Auditory Perspective Taking

Eric Martinson

U.S. Naval Research Labs, Washington, D.C.
Georgia Institute of Technology, Atlanta, Ga.
ebeowulf@cc.gatech.edu

Derek Brock

U.S. Naval Research Labs,
Washington, D.C.
brock@itd.nrl.navy.mil

ABSTRACT
Auditory perspective taking is imagining being in another's place, and predicting what they are able to hear and how it will affect their general comprehension. From this knowledge of another's auditory perspective, a conversational partner can then adapt his or her auditory output to overcome a variety of environmental challenges and insure that what is said is intelligible. In this poster presentation, we explore this concept of auditory perspective taking for a robot speech interface.

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

General Terms
Reliability, Human Factors

Keywords
Auditory Scene, Auditory Interface, Human-Robot Interaction

1. INTRODUCTION
Imagine, for a moment, the auditory scene in a mobile military command center of the future. Within this scenario, a robot is reporting the latest results of robotic surveillance efforts to a human operator who is visually monitoring a number of other systems. Consequently, the robot uses speech to convey most of its information to limit the number of times the operator needs to look away from the visual displays. While the robot is reporting, a group of helicopters flies overhead. As long as those helicopters are nearby, the human operator cannot hear what the robot is saying. What is the appropriate response by the robot in this situation? At the very least, we know the inappropriate response. Ignoring the helicopter noise and continuing to talk normally means that anything spoken aloud while the noise is present will probably not be heard.

The robot's challenges in this scenario are problems that require auditory perspective taking. In its broadest sense, perspective taking is the ability to construe comprehension and perception from a non-egocentric frame or point of view [1]. In collaborative activities, perspective taking skills greatly facilitate the effort participants must make in communicating with each other. 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 change its auditory presentation behavior accordingly. In the case of the helicopter flying overhead, for instance, the robot should be able to predict the inability of its addressee to adequately hear what it is saying and either try to talk louder or simply pause until noise levels return to normal. Either option is something that is easily implemented, and which could go far towards improving the quality of human-robot interaction.

Figure 1. The B21R combines audio and visual interfaces

In general, a robot capable of auditory perspective taking could pay attention to a number of different factors that might affect human robot communication:

- Masking Noises
- Interruptions / Distractions
- Changes in Operator or Robot Position
- Individual differences between different operators

As people can adapt to each of these naturally, a robot speech interface unable to similarly adapt is likely to prove frustrating to a human partner. Furthermore, many of these concerns can be addressed in part by relatively straightforward approaches.

2. ROBOTIC IMPLEMENTATION
To demonstrate this process of auditory perspective taking, an audio-visual interface was implemented for a real robot (Figure 1) that could handle two types of changes to the auditory scene: (1) masking noise, and (2) interruptions. In particular, the system works successfully in environments exposed to HVAC noise, robot ego-noise (i.e. motors and fans on the robot), radios, and human speech, both in the background and as a nearby

Figure 3. The visual interface displays news reports that the information kiosk can repeat. A new user reads aloud the title of the news story to listen the report.

interruption. The hardware used for this work is our B21R robot equipped with:

- Microphones for monitoring ambient noise (overhead array) and speech recognition (mounted below monitor for better speech pickup)
- Monitor mounted at eye-level to display for new users the available topics the robot may talk about.
- Speaker and internal amplifier to allow the robot to speak at a variety of volumes to a human listener.

The implementation follows the robot information kiosk scenario described in the introduction. The robot has a set of available reports that a human can listen to. For simplicity, this implementation used current news headlines, but they could be any arbitrary set of reports relevant to a military or commercial user. When a person first approaches the information kiosk, they will first see a visual interface displayed on the monitor listing all of the reports that the robot can read for the human user. Figure 3 is an example of such an interface displaying news reports. To hear a report, the user then reads aloud the title of the report (i.e. news story) in which they are interested. Then, speech recognition software on the robot identifies the report title, and calls a text-to-speech engine (TTS) to read aloud the report sentence by sentence. Both the TTS and speech recognition software used were developed using the Microsoft Speech Application Programmers Interface (SAPI 5.1).

As the robot is reading the report aloud, the auditory scene may change (e.g. the HVAC system starts up), making it difficult for a listener to understand the robot. This is where the auditory perspective taking comes in, as the robot must, in addition to speaking, also listen to the environment in which it is communicating so that it can change its auditory output appropriately. In our implementation, the listening is done by sampling 250-ms of audio data from the microphone array at the end of each sentence. Sampling more often is difficult, as the robot contributes to ambient noise levels while it is speaking.

To select the appropriate action from this data, a classifier is used to identify the sample as being clean, noisy, or containing speech. If the signal is clean, then the next sentence is read immediately. If, however, noise is present, but not speech, then ambient noise may be masking the speech output of the robot. In that case the robot can either raise the volume of its speech output, or, if the noise is just too loud (e.g. a helicopter flying overhead), it can pause until the noise levels abate.

Lastly, if the sample is classified as containing speech, then we assume that another human speaker has started talking to the human listener. By assuming that the robot is less important than a human conversant, we expect that the human listener's attention will be taken away from the robot and refocused on the human speaker. This constitutes an interruption. If the robot were to continue while somebody else is talking, then the human listener would not hear what the robot said. In this case, the robot should stop speaking until directed to continue by the listener once the interruption is over.

A problem with pausing in the presence of an interruption is determining from where the robot should resume speaking. If the robot did not stop speaking immediately (i.e. speech was not quickly classified), then the human may have missed something. Also, if the interruption was lengthy, then the person may not remember where the conversation stopped. For this reason, the set of speech commands available to the human user when asking the robot to continue, are designed to allow control over where the robot should continue speaking from. They may ask the robot to continue from the beginning, from the last sentence, from where it stopped, or even change to a new subject.

3. SUMMARY

What we have presented in this paper are some relatively simple solutions to the complex problem of changes in the auditory scene. Even with just a TTS output, a robot does not have to remain passive in dealing with noise or interruptions from the surrounding environment. A robot can raise the volume at which it speaks, pause during very loud noises, or recognize interruptions and wait for the user to refocus their attention. These actions are ones that people often do naturally while speaking to each other, because they at some level want to insure that what they say is being understood, and they are issues at the heart of this notion of auditory perspective taking.

Future goals in exploring this notion auditory perspective taking are twofold. First, we intend to obtain experimental results with human subject testing to determine the robot's success in adapting to changes in the auditory scene. Second, we intend to further explore auditory perspective taking by looking at more human-human interaction under changing environmental noise conditions, and applying it to human-robot interaction.

4. ACKNOWLEDGEMENTS

This research has been funded by the ONR Intelligent Systems Program (Work Order #N0001405WX30022).

5. REFERENCES

[1] L. Hiatt, J. G. Trafton, A. Harrison, and A. Schultz, "A Cognitive Model for Spatial Perspective Taking," presented at International Conference on Cognitive Modeling, Mahwah, NJ, 2004.