Using the Concept of Auditory Perspective Taking to Improve Robotic Speech Presentations for Individual Human Listeners

Derek Brock[1] **and Eric Martinson**[1,2]

[1]Naval Research Laboratory, Washington, DC 20375
[2]Georgia Institute of Technology, Atlanta, GA 30332
[1]brock@itd.nrl.navy.mil, [2]ebeowulf@cc.gatech.edu

Abstract

In this short paper, we introduce the concept of "auditory perspective taking" and discuss its nature and utility in aural interactions between people. We then go on to describe an integrated range of techniques motivated by this idea we have developed for improving the success of robotic speech presentations for individual human user/listeners in relevant circumstances.

Introduction

Auditory perspective taking can be thought of as the ability to reason about and facilitate the perception of speech and sound at another person's point of view. People, in general, are quite good at doing this, readily grasping, for instance, how well their conversational partners can hear what they are saying in changing circumstances and adjusting their presentations accordingly. The presence of ambient noise and people's proximity to each other are two of the most common conditions speakers adjust for, but other factors such as social considerations can also be involved. As a result, people speak louder, move closer, or move to some other location to overcome noise, pause until particularly loud, interrupting sounds have passed, speak quietly when others might hear what they are saying or wish not to be disturbed, and even suspend conversation or resort to other ways of communicating when talking is inappropriate or not feasible.

Although presenting speech in a way that meets a listener's perceptual needs is largely a transparent skill for people in most settings, little or no research has formally addressed this adaptive function in robotics, where the emphasis in speech interfaces has focused primarily on dialogue and initiative. Robust human-robot speech and auditory interaction designs, though, will ultimately require machine strategies for coping with dynamic auditory circumstances. People in conversation naturally rely on each other to collaborate in the effort it takes to hear and be heard, and especially so in challenging conditions. They do this, in part, by modulating their delivery to succeed on the basis of what they perceive are their addressee's requirements. Without this auditory perspective-taking function, the utility of robots intended to carry out collaborative activities with people in face-to-face settings coordinated primarily by speech is likely be checked by a range of common acoustic and task-dependent factors.

The goal of the effort described in this short paper, then, is to automate the coordination of robotic presentations of speech to meet the perceptual expectations of human users. In contrast to the notion of organizing or refashioning what a system says so that its meaning can be better understood, the work outlined here focuses on the problem of ensuring that the system's speech signal itself can be appropriately heard—by countering the masking effects of noise and/or adjusting for the listener's proximity—and is therefore intelligible in various contexts. In the material that follows, we briefly cover areas of related work and then outline several of the adverse conditions an adaptive auditory display might handle, identifying both what such a system should be aware of and the actions it might take to improve its presentations. We next describe our work on a prototype auditory perspective-taking system for a robot configured to function as a mobile information kiosk and give a synopsis of its current abilities. We close with a short critique of our approach and outline additional work on this effort we are currently pursuing.

Related Work

Our work on auditory perspective taking draws from two related areas of user interface research at the Naval Research Laboratory (NRL), visual perspective taking in human-robot interaction and adaptive auditory information displays.

Reasoning about Visual Perspectives

Much of NRL's recent human-robot interaction research has focused on the development of computational methods for reasoning about a user's visual frame of reference in collaborative tasks for platforms such as NASA's Robonaut (Sofge *et al*. 2005). Working from the premise that modes of human-human interaction are the proper

bases for implementing computational collaboration skills, NRL has studied patterns of perspective taking among astronauts in training for extravehicular activities and developed an integrated architecture in which cognitive modeling techniques are employed to allow robots to make inferences about user requests involving non-egocentric frames of reference in a range of circumstances (Hiatt *et al*. 2004; Trafton *et al*. 2005). The notion of auditory perspective taking extends the conceptual premise of this work into the domain of sound.

Adaptive auditory displays

NRL is also investigating the advanced design and use of auditory information in conjunction with visual displays with the goal of reducing effort and improving situation awareness in complex multitask settings such as Navy combat information centers (Brock *et al*. 2004). Since mixed informational uses of sound for individuals must function in the context of external operational uses of sound, NRL begun to consider how individual auditory displays can adaptively self-organize presentations to counter the effects of incidental energetic and informational masking (Brock and Ballas 2005). Adaptations on the basis of *a priori* knowledge and computational monitoring of parameters in the task and listening environments could include modulated rendering of critical sound information (McClimens *et al*. 2005) and prioritized, time-scaled serialization of concurrently arriving audio streams (McClimens and Brock 2006). Although further work on these ideas has been limited both by the need for additional perceptual research and the complex processing requirements of such a system, several of the essential concepts for an adaptive auditory display are demonstrated in the prototype auditory perspective taking system we describe below.

Challenges for Auditory Presentations

Given the utility and the privileged role of spoken language as a medium for interaction and communication between people, and the facility of its use, even in perceptually demanding everyday settings, the design of speech and audio capacities for collaborative robots should arguably incorporate inferential methods for meeting the user's contingent needs as a listener. In this section we consider some of the elements of this problem in terms of sensing and compensatory responses.

Monitoring Auditory Circumstances

Reasoning about a listener's auditory perspective presupposes knowledge of the interaction environment, which, in real-world settings, is frequently subject to change. In this regard, the primary issues an adaptive auditory presentation system must be concerned with are 1) where the addressee is and 2) the presence and nature of competing ambient sounds.

Detecting and tracking the user. Since humans are able to easily move about, they are perceptually skilled at determining where their interlocutors are. This information, coupled with experience, allows people to make basic judgments about how loud their speech should be and where it should be directed.

A robotic auditory display should thus be designed to recognize when a listener is present and to assess the listener's position, relative to its own, in terms of direction and proximity. In addition to standard desktop command mechanisms such as touch screen interaction, this can be accomplished aurally though detection, interpretation, and localization of a user's utterance or visually by identifying the user's proximity and interactive posture, or, preferably, through a combination of information obtained from both of the latter modalities. Once the system has established where the user is, it can make use of *a priori* knowledge to select an appropriate speaking volume. The system's monitoring function should also include a mechanism for tracking the user during an auditory presentation of any length or over the course of a series of speech interactions.

Assessing the auditory scene. People are also extraordinarily skilled at what Bregman refers to as "auditory scene analysis" (Bregman 1990). Not only are they able to listen for salient auditory events in their environment while they are talking, but they also quickly develop and maintain an understanding of how the ambient soundscape is organized and what its properties imply for the success of their aural interactions. In particular, people use this knowledge to augment their basic judgments about the proper volume for their speaking voice.

Since ambient noise can be loud enough to obscure or mask what a user can hear, a robotic auditory display should also include a machine listening function that both monitors the immediate auditory scene for onsets of competing sounds and maintains an evolving representation of sound sources and areas of relative quiet in its operating space. The primary purpose of the monitoring function is to assess the level of ambient sounds and, so, infer their additive effect on the user's listening requirements. A practical difficulty for measuring levels on a moment-by-moment basis, though, is the problem of discounting the robot's own production of incidental and intentional sound. An additional monitoring role is to recognize aurally signaled interruptions.

Keeping track of where sounds occur in the operating space allows the robot to make informed changes in location to minimize the impact of noise in its aural environment. This knowledge can be acquired through an exploratory process of auditory sampling and mapping. Although the maintaining noise map falls short of the ideal of a real-time analysis of the auditory scene, knowledge of the shape and directivity of persistent sound fields could be determined with this technique, which, in turn, would allow the robot to make judgments about where ambient noise can be minimized that are potentially more convenient for the user.

Actions an Adaptive System Might Take

Given a representation of the auditory interaction environment—specifically, knowledge of where the user is and a continuous awareness of ambient sound levels—and corresponding inferences about the user's auditory perspective, there are several adaptive actions a robot's auditory display can take that are likely to meet most of its addressee's listening requirements. The robot can first turn to face its user, which has the added benefit of drawing attention to itself. As it speaks and tracks the listener, it can also raise or lower its voice on the basis of their mutual proximity. Similarly, in the presence of varying levels of ambient noise, the robot can, up to a point, raise its voice to compensate for the difficulty of being heard and, conversely, lower it when competing sounds diminish. When overwhelmingly loud sounds occur, the robot can simply pause until the noise passes and then resume where it left off. If, on the other hand, the immediate setting evolves into a persistently loud situation, the robot can either stop and let the user decide what to do next or it can pause and consult its aural knowledge of the operating space and, with the user's consent, move to, and resume its presentation, in a quieter location. And finally, if the user or another person chooses to vocally interrupt the robot while it is speaking, or an interrupting auditory alert the robot is capable of identifying should sound—for instance, a fire alarm—the robot can yield, by pausing indefinitely, and await further instructions in the new context.

An Auditory Perspective-taking Prototype

There are other aspects of auditory perspective taking that have been notably overlooked in the preceding discussion, particularly those of a social nature. However, the contextual knowledge a system must assess to make these richer forms of auditory inferences is far more semantically complex than the conditions we have outlined here, and fall into the realm of discourse processing and cognitive modeling. In the implementation we describe next, we have, for now, intentionally sidestepped the use of linguistic and cognitive processing techniques and focused on inferential adaptations that can be computed primarily from the robot's environmental sensing technologies.

System Description

As a proof of concept, we are using an iRobot B21R as the basis for a mobile robotic information kiosk that ideally might be placed in the lobby of a conference center or a museum or perhaps in a department store. The kiosk is prominently equipped with a flat-panel display that visually depicts a set of topics and invites a user to make a selection. It then provides a paragraph or two of verbal information on the subject the user has chosen, after which, its waits for further prompts.

As the robot-kiosk presents its information aloud, it exercises a range of sensory functions that are designed to inform the coordination of its output to meet the addressee's inferred listening requirements. The system employs an overhead, four-element array of microphones to detect and localize speech and monitor ambient sound levels. The microphone array is also used by the system in its auditory mapping mode to measure and localize persistent sources of sound in its operational space. As an expedience for command purposes, speech recognition (as opposed to speech detection and localization) is currently handled with a wireless microphone setup and Microsoft's free Speech Application Programming Interface (SAPI 5.1). This API's text-to-speech service is also used to implement the system's spoken output, which is rendered with an internally amplified loudspeaker that faces forward. Last, a stereo vision system (a TRACLabs BiClops) in combination with face detection software (Intel OpenCV) and a laser measurement system (a SICK LSM200) in combination with continuous localization software developed at NRL (Schultz and Adams 1998) are used by the system to visually find and track a user (both direction and proximity) whose voice has been initially detected and localized with the microphone array. More information about the system and an outline of the techniques used to implement speech detection and localization and auditory mapping is given in (Martinson and Brock, submitted).

Synopsis of Present Adaptive Capacities

When our prototype senses a change in its aural relationship with the user, and infers the need for an adjustment in its presentation from the user's listening perspective, there are four adaptive strategies it can currently carry out.

Face the listener. First, on the basis of its speech localization and visual tracking mechanisms, the system can turn the robot to face, and, while it talks, maintain, its physical orientation to the user. This strategy keeps the robot's visual display available for inspection and ensures that the loudspeaker rendering the robot's speech is directionally efficient.

Adjust the robot's speaking volume. Second, the system can raise or lower the robot's voice in response to corresponding changes in the user's proximity and/or the level of ambient noise. Auditory monitoring for this purpose is done while the robot is silent between sentences.

Pause-and-resume or stop-and-wait. Third, the system can pause and resume the robot's speech presentations when it determines that the ambient noise level is momentarily too loud to be heard over. If the noise remains too loud for too long, the system can treat it as an interruption. In this case, it forgoes trying to resume and waits for the user's next instruction. The system will also stop and wait in the same manner if it detects an instance of user speech while a verbal presentation is being made. However, cannot currently detect non-speech auditory alerts. In conjunction with these adaptive moves the system uses the robot's visual display to convey its status

(either pausing or interrupted) and the current set of commands it can execute.

Move to another location. Fourth, when sound levels remain too high where the user and the robot are presently located, the system can make use of its auditory mapping function (Martinson and Schultz Accepted; Martinson and Arkin 2004), in conjunction with the robot's path-planning algorithm and its knowledge of the physical layout of the operating space, to identify and move the robot to a new location that is sufficiently quiet for the listener's needs. Because this is the most recent adaptive strategy we have implemented, we have only evaluated this function for reduced levels of ambient noise at quiet locations identified by the system and have yet to implement additional interaction steps that would collaboratively engage the user in a decision to proceed to a new location.

Discussion and Ongoing Work

The merit of our present sensor-informed approach to the problem of coordinating robotic speech presentations to meet the listening requirements of individual users is that it computationally demonstrates a powerful intuition about aural interactions, namely that interlocutors not only listen but also watch to construe each other's complementary perceptual needs.

Still, a substantial number of challenges to the realization of a truly robust machine implementation of auditory perspective taking skills remains to be addressed. People, for example, frequently have lapses of attention or fail to effectively infer their addressee's perspective, but our system makes provision neither for efficient conversational repairs nor for the user to simply ask the robot to alter its volume. Other concerns, listed roughly in order of increasing complexity, include mapping the directivity of sound-sources, implementing the ability to walk along and talk with the user, solving the problem of aural monitoring while the robot is speaking, developing computational theories of more complex forms of auditory perspective taking, and semantic analysis of individual sound sources relative to auditory maps.

In addition to making iterative refinements to the system, we are presently finalizing plans for a two-part study that is designed to empirically assess whether user recognition and comprehension of robotic presentations of spoken information in noisy environments will be significantly better when adaptive strategies are employed—specifically adaptive changes in speaking volume and the use of pausing strategies—than when they are not in otherwise equivalent circumstances. To evaluate ease of use, in the second part of the study, the utility of these same adaptive functions in the actual system will be compared, again in noisy environments, with two non-adaptive versions of the system, one that allows the user to interactively control the presentation of spoken information and one that provides no such interactive control. Participant performance measures will include counts of correctly identified spoken phrases in a list of targets and foils, time-coded histories of interactions with the system in each condition, answers to subsequent multiple-choice questions about the spoken information, and subjective ratings of workload.

Conclusion

The overarching goal of the work described in this brief report is to address the importance and utility of auditory perspective taking in human-robot interaction. Because demanding sounds and sound ecologies are everywhere, people readily exercise a range of auditory sensibilities in their everyday encounters with each other. Eventually, they will expect collaborative robotic platforms to possess communication skills that are at once familiar and comparable in power to those that their human collaborators employ. For now, our information kiosk demonstrates that even with a relatively modest computational approach, a robotic auditory display can be designed to monitor the shared aural environment and accommodate its user's listening requirements when mundane contingencies such as changes in the user's proximity or increased levels of ambient noise threaten undermine the successful delivery of auditory information.

One of our ambitions is to extend the adaptive abilities of robotic auditory displays by expanding the system's knowledge about the auditory scene that surrounds it. Knowledge of source types, for instance, will allow the system to anticipate how a particular sound source is likely to change in time and, thus, make more experientially informed inferences about its listener's auditory perspective. However, our eventual aim is to explore more complex aspects of auditory perspective taking by augmenting the system's inferential resources with adaptive linguistic and cognitive processing techniques that complement our current sensor-based approach.

Acknowledgments

We would like to thank Bill Adams, Magda Bugajska, and Dennis Perzanowski for their comments and technical assistance in the development of the information kiosk. This research was funded by the Office of Naval Research under work order number N0001406WX30002.

References

Bregman, A. S. 1990. *Auditory Scene Analysis.* Cambridge, MA: MIT Press.

Brock, D. and Ballas, J. A. 2005. Audio in VR: Beyond Entertainment Setups and Telephones. *Proceedings of the 1st International Conference on Virtual Reality.*

Brock, D., Ballas, J. A., Stroup, J. L., and McClimens, B. 2004. The Design of Mixed-use, Virtual Auditory Displays: Recent Findings with a Dual-task Paradigm.

Proceedings of the 10th International Conference on Auditory Display.

Hiatt, L., Trafton, J. G., Harrison, A., and Schultz, A. 2004. A Cognitive Model for Spatial Perspective Taking. *International Conference on Cognitive Modeling*. Erlbaum, Mahwah, NJ.

Martinson, E. and Arkin, R. C. 2004. Noise Maps for Acoustically Sensitive Navigation. *Proceedings of SPIE, 5609.*

Martinson, E. and Brock, D. Improving Human-robot Interaction through Adaptation to the Auditory Scene. *2nd International Conference on Human-Robot Interaction.* Submitted.

Martinson, E. and Schultz, A. Auditory Evidence Grids. Proceedings of the 2006 International Conferance on Intelligent Robots and Systems (IROS 2006). Accepted.

McClimens, B, Brock, D, and Mintz, F. 2006. Minimizing Information Overload in a Communications System Utilizing Temporal Scaling and Serialization. *Proceedings of the 12th International Conference for Auditory Display*.

McClimens, B., Nevitt, J., Zhao, C., Brock, D., and Ballas, J. A. 2005. The Effect of Pitch Shifts on the Identification of Environmental Sounds: Design Considerations for the Modification of Sounds in Auditory Displays. *Proceedings of the 11th International Conference on Auditory Display*.

Sofge, D., Bugajska, M., Trafton, J. G., Perzanowski, D., Thomas, S., Skubic, M., Blisard, S., Cassimatis, N. L., Brock, D. P., Adams, W., and Schultz, A. 2005. Collaborating with Humanoid Robots in Space. *International Journal of Humanoid Robotics*, vol. 2, pp. 181-201.

Trafton, J. G., Cassimatis, N. L., Bugajska, M., Brock, D. P., Mintz, F. E., and Schultz, A. 2005. Enabling Effective Human-robot Interaction Using Perspective-taking in Robots. *IEEE Trans. on Systems, Man and Cybernetics, Part A*, vol. 35, pp. 460-470.

Schultz, A. and Adams, W. 1998. Continuous Localization Using Evidence Grids," *IEEE International Conference on Robotics and Automation*.