EVALUATING LISTENERS' ATTENTION TO AND COMPREHENSION OF SERIALY INTERLEAVED, RATE-ACCELERATED SPEECH

Derek Brock and Brian McClimens

U.S. Naval Research Laboratory,
4555 Overlook Ave., S.W.,
Washington, DC 20375 USA
derek.brock@nrl.navy.mil

S. Camille Peres

University of Houston-Clear Lake,
2700 Bay Area Blvd.
Houston, TX 77058 USA
peressc@uhcl.edu

ABSTRACT

In Navy command operations, individual watchstanders must often concurrently monitor two or more channels of spoken communications at a time, which in turn can undermine information awareness and decision performance. Recent basic work on this operational challenge has shown that a virtual auditory display solution, in which competing messages are presented one at a time at faster rates of speech, can achieve large and significant improvements on diminished measures of listening performance observed in concurrent monitoring at normal speaking rates with equivalent materials. In the third of a series of experiments developed to address performance questions the parameters of this framework raise for listeners, dependent measures of attention and comprehension were compared in a two factor design that manipulated how serial turns among four talkers were organized and their rate of speech. Although both factors impacted performance, the resulting measures remained substantially higher than corresponding measures of performance with concurrent talkers in an earlier study.

1. INTRODUCTION

In Navy command operations, individual watchstanders must often interact with and monitor two or more concurrent channels of spoken radio communications, and this, coupled with the demands of visual tasks, can easily impact information awareness and decision performance [1]. Even so, efforts to increase productivity and streamline operational requirements, have recently raised the possibility of giving watchstanders a range of new display technologies and enlarging their responsibilities to as many as four active communications circuits. A 2001 operational study with a diverse group of experienced watchstanders, however, found that overall message comprehension and awareness of time-critical events fell significantly in a realistic tactical scenario when communications monitoring involving only three channels of competing speech was tasked [2]. This outcome and other findings in the same study suggest that the challenge of attending to multiple streams of concurrent aural information can quickly become overwhelming in high-paced operations.

Monitoring voice communications serially (one at a time) could reduce the considerable requirements of the watchstander's listening task, but would almost certainly result in cumulative and, in some cases, unacceptable presentation delays during periods of high volumes of message traffic. Digitally buffered and recorded speech, however, can be artificially sped up with signal processing techniques that allow the essential timbral features needed for intelligibility and other expressive and informational factors to remain intact. Synthesizing a faster version of what is said on a given radio channel naturally requires a processing delay before it can be aired for the listener—minimally, the time required to receive the original transmission plus a marginal amount of additional processing time. But since competing messages can be processed in parallel, speech rate acceleration techniques can be used to limit the accumulating cost of serial presentation delays and, therefore, provide an opportunity to study serial monitoring as an effective alternative to current communications monitoring practices.

A straightforward model of just under three minutes of activity on four concurrent channels, for instance, would take approximately five minutes to listen to serially, assuming a relatively busy, mean use rate of 40% on each channel and a nearly continuous overlap of two or more messages (see Figure 1a and b). Just doubling the speed of all but the initial message, however, (assuming the first message would be monitored in real-time while competing messages are concurrently buffered and accelerated in parallel) substantially reduces the extent of accumulating delays. Under the acceleration scheme shown here, serialization never adds more than half of the running time required to monitor all four channels concurrently, and

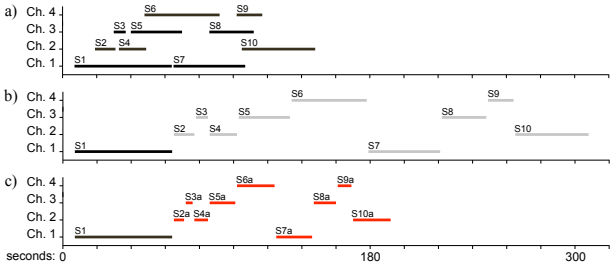


Figure 1: a) Randomized 3-minute model of ten spoken radio transmissions (numbered from S1 in the order they are received) on four concurrent channels. b) Time required to buffer and display the same ten signals serially, in the same order and at the same speaking rate as received (i.e., with no acceleration). c) Time required to buffer, process, and display the same signals at an accelerated speaking rate of 100% (i.e., twice as fast as the original speaking rate). Unserialized messages in the figure (messages presented as they are received) are shown in black. The letter "a" indicates accelerated signals, as in "S2a."

This work was supported by the Office of Naval Research under work request N0001410WX20448.

total listening time is just over three minutes (Figure 1c). More efficient accelerated monitoring and message organization tactics are also possible, but might only rarely be needed due to routine lulls in most real-world patterns of communications traffic.

1.1. Performance concerns for listeners

Although the idea of synthetically accelerating concurrently received messages for subsequent display makes serial monitoring an operationally feasible concept—at least in the sense that it allows serial monitoring to be carried out in nearly the same amount of overall time concurrent monitoring requires—it also raises specific human performance questions for listeners. The most important practical concerns are: a) the performance strengths and weaknesses of human auditory attention; b) performance differences associated with listening to different rates of accelerated speech; and c) the impact of having to shift between communications contexts in an "interleaved" manner, as depicted in Figure 1b and c, or as dictated by some other prioritization scheme.

1.1.1. Auditory attention

Intuitively, listening for content from two or more talkers is harder to do when the parties speak at the same time, as opposed to when they speak individually. Listening to competing talkers requires what is called "divided attention." Both Broadbent [3], and, more recently, Shinn-Cunningham [4], attribute the difficulty of divided listening to an essential limitation of the human auditory attention resource. When divided attention is required, despite anecdotal claims to the contrary, it appears listeners are not really able to focus on two or more auditory streams simultaneously. Instead, while they may be aware of multiple sources and alert to salient features of those sounds, they can only give their attention, selectively, to one coherent stream at a time, and consequently must resort to ad hoc, though possibly practiced, listening strategies that entail rapidly switching their focus back and forth between competing threads of information. What makes giving divided attention to competing auditory streams more difficult than giving sustained attention to one at a time is the mental effort that switching between aural information contexts requires.

As part of a series of experiments that includes the study reported here, Brock et al. [5] examined the question of divided and undivided listening in a quasi-applied context. Working with a corpus of spoken commentaries on everyday topics, inferential measures of auditory attention and comprehension were used to compare listening performance in four manipulations involving either concurrent or serial talkers. The manipulations with concurrent talkers (two talkers in one condition and four in the other, and both at normal speaking rates) reflected current and proposed Navy communications monitoring practices. The serial talker manipulations (one at normal speaking rates, the other at an accelerated rate of 75%, and both with four talkers) allowed serial monitoring to be compared directly with concurrent listening and provided an initial look at the impact of accelerated speech on serial listening performance. The resulting measures of attention and comprehension proved to be highly correlated with each other, and all pairwise comparisons between the manipulations were significant. Listening performance was respectively poor and poorest in the two and four concurrent talker conditions, and better and best in the accelerated and normal serial conditions. The outcome was thus consistent with the current understanding of auditory attention and demonstrated a clear performance advantage for serial monitoring over current practice, even with faster speech.

1.1.2. Rate-accelerated speech

Techniques for synthetically compressing (and, therefore, accelerating) the nominal speaking rate of normal, recorded speech—without altering its pitch—were first studied in the early 1950s. Research by Miller and Licklider [6] showing that brief segments of continuous speech could be either systematically blanked out ("interrupted") or masked with only modest impacts on perceived intelligibility led to the idea of splicing together what remained to reduce listening time [7]. Eventually, as interest in accelerated speech grew, and access to digital signal processing technology became widespread, more sophisticated speech-rate modification techniques were developed that are capable of preserving most, if not all, of the vocal features involved in clear enunciation at rates of compression that exceed 200% (see [8] for an outline of research up to the beginning of the 1990s). The technique used in the work reported here is a computationally efficient method for modulating the time scale of speech known as "pitch synchronous segmentation" (PSS) that was developed by the Navy in 1994 [9]. Human performance and perceptual studies associated with rate accelerated speech have focused primarily on the intelligibility of individual words and the practical limits of acceleration, as well as the impacts of acceleration and prosodic modifications (particularly, the removal of pauses) on the more practical question of comprehension performance. Additional work has also explored the impacts of training and practice and, more recently, performance differences associated with aging (see, e.g., [10]).

Since varied pacing might be needed to accommodate changing amounts of message traffic in a serial communications monitoring scheme, two experiments ([1] and [11])—one planned as a follow on for [5] and another developed by Wasylyshyn—were recently conducted to examine listening performance with different rates of accelerated speech using the PSS technique [9]. Although different materials and exposure regimes were used in each protocol, the outcomes of both studies are in general agreement with the findings of earlier research on this question using other speech rate compression methods. Brock et al. [1] found comprehension of compressed speech up to a 100% increase in ordinary speaking rates (i.e., twice as fast) to be essentially equivalent to listening to normal speech. Similar equivalence was observed in Wasylyshyn's study [11] up to a rate of 80%, or 1.8 times as fast as normal speech. Above these levels, as in other research, performance was found to slowly but significantly decline in a relatively steady manner as the degree of acceleration grows. In both studies, however, even at the highest levels of accelerated speech rates (175% in [1] and 140% in [11]), mean comprehension was much better than, or as good as, the listening conditions involving two and four concurrent talkers in [5]. The consistency of these performance

outcomes with the findings of others suggests that the ability of listeners to follow and make verifiable sense of synthetically accelerated speech at speeds up to and beyond a 100% increase in normal speaking rates is a readily acquired skill.

1.1.3. Listening to serially interleaved communications

Questions concerned with interleaving, specifically, shifting back and forth between communication contexts, are motivated by the insight that competing communications are just that. If one message is more timely or important than another, the listener will want to give its presentation priority, even if this means withdrawing attention from or suspending the less urgent of the two and returning to it later. Suspension would be the case in a serial monitoring scheme, and the issue then becomes, what is the likely impact of system-imposed interruptions on listening performance when messages are subsequently resumed. Even more to the point in a communications setting is the fact that what is said on most radio nets is not just one individual talking, but discourse among multiple talkers. Thus, upon resuming a suspended channel, the listener not only faces the problem of attending while reengaging with the channel's operational context and recalling its state, but also of recognizing who the talker is and/or what the talker's role in the current communications context is. Mastering these additional aspects of the serial listening task may well be made more difficult by accelerating what is displayed for the listener, even if the increase falls within the range of equivalent-to-normal comprehension performance.

However, other factors may measurably impact listening performance, too. The most important concerns are: message complexity; where suspensions occur within a message stream; and whether or not the pace of display provides opportunities to reflect on or rehearse a suspended context before it is resumed. For instance, in addition to difficulties that ordinarily arise for listeners when speech materials in any format are syntactically complex (e.g., [10]), listening performance is known to be hurt when unexpected pauses occur in sentences, as opposed to at grammatical clause boundaries [12]. From this, it follows that listeners are likely to find arbitrary suspension points more difficult to work with than suspensions that occur at the end of clauses or on sentence boundaries, or perhaps breaks that occur between different talkers on a given channel.

As for the pace of display, listening that involves interrupting one informational context and attending to another can be likened to a sequential multitasking paradigm [13]. Current cognitive theories of multitasking model the ability to juggle more than one task at a time as interacting goal hierarchies [14] and, more recently, as separate "threads" of goal directed activity [15]. For comprehension tasks, people often need to maintain an informational context or "problem state"—a small amount of applicable knowledge, and/or intermediate results, that is temporarily buffered for working access. Recent work by Borst et al. [16] has concluded that the cognitive resource for this intermediate store can only be used by one task thread at a time. Thus, part of the difficulty of managing even two ongoing comprehension tasks at once, whether they are perceptually concurrent or sequentially paced, is explained by the mental effort that is needed to repeatedly reinstate their respective contexts. The time this requires can become an issue, too, if the wait before the next episode of attention to a task becomes too long. In a serial monitoring scheme, a progression of different channels may intervene before a given interrupted channel is resumed, depending on how the incoming spoken information is prioritized and segmented. As the pace of imposed switching between suspended contexts slows, listeners will have increasing difficulty recalling each channel's respective problem state [17]. Empirical studies and related modeling work by Trafton et al. [18] and Altmann and Trafton [19] have shown that to counteract this quantifiable tendency to forget, listeners need to rehearse an interrupted context—ideally, at the point when the interruption occurs. Consequently, if in addition to relatively slow pacing, switches between channels are effectively immediate (with no gap to briefly think about what was just interrupted), listening performance can be impacted in two ways. Either, listeners will try to rehearse the previous context anyway, and initial attention to the new context will be impaired, or listeners will fail to think about the previous context and have greater difficulty recalling it later, which will also impair initial attention to the new context.

1.2. Listening study

The listening study reported here—an initial 2x2 comparison of interleaved and non-interleaved listening with normal and rate-accelerated speech—is the third in a series of experiments that includes the work presented in [5] and [1]. For consistency with the previous studies, the speech materials used for auditory display in the present experiment were again developed from a public radio archive of spoken essays by four professional commentators. (An essay from an additional commentator was also used for training purposes—see Section 2.1.3 below). This category of talk sidesteps potential confounds and has specific advantages for the population of non-specialist listeners recruited to participate in the study. In particular, each commentary is presented by a single talker and, so, avoids contextual confusions that could arise from the presence of more than one voice on the same channel. Each commentary also covers a single everyday topic in ordinary conversational language that is easy to follow and quickly establishes an easily recognized contextual theme for the channel it corresponds to during its presentation.

A serially interleaved communications display in an operational setting would probably exhibit some of the characteristics depicted in Fig. 1c, notably, a mix of communications sounded at normal and accelerated rates and a mixed range of message lengths. The present study's chief aim, however, was to examine the impact of interleaving itself on listening performance with normal and faster speech, as opposed to other issues interleaved listening designs may raise. Consequently, the main questions addressed here are: a) Is the problem of having to follow and understand four different spoken information contexts harder to do when, instead of being allowed to listen to each talker's full presentation, one at a time, what each talker has to say is broken into an ordered series of utterances that are displayed as a randomized sequence of turns among the talkers? And b) how does making the speech materials in these contrasting conditions much faster affect the ability of listeners to follow and understand all of what each talker has to say?

Because serially interleaved listening can be characterized

as an example of sequential multitasking, the performance concerns raised in Section 1.1.3 related to the interruption of contexts are addressed in the experimental design by the insertion of a brief gap after each talker's turn in the manipulations that involve interleaved listening. The intent in doing this, though, was not to measure the impact of remedial measures for interruptions, but rather to organize the design of the interleaved listening task in a theoretically principled way. To ensure talkers had equal priority throughout, each commentary was edited to approximately the same length and segmented into a congruent (equally numbered) sequence of utterances or "turns." Four commentaries (one per talker) were presented in each of the listening exercises, and in those with interleaved utterances, the order of sequential turns among the talkers was randomized for each listener (see Section 2.1.3 below for additional details). As a result, the wait between a given talker's completed turn and that talker's next turn in the interleaved listening exercises might be short or long, but, on average, entailed the span of time defined by the first inserted gap, plus three turns from the other talkers, plus the gaps inserted after each of these intervening turns. Each gap thus provided a moment to think about the completed turn's context, but for consistency with the non-interleaved portion of the study, no constraint was placed on how listeners were expected to manage their time during any of the listening exercises.

Other aspects of the experimental task design that were similarly informed by current theory are the use of separate virtual locations for each talker in the auditory display and the manner in which commentaries were divided into turns. Giving the apparent source of each talker's voice its own virtual location, and keeping this constant throughout the study, provided two, closely related theoretical benefits for listeners. First, it capitalized on the spatial skills listeners routinely use to discriminate between sources of auditory information in selective attention (cf. [20]). And second, it provided an external set of talker-specific, contextual cues in the aural information environment. That is, listeners could use each talker's virtual location as an aural reminder for returning to that talker's corresponding problem space during the serially interleaved listening exercises. (Listeners were also able to exploit an external set of visual cues in these exercises; see Section 2.1.1 and 2.1.2 below.) As for turns, speech on competing radio channels could no doubt be broken into separate utterances in several different ways for interleaved display in an operational serial monitoring scheme. Empirical findings such as those in [12], however, suggest that forming an understanding of interrupted speech is facilitated when interruptions occur on grammatical and/or conceptually complete boundaries, and that listeners perform best when this is the case. Thus, to minimize performance confounds related to encoding, in addition to dividing each commentary into an equal number of turns, all of the partitions were made so utterances were sentences or complete phrases.

2. METHOD

Sixteen participants, two female and fourteen male with a mean age of 29.3 years (s.d. = 10.7), all personnel at NRL, and all claiming to have normal hearing, took part in the experiment, which employed a within-subjects design. The visual part of the study was displayed on an NEC MultiSync LCD 2090UXi flat-panel monitor and the aural component was rendered with VRSonic Vibestation runtime spatial audio software, Sony MDR-600 headphones, and an InterSense InertiaCube3 for head tracking. The main experiment consisted of four listening exercises, which were performed by all participants in counterbalanced order. A brief introductory session before the study explained each of the ways participants were asked to respond and described what they would hear and see in the study. Each condition in the main experiment was preceded by a short training session that resembled the format of the listening exercise that followed. These sessions allowed participants to become familiar with the auditory manipulations and their corresponding listening requirements.

2.1. Apparatus

Listeners were asked to make two types of responses in the experiment—the first while listening and the other performed immediately after. Both of these tasks are largely the same as those used to assess listening performance in [5] and [1].

2.1.1. Response tasks

In the first response task, participants were instructed to mark items in a set of lists that were displayed on the flat-panel monitor during the auditory portions of both the training sessions and the main listening exercises. Each list (as well as its left-to-right position onscreen) corresponded to one of the commentaries being presented in the current segment of the experiment and was composed of an ordered set of noun phrases. There were four lists and four commentaries in each of the main experimental manipulations and two lists and two commentaries in each of the respective training sessions. Each list functioned as a visual contextual cue when its talker's commentary was active. Participants were asked to use a mouse to successively check off exactly worded phrases if they heard them spoken (targets) and to ignore any intervening, though topically similar, phrases they did not hear (foils). Lists in the main listening exercises were each composed of twenty targets and an equal number of foils, with zero to three intervening foils placed at random between targets, and no more than three targets in a row. (Shorter lists were used in the training sessions.) In part, because participants were not made aware of the arrangement of targets and foils, and in part because of the potential to become lost while trying to perform the phrase recognition task (thus, undermining the overriding goal of listening), a portion of the currently active list was highlighted as a pale blue region that functioned as a position marker corresponding roughly to the utterance that was currently being presented in the active commentary (see Figure 2a). To ensure that listeners could not game the task, the highlighted area moved continuously and always encompassed several phases in the active list.

In the second response task, which is derived from a technique for measuring reading comprehension developed in [21], participants were given a series of representative sentences to read and asked to judge whether each contained "old" or "new" information based on the spoken materials they had just listened to. "Old" sentences were of two types: verbatim renderings and synonymous paraphrases of sentences in the commentaries. "New" sentences were similarly of two

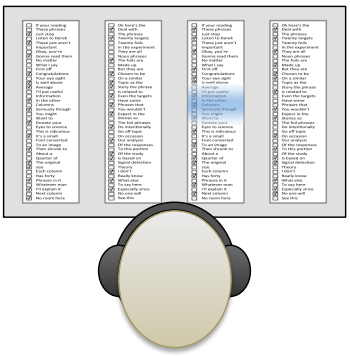
types: "distractors"—sentences stating something that was not implied or said—and commentary sentences with one or two words changed to make the meaning clearly different from what was said. Participants were also given the option to indicate that they did not know whether a sentence they were asked to evaluate was old or new. In the training sessions, participants were given only two sentences per commentary to assess, one old and the other new. Eight sentences per commentary (two of each of the old and new sentence types) were given in the main listening exercises.

In the present study, participants were also asked to indicate how confident they were in their judgments. They did this with an appropriately labeled onscreen widget resembling a slider, with end points labeled "Low" and "High." When participants indicated they could not evaluate a particular sentence, the confidence scale was grayed out.

2.1.2. Auditory display

All of the auditory manipulations were presented in a virtual listening environment organized somewhat similarly to the auditory displays used in [5] and [1]. In this experiment, however, head tracking was also used to implement an augmented auditory reality display, meaning that the apparent referential frame of the virtual aural setting remained the same as that of the actual visual setting, regardless of how participants moved their heads. Each of the normally spoken and rate-accelerated commentaries was binaurally filtered and rendered with headphones using a non-individualized head-related transfer function. To ensure that participants could quickly focus their aural attention on the active commentary (cf. [20], [22]), the apparent locations of the four talkers in each of the main listening exercises were positioned, from left to right on the virtual horizontal plane in front of the listener, at -75°, -25°, 25°, and 75°, with 0° being straight ahead in the visual environment (see Fig. 2b). In the training sessions, only the -25° and 25° positions were used. Each talker's virtual location was maintained across all manipulations and, as was noted above, each of these locations corresponded in a left-to-right manner to the visual location of its matching phrase list in a given exercise on the flat screen monitor.

a)



b)

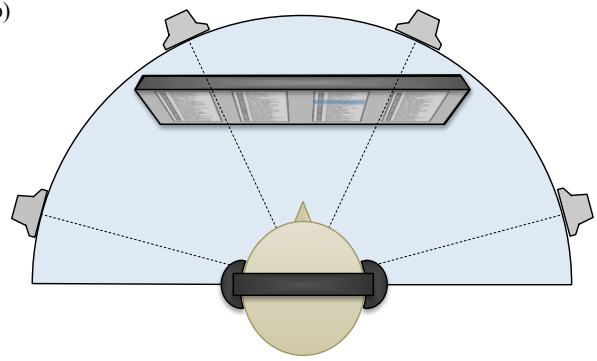


Figure 2: a) Illustration of the visual display showing the four lists of target and foil noun phrases used for the phrase recognition task in the listening exercises. Each list corresponds to a talker, and the pale blue region about midway down the third list indicates that the middle-right talker is speaking. Listeners were asked to mark any noun phrases in each list they heard the corresponding talker say. b) Diagram of the runtime spatial audio environment showing the virtual locations of the four talkers in the listening exercises and their left-to-right correspondence with the onscreen lists used for the phrase recognition task.

2.1.3. Listening materials and experimental manipulations

Each participant in the study listened to a total of 18 spoken essays by two female and three male commentators selected from an internet archive of public radio broadcasts. Both of the women and two of the men were designated as the set of talkers participants would hear in each of the study's main listening exercises. Four pieces from each of these individuals were chosen and edited to remove music and other non-speech sounds. The resulting 16 commentaries ranged from 2 min. 9 sec. to 2 min. 32 sec. in length, with a mean length of 2 min. 19 sec. Listeners heard one commentary per talker in each of four experimental manipulations in the main body of the experiment. In addition to these commentaries, two shorter pieces were also selected and similarly edited for the study's training sessions. Both were spoken by male talkers, of whom, one was the remaining male commentator from above. Participants trained with appropriately manipulated versions of these two commentaries before each of the main listening exercises. These short training sessions allowed participants to become acquainted with the format of each of the auditory display manipulations and practice the listening requirements.

All of the commentaries were further edited into ordered sequences of successive, non-overlapping clips, with each clip corresponding to an utterance. The edits were made so that utterances were either complete sentences or grammatically complete clauses. Additionally, each utterance was edited to start and end with its talker's voice, meaning that any preceding or trailing silence at these specific points was removed. The 16 commentaries used in the main listening exercises were divided in to 15 clips each, with utterances ranging from 4 to 16 sec. and averaging about 9.5 sec. The short commentaries used for the training sessions were also similarly divided into six clips each. Next a version of each clip at double the rate of its original speech was generated with the PSS algorithm [9]. 100% acceleration was chosen for the study because listening performance at progressively faster rates of speech markedly declined above this point in [1].

Each of the four main listening exercises implemented a separate treatment within a two-factor, 2x2, repeated measures design. The first factor, presentation, manipulated the serial organization of talker turns (two levels: **Full** turns, with each

Condition	Description
FN	**Full** turns, **Normal** speech
FA	**Full** turns, **Accelerated** speech (100% faster)
IN	**Interleaved** turns, **Normal** speech
IA	**Interleaved** turns, **Accelerated** speech (100% faster)

Table 1: A summary of the four experimental conditions and their coded designations.

turn being a full commentary vs. **Interleaved** turns, with each turn being an utterance). The second factor manipulated each talker's speaking rate (two levels: **Normal** speech vs. **Accelerated** speech). Table 1 summarizes the manipulations in each of the four conditions and serves as a key for their coded designations in the remainder of the paper. Overall listening performance was predicted to be best in condition **FN**, and progressively worse in conditions **FA**, **IN**, and **IA,** in that order.

The treatments and listening materials were organized in the following way. The 16 commentaries developed for the main listening exercises were divided into four groups of four commentaries with one from each of the four talkers. These four groups were used for the four listening exercises each participant carried out. Participants were assigned to one of four different treatment orders based on a 4x4 latin square, in the order of their enrollment. Further, to ensure that all pairings of treatments and commentaries appeared in the study an equal number of times, each order of treatments was combined with a different ordering of the four commentary groups (also based on a 4x4 latin square).

Silent pauses of pre-defined lengths were inserted between clips at runtime in each of the listening exercises, as well as in the training sessions, to simulate natural pauses talkers frequently add between clauses and sentences in normal speech. The lengths of inserted pauses were proportional to the speed of the speech materials: 400 ms was used for pauses in normal speech and 200 ms for pauses in accelerated speech.

In each of the listening exercises involving full turns (the **FN** and **FA** conditions), commentaries were presented from left to right. Thus, the sequence of clips corresponding to the first talker's full commentary were played in order, with pauses inserted between them, followed by the next talker's full set of clips and inserted pauses, and so on, until all four commentaries had been aired. In contrast, in the listening exercises involving interleaved turns (the **IN** and **IA** conditions) the following algorithm was used to alternate among each of the talkers' commentaries: The first talker was chosen at random, and the first clip from the corresponding commentary was removed from its sequence of utterances, played with a pause inserted at the end, and followed by an additional gap of 300 ms (this is the "brief gap" discussed in Section 1.2 above). This set of actions completed the first "interleaved" turn. Each successive clip was then selected from the commentary with the greatest amount of time remaining and played in the same manner as the first clip. In the event of a tie (e.g., two or more commentaries had an equal amount of time remaining), the next talker was again chosen at random. This procedure continued until all four sequences of utterances were exhausted. The addition of the 300 ms gap after each interleaved clip and its inserted pause made the net pause between interleaved utterances 700 ms in the **IN** condition and 500 ms in the **IA** condition. 300 sec. gaps were not added in the **FN** and **FA** manipulations because full turns allowed listeners to focus on each talker for over two minutes at a time, and all of the commentaries had a clear beginning, middle, and end.

2.2. Dependent Measures

In the series of studies this experiment is part of, the participant's task of listening for information is regarded as having two successive stages of perceptual performance: aural attention and aural comprehension. Neither of these functions is directly observable, so indirect techniques are needed to estimate how well the listener discharges them. As in [5] and [1], phrase recognitions and sentence judgments are used for this purpose.

2.2.1. Attention

The first response task, which required participants to recognize specific noun phrases in the speech materials (see Section 2.1.1), is used here as a measure of attentional performance—specifically, how well listeners were able to attend to and identify what each of the talkers said during the listening exercises. The use of targets and foils in this task allows performance to be scored in two ways—as a proportion of correctly identified targets and rejected foils and, alternatively, as a *d′*. The latter, which is reported here, is a signal detection sensitivity score derived from the respective rates of "hits" (targets correctly identified) and "false alarms" (foils marked as targets). *d′* can be thought of as the distance between the means of the observed distributions of hits and false alarms. Higher values for this measure indicate that listeners marked many of the targets and very few of the foils[1].

2.2.2. Comprehension

Aural comprehension performance is measured here as the combined proportion of sentences participants correctly judged to be consistent or inconsistent (i.e., "old" or "new"; see Section 2.1.1) with the speech materials they had just listened to in each of the experimental manipulations. Because a strong correspondence between respective patterns of attention and comprehension performance was previously observed in this series of experiments (see [5] and [1]), a similar correspondence was expected in the present study. Other measures associated with listeners' sentence judgments are their confidence scores—a self-reported measure of how certain they were about each judgment, ranging from "not at all" to "very" (see Section 2.1.1)—and the number of "I don't know" responses each listener made. Analyses of these data will be reported elsewhere.

3. RESULTS

A two-factor, repeated measures analysis of variance, with two levels for each factor (presentation: Full vs. Interleaved turns;

[1] *d′* was calculated with substitute fractional rates of 1-(1/(2N)) and 1/(2N) for listeners with a perfect hit rate of 1 and/or a false alarm rate of 0, using the number of targets or foils for N.

and speaking rate: Normal vs. Accelerated speech), was performed for each of the dependent measures derived from the response task data. Performance in each of the treatments was largely consistent with the expected pattern of differences.

3.1. Attention

There were significant main effects of speaking rate and presentation on participants' *d'*s, which index aural attention performance (the ability to follow what was said) in terms of how often participants chose targeted noun phrases and incorrectly chose foils as they listened to the commentaries. Specifically, phrase discrimination was hurt by accelerating the rate of speech, regardless of whether talkers took full or interleaved turns (F (1, 15) = 98.15, $p < 0.001$, $\eta^2 = 0.867$). Additionally, performance fell when talkers took interleaved turns, regardless of the rate of speech (F (1, 15) = 4.98, $p = 0.041$, $\eta^2 = 0.249$). There was no interaction between the factors ($p = 0.72$). Figure 3 shows mean d' scores plotted by presentation and speech rate.

3.2. Comprehension

Participants' mean scores in the comprehension response task are given in Figure 4. The proportion of correct sentence judgments participants made after listening to the commentaries in a given exercise dropped significantly when the rate of speech was doubled (F (1, 15) = 37.8, $p < 0.001$, $\eta^2 = 0.716$). As the plots in the figure show, accelerated speech undermined how well participants were able to decide if representative sentences were consistent with their understanding of the speech materials when talkers took full and interleaved turns. In contrast, there was no main effect of presentation—comprehension performance was not significantly impacted when commentaries were displayed as a series of interleaved turns among talkers ($p > 0.10$). Additionally, there was no interaction between factors ($p > 0.10$).

4. DISCUSSION AND CONCLUSION

The first and most pressing question the present study intended to address is the effect of serial interleaving (dividing what multiple talkers have to say at the same time into an alternating sequence of turns) on the ability of listeners to keep track of and

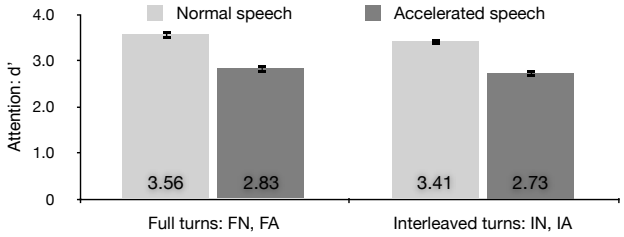


Figure 3: Mean aural attention performance, indexed by the signal detection score *d'*, showing the extent of participants' ability to recognize targeted noun phrases and minimize the selection of foils (phrases not present in the speech materials) while listening in each of the experimental treatments. Higher scores indicate better performance. Error bars show the standard error of the mean (s.e.m.).

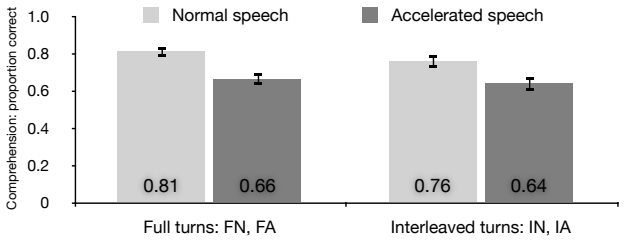


Figure 4: Mean aural comprehension performance as measured by the proportion of representative sentences participants correctly judged as consistent or inconsistent with their understanding of the spoken materials after listening in each of the experimental treatments. Error bars show the s.e.m.

understand the import of each thread of spoken information. The motivations for examining this way of intercepting and organizing competing contexts of speech are the inherent performance costs of attending to them at the same time and, conversely, the likely operational drawbacks of listening to each at length and one-at-a-time.

Serially interleaved listening reconciles the requirements of competing information priorities and alleviates the more difficult work of divided attention by allowing one context to be interrupted by another and resumed later. However, it also poses all of the challenges of sequential multitasking for listeners. Consequently, listening performance in the study's comparison of commentaries spoken in full turns and in interleaved turns was expected to be somewhat worse in the latter two manipulations because of the disruptive effects of repeated interruptions. As it turned out, though, while interleaving did have a significant impact on listeners' aural attention scores, the effect was not large, and, surprisingly, there was no corresponding effect of interleaving on listeners' comprehension performance at all.

Several theoretically motivated elements in the design of the listening task (outlined above primarily in Section 1.2) may have contributed to this outcome, including: the insertion of 300 ms gaps between interleaved turns; the external contextual cues provided by each talker's aural location and corresponding onscreen phrase lists (as well as linguistic cues in these displays); how the commentaries were divided into separate utterances; and the wide spatial separations between talkers in a stable virtual listening environment. If this is the case, it suggests that while serial interleaving necessarily imposes attentional costs on listeners, it can, in fact, be designed and displayed in ways that help to ameliorate the more decisive performance tolls sequential multitasking can potentially levy on tasks, particularly, functional loss of contextual understanding.

Although the second outcome of the study—the significant impact of accelerated speech on both measures of listening performance, regardless of how turns were organized—was not wholly unexpected, it also included an unanticipated development that may be a consequence of workload and how performance was measured. The decision to compare listening to normal and 100% faster speech in the study's design was made on the premise that acceleration rate should be at or just above the range where empirical performance begins to fall (e.g., per [1]). Moreover, because interleaving and accelerated speech were both expected to produce performance effects, an

important aim of the study was to evaluate how profoundly the upper end of effective accelerated speech might hurt serially interleaved listening performance. What was unexpected was that accelerated speech, rather than interleaving, would be responsible for the largest effects in the study (thus the anticipated order of performance declines across manipulations given in Section 2.1.3). A plausible explanation for this result, though, may be tied to differences in the respective ways aural attention and comprehension were measured here and in [1] and [11]. In [1], in particular, the method and specific manipulations were much the same as the **FN** and **FA** treatments above. However, in [1], participants only listened to two commentaries per exercise, which suggests that the use of four talkers per exercise here may have increased the workload associated with the response tasks enough to impair both measures of performance with faster speech at or near previously observed ceilings.

Still, to place the study's key performance result in context, it is worth noting that while the combined impacts of interleaving and accelerated speech respectively reduced attention and comprehension performance in the **IA** manipulation to a mean *d′* of 2.73 and to a mean proportion of correct sentence judgments of 0.64, both of these scores are substantially higher than the corresponding scores for the two and four concurrent talker conditions reported in [5]. In those manipulations, listeners' mean *d′*s were 1.93 and 1.45, respectively, and their mean proportions of correct sentence judgments were respectively 0.47 and 0.25.

Analyses of additional measures collected in the study will be reported at a later date. Future research on the applied use of this framework should begin with issues raised by more operationally realistic speech materials and performance questions raised by its integrated use in a mixed-purpose auditory display.

5. REFERENCES

[1] D. Brock, C. Wasylyshyn, B. McClimens, and D. Perzanowski, "Facilitating the watchstander's voice communications task in future Navy operations," in *Proceedings of the 2011 IEEE Military Communications Conference (MILCOM)*. Baltimore, MD. November 7-10, 2011.

[2] 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.

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

[4] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in Cognitive Sciences*, 12, 182-186, 2008.

[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] G. A. Miller and J. C. R. Licklider, "The intelligibility of interrupted speech," *J. Acoustical Soc. Am.*, 22(2):167–173, 1950.

[7] W. D. Garvey, "The intelligibility of speeded speech," *J. Exp. Psychol.*, vol. 45, no. 2, pp. 102–108, 1953.

[8] B. Arons, "Techniques, perception, and applications of time-compressed speech," *Proc. 1992 Conf., Am. Voice I/O Soc.*, pp. 169-177, 1992.

[9] 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.

[10] A. Wingfield, J. E. Peelle, and M. Grossman, "Speech rate and syntactic complexity as multiplicative factors in speech comprehension by young and older adults," *Aging, Neuropsychology, and Cognition, 10,* 310 –322, 2003.

[11] 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, June, 2010.

[12] S. S. Reich, "Significance of pauses for speech perception," *J. of Psycholinguistic Res.*, 9(4):379–389, 1980.

[13] D. D. Salvucci, N. A. Taatgen, and J. P. Borst, "Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption," in *Human factors in computing systems: CHI 2009 conference proceedings*, New York, NY: ACM Press, 2009, pp. 1819–1828.

[14] D. E. Kieras, D. E., Meyer, J. A. Ballas, and E. J. Lauber, "Modern computational perspectives on executive mental processes and cognitive control: Where to from here?" in S. Monsell & J. Driver (Eds.), *Control of Cognitive Processes*, Cambridge, MA: MIT Press, 2000 pp. 681-712.

[15] D. D. Salvucci and N. A. Taatgen, *The Multitasking Mind*, Oxford University Press, 2011.

[16] J. P. Borst, N. A. Taatgen, and, H. Van Rijn, "The problem state: A cognitive bottleneck in multitasking," *J. Exp. Psychol.: Learning, Memory, & Cognition*. vol. 36, no. 2, pp. 3363-382, 2010.

[17] E. M. Altmann and J. G. Trafton, "Memory for goals: An activation-based model," *Cognitive Science*, 26, 39-83, 2002.

[18] J. G. Trafton, E. M. Altmann, D. P. Brock, and F. E. Mintz, "Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal," *Int. J. of Human-Computer Studies*, 58, 583-603, 2003.

[19] E. M. Altmann and J. G. Trafton, "Timecourse of recovery from task interruption: Data and a model," *Psychonomic Bulletin & Review*, 14, 1079-1084, 2007.

[20] V. Best, F.J. Gallun, A. Ihlefeld, and B.G. Shinn-Cunningham, "The influence of spatial separation on divided listening," *J. Acoust. Soc. Am.*, vol. 120, no. 3, pp. 1506–1516, Sept. 2006.

[21] J.M. Royer, C.N. Hastings, and C. Hook, "A sentence verification technique for measuring reading comprehension," *J. Reading Behavior*, vol. 11, no. 4, pp. 355–363, 1979.

[22] A.W. Mills, "On the minimum audible angle," J. Acoust. Soc. Am., vol. 30, no. 4, pp. 237–246, Apr. 1958.