Auditory Perspective Taking

Eric Martinson and Derek Brock

***Abstract*—Effective communication with a mobile robot using speech is a difficult problem even when you can control the auditory scene. Robot self-noise or ego noise, echoes and reverberation, and human interference are all common sources of decreased intelligibility. Moreover, in real-world settings, these problems are routinely aggravated by a variety of sources of background noise. Military scenarios can be punctuated by high decibel noise from materiel and weaponry that would easily overwhelm a robot's normal speaking volume. Moreover, in nonmilitary settings, fans, computers, alarms, and transportation noise can cause enough interference to make a traditional speech interface unusable. This work presents and evaluates a prototype robotic interface that uses perspective taking to estimate the effectiveness of its own speech presentation and takes steps to improve intelligibility for human listeners.**

***Index Terms*—Acoustic propagation, auditory displays, human–robot interaction, robot sensing systems.**

I. Introduction

IDENTIFYING and applying human factors that promote utility and usability are an overarching concern in the design of auditory displays [1]. The importance of this tenet is particularly relevant for robotic platforms that are intended to be actors in social settings. The public naturally wants to interact with robots via means that are already familiar, and aural communication is arguably the mode many would expect to be the most intuitive and efficient for this purpose.

Implementing an auditory user interface for a robot calls for complementary machine audition and auditory display systems. These are both multifaceted functions that present a number of 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 localization, 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 exercise flexible interactive behaviors that are guided by the perception, interpretation, and production of sound-based information lies far behind the generally effortless skills of human beings. The computational challenges of auditory scene analysis and many aspects of natural language dialog are two of the key reasons for this shortcoming. Surprisingly, though, little work has addressed the kinds of practical situational reasoning robots will also need for successful auditory interactions in everyday sound-rich environments.

For example, in speech and auditory interactions with each other, people typically take into account a number of 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 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 that 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 the 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. Moreover, 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.

Prompted by this insight, the authors have designed and implemented a computational auditory-perspective-taking system for a mobile robot. Using ray-tracing-based room acoustic simulations, a robot estimates signal-to-noise ratio (SNR) at a

Manuscript received October 14, 2011; revised May 12, 2012; accepted September 1, 2012. This work was supported in part by the Office of Naval Research under Contracts N0001409WX30013 and N0001411WX30017. This paper was recommended by Editor V. Murino.

The authors are with the U.S. Naval Research Laboratory, Washington, DC 20375 USA (e-mail: eric.martinson.ctr@nrl.navy.mil; brock@itd.nrl.navy.mil).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2012.2219524

detected listener's location and acts to maintain intelligibility by inferentially altering the level and/or progress of its speech or, when necessary, offering to move to a quieter location. In contrast to previous work on this subject [2], the auditory perspective of the listener is described by the theory of sound flow through a room, integrating knowledge of environmental obstacles, sources of masking noise, and listener characteristics into its decision-making process. In the following material, we describe the technical details, core functions, and an evaluation of key facets of the system's performance. In addition, we summarize the results of a recent empirical study in which changes in the level and progress of synthetic speech predicated on the system's auditory-perspective-taking scheme in the presence of noise were evaluated for their impact on measures of user listening performance.

II. Adaptive Auditory Interface for a Mobile Robot

The purpose of an information kiosk, traditionally, has been to provide information about the environment to interested people. The types of kiosks differ dramatically. A very simple kiosk might just relate the day's weather conditions or list the set of departing flights at an airport. A more advanced kiosk could be a computerized map, where people use a mouse, keyboard, or touch screen to read reports about different objects on the map. At the farthest end of the spectrum, even people could be considered as a type of mobile information kiosk prepared to answer an arbitrary set of questions to the best of their abilities. Within this large range, our implementation of a robotic information kiosk fits somewhere between a stationary computerized map and the extreme of a person. An interested participant speaks the title of a story or object that he or she would like to have information about, and then, the robot uses text to speech (TTS) to read aloud the precompiled story that matched to that title.

Effective communication with a mobile robot using speech, however, is a difficult problem, particularly in real environments with significant ambient noise. To maintain speech intelligibility under such dynamic noise conditions requires auditory perspective taking. Similar to perspective taking in spatial reasoning [3], auditory perspective taking means having the robot use its knowledge of the environment, both *a priori* and sensed, to predict what its human counterpart can hear and effectively understand and then adjusting its spoken output or altering its position to maximize intelligibility and ease of use. It should be noted that this does not replace the theory of mind research, where the mental state of the listener is modeled to resolve misunderstandings. By modeling what is physically possible for an observer to hear, an auditory-perspective-taking system reacts in real time to prevent misunderstandings caused by poor perceptual conditions.

The remainder of this section is thus organized into three subsections. First, we outline the components that we used to support an adaptive "information kiosk" application involving auditory perspective taking. Next, we introduce a set of functions that the robot can use to learn and reason about sound in surrounding environment. Last, we cover the robot's aural adaptations and their integration into the information kiosk task.

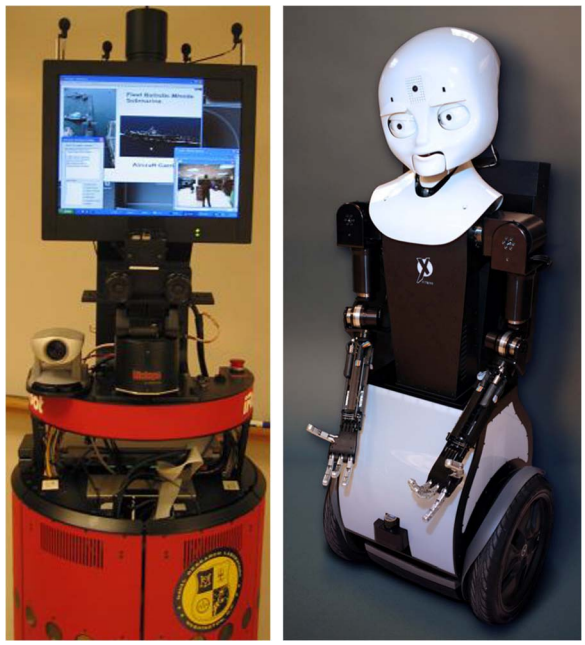

Fig. 1. (Left) B21r robot from iRobot and (right) MDS humanoid robot.

A. Robotic Hardware

Auditory perspective taking has been implemented on two robotic platforms: an iRobot B21r (the original prototype [2]) and a mobile dexterous social (MDS) humanoid robot with similar sensing capabilities (Fig. 1).

1) An array of microphones for monitoring ambient noise (B21r: overhead; MDS: on the backpack) composed of four lavalier microphones routed to separate battery-powered preamplifiers and, then, to an eight-channel data acquisition board.
2) A loudspeaker to allow the robot to speak at different volumes to its listener.
3) A vision system for tracking users. The B21r uses stereovision, while the MDS uses a combination of color and time-of-flight cameras.
4) A laser measurement system (B21r: SICK LMS200; MDS: Hokuyo) used with continuous localization [4] to provide reliable robot pose (position and orientation) estimates. (Additional pose information was gathered with external cameras when using the MDS humanoid platform.)

A wireless microphone headset, worn by the primary user (i.e., the listener), is utilized in both implementations to speak to the robot. Spoken commands are processed and recognized by Nuance's Dragon Naturally Speaking software. Microphone arrays can also be used for input to speech recognition systems, e.g., [5], and ultimately, this will be a more acceptable speech interaction solution. However, given the focus on perspective taking in this work and the requirements of other research, our microphone arrays are not presently targeted for this purpose.

B. Auditory Scene Modeling

Taking the perspective of a human listener first requires a basic understanding of the auditory scene. In this section, we

briefly cover the tools, techniques, and information that our system integrates to determine how the intelligibility of its spoken output can be maintained for its user.

1) Tracking Users Visually: Visual tracking is provided using two different vision systems. On the B21r, stereovision is realized with a rotatable TRACLabs Bi-Clops that provides dual color images from which depth information can be extracted. Combined with face detection software developed with OpenCV [6], the camera can localize and follow a detected person through a 180° arc in front of the robot. On the MDS humanoid, a detected face initiates a track. Depth information from the SR3100 time-of-flight camera is then utilized in conjunction with a color histogram to determine whether the individual is still within view of the camera [7]. Both systems are used to indicate the presence of a human and estimate relative distance and angle from the robot.

2) Speech Localization: Understanding the surrounding auditory scene and estimating its implications for human–robot interaction both require effective sound source localization. In general, there are at least two types of sound sources that a robot should be able to localize. The first type is people using speech. Before the robot's vision system can be used to find someone who is speaking, its audition system needs to detect and localize the source of the speech. Only then can the camera be oriented correctly. Speech detection is accomplished by calculating mel-cepstrum coefficients [8] for sampled audio and comparing the first coefficients to an environment-dependent threshold. Given a 10-s training period, over which the mean and maximum MFCC_0 are identified, the threshold

$$\mathrm{threshold} = 0.5 * (\mathrm{mean\_mfcc}_0 + \mathrm{max\_mfcc}_0).$$

Applying this threshold classifies wide spectrum sounds as possible speech. Such samples are processed using generalized cross-correlation (GCC) to estimate their direction of origin and combined over time using a 1-D auditory evidence grid [9]. A person is localized when 1–2 s of speech evidence accumulates for a particular direction.

3) Ambient Noise Localization: People, however, are only part of the auditory scene and not its entirety. Many common items such as televisions, radios, and air vents are significant stationary sources of sound that can easily raise ambient noise levels and disrupt aural interactions by reducing the intelligibility of speech. Given enough time, a mobile robot can investigate its operating environment and localize all fixed sources of sound in 2- or 3-D space using the same auditory evidence grid formulation [10] that is used for speech localization. While it moves, the robot samples data from loud areas, analyzes these data using GCC to estimate the direction of incoming sound, and adds the results to a 3-D evidence grid containing estimates from other locations. Over time, regions in the grid containing real sound sources (versus reflections or transient noise) increase in likelihood while reflections are suppressed, effectively triangulating on active sound sources. Fig. 2 shows a completed grid localizing three sound sources from robotic sampling.

In general, however, completing a full mapping before each interaction begins is not a realistic assumption. Interactions often occur too quickly, or environment circumstances could make autonomous movement impractical. The current solution is only approximate. It assumes that the result of the last exploration is good enough at interaction time and that no significant changes have occurred in the aural environment. A better solution could incorporate a history of discovered sound sources. Sources repeatedly discovered in the same location could be acoustically modeled. Then, when new sounds are locally detected during an interaction, the robot could recognize sources in its history and update its source list without having to move around the environment.

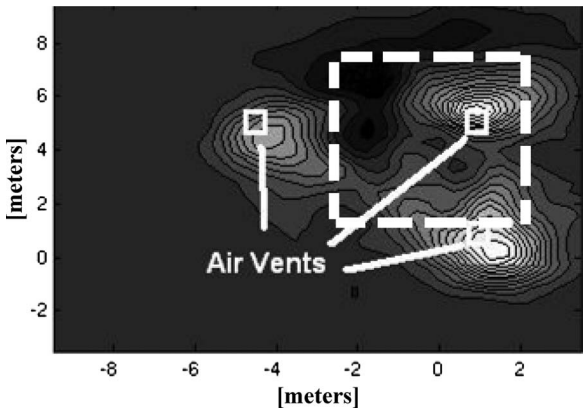

Fig. 2. After exploring the area inside the dashed square, the robot was able to identify the location of three low-volume air vents. Accuracy decreased with the air vents' distance from the area explored by the robot.

4) Noise Volume Estimation: Finally, the combined effect of persistent noise sources in the robot's operating space can be estimated for online use through acoustic ray tracing [11]. The basic idea behind ray tracing is to generate and follow a large number of virtual rays as they travel from their origin at a known sound source into the surrounding environment. Each ray begins with a known power relative to the power of the originating sound source. This power then decays with the distance from the ray origin and each reflection until the power of the ray falls below some minimum threshold, and the ray is discarded. Sound volume is calculated using these rays by placing virtual sensors in the simulation environment. Every time a ray passes through the region of space occupied by a sensor, the ray is added to the measured sound at that location.

In contrast to simple linear decay models of sound volume propagation used in the original prototype [2], ray tracing allows an adaptive system to include reverberations from walls and other obstacles in its noise volume estimations. This increased ability, however, requires a map of the environment. Occupancy grids are a map representation that can be quickly constructed either by hand or autonomously by the robot itself [11]. To use these, however, some simplifying assumptions are required.

1) Surface diffusion is assumed to be constant across the entire environment. As the robot has no sensors for measuring this value, 0.25 was picked as a good average across both large and small surfaces [10].
2) Reflections (specular and diffuse) occur at the edges of occupied grid cells. As grid cells only indicate the presence of a reflecting surface, not its orientation, deviations in surface orientation from a major axis are directly

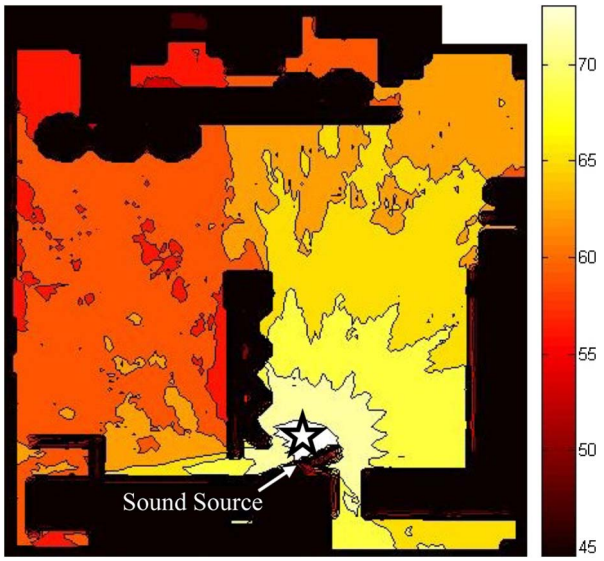

Fig. 3. Two-dimensional slice of the volume estimates produced by acoustic ray tracing for an 80-dBA sound source.

related to error in reflected ray directions. This assumption has the greatest impact on phase estimation, which is not being used to create sound volume maps.

3) All surfaces are assumed to have the same absorption coefficient α

$$\alpha = 0.161 * V/(T_{60} * S).$$

This value is determined empirically from the measured room reverberation time (T_{60}), room volume (V), and total surface area (S) [12]. S and V are estimated directly from the occupancy grid.

For a typical application of acoustic ray tracing, one or more virtual sensors would be created within the bounds of the described obstacle map. A robotic system, however, needs a map to identify regions and plan movement through the environment. To generate a map of sound volume, sensors are selected such that there exists one sensor for every grid cell in the occupancy grid. Sensors are modeled as cubic volumes. These sensors then store and record the set of all rays generated by the source or reflections off obstacles that pass through them. In this work, we used different numbers of rays for different situations. Prior to an interaction, the robot can model known sound sources with 200 000 rays. During an interaction, however, newly discovered sources are modeled with only 50 000 rays to speed up the modeling process. Fig. 3 shows an example noise map created through acoustic ray tracing.

C. Implementation of Aural Adaptations

Our implementation of an adaptive information kiosk focuses on maintaining the intelligibility of the application's TTS output. When a user indicates that he or she wants to use the kiosk, the robot effectively takes the listener's auditory perspective by determining what should be intelligible to someone at the detected location and then executes steps to ensure that its synthetic speech can be easily understood.

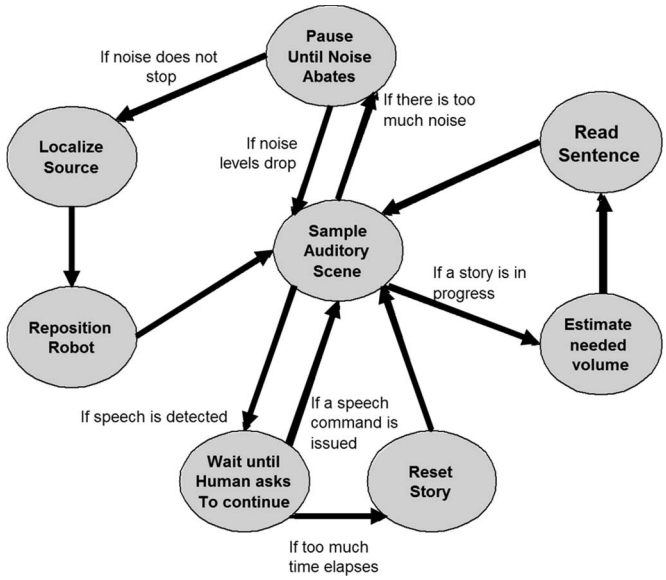

Fig. 4. Sequence of steps that the robot takes while reading a story to a human listener. Starting with the step in the center, the robot samples the auditory scene, estimates the current SNR, adapts its auditory output, reads a sentence, and repeats.

The system resources outlined in the previous section allow the robot to implement four basic courses of action to maintain the intelligibility of its speech output and mitigate aural problems that the dynamics of ordinary auditory settings can pose for listeners: 1) The robot can face the listener; 2) the robot can modulate its speaking volume; 3) the robot can pause or stop in the midst of speaking; and 4) the robot can move to quieter location. The latter three of these actions are coordinated via a finite-state automaton, which is shown in Fig. 4. Orienting to face the listener relies on the vision system after the initial determination of his or her position, so this process is run in parallel with the other actions.

1) Facing the Listener: Loudspeakers are directional sources of sound. They are loudest to the front and fade to the sides in a cardioid pattern. As the properties of loudspeakers are known *a priori*, a perspective-taking interface could include directivity when modeling speech output and change volume levels as listeners change the angle of their position relative to the robot's orientation. A simpler human-inspired approach, however, is to continuously reorient the loudspeaker to face the listener. This has two advantages. It simplifies modeling by approximating sources as omnidirectional, and it allows visual (B21r) or gestural (MDS) communications to be presented more effectively.

Rotating to face the listener requires a combination of speech localization and visual tracking. Since the kiosk can be approached from any direction, the robot waits for a potential kiosk user to say something and then uses its speech detection and localization tools to determine which way it should face. Thereafter, visual tracking takes over. The kiosk reorients whenever the user moves more than 30° to the left or right.

2) Modulating Its Speaking Volume: The information kiosk is designed to relay textual information to an interested individual. In a noisy environment, people naturally adjust not just the volume but also the properties of their vocal tract

to accommodate the ambient conditions [13]. Unfortunately, computer-synthesized speech has not yet reached that level of naturalness [14]. Our robotic adaptations, therefore, are focused on volume adjustments relative to ambient noise conditions.

Once a user has selected a topic, recognized as a verbal command by a commercial speech recognition engine, the robot estimates the user's listening requirements and reads corresponding text aloud at a speaking volume that balances intelligibility and social norms. Between each sentence, the system reassesses its volume on the basis of estimated ambient noise at the detected listener's physical location. Noise levels are determined by detecting changes to ambient volume with the robot microphone array and adding an offset to modeled active source sound levels. As a result, the kiosk uses different volumes for different locations in the room, raises its voice in response to new ambient noise, and/or the listener steps back and speaks more quietly when ambient noise abates and/or the user moves closer.

3) Pausing for Interruptions: Some environments are occasionally subject to unreasonably loud noise events. In military settings, for example, users of speech interfaces can periodically expect to encounter overwhelming levels of noise due to aircraft, vehicles involved in logistical operations, and even weapon fire. Excessively loud sounds also occur on city streets and in various kinds of public spaces. In these circumstances, speech can become unintelligible, and people usually stop speaking and wait until the sound passes. In anticipation of applications where this situation can easily arise, the kiosk uses its perspective-taking system to determine when the level ambient noise is likely to overwhelm the listener's ability to hear what it is saying and will pause if necessary until the intervening sound has subsided. When the kiosk resumes, it begins with the phrase, "As I was saying. . ."

When the kiosk is speaking, it will also pause to accommodate the listener's or another person's speech—in fact, it brings its presentation to a full stop. Unlike noise and other kinds of auditory events, secondary speech that arises in the midst of speaking that is already in progress does not necessarily impact the latter's intelligibility; it may, however, be intended to interrupt and draw the listener's attention elsewhere. As the kiosk is unable to determine the meaning or intent of such speech, it relies on proximity to determine whether an interruption is in order. The perceptual system for speech detection and localization (Section II-B2) uses an environment-dependent threshold to separate speech from background noise. To be identified as interrupting speech, a talker must either be nearby or substantially louder than normal. When this occurs, the kiosk stops speaking until it regains its audience under the assumption that its presentation is no longer the focus of its listener's attention. This action functions as a simplified form of social decorum in which the robot is regarded not as a peer but as an appliance. In particular, the auditory-perspective-taking scheme has no computational facility for theory of mind functions such as inferring user intentions and perceptual abilities. Consequently, since it cannot infer when its listener's attention has returned on the basis of the kind of visual or indirect speech-based (semantic) evidence that people ordinarily exhibit with each other, the robot uses its screen (B21r) or raises an eyebrow (MDS) to indicate that it has been interrupted and is waiting for a new command. When the user is ready, he or she can choose from a set of commands, including "Continue where you stopped," "Repeat from the beginning," "Repeat the last line," and "Change to a new subject." These phrases allow users to control the contents of the TTS output depending on how much they remember from before the interruption.

4) Relocating the Robot: The final action in response to a noisy auditory scene is to move someplace else to reduce the masking effect of persistent noise on the intelligibility of the kiosk's speech. For instance, if it is located next to a loud air vent when the ventilation system starts up, it does not have to stay there. If this situation occurs, when the robot is not engaged with a user, it is free to move autonomously. Otherwise, it first asks its listener if he or she would like to relocate.

To determine where to relocate, our system estimates noise volumes across the entire region of reachable space. When a new source of ambient noise is detected and localized, its effects on the auditory scene are estimated using ray tracing, and the results are added to the robot's internal noise map. This then allows the robot to identify relative volume levels throughout the environment and identify a suitable alternative location within acceptable ambient noise levels for both the human and robot participants.

D. Interface Discussion

In general, our adaptive auditory interface for a mobile information kiosk integrates a wide range of robotic capabilities to enhance speech intelligibility. It models the auditory scene and its effects on a listener, taking the human's perspective by integrating prior work in sound source localization and visual detection/tracking of humans into acoustic models of the surrounding environment (Section II-B4). Then, when the modeled human perspective suggests interference, the robot can act to maintain intelligibility—rotating to keep the listener in front of the loudspeaker, changing the volume of its speech output, pausing for interruptions, or offering to relocate the conversation to somewhere with a higher SNR. Altogether, this diverse skill set allows a robot to respond to a dynamic auditory scene in a way that, as will be demonstrated, is both effective and expected by a human partner.

III. Robotic Perception Evaluation

The application of perspective-taking skills to a working robotic system has followed the general goals of continuous adaptation to promote intelligibility and ease of use, thereby improving overall knowledge transfer between the robot and the human listener. These adaptations are dependent upon reliable robotic perception. Facing the listener (and pausing for interruptions) requires first recognizing speech in noisy environments. Adequately raising speech volume depends upon accurately factoring environmental noise into models of the auditory scene in order to predict intelligibility. Finally, relocating the robot to a quieter place requires accurate sound volume maps. In this section, we focus on evaluating the information kiosk's perceptual capabilities, specifically speech detection

TABLE I
EXPERIMENTAL RESULTS OF SPEECH DETECTION AND SOUND SOURCE LOCALIZATION

Ambient Noise Level	52 dB (No Sources)	63-65 dB, (Seamer)	67 dB (Rattle)
Maximum Detection Distance	6-m	3-m	2-m
Mean Localization Error	4.6 deg	6.1 deg	12.4 deg

and localization, the accuracy of acoustic ray-tracing estimates, and the effectiveness of relocating the robot. Together with a human study examining the effectiveness of adjusting speech volume and pausing between sentences (Section V), these experimental results demonstrate the advantage of an auditory-perspective-taking system.

The perceptual capabilities of the information kiosk were evaluated using recordings made in a working machine shop on the Naval Research Laboratory. The machine shop environment was selected for both its potential future applicability and its similarity to other scenes involving Navy hardware. Jet engines, trucks, and other mechanical systems can generate significant volumes of noise that people must be able to communicate around.

A. Speech Detection and Localization

Speech detection among ambient noise is a critical component of an auditory-perspective-taking system. Aside from maintaining the maximal efficiency of loudspeakers (Section II-C1), detecting and rotating to face speech let the robot track an individual interacting with the system (Section II-B1) and build models of what someone at their location would be hearing (e.g., their auditory perspective). Speech detection also assists when pausing for interruptions (Section II-C3). When people talk nearby or loudly, it is disruptive, and the system should pause for the interruption to pass.

To evaluate speech detection and localization, the robot was exposed to three different auditory scenes created from the machine shop recordings: 1) no active sound sources (52 dB ambient); 2) an electric seamer at moderate distance (63–65 dB ambient); and 3) a nearby machine rattle noise (67 dB ambient). In each environment, the robot was given 10 s to learn the ambient noise characteristics. Then, a loudspeaker was used to play a 5-s speech utterance in: 1) 1-m increments along a line away from the robot to determine the maximum distance at which speech could still be detected and 2) 20° increments at 2 m from the robot to estimate localization accuracy for the detected speech. The results of these tests are shown in Table I.

As expected for a threshold-based system, the radius around the robot in which speech is detectable shrinks with the volume of masking noise in the environment. This is the desired effect for determining the disruptiveness of speech in the environment. In louder environments, speech must be significantly closer in order to be equally disruptive. Localization accuracy also decreases with higher ambient noise levels but remains high enough under all noise conditions to correctly orient visual tracking systems that have 30° field of view.

B. Volume Estimation Accuracy

Using ray tracing for sound volume estimation has been tested previously in the literature [10] and even migrated to professional acoustical engineering tools [15]. This work does not need to revalidate the algorithm. However, the use of occupancy grids to facilitate robotic collection of environmental maps does introduce additional error into the estimation process. In this section, we demonstrate that, even with this additional error, ray tracing shows significant estimation improvement over spherical spreading from an ideal source [12]. For the following tests, acoustic ray tracing uses hand-created obstacle maps of the environment. These maps include walls and other large furniture but leave out smaller objects. The spherical spreading method, used in previous robot mapping work to estimate volume decay [11], ignores the effects of individual obstacles. It assumes that sound pressure decays inversely with distance from the sound source and treats reverberation as constant over the entire room.

The effectiveness of sound volume predictions was evaluated using a loudspeaker, a camera, and a hand-moved microphone in two different-size rooms, i.e., a 3×5 m^2 office and an 8×10 m^2 laboratory. For each microphone position, the stationary loudspeaker played a short sound pulse. The microphone then recorded the SNR between the recorded impulse and the background noise. A visual marker of known size located under the microphone was tracked by the camera to measure the distance from the sound source.

Fig. 5 compares the measured results to the SNR predicted by each method. The two graphs show the difference between the different environments. In the small office, reverberation is more significant. As expected, acoustic ray tracing more closely follows the measured SNR than the ideal source model which ignores reverberant effects. Also, because of the dominance of reverberation in the small environment, it is clear that absorption was overestimated in this case as the measured values are all greater than the estimated ones. In the larger environment, we see a closer match between measured and estimated values. Ray tracing still outperforms the ideal source model but only at greater distances from the source. In general, incorporating the effects of obstacles into sound propagation, even if using a coarse representation of objects like an evidence grid, is a better alternative to the constant reverberation approximation.

C. Relocation Effectiveness

Unlike adjusting the robot's speech output, repositioning the robot has an objective measurable goal, specifically, to reduce the level of detectable masking noise. Therefore, we evaluated the effectiveness of this action by allowing the robot to move itself from an initially noisy location to a quieter location and then measuring the difference in noise levels. The auditory scene, shown in Fig. 6, uses three recorded sounds from the machine shop (lathe, electric seamer, and a rattle noise from a loose part in one machine) and a recording of human speech to create a dynamic environment in which a robot might operate. All of these sources represent ambient noise that might interfere, but no more than three sources were active at any time.

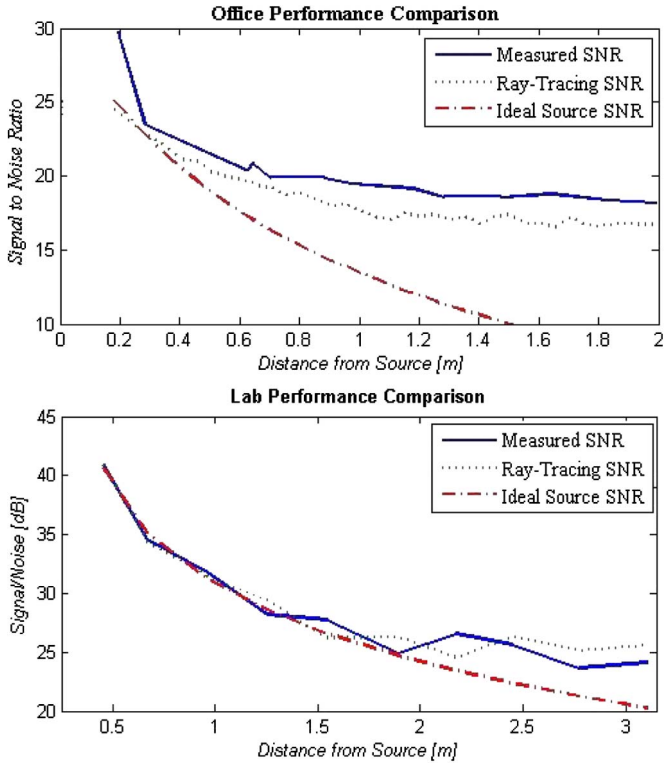

Fig. 5. SNR estimate versus measured graph for (top) an office environment, and (bottom) a larger laboratory environment.

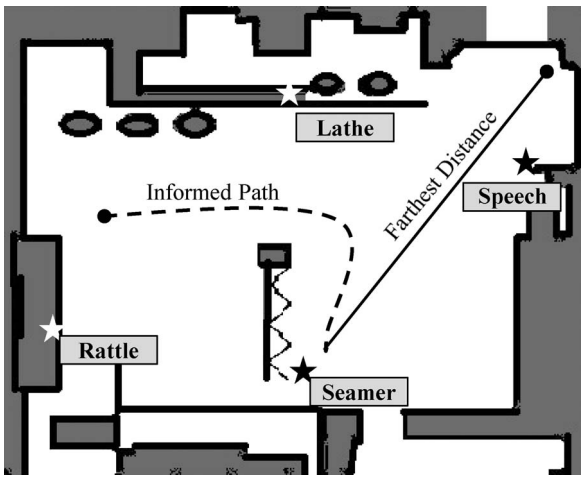

Fig. 6. Map of the room illustrating differences between movement strategies when relocating the robot. When two sources are active (seamer and speech), the informed path leads to a 7-dB improvement.

The noise reduction strategies evaluated were as follows: 1) Identify the farthest distance from the source, and relocate, and 2) add the source effects to an acoustic noise map, and identify the quietest location. The robot began in a quiet section of the scene that is interrupted by a "new" nearby source of noise. When ambient noise levels recorded by the robot exceed 60 dBA, it seeks a new improved location. For both noise avoidance strategies, the robot identifies the direction to the new sound source and assumes that the sound source is 1 m away in the detected direction. The new source's volume is calculated from the detected change in ambient volume level. The farthest distance strategy then uses that position to identify

TABLE II
COMPARATIVE IMPROVEMENT IN AMBIENT NOISE LEVEL IN A-WEIGHTED DECIBELS OBSERVED BY ROBOT AND HUMAN USER AFTER "FARTHEST DISTANCE" AND "INFORMED AVOIDANCE" RELOCATION STRATEGIES

Means for each "new" source (4-5 trials / source).					
New Source	Source Volume	Farthest Distance		Informed Avoidance	
		Robot	*Human*	*Robot*	*Human*
Lathe	72	-1.4	1.4	4.9	4.9
Rattling	82	10.7	10.2	13.1	12.7
Speech	74	10.3	8	10.9	7.6
Seamer	80	7.2	9.4	10.4	13.7
Statistics across all 18 trials					
	Max	14	13.5	15.5	17
	Min	-10	-5	3.5	0.5
	Mean	7.0	7.5	10.0	10.1

the region of clear space in the obstacle map that is farthest away from the new source. The informed avoidance strategy uses acoustic ray tracing to estimate the effects of the new sound source on the environment and adds to the noise map created for other known sources. It then moves to the quietest reachable location in the map.

A total of 18 trials was completed using each noise reduction strategy: four trials with only one active source, 12 trials with two active sources, and two trials with three active sources. Table II summarizes the improvement in ambient noise levels due to each relocation strategy. Columns marked with "Robot" indicate the difference in noise levels between robot positions, while the human column indicates noise differences at a position 1 m from the front of the robot. Results per source are the mean improvement across all trials in which that source was turned on near the robot to raise sound levels above 60 dBA. A negative number indicates an increase in ambient noise levels.

In all but three trials, an informed approach to relocating the robot outperformed the farthest distance relocation, averaging a 3-dB improvement for the robot and 2.5 dB for the human. Of those three trials, the greatest difference between relocation strategies was only 2 dB for either human or robot locations, and two of those trials had a difference of < 1 dB. The difference between the two strategies is best shown, however, not by the average but by the worst case situations. In the worst case relocation, simply moving away from the target meant a negative performance improvement. When activating the relatively quiet lathe noise in the presence of other sources, the simple approach placed the robot next to other louder sound sources twice and, in the best case, improved the robot's noise levels by only 5 dB. Using an informed approach, however, the robot never ended up next to another sound source. It improved the ambient noise levels of both the human and the robot in all trials. Even when relocating with only one active source in the environment, meaning that moving farther away could not make the auditory scene worse, incorporating an obstacle map into the sound propagation estimates meant a 1.5-dB average improvement for both the human and the robot.

D. Perception Evaluation Discussion

This section empirically evaluated robotic perception in support of auditory perspective taking. It examined three system components: speech detection and localization,

ray-tracing-based volume estimation, and relocating the robot in response to poor SNR. The first component, i.e., speech detection and localization, was examined under three different noise conditions. While increased noise levels did adversely affect performance, it decayed as theoretically desired, reducing the range at which speech would interrupt an interaction.

Second, ray-tracing-based estimates for sound volume were compared to hand-collected information about the auditory scene. Volume estimates closely followed real measurements in two different environments. Importantly, this component depends only on data that can be autonomously collected, leaving room for a fully robotic investigation of any environment in which the interface is being deployed.

Finally, having these estimation capabilities enables a smarter relocation strategy as part of perspective taking. Equipped with a noise volume map, the robot outperforms pure avoidance-based strategies for improving ambient noise levels. These ray-tracing generated maps correctly indicated quieter regions with better SNR, and the robot could relocate effectively. Given an accurate map, alternative map-based strategies in the future could focus on closer but quiet enough positions in which to continue the interaction. The same map could identify regions in which the expected SNR is high enough but which are not too far out of the way. In general, estimating noise throughout the environment is an important auditory-perspective-taking ability, informing the robot of its human partner's acoustic circumstances and of expected performance elsewhere.

IV. Adaptive Interface Study

An empirical usability study was also carried out to evaluate the merit of the auditory-perspective-taking scheme's key functions, specifically its ability to make changes in the level and progress of the presented speech (originally reported in [16]). Usability was construed in terms of the impact of these adaptive behaviors on listening performance. To ensure that the experimental setting was the same for all participants, the auditory materials and adaptive actions in the study's manipulations were simulated in a controlled studio environment where response tasks could be executed while seated. Similarly, a small number of different types of broadband noise were employed as acoustic maskers, as opposed to a less generalizable set of real-world noise environments (e.g., urban traffic, factory floor, busy building lobby, stadium crowd, etc.). The speaking materials used in the study were developed from a corpus of public radio commentaries and converted to "robot" speech with a commercial TTS engine [17]. To avoid making the use of different voices, an additional experimental factor (see, e.g., [18] and [19]), a single "standard" synthetic male voice, was used throughout.

A. Study Methods

Five female and nine male listeners, all claiming to have normal hearing, volunteered for the study, which employed a within-subject design. The timing and presentation of all sounds and response materials were coordinated by software coded in Java. The auditory materials (synthetic speech and episodes of masking noise) were rendered with three powered studio loudspeakers placed directly left, right, and in front of the listener, at a distance of approximately 1.32 m. The response tasks were visually displayed on a 0.61-m (diagonal) flat-panel screen, and all sound was limited to a maximum of 85-dB sound pressure level.

1) Listening Materials and Experimental Manipulations: The speech materials, which are ten short commentaries on topics of general interest, were equated for length and randomly assigned to three training sessions and to seven formal listening exercises making up the main body of the experiment. The commentaries for the training sessions were each further reduced to about a minute of continuous speech, and these materials were used to allow participants to become familiar with the listening and response tasks. The main exercises lasted between 2.5 and 3.5 min, depending on the particular manipulation (see hereinafter), and a check of these commentaries for uniformity showed no significant differences across a number of lexical parameters (number of sentences, words, and syllables, etc.).

Since most real-world noise environments have variable characteristics that make their effectiveness as maskers difficult to systematically control, the study employed broadband noise to simulate different types of speech-masking events. Systematic episodes of brown noise were used as maskers in two of the training sessions. In each of the main exercises (except the Baseline condition), speech was masked either by pink noise, white noise, or "Fastl" noise (white noise modified to simulate the average spectral distribution and fluctuating temporal envelope of speech [20]). Four kinds of masking events—two "short" (5 s) and two "long" (30 s)—were defined. These pairs each included a "quiet" (-26 dB) event and a "loud" (-19 dB) event, with linear onset and offset ramps lasting 0.51 s for short episodes and 7.56 s for long ones. Only one type of broadband noise was used in each of the main exercises with masking events, with each of the four kinds of events occurring twice in random order.

a) Design: The study combined a Baseline listening exercise with a two factor, 2×3 repeated-measure design, presented in counterbalanced order. The first factor (two levels) manipulated the use and nonuse of the system's adaptive speech behaviors in the presence of noise, and the second factor (three levels) examined the use of pink, white, or Fastl noise. 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. All spoken material was rendered by the loudspeaker in front of the listener, and noise events were rendered by the loudspeakers on the listener's left and right. Coded designations (which are used in the remainder of this paper) and a summary of the seven listening exercises in the main part of the experiment are given in Table III.

b) Planned comparisons: The study's seven conditions were motivated by a set of anticipated outcomes. Baseline measures of listening performance were expected to be best in the study but not optimal due to the use of synthetic speech. The listening performance in the Nonadaptive conditions (NA-white, NA-pink, and NA-Fastl) was expected to be the

TABLE III
SUMMARY OF THE SEVEN EXPERIMENTAL CONDITIONS AND THEIR CODED DESIGNATIONS. PARTICIPANTS HEARD ALL SEVEN CONDITIONS IN COUNTERBALANCED 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

lowest. The performance in the 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 Nonadaptive conditions. It was unclear how the broadband noise manipulations would affect performance, however, because the auditory-perspective-taking system only takes relative loudness into account and makes no distinction between noise types. Each class of noise in the study differs in key ways from the other but should all be good maskers of speech. Accordingly, planned contrasts are used in the following to explore how the performance in the two presentation strategy manipulations differs from the performance in the Baseline condition across the three types of noise.

c) Adaptive auditory display behaviors: To emulate what the auditory-perspective-taking system does to ensure that its speech can be heard by its listener, the commentaries in the Adaptive auditory display conditions—A-white, A-pink, and A-Fastl—were modified as follows. Each was aligned with the eight randomly ordered noise events in its manipulation, and its amplitude envelope was modulated appropriately to compete in parallel with the maskers. The resulting envelope modifications were then shifted in time to simulate the delay that it takes for the onset of a noise event to cross the system's response threshold—3.0 and 1.0 s for long and short "quiet" events, and 2.0 and 0.8 s for long and short "loud" events, respectively (the latter with correspondingly steeper onset ramps). Next, periods of silence corresponding to the system's pause response for loud maskers were inserted, thus lengthening the commentary's running time. During short episodes of loud noise, pauses were placed at the first word boundary following 1.2 s of the loudness response and, during long episodes, at or just beyond the 5.0-s mark. The commentaries were resumed, where each loud noise event drops below the pause threshold by reuttering the interrupted sentence or phrase or, in the case of long pauses, resuming first with the words, "As I was saying…" The idea of resuming interrupted synthetic speech in this latter manner arose during the development of the prototype and was found to be consistent with listeners' intuitions about long verbal pauses in piloting for the study. A schematic of the auditory display's four adaptive behaviors showing level changes and pauses is given in Fig. 7. (Edited examples of the sound materials used in the study are available from the authors.)

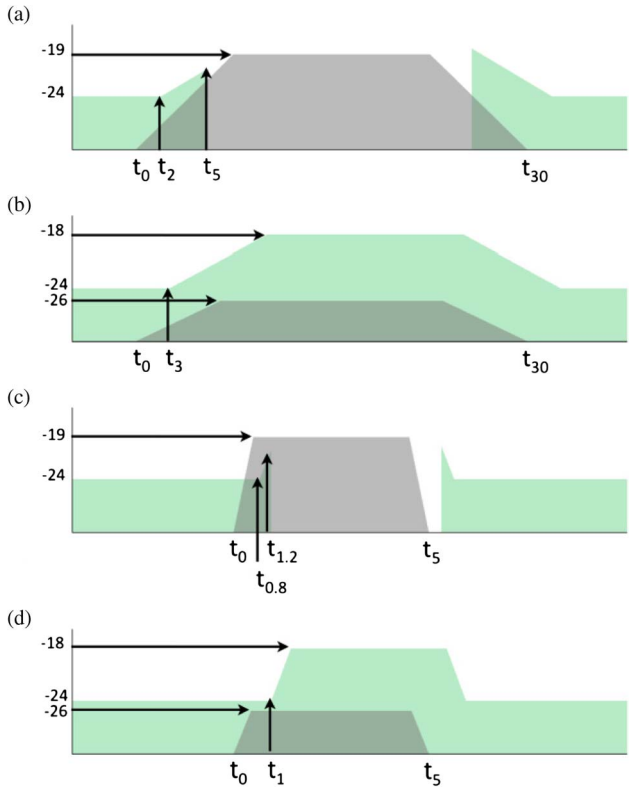

Fig. 7. Schematic diagrams showing actions taken by the auditory display in the experiment's Adaptive conditions to counter noise events with the potential to mask speech from the listener's perspective: (a) Long–loud, (b) long–quiet, (c) short–loud, and d) short–quiet. Time in seconds is shown on the horizontal axis (note differences in scale for long and short events), and level in decibels 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) Response Tasks and Dependent Measures: Participants were asked to carry out two response tasks: one while listening and the other immediately after. They were also asked to rate their preference for the way the synthetic speech was presented after each exercise on a seven-point scale.

The response task while listening involved monitoring for noun phrases in the spoken material and marking them off in an onscreen list that contained both the targeted phrases and an equal number of foils from commentaries on similar but not identical topics. Targets were listed in the order that they occurred and were randomly interleaved with no more than three intervening foils. Listeners rarely mistook foils for utterances in any of the commentaries, regardless of their ability to verify targets, and because of the extremely low incidence of false alarms (a total of 4 out of 1960 possible correct rejections), performance in this task was measured only as the percentage correctly identified target noun phrases. In the results and discussion sections, this measure is referred to as $p(\mathrm{targets})$.

In the after-listening 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 [21]. "Old" sentences were either *original* word-for-word transcriptions or semantically equivalent

TABLE IV
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 DO NOT 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

paraphrases of commentary sentences. "New" sentences were either "*18istracters*"—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 IV. In addition to responding "old" or "new," participants could also demur (object to either designation) by responding, "I do not know." Eight sentences (two of each of the old and new sentence types) were presented after each of the main listening exercises. Two measures from this response task in each condition are reported as follows: $p(\text{sentences})$ is the proportion of sentences correctly judged as old or new, and $p(\text{demurs})$ is the proportion of "I do not know" responses.

B. Study Results

The performance measures for both response tasks were mostly consistent with the anticipated pattern of listening performance. Listeners' abilities to recognize targeted noun phrases $p(\text{targets})$ and judge sentences as old or new $p(\text{sentences})$ were both highest in the Baseline condition and lowest in the Nonadaptive (NA) conditions. Moreover, scores for the target phrase task were only slightly lower than Baseline in the three Adaptive (A) conditions, as predicted. Scores for the sentence task in the Adaptive conditions, though, were not as high as predicted. However, the correlation between $p(\text{targets})$ and $p(\text{sentences})$ is significant [Pearson's $r = 0.573$, $p = 0.05$ (two-tailed)]. Plots of the mean proportions of correctly identified target noun phrases $p(\text{targets})$ and sentences correctly judged as "old" or "new" $p(\text{sentences})$ in all seven conditions are shown in Fig. 8(a) and (b), respectively.

To determine the respective effects of presentation strategy and noise type, a two-by-three repeated-measure analysis of variance (ANOVA) for each of the dependent performance measures was performed on the manipulations involving noise events. Listeners were significantly better at the target and sentence tasks in the Adaptive presentation conditions (versus the Nonadaptive conditions)—respectively, $F(1, 13) = 190.7$, $p < 0.001$ and $F(1, 13) = 5.077$, $p = 0.042$—but there was no main effect for noise type. A significant interaction between the two factors was also found in the analysis for $p(\text{targets})$ ($F(2, 12) = 8.306$, $p = 0.005$).

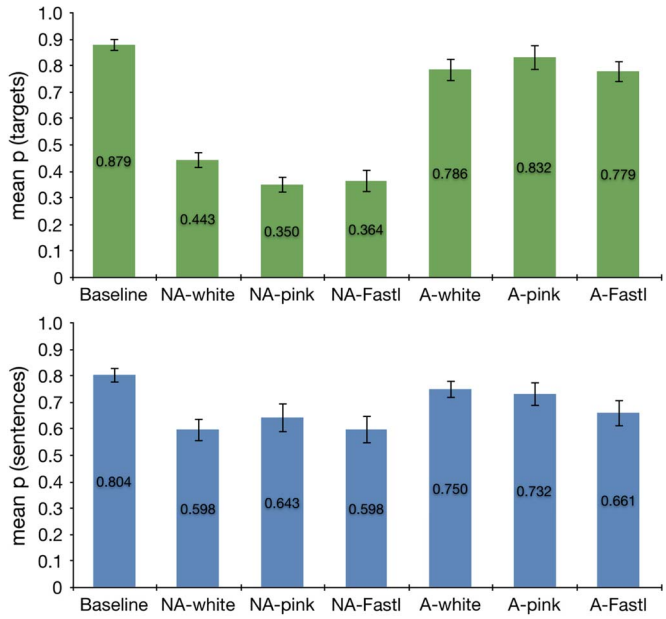

Fig. 8. (a) Plot of the mean proportion of correctly identified target noun phrases $p(\text{targets})$ in each condition. (b) Plot of the mean proportion of sentences correctly judged as "old" or "new" $p(\text{sentences})$ in each condition. The y-axis in both plots shows proportion. The error bars show the standard error of the mean.

TABLE V
F STATISTICS FOR THE PLANNED CONTRASTS BETWEEN THE BASELINE AND ADAPTIVE CONDITIONS FOR THE $p(\text{targets})$ AND $p(\text{sentences})$ 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	$F(1, 13) = 5.438$, $p = 0.036$*

Planned contrasts with performance in the Baseline condition were also used to evaluate how each type of noise impacted the dependent measures. As expected, all types of noise were significant maskers of speech in the respective nonadaptive manipulations. More interestingly, this was also the case in some of the Adaptive conditions, meaning that, while the Adaptive presentation strategy helped listeners hear significantly better than they could in the nonadaptive manipulations, it was not quite as good as listening in the absence of noise. In particular, Fastl noise impacted both performance measures in the Adaptive manipulations, and white noise impacted $p(\text{targets})$. Table V summarizes the results of the planned comparisons in the Adaptive conditions.

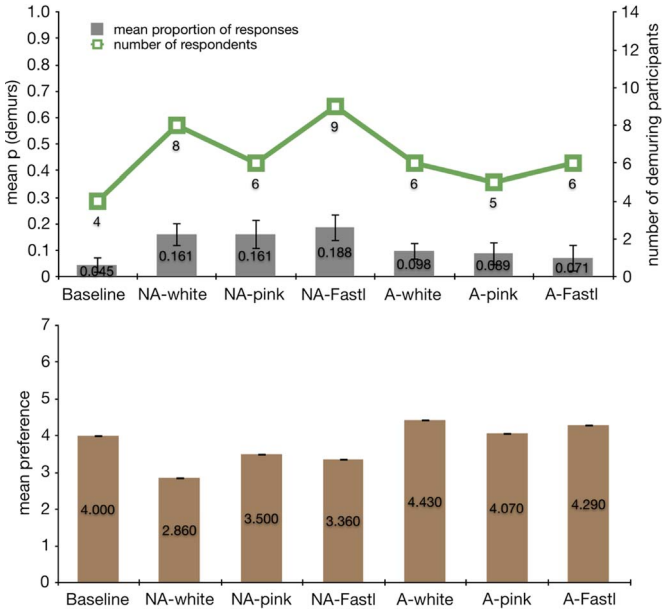

Fig. 9. (a) Plot of the mean proportion of "I do not know" responses $p(\text{demurs})$ (the gray columns correspond to the y-axis on the left) in the sentence judgment task and the number of participants in each condition choosing to demur one or more times (the squares correspond to the y-axis on the right). (b) Plot shows 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." The error bars in both plots show the standard error of the mean.

An ANOVA of the six non-Baseline conditions and planned contrasts with the Baseline were also performed for $p(\text{demurs})$—the proportion of "I do not know" responses that participants made in each condition—and a plot of this measure in each of the seven conditions is shown in Fig. 9(a) (respective counts of listeners who demurred one or more times are also shown). The trends in these data are consistent with the anticipated results, but no main effects were found in the two-by-three analysis; moreover, only the contrast between the NA-Fastl and Baseline conditions was significant. Finally, the mean subjective preference for the way that the synthetic speech was presented in each exercise is plotted in Fig. 9(b). As mentioned earlier, a seven-point scale was used, and interestingly, preference for the Adaptive presentations was slightly greater than preference for the Baseline condition. Planned contrasts with the Baseline were not significant, but an ANOVA of the non-Baseline manipulations showed a significant preference for the Adaptive presentation strategy ($F(1, 13) = 10.538$, $p = 0.006$).

C. Study Discussion

Collectively, the results of the study provide significant empirical evidence of the utility of simulated auditory perspective taking, represented here by 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 markedly better than unaided listening in the presence of noise. In addition, when asked, listeners indicated a significant subjective preference for the adaptive style of synthetic speech over the nonadaptive style.

This outcome has several important implications for the design of auditory interfaces for robots and, more generally, for Adaptive auditory display research. First, it is worth noting that the performance in the noise-free Baseline condition—$p(\text{targets}) = 0.88$, $p(\text{sentences}) = 0.80$—was poorer than might be expected and is likely a consequence of the use of a synthetic voice. Brock *et al.* observed better scores ($p(\text{targets}) = 0.91$, $p(\text{sentences}) = 0.87$) for unmanipulated individual human speech in a separate, but somewhat similar, experiment with a longer listening task [22], and even larger performance differences have been found in other studies with other paradigms, e.g., [18] and [19]. This performance disparity and the fact that there are no alternatives to synthetic speech for conversational purposes in robotic systems are further motivation for accommodating users' listening requirements in noisy settings. Second, the different impacts of each type of noise used in the study suggest that additional techniques may be needed to effectively accommodate the user's auditory perspective in certain kinds of environments. While pink, or more broadly, "$1/f$" noise, which occurs widely in nature, was the most successfully adapted for in the study, white and particularly Fastl noise events clearly hampered listening in spite of the system's adaptive strategies. The pronounced impact of Fastl noise, particularly with its speechlike properties, suggests that machine categorization of the type of ambient noise present in an auditory could be used to augment the adaptive techniques explored here. Third, in light of the listeners' significant subjective preference for the way that synthetic speech was presented in the study's Adaptive manipulations, it is essential to underscore that each of the ways that the robot's auditory display is designed to respond to noise onsets is modeled on a human solution. Participants in conversation generally try to be aware of, and act on, their addressees' listening needs transparently and in ways that meet each other's expectations. Thus, simulating this type of perspective taking in auditory interaction design for robotic systems, as well as others, has important collaborative utility and merits further development. Fourth, the performance improvements associated with speaking louder to overcome low levels of masking noise suggest that the same type of adaptation will successfully extend to situations in which the proximity between the robot and its user is likely to vary with any frequency. This capability has already been implemented but has not been formally tested.

V. Conclusion

The notion that robots will eventually assume collaborative roles involving aural interactions in social settings has already materialized in the form of self-serve checkout 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 that users will expect robots and other systems to be capable of, such as gaze following, contextual awareness, and implied goal recognition.

With these insights in mind, this paper has presented a prototype computational auditory-perspective-taking scheme for a mobile robot. It monitors the auditory scene through multimodal sensing, tracking human user positions, listening for speech, and localizing noise sources to estimate the intelligibility of its own auditory presentation. As necessary, the robot can then use that knowledge to alter the level and/or progress of its speech or move to a quieter location to accommodate its user's listening needs.

Evaluation of our perspective-taking interface has focused on three different core abilities: speech detection and localization, estimating noise volume, and modifying speech output in response to external stimuli. Section III demonstrated the accuracy of the first two, correctly identifying speech and its incident angles acoustically under differing noise conditions, building maps of the environment, and using those maps to relocate the robot to areas with lower ambient noise. Hand-collected measurements verified the substantial improvement, even over alternative relocation strategies. Section IV then demonstrated in a human study the importance of adapting a speech interface to external noise, through volume adjustments and effective pausing. When noise levels were moderate, and did not require relocation, users both preferred an adaptive interface and showed enhanced listening performance to a nonadaptive speech presentation.

In conclusion, the success of the Adaptive auditory display strategies evaluated confirms the importance of this emerging direction in user interface design, and the developed prototype demonstrates the wide range of skills available to a perspective-taking robotic interface. Future auditory-perspective-taking research, however, has significant room for growth. In particular, adaptive behavior by a perspective-taking system should incorporate knowledge or inference of users' privacy concerns and other socially motivated considerations. Also, enunciation and, more ambitiously, conversational repair can be explored as techniques for countering informational masking. Finally, there is a range of nonspeech applications for robot auditory interfaces, such as aural monitoring and playback and sonification of process or sensor data that are even more challenging than speech displays to use in the presence of ambient noise. Nevertheless, while adaptive presentation strategies for nonspeech sound information in noisy settings will likely require approaches that are somewhat different from the techniques evaluated here, it remains important to keep the listener's perspective in mind when developing an auditory display.

ACKNOWLEDGMENT

The authors would like to thank H. Fouad and B. McClimens for the technical help with the production of audio files for this research.

REFERENCES

[1] S. C. Peres, V. Best, D. Brock, B. Shinn-Cunningham, C. Frauenberger, T. Hermann, J. G. Neuhoff, L. V. Nickerson, and T. Stockman, "Auditory interfaces," in *HCI Beyond the GUI*, P. Kortum, Ed. San Francisco, CA: Morgan Kaufman, 2008, pp. 145–195.

[2] E. Martinson and D. Brock, "Improving human–robot interaction through adaptation to the auditory scene," in *Proc. ACM/IEEE Int. Conf. Human–Robot Interact.*, Arlington, VA, 2007, pp. 113–120.

[3] D. Sofge, M. Bugajska, J. G. Trafton, D. Perzanowski, S. Thomas, M. Skubic, S. Blisard, N. Cassimatis, D. Brock, W. Adams, and A. Schultz, "Collaborating with humanoid robots in space," *Int. J. Humanoid Robot.*, vol. 2, no. 2, pp. 181–201, 2005.

[4] A. Schultz and W. Adams, "Continuous localization using evidence grids," in *Proc. IEEE ICRA*, Leuven, Belgium, 1998, pp. 2833–2839.

[5] J. M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, and H. G. Okuno, "Robust recognition of simultaneous speech by a mobile robot," *IEEE Trans. Robot.*, vol. 23, no. 4, pp. 742–752, Aug. 2007.

[6] G. Bradski and A. P. Kaehler, "Learning-based computer vision with Intel's open source computer vision library," *Intel Technol. J.*, vol. 9, no. 2, pp. 119–130, May 2005.

[7] B. Fransen, V. Morariu, E. Martinson, S. Blisard, M. Marge, S. Thomas, A. Schultz, and D. Perzanowski, "Using vision, acoustics, and natural language for disambiguation," in *Proc. ACM/IEEE Int. Conf. Human–Robot Interact.*, Arlington, VA, 2007, pp. 73–80.

[8] T. Quatiri, *Discrete Time Speech Signal Processing*. Delhi, India: Pearson Education Inc., 2002.

[9] E. Martinson and A. Schultz, "Discovery of sound sources by an autonomous mobile robot," *Autonom. Robots*, vol. 27, no. 3, pp. 221–237, Oct. 2009.

[10] D. O. Elorza, "Room acoustics modeling using the ray-tracing method: implementation and evaluation," M.S. thesis, Dept. of Physics, Univ. Turku, Turku, Finland, 2005.

[11] E. Martinson, "Simulating robotic auditory environments," in *Proc. Int. Conf. Intell. Autonom. Syst.*, Ottawa, ON, Canada, 2010.

[12] D. Raichel, *The Science and Applications of Acoustics*. New York: Springer-Verlag, 2000.

[13] J. C Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Amer.*, vol. 93, no. 1, pp. 510–524, Jan. 1993.

[14] B. Langner and A. Black, "Using speech in noise to improve understandability for elderly listeners," in *Proc. ASRU*, San Juan, Puerto Rico, 2005, pp. 392–396.

[15] J. H. Rindel, "The use of computer modeling in room acoustics," *J. Vibroeng.*, vol. 3, no. 4, pp. 219–224, 2000.

[16] D. Brock, B. McClimens, C. Wasylyshyn, J. G. Trafton, and M. McCurry, "Evaluating the utility of auditory perspective-taking in robot speech presentations," in *Lecture Notes in Computer Science*, vol. 5954/2010. Berlin, Germany: Springer-Verlag, 2010.

[17] Cepstral, Sep. 2011. [Online]. Available: http://cepstral.com

[18] J. B. Hardee and C. B. Mayhorn, "Reexamining synthetic speech: Intelligibility and the effect of age, task, and speech type on recall," in *Proc. Human Factors Ergonom. Soc. 51st Annu. Meet.*, Baltimore, MD, 2007, pp. 1143–1147.

[19] C. Stevens, N. Lees, J. Vonwiller, and D. Burnham, "Online experimental methods to evaluate text-to-speech (TTS) synthesis: Effects of voice gender and signal quality on intelligibility, naturalness, and preference," *Comput. Speech Lang.*, vol. 19, no. 2, pp. 129–146, Apr. 2005.

[20] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, 3rd ed. Berlin, Germany: Springer-Verlag, 2007.

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

[22] 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 *Proc. 14th ICAD*, Paris, France, 2008.

Eric Martinson received the Ph.D. degree in computer science from the Georgia Institute of Technology, Atlanta, in 2007.

Following his graduation, he was awarded a Fulbright Fellowship to the Kharkov National University of Radio-Electronics, Kharkiv, Ukraine, where he expanded upon his dissertation work in robotic auditory perception. He was also engaged in research in learning by demonstration while employed by HRL Laboratories and General Motors. He is currently a Postdoctoral Fellow with the Navy Center for Applied Research in Artificial Intelligence, U.S. Naval Research Laboratory, Washington, DC, where he explores multimodal perception as applied to humanoid and mobile robots. His research interests include autonomous robotics, human–robot interaction, and multimodal perception.

Derek Brock received the M.S. degree in computer graphics and multimedia systems from The George Washington University, Washington, DC.

Since 1991, he has been with the Navy Center for Applied Research in Artificial Intelligence, U.S. Naval Research Laboratory, Washington, where he is currently a Computer Scientist who has specialized in human–computer interaction research. His current interests include the application of sound and models of human language use to the design of standard and novel user interfaces for computational systems.

Mr. Brock is a member of the International Community for Auditory Display.