “no-response” key. Fifty percent of the experimental trials included the critical call sign.

Prior to data collection, participants completed five practice sessions, one practice session for each of the five spatialization conditions. Participants rated the perceived mental workload of the task by completing the NASA Task Load Index (Hart & Staveland, 1988) after completing the non-spatialized control (C) and full 360° (F) experimental sessions.

Results

Performance Efficiency

Several indices of performance efficiency were calculated, including percent correct detections, percent errors of commission, response time of correct detections, and percent correct identifications.

Correct Detections (HITS). Mean percent correct detections - i.e., detecting the presence of a critical call sign when it was present - were calculated for all experimental conditions and subjected to a 5 (condition) x 8 (talker) x 2 (sex of critical call sign) repeated measures analysis of variance.

The results of the analysis revealed that the main effects of *CONDITION*, *TALKER*, and *SEX OF CRITICAL SIGNAL* were statistically significant, $F(4,28) = 4.74$, $p < .05$, $F(7,49) = 48.47$, $p < .05$, and $F(1,7) = 8.14$, $p < .05$, respectively. In addition, the Talker x Sex of Critical Signal interaction was statistically significant, $F(7,49) = 3.51$, $p < .05$. All other sources of variance in the analysis lacked significance ($p > .05$).

The Talker x Sex of Critical Signal interaction is presented in Fig. 3, which shows mean percent correct detections plotted for the male

and female spoken critical signals within each of the eight talker conditions. It is evident in Fig. 3 that correct detections varied inversely with the number of simultaneous talkers and that female spoken critical signals were detected more often than male spoken critical signals when four or more talkers were presented simultaneously. Post hoc pairwise comparisons confirmed these impressions ($p < .01$).

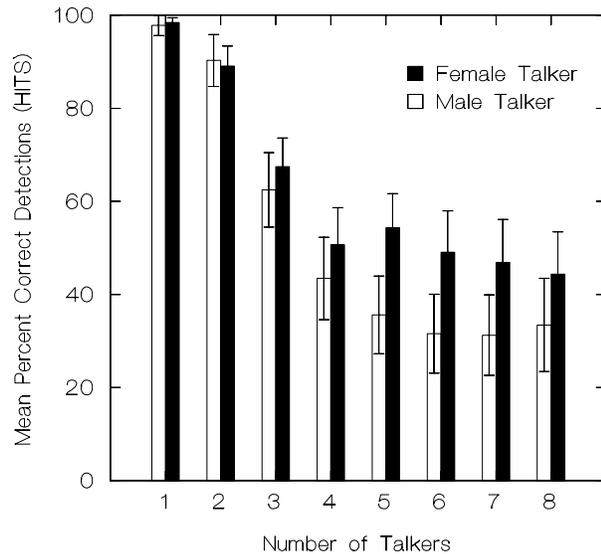


Fig. 3 Percent correct detections for the male and female spoken critical call signs plotted as a function of the number of simultaneous talkers.

The significant main effect for spatialization condition is depicted in Fig. 4, which shows mean percent correct detections under each of the five spatialization conditions. As can be seen in Fig. 4, the spatialized conditions (RQ, FH, RH, F) were associated with higher detection scores as compared with the non-spatialized control condition (C). Post-hoc pairwise comparisons indicated that detection scores associated with the control condition were significantly ($p < .01$) lower than each of the four spatialized conditions. However, the spatialized conditions did not differ from each other ($p > .01$).

Errors of Commission (False Alarms). Mean percent errors of commission - i.e., responses that indicate the detection of a critical signal when in fact none was present - were calculated for all experimental conditions and submitted to a 5 (condition) x 8 (talker) x 2 (sex of critical signal) repeated measures analysis of variance. All sources of variance lacked statistical significance ($p > .05$). The absence of significant main effects and/or interactions implies that the experimental factors did not affect participants' response criteria.

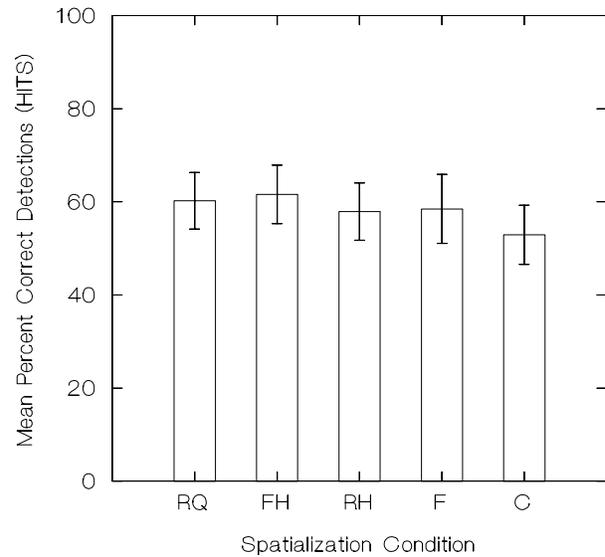


Fig. 4 Mean percent correct detections for each of the five spatialization conditions (RQ = right quadrant; FH = front hemifield; RH = right hemifield; F = full 360°; C = control)

Response Time. Mean response times of correctly detected speech signals were calculated for all experimental sessions and analyzed using a similar 5 (condition) x 8 (talker) x 2 (sex of critical signal) repeated measure analysis of variance. The analysis of these data revealed a significant *TALKER* main effect, $F(7,14) = 9.06$, $p < .05$, which is illustrated in Fig. 5. None of the remaining components of variance in the analysis were significant ($p > .05$). Perusal of Fig. 5 indicates that response times increased steadily between one and six talkers.

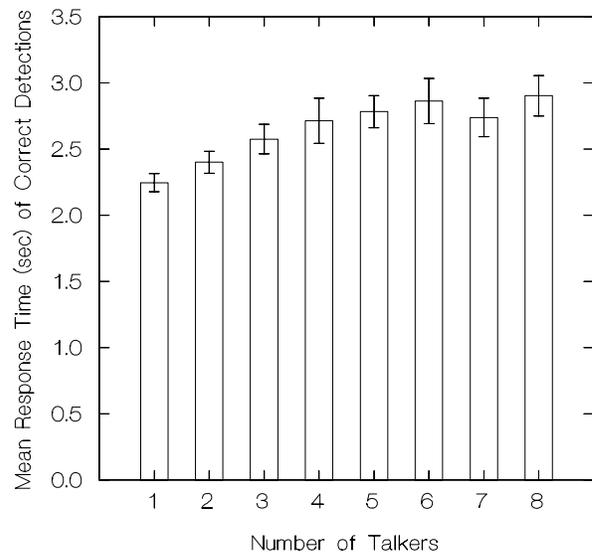


Fig. 5 Mean response time of correct detections as a function of number of simultaneous talkers.

Correct Identifications. Mean percent correct identifications - i.e., correct detection of the call sign and the correct identification of the color and number combination - were calculated for all experimental conditions. These data were analyzed with a 5 (condition) x 8 (talker) x 2 (sex of critical signal) repeated measures analysis of variance, which revealed significant main effects for *CONDITION*, $F(4,28) = 14.78$, $p < .05$, and *TALKER*, $F(7,49) = 580.12$, $p < .05$. All other sources of variance in the analysis lacked significance. The *CONDITION* main effect, which is depicted in Fig. 6, can be explained by noting that identification scores associated with each of the four spatialized auditory conditions (RQ, FH, RH, F) were superior to the non-spatialized control (C). These impressions were supported by post hoc pairwise comparisons ($p < .01$). However, the scores associated with the four spatialized conditions did not differ significantly from each other ($p > .01$).

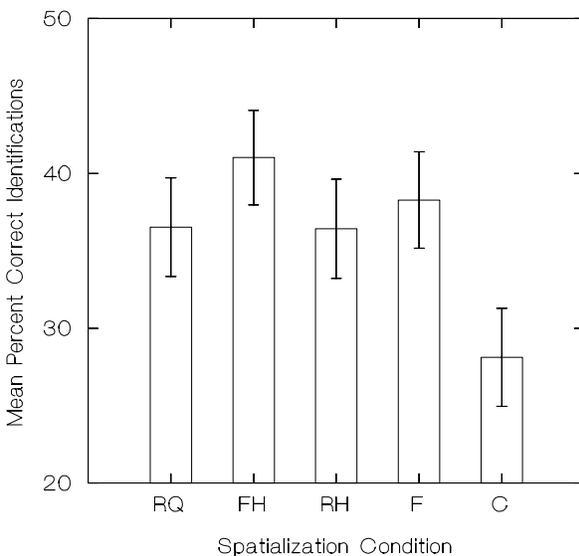


Fig. 6 Mean percent correct identifications for each of the five spatialization conditions (RQ = right quadrant; FH = front hemifield; RH = right hemifield; F = full 360°; C = control)

The *TALKER* main effect is illustrated in Fig. 7. As can be seen in the figure, increases in the number of simultaneous talkers produced dramatic decrements in performance efficiency. Identification scores declined 87.65 % between the one and eight talker conditions, and approximately 88% of this decline occurred across the first four talker conditions.

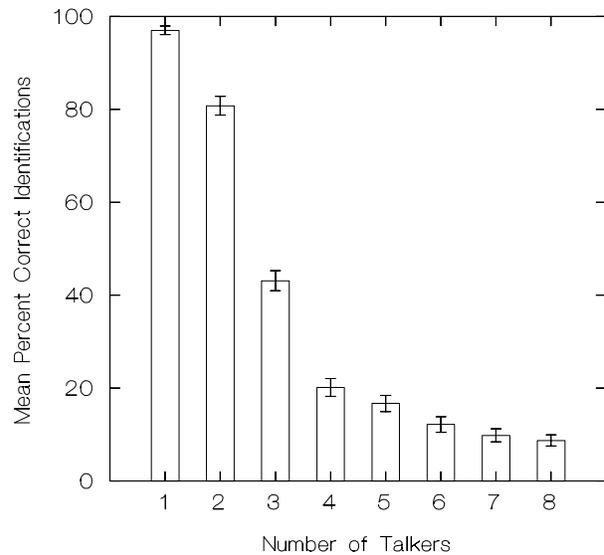


Fig. 7 Mean percent correct identifications as a function of the number of talkers.

Workload Ratings

The NASA Task-Load Index (NASA-TLX; Hart & Staveland, 1988), a multidimensional scale of perceived mental workload, was used to provide subjective estimates of the information processing demands associated with the experimental task. The NASA-TLX provides a global measure of overall workload (on a scale of 0 to 100), and also identifies the relative contributions of six sources of workload: (1) Mental Demand, (2) Physical Demand, (3) Temporal Demand, (4) Performance, (5) Effort, and (6) Frustration.

Overall Workload. Mean overall workload ratings for the full 360° (F) and the non-spatialized control (C) conditions are presented in Fig. 8. As can be observed, overall levels of perceived mental workload were moderate, falling, on average, at the midpoint of the NASA-TLX scale. An analysis of the workload ratings indicated no significant difference between the full 360° (F) and control (C) conditions, $t(7) = .756$, $p > .05$.

Subscales. Mean weighted ratings for the NASA-TLX subscales are presented in Fig. 9 for the full 360° (F) and the non-spatialized control (C) conditions. As can be seen in the figure, the *Effort*, *Performance*, and *Mental Demand* components were associated with the highest weighted ratings. Inspection of the figure also revealed the ratings on the *Effort* and *Temporal Demand* subscales were higher in the full 360° (F) condition as compared with the non-spatialized control (C). These observations were

confirmed with post hoc pairwise comparisons ($p < .05$).

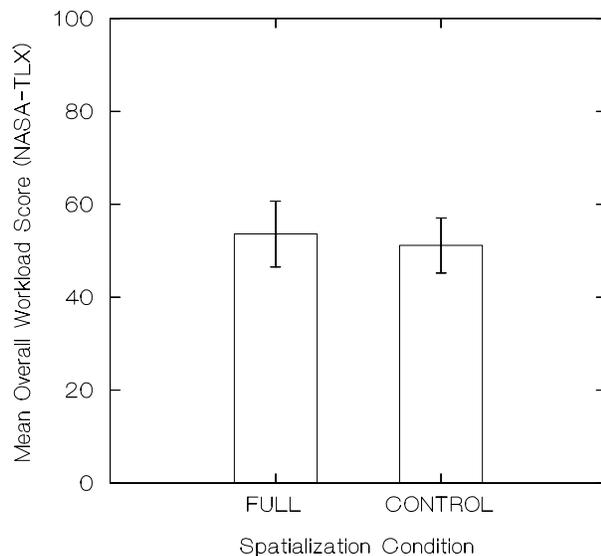


Fig. 8 Mean overall workload scores for the spatialized full 360° (F) and non-spatialized control (C) conditions.

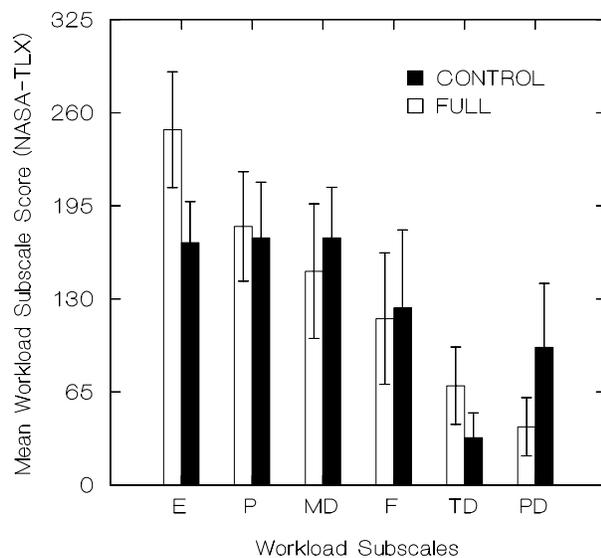


Fig. 9 Mean workload ratings for the six components of workload (E = effort; P = performance; MD = mental demand; F = frustration; TD = temporal demand; PD = physical demand) in the context of the full 360° (F) and non-spatialized control (C) conditions.

Conclusions

The present investigation represents an initial experimental effort to determine the effects of spatial auditory technology on listeners' ability to detect and identify critical speech signals presented in a multi-talker competing message environment

The principle conclusion that emerges from the present experiment is that the *spatialization*

of speech signals enhances one's ability to detect and identify critical speech signals. Detection and identification scores associated with the spatialized speech conditions were significantly higher than those in the non-spatialized control. Moreover, as evidenced by the lack of significant interactions with the other two experimental factors, the *spatialization* effect was not mediated by the sex of the talker or the number of simultaneous speech signals. Finally, the analysis of the false alarm data indicated that the performance benefits associated with the spatialized speech conditions could not be attributed to participants adopting unusually high response rates or lenient response criteria.

In spite of the overall performance advantage for the spatial separation of the speech signals, no significant differences were found between the various spatialization conditions. Such an outcome is important for a couple of reasons. First, given that location per se does not determine the efficacy of the *spatialization* effect, designers will not be constrained by a limited spatial area for displaying speech. Consequently, designers can "place" the speech signals anywhere along the horizontal plane with the assurance that the associated performance benefits will not be jeopardized. Second, from an engineering perspective, the non-specific nature of the spatialization effect may afford significant reductions in the number of HRTFs and associated FIR filters required by the 3-D audio device. For example, the number of HRTFs is reduced by a factor of four when speech signals are restricted to emanating from the right front quadrant as compared to the entire horizontal plane.

While the spatialization of the speech signals enhanced detection and identification scores, no main effects or interactions involving the spatialization factor were revealed for response time to correct detections or workload ratings. In fact, inspection of the workload subscale ratings (see Fig. 9) revealed that the full-360° condition was associated with higher ratings of *Effort* and *Temporal Demand* as compared to the non-spatialized control. Apparently, the performance enhancing effects of the spatial speech, were accompanied by additional information processing demands. Collectively, these results have important implications for the use of spatialized speech interface technology, especially in application domains in which operators are required to issue time-critical decisions in high workload environments.

Acknowledgments

The authors wish to acknowledge the numerous technical contributions of the following Veridian personnel: Dennis L. Allen who provided technical support to all aspects of the experiment and generated figures and illustrations, Michael L. Ward who conducted data collection and assisted with the preparation of the speech signals, and Ronald C. Dallman who provided technical assistance with the virtual auditory display generator.

Author Biographies

W. Todd Nelson is an engineering research psychologist in the Air Force Research Laboratory's Crew System Interface Division, Wright-Patterson Air Force Base, Ohio. He earned the Ph.D. in experimental psychology from the University of Cincinnati.

Robert S. Bolia is a research scientist with Veridian. Currently, he is pursuing a doctoral degree at Wright State University, Dayton, Ohio.

Mark A. Ericson is an electrical engineer in the Air Force Laboratory's Crew Survivability and Logistics Division, Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio. He is a doctoral candidate in hearing science at The Ohio State University.

Richard L. McKinley is the Technical Director of the Crew Survivability and Logistics Division, Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio. He received his M.S. in bioengineering/digital signal processing in 1985 from the Air Force Institute of Technology, Dayton, Ohio.

References

- Begault, D. R. (1993). Head-up auditory displays for traffic collision avoidance system advisories: A preliminary investigation. *Human Factors*, 35, 707-717.
- Begault, D. R., & Pittman, M. T. (1996). Three-dimensional audio versus head-down traffic alert and collision avoidance system displays. *The International Journal of Aviation Psychology*, 6(1), 79-93.
- Begault, D. R., & Wenzel, E. M. (1993). Headphone localization of speech. *Human Factors*, 35(2), 361-376.
- Brickman, B. J., Hettinger, L. J., Haas, M. W., & Dennis, L. B. (1998). Designing the supercockpit. *Ergonomics In Design*, 6(2), 15-20.
- Bronkhorst, A. W., Veltman, J. A. H., & van Breda, L. (1996). Application of a three-dimensional auditory display in a flight task. *Human Factors*, 38(1), 23-33.
- Doll, T. J. (1986). Synthesis of auditory localization cues for cockpit applications. *Proceedings of the Human Factors Society's 30th Annual Meeting* (pp. 1172-1176). Santa Monica, CA: Human Factors Society.
- Ericson, M. A., & McKinley, R. L. (1997). The intelligibility of multiple talkers separated spatially in noise. In R. H. Gilkey, & T. R. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments*. (pp. 701-724). Mahwah, NJ: Lawrence Erlbaum Associates.
- Furness, T. A. (1986). The super cockpit and its human factors challenges. *Proceedings of the Human Factors Society 30th Annual Meeting* (pp. 48-52). Santa Monica, CA: Human Factors Society.
- Koehnke, J., Besing, J. M., Abouchacra, K. S., & Tran, T. V. (February, 1998). Speech recognition for known and unknown target message locations. *Poster presented at the 1998 Mid-Winter Meeting of the Association for Research in Otolaryngology*. St. Petersburg, FL.
- McKinley, R. L., Ericson, M. A., D'Angelo, W. R. (1994). 3-dimensional auditory displays: Development, applications, and performance. *Aviation, Space, and Environmental Medicine, May*, 31-38.
- Moore, T. J. (1981). Voice communication jamming research. *AGARD Conference Proceedings 311: Aural Communication in Aviation* (pp. 2:1-2:6). Neuilly-Sur-Seine, France.
- Nelson, W. T., Bolia, R. S., Ericson, M. A., & McKinley, R. L. (1998). Monitoring the simultaneous presentation of multiple spatialized speech signals in the free field. *Journal of the Acoustical Society of America*, 103(5), 3019.
- Perrott, D. R., Cisneros, J., McKinley, R. L., & D'Angelo, W. R. (1996). Aurally aided visual search under virtual and free field listening conditions. *Human Factors*, 38(4), 702-715.
- Ricard, G. L., & Meirs, S. L. (1994). Intelligibility and localization of speech from virtual directions. *Human Factors*, 36, 120-128.
- Wenzel, E. M. (1992). Localization in virtual acoustic displays. *Presence*, 1, 80-106.
- Yost, W. A., Dye, R. H., & Sheft, S. (1996). A simulated "cocktail party" with up to three sound sources. *Perception & Psychophysics*, 58(7), 1026-1036.
- Yost, W. A. (1997). The cocktail party problem: Forty years later. In Gilkey, R. H. & Anderson, T. R. (Eds.) *Binaural and spatial hearing in real and virtual environments* (pp. 329-347). Mahwah, NJ: LEA.