Facilitating the Watchstander's Voice Communications Task in Future Navy Operations

Derek Brock, Christina Wasylyshyn, Brian McClimens, and Dennis Perzanowski
Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory, Code 5512
Washington, DC 20375
Email: derek.brock@nrl.navy.mil

Abstract— **Recent human performance research at the Naval Surface Warfare Center, Dahlgren Division (NSWCDD) has shown that increasing the number of concurrent voice communications tasks individual Navy watchstanders must handle is an uncompromising empirical barrier to streamlining crew sizes in future shipboard combat information centers. Subsequent work on this problem at the Naval Research Laboratory (NRL) has resulted in a serialized communications monitoring prototype (U.S. Patent Application Pub. No. US. 2007/0299657) that uses a patented NRL technology known as "pitch synchronous segmentation" (U.S. Patent 5,933,808) to accelerate buffered human speech up to 100% faster than its normal rate without a meaningful decline in intelligibility. In conjunction with this research effort, a series of ongoing human subjects studies at NRL has shown that rate-accelerated, serialized communications monitoring overwhelmingly improves performance measures of attention, comprehension, and effort in comparison to concurrent listening in the same span of time. This paper provides an overview of NRL's concurrent communications monitoring solution and summarizes the empirical performance questions addressed by, and the outcomes of, the Lab's associated program of listening studies.**

Keywords: concurrent voice communications; watchstander; serialized communications monitoring; speech rate acceleration; intelligibility; attention; comprehension; effort

I. INTRODUCTION

In late 2001, the Naval Surface Warfare Center, Dahlgren Division (NSWCDD) conducted a realistic tactical operations study [1] to assess certain practical limitations that could potentially stand in the way of Navy plans to implement operational efficiencies and crew optimizations on future platforms in the 21st century (see, e.g., [2]). The experiment focused primarily on voice communications tasks future watchstanders are expected to face in an optimized combat information center (CIC). Measures of performance and situation awareness were used to evaluate how well participants were able to handle radio communications on as many as four concurrently active circuits while interacting with an evolving tactical scenario. In spite of the use of virtual audio and speech-to-text displays, it was found that, short of the development of a new technical capability, an increase in per-person, multichannel communications requirements would adversely impact the Navy's operational objectives [1].

Subsequent research on this problem beginning in 2004 at the Naval Research Laboratory (NRL) led to the proposal and implementation of a serialized multichannel communications monitoring technique (U.S. Patent Application Pub. No. US. 2007/0299657), using a patented NRL technology known as "pitch synchronous segmentation" (PSS; U.S. Patent 5,933,808). PSS allows the normal utterance rate of buffered human speech to be synthetically scaled (accelerated or slowed) without meaningful degradations in factors that affect intelligibility [3], [4]. In addition, a series of ongoing human subjects studies at NRL has shown that a serialized, rate-accelerated communications approach overwhelmingly improves performance measures of attention, comprehension, and effort in comparison to concurrent listening in the same span of time [5], [6].

This paper provides an overview of NRL's communications monitoring solution and summarizes the related empirical studies the lab has carried out to date. The goals of additional, ongoing listening studies are also discussed and critical issues that warrant further research are outlined in the conclusion.

II. USING TEMPORAL SCALING AND SERIALIZATION TO MONITOR CONCURRENT VOICE COMMUNICATIONS

A. Motivations

Despite the increasing use of chat and other technologies for disseminating information, voice communications continue to play a critical role in Navy tactical operations. Modern watchstanding can involve over twenty radio-telephone, satellite, and internal circuits, and in the present division of labor, system coordinators typically manage two or more concurrent channels of voice transmissions that are central to the functions of their positions. Watchstanders readily admit to the challenges of this auditory task. When asked, they confide that success is only possible because of domain knowledge, information predictability and repetition, the intermittent nature of communications concurrencies, and, significantly, redundant monitoring by other members of their team. In light of these factors and the imperatives of downsizing, the Navy has good cause to look for alternatives to current voice communications practices in its CICs.

B. NRL's Prototype for Communications Monitoring

Most of the performance declines observed in the NSWCDD study [1] were directly attributable to the intuitive difficulty of trying to listen to as many as four simultaneous talkers and having to keep track of what was being said on each

This work was suppored by the Office of Naval Research under work orders N0001408WX30007 and N0001410WX20273.

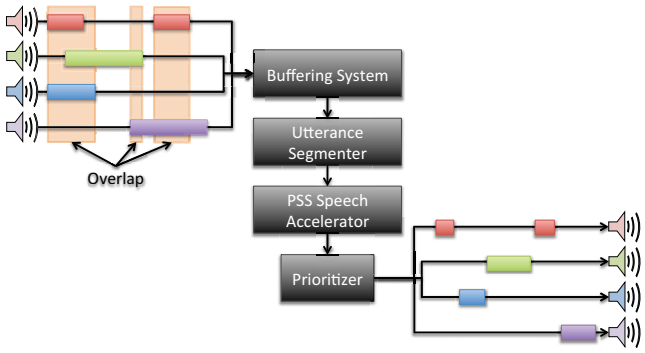

Figure 1. Architecture of NRL's concurrent audio serialization workbench (CASW). Content that overlaps in time on concurrent radio circuits is indicated on the left. Individual circuits are buffered and partitioned into utterance segments. Utterances are then accelerated and presented serially to the listener in a prioritized manner (e.g., first-come-first-serve). AFRL's MMC system (not shown) spatializes the resulting output with a virtual audio engine and provides a speech-to-text display.

radio circuit. The factors that make this sort of task (known as "divided" attention in the human performance literature) challenging for listeners were first studied in the early 1950s by Broadbent and others. Concluding that people are inherently best at "selective" listening—giving their aural attention to a single voice when more than one person is talking at the same time—Broadbent, in [7], proposed that when divided attention is required, people do not actually focus on two or more auditory streams at the same time. Instead, they employ selective listening strategies that involve rapid, serial switches of attention between competing sources. Additionally, each momentary episode of attended information and its context has to be understood, and this limits how much people can follow. Broadbent's analysis thus accounts not only for the effort involved in listening to two simultaneous talkers, but also for why this task becomes more difficult as more voices are added.

At NRL, the sizeable costs of lost information and diminished performance associated with larger concurrent communications workloads were weighed against the costs of sacrificing the timely receipt of some classes of spoken information. This led to the proposal of a serialized monitoring scheme in which the normal rate of speech on buffered circuits could be artificially sped up to compensate for delayed access.

The theoretical advantage of this idea is that it allows listeners to focus on one audio stream at a time—and, thus, accords with their native aural attentional strengths. It does risk the possibility that comprehension of faster-than-normal speech may not be entirely as robust as it is for ordinary speech, and this concern is a primary focus of the human performance studies associated with this program of research. The idea also poses certain time-related risks. However, since voice circuits and are seldom continuously active, accelerating their content means that serial monitoring can be prioritized and accomplished in roughly the same amount of time concurrent monitoring requires, albeit with an inherent processing delay.

As a proof of concept, NRL has developed a prototype system for serializing competing radio circuits, referred to as the "concurrent audio serialization workbench" (CASW) [8]. To fill out its conceptual goals, CASW is integrated with a net-centric multimodal communication (MMC) management suite developed by the Air Force Research Laboratory (AFRL) at Wright-Patterson [9]. MMC incorporates virtual audio and speech-to-text, which have both been shown to be beneficial adjuncts in communications operations [1], [9], [10], [11].

The basic architecture of the NRL's prototype is illustrated in Fig. 1. CASW buffers up to four concurrent radio communications circuits and uses a threshold-based scheme to partition each stream into individual utterances. Each utterance is then processed by NRL's PSS technology [4], which allows its corresponding rate of speech to be synthetically accelerated to an arbitrary speed. Much like musical temporal scaling algorithms, PSS specifically preserves the acoustic pitch and timbre of speech and thus limits changes in factors that impact intelligibility for listeners. The accelerated utterances are then prioritized and passed to the MMC system, which handles the virtual audio and speech-to-text components of the system.

III. Human Performance Studies

A. Issues and Concerns

Although performing a comparative, baseline evaluation of listening performance in concurrent and serial talker paradigms is an essential first step, serial monitoring—especially serial monitoring involving compressed rates of speech—raises a range of issues for human performance that differ in various ways with the demands of concurrent monitoring.

Previous research has shown that virtual listening techniques (i.e., 3D audio) significantly improve the ability of operators to selectively attend to individual talkers when multiple voices are present (e.g., [10]). Serializing the presentation of talkers on multiple circuits is also expected to require virtual listening techniques to reduce the potential for source confusions, particularly at transition points between talkers on different channels. This concern may be even more important in mixed-use auditory displays where additional, non-speech, auditory information may be present (c.f. [12]).

Another concern is that in time-critical contexts, processing delays may be unacceptable and serial monitoring will not be an option. In other contexts, some level of delay may be tolerable, but message prioritization will be an issue.

Perhaps most importantly, accelerating the rate of speech for monitoring purposes may have adverse consequences for objective measures of listening performance. Certainly, any decrement must be weighed against performance in concurrent monitoring paradigms. This issue, alone, should be studied from a number of perspectives. Much as a straightforward comparison between concurrent and serial listening is needed, so, too, is a direct comparison between listening involving normal and sped up speech. Orthogonal to this is the need to measure the impact of different rates of acceleration on measures of listening performance. Directly related to this latter question is the need to evaluate the potential contribution of training and practice effects. Another crucial concern is the impact of forced switching of attention between contexts (i.e., circuits) that will naturally arise whenever what is being said on competing channels is segmented into utterances as a consequence of automated partitioning and, as a result, must be

interleaved. A further complexity interleaving poses for listeners is the fact that there can be more than one talker on a given channel.

A small number of listening studies, designed for an initial examination of some of these questions, have been carried out at NRL. Methods used to evaluate listening performance in this work are discussed next, and then the outcomes of these experiments are profiled in the remainder of this section.

B. Measuring instruments

A common difficulty in applied human factors research is obtaining access to target user populations. Due to its location in Washington, DC, for instance, NRL lacks ready access to sufficient numbers of experienced Navy watchstanders. A related challenge is the problem of developing operationally appropriate performance scenarios—e.g., a realistic body of tactical communications. NRL has thus developed stimuli and performance paradigms for non-specialists that are structured to obtain generalizable results. Accordingly, the speech materials used to evaluate concurrent and serialized listening in the experiments summarized below have been drawn from source materials that are taken to be easy to follow by a typical member of the laboratory's general staff. Two different corpora have been developed. The first is a set of broadcast radio commentaries on familiar topics made by four professional commentators, each edited to be the same length when heard at a normal rate of speaking. The second is a battery of narratives, each describing an event in an ordinary person's life, all spoken by the same talker. The materials of this latter corpus have been normalized over parameters such as length, speaking rate, numbers of words and propositions, etc. (For respective details, see, [5] and [6].)

Listening performance can be thought of as having two critical stages: what the listener attends to and what is comprehended. Since aural attention cannot be physically tracked, a secondary means for estimating this aspect of listening is needed. A somewhat similar problem arises for determining what is understood. Comprehension cannot be directly observed and any method for eliciting evidence of understanding as it is taking place risks interfering with what is being measured. Taking these considerations into account, NRL has employed two instruments for measuring attention and comprehension in its voice communications monitoring studies. Aural attention performance is estimated by displaying a set of discourse-ordered target and foil phrase lists, corresponding to each of the auditory streams (i.e., talkers) that are presented. Listeners are asked to actively track and mark the phrases they hear in these lists during the aural monitoring portion of each experimental condition. Comprehension is estimated after episodes of aural monitoring by asking listeners to respond to a series of content-related statements that correspond either to information they heard or that was not present in the spoken material. (See, [5] and [6] for additional details about these instruments and their use in each study.)

C. Comparing concurrent and serial listening

NRL's initial evaluation of the potential for serial aural monitoring was conducted with a group of twelve listeners from NRL, and is reported in [5]. Listening performance in this within-subjects study (all of the listeners were enrolled in all of the manipulations) was measured in four conditions. Two conditions respectively involved listening to two and four concurrent talkers speaking at normal rates, with the two-talker manipulation serving as a comparative baseline corresponding to current Navy practice. In the other two conditions, listeners respectively heard four serialized talkers speaking normally and at a rate made 75% faster by the PSS algorithm. A summary of these four manipulations and their coded designations is given in Table 1. All of the listening materials were taken from the corpus of radio commentaries described in section IIIB. In each manipulation, talkers were horizontally arrayed at separated positions in front of the listener in a virtual listening space rendered with headphones. Aural attention and aural comprehension were measured, respectively, with the phrase list and statement verification response tasks described above. Accordingly, an interactive list of randomly interleaved targets and foils (phrases not present in the stimuli) for each talker was displayed during the presentation of the auditory materials in each condition, and the comprehension response task was given afterwards. (In the serial manipulations, this meant that listeners heard all four talkers, back to back, before being asked to respond to the comprehension statements.)

Fig. 2 presents the resulting mean proportional data for both response tasks in [5]. As one might expect, attention and comprehension were best in the serial monitoring condition in which talkers spoke at normal rates (4S). In contrast, both of these measures were undermined by the 75% accelerated speaking rate in the other serial manipulation (4SF). Clearly, though, listeners were able to attend to and understand most of the spoken information in both of these conditions. The comparative performance declines in the other two conditions (2C and 4C) are indicative of the substantial challenges divided listening poses. Proportional performance was dramatically impacted by the manipulation involving four competing talkers (4C), showing the difficulty of attending to four voices speaking at the same time. More importantly, this manipulation had a relatively severe impact on listeners' ability to form an understanding of what was said. Indeed, the numbers underlying performance in the baseline manipulation (2C) reflect a surprising fact. Since 2C involved only two talkers, listeners were given half as many phrases and comprehension statements as in each of the four-talker manipulations. Listeners attended to and understood much more of what was said in 2C than in 4C. But the numbers underlying the proportions in 2C are smaller than the corresponding data in 4C. Thus, listeners actually worked harder in 4C, but scored worse on what was said than in any of the other conditions.

A final comparison worth noting involves the performance difference between the 4SF and 2C manipulations. Clearly, the challenges listeners faced in the concurrent talker

TABLE I. A SUMMARY OF THE FOUR MANIPULATIONS IN [5]

Condition	Description
2C	***2*** talkers, ***concurrent*** presentation (baseline)
4C	***4*** talkers, ***concurrent*** presentation
4S	***4*** talkers, ***serial*** presentation
4SF	***4*** talkers, ***serial*** presentation, 75% ***faster*** than normal rate of speech

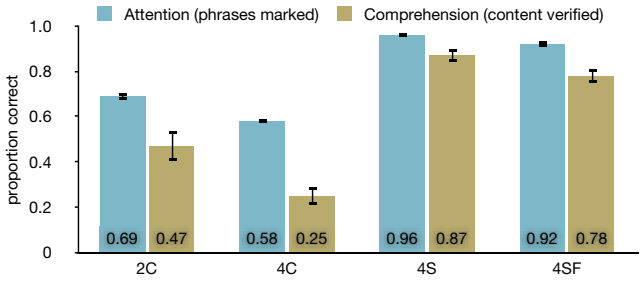

Figure 2. Combined plots of the mean proportion of correctly identified target phrases and rejected foils (blue bars) and the mean proportion of correctly identified comprehension statements (tan bars) in [5]. Error bars show the standard error of the mean (s.e.m.).

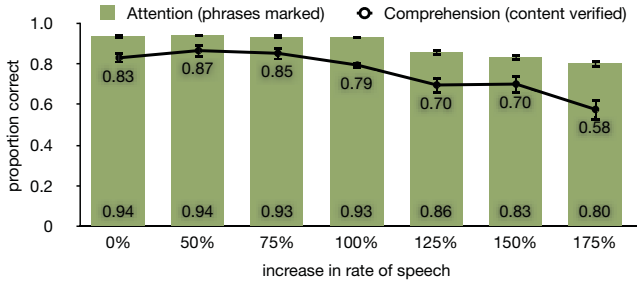

Figure 3. Combined plots of the mean proportion of correctly identified target phrases and rejected foils (green bars) and the mean proportion of correctly identified comprehension statements (line) in NRL's previously unreported accelerated speech study. Error bars show the s.e.m.

manipulations were much greater than the difficulties posed by listening to accelerated speech. More to the point, utterances in the accelerated serial talker condition (4SF) were 57% shorter than ordinary speech, so the duration of the first half of this condition is only 14% longer than the full duration of the two-concurrent-talker-condition (2C). Taking into account the fact that all of the stimuli in the study involved continuous speech, this rough equivalence in listening times emphasizes the limited costs and appreciable performance advantages serial monitoring of accelerated speech offers over current practices. Still, if accelerated speech is to be used in future Navy operations, the rate at which speech can be presented without a considerable loss of information needs to be determined.

D. Listening to different rates of accelerated speech

Serial listening performance was next evaluated in a previously unreported study at normal and progressively faster rates of speech with the same response instruments as described earlier. Fourteen new NRL listeners were enrolled and, again, a within-subjects design was used. In this experiment, two commentaries were serially presented, back-to-back, in each of seven conditions. The experiment entailed a baseline of unmodified speech and six progressive manipulations in which the rate of speech was respectively made 50% to 175% faster than normal in 25% increments, using the PSS algorithm. Commentaries were rendered in the same virtual listening environment as before. Listeners carried out the baseline exercise first and performed the remaining six in randomized orders. As in the previous study, aural attention was measured by asking listeners to look for, and mark, phrases in the auditory materials. Similarly, comprehension in each condition was measured with responses to content-related statements presented after both commentaries were complete.

Fig. 3 shows the resulting mean proportional data for both response tasks across the seven speech rate manipulations. As attention performance declines across increasing speed, so does comprehension performance. This correspondence between the two measures was also seen in the previous study (cf. Fig. 2). Both measures of performance in Fig. 3 markedly declined after the rate of speech was accelerated beyond 100%, and pairwise comparisons confirm that these measures in the last three conditions (125%, 150%, and 175%) are significantly different from performance in the baseline condition. The differences among the first four manipulations, however, are not significant. This latter result seems to contradict the results of the previous study, which found larger differences in performance between the normal and 75%-accelerated serial conditions. There are methodological differences between the studies, though, and the disparity may be an exposure effect. Four talkers were presented back-to-back in the prior study's serial manipulations (4S and 4SF) before comprehension performance was assessed (as opposed to two talkers in all of this study's manipulations), and unlike the prior study, the baseline manipulation in this experiment was given to all listeners first. Even so, performance in the first four conditions here is consistent with the corresponding ranges of serial listening measures that were previously observed. It is also worth noting that serial listening performance in the last three conditions in Fig. 3 is better than performance in the previous study's two concurrent-talker condition (2C in Fig. 2).

E. Evaluation of Training and Practice Effects

Encouraged by these findings, another accelerated speech study was developed at NRL to investigate whether or not listeners' comprehension performance improves with training and is subject to practice effects; this study is reported in [6]. A corresponding set of speech materials—the corpus of narratives about events in people's daily lives described above—was developed for this study, and comprehension performance was measured. Twenty NRL employees were randomly divided into two equal-sized groups. Each participant carried out a baseline listening task involving two narratives at a normal speech rate and no between-group differences in comprehension performance were observed. Each group was then asked to complete a series of exercises that respectively entailed listening to seven progressively faster rates of accelerated speech, ranging from 50% to 140% in 15% increments. In each exercise, listeners heard three narratives, all rendered at the same rate of acceleration. Comprehension was measured immediately after each narrative was presented to test for within-speed practice effects. All of the listeners in the first group were designated as the "training" group and carried out the seven exercises in a progressively faster order. The remaining listeners—designated as the "random" group—were given the exercises in a randomized order.

The plot in Fig. 4 shows the outcome of the study. The main analysis compared performance in the "training" group with performance in the "random" group. As expected, across groups, comprehension performance declined as speech rate increased. The "training" group performed significantly better

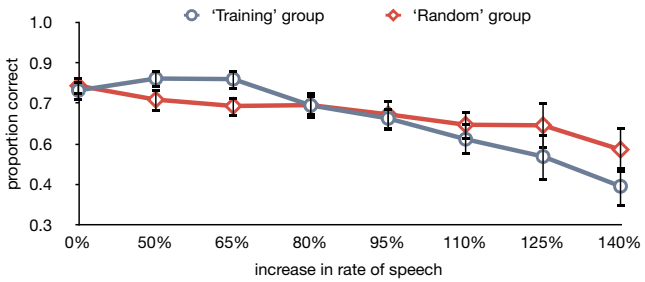

Figure 4. Combined plots of the mean proportion of correctly identified comprehension statements in the "training" group (blue line) and in the "random" group (red line) in [6]. Error bars show the s.e.m.

than the "random" group, but only at the lower accelerated speech rates (i.e., 50% and 65%). What was not expected, however, was how quickly participants adapted to listening to the accelerated speech in each exercise. This was demonstrated by the lack of practice effects within each listening exercise. Figure 4 also seems to suggest that the "random" group performed significantly better at the higher accelerated speech rates (i.e., 125% and 140%) than the "training" group. However, there were no significant differences between the groups. The way in which the narratives were presented to the "training" group (i.e., at progressively faster rates) may have induced fatigue over the course of the experimental session, further supporting the notion that training may only be effective at lower accelerated speech rates.

IV. Conclusions

The outcomes of the three studies summarized above make a compelling case for the performance advantages of listening to competing voice communications serially as opposed to the Navy's current voice communications task practices, which entail listening to everything concurrently. The use of speech rate acceleration technology offers a way to compensate for most of the operational costs listening to one circuit at a time would introduce. And notably, even highly accelerated speech appears to have a less substantial impact on measures of aural attention and comprehension than concurrent listening does.

The hope is that serializing the watchstander's voice communications task will be a crucial, facilitating technology in the Navy's future operations. Still, a range of additional performance questions remain to be addressed before this promising paradigm can be judged to be ready to move to the next level of evaluation. These include further research on the use of virtual listening environments for managing voice communications, operational research on message prioritization, and the procedural costs of processing delays that serialization inherently imposes. Furthermore, performance studies with more realistic corpora, patterns of communications traffic, and levels of operator interaction—both with communications and other tasks—also need to be conducted. For instance, it was found in [1] that increasing an individual's concurrent communications load directly impacts performance on other, non-communications tasks. How serial communications monitoring will impact a watchstander's performance of other operational responsibilities is unknown.

In concert with this research agenda, NRL is currently engaged in the conduct of two additional studies. One critically addresses listening performance when content on competing channels is serially interleaved as a result of automated segmentation and/or prioritization schemes. The other is an extension of the third study outlined above and is investigating strategies for improving listeners' comprehension performance on the basis of natural concept boundaries. This work is expected to be complete in 2011.

Acknowledgment

The authors would like to acknowledge and thank Philip Moore of Knexus Research Corp. for his work on the CASW prototype developed as part of this research, as well as the Battlespace Acoustics Branch at AFRL, Wright-Patterson AFB for their work on the MMC management suite and Paul Bello for comments on an earlier draft of this paper.

References

[1] D. Wallace, C. Schlichting, and U. Goff, Report on the Communications Research Initiatives in Support of Integrated Command Environment (ICE) Systems, Naval Surface Warfare Center Dahlgren Division, TR-02/30, January, 2002.

[2] C. T. Bush, J. R. Bost, P. S. Hamburger, and T. B. Malone, "Opimizing manning on DD21," Proceedings of the Association of Scientists and Engineers (ASE) 36th Annual Technical Symposium, April, 1999.

[3] B. McClimens, D. Brock, and F. E. Mintz, "Minimizing information overload in a communications system utilizing temporal scaling and serialization," in Proceedings of the 12th International Conference on Auditory Display (ICAD). London, UK, June, 2006.

[4] G. S. Kang and L. J. Fransen, Speech Analysis and Synthesis Based on Pitch-Synchronous Segmentation of the Speech Waveform, Naval Research Laboratory, TR-9743, November, 1994.

[5] D. Brock, B. McClimens, J. G. Trafton, M. McCurry, and D. Perzanowski, "Evaluating listeners' attention to and comprehension of spatialized concurrent and serial talkers at normal and a synthetically faster rate of speech," in Proceedings of the 14th International Conference on Auditory Display (ICAD). Paris, France, June, 2008.

[6] C. Wasylyshyn, B. McClimens, and D. Brock, "Comprehension of speech presented at synthetically accelerated rates: Evaluating training and practice effects," in Proceedings of the 16th International Conference on Auditory Display (ICAD). Washington, DC, USA, France, June, 2010.

[7] D. W. Broadbent, Perception and Communication, Pergamon Press, New York, NY, USA, 1958.

[8] P. Moore, CASW final report, Knexus Research Corporation, 2009, unpublished.

[9] V. S. Finomore, D. K. Popik, B. D. Simpson, D. S. Brungart, C. E. Castle, and R. C. Dallman, "Development and evaluation of the multi-modal communication management suite," Proceedings of the 15 th. International Command and Control Research and Technology Symposium (ICCRTS), Santa Monica, CA, USA, June, 2010.

[10] D. S. Brungart & B. D. Simpson, "Optimizing the spatial configuration of a seven-talker speech display," ACM Transactions on Applied Perception, vol. 2, pp. 430-436, 2005.

[11] W. T. Nelson, R. S. Boila, & M. A. Vidulich, "User-centered evaluation of multi-national communication and collaborative technologies in a network-centric air battle management environment," Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting, pp. 731-735, 2004.

[12] D. Brock, J.A. Ballas, J.L. Stroup, and B. McClimens, "The design of mixed-use, virtual auditory displays: Recent findings with a dual-task paradigm," Proceedings of the 10th International Conference on Auditory Display, Sydney, Australia, July, 2004.