Evaluating the Utility of Auditory Perspective-Taking in Robot Speech Presentations

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

Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, 4555 Overlook Ave., S.W.
Washington, DC, 20375 USA
{Derek.Brock,Brian.McClimens,Christina.Wasylyshyn, Gregory.Trafton,Malcolm.McCurry}@nrl.navy.mil

Abstract. In speech interactions, people routinely reason about each other's auditory perspective and change their manner of speaking accordingly, by adjusting their voice to overcome noise or distance, or by pausing for especially loud sounds and resuming when conditions are more favorable for the listener. In this paper we report the findings of a listening study motivated both by this observation and a prototype auditory interface for a mobile robot that monitors the aural parameters of its environment and infers 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.

Keywords: auditory perspective-taking, adaptive auditory display, synthetic speech, human-robot interaction, auditory interaction, listening performance.

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 interact with robots in ways that are already familiar to them, 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

S. Ystad et al. (Eds.): CMMR/ICAD 2009, LNCS 5954, pp. 266–286, 2010.

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 simplistic displays of information in auditory form—these complementary listening and display 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 factors 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 a listener can easily hear and usually try not to adjust their manner of speaking much beyond what is needed to be effective. Certainly, 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 more carefully 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.

All together, these observations show that achieving effectiveness in aural interactions often involves more than just presenting and listening. Given this perspective, it is not difficult to imagine that people are likely to find speech and other forms of auditory information an unreliable medium for human-robot interaction if the 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 if this expectation is not met, they must redouble their listening effort, or ask talkers to speak louder, and so on. Giving an auditory user interface the ability to diagnose and adapt to its listeners needs, then, is a practical imperative if the underlying platform is targeted for social roles in everyday environments or 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, system parameters must be tuned for specific environments and there is limited integration with 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 in 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, pauses if necessary, and can even propose moving 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 casual experience with this prototype interface in demonstration runs confirms that it functions as intended, it is nevertheless important to formally show whether adaptive auditory display techniques in human-robot interaction, or in other user interaction paradigms, can, in fact, meet listeners' expectations and improve their listening performance in difficult auditory situations. This paper describes the objectives and method of an initial empirical evaluation of this question and presents the findings of the resulting experiment. Additionally, implications for the design of auditory interfaces for robotic platforms and future adaptive auditory display research are discussed.

A study of the effectiveness of an automated auditory perspective-taking scheme could be approached in a number of ways, the most obvious being an *in situ* evaluation. Consideration of both the number of interrelated parameters used by the system outlined in [4] and the range of its adaptive actions, however, argued here for the design of a smaller, more constrained initial experiment. Moreover, it was recognized that the system's key auditory behaviors, namely, its ability to make changes in the level and progress of presented speech, were essentially the most important actions to evaluate in terms of usability and 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.

Focusing solely on the utility of changes in auditory level and the use of pauses made it unnecessary to employ the robotic implementation in the experiment. All of the sound 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.

Last, the expository materials and techniques employed here to measure participants' listening performance were adapted from, and 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 substantially varied and are known to be more difficult for listeners to process than natural speech (see e.g., [8][9]). Hence, to remove voice type as a factor, a single, “standard” synthetic male voice was used for all of the information 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. A within-subjects design was employed. 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 (this layout is shown below in Fig. 1). 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

A corpus of expository materials derived from an archive of editorials on topics of general interest originally broadcast on public radio was developed for the study. 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 audio 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 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), and white, pink, and “Fastl” noise [11]. The latter of these 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 “forte” (-26 dB) and the other “fortissimo” (-19 dB)—and two long events (30 sec.) differing in loudness in the same manner were created. (The terms “forte” (loud) and “fortissimo” (very loud) are used here for expository purposes and are borrowed from the lexicon of musical dynamics for indicating relative loudness in performance.) Onset and offset ramps were linear fades lasting 0.51 sec. for short events and 7.56 sec. for long events. Six of the experimental conditions featured noise, and listeners heard each of the four kinds of masking events in these manipulations twice in random order (for a total of eight events). A slightly different matrix of brown noise events was created for use in two of the training sessions.

Design. The experiment's scheme of manipulations involved a Baseline listening exercise and a two factor, two-level by three-level (2x3) design with repeated measures. Participants heard the Baseline manipulation first and then each of the remaining six manipulations in counterbalanced order. In the Baseline condition, participants simply listened to one of the commentaries and carried out the associated response tasks (see Section 2.2 below). In the other six conditions, they performed equivalent listening and response tasks with the addition of eight intermittent noise events. As is shown in Fig. 1, the commentaries were 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.

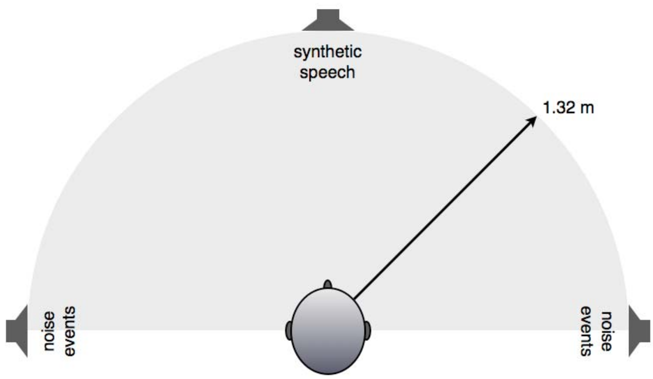

Fig. 1. Layout of the experimental listening environment, showing the location and purpose of each audio monitor and its distance from the listener

The main intent of the experiment was to evaluate the combined utility of automated pauses and level changes when ambient noise with the potential to mask the auditory display arises; interest in the comparative effects of one type of noise versus another was secondary. Accordingly, the two levels of the first factor in the design of the non-baseline manipulations entailed the respective non-use and use of the automated adaptive presentation strategies, and the second factor (three levels) involved the respective use of white, pink, and Fastl noise events. Consistent with this focus on the utility of adaptivity, the three training sessions highlighted only the first factor by introducing abbreviated versions of the "baseline" manipulation and then the contrast between "non-adaptive" and "adaptive" presentations of synthetic speech during episodes of brown noise rather than any of the three types of noise used in the formal study. A summary of the seven listening exercises participants carried out in the body of the experiment is given in Table 1.

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

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

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 poorest 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 very 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 auditory cognition is perceptually tuned to attend to voices, Fastl noise 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 evaluate how performance in the two presentation strategy manipulations differ from performance in the Baseline condition across the three manipulations of noise-type.

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 (increased by 6 dB) to compete in parallel with the eight randomly ordered noise events in its particular manipulation. The resulting modulations were then appropriately delayed (i.e., shifted forward in running time) to simulate the time it takes for the onset of a noise event to cross the system's response threshold. Thus for noise events at the forte level (-26 dB), the synthetic speech starts to become louder 1.0 sec. after the short event begins and 3.0 sec. after the long one begins, the difference being due to the more gradual onset ramp of long events (see Section 2.1 above). The short and long episodes of noise at the fortissimo level (-19.dB) have correspondingly steeper onset ramps, so the response for these events begins at 0.8 and 2.0 sec., respectively. Fortissimo 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 fortissimo 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 kinds 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-forte 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-fortissimo 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-fortissimo 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 Fig. 2.

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-fortissimo/NA-white, long-forte/NA-pink, short-fortissimo/NA-Fastl, and short-forte/NA-Fastl. ADAPT presents an instance of each of the four noise event types in the Adaptive manipulations: long-fortissimo/A-white, long-forte/A-pink, short-fortissimo/A-Fastl, and short-forte/A-Fastl.

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

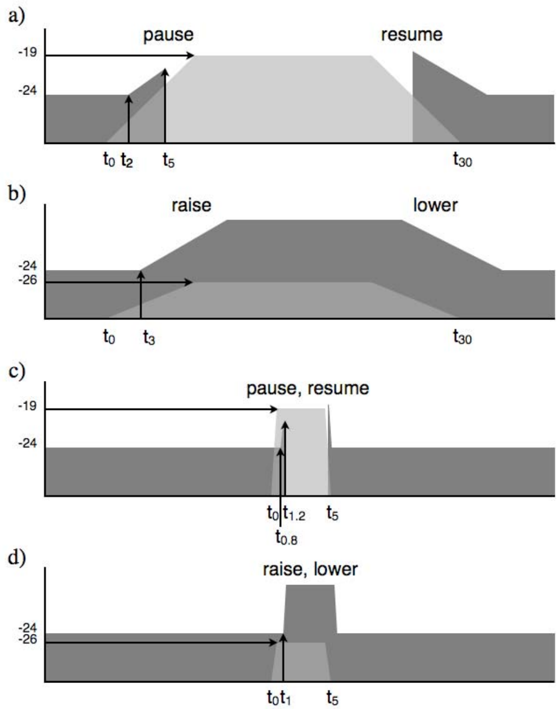

Fig. 2. Schematic diagrams showing actions taken by the auditory display in the experiment's Adaptive conditions to counter noise events with the potential to mask speech from the listener's perspective: (a) long-fortissimo, (b) long-forte, (c) short-fortissimo, and (d) short-forte. Time in seconds is shown on the horizontal axis, and level in dB is shown on the vertical axis. Noise event envelopes are shown as semi-transparent light gray trapezoids. Envelopes of continuous speech are shown in dark gray. See the text for additional details.

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. Each list contained both the targeted noun phrases and foils in equal numbers (eight targets in each of the training sessions and twenty targets per manipulation in the formal listening exercises). Targets were listed in the order of their aural 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 the commentary 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 commentary 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.

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.

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

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 largely consistent with the pattern of listening performance that was expected to arise between the noise-free Baseline condition and the six conditions featuring noise events and the non-use or use of the adaptive auditory display. 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 Fig. 3 and Fig. 4.

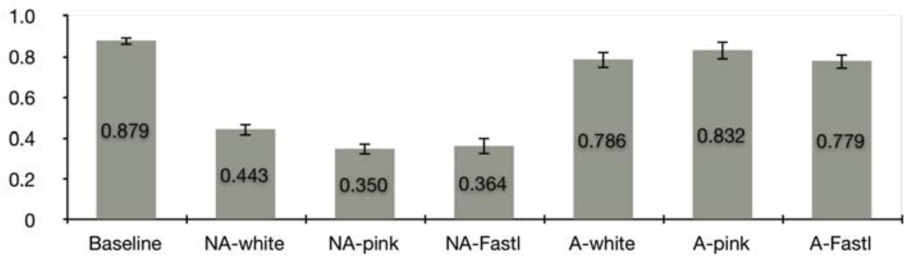

Fig. 3. Plot of the mean proportion of correctly identified target noun phrases, p(targets), in each condition. The y-axis shows proportion. Error bars show the standard error of the mean.

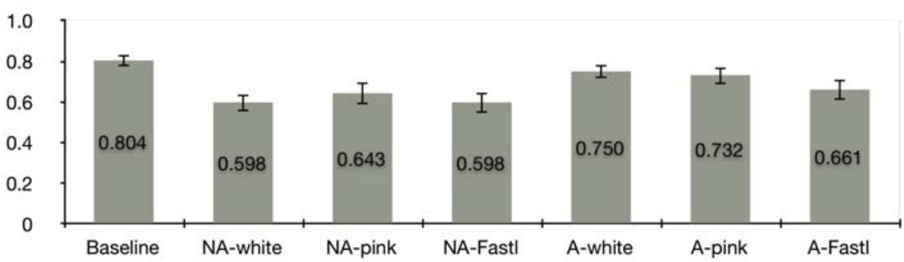

Fig. 4. Plot of the mean proportion of sentences correctly judged as "old" or "new," p(sentences), in each condition. The y-axis shows proportion. Error bars show the standard error of the mean.

To evaluate effects of presentation strategy—non-adaptive vs. adaptive—and the three types of noise on listening performance, the six conditions involving noise events were construed as a factorial design, and a two-level by three-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 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 auditory display strategies were used, planned contrasts were carried out to evaluate how the dependent measures in these manipulations differed with performance in the Baseline condition. All of these comparisons involving the Non-Adaptive manipulations were significant, meaning that both performance measures, *p*(*targets*) and *p*(*sentences*), were meaningfully hurt by the noise events in these conditions. In other words, as was expected, all three types of noise proved to be good maskers of synthetic speech. The *F* statistics for the contrasts involving the Non-Adaptive conditions are summarized in Table 3.

Table 3. *F* statistics for the planned contrasts between the Baseline and Non-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.

Measure	Contrast	*F*
p(*targets*)	**NA-white** vs. **Baseline**	$F(1, 13) = 200.718$, $p < 0.001$*
	NA-pink vs. **Baseline**	$F(1, 13) = 354.169$, $p < 0.001$*
	NA- Fastl vs. **Baseline**	$F(1, 13) = 232.386$, $p < 0.001$*
p(*sentences*)	**NA-white** vs. **Baseline**	$F(1, 13) = 12.526$, $p = 0.004$*
	NA-pink vs. **Baseline**	$F(1, 13) = 5.692$, $p = 0.033$*
	NA-Fastl vs. **Baseline**	$F(1, 13) = 13.200$, $p = 0.003$*

A more interesting set of results emerged from the contrasts involving the Adaptive manipulations. Some of the contrasts in this set were not significant, meaning that the corresponding measures of performance were not substantially worse than the Baseline. This was the expected result, but it 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 4.

Table 4. 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.

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

Since the study, as it was primarily conceived, can be characterized as a test of the merit of adaptively improving the signal-to-noise ratio (SNR) of an auditory display, the outcome of these latter contrasts raises an important concern about the underlying uniformity of the experimental treatments. In particular, because of the fact that the auditory signals corresponding to both of the primary response measures—target noun phrases and sentences—were synthetic speech, it was not feasible to tightly equate the SNRs of the stimuli across the manipulations involving noise events. Noun phrases, such as "hammy lines," "hour-long interview," and "the whole point," for example, have different signal powers due to their differing lengths and differing patterns of phonemes, and the same can be said for each of the sentences that formed the basis of the sentence judgment task. A fair degree of preparatory attention was given to this matter, but as can be seen by examining the dark bars in Figs. 5 and 6, this key dimension of the stimuli was not distributed in a strictly uniform manner. (The underlying SNRs in these figures were calculated with a noise floor of -96 dB; the plots show the mean SNR of signals occurring both with and without noise that required a response in each condition.) To determine whether or not there were significant differences between the stimulus SNRs in each of the six conditions involving noise events, a 2x3 ANOVA (two presentation strategies by three types of noise) was carried out for a) the SNRs of the target phrases and b) the SNRs of the commentary sentences that were selected for listeners to evaluate as "old" or "new." As expected, the SNRs in the adaptive presentations were higher than in the non-adaptive presentations, (targets: $F(1, 114) = 21.236$, $p < 0.001$; sentences: $F(1, 30) = 7.44$, $p = .011$. Note that this was by design. Critically, there were no differences in the SNRs between noise types (targets: $F(2, 114) < 1$, n.s.; sentences: $F(2, 30) < 1$, n.s.) and no interactions between the two factors (presentation and noise) (targets: $F(2, 114) < 1$, n.s.; sentences: $F(2, 30) < 1$, n.s.).

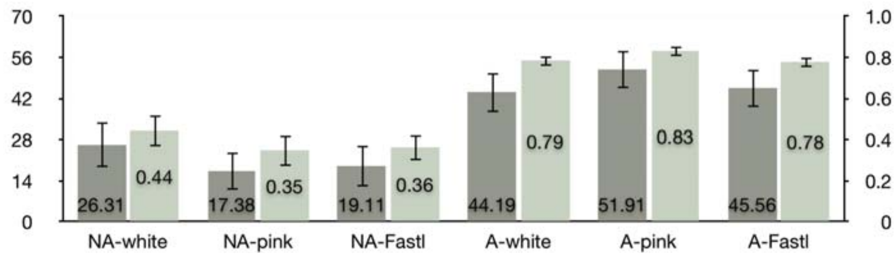

Fig. 5. Comparative plots of a) the mean signal-to-noise ratio (SNR) of the 20 target noun phrases in each of the manipulations involving noise events (dark bars) and b) the mean proportion of correctly identified target noun phrases (*p*(*targets*), light bars) in each condition. The y-axis on the left shows SNR in dB; the y-axis on the right shows proportion. Error bars show the standard error of the mean.

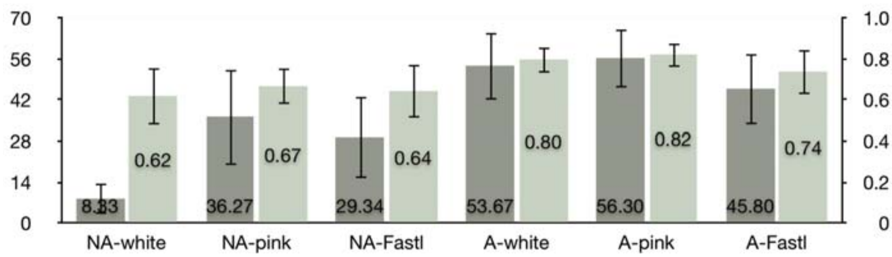

Fig. 6. Comparative plots of a) the mean signal-to-noise ratio (SNR) of the six sentences that corresponded to judgments entailing "original" or "paraphrased" content or a "meaning change" in each of the manipulations involving noise events (dark bars) and b) proportion of just these six sentences correctly judged as "old" or "new," (different from p(sentences), light bars) in each condition. (Note that the "distractor" sentences participants were asked to judge as old or new were entirely made up and did not correspond to specific commentary sentences.) The y-axis on the left shows SNR in dB; the y-axis on the right shows proportion. Error bars show the standard error of the mean.

Also plotted in Figs. 5 and 6, however, are the comparative mean proportions of correct responses for each type of stimuli (light bars). In Fig. 5, these are simply the values of *p*(*targets*) reported in Fig. 3. The mean proportional scores shown in Fig. 6 differ from *p*(*sentences*) in that they reflect only judgments involving sentences that were actually presented and potentially heard. (Only six of the eight sentences participants were asked to judge in each condition—specifically, those that were either "original" (verbatim), a "paraphrase," or that involved a conspicuous "meaning change"—meet this criterion (see Table 2). The other two sentences—the "distractors"—were topically-related fabrications that were not present in the commentaries; participants were expected to recognize that they had not heard distractors and, thus, mark them as "new.") A consistent correspondence between the response measures and the SNRs in these plots is readily apparent and the respective correlations are significant (targets: $R^2(118) = 0.191$, $p < 0.001$); sentences: $R^2(34) = 0.31$, $p < 0.001$).

Overall, then, 19% of the variability in p(*targets*) and, in Fig. 6, 31% of the variability in the proportion of correct, stimuli-based sentence judgments can be explained by the corresponding SNRs of the stimuli, and these correlations should be taken into account in the material below assessing differences in listening performance relative to the three types of noise used in the study.

In addition to the primary performance measures, an additional measure was associated with the sentence judgment response task, specifically, 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 overall mean proportion of demurs in all seven experimental conditions, shown in Fig. 7 (columns with error bars), exhibits, inversely, the same broad pattern as that seen for both p(*targets*) and p(*sentences*) in Figs. 2 and 3. 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 the 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 Fig. 7 are the corresponding counts of participants in each condition who chose to demur one or more times (inverted triangles), and the mean proportion of demurs made by just these individuals (square bracketed values; the lower bound of this proportion is 1/8 or 0.125). Not surprisingly, both of these series are significantly correlated with the overall mean p(*demurs*) values (respectively, Pearson's $r = 0.864$, $p < 0.02$ (2-tailed) and Pearson's $r = 0.851$, $p < 0.02$ (2-tailed)). These augmenting data are given to provide additional perspective on the nature of the comprehension task as a measure of human listening performance. In particular, the values in the more aurally favorable Baseline and Adaptive conditions show that a number of listeners (though less than half) did not recall what was said well enough, to various degrees, to be confident in their judgment of sentences as old or new information. In contrast, the corresponding numbers in the Non-Adaptive conditions reveal that several listeners (though again, less than half) chose not to demur in spite of noise events that masked sentences they were asked to judge.

Finally, a plot of the listeners' mean level of subjective agreement in each condition with the statement, "I prefer the way the synthetic speech was presented in this listening exercise," is shown in Fig. 8. As mentioned above, the range of this data corresponds to a seven point Likert scale. The resulting pattern of ratings across manipulations is somewhat similar to the patterns seen in Figs. 3 and 4. 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 ($F(1, 13) = 10.538$, $p = 0.006$).

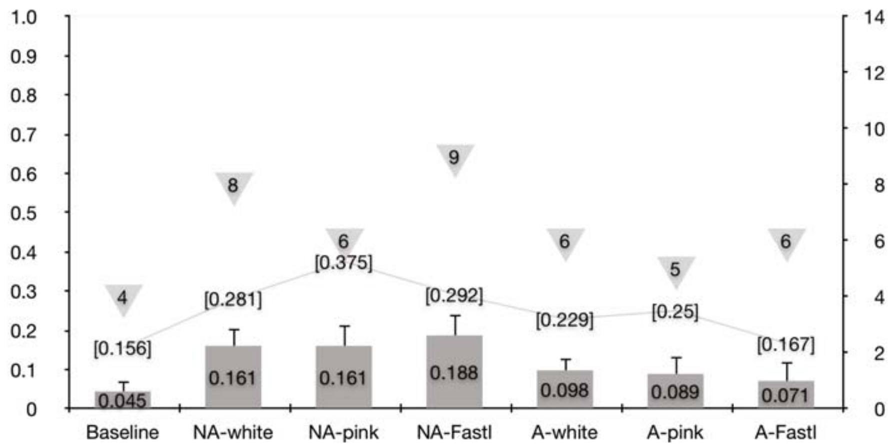

Fig. 7. Combined plot showing a) p(demurs), the mean proportion of "I don't know" responses across all listeners in each condition, (gray columns with error bars (showing standard error of the mean; negative error bars not shown to reduce clutter) corresponding to y-axis on left), b) number of listeners in each condition who demurred one or more times (inverted triangles corresponding to y-axis on right), and c) the mean proportion of "I don't know" responses in each condition made by listeners who demurred one or more times (square bracketed values corresponding to y-axis on left). See text for additional information.

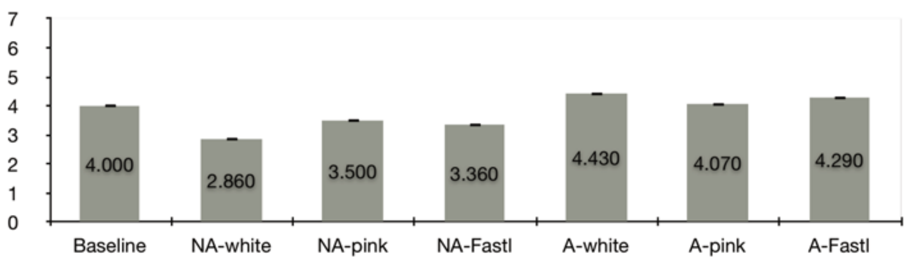

Fig. 8. Plot showing the mean level of participants' agreement with the statement, "I prefer the way the synthetic speech was presented in this listening exercise," in each condition. The y-axis in this plot reflects a seven-point Likert scale ranging from 1 = "strongly disagree" to 7 = "strongly agree." Error bars show the standard error of the mean.

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 elicited a noise-free baseline of listening performance 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.

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 non-individualized 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 (e.g., [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 is little alternative to synthetic speech in the less constrained aural interaction models that are called for if robots are to be accepted as credible actors in social settings.

4.2 The Impact of Noise Type on Listening Performance

Although the use of three different types of broadband noise, as surrogates for real-world noise capable of masking synthetic speech, was a secondary consideration in the design of this study, its outcome, with respect to maskers, suggests that one type of noise may be more difficult to adapt for than another.

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 contrasts reported in Table 4 provide a degree of evidence that some forms of noise—presumably, because of their particular characteristics (see Section 2.1.2)—can potentially undermine a listener's auditory concentration. First, note that the *p*(*targets*) interaction, which can be seen in both Fig. 3 and Fig. 5, 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. Although this pattern is correlated with the SNRs of the stimuli, the latter does not reflect an interaction and the observed variance in *p*(*targets*) is much smaller, particularly in the adaptive manipulations. Thus, pink noise was a good masker of synthetic speech, but it was also the best type of noise to successfully adapt for. Add to this the pattern of significant contrasts in the *p*(*targets*) portion of Table 4, and it is plausible that in spite of the adaptive auditory display, both white and Fastl noise impacted listeners in a way that pink noise did not.

Did these effects happen for similar reasons? Arguably not, when the contrasts in the *p*(*sentences*) portion of Table 4 are also taken into consideration. These contrasts show that listeners demonstrated a good understanding of the commentaries in the Adaptive manipulations involving white and pink noise events—relative to their performance in the Baseline condition—but were unable to do so in the A-Fastl condition. Given the notable descriptive 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 challenging type of masker to consistently overcome.

In seeming opposition to this interpretation, though, is the pattern of *p*(*demurs*) data shown in Fig. 5 and the associated data for participants electing to respond in this way in the post-listening sentence judgment task. If Fastl noise does impair auditory concentration, observing a substantial number of demurs would be good supporting evidence. However, the value of *p*(*demurs*) in the A-Fastl condition is essentially no different from this measure in the other two Adaptive conditions. Instead, the largest number of demurs occurs in the NA-Fastl condition, and furthermore, only this contrast with the Baseline value of *p*(*demurs*) is significant. In conjunction with the low value of *p*(*sentences*) in this condition, Fastl noise is the most successful masker of non-adapting synthetic speech in the context of post listening measures.

But is this pattern in *p*(*demurs*) data 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 (as opposed to simply overwhelming a target sound energetically), then it could be an even better masker than, say, unvarying continuous noise. Certainly, more participants in the NA-Fastl condition than in any other appear to have decided that they had a poor 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 fully masked. If this is so, then there should be a greater mean proportion of sentence judgment errors relative to the other Adaptive conditions, and this turns out to be the case (see Fig. 4). In fact, the mean proportion of sentence judgment errors in the A-Fastl condition (calculated as the remainder when *p*(*sentences*) and *p*(*demurs*) are deducted from a perfect score) is greater, at 0.268, than the corresponding proportion of errors in any of the other conditions in the study.

Thus, more so than white or pink noise, Fastl noise appears to be a good masker of synthetic speech, both during and after listening and even when adaptive changes in loudness and the use of pauses are employed. If this conclusion is right, it has implications for the design of auditory human-robot interaction in social settings because Fastl noise is taken to be a type of analog for speech noise. However, because the underlying pattern of stimulus SNRs in the study is significantly correlated with the primary performance measures (see Section 3), this result will require additional study.

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 mean preference data shown in Fig. 6 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 rating for the Baseline condition, though, turns out to be exactly midway between the two ends of the Likert scale used for this measure. On balance, then, listeners seem to have been indifferent to the use of synthetic speech by itself. But two factors may have contributed to this outcome. First, this condition was always heard first by listeners in the full experiment and, second, nothing of consequence (i.e., no interruptions, etc.) occurs in this manipulation, which, after having gone through the training exercises, would make the response tasks seem relatively straightforward.

Not knowing what might be coming next and having little additional basis for expressing a preference may well be the best explanation for this neutral outcome.

In the Adaptive manipulations, on the other hand, 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 speculated that their agreement with the preference statement was somewhat higher in the Adaptive conditions because noise events were stimulating and 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. Recognizing the need for enunciation could perhaps be informed by machine classification of the ambient noise environment. Yet, 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.

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.

References

1. Peres, S.C., Best, V., Brock, D., Frauenberger, C., Hermann, T., Neuhoff, J., Nickerson, L.V., Shinn-Cunningham, B., Stockman, A.: Auditory Interfaces. In: Kortum, P. (ed.) HCI Beyond the GUI, pp. 145–195. Morgan Kaufman, San Francisco (2008)
2. Brock, D., Martinson, E.: Exploring the Utility of Giving Robots Auditory Perspective-Taking Abilities. In: Proceedings of the 12th International Conference on Auditory Display (ICAD), London (2006)
3. Thrun, S., Beetz, M., Bennewitz, M., Burgard, W., Cremers, A.B., Dellaert, F., Fox, D., Hähnel, D., Rosenberg, C., Roy, N., Schulte, J., Schulz, D.: Probabilistic Algorithms and the Interactive Museum Tour-guide Robot Minerva. Intl. J. Robotics Res. 19, 972–999 (2000)
4. Martinson, E., Brock, D.: Improving Human-Robot Interaction through Adaptation to the Auditory Scene. In: HRI 2007: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Arlington, VA (2007)
5. Brock, D., 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. Kagami, S., Sasaki, Y., Thompson, S., Fujihara, T., Enomoto, T., Mizoguchi, H.: Loudness Measurement of Human Utterance to a Robot in Noisy Environment. In: HRI 2008: Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction, Amsterdam (2008)
7. Brock, D., McClimens, B., Trafton, J.G., McCurry, M., Perzanowski, D.: 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 (2008)

8. Hardee, J.B., Mayhorn, C.B.: 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, pp. 1143–1147 (2007)
9. Stevens, C., Lees, N., Vonwiller, J., Burnham, D.: Online Experimental Methods to Evaluate Text-to-speech (TTS) Synthesis: Effects of Voice Gender and Signal Quality on Intelligibility, Naturalness, and Preference. Computer Speech and Language 19, 129–146 (2005)
10. Cepstral, `http://cepstral.com`
11. Fastl, H., Zwicker, E.: Psychoacoustics: Facts and Models, 3rd edn. Springer, Berlin (2007)
12. Royer, J.M., Hastings, C.N., Hook, C.: A Sentence Verification Technique for Measuring Reading Comprehension. J. Reading Behavior 11, 355–363 (1979)