Improving Human-Robot Interaction through Adaptation to the Auditory Scene

Eric Martinson  
Georgia Institute of Technology, Atlanta, Ga  
Naval Research Laboratory, Washington D.C.  
ebeowulf@cc.gatech.edu

Derek Brock  
Naval Research Laboratory, Washington D.C.  
brock@itd.nrl.navy.mil

ABSTRACT
Effective communication with a mobile robot using speech is a difficult problem even when you can control the auditory scene. Robot ego-noise, echoes, and human interference are all common sources of decreased intelligibility. In real-world environments, however, these common problems are supplemented with many different types of background noise sources. For instance, military scenarios might be punctuated by high decibel plane noise and bursts from weaponry that mask parts of the speech output from the robot. Even in non-military settings, however, fans, computers, alarms, and transportation noise can cause enough interference that they might render a traditional speech interface unintelligible. In this work, we seek to overcome these problems by applying robotic advantages of sensing and mobility to a text-to-speech interface. Using perspective taking skills to predict how the human user is being affected by new sounds, a robot can adjust its speaking patterns and/or reposition itself within the environment to limit the negative impact on intelligibility, making a speech interface easier to use.

Categories and Subject Descriptors
H.5.2 [Information Interfaces and Presentation]: multimedia information systems, sound and music computing, user interfaces

General Terms
Performance, Human Factors.

Keywords
Acoustics, Auditory Scene, Auditory Perspective Taking, Human-Robot Interaction

1. INTRODUCTION
Speech is a highly effective communication medium through which people can express a wide range of concepts. Ideas ranging from the very abstract to the smallest level detail can be communicated to another person with a high degree of comprehension and reliability. Perhaps most impressively, people can maintain these high comprehension levels in their conversations over a wide range of auditory scenes. Within a single auditory scene filled with many types of noise sources (including other human conversations, cars, telephones, and machinery hums), people can still communicate, adapting their conversation to compensate for the interference. The ways in which people compensate generally fall into three categories. The first category of adaptation is adding more redundancy to their speech output. By watching the listener for comprehension and asking questions, people can judge when more information is needed and adapt their word choice to provide greater redundancy under poor acoustic conditions. The second category of adaptation is to improve the intelligibility of the speech waveform [1]. People reflexively adapt their own speech in the presence of noise or stress, changing the shape and tightness of their vocal tract to produce a different volume, prosody, pitch, and/or timbre of their speech. The resulting signal is more intelligible to the human auditory system than an unmodified signal under the same noise conditions. The final category of adaptation comes from adjusting the position of the talker relative to the listener and/or the noise sources. People may employ gestures, move closer to the listener, face the listener as much as possible, and if all else fails, move to someplace else where there is less interference. In the presence of noise, all of these adaptations serve to improve intelligibility and make reliable speech communication possible. For a robotic interface using speech, this same adaptive behavior in the presence of noise is also necessary for effective communication with a human partner.

In this work, it is our supposition that successfully applying these types of adaptations to a speech interface requires perspective taking in the auditory domain [2]. In its broadest sense, perspective taking is the ability to construe comprehension and perception from a non-egocentric frame of point of view. By focusing on perspective taking in the auditory domain, we seek to allow a robot to use its knowledge of the environment, both a priori and sensed, to predict what its human counterpart can hear and effectively understand. Equipped with this knowledge, the robot can make decisions about how best to change its auditory presentation behavior to improve intelligibility and ease of interaction in the presence of adverse auditory conditions. Implementing the latter two categories of auditory adaptations (altered speech characteristics and interface movement) with a text-to-speech interface, our robot gathers information about the auditory scene, and then uses knowledge of ambient noise conditions and the position of the human listener to: (1) dynamically adapt the volume of its speech output in response to noise levels and the distance of the listener; (2) insert pauses into
the speech signal to avoid extraordinary interference by ambient noise sources; (3) rotate to always face the user; and (4) reposition itself when the noise interference becomes too large.

The remainder of this paper is broken up as follows. Section 2 covers related work in the areas of perspective taking and adaptive auditory displays. Section 3 describes the algorithmic tools that are used to find and combine knowledge about the auditory scene. Section 4 describes the Mobile Information Kiosk task under which this work is implemented. Finally, Section 5 describes results to date as well as ongoing work to evaluate the interface.

2. RELATED WORK

This work on auditory perspective taking grows out of two related fields of research: perspective taking by a robotic interface, and adaptive auditory interfaces. The first area, perspective taking, is focused on providing robots with the ability to reason about non-ego-centric frames of reference in order to improve human-robot collaboration, such as with NASA's Robonaut [3]. In a case study of astronaut extravehicular training, Trafton et al [4], found that over 85% of astronaut utterances involving spatial frames of reference were either object-centered, exocentric, addressee-centered, or deictic, as opposed to ego-centric. In roughly 25% of these interactions it was necessary for collaborators to take each other's visual perspective, with speakers changing perspective roughly once every other utterance. Motivated by these and other findings, a comprehensive architecture for human-robot interaction has been developed at the Naval Research Laboratory that facilitates the use of cognitive models to reason about non-ego-centric perspectives in a range of circumstances [4, 5, 6].

In this work, we are augmenting to the notion of robotic perspective taking by working with an adaptive auditory interface. Up till now, work in this area has remained fairly limited, as such research requires a large number of technical solutions to achieve a comprehensive, adaptive auditory interface, such as intelligent prioritization of auditory information, the use of machine listening and computational auditory scene analysis [7] for real time auditory domain monitoring [8]. However, one particular area of adaptive audition that has received significant attention is “clear speech” or “Lombard speech” [1]. Better known in the speech recognition community, Lombard speech describes the change in people’s voices in response to stress, as well as subconscious adaptations to the auditory scene. By altering the properties of their vocal tracts, people can actually improve the intelligibility of their speech for other human listeners [9]. Unfortunately, the same cannot be said yet for computerized speech. Although attempts have been made to duplicate the Lombard effect in speech synthesis systems for improved intelligibility under noisy conditions, efforts at the waveform generation level have so far demonstrated only mixed success [9, 10].

To tackle the same adaptive speech problem from a different direction, we are focusing on applying robotic advantages of sensing and mobility to the domain. Applying perspective taking skills to combine knowledge about the listener with knowledge about the environment, the robot makes decisions about when and where to speak to achieve the best intelligibility, rather than simply improving the signal quality at the source.

3. MONITORING THE AUDITORY SCENE

Taking the perspective of a human listener first requires some knowledge about the shape of the auditory scene. Given some a priori knowledge, the experiential knowledge gathered by the robot, and the local auditory scene currently detected by the robot, what kind of predictions can the robot make about what the human can or can not hear? In this section, we discuss the tools and information available to the robot that can allow it to make decisions about the intelligibility of what it is or will be saying to a human listener. We first cover the sensors on the robot itself, and then address speech detection, speech localization, sound source localization, and mapping the auditory scene.

3.1 Robot

The hardware used for this work is a B21R (Figure 1) robot equipped with:

- An overhead microphone array for ambient noise monitoring. This array is composed of four Audio-Technica AT831b lavaliere microphones mounted at the top of the robot. These microphones are each connected to battery-powered preamps mounted inside the robot body and then to an 8-Channel PCMCIA data acquisition board.
- A flat-panel monitor mounted at eye-level to display topics the robot can talk about, and list the speech commands the robot can understand.
- A loudspeaker and internal amplifier to allow the robot to speak at a variety of volumes to a human listener.
- A stereo-vision system for person tracking.

Figure 1. B21R robot from iRobot, outfitted with a four microphone overhead array, bi-clops stereo vision system, and monitor for visual feedback.
A SICK laser measurement system (LMS200) used with continuous localization [11] to provide reliable robot pose (position and orientation) estimates.

Currently, in addition to the above hardware we are also using a separate wireless microphone headset to capture speech for speech recognition tasks and freely available speech recognition software (Microsoft SAPI 5.1). For now, this separate microphone is necessary to get reliable speech recognition results. Although a microphone array can often improve speech recognition [12], our overhead microphone array is not appropriate for speech recognition tasks. In future implementations, however, the intent is to replace the wireless microphone with a directional microphone mounted on the robot body. Combined with other efforts by the robot to always face the user, a directional microphone should provide reasonable speech recognition results with a minimum of additional effort by a user.

3.2 Visual Person Tracking

The vision system on the robot is an actuated TRACLabs BiClops. The rotatable stereo camera provides dual color images from which depth information (Figure 2, top) can be extracted. Combined with face detection software (created using OpenCV [13]), the robot can use the camera to track, localize, and follow a detected person through a 180 degree arc in front of the robot (Figure 2, bottom). To start tracking a person, the individual’s face needs to be at least 20 pixels in width, which corresponds to a distance of roughly 1.5m from the robot. After initializing a track, however, the camera can continue to provide depth information up to 3 meters away from the robot.

3.3 Speech Detection

Before the cameras can be used to find a person, however, the robot needs to first identify and localize on speech sounds in the environment. To detect speech sounds, we calculate the first 2 mel-cepstrum coefficients [14] for each microphone in the overhead array. Each coefficient is averaged across all microphones, and then compared to an environment dependent threshold. While this speech detection system is relatively simple, and prone to errors when classifying a single sound sample, it works well over time to augment other auditory and vision sensors tracking humans in the environment.

3.4 Speech Localization

Once speech has been detected in the environment, the robot also needs to know from where the speech sounds are originating. Even if there is only one person in the room, unless that person is standing directly in front of the robot, the robot cannot know the individual is there using just the stereo vision system. But audition can be used to identify where the detected speech is likely to be coming from and allow the robot to reorient itself to face the speaker. If there are multiple people, then audition can also be used to disambiguate the vision results. Which of the multiple people in the room was actually talking? There are multiple methods by which this can be accomplished [12,15]. For this task we used spatial likelihoods [15].

Spatial likelihoods are based on the principle of time difference on arrival. As the speed of sound can be assumed constant, and the microphones are physically separated in space, the signal received by each microphone from a single source will be offset by some measurable time. If the value of these offsets can be determined, then the location of the sound source will be constrained to all positions in the room whose geometry relative to the array corresponds to the measured time differences. Spatial likelihoods are then a maximum likelihood approach utilizing these time differences to estimate the likelihood associated with every possible location in the room. Figure 3 shows the spatial likelihood output of a sample containing speech, plotted on a contour plot.

In theory, given enough microphones in an array, it should be possible to exactly localize upon the source of a speech signal.
3.5 Sound Source Localization

With a human speaker, the speech sounds the robot needs to localize on are sporadic and short in duration. As a result, the robot needs to make a best guess on localization using the few samples it can get. With longer duration sources, however, there is a robotic solution to overcoming the localization error. When a sound source can be relied on to keep generating sound for a few minutes, then the robot can move about the environment to collect measurements from different angles towards the source. Combined together, those measurements effectively triangulate upon the sound source, providing 2 or 3 dimensional coordinates relative to the robot’s position. The approach we use for combining these measurements is Auditory Evidence Grids [16]. Auditory evidence grids use the spatial likelihood measurements collected over time, along with an estimate of the change in robot pose (i.e., position and orientation), to determine the likelihood of a sound source being located at any given location within the room. As the robot moves through the environment, other noises such as robot ego-noise (motors, wheels, fans, etc.) and echoes from the surrounding walls are effectively filtered out, and only the evidence indicating real sources will appear to remain stationary over time (Figure 4). In practice, a robot can localize a single source using an auditory evidence grid with samples taken from as few as 2 positions, provided the positions are separated enough in space. However, with more data, auditory evidence grids can easily be used to localize more than one source. Therefore, prior to running the information kiosk experiment described below, the robot should first wander around the environment to localize the set of loudest currently active sources.

3.6 Noise Mapping

Once a set of sources has been identified, the way to estimate its combined effect on the auditory scene is through Noise Maps. In general, noise maps are a graphical tool commonly employed by acousticians to plot the average levels of noise found throughout an environment. They can be measured by hand, measured autonomously by a robot, estimated from models of ideal sources (Figure 5), or created through a combination of these techniques [17]. In this work, we will be modeling sound sources as ideal
point sources in an open environment. Although the resulting ideal source map has inconsistencies with the sampled real world data, particularly near walls, it is quick to create in real time, and is still useful for noise avoidance and repositioning.

To create the map, we model each source as an ideal point source, where the amplitude of the sound wave decreases linearly with the distance from the source. After localizing a source, the robot collects 5 seconds worth of data to estimate volume at a known distance from the source, and then uses that to estimate the effect of that sound source on the remainder of the environment. Figure 5 (top) shows an example of the spherical spreading due to a single source. As additional sources are discovered (Figure 5, middle) their effects can be added to the environment using linear unweighted summation. The result (Figure 5, bottom) is an estimate of the effect of all known sound sources on the environment, illustrating zones of quiet and loud volume.

4. THE INFORMATION KIOSK

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 touchscreen 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 current 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 to read aloud the pre-compiled story matched to that title. While the robot is reading, it takes the perspective of the human listener by predicting what is intelligible to the human listener, and then adjusts its spoken output or alters its position to maximize intelligibility and ease of use.

Using the tools discussed in the previous section, there are four general actions the robot can take in response to predicted intelligibility deficiencies due to the dynamic auditory scene: (1) the robot can rotate to face the listener, maintaining the interaction and orienting its loudspeaker in the correct direction; (2) the robot can adapt its volume before reading each sentence, maintaining a steady volume at the listener's location; (3) the robot can pause for speech and excessive noise that interrupt its reading and distract the listener; and (4) the robot can move to another location when sound levels stay too high in its current location. Figure 6 shows the general finite state machine used for selecting each of the latter three actions in response to a changing auditory scene. Note that rotating to face the listener is not listed in this diagram. Since the stereo vision system rather than the auditory array is used to maintain the robot's orientation once a person has initially been detected, the process of rotating to face the listener is run in parallel with all other actions.

4.1 Rotating To Face the Listener

A person arriving at the information kiosk might approach from any angle. Although the robot ultimately uses the vision system to track its listener, it first waits for the person to say something and uses the speech detection and localization tools discussed earlier to determine the direction it should turn to face. Then the vision system is initialized and the biclops camera takes over the job of continuous tracking. As the camera is actuated, it can rotate independent of the robot body to follow the person through arcs of up to 90 degrees in each direction. However, for intelligibility and ease of use, it is best to restrict this range to 30 degrees or less in each direction, and rotate the robot body when the person moves too far to one side or another.

The purpose of rotating the robot is twofold. First, it promotes ease of use because it frees the listener of the need to remain in place while interacting with the kiosk and it places the flat-panel monitor, which displays information topics and speech commands for using the system, in front of the user. An example of this visual interface featuring two kinds of Navy ships the robot can talk about is shown in Figure 7. Other interfaces featuring current news briefs, biographies of interesting people, and NRL robotics laboratory projects have also been developed.

The second purpose of rotating the robot is to maintain the desired intelligibility levels. The loudspeaker on the robot is not omnidirectional, meaning that its apparent volume changes with the angle of the perceiver. Consequently a person standing to the side of the robot will not hear its speech output as well as a person standing directly in front of it. By not allowing the listener to stand too far to either side of the robot (i.e., by rotating the robot to face the listener after more than 30 degrees of angular displacement), the system minimizes the effects of loudspeaker directionality on volume levels and general intelligibility at the listener's location.

4.2 Changing the Volume

After the user selects a topic the robot reads a corresponding paragraph or two of information aloud, sentence by sentence.
Before each sentence, the robot measures the current level of ambient noise in the room, and the distance at which the listener is standing to estimate an ideal volume at which to speak in order to maintain the desired intelligibility levels. The louder the ambient noise in the environment, the louder the robot needs to speak. Similarly, the farther away the listener stands, the louder the robot needs to speak. Conversely, if the ambient noise volume or the listener’s distance decreases, the robot should lower its volume to avoid being excessively loud. Ambient noise levels in the room are measured by the microphone array, and the distance to the user is measured by the stereo vision system.

4.3 Pausing for Interruptions

Sometimes, auditory events in the environment will require that the robot stop reading for some period of time. In a military environment, for instance, users of speech interfaces might commonly be exposed to planes flying overhead, helicopters, and in general, loud vehicular engine noise. If such an event occurs while the robot is reading a text, then the robot speech will not be intelligible during the event, thereby losing any knowledge being transferred at that time and frustrating the listener. A robot with perspective taking abilities, however, knows the maximum volume at which it can speak, can estimate how much the listener can hear, and can then choose to pause during the period(s) of excessive noise. When ambient noise levels finally return to a reasonable level (i.e., intelligible to the listener), the robot can resume speaking. Currently, to alert the listener to the fact that it is about to continue speaking, the robot starts the next sentence by saying, “As I was saying…”

Another source of interruptions for a robot speech interface are other people. Unlike auditory events, which reduce intelligibility by masking the speech signals, other speech in the environment does not necessarily reduce intelligibility below acceptable levels for a human listener. However, if that speech is directed at the listener, then the user’s attention will be diverted from the robot and focused on the new human speaker. In this case, a robot capable of perspective taking should recognize that it is no longer the focus of attention and should pause until it regains its audience. With speech interruptions, however, the timing of the pause adaptation is not clear. Should the robot simply continue to speak when it thinks the conversation is over? What if the conversation was very long and the person no longer remembers the last thing the robot was talking about? What if the robot misclassifies the end of the conversation and starts speaking in the middle of the new conversation, thereby frustrating everyone? To satisfy these conditions, the current implementation of the information kiosk does not resume automatically from a speech interruption. On the screen, it states that it has been interrupted and is waiting for a new command. Then 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.

4.4 Relocating the Robot

The final action that a robot can take in response to a changing auditory scene is to move someplace else. For instance, if the robot is located next to an air vent when the ventilation system starts up, then it does not have to stay in that location. The same is true of other machinery, radios, and even disruptive human to human conversations. If the disruption happens while the robot is not talking to anyone, then it can move without any reservations. If, however, the robot is talking to somebody, it first asks its listener if he or she would like to move due to the noise.

The goal of the relocation process is to reduce the effect of masking noise on speech intelligibility, but there are at least two possible strategies for determining where to relocate the robot. The simplest relocation strategy is avoidance. Since loudness diminishes with distance, the robot can decrease its exposure to an source of ambient noise by moving as far away from it as possible, while remaining within a specified area. First using an auditory evidence grid to localize the source, the robot looks at an obstacle map to identify the set of reachable locations, and picks the one farthest away. The dashed line in Figure 8 shows an example of this as the robot moves all the way across the room. At this point, if there were no other sound sources, the volume would be low enough that the robot could continue speaking again.

The trouble with this farthest-distance-removed approach is that there might be additional sound sources on the other side of the room opposite the newly interfering sound source. Therefore, by simply picking the farthest location away from the present sound source, the robot may not actually be decreasing its noise exposure. Using a smarter strategy that keeps track of all existing sound sources, a robot can reduce the amount of noise interference much more reliably. The way to track all of these existing sound sources, and measure their combined effect, is through Noise Maps (Sec. 3.4). Once the robot discovers the sound source location, or makes a best guess, it adds the effects of the new sound source into its Noise Map, picks the quietest remaining location, and moves to that goal using a path planning algorithm.

5. Discussion

The application of these perspective taking skills to a working robotic system has largely 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. Some of these adaptations, however, work better than others, and in this section we address what is being done to test the validity of the system and how the system might be improved.

5.1 Speech Recognition
The speech recognition capabilities of the Information Kiosk are limited. Using a commercial speech recognition package (SAPI 5.1), the robot recognizes a small set of phrases spoken into the wireless microphone (to be replaced in the future). These phrases include titles of stories the robot can talk about plus a set of phrases for disabling options in the interface and controlling the flow of text during an interruption (Figure 7; see 4.2.3). In the future, we hope to incorporate larger linguistic grammars to provide a more natural interface for speech communication.

5.2 Robot Speech Feedback
The largest technical difficulty we encountered while implementing this system arose from the robot listening to itself talk. While humans may actually depend on this type of internal feedback to adjust their voice as necessary to produce the desired sounds, our robotic system does not have the capability to recognize its speech output as being its own. Therefore, when the robot hears itself talk, it may incorrectly pause for an interruption or raise volume. In the short term, we have overcome this problem by only sampling for loud noises between sentences (see Figure 5). Between sentences, we know that the robot will not hear itself talk, and therefore will not improperly react to its own auditory feedback. However, the drawback is that if the robot is in the middle of a sentence when a loud noise goes off, it will not stop speaking until the sentence has finished.

We believe that final solution to this problem lies in incorporating knowledge about what is being said. Ideally, the robot should know what it is supposed to be saying and be able to recognize its own speech. Short of that, however, filters could be applied to reduce the magnitude of the robot’s dominant speech frequencies, and the system could listen for loud sounds on other frequencies. This strategy, combined with sampling between sentences, should make it possible for the robot to stop in the middle of sentences.

5.3 Volume and Pause Effectiveness
To evaluate the success of our system in improving intelligibility, we are finalizing a human-robot interaction study. The study is being designed to address 2 primary questions: (1) Does the incorporation of adaptive volume and pause improve intelligibility over a non-adaptive speech interface; and (2) Do the same adaptations improve the system’s ease of use. The adaptive system will then be compared to two types of non-adaptive speech interfaces: (1) a speech interface with no control over the flow of speech; and (2) a speech interface where the user can stop the robot, and ask it for repetition. The study is to be performed on a stationary robot inside of the NRL auditory array, where a full 3D auditory scene can be generated to immerse a human listener.

Given the poor performance of other speech synthesis systems at improving intelligibility in the presence of noise, it is imperative that we determine the affects of our adaptations on intelligibility. It is our hypothesis that, as adjusting volume and adding pauses has a minimum effect on the speech waveform itself, these changes should, at least, not have a negative effect on intelligibility and, ideally, should demonstrate a positive improvement. Where we expect to see the greatest improvements, however, is in ease of use. By reducing speech output during competing auditory events, we expect to see reduced overall frustration with the speech interface as people will not have to make an effort to filter out the interfering ambient noise.

5.4 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 strategy by allowing the robot to move itself from an initially noisy location to a quieter location and then measuring the difference in noise levels. Both the strategy of moving the farthest distance away from the source, and using an informed decision making process were tested.

The auditory scene in which these strategies were tested can be seen in Figure 8. This shows a layout of the laboratory, with all of the sound sources that influenced the robot. Three of the sources are fan type sources of different volumes (52dB, 54dB, 59dB) that are always on (these three sources are shown in Figure 5, middle). For a single test, the robot would start near a single enabled music source (the other two music sources would be off), detect that source, and then move to another location using one of the two possible relocation strategies. This test was repeated five times for each of the music sources and relocation strategies, for a total of 30 runs. Table 1 shows the average improvement in the auditory scene while avoiding each of the three sources.

<table>
<thead>
<tr>
<th>Source</th>
<th>Farthest Distance</th>
<th>Informed Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source 1</td>
<td>-9dB</td>
<td>-10dB</td>
</tr>
<tr>
<td>Source 2</td>
<td>-13dB</td>
<td>-12dB</td>
</tr>
<tr>
<td>Source 3</td>
<td>-2dB</td>
<td>-5dB</td>
</tr>
</tbody>
</table>

The results demonstrate the benefit of avoiding sources and the difference in relocation strategies. The three music sources...
ranged from 60 to 65dB in average volume, varying slightly from test to test depending on the music that was currently playing. To avoid the first two music sources, both relocation strategies led to similar final positions in the quietest part of the room (the upper right in Figure 8), resulting in comparable performance improvements. The relocation strategies selected notably different paths, though, to avoid the third music source, which was located in the quiet area both strategies selected to avoid the first two music sources (again, the upper right in Figure 8). With that part of the room now filled with noise, the robot was not able to demonstrate as much of a drop in average noise levels as the other scenarios. However, the farthest distance strategy sent the robot to one of the fan sources where noise levels were only slightly less than the original position (the dashed-line path in Figure 8). The informed decision strategy resulted in a location closer to the middle of the room where the robot could still talk over the noise.

6. CONCLUSION
In conclusion, what this work has demonstrated is a set of actions a robotic speech interface can make to improve interaction with a human listener. By taking into account the affects of a dynamic auditory scene on a listener’s comprehension of its speech output, a robot can act in human inspired ways to maintain intelligibility and, ideally, improve its user’s interaction experience. With just the initial analysis of the robot’s adaptive abilities completed, we can already see large improvements in reducing noise exposure by simply relocating the robot. We expect to see similar improvements due to intelligent pausing and dynamic volume selection once our empirical study has been completed.

In the future, we hope to further extend the adaptive abilities of the robot by expanding its knowledge about the auditory scene that surrounds it. By taking into account source directivity, rather than simple locations, the robot can better predict where to move with the least amount of inconvenience to a human listener. By adding knowledge of source types, the robot can better predict how a source is likely to change in the future and so better select actions to mitigate the affect of that source on its speech presentations. Ultimately, however, the goal is to integrate this work into a full natural language system with adaptive linguistic abilities to complement the current robotic advantages.

7. ACKNOWLEDGMENTS
This research was supported by the Office of Naval Research under work order number N0001406WX30002. Thanks also to Vlad Morariu who developed the vision system used on the B21R

8. REFERENCES