EXPLORING THE UTILITY OF GIVING ROBOTS AUDITORY PERSPECTIVE-TAKING ABILITIES

Derek Brock

Naval Research Laboratory
Washington, DC 20375
brock@itd.nrl.navy.mil

Eric Martinson

Georgia Institute of Technology
Atlanta, GA 33302
ebeowulf@cc.gatech.edu

ABSTRACT

This paper reports on work in progress to develop a computational auditory perspective taking system for a robot. Auditory perspective taking is construed as the ability to reason about inferred or posited factors that affect an addressee's perspective as a listener for the purpose of presenting auditory information in an appropriate and effective manner. High-level aspects of this aural interaction skill are discussed, and a prototype adaptive auditory display, implemented in the context of a robotic information kiosk, is described and critiqued. Additionally, a sketch of the design and goals of a user study planned for later this year is given. A demonstration of the prototype system will accompany the presentation of this research in the poster session.

1. INTRODUCTION

Gregory Kramer notes that, "auditory display research applies the ways we use sound in everyday life to the human/machine interface and extends these uses via technology." [1]. One area that is particularly appropriate for this characterization is human-robot interaction. Already, robots designed for everyday and specialized environments are being marketed as toys, home appliances, and battlefield support. And other platforms, intended to function as human surrogates in schools, museums, and hospitals, are moving from the realm of research projects to fully deployed robotic systems. To foster intuitive interactions, some of these systems incorporate rudimentary speech and sound displays, and others are engineered for specific types of listening and aural communication tasks (e.g., [2][3]). Generally, though, the ability of robots to flexibly exercise interactive behaviors informed by the interpretation and production of auditory information remains far behind the broad and mostly transparent skills of human beings.

One of the more unique skills people possess with regard to their everyday use of sound is the ability to imagine and reason about another's auditory perspective. People use this important capacity to coordinate their speech and other productions of sound for addressees in a variety of ways that both take into account the auditory scene in their shared environment and satisfy public and private concerns about sound that attend social interactions. Auditory perspective taking, for instance, makes it possible to reasonably ensure another's auditory comprehension in a noisy space, to minimize overhearing in a quiet setting, to judge whether the sound resulting from something one does will be apprehended by others, and even to determine the precedence of interrupting speech or sound.

As collaborative encounters with robotic systems become more common in the coming years, people will increasingly expect the communication skills of these platforms to match the native abilities of humans. Perceptual perspective taking is one of a number of skills people use to facilitate the effort needed to converse with each other, and it is nearly indispensable for the coordination of shared actions. Giving a robot the ability to ascertain and assess aural parameters that are relevant to ensuring an addressee can adequately hear a display of sound information is an essential prerequisite for implementing the auditory dimension of this skill. Coupled with an *a priori* set of rules, a robot with even limited, semantically impoverished knowledge of the auditory scene will be better equipped to adapt its speech or other sonic output to meet the ordinary sorts of challenges that occur in dynamic settings.

In this report on work in progress, the authors describe a prototype computational auditory perspective taking system that is implemented in the context of a robotic information kiosk. Functional and theoretical aspects of the auditory display system and the design of a user study that will be carried out later this year are presented.

2. RELATED WORK

The authors' work on robotic auditory perspective taking grows out of two areas of user interface research at the Naval Research Laboratory (NRL) concerned, respectively, with the design of human-robot interaction and the display of auditory information.

2.1. A Computational Framework for Visual Perspective Taking

Following on work that focused on the development of intuitive speech and gestural modes for user control of robots [4], NRL's robotics group recently turned its attention to the implementation of collaborative perspective-taking abilities in robotic assistants, such as NASA's Robonaut [5]. In a case study of extravehicular training, it was found that over 85% of astronauts' utterances involved non-egocentric frames of reference, and that in about 25% of their interactions, it was necessary for the collaborators to take or use each other's perspective [6]. Motivated by these and other findings, NRL has developed a framework for human-robot interaction that employs computational cognitive models to reason about non-egocentric perspectives in a range of collaborative circumstances [6]. Although this work has so far only addressed spatial reasoning in the context of visual perspective taking, the broader aim is to develop an integrated computational architecture that also supports auditory and haptic modes of perceptual cognition and perspective taking.

2.2. Adaptive Auditory Displays

NRL's auditory display group studies advanced uses of sound in conjunction with visual displays in complex decision environments (e.g., [7]). Operational goals for this work include transparent management of attention in multitask settings, reductions in required effort, increased levels of situation awareness, and improved monitoring capacities for voice communications. A recent performance study involving mixed, heterogeneous uses of sound, though, raised concerns about task attribution and priority [8]. Partly in response to this issue, NRL is considering how auditory displays might self-organize their presentations on the basis of *a priori* knowledge and computational monitoring of various parameters in the task and listening environments. Display adaptations could include prioritized serialization of concurrently arriving audio streams and, when appropriate, modulated rendering of sound information through parameter adjustments such as amplitude, fundamental frequency, timbre, and rhythm. The latter of these adaptations exploits the notion of vacant auditory niches, but depends critically on the ability of listeners to recognize the underlying identity of altered sounds [9][10]. A brief overview of these adaptive display ideas can be found in [11].

3. A PROTOTYPE ROBOTIC AUDITORY PERSPECTIVE-TAKING SYSTEM

Both computational reasoning about perspectives and the challenge of reactively adapting the presentation of auditory information to compensate for various concerns are particularly relevant for robotic platforms that are intended to do things with people. Because humans easily move about, they typically employ a range of tactics for sustaining face-to-face conversation that includes reorienting their direction of address, moving together, and speaking appropriately louder or softer, all of which are relevant for mobile robots as well as certain kinds of stationary platforms. Noise, too, routinely makes it difficult to be heard, and people readily use the same tactics and others to compensate, such as pausing until the noise has passed or, in the event of constant noise, moving to a quieter place or even resorting to a different medium such as writing. People also exercise intuitions about interruptions when they are speaking that can be construed as a reactive adaptation that involves perspective taking as well. The source of the interruption may be an addressee or another individual or an instance of sound information that is collaboratively known to have precedence. By pausing until it is appropriate to resume, a speaker yields the floor, as it were, to the importance of the interrupting concern. The authors' goal in the work reported here is to implement a computational capacity for these sorts of interactive, auditory adaptations in a robot acting ultimately as both a speaker and a listener.

3.1. Performance Adaptations and Implementation Details

Several of these adaptive strategies are currently implemented in a prototype interactive auditory display for a robot the authors have devised that functions as an information kiosk. The setup, which is shown in Figure 1, involves a mobile iRobot B21R, configured with an overhead microphone array to monitor ambient sound, a microphone placed below this array for speech commands, a visual display for communication with images and text, and an internally amplified loudspeaker for delivering information aurally. Using speech recognition and text-to-speech services available through Microsoft's Speech Application Programmers Interface (SAPI 5.1), the system reads information briefs aloud to users after being verbally prompted from a menu of titled images depicted on the visual display. As it speaks, between sentences, the robot samples 0.25 sec. of ambient sound from the microphone array and passes this to a classifier that determines whether the auditory data is "clean" or "noisy" or contains speech. The robot then decides how it should continue on the basis of a set of rules and thresholds that function as an *a priori* assessment of the listener's auditory perspective.

Figure 1. *A mobile iRobot B21R configured with microphones, an amplified loudspeaker, and a visual display to function as an information kiosk.*

Up to a point in the presence of occasionally elevated ambient sounds, the robot's first tactic is simply to speak louder (cf. the "Lombard reflex," e.g., [12]). Using different high and low octave-band thresholds, the classifier is tuned to disregard a variety of nearby, indoor sounds with low magnitudes, including heating, ventilation, and air conditioning (HVAC) sounds, radios, and people conversing in the background, as well as sounds associated with the robot's own operation. However, if any of these sounds happens to become loud enough, the potential arises to partially mask, and thereby undermine the intelligibility of, the robot's speech from the listener's perspective. Thus, to maintain intelligibility, and also to subsequently return to a normal speaking voice, the prototype auditory display shifts the volume of its text-to-speech output up or down, in a linear manner derived from [13], as the level of background sounds fluctuate below levels that warrant a different adaptive response.

When one or more ambient sounds exceed the classifier's thresholds for noise, the robot's next tactic is either to pause until the masking sound abates (rather than

to speak even louder) or, when the noise is prolonged or speech is detected, to regard the interference as an interruption and suspend its spoken presentation indefinitely. The classifier decides if interrupting voices are present at the same time it checks for noise with a method for detecting speech onsets that is similar to a technique for onset detections described in [14]. Since a degree of variability is often present in the power of masking sounds, the system dynamically increases its sensitivity when a noisy signal is first identified by lowering its octave-band thresholds; it then resumes speaking as soon as it is able to classify 0.75 sec. of ambient sound as clean. If the sound, however, fails to abate and the robot is unable to continue after 10 seconds, the pause is construed as a suspending interruption. To coordinate how the listener wishes to proceed when the time is next appropriate or favorable listening conditions return, the robot resorts to its visual display and exhibits a menu of verbal prompts it can carry out that allow the user to have the presentation resumed from where it stopped or from the last sentence, begun again, or dropped in favor of a different information brief.

3.2. Implementation Critique

In spite of a modest repertoire of performance adaptations and a current presumption that shared sound conditions are approximately the same for the robot and its addressee, the prototype information kiosk demonstrates that an interactive auditory display capable of changing its presentation to accommodate the consequences of changes in the shared auditory scene from the listening perspective of its user is feasible on a robotic platform. The present implementation of the system's adaptive responses to interrupting noise or speech, i.e., pausing briefly or suspending its presentation and soliciting further interaction through its visual display, balances the need to account for the addressee's perspective—particularly, his or her inability to hear what is being said in the presence of excessive noise—against the prototype's current limitations, especially, its pragmatic approach to the problem of sampling and classifying ambient sound and its lack of a natural language dialog system for verbally negotiating how its presentation should continue. Although it is a substantial machine listening problem, presentation adaptations at the granularity of words or phonemes require classification of ambient noise while the robot is speaking. Similarly, the problem of system comprehension of arbitrary user references to presented information requires computational methods for collaborative discourse. In addition to moving forward on these two issues, more robust auditory perspective taking adaptations will result from giving the system a range of additional abilities including a capacity to locate, track, turn to, and move to an addressee's position, a capacity to map and subsequently infer noise levels at specific locations in auditory environment (e.g., [15]), and a capacity to distinguish between users and their needs as listeners.

4. PLANNED USER STUDY

Although the ability to reason about one another's auditory perspective facilitates people's aural interactions and informs their adaptive strategies when sonic, spatial, and/or social constraints are present, it is likely that additional factors such as visual interactions and language use skills also play important roles in people's successful coordination of speech and other sound information in challenging circumstances. On the other hand, machine systems that are designed to do things with people are inherently limited by the state of the art in this regard and, at best, can only account for a subset of the full range of factors at work in native interaction skills. Because of this, user testing is needed to evaluate the efficacy of interaction designs and to indicate theoretical and technical directions for making further improvements.

Due to the limited interactive nature of the information kiosk, and also because the notion of adaptive auditory interfaces has been little studied, an experiment is planned to evaluate high-level, user performance issues that are relevant to the set of presentation adaptations the prototype auditory perspective-taking system is currently capable of making. Using a counter-balanced, within-subjects design, participants immersed in a scripted auditory scene with the robot will be asked to select and listen to information briefs in a baseline, non-adaptive auditory display condition and in a second condition in which the prototype's adaptations are exercised. The scene will entail a range of characteristic ambient sounds, with levels varying at times from partly to fully masking and vocal interruptions from the experimenter for answers to simple questions, Although piloting for the design of the study's materials will make use of an immersive, virtual auditory simulation, ambient sounds in the actual study will be rendered with a 28-source, loudspeaker array installation at NRL [16].

Objective measures of intelligibility and information retention, and a subjective measure of workload will be collected. The first of these data will be gathered by asking participants to select phrases present in the spoken presentation from a list of targets and foils as they carry out each condition. Workload will be measured immediately after each condition with an instrument such as the NASA Task Load Index (TLX) [17]. After the measurement of workload, information retention will be gauged by asking participants to write down what they can recall about the information brief they heard. While it is expected that each of these measures will differ between conditions and show a substantial advantage for the "adaptive" condition, the ambient auditory events and the experimenter interruptions will be scripted in such a way that correspondences between targets and classes of auditory interference in the intelligibility data can also be explored to evaluate both the relative costs of different types of interference and the relative benefits of each of the presentation adaptations. Additionally, participants may be asked through an interview process after their last session to assess the performance of the adaptive auditory display and to characterize presentation strategies they might employ in the same circumstances if they were presenting an information brief.

5. CONCLUSIONS

The overarching goal of the work in progress described in the preceding material is to address the importance and utility of auditory perspective taking in human-robot interaction. People routinely exercise a range of auditory sensibilities in their everyday encounters with each other, and they will ultimately expect collaborative robotic platforms to possess similar communication skills. The authors' 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 perspective when noisy or sonically disruptive events occur.

In addition to the user study planned for later this year, the authors and their colleagues intend to extend the present system in several ways to further explore the computational dimensions of auditory perspective taking. The incorporation of a cognitive modeling component and NRL's natural language understanding system [18] are contemplated, and a technique for locating addressees is needed to allow robotic movement in response to user repositioning. Given knowledge of the user's location relative to the robot and a representation of the interaction semantics obtained from the natural language system, the cognitive modeling component will allow the system to reason in a more flexible and psychologically plausible way about its user's auditory perspective. Additionally, this approach will also allow the system to reason about social constraints on the collaborative use of auditory information.

6. DEMONSTRATION

A short, 10 to 15 min., demonstration of the prototype information kiosk will accompany the presentation of this research in the conference's poster session. An explanation of the role of auditory perspective taking in human-robot interaction and a presentation of the system's adaptive strategies will be given. For practical reasons, the demonstration will be a table-top setup and will not make use of the iRobot B21R platform. In addition to ambient noises in the exhibit area of the conference, recorded sources of ambient noise will be used to exercise the system.

7. ACKNOWLEDGEMENTS

The authors 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 N0001405WX30022.

8. REFERENCES

[1] G. Kramer, "An introduction to auditory display," in *Auditory display: Sonification, audification, and auditory interfaces*, G. Kramer, Ed., Santa Fe Institute Studies in the Sciences of Complexity, Proceedings Volume XVIII, Addison-Wesley, Reading, MA, 1994, pp. 1-77.

[2] C. Breazeal, "Emotive qualities in lip-synchronized robot speech," *Advanced Robotics*, vol. 17, no. 2, pp. 97-113, 2003.

[3] S. Yamamoto, K. Nakadai, H. Tsujino, T. Yokoyama, and H. G. Okuno, "Improvement of robot audition by interfacing sound source separation and automatic speech recognition with missing feature theory," in *Proceedings of the 2004 IEEE International Conference on Robotics and Automation*, 2004, pp.1517-1523.

[4] D. Perzanowski, A. Schultz, W. Adams, M. Bugajska, E. Marsh, J. G. Trafton, D. P. Brock, M. Skubic, and M. Abramson, "Communicating with teams of cooperative robots," in *Multi-Robot Systems: From Swarms to Intelligent Automata*, A. C. Schultz and L. E. Parker, Eds. Kluwer, The Netherlands, 2002, pp. 16-20.

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

[6] J. G. Trafton, N. L. Cassimatis, M. Bugajska, D. P. Brock, F. E. Mintz, and A. Schultz, "Enabling effective human-robot interaction using perspective-taking in robots," *IEEE Trans. on Systems, Man and Cybernetics, Part A*, vol. 35, pp. 460-470, 2005.

[7] D. Brock, J.L. Stroup, and J.A. Ballas, "Effects of 3D auditory cueing on dual task performance in a simulated multiscreen watchstation environment," in *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, 2002.

[8] D. Brock, J.A. Ballas, J.L. Stroup, and B. McClimens, "The design of mixed-use, virtual auditory displays: Recent findings with a dual-task paradigm," in *Proceedings of the 10th International Conference on Auditory Display*, 2004.

[9] D. Brock, J.A. Ballas, J.L. Stroup, and B. McClimens, "Perceptual issues for the use of 3D auditory displays in operational environments," in *Proceedings of the International Symposium on Information and Communication Technologies*, pp. 470-473, 2003.

[10] B. McClimens, J. Nevitt, C. Zhao, D. Brock, and J. A. Ballas, "The effect of pitch shifts on the identification of environmental sounds: Design considerations for the modification of sounds in auditory displays," in *Proceedings of the 11th International Conference on Auditory Display*, 2005.

[11] D. Brock and J. A. Ballas, "Audio in VR: Beyond entertainment setups and telephones," in *Proceedings, 1st International Conference on Virtual Reality*, 2005.

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

[13] J. E. Hawkins, Jr. and S. S. Stevens, "The masking of pure tones and speech by white noise," *J. Acoust. Soc. Am.*, vol . 22, no., 1, pp. 6-13, 1990.

[14] M. Goto and Y. Muraoka, "A real-time beat tracking system for audio signals," in *Proceedings of the International Computer Music Conference*, 1995.

[15] E. Martinson and R. C. Arkin, "Noise Maps for Acoustically Sensitive Navigation," *Proceedings of SPIE*, vol. 5609, pp. 50-60, 2004.

[16] J. A. Ballas, H. Fouad, D. Brock, and J. Stroup, "The effect of auditory rendering on perceived movement: Loudspeaker density and HRTF," in *Proceedings of the 7th International Conference on Auditory Display*, Espoo, Finland, 2001.

[17] V. J. Gawron, *Human Performance Measures Handbook*, Lawrence Erlbaum, Mahwah, NJ, 2000.

[18] K. Wauchope, "Eucalyptus: Integrating natural language input with a graphical user interface," Naval Research Lab., Washington, DC, Tech. Rep. NRL/FR/5510-94-9711, 2000.