EVALUATING LISTENERS' ATTENTION TO AND COMPREHENSION OF SPATIALIZED CONCURRENT AND SERIAL TALKERS AT NORMAL AND A SYNTHETICALLY FASTER RATE OF SPEECH

Derek Brock, Brian McClimens, J. Gregory Trafton, Malcolm McCurry, and Dennis Perzanowski

U.S. Naval Research Laboratory
Washington, DC 20375
[derek.brock, brian.mcclimens, greg.trafton, malcolm.mccurry, dennis.perzanowski]@nrl.navy.mil,

ABSTRACT

Concurrent voice communications workload has been identified as a pivotal issue for desired reductions in the size of Navy watchstanding teams on future platforms. Without effective augmenting technologies, real increases in current per-person communications monitoring requirements will lead to unacceptable reductions in operator performance. A proposal to buffer voice communications and monitor them serially at synthetically increased rates of speech has recently been put forward as an alternative to concurrent monitoring. However, any decrements in listening performance associated with temporal scaling must be weighed against the costs of current practices. A comparative study reported here examines measures of auditory attention and comprehension in different multitalker contexts using long blocks of continuous speech. In four conditions, listeners respectively heard two and four concurrent talkers and four serial talkers (i.e., one at a time) speaking normally and 75% faster. With only a few exceptions, all pairwise differences between measures were significant. Performance in the faster serial condition was lower than in the normal serial condition, but was found to be greater than in either of the concurrent conditions by a substantial margin.

1. INTRODUCTION

Auditory display research at the Naval Research Laboratory (NRL) is principally motivated by the increasingly complex information workload Navy watchstanders routinely face. In addition to responsibilities imposed by tactical displays and newly introduced "net-centric" information technologies, speech communications are a vital component of watchstanding operations, and will continue to be so for the foreseeable future. Shipboard command operations actively work with over twenty radio-telephone, satellite, and internal circuits, and in the division of labor, present-day systems coordinators frequently must manage two or more concurrent channels of voice communications that are central to the functions of their position. When asked, watchstanders readily admit to the challenges of this auditory task, but confide that success is possible because of domain knowledge, information predictability and repetition, the intermittent nature of communications concurrencies, and redundant monitoring by other members of their watchstanding team.

As the Navy strives to optimize command operations in the design of its new platforms through automated decision support and human-centric control strategies, voice communications workload has been identified as a pivotal issue for reductions in the size of watchstanding teams. In particular, it is recognized that in the absence of effective augmenting technologies, real increases in current per-person communications management requirements will lead to unacceptable reductions in operator performance. Spatialized audio and speech-to-text methods were introduced in a 2001 Navy human-systems integration study that examined the effects of monitoring up to four critical voice communications circuits at a time on the ability of operators to maintain situation awareness in a realistic tactical scenario. Both technologies were found to have valuable operational utilities, but neither proved to be effective for countering performance declines associated with increased communications workload [1].

Another proposal, recently advanced at NRL, is to explore technical strategies for serialized communications monitoring as an alternative to concurrent monitoring [2]. Because of varying rates of service, parallel channels of spoken information can be readily buffered and presented to listeners in a serially interleaved manner. Additionally, buffering allows normal rates of speech to be synthetically increased. Depending on the contributing rates of service, this has the potential to allow serial monitoring to be accomplished in roughly the same amount of time concurrent monitoring would require, albeit with an inherent processing delay. NRL has thus mounted a multidisciplinary effort to prototype a communications serialization scheme and study its potential for operational use.

Serial monitoring involving compressed rates of speech raises issues for human performance that differ with the demands of concurrent monitoring. Presentation in a possibly mixed-use immersive auditory display will be required to ameliorate the potential for source confusions particularly at transition points between talkers on different circuits (c.f. [3]). 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. Just as importantly, temporal scaling of speech rates may adversely affect objective measures of listening performance, and any decrement must be weighed against other monitoring strategies. An initial comparative analysis of this last issue is the focus of this paper.

The speech materials used here are edited segments of radio broadcasts made by four professional commentators. This category of talk has several advantages over other materials that were considered, both for the aims of the study and for the targeted population of listeners. In particular, the speech is even throughout and well enunciated, and none of the talkers have strong regional accents. Further, each segment covers a familiar, easily understood topic and is spoken by a single commentator. These regularities and the use of non-specialist language simplified the study's performance design and helped to minimize a number of potential confounds. Corpora of Navy, air traffic control, and emergency services communications differ from these straightforward materials in important ways that are not addressed here. Among the issues that must be

explored in future studies are varying rates of service, more than one talker on a channel, and rapid transitions between circuits.

2. METHOD

Twelve participants, three female and nine male, 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 a large flat-panel monitor and the auditory component was rendered binaurally in Sony MDR-600 headphones. Four listening exercises—one for each experimental condition—were presented to all of the participants in counterbalanced order. Each was preceded by a short training session that resembled the format of the formal exercise that followed. The training sessions allowed listeners to become familiar with each of the auditory manipulations and the listening requirements.

2.1. Apparatus

2.1.1. Response Tasks

In both the training sessions and the formal listening exercises, participants carried out two response tasks, one while listening and the other immediately after. All of the listening materials were edited segments of broadcast radio stories available on the internet. The first response task involved hearing noun phrases in the spoken material and marking them off in onscreen lists that corresponded to each of the stories being presented in the current segment of the study. The lists contained both targeted phrases and foils in equal numbers (eight targets per story in the training sessions and twenty targets per story in the formal exercises). Targets were listed in the order of their spoken occurrence and were randomly interleaved with no more than three intervening foils; foils were selected from radio stories on similar but not identical topics. In the second response task, participants were given a series of sentences to read and asked to indicate whether each contained "old" or "new" information based on what they had just heard [4]. "Old" sentences were either word-for-word transcriptions or semantically equivalent paraphrases of story sentences. "New" sentences were either "distractors"—topic-related sentences asserting novel or bogus information—or story sentences altered to make their meaning inconsistent with the content of the spoken material. An example of each sentence type developed from a story on home buying in Washington, DC is provided in Table 1. In addition to responding "old" or "new," participants could also demur (object to either designation), by responding "I don't know." Only two sentences, one old and the other new, were presented for each story in the training sessions. In the formal exercises, eight sentences per story (two of each of the old and new sentence types) were presented.

2.1.2. Listening Materials and Experimental Manipulations

Each participant listened to a total of 28 edited stories selected from the broadcast archives of two male and two female talkers. Half of the stories were used for the training sessions and the remainder for the formal exercises. The stories —and thus the talkers—chosen for each manipulation were the same for all participants. Music and other non-speech sounds were removed from the audio files with a sound editing tool and the spoken material was edited to make the length of each story appropriate for its use—short for training sessions (1 min.) and long (2 min. 45 sec.) for the formal listening exercises.

Stories in two of the four conditions were presented concurrently; in the remaining two, they were presented serially.

Sentence type	Example sentence	Designation
Original	For anyone, purchasing real estate in the nation's capital would be an ordeal.	Old
Paraphrase	For anybody, buying a home in Washington, DC would be an ordeal.	Old
Meaning change	For most people, purchasing real estate in the nation's capital is fun and easy.	New
Distractor	Real estate sales are down this year in the nation's capital.	New

Table 1: An example of each of the four types of sentences participants were asked to judge as "old" or "new" immediately after each listening exercise. Listeners were also allowed to demure by selecting "I don't know" as a response.

Equal numbers of male and female talkers were used in each condition. In the concurrent story conditions, listeners respectively heard two and four different talkers speaking at the same time. These conditions are referred to as "2C" and "4C" below. In the serial conditions, stories from each of the four talkers were played sequentially. The serial conditions were differentiated by their rate of speech. Talkers spoke at a normal rate in one and 75% faster in the other. These conditions are referred to as "4S" and "4SF" below. Listening performance was predicted to be best in condition 4S and progressively worse in conditions 4SF, 2C, 4C. The rate modulation in condition 4SF was synthesized with a speech analysis and synthesis technique developed at NRL known as "pitch synchronous segmentation" (PSS) [5]. A particular strength of the PSS method is that it largely preserves intelligibility, pitch, and other informational characteristics of human speech even at relatively high rates of compression. Table 2 summarizes the manipulations in each of the four conditions and serves as a key for their coded designations in the remainder of the paper.

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

Table 2: A summary of each the the four experimental conditions and their coded designations.

2.1.3. Auditory Display

When competing voices in a communications system are functionally co-located, as they are when their signals are mixed and presented monaurally or diotically in headphones, various measures of targeted listening performance are negatively impacted. In particular, the range of aural cues listeners ordinarily exploit to segregate different sources of speech is impoverished [6]. Thus, to ensure that participants could easily focus their aural attention on each of the stories in all four conditions, the talkers were rendered in a virtual listening space. Specifically, in each condition the talkers were respectively heard to come from separate apparent locations on the

horizontal plane in front of the listener that corresponded to the ordered positions of the onscreen lists of target and foil phrases. Binaural filtering with a non-individualized head-related transfer function was used for this purpose. In the three conditions in which four stories were presented, the talkers, from left to right, were respectively located at -60°, -10°, 10°, and 60°, with 0° being straight ahead. In the two-talker condition, only the -10° and 10° positions were used. This set of positions was chosen to exploit the pattern of human sensitivity to horizontal changes in the placement of sounds [7] and to make the location of each talker easy to discriminate and remember. (See also the use of a similar configuration in [8]). Shortened examples of what participants heard in each condition are given in the following sound files. The files for conditions 2C and 4C (O_46-1.WAV and O-46-2.WAV) are 15 second clips. The files for conditions 4S and 4SF (O_46-3.WAV and O_46-3.WAV) have been edited so that only one sentence from each talker is presented.

2C [O_46-1.WAV]

4C [O_46-2.WAV]

4S [O_46-3.WAV]

4SF [O_46-4.WAV]

2.2. Dependent Measures

Participants' listening performance in each condition was evaluated on the basis of phrase identifications and post-listening sentence judgments. Both response tasks can be viewed as two-choice discrimination problems.

2.2.1. Phrase Discrimination While Listening

The phrase identification task is primarily intended to be an attentional measure, that is, a measure of how well the listener is able to attend to and identify what each of the talkers is saying during the listening exercises. To some extent, it may also be construed as a measure of intelligibility due to the interference of speech that is not being attended to in the concurrent talker conditions and the participants' presumed lack of substantial experience with faster speech (i.e., speech modulated by the PSS technique) in one of the two serial conditions. Interference from other sources of noise was not a factor in the study and no consideration was given to phonetic properties in the choice of noun phrases used for targets.

Respective counts of marked and unmarked target and foil phrases in each condition were binned as a signal detection paradigm [9]. Two performance measures were calculated from these counts: the combined proportion of correctly identified targets and rejected foils, denoted $p(c)$, and d', a signal detection sensitivity score derived from the respective rates of "hits" (targets correctly identified) and "false alarms" (foils marked as targets).

2.2.2. Post-listening Sentence Judgments

Complementing the phrase identification task, the sentence judgment task is intended to be a measure of comprehension. A potential problem with the use of phrase lists while listening is that the procedural effort associated with phrase identifications might interfere with the ability of listeners to encode and retain understanding of content at the discourse level. If this is the case, post-listening measures of content understanding should be comparatively poor across conditions. Otherwise, measures of comprehension can be expected to be well correlated with phrase identification performance.

Three measures were calculated from participants' sentence judgments in each condition. The first, denoted $p(c)$, is the proportion of sentences correctly judged as "old" and "new." The second, d_a, is an alternative signal detection sensitivity parameter that is appropriate for use when listeners are allowed to give intermediate responses, such as the option "I don't know" used in the present study [10]). And the last, denoted $p(i)$, is the proportion of "I don't know" responses (demurs). This is calculated as a percentage of sentences presented for verification in each condition.

3. RESULTS

A four-level, single-factor analysis of variance was performed for each of the dependent measures derived from the response task data. Performance in each manipulation was consistent with the expected pattern of differences. The correlation between the proportion of correctly identified target and foil phrases and the proportion of sentences correctly judged as "old" and "new" was $r = 0.87$. Similarly, the correlation between the respective signal detection sensitivity measures for each response task was $r = 0.82$. Correlations of this size suggest that listeners' understanding of the content of the spoken material was comparatively unaffected by the demands of the target phrase identification task.

3.1. Phrase Identifications

Both measures of the participants' ability to carry out the phrase identification task varied significantly across the manipulations.

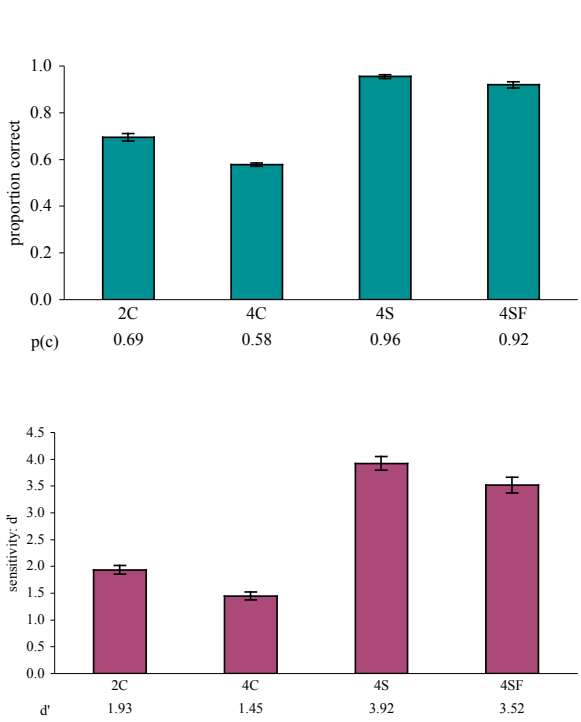

Figure 1. a) *Plot of the mean combined proportion of correctly indentified target noun phrases and rejected foils in each condition.* b) *Plot of the corresponding mean d' for each condition. Error bars show the standard error of the mean.*

For $p(c)$, $F(3, 33) = 335.29$, $p < 0.0005$, and for d', $F(3, 33) = 156.28$, $p < 0.0005$. The respective plots for these scores in each condition are shown in Figure 1a and b. In separate pairwise comparisons, all mean differences for $p(c)$ and d' were significant except for d' between conditions 4S and 4SF (0.405, $p < 0.078$).

3.2. Sentence Judgments

Like the phrase task, each of the post-listening measures of comprehension derived from the sentence judgment task varied significantly across the manipulations. For $p(c)$, $F(3, 33) = 64.09$, $p < 0.0005$ and all pairwise differences between means were significant. For d_a, $F(3, 33) = 30.53$, $p < 0.0005$. All

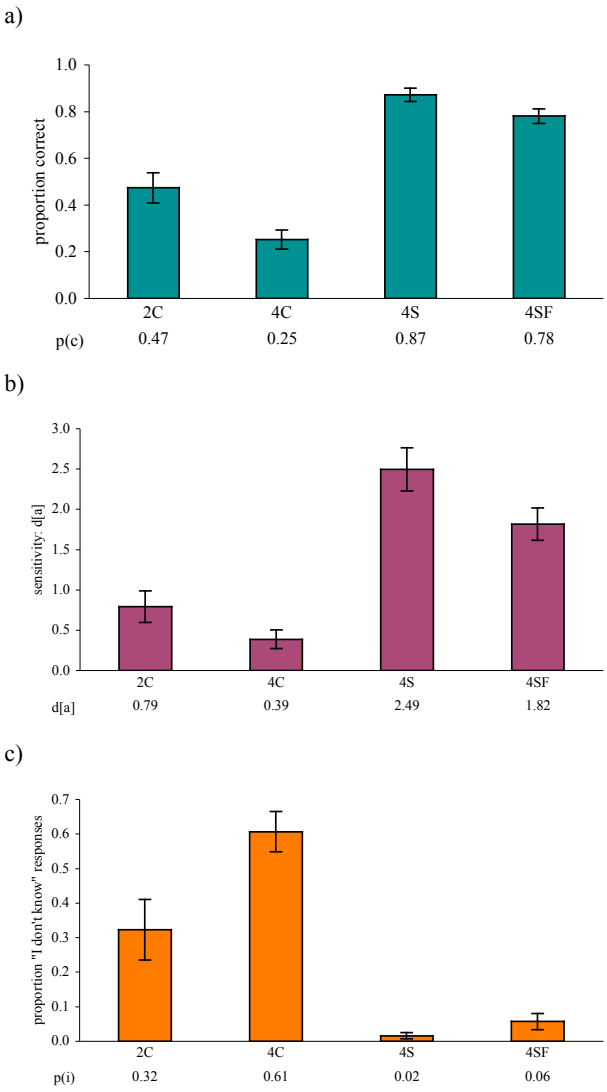

Figure 2. a) *Plot of the mean combined proportion of sentences correctly indentified as "old" and "new" in each condition.* b) *Plot of the corresponding mean d_a for each condition.* c) *Plot of mean proportion of "I don't know" sentence judgments in each condition (demurs). Error bars show the standard error of the mean.*

pairwise differences between the d_a means were significant except for conditions 2C and 4C (0.404, $p < 0.499$) and conditions 4S and 4SF (0.679, $p < 0.114$). For $p(i)$, $F(3, 33) = 44.89$, $p < 0.0005$. All pairwise differences between the $p(i)$ means were significant except for conditions 4S and 4SF (0.042, $p < 0.148$). The respective plots for these scores are shown in Figure 2a, b, and c.

4. DISCUSSION

The primary goal for this listening study was to assess how well listeners can usefully comprehend—or "ground," in the sense of encoding the meaning of spoken information for later use—what multiple talkers, speaking on separate topics, are saying in different presentation contexts. To summarize: the contexts employed here, rendered in a virtual listening space with the talkers arrayed from left to right, were concurrent speech involving two and four talkers, and serial speech involving four talkers speaking normally and 75% faster. To gauge what listeners are able to ground in these contexts, participants were asked first to listen for and indicate the presence of selected noun phrases in the spoken material. Second, after listening, they were asked to judge whether four types of sentences, derived from what was said, were "old" or "new," in the sense that the meaning of each sentence was or was not consistent with what was declared in the spoken material.

The resulting measures of performance in each presentation context provide clear empirical evidence of a) the range of difficulties listeners experience when asked to attend to and comprehend what more than one person is saying at the same time, b) the crucial difference between overall listening performance in concurrent and serial talker settings, and c) the intermediate effect an increased rate of speech has on auditory attention to and comprehension of serial talkers.

4.1. Listening to Multiple Concurrent Talkers

To show overall accuracy in the phrase identification task, the proportions plotted in Figure 1a are the sum of two measures: correctly indentified target noun phrases (hits) and correctly rejected foils. Performance in both of the concurrent talker conditions, 2C and 4C, is above 50%, but was undermined in the latter by the increase from two to four talkers. There was an almost uniformly low rate of false alarms (foils marked as targets) in each of the manipulations. Consequenty, nearly all of the variance in these measures is due to respective differences in hit rates. This very low incidence of false alarms was unexpected, but it is likely that the methodological rejection of foils by default was a contributing factor, especially in the concurrent speech manipulations where greater information densities and larger response lists made additional demands on performance. However, given the nearly total absence of false alarms in this response task, it appears that even in contexts that require substantially divided auditory attention, listeners are quite good at distinguishing between verbal information that is and is not present in their immediate auditory memory.

The application of signal detection theory to the target-foil response data is an attempt to quantify the impact of each presentation context on the auditory demands of the experimental task, specifically, the degree to which listeners were able to confirm that selected spoken phrases were present in the auditory materials. Although assessment of each item in the phrase lists that participants marked is functionally a two-choice discrimination problem, the organization of the task they were asked to perform does not conform to the ordinary assumptions and step-by-step order of a standard "yes-no" experiment. In particular, the notion of a "trial" is ill-defined: listeners were given unconstrained access to ordered decision lists and the corresponding targets were embedded in a streaming medium with different information rates and densities

in each condition. Thus, d' should not be construed here as a classically derived measure of sensitivity, but rather as an index of context-mediated attentional performance derived from response data binned as a yes-no discrimination task.

To elaborate, a summary of the participants' apparent allocation of attention and mean target detection counts in conditions 2C and 4C is given in Table 3. In both of these conditions, the listening task required participants to divide their auditory attention between mulitple talkers, each located at a separate position in the listening space. The averages in the right-most column show that the listeners were unable to detect as many targets in condition 4C as they did in 2C. (Only one false alarm was made in each condition.) However, the similar size of these counts suggests that participants may have been performing near the limit of their attentional abilities. This is reinforced by examining how and where their attention was allocated in each condition. In contrast to a roughly even division of attention when only two talkers were present, listeners resorted to a predominately far-left-and-far-right listening strategy when confronted with four talkers—presumably, because this was the easy thing to do—and left attention to the middle to whatever else they could manage. Not surprisingly, the contrasts in this data provide evidence of a performance cost associated with the additional attention switching and larger response apparatus that were mandated by a doubling of the concurrent task load in condition 4C (c.f., [11]). Hence, what d' indexes in these conditions (relative to performance with a single talker) is the combined costs of information density, attention switching, and response load on detection performance. Were there little or no cost for having to rove among talkers and phrase lists, the change in d' from condition 2C to 4C would primarily reflect the difference in the number of concurrent targets to detect. In particular, a mean hit rate in 4C equal to half the rate in 2C would only decrease d' by about 0.3. But participants were unable to make as many detections in 4C as in 2C—despite the availability of twice as many targets—resulting in a larger reduction in d' (approximately 0.5 as opposed to 0.3). This larger difference suggests that additional executive and physical overhead is required when auditory attention and response effort must be divided among a greater number of sources.

Like the target phrase identification measures for conditions 2C and 4C, the corresponding measures of comprehension from the sentence judgment response task are comparatively poor. However, the $p(c)$ measure for these conditions, plotted in Figure 2a, can arguably be interpreted as response data corresponding to the grounding, or comprehension, of roughly "one talker's worth" of spoken information. This interpretation is compelling in part because it is consistent with Broadbent's selective model of auditory attention [12], which posits that semantic processing of simultaneous auditory signals is limited to serial episodes of exclusive attention to each stream. In effect, only one auditory stream can be regarded at a time. Listeners also have fleeting access to receding portions of competing streams via traces in immediate memory, but selective attention to memory precludes selective attention to an incoming signal. (See [13] and [11] for recent studies addressing the role of immediate auditory memory in divided attention.) Thus, auditory attention to spoken information involving understanding or other forms of semantic processing must be interleaved between live sources and immediate memory when more than one talker must be attended to at the same time.

If this explanation is correct, then listeners, on average, should be able to ground little more than 1/N of a volume of continuous spoken information presented by N concurrent talkers because of persistent competition, evanescing auditory memory, and the imperative to keep up with what is being said. Conditions 2C and 4C are instances of this particular context.

Condition	Mean detection location counts				Mean number of detections
	talker-1	talker-2	talker-3	talker-4	
2C		8.42	7.33		15.75
4C	5.67	0.67	2.42	4.08	12.83

Table 3: Mean counts of target phrase detections made by listeners at each location in the two conditions involving concurrent talkers, and the mean number of detections made in each condition. There were 40 targets to detect in condition 2C and 80 in 4C.

While 1/N corresponds to approximately "one talker's worth" of information, it need not all be from the same source. An earnest listener is likely to have selectively processed information from each of the N sources in varying amounts and come away with a mosaic of partial understandings that, in sum, correspond to roughly 1/N of the information presented.

To quantify comprehension, listeners can be asked to make use of, and/or verify, what they have and have not grounded in comparable samples of information from each talker. The sentence judgment response task used here relies on the selection and semantic manipulation of an equal number of roughly overlapping sentences drawn from each talker plus the addition of invented topical sentences that are inconsistent with the declarative information conveyed in the spoken material (see Section 2.1.1 and Table 1). Correct assessment of a sentence as "old" or "new" demonstrates use of encoded meaning. Demurs (responding "I don't know") serve to verify what a listener deems he or she does not know or, at least, cannot judge. Allowing demurs also helps to reduce unintended response entries and minimizes the inclusion of empty guesses in the measure of $p(c)$.

The respective proportions of demurs in conditions 2C and 4C, i.e., the measures of $p(i)$, shown in Figure 2c, are substantial relative to the corresponding proportions of correct sentence judgments, $p(c)$, in these two conditions. The size and pattern of these measures, 0.32 in 2C and nearly double at 0.61 in 4C suggest that listeners were systematically unable to usably retain (much less ground) significant amounts of competing speech. As the number of talkers increased, so did the amount of information that was lost. The complementary pattern of $p(c)$ measures, 0.47 in 2C and 0.25 in 4C is similarly systematic with respect to the task load. Roughly one half of the sentences were correctly judged in 2C and one quarter in 4C. Again, these numbers notably correspond to 1/N of the information presented in each condition.

But participants in both conditions also made a number of incorrect sentence judgments—judging some old sentences as new and vice versa. In condition 2C, the mean proportion of errors is 0.20 and in 4C it is 0.14. These responses demonstrate that the ability of listeners to know and use what they have attended to is often graded. Indeed, grounding is generally understood to be subject to this qualification. Thus, any sentence a participant chose not to demur on arguably represents information that that listener managed to selectively attend to *to some degree*. Some judgments will be the result of secure knowledge and others will be the result of an informed guess. Some proportion of the latter will be wrong, hence the residual mean error rates shown above. Any informed guesses that prove to be right, then, are included in the count of correct judgments, the point being that $p(c)$ cannot be construed as a pure measure of what the listener fully grounded. Lacking more finely graded response data with which to partition the range of judgments, e.g., a Lickert scale, a reasonable strategy is to assume that participants made equal numbers of right and wrong responses that qualify as informed guesses. Designating

the mean error rates from above as $p(e)$, a more realistic, if conservative, measure of mean reliable comprehension would then be $p(c) - p(e)$. In contrast, the sum $p(c) + p(e)$ provides a more informative measure of how much information listeners, on average, were able to selectively attend to, regardless of their encoding success. This later formula may partially account for why people often subjectively feel that listening contexts similar to 2C are feasible.

The signal detection measure d_a plotted in Figure 2b addresses the question of grounding from a different perspective. Here, demurs are assigned an alternative interpretation and treated as an intermediate response between sentences judged to be "old" and "new." Sentences correctly judged as old are counted as hits and those that are accurately marked as new are counted as correct rejections. The resulting sensitivity measure is not an index of what was attended to. Instead, it is an index of how well the listener, after the fact, is able to recognize verbatim or semantically equivalent representations of what he or she is presumed to have internalized. Unlike the methodological caveats for the target-foil d' analysis above, the basis for the present analysis is a true one-interval design. An important consequence of this difference is that time and the size of the response apparatus are not limiting factors, so there are no default responses, as there assuredly are in the target-foil analysis. Hence, the values of d_a reported here are diagnostic of grounding success, although there do appear to be substantive differences in the way listeners' internal processes for recognizing old and new information are distributed, and especially so after listening to concurrent speech. Comparison of the measures for conditions 2C and 4C show that the mean detectability of quotes and paraphrases from each of the talkers fell as the number of talkers was doubled; however, the pairwise difference between these scores does not reach significance. The relative proximity of both scores to zero (i.e., chance) suggests that it is generally quite difficult for listeners to distinguish between spoken information that was correctly encoded in a context requiring divided attention and lexically similar, but semantically different information that was supposedly uttered instead.

4.2. Concurrent vs. Serial Talkers

All of the plots in Figures 1 and 2 show a pronounced performance difference between the concurrent talker conditions, 2C and 4C, and the serial talker conditions, 4S and 4SF. The pattern of differences is not surprising, but the point of the manipulations was to develop a baseline of comparative measures involving relatively long blocks of continuous speech in each context. Continuous speech was chosen over intermittent speech because it encouraged listeners to divide their attention as equally as possible throughout the concurrent exercises. In particular, fully simultaneous speech tends to occur episodically rather than continuously in real-world contexts where multiple talkers compete, and this naturally affords intermittent opportunities for listeners to both divide their attention among fewer talkers and to use their immediate auditory memory more frequently to catch up. Using continuous speech minimizes perceptual advantages of this sort and places emphasis on systematic properties of divided listening that might otherwise be obscured by their presence, particularly attentional strategies and grounding difficulties.

With the exception of condition 2C, equivalent numbers of stimuli were used in the serial talker conditions to make principled comparisons with the concurrent talker manipulations and to evaluate performance with a preliminary instance of the proposed multitalker monitoring solution involving synthetically faster rates of speech. However, the four manipulations do not vary as a group across a single, unifying dimension. Instead, they are best conceptualized in terms of two and three member subgroups. 4C and 4S, in particular, together with 4SF, represent manipulations in which equivalent amounts of spoken information were presented to listeners. 2C was included in the study because its analogue is commonly encountered in operational practice and it provides a relevant intermediate manipulation between one and four talkers at a time.

The extent of the differences between each of the performance measures for conditions 4C and 4S underscores the acute difficulty of highly divided listening. When participants were able to listen to each of the four talkers speak one at a time in 4S, their identification of target phrases (i.e., their hit rate) averaged 91% as opposed to 16% in 4C. This large change in hit rates while listening accounts for the substantially different measures of d' in each condition. Serial presentation also had a substantial impact on listeners' grounding success. The conservative grounding measure, $p(c) - p(e)$, developed in section 4.1, is 0.76 for 4S as opposed to 0.11 in 4C. Corresponding to this result are an extremely low proportion of demurs in 4S at 0.02, and a d_a of 2.49, which reflects a moderately high sensitivity for recognizing quotes and paraphrases from each of the talkers.

A valid critique of the methodology underlying the comparison of divided and serial listening contexts in this study is that the response demands of the phrase identification task were likely a factor in the size of the contrasts. The purpose of this response procedure was to measure attentional performance in each of the listening exercises, and in the concurrent conditions, this meant that phrase lists for all of the talkers had to be displayed simultaneously for listeners to mark with a pointing device. This added a substantial visual and motor component to the listening task in all of the manipulations, but particularly in condition 4C, where four phrase lists requiring point and click responses were arrayed from left to right on a cinema-style display. In 4S, by comparison, only one talker spoke at a time, each to completion, and the corresponding phrase lists were displayed separately, one after the other. Thus, the response demands in 4C were more than four times as great as in 4S and arguably may have interacted with the conduct of listeners' auditory attention.

If this interaction was substantial, the contrasts reported here likely do not fully reflect the empirical differences in attention and grounding that listeners may ordinarily realize. In particular, the respective differences in performance measures for 4C and 4S may not be quite as large. A methodological alternative for measuring attentional performance that was not considered before the study was run would be to enlarge the sentence judgment task, or alternatively, to ask participants to carry out the phrase identification task immediately after each listening exercise. This merits consideration for a future study, but it fails to address a crucial premise for measuring phrase identification performance in the midst of listening, which is that auditory attention is a precursor to encoding. Once encoding has occurred, it is subject to both erroneous reports and decay; in particular, any measure of attention taken after the fact of listening depends on encoding and is likely to include some level of guessing. Hence the practical questions addressed by the target phrase identification task appear to involve a necessary trade off.

4.3. Increasing the Rate of Speech in Serial Presentations

Listening performance in the 4SF manipulation was not as good as in 4S; moreover, the pairwise decline in $p(c)$ measures for phrase identifications and sentence judgments between these conditions were both significant. In the phrase identification task, in addition to a lower mean hit rate, there was a slight

uptick in the number of foils marked as targets. In the sentence judgment task, the mean proportion of demurs increased somewhat, as did response errors, with the latter reducing the derived measure of grounding success in section 4.1, $p(c)$ - $p(e)$, to 0.62. The corresponding pairwise differences between both mean signal detection measures, d' for phrase identifications and d_a for sentence judgments, did not reach significance, but the correlated pattern of respective declines indicates that spoken information presented serially, but rendered at synthetically faster rates of speech, is to some extent harder for listeners to attend to and harder to internalize. In contrast to these differences with 4S, all of the 4SF performance measures are significantly better than the corresponding measures in 4C and 2C by respectively wide margins.

4.4. Summary

The four conditions in the study address a variety of multitalker information contexts involving primarily concurrent and serial speech communications. The goal was to develop a set of comparative measures of attention and grounding in each of the listening contexts. Relatively long blocks of continuous speech were used in all of the manipulations to minimize opportunities for participants to exploit periods of silence or other irregularities that might allow them to alter or enhance their listening performance. The measures derived from the two response tasks employed to gauge listening performance vary significantly and in a correlated manner across the four conditions. With only a few exceptions, all of the pairwise differences between the measures were significant.

The measures for the phrase identification task participants carried out during each listening exercise were the mean proportion of correct responses and mean d'. d' is taken here to be an index of the net attentional demands in each listening context. However, it should be interpreted with substantial caution due to important concerns with its underlying assumptions. The measures for the post-listening sentence judgement task were similarly the mean proportion of correct responses, mean d_a, which is interpreted as an index of how well listeners were able to internalize the correct meaning of what was said, and mean proportion of demurs.

In all four conditions, there was an unexpectedly low rate of phrases incorrectly identified as present in the spoken material, which suggests that listeners tend to make reliable use of their immediate auditory memory. However, the default method of response may also be a factor in this result. Most of the variance in the $p(c)$ for the phrase identification task is thus due to the proportion of hits in this measure. The visual and motor demands of phrase identification responses may also have been a factor in certain contrasts. In particular, working with a larger response apparatus in the concurrent talker conditions may have interacted with the exercise of divided attention and thus diminished the proportion of hits. It is likely that differences in both d' and $p(c)$ for the phrase identification task, particularly those between 4C and the other conditions, reflect this interaction to some degree. The other contrasts in these measures are less subject to this concern. More broadly, the contrasts in the phrase response measures suggest that overhead associated with attention switching systematically limits selective listening performance in contexts that require progressively higher degrees of divided attention among concurrent talkers. When divided listening is not a concern, these same measures of attentional performance are exceptionally high. However, attentional performance in this latter context is less nearly optimal when normal rates of speech are synthetically increased.

The performance measures drawn from the sentence judgment task are interpreted as parameters of participants' information encoding success in each of the listening contexts. On the surface, contrasts in the $p(c)$ for these responses suggest that listeners are only able to correctly use about 1/N of the spoken information N competing talkers can continuously present in a given amount of time. A similarly systematic but inversely correlated pattern of contrasts is present in the proportion of demurs, $p(i)$, which is taken to be a measure of information that either was not attended to or that cannot be recalled well enough to judge, and was thus lost. The mean proportion of sentence judgment errors, which can be inferred from the proportional sum of correct responses and demurs, is interpreted as an indicator of the number of judgments that are informed guesses. Taking the average number of errors and correct guesses to be approximately equal, $p(c)$ - $p(e)$ and $p(c)$ + $p(e)$ are respectively proposed as mean proportional measures of reliable comprehension and information attended to but not necessarily grounded. These measures are shown below in Table 4. Construal of the mean d_a as an index of internal access to what was supposedly conveyed in each condition corroborates the broader implications of this analysis. In particular, the contrasts in the sentence judgment data suggest that, by a wide margin, spoken information is significantly less reliably encoded for later use in contexts requiring divided listening than in contexts where divided listening is avoided. Encoding reliability in the latter context is also significantly impacted by a synthetic increase in the rate of speech, but is still significantly better than in concurrent multitalker environments.

Condition	$p(c)$ - $p(e)$	$p(c)$ + $p(e)$
2C	0.27	0.68
4C	0.11	0.39
4S	0.76	0.98
4SF	0.62	0.94

Table 4: Indicators of post-listening information encoding performance derived from the sentence judgement response data. $p(c)$ - $p(e)$ and $p(c)$ + $p(e)$ are respectively proposed as mean proportional measures of reliable comprehension and information attended to but not necessarily grounded.

5. CONCLUSIONS

If the Navy is to optimize command operations on future platforms, limitations imposed by current voice communications workloads must be overcome. The present findings suggest that serialized monitoring at synthetically faster rates of speech deserves further exploration as a possible alternative to concurrent monitoring. In addition to its methodological and analytical contributions, the baseline of listening performance in continuous concurrent and serial multitalker contexts developed here provides a foundation for further investigations. The reduced level of reliable grounding that was observed to occur in condition 4SF in the present study instantly suggests that there are practical limits on the efficacy of temporal scaling. The authors plan to address this question next.

6. ACKNOWLEDGMENTS

The authors thank Astrid Schmidt-Neilsen and Christina Wasylyshyn for invaluable comments on drafts of this paper. This research was supported by the Office of Naval Research under work order number N0001408WX30007.

7. 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, Jan. 2002.

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

[3] 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.

[4] 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.

[5] 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, Nov. 1994.

[6] D.S. Brungart, "Informational and energetic masking effects in the perception of multiple simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 110, no. 5, pp. 2527-2538, Nov. 2001.

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

[8] D.S. Brungart and B.D. Simpson, "Optimizing the spatial configuration of a seven-talker speech display," *Proceedings of the 2003 International Conference on Auditory Display*, Boston, MA, July 2003.

[9] D.M. Green and J.A. Swets, *Signal Detection Theory and Psychophysics*. Wiley, New York, 1966.

[10] N.A. Macmillan and C.D. Creelman, *Detection Theory: A User's Guide*. Cambridge Univ. Press, Cambridge, UK, 1991.

[11] 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.

[12] D.W. Broadbent, *Perception and Communication*. Pergamon Press, NY, 1958.

[13] A.R. Conway, N. Cowan, and M.F. Bunting, "The cocktail party phenomenon revisited: The importance of working memory capacity," *Psychon. Bull. Rev.*, vol. 8, pp. 331–335, 2001.