# EVALUATING THE UTILITY OF AUDITORY PERSPECTIVE-TAKING IN ROBOT SPEECH PRESENTATIONS

*Derek Brock, Brian McClimens, Christina Wasylyshyn, J. Gregory Trafton, and Malcolm McCurry*

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

## ABSTRACT

In speech interactions, people routinely reason about each other's auditory perspective and adjust their manner of speaking accordingly by raising their voice to overcome noise or distance, and sometimes by pausing and resuming when conditions are more favorable for their listener. In this paper we report the findings of a listening study motivated by both this observation and a prototype auditory interface for a mobile robot that monitors the aural parameters of its environment to infer its user's listening requirements. The results provide significant empirical evidence of the utility of simulated auditory perspective taking and the inferred use of loudness and/or pauses to overcome the potential of ambient noise to mask synthetic speech.

## 1. INTRODUCTION

The identification and application of human factors that promote utility and usability is an overarching concern in the design of auditory displays [1]. The importance of this tenet is especially relevant for robotic platforms that are intended to be actors in social settings. People naturally want to be able to do things with robots in ways that are readily familiar, and aural communication is arguably the medium that many would expect to be the most intuitive and efficient for this purpose.

Implementing an auditory interface for a robot requires the integration of complementary machine audition and auditory display systems. These are ideally multifaceted functions and consequently pose a variety of interdisciplinary challenges for roboticists and researchers with related concerns. Audition, for instance, requires not only an effective scheme for raw listening, but also signal processing and analysis stages that can organize and extract various kinds of information from the auditory input. Important tasks for a robot's listening system include speech recognition and understanding, source location, and ultimately, a range of auditory scene analysis skills. The auditory display system, in contrast, should be capable of presenting speech and any other sounds that are called for by the robot's specific application. To support aurally based interactions with users and the environment—and thus be useful for more than just the output of information in auditory form—these systems must be informed by each other (as well as by other systems) and coordinated by an agent function designed to implement the robot's auditory interaction goals.

In practice, the current ability of robots to flexibly exercise interactive behaviors informed by the interpretation and production of sound-based information remains far behind the broad and mostly transparent skills of human beings. The computational challenges of auditory scene analysis and certain aspects of natural language dialogue are two of the primary reasons for this, but it is surprising that little attention has been given to some of the practical kinds of situational reasoning robots will need for successful auditory interactions in everyday, sound-rich environments.

For example, in speech and auditory interactions with each other, people typically account for circumstances that affect how well they can be heard from their listener's point of view and modify their presentations accordingly. In effect, they reason about their addressee's auditory perspective, and in most situations, their exercise of this skill markedly improves communication and reduces shared interactional effort. Talkers learn from experience that an addressee's ability to successfully hear speech and other sorts of sound information depends on a range of factors—some personal and others contextual. They form an idea of what their listener can easily hear and usually try not to adjust their manner of speaking much beyond what is needed to be effective. One of the most common accommodations talkers make is to raise or lower their voice in response to ambient noise or to compensate for distance or changes in a listener's proximity. If an ambient source of noise becomes too loud, talkers will often enunciate their words or move closer to their listener or pause until the noise abates, and then will sometimes repeat or rephrase what they were saying just before they stopped.

Taken together, these observations show that effectiveness in sound based interactions often involves more than just presenting and listening, so it is not hard to imagine that people are likely to find speech and other forms of auditory information a poor medium for human-robot interaction if a robot is unable to sense and compensate for routine difficulties in aural communication. Listeners count on talkers to appreciate their needs when circumstances undermine their ability to hear what is being said. And when this expectation is not met, they must redouble their listening effort, or ask talkers to speak louder, and so on. The ability to implement a comparable set of adaptive functions, then, is arguably a practical imperative for any auditory interface that is targeted for social interactions with users in everyday environments and noisy operational settings.

Motivated by this insight, the first author and a colleague recently demonstrated a prototype computational auditory perspective-taking scheme for a mobile robot that monitors both its user's proximity and the status of the auditory scene, and inferentially alters the level and/or progress of its speech to accommodate its user's listening needs [2]. The hardware and software framework for this system is primarily a proof of concept rather than a full solution. In particular, its parameters must be tuned for specific environments and there is limited integration with other, non-auditory sensors and functions that can play important roles in sound-related behaviors. The prototype's conduct involving auditory perspective taking is demonstrated in the context of an interactive auditory display that might be used as a mobile information kiosk in a lobby or a museum or exhibit hall where groups of people and other sources of noise are expected to be present on an intermittent but frequent basis (cf. [3]). Speech-based user interactions are limited to a few fixed phrases, and the auditory display is essentially a text-to-speech system that reads selected paragraphs of information with a synthetic voice. The system develops a map of auditory sources in its immediate surroundings, detects and localizes its user's voice, faces and follows the user visually, and monitors the user's proximity and the varying levels of ambient noise at its location. It then judges how loudly it needs to speak to be easily heard, and pauses if necessary, and can even propose to move to a quieter location. Further details about the system and its implementation are described in [4], and a more thorough development of the idea of auditory perspective taking is given in [5].

Although it is apparent that auditory perspective taking can be instrumental in the success of speech and aural interactions between people, it is unclear whether the use of adaptive auditory display techniques in collaborative paradigms such as human-robot interaction can, in fact, meet users' expectations in adverse auditory circumstances and improve the overall effectiveness of their listening experience. In this paper, the empirical results of an initial formal evaluation of this question are presented and implications for the design of auditory interfaces for robotic platforms and future adaptive auditory display research are discussed.

An empirical study of the utility of an automated auditory perspective-taking scheme can be approached in a number of ways, the most obvious being an *in situ* evaluation. However, the range of parameters used by the system outlined in [4] and its corresponding set of actions, together with the substantial number of manipulations this implies for a formal study, argued for the design of a smaller, more constrained initial experiment. Moreover, it was recognized that the system's key aural behaviors—its ability to make changes in the level and progress of presented speech—were also the most important actions to evaluate in terms of usability and their impact on users' listening performance. Consequently, several of the interactions the prototype addresses were not incorporated in the present study, particularly, changes in listener proximity (cf. [6]) and the role and utility of speech-based user controls.

Deciding to focus solely on changes in speaking level and the use of pauses made it unnecessary to employ the robotic implementation in the experiment. All of the auditory materials and adaptive actions could be simulated in a studio setting where participants could comfortably perform the response tasks used to measure their listening performance while seated. Similarly, to avoid the artificial manipulation and seemingly arbitrary selection of one set of noisy real-world environments over another (e.g., urban traffic, factory floor, busy theatre lobby, stadium crowd, etc.), a small number of broadband noise types was used for maskers.

Finally, the expository materials and techniques used here to measure participants' listening performance are largely the same as those developed by the authors for a previous but unrelated study involving a somewhat similar set of issues [7]. Here, though, the spoken information used in the earlier study—short segments of public radio commentaries—has been converted to "robot" speech with a commercial speech-to-text engine. Synthetic voices of both genders are now in relatively wide use, but they are known to be more difficult for listeners to process than natural speech (see e.g., [8][9]). Hence, to remove voice as a factor, a single, "standard" synthetic male voice was used for all of the speech materials presented to listeners.

## 2. METHOD AND APPARATUS

Fourteen participants, five female and nine male, all personnel at the authors' institution, and all claiming to have normal hearing, took part in the experiment, which employed a within-subjects design. The timing and display of all sounds and response materials were coordinated by software, coded in Java by one of the authors, running on a laboratory PC. The auditory component was rendered with three Yamaha MSP5 powered studio monitors placed directly left, right, and in front of the listener, all at a distance of approximately 1.32 m. Sound was limited to a maximum of 85 dB SPL. The response tasks were presented visually on a 0.61m (diagonal) Samsung SyncMaster 243T flat-panel monitor.

### 2.1. Listening Materials and Experimental Manipulations

The spoken information developed for the study was derived from an archive of short commentaries on topics of general interest that were originally broadcast on public radio. Ten commentaries were transcribed, and in some cases edited for length, and then re-recorded as synthesized "robot" speech using the Cepstral text-to-speech engine [10] and a standard male synthetic voice named "Dave." The resulting speech materials were randomly assigned to three training sessions, which allowed participants to become familiar with the listening and response tasks, and to seven formal listening exercises that made up the body of the experiment. The assignments were the same for all listeners. Additionally, a test for uniformity among the commentaries assigned to the listening exercises was made and showed no significant differences between a number of lexical parameters (number of sentences, words, and syllables, etc.). The training sessions were each about a minute in length and the listening exercises lasted between 2.5 and 3.5 minutes, depending on the particular manipulation (see below).

Most real-world noise environments have notably different and variable time-frequency characteristics, which in turn make their effectiveness as maskers difficult to systematize in a controlled experiment. To avoid this potential confound, four types of broadband noise were selected to simulate the occurrence of ambient, potentially speech-masking, noise events in the study: brown noise (used only in the training sessions), pink noise, white noise, and "Fastl" noise [11]. The last is white noise filtered and modulated to simulate the average spectral distribution and fluctuating temporal envelope of an individual's speech. A digital audio editing tool was used to normalize and create a matrix of four masking events for each kind of noise. For white, pink, and Fastl noise, two short events (5 sec.)—one "quiet" (-26 dB) and the other loud (-19 dB)—and two long events (30 sec.) differing in loudness in the same manner were created. Onset and offset ramps were linear fades lasting 0.51 sec. for short events and 7.56 sec. for long events. A slightly different matrix of brown noise events was used in two of the training sessions, and the matrices corresponding to the other three noise types, as just described, were used in six of the experimental manipulations. Listeners heard two instances of each of the four masking event types in random order in these exercises.

#### 2.1.1. *Design*

The scheme of the study combined a Baseline listening exercise and a two factor, 2x3 design with repeated measures. Participants heard each of the seven manipulations in counterbalanced order. In the Baseline condition, participants simply listened to one of the commentaries and carried out the associated response tasks. In the other six conditions, they performed functionally equivalent listening and response tasks with the addition of eight intermittent noise events. Commentaries were always rendered by the audio monitor in front of the listener, and instances of broadband noise were rendered by the monitors on the listener's left and right.

The chief goal of the experiment was to evaluate whether automatic pausing and resumption and correlated changes in the level of sound presented by an auditory display (in this case speech) can benefit users' listening performance when ambient noise arises. Accordingly, the first factor in the non-baseline manipulations entailed the combined non-use or use of these presentation strategies and the second factor involved the use of pink, white, or Fastl noise events. The three training sessions emphasized the first factor by introducing the "baseline" manipulation and then the contrast between "non-adaptive" and "adaptive" presentations of synthetic speech during episodes of brown noise. A summary of the seven listening exercises participants carried out in the body of the experiment is given in Table 1.

| Condition | Description |
|---|---|
| **Baseline** | **Baseline** synthetic speech, no noise events |
| **NA-white** | **Non-adaptive** synthetic speech and **white** noise events |
| **NA-pink** | **Non-adaptive** synthetic speech and **pink** noise events |
| **NA-Fastl** | **Non-adaptive** synthetic speech and **Fastl** noise events |
| **A-white** | **Adaptive** synthetic speech and **white** noise events |
| **A-pink** | **Adaptive** synthetic speech and **pink** noise events |
| **A-Fastl** | **Adaptive** synthetic speech and **Fastl** noise events |

Table 1: A summary of the seven experimental conditions and their coded designations. Participants heard all seven conditions in counter-balanced order.

### *2.1.2. Predictions and planned comparisons*

The seven conditions chosen for the study were motivated by a specific set of anticipated outcomes. First, it was expected that measures of listening performance (see Section 2.2) in the Baseline condition would be the best in the study, but would fail to approach perfect performance due to the use of a synthetic voice. In contrast, listening performance in the three Non-Adaptive conditions (those in which broadband noise events were allowed to mask portions of the spoken commentary: NA-white, NA-pink, and NA-Fastl) was expected to be lowest in the study, both collectively and individually. More importantly, and the focus of the experiment, listening performance in the three Adaptive auditory display conditions (A-white, A-pink, and A-Fastl) was expected to be nearly as good as the Baseline and substantially better than in the non-adaptive conditions.

Since the prototype auditory perspective-taking system makes no distinction between one type of noise and another, and only tries to infer listening needs on the basis of amplitude, it was unclear how each of the broadband noise manipulations would affect participants' comparative performance, particularly in the three Adaptive conditions. White noise and pink noise are both continuous at a given volume and are both effective auditory maskers. But white noise, with equal energy in all frequencies, is the more comprehensive masker of the two and, for many individuals, it may also be the more attentionally and cognitively disruptive under any circumstances, but especially when it is loud. Fastl noise, on the other hand, because of the shape of its underlying spectral power density and fluctuating amplitude envelope, provides the least comprehensive coverage as a masker. However, if cognition is perceptually tuned to attend to voices, it may be more distracting than either white or pink noise due to its speech-like properties. Nevertheless, all three types of noise should be good maskers of speech. Because of these qualified differences and the difficulty of predicting how broadband noise events may interact with auditory concentration in various circumstances, planned comparisons (contrasts) are used below to explore how performance in the two presentation strategy manipulations differ from performance in the Baseline condition across the three manipulations of noise-type.

### *2.1.3. Adaptive auditory display behaviors*

To approximate the prototype auditory perspective-taking system's response to different levels of ambient noise in the Adaptive auditory display manipulations (i.e., A-white, A-pink, and A-Fastl), the three commentaries respectively assigned to these conditions were modified in the following ways. First, as in the Non-Adaptive conditions, they were appropriately aligned with noise events on separate tracks in a sound editor. Next, using linear onset and offset ramps, the amplitude envelope of each commentary was modulated to compete in parallel with the eight randomly ordered noise events in its particular manipulation. The resulting modulations were then shifted forward (i.e., later in time) to simulate the time it takes for the onset of a noise event to cross the system's response threshold. Thus for quiet noise events, the synthetic speech starts to become louder 1.0 sec. after the short event begins and 3.0 sec. after the long one starts, the difference being due to the more gradual onset ramp of long events (see above). The short and long episodes of loud noise have correspondingly steeper onset ramps, so the response for these events begins at 0.8 and 2.0 sec., respectively. Loud events, though, are intended to trigger the prototype's pause response. To mimic this effect, corresponding periods of silence were inserted in the commentaries with the sound editor (thus increasing their length). During short episodes of loud noise, pauses begin at the first word boundary following 1.2 sec. of the loudness response; during long episodes they begin similarly at or beyond the 5.0 sec. mark. The commentaries were then edited to resume at the point where the noise event drops below the pause threshold by re-uttering the interrupted sentence or phrase. Long pauses, however, first resume with the words, "As I was saying…" The idea of resuming interrupted synthetic speech in this manner arose during the development of the prototype and was found to be consistent with listeners' intuitions about verbal pauses in piloting for the study.

To summarize, eight noise events (two of each of the four types outlined in Section 2.1) occurred in each of the three Adaptive conditions, and the auditory display took the following actions to overcome the potential for its presentation of synthetic speech to be masked from its listener's perspective. When the respective short- and long-quiet events occurred, the level of the speech rose by 6 dB to be easy to hear over the level of the noise and then fell to its previous level as the noise abated. When the respective short- and long-loud events occurred, the speech became louder to a point and then paused. After the noise abated, the auditory display resumed from the beginning of the phrase or sentence it interrupted, but in the case of the long-loud event prefaced its resumption with the words, "As I was saying." A schematic of the auditory display's four adaptive behaviors showing level changes and pauses is given in Figure 1.

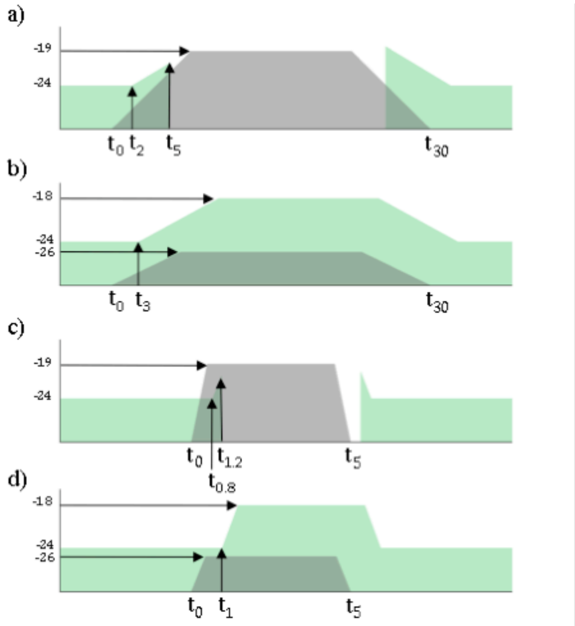

Figure 1. *Schematic diagrams showing actions taken by the auditory display in the experiment's Adaptive conditions to counter noise events with the potential to mask speech from the listener's perspective:* a) *long-loud,* b) *long-quiet,* c) *short-loud, and* d) *short-quiet. Time in seconds is shown on the horizontal axis (note differences in scale for long and short events), and level in dB is shown on the vertical axis. Noise event envelopes are shown as gray trapezoids. Envelopes of continuous speech are shown in green. See the text for additional details.*

### 2.1.4. Auditory examples

Edited examples of the sound materials used in the study are given in the binaural recordings listed below, which are available by email from the first author as .wav or .mp3 files. NADAPT presents an instance of each of the four noise event types in the Non-Adaptive manipulations: long-loud/NA-white, long-quiet/NA-pink, short-loud/NA-Fastl, and short-quiet/NA-Fastl. ADAPT presents an instance of each of the four noise event types in the Adaptive manipulations: long-loud/A-white, long-quiet/A-pink, short-loud/A-Fastl, and short-quiet/A-Fastl.

NADAPT: example noise events in Non-Adaptive conditions
ADAPT: example noise events in Adaptive conditions

## 2.2. Response tasks and dependent measures

In both the training sessions and the listening exercises, participants carried out two response tasks, one while listening and the other immediately after. After each training session and listening exercise, participants were also asked to rate their preference for the way the synthetic speech was presented.

The first response task involved listening for noun phrases in the spoken material and marking them off in an onscreen list that corresponded to the current commentary in the study. Each list contained both the targeted noun phrases and foils in equal numbers (eight targets per story in the training sessions and twenty targets per story in the listening 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 commentaries on similar but not identical topics.

Participants proved to be quite good at discriminating between target phrases and foils on the basis of the speech materials, and only rarely mistook foils for utterances in any of the commentaries, regardless of their ability to verify targets. Thus, because of an extremely low incidence of false alarms, (a total of 4 out of 1960 possible correct rejections), performance in the phrase identification task was measured only as the percentage correctly identified target noun phrases. In the results and discussion sections below, this measure is referred to as *p*(*targets*).

In the second response task, participants were given a series of sentences to read and were asked to indicate whether each contained "old" or "new" information based on what they had just heard [12]. "Old" sentences were either *original*, word-for-word transcriptions or semantically equivalent *paraphrases* of commentary sentences. "New" sentences were either "*distractors*"—topic-related sentences asserting novel or bogus information—or commentary sentences *changed to make their meaning* inconsistent with the content of the spoken material. An example of each sentence type developed from a piece on the ubiquitous popularity of baseball caps is provided in Table 2. 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 commentary in the training sessions. In the formal exercises, eight sentences per commentary (two of each of the old and new sentence types) were presented.

| Sentence type | Example sentence | Designation |
|---|---|---|
| Original | Baseball caps are now bigger than baseball. | Old |
| Paraphrase | Baseball caps have become more popular than the game of baseball. | Old |
| Meaning change | Baseball caps are now bigger than football. | New |
| Distractor | Most baseball caps are now made in China. | New |

Table 2: 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 demur by selecting "I don't know" as a response.

Two measures were calculated from the participants' sentence judgments in each condition. The primary measure, denoted *p*(*sentences*), is the proportion of sentences correctly judged as old or new. The second measure, denoted *p*(*demurs*), is the proportion of "I don't know" responses. Both measures are calculated as a percentage of the eight sentences presented for verification in each condition.

Last, to gage participants' subjective impressions, after completing the sentence judgment task in the training sessions and in each of the experimental conditions, they were asked to rate their preference for the auditory display. They did this by indicating their agreement with the statement, "I prefer the way the synthetic speech was presented in this listening exercise," on a seven point Likert scale, with 1 = "strongly disagree" and 7 = "strongly agree."

## 3. RESULTS

The performance measures for both response tasks were mostly consistent with the pattern of listening performance that was expected to arise between the noise-free Baseline condition and the use/non-use of the adaptive auditory display when noise events capable of masking speech were present. In particular, participants' abilities to correctly recognize targeted noun phrases, *p*(*targets*), and judge sentences as old or new, *p*(*sentences*), were both highest in the Baseline condition and lowest in the three Non-Adaptive conditions (NA-white, NA-pink, and NA-Fastl). Mean scores for the target phrase recognition response task were only slightly lower than Baseline in the three Adaptive conditions (A-white, A-pink, and A-Fastl), as predicted. Scores for the sentence judgment task in the Adaptive conditions, however, were not as high as expected (see Section 2.1.2), and fell in a more intermediate position between the scores for the respective Baseline and Non-Adaptive conditions. Even so, the correlation between *p*(*targets*) and *p*(*sentences*) is significant (Pearson's $r = 0.573$, $p = 0.05$ (2-tailed)). Plots of the mean proportions of correctly identified target noun phrases, *p*(*targets*), and sentences correctly judged as "old" or "new," *p*(*sentences*), in all seven conditions are respectively shown in Figure 2a and b.

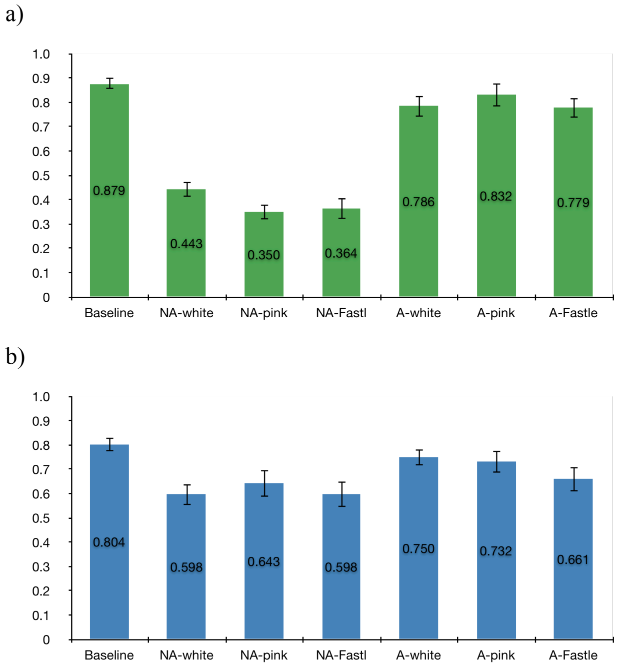

Figure 2. a) *Plot of the mean proportion of correctly identified target noun phrases, p(targets), in each condition*. b) *Plot of the mean proportion of sentences correctly judged as "old" or "new," p(sentences), in each condition. The y-axis in both plots shows proportion. Error bars show the standard error of the mean.*

To evaluate effects of presentation strategy—non-adaptive vs. adaptive—and noise type on listening performance, the six conditions involving noise events were construed as a factorial design, and a 2 level by 3 level, repeated measures analysis of variance was performed for each of the dependent measures. In these analyses, there was a main effect for presentation strategy but not for noise type. Specifically, the 2x3 ANOVA for *p*(*targets*) showed that participants were significantly better at the target phrase task when Adaptive presentations were used to counter all three types of noise events ($F(1, 13) = 190.7$, $p < 0.001$). The corresponding ANOVA for *p*(*sentences*) showed similarly that performance of the sentence judgment task was significantly better in the conditions involving Adaptive presentations ($F(1, 13) = 5.077$, $p = 0.042$). Additionally, there was a significant interaction between presentation strategy and noise type in the analysis for *p*(*targets*) ($F(2, 26) = 4.518$, $p = 0.021$), but not in the analysis for *p*(*sentences*).

Because it was unclear how each type of noise might impact listening performance when the respective Non-Adaptive and Adaptive speech strategies were used, planned contrasts were used to evaluate how the dependent measures in these conditions differed with performance in the Baseline condition. All of these contrasts were significant when the speech was Non-Adaptive, meaning that both performance measures, *p*(*targets*) and *p*(*sentences*), were meaningfully hurt regardless of the type of masker. In other words, as expected, each type of noise proved to be a good masker of speech.

A more interesting set of results emerged from the contrasts involving Adaptive speech. Here, as expected, some of the contrasts are not significant, meaning that the corresponding measures of performance were not substantially worse than the Baseline. However, this was only the case for *p*(*targets*) and *p*(*sentences*) with pink noise events and for *p*(*sentences*) with white noise events. The other three contrasts were all significant: in spite of the Adaptive presentation strategies, both white noise and Fastl noise had a meaningful impact on listeners' ability to perform the target phrase recognition task, and Fastl noise significantly hurt their corresponding ability to perform the sentence judgment task. The *F* statistics for the contrasts involving Adaptive speech are summarized in Table 3.

| Measure | Contrast | *F* |
|---|---|---|
| *p*(*targets*) | **A-white** vs. **Baseline** | $F(1, 13) = 10.876$,<br>$p = 0.006$* |
| | **A-pink** vs. **Baseline** | $F(1, 13) = 1.441$,<br>$p = 0.251$ |
| | **A- Fastl** vs. **Baseline** | $F(1, 13) = 7.280$,<br>$p = 0.018$* |
| *p*(*sentences*) | **A-white** vs. **Baseline** | $F(1, 13) = 1.918$,<br>$p = 0.189$ |
| | **A-pink** vs. **Baseline** | $F(1, 13) = 2.537$,<br>$p = 0.135$ |
| | **A-Fastl** vs. **Baseline** | $F(1, 13) = 5.438$,<br>$p = 0.036$* |

Table 3: *F* statistics for the planned contrasts between the Baseline and Adaptive conditions for the *p*(*targets*) and *p*(*sentences*) performance measures. Statistics showing that a lower performance measure in a particular condition is significantly different from the corresponding measure in the Baseline condition are indicated with an asterisk.

The other measure associated with the sentence judgment response task was the proportion of "I don't know" responses participants made in each condition, denoted *p*(*demurs*). Giving participants the option to make this response allowed them to indicate they felt they had no basis to judge a particular sentence as old or new information. Intuitively, a greater percentage of demurs should be expected in the Non-Adaptive manipulations because of the masking effects of noise. This proved to be the case, and a plot of the mean proportion of demurs in all seven experimental conditions, shown in Figure 3a, exhibits, inversely, the same broad pattern as that seen for both *p*(*targets*) and *p*(*sentences*) in Figure 2. A 2x3 ANOVA of the six non-Baseline conditions for *p*(*demurs*), however, showed no main effect for either factor and no interaction. Out of six planned contrasts, only NA-Fastl vs. Baseline was significant (F(1, 13) = 7.495, p = 0.017), meaning that the number of demurs in each of the other five conditions was not meaningfully greater than in the Baseline condition. Also shown in Figure 3a are the corresponding numbers of participants in each condition that chose to demur one or more times. These counts are significantly correlated with the mean *p*(*demurs*) values (Pearson's r = 0.864, p = 0.02 (2-tailed)), but the rate of demurs per demurring respondent is comparatively higher in the Non-Adaptive conditions.

Last, a plot of the participants' mean level of subjective agreement with the statement, "I prefer the way the synthetic speech was presented in this listening exercise," in each condition is shown in Figure 3b. As mentioned above, the range of this data corresponds to a seven point Likert scale. The emergent pattern of ratings across manipulations is somewhat similar to the correlated pattern seen in the plots in Figure 2. However, there is an interesting difference here in that while participants' mean preference for the Baseline presentation is greater than their preference for any of the Non-Adaptive presentations, it is not greater than their preference for any of the Adaptive presentations. Planned contrasts with the Baseline condition were not significant, but a two factor ANOVA of this data in the non-Baseline conditions showed a main effect for presentation strategy (*F*(1, 13) = 10.538, *p* = 0.006).

## 4. DISCUSSION

The chief motivation for this experiment was to evaluate the combined utility of two adaptive auditory display techniques for individual listeners in noisy settings, namely automated changes in loudness and the use of pauses. In the application context of the study—human-robot interaction involving synthetic speech—both of these flexible presentation strategies are intended to anticipate listening requirements from the user's auditory perspective and improve the overall effectiveness of his or her listening experience. To test these ideas, participants were asked to listen to seven short commentaries spoken by a synthetic voice and for each commentary carry out two response tasks designed to measure a) their ability to attend to the content while listening and b) the consistency of their understanding of the content afterwards. The commentaries were randomly assigned to a set of experimental conditions that provided a noise-free, listening performance baseline and, in six additional manipulations, tested how the non-use and combined use of the two adaptive aural presentation techniques affected listening performance in the presence of eight coordinated episodes of three types of broadband noise.

Collectively, the results of the study provide significant empirical evidence of the utility of simulated auditory perspective taking and the inferred use of loudness and/or pauses to overcome the potential of noise to mask synthetic speech. In particular, while measures of listening performance aided by the adaptive techniques in the presence of noise were not as robust as listening in the absence of noise, they were demonstrably better than unaided listening in the presence of noise. Additionally, when asked, listeners indicated a significant subjective preference for the adaptive style of synthetic speech over the non-adaptive style.

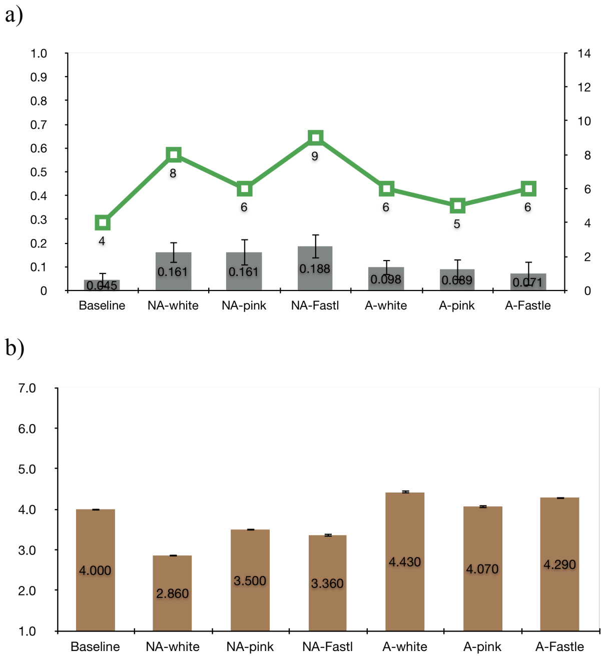

Figure 3. a) *Plot of the mean proportion of "I don't know" responses, p*(*demurs*) *(gray columns corresponding to y-axis on left) in the sentence judgment task and the number of participants in each condition choosing to demur one or more times (green squares corresponding to y-axis on right)*. b) *Plot showing the mean level of participants' agreement with the statement, "I prefer the way the synthetic speech was presented in this listening exercise," in each condition. The y-axis in this plot reflects a seven-point Likert scale ranging from 1 = "strongly disagree" to 7 = "strongly agree." Error bars in both plots show the standard error of the mean.*

Overall, this finding has implications for the design of auditory interfaces for robots and, more generally, for adaptive auditory display research, some of which will be covered below. Certain aspects of the study, however, warrant further consideration and/or critique. Among these are how Baseline performance in the study compares to listening performance involving human speech, the impact of noise type on listening performance in the Adaptive conditions, and listeners' subjective preferences

### 4.1. Listening to synthetic and human speech

Although listening performance in the Baseline condition, as measured by *p*(*targets*) and *p*(*sentences*), was expected to be the best in the study, it was also expected to fail to approach perfect performance due to the use of a synthetic voice. No test of this conjecture was made here, but a specific manipulation in the concurrent vs. serial talker experiment by Brock et al. in [7] offers a useful, if imperfect means for comparison.

In the cited experimental condition, a different group of participants from those in the present study listened to a serial presentation of four commentaries that were drawn from the same source as those used here. The commentaries were spoken by human talkers and were rendered with headphones at separate locations in a virtual listening space using a nonindividualized head-related transfer function. During the listening exercise, the same target phrase and sentence judgment methods used in the present study were employed to measure listening performance, but all four commentaries were presented before the corresponding sentence judgment tasks were given to listeners.

The resulting mean proportion of correctly identified target phrases was 0.91, and the corresponding mean proportion of correctly judged sentences was 0.87. When these numbers are compared with their counterparts in the present Baseline condition (respectively, 0.88 and 0.80), it can be seen that listening performance, in spite of a number of experimental differences, was somewhat poorer when the information medium involved synthetic speech.

The purpose in making this rough comparison is not to claim significance, which has been shown elsewhere (see [8][9]), but rather to stress the aurally anomalous properties of current synthetic voice technology and thus point to a further motivation for accommodating users' listening requirements when this technology is used in noisy settings. Canned human speech can be used for limited purposes, but there are no alternatives to synthetic speech for broader conversational applications, which is a key technical objective for robots targeted for roles in social settings.

### 4.2. The impact of noise type on listening performance

While the use of three different types of broadband noise as surrogates for real-world noise capable of masking speech was a secondary consideration in the design of this study, several of the specific results suggest that some types of noise are more difficult to effectively adapt for than others.

Taken together, the significant interaction between the presentation and noise factors in the *p*(*targets*) data and the pattern of significant performance differences among the planned contrasts involving the Adaptive conditions shown in Table 3 are good evidence that some forms of noise, as discussed in Section 2.1.2, can, in fact, undermine a listener's auditory concentration because of their inherently annoying and/or distracting properties. In particular, the *p*(*targets*) interaction arises primarily from the fact that listening performance in the NA-pink and A-pink manipulations are respectively lower and higher than listening performance in the other Non-Adaptive and Adaptive conditions. In other words, pink noise was a very effective masker of synthetic speech, but it was also the best type of noise to successfully adapt for. When the pattern of significant *p*(*targets*) contrasts in Table 3 is added to the picture, it is apparent that in spite of the adaptive auditory display, both white and Fastl noise disconcerted listeners in a way that pink noise did not.

Did these effects happen for similar reasons? Probably not, because, Fastl noise also engendered a significant contrast in the *p*(*sentences*) data shown in Table 3, which implies that it, unlike white noise, also meaningfully undermined the ability of listeners' to form a good understanding of the commentary in the A-Fastl condition—relative to the understanding listeners achieved in the Baseline condition. Given the differences between white and Fastl noise, the one being continuous and spectrally uniform and the other having fluctuating, speech-like properties, it would appear that competing ambient noise with speech-like qualities may be a particularly challenging type masker to consistently overcome.

In apparent contrast with this interpretation is the pattern of mean *p*(*demurs*) data shown in Figure 3a and the corresponding counts of participants electing to respond in this way in the sentence judgment task. If Fastl noise impairs auditory concentration, observing a correspondingly substantial proportion of demurs would seem to be good supporting evidence. However, the number of demurring participants and the mean value of *p*(*demurs*) in the A-Fastl condition is essentially no different than the corresponding values in the other two Adaptive conditions.

Oddly, though, the conspicuously large values in this plot appear in the NA-Fastl condition, and furthermore, only this planned contrast with the mean Baseline value of *p*(*demurs*) was significant. What this implies is that Fastl noise was a particularly effective masker of unmodified synthetic speech. But note that this is not inconsistent with the premise that some forms of noise can substantially undermine a listener's auditory concentration. If an aural masker has this additional cognitive effect, then it should be an even better masker than, say, unvarying continuous noise. Certainly, more participants in this condition than in any other appear to have recognized the poverty of their understanding of the commentary they had just heard, and thus responded appropriately. So in the A-Fastl condition, it may only be the case that listeners were unaware of the extent of their impaired understanding because the adaptive auditory display ensured that none of the commentary was aurally masked. If this is so, then there should be a large mean proportion of sentence judgment errors relative to the other Adaptive conditions, and this turns out to be the case. In fact, the mean proportion of sentence judgment errors in the A-Fastl condition is greater, at 0.268, than the corresponding proportion of errors in any of the other conditions in the study. Fastl noise thus turns out to be an exceptionally effective masker of synthetic speech even when adaptive changes in loudness and the use of pauses are employed.

### 4.3. Listeners' subjective preferences

The purpose of asking participants after each listening exercise to rate their agreement with the sentence, "I prefer the way the synthetic speech was presented in this listening exercise," was to determine, in a relatively unbiased way, how much they liked or disliked the particular auditory display they had just worked with. Ratings of this sort are inherently subjective, but can nevertheless provide useful insights and/or reveal unanticipated issues.

The preference data shown in Figure 3b shows a significant main effect in favor of the Adaptive auditory display, and it also reveals a consistently greater preference for the adaptive manipulations over the Baseline condition. The contrasts are not significant, but the trend is conspicuous and unexpected: it seems counter-intuitive that an uninterrupted presentation in the quiet would be less preferable than adaptively modified presentations accompanied by multiple noise events.

The mean rating for the Baseline condition, however, is exactly midway between the two ends of the Likert scale used for this measure, so listeners appear to have been basically indifferent to the use of synthetic speech by itself. This could easily be due to the fact that nothing of consequence occurs in the manipulation, which, in turn, makes the response tasks seem relatively straightforward. In the Adaptive manipulations, though, substantial impediments to listening arise and the auditory display responds to the intruding noise effectively and with dispatch. More importantly, it does this in ways that are modeled on human solutions. Without corroborating data to specifically indicate why participants rated each manipulation as they did, it can only be surmised that their agreement with the preference statement was somewhat higher in the Adaptive conditions because the synthetic voice acted on their listening needs transparently and in ways that met their expectations or, at the very least, facilitated their performance of the response tasks. If this interpretation is correct, it shows that simulated perspective taking in this type of auditory interaction design has important collaborative utility and merits further development.

### 4.4. Implications for design and research

The outcome of the study supports the idea that auditory interaction designs for robotic platforms can and should account for their users' listening requirements, especially in operational settings where ambient noise is likely to be an issue. This idea also extends to situations in which the proximity between the robot and its user is likely to vary with any frequency. The small but measurably different impact that Fastl noise had on listening performance in the study suggests that additional adaptive strategies such as enunciation and repair may be needed in some circumstances to cope with the distracting and informational masking effects of extraneous speech. This mode of operation could perhaps be informed by machine classification of the ambient noise environment. Another aspect of auditory perspective-taking that will need to be addressed in future research involves inferences made on the basis of users' privacy concerns and other socially motivated considerations.

It is also possible to imagine a range of non-speech applications for robot auditory interfaces such as aural monitoring and playback and sonification of process or sensor data. Auditory displays of this sort on robots or in other formats may be even harder to use in the presence of ambient noise than speech displays precisely because of the way they represent information. Real-world noise is likely to be a good informational masker of non-speech sounds in much the same way that speech and speech-like noise can be an informational masker of speech. Ambient speech may also have masking effects on non-speech auditory displays, especially if sonifications are involved, because of the nature of their information content and the sustained auditory attention they require. Effective adaptive presentation strategies in these circumstances will require additional research and may prove to be different from the techniques evaluated here.

## 5. CONCLUSIONS

The notion that robots will eventually assume collaborative roles involving aural interactions in social settings has already materialized in the form of self-serve check out registers at stores, automated telephone support, and toys that talk and respond to voice commands. In the relatively near future, it is widely expected that mobile robotic platforms capable of far greater autonomy than is technically feasible today will be deployed for a wealth of interactive societal purposes ranging from service and caretaking to military and logistical applications. Soon, people will not only expect to be able to interact with robots in much the same way they interact with each other in face-to-face activities, but they will also expect these advanced systems to understand their communicative needs. The idea of auditory perspective taking—inferring what an addressee's listening requirements are on the basis of ambient sound, proximity, and, ultimately, social constraints—is just one element of this understanding, albeit an important one, that will eventually be joined with other communication skills users will expect robots and other systems to be capable of, such as gaze following, contextual awareness, and implied goal recognition. The success of the adaptive auditory display strategies evaluated in the present study confirms the importance of this emerging direction in user interface design.

## 6. ACKNOWLEDGMENTS

This research was supported by the Office of Naval Research under work order number N0001409WX30013. The authors would like to thank Hesham Fouad for technical help with the production of the accompanying audio files.

## 7. REFERENCES

[1] S.C. Peres, V. Best, D. Brock, C. Frauenberger, T. Hermann, J. Neuhoff, L.V. Nickerson, B. Shinn-Cunningham, and A. Stockman, “Auditory interfaces,” in P. Kortum (Ed.), *HCI Beyond the GUI*. Morgan Kaufman, San Francisco, CA, 2008.

[2] D. Brock and E. Martinson, “Exploring the utility of giving robots auditory perspective-taking abilities,” in *Proc. 12th International Conference on Auditory Display*, London, UK, 2006.

[3] S. Thrun, M. Beetz, M. Bennewitz, W. Burgard, A.B. Cremers, F. Dellaert, D. Fox, D. Hähnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz, “Probabilistic algorithms and the interactive museum tour-guide robot Minerva,” *Intl. J. Robotics Res.*, vol. 19, no. 11, pp. 972-999., Dec. 2000.

[4] E. Martinson and D. Brock, “Improving Human-Robot Interaction through Adaptation to the Auditory Scene,” in *HRI '07: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction,* Arlington, VA, Mar. 2007.

[5] D. Brock, and E. Martinson, E., “Using The Concept of Auditory Perspective Taking to Improve Robotic Speech Presentations for Individual Human Listeners,” in *AAAI 2006 Fall Symposium Technical Report: Aurally Informed Performance: Integrating Machine Listening and Auditory Presentation in Robotic Systems*, Washington, DC, 2006.

[6] S. Kagami, Y. Sasaki,S. Thompson, T. Fujihara, T. Enomoto, and H. Mizoguchi, "Loudness measurement of human utterance to a robot in noisy environment," in *HRI '08: Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction,* Amsterdam, The Netherlands, Mar. 2008.

[7] 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 24-27, 2008.

[8] J.B. Hardee and C.B. Mayhorn, “Reexamining synthetic speech: Intelligibility and the effect of age, task, and speech type on recall,” in *Proceedings of the Human Factors and Egonomics Society 51st Annual Meeting*, Baltimore, MD, October 1-5, 2007, pp. 1143-1147.

[9] C. Stevens, N. Lees, J. Vonwiller, and D. Burnham, “Online exerpimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness, and preference,” *Computer Speech and Language*, vol. 19, no. 2, pp.129-146, 2005.

[10] http://cepstral.com.

[11] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models, Third Ed*. Springer-Verlag, Berlin, Germany, 2007.

[12] 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.