Integrating Vision and Audition within a Cognitive Architecture to Track Conversations

J. Gregory Trafton
Naval Research Lab
Code 5515
Washington, DC 20375-5337
trafton@itd.nrl.navy.mil

Magdalena D. Bugajska
Naval Research Lab
Code 5515
Washington, DC 20375-5337
magda@aic.nrl.navy.mil

Benjamin R. Fransen
Naval Research Lab
Code 5515
Washington, DC 20375-5337
fransen@aic.nrl.navy.mil

Raj M. Ratwani
Naval Research Lab
Code 5515
Washington, DC 20375-5337
ratwani@itd.nrl.navy.mil

ABSTRACT

We describe a computational cognitive architecture for robots which we call *ACT-R/E* (ACT-R/Embodied). ACT-R/E is based on ACT-R [1, 2] but uses different visual, auditory, and movement modules. We describe a model that uses ACT-R/E to integrate visual and auditory information to perform conversation tracking in a dynamic environment. We also performed an empirical evaluation study which shows that people see our conversational tracking system as extremely natural.

Categories and Subject Descriptors

H.1 [**Models/Principles**]: Human information processing

General Terms

Theory

Keywords

Cognitive modeling, ACT-R, human-robot interaction, Conversation following

1. INTRODUCTION

One of the goals of Human-Robot Interaction (HRI) research is to have natural conversation partners. The complete solution to this large goal is not currently within our reach. However, conversation tracking, one of the subcomponents of full conversation ability, has recently made large strides within both the agent community and the robot community.

Conversation tracking is based on several core human competencies. First, and perhaps most importantly, people who hold conversations in dyads, small groups (˜5 people) and large groups (10+ people) take turns [10, 11, 34]. Most transitions between people occur at transition relevant places (TRPs) with a slight gap between speakers [34].

Second, the interval between turns ranges between 450-650ms [7]. Of course, people can and do interrupt speakers [24, 34], but that is not the focus in this paper.

Third, people generally look at the person who is speaking [3, 26, 44]. While the general finding is clear, the amount that people actually look at the speaker varies a bit from study to study. Argyle and Cook, for example, found that people looked at the speaker 75% of the time, while Nielsen found that people looked at the speaker 62% of the time. In a carefully controlled study using four-person groups and eye-tracking, Vertegaal et al. found that listeners looked at the person who was speaking 88% of the time.[1] From their study, Vertegaal et al. concluded that gazing at faces is an excellent predictor of conversational attention in multi-party conversations. They claim "Overall, our results mean that the user's eye gaze can form a reliable source of input for conversational systems that need to establish whom the user is speaking or listening to" (p. 307).

Other researchers have taken this claim to heart. There are now a large number of agent and robotic systems that can follow a user's gaze to help determine who a user is talking to, and who is speaking. For example, Qvarfordt and Zhai [31] used eye gaze to sense user interest. Raidt, Bailly, and Ellisei [32] created an agent that responded to a user's gaze. They found in an evaluation experiment that people responded to eye gaze cues from a computer-generated agent.

Other researchers have combined gaze with speech information to identify the addressee or speaker. For example, Katzenmaier et al. [14] used head pose and (non-spatial) speech features (e.g., the existence of names and syntactic and semantic features) to find which person was being addressed in a multi-person conversation. Similarly, Ou and colleagues [28, 29] used audio and video information to pre-

This paper is authored by an employee(s) of the United States Government and is in the public domain.
HRI'08, March 12–15, 2008, Amsterdam, The Netherlands.
ACM 978-1-60558-017-3/08/03.

[1] Note also that the task itself can influence the amount of speaker-gaze that occurs. Argyle and Graham [4] found that when a dyad used a central map to help plan a holiday, mutual gaze dropped to 6%.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE
2008

2. REPORT TYPE

3. DATES COVERED
00-00-2008 to 00-00-2008

4. TITLE AND SUBTITLE
Integrating Vision and Audition within a Cognitive Architecture to Track Conversations

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Research Laboratory,Code 5515,4555 Overlook Avenue SW,Washington,DC,20375

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
HRI'08, March 12-15, 2008, Amsterdam, The Netherlands

14. ABSTRACT
We describe a computational cognitive architecture for robots which we call ACT-R/E (ACT-R/Embodied). ACT-R/E is based on ACT-R [1, 2] but uses different visual, auditory and movement modules. We describe a model that uses ACT-R/E to integrate visual and auditory information to perform conversation tracking in a dynamic environment. We also performed an empirical evaluation study which shows that people see our conversational tracking system as extremely natural.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

dict where a user's focus of attention is and will be. They found that hand-coded audio information combined with visual information was a better predictor than audio alone.

Finally, there are many researchers who have given robots the capability to gaze at people's faces [13, 18] and track where a person is looking [6, 35]. In the past few years there have been several projects that integrate auditory and visual information for human-robot interaction [9, 17, 27]. However, only a much smaller group actually tries to follow a conversation [37, 45].

Yoshikawa et al, for example, gave their robot the capability to gaze in response to where a person was looking during dyadic conversations. They found that when the robot used a responsive gaze (e.g., mutual gaze between robot and human or following the person's gaze) the human felt like s/he was being looked at more than when the robot used a non-responsive gaze (e.g., a random or staring gaze). Similarly, Sidner et al. found that when their robot gestured and gazed at people during a conversation it was perceived as more engaging than a robot that did not move or gaze at all, especially for female participants.

This brief introduction suggests that not only is it possible for a robot to follow a conversation using a human's gaze, but that people like and expect a robot to visually follow a conversation.

One aspect of this review that is particularly interesting is that, while gaze is an excellent predictor of who is talking (ranging from 62%-88% in goal-directed conversations), there exists an even better predictor of who is talking: the spatial location of the sound source. That is, when people speak, people are able to perform sound localization and find the origin of the sound in space. Humans are able to accomplish this task by the very brief time delay of sound hitting the two ears.

It is clear, however, that people can use both vision and audition to locate distal objects. These different modalities provide convergent spatial cues which must be integrated into a coherent and unified perception [20].

Our goal in this paper is to use sound localization to guide vision to find the speaker. By using auditory information, we can integrate the audition and visual streams to form a coherent representation of the speaker. This information can then be used to track conversations.

2. ROBOT SYSTEM

The robot, named George, is a commercial iRobot B21r suited to operation in interior environments. It has a zero turn radius drive system, an array of range, image, and tactile sensors, and an on-board network of Linux and Windows computers with a wireless Ethernet link to the external computer network. The robot is shown in Figure 1. In the following sections we describe systems that are specific to conversation tracking.

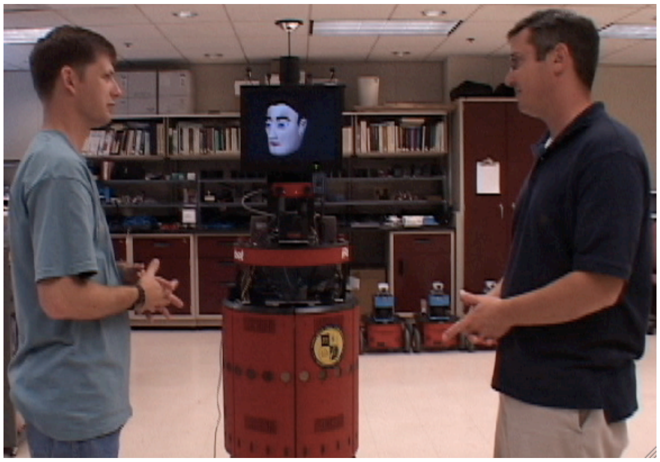

Figure 1: The robot on which we have implemented ACT-R/E and conversation tracking.

2.1 Omni-Directional Person Tracking

An omni-directional camera is used to visually record the radial position of people around the robot. Placed atop the robot, the camera is allotted full 360° visibility. Without recovering depth, such a camera must focus on the use of color information for the entire tracking process. To handle the case of a moving camera (robot) or very dynamic environments, we utilize maximal discriminative data to build a model for each person that maximizes the difference between color information of their skin and clothes with the surrounding environment and other people.

Person tracking begins with a detection system that distinguishes people from other objects in a room. To detect people a boosted cascade detection algorithm has been utilized [21]. The system is based on integrating the results from a large number of weak classifiers to form a strong classifier.

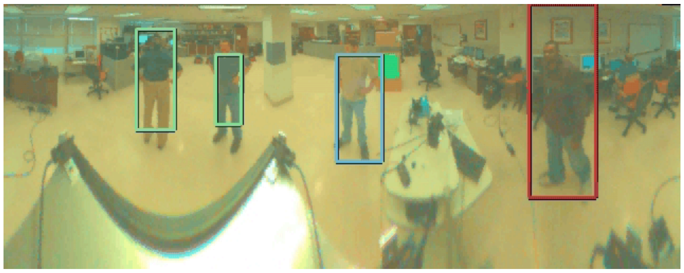

Figure 2: Multi-person, omni-directional tracking.

Once a person is detected, a maximal discriminative model is generated and then tracked using a particle filter [12]. While the maximum number of tracked people is not limited by the logical implementation, performance does decrease linearly with each additional person tracked.

2.2 Robot Audition

To facilitate speaker identification, robot audition is utilized to spatially locate sound sources in the robot's surrounding environment. The robot's auditory system consists of four omni-directional microphones placed just under the omni-directional camera. The four microphones form six pairs that are evaluated as pairs and then combined to estimate sound source positions. Time of arrival delay is estimated for sounds arriving at two microphones by computing their cross correlation in the frequency domain [23, 22]. The microphone pairs are then combined to form a prediction for sound source locations. One such prediction generated from a speaker is located in figure 3.

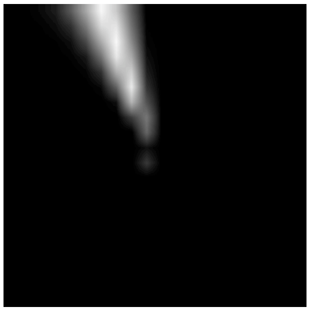

Figure 3: Speaker localization.

To avoid swamping higher level logic with an excessive number of events, auditory information must be pruned prior to the forwarding of data from the sensors to the cognitive model. One such technique is the use of frequency filters, limiting the range of sounds evaluated by the robot audition system to speech-only frequency ranges or other event specific frequency bands. Limiting the frequency, however, greatly restricts the types of events that can be output by the auditory system. Rather than limit the types of sounds, we choose to generate a background model for sound events, such that significant sounds will be localized, whether they are speech or not, and higher level logic will be used for the identification of speech or other events. To model the background sound, an adaptive background model was generated that labels sound events as background or foreground based on volume level. To perform the labeling, a k-means algorithm is applied to sound recorded during a period of time containing user interaction. The sound level is automatically segmented into three classes by the k-means algorithm. To classify sound events as significant, the median of the middle class is used. Sound events above the median are labeled significant, with locations passed onto the cognitive model. For events below the median, no information is passed to higher level reasoning units. Performing significance labeling of sound events facilitates the removal of background noise such as computer fans and heating vents.

2.3 Face

For interaction with humans, the robot displays an expressive, animated face on its LCD "head." The face is used for looking at conversationalists. Specifically, the face can turn to look at any location in a 360° area (when the face looks behind the screen, the back of the head is shown). The face algorithm is more fully described in [30, 38].

3. ACT-R/E

Our approach to human-robot interaction focuses on the hypothesis that a robot that is able to think like a person is better able to interact with a person than a robot that does not [41, 42, 43]. We believe that the best way to create systems that think and reason like people do is to use computational cognitive architectures like ACT-R [1, 2], Soar [19, 25], or EPIC [16].

Our choice of cognitive architectures is ACT-R (Adaptive Control of Thought-Rational). The ACT family of theories has a long history of integrating and organizing psychological data and has been broadly tested in psychological and computational terms.

One aspect of ACT-R, however, is that the connections to the outside world are currently limited to a relatively small number of sensors and effectors (though see Ritter's work for reducing this limitation [33] and Best and Lebiere's work on connecting ACT-R to Unreal Tournament [5]). In order to allow ACT-R to work in a robotic environment, we needed to enhance several aspects of ACT-R. Our version of ACT-R, which we call **ACT-R/E** (ACT-R: Embodied), is shown in Figure 4.

ACT-R is a hybrid symbolic/sub-symbolic production-based system. ACT-R consists of a number of modules, buffers, and a central pattern matcher. Modules in ACT-R contain a relatively specific cognitive faculty usually associated with a specific region of the brain. For each module, there is one or more buffers that communicates directly with that module as an interface to the rest of ACT-R. At any point in time, there may be at most one item in any individual buffer; thus, the module's job is to decide what and when to put a symbolic object into a buffer. The pattern matcher uses the contents of the buffers to match specific productions.

Standard ACT-R interfaces with the outside world through the visual module, the aural module, the motor module, and the vocal module. Other current modules include the intentional, imaginal, temporal and declarative modules.

For ACT-R/E, we have added two new modules (pedal and spatial) and modified the visual and aural modules to work with our robot and to use real-world sensor modalities. We did not modify other parts of the architecture itself.

3.1 Intentional Module

The intentional module is responsible for the current goal state; most ACT-R models use the intentional module and the associated goal buffer to control the order that productions fire.

3.2 Imaginal Module

The imaginal module and the associated imaginal buffer maintains current context relevant to the current goal as well as providing a rudimentary imagination.

3.3 Declarative Module

The declarative module is the core memory system of ACT-R and is responsible for what items can be remembered and how long it takes to retrieve specific items in memory. In simplest terms, it is a sophisticated semantic network.

3.4 Temporal Module

ACT-R's temporal module provides the capability for a model to perform prospective time estimation (i.e., determining when a given time interval has passed). Complete details about the temporal module are given in [39].

3.5 Spatial Module

We have given our robot access to a "cognitive map" so that it can talk about spatial components in a manner that is compatible with the way people think about space. Complete details about the spatial component can be found in [15].

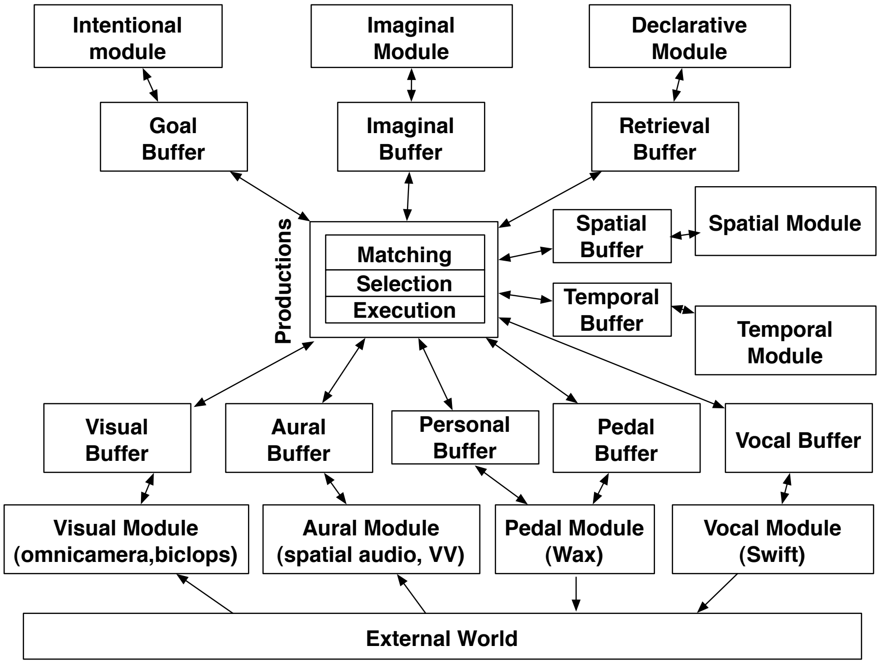

Figure 4: Our ACT-R/E architectural diagram. The original ACT-R architecture and diagram is based on Anderson (2007). Systems and hardware shown in parentheses are robot-specific additions.

3.6 Visual Module

The Visual Module is used to provide a model with information about what can be seen in the current environment. We modified the original visual module to accept input from the omni-directional camera rather than through ACT-R's typical input, a computer monitor. The input from the omnicam is projected onto a virtual wrap-around screen centered around the robot, which allows us to preserve the majority of the original logic of the module; the modification simply accounts for contiguity of the screen. The virtual screen is updated at a constant rate using tracks obtained by the omnicam software and processed according to the internal ACT-R schedule. The visual module allows access to both the location of a tracked person (the "where" system) and a more detailed representation (the "what" system). Obtaining additional information about an object or person requires declarative retrieval(s). Finally, the default requests (i.e. buffer stuffing) are biased towards the location of the previously attended person.

3.7 Aural Module

The Aural Module provides a model with rudimentary auditory perception abilities and is very similar to the Visual Module. We modified the original module to take input from our sound-source localization system and insert it into ACT-R's audicon. The audicon can be considered what ACT-R can hear. We augmented the original symbolic representation used by the aural "where" system to include the peak direction and the two tails of the Gaussian distribution of the sound's location. Thus, for each sound that the sound localization system detects, a symbol is inserted into the audicon which ACT-R can then reason about.

The auditory scene is updated at a constant rate using the results of the scene analysis by the sound-source localization software and processed instantaneously to a level of individual sound sources. As was the case with the visual module, the aural module allows access to both the location of a detected sound and a more detailed "what" representation. Additional information about an aural object (e.g., the semantics of a specific word) can then be retrieved through the declarative module. Finally, the default requests are biased towards the location of the previously attended sound as well as newest and loudest sounds.

3.8 Pedal Module

The pedal module allows commands to be issued to the navigation and mobility system, as well as providing self-localization knowledge. The pedal module interfaces directly with our WAX system that provides localization, navigation, and path planning [36].

3.9 Vocal Module

The Vocal Module gives ACT-R a limited ability to speak. The speech requests processed by this module are forwarded to robot's output system based on Cepstral's Swift system [8].

4. TASK

The robot's task is to "simply" look at the person who is speaking. The task itself is in actuality quite difficult due to multiple auditory and visual distractors such as ambient noise, overlapping conversations (e.g., attempts to interrupt the speaker), multiple people and their motion, etc. To be consistent with human behavior, the robot should note the presence of the distractors, but only attend to them as appropriate. This task is highly dynamic and requires contin-

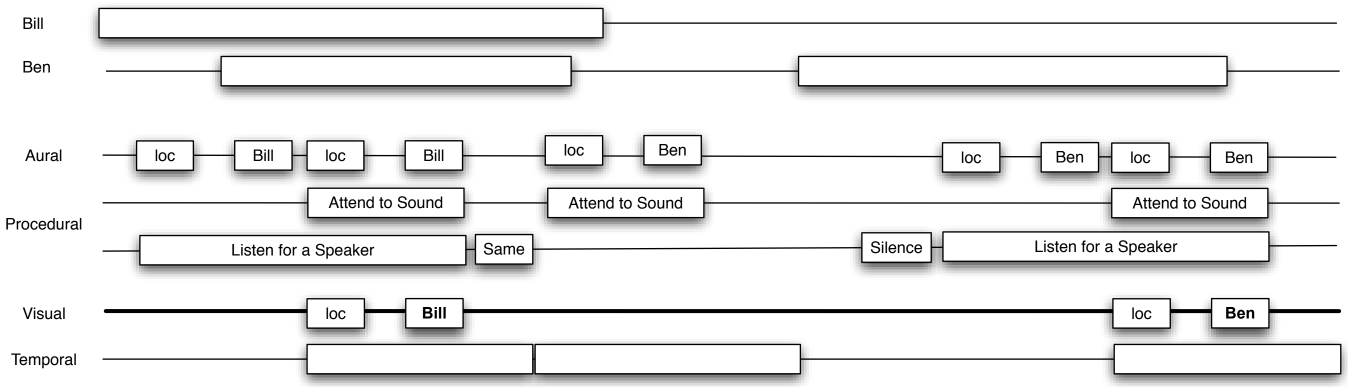

Figure 5: Time-Graph of a sample run of the conversation tracking model.

uous reevaluation of visual and auditory inputs and their correspondence, as well as awareness of any change in auditory source (e.g., which speaker is talking) over time (e.g., the cocktail party effect). We have modeled conversation tracking using the ACT-R/E architecture.

4.1 ACT-R/E Model

Within ACT-R, module operations occur in parallel, but production firing occurs serially. The implication of this is that, like a person, ACT-R can look and listen to two different things at any instant. It can only think about one thing at a time, though. Usage of the aural and visual buffers typically consists of finding a sound (or visual object), attending to it, and then retrieving some information about it to further process it. This is called a "find-attend-harvest" cycle in ACT-R.

A very simple model hears a sound which goes into the audicon. The existence of a sound in the audicon then causes a bottom-up production to fire that notices that sound and localizes it in its environment. Next, a production fires that attends to the sound with further processing as appropriate (e.g., recognizing that a command has been spoken). Finally, a production fires that changes visual attention to the spatial location of the sound. Visual attention is shown to the user by having the face look at the spatial location; this is controlled by a request to the personal buffer. This simple model results in behavior that is extremely distracted: attention jumps from sound to sound as it enters the audicon (default ACT-R parameters like *recency* and *nearest* control what sound is selected from the audicon).

To achieve a more natural degree of attention switching during a conversation, we added two components to our model: a temporal-based mechanism and a more goal-directed component. Because there is typically a 450-650ms delay between speakers [7], the model does not switch visual attention if the last speaker has said something in that timeframe. To implement this we used ACT-R's temporal module which waits for approximately 500ms of silence from the attended speaker before it will switch to a different speaker, resetting the temporal module every time a new word from the original speaker is attended to.

Our model also implements an intentional version of the basic model with an aim to locate a speaker *(listen-for-a-speaker)* in addition to the basic auditory information processing *(attend-to-a-sound)*. Since the *attend-to-a-sound* process makes no use of the visual module, it can run in parallel with the *listen-for-a-speaker* process. This allows the visual module to focus visual attention on the speaker while listening to (and for) other sounds and words in the environment. If the current speaker stops talking for half a second, the system actively looks for another sound, which then starts the entire cycle over again. An example trace of the model is shown in Figure 5.

In the current version, we do not process anything that is actually said; there are no semantics or pragmatics built into the existing system (though the model can start a new conversation if an especially long time passes without any aural input). This is clearly a weakness of the current model, but because the model has such a large bottom-up component to it, we expect to be able to process semantic information in future work.

In summary, our model, based on the ACT-R/E architecture, makes use of the auditory information to direct visual attention. In order to achieve the degree of gaze control exhibited by humans, we have provided some intentional controls over the task, allowed for anticipation of further input from the speaker, and accounted for the perception of time. With minor modifications, our model is capable of responding to high saliency words (e.g., its own name).

This model acts quite natural in informal testing: if a speaker is talking, visual attention (and George's gaze) focuses on the speaker. If someone attempts to interrupt the speaker, visual attention does not change. Once the current speaker stops talking for half a second or so and a new person starts speaking, the model moves its visual attention (and gaze) to the new speaker. We explore how natural naive participants view our system in the following section.

5. EMPIRICAL EVALUATION

Our goal in this experiment was to evaluate the naturalness of our conversation tracking system. Our hypothesis was that a system that acted more human would appear to be more natural to participants.

5.1 Method

In this experiment, we were interested in exploring our conversation tracking system in both absolute and relative terms. Toward this end, we filmed two different scenarios and asked participants to report both a "naturalness" score for both films and to make an explicit choice of which scenario was the most natural. By using a Likert rating scale of naturalness, we could see how natural each system was in an absolute sense. By using a ranking task, we could determine which of the two systems was seen as more natural. Note that by using both methods we can determine several aspects which are not available by using either method alone. For example, if we used only rankings, people may prefer one or the other, but think that both are quite bad overall. Similarly, by using a Likert scale, it is possible that people could see both behaviors as equally natural, but actually have a preference for which is slightly more natural (in direct comparison). We have used a variation of this method in previous evaluation work [40].

5.1.1 Participants

Fourteen participants from George Mason University participated for course credit.

5.1.2 Materials

Both scenarios were filmed in our robot laboratory. Each video lasted 17 seconds. For both conditions, two speakers talked to each other. Their physical location was slightly in front of and on either side of George, the robot (see Figure 1). The "attentive" system used the model described above in which the model waited approximately 500ms without hearing a sound from the current speaker before switching to a different speaker. The "distracted" system used exactly the same model and sensor system, but did not wait 500ms between speakers, instead switching any time it heard a relevant noise. Recall that Bull and Aylett showed that people typically wait approximately 500ms before switching speakers [7]. The effect of changing the switch interval from 500ms to 0ms was that the model appeared highly distracted, switching any time someone made any backchannel comments like "uh-huh" or made some other noise. Note that all other components of the system were identical, including the threshold for sound to enter ACT-R/E's audicon, the physical location of the speakers and the robot, the visual and auditory localization software, and all associated parameters. Thus, any differences between conditions can not be due to sensor level information, but only due to cognitive differences.

5.1.3 Procedure

Participants sat at a computer desk. They were told that they would see two people having two conversations about George the robot. They then watched both videos; order was counterbalanced across participants. Videos were not labeled in any way (they were referred to during the questions as first or second). They were then asked to rate how natural each video was on a scale of 1 (completely unnatural) to 7 (completely natural). After performing this rating task, participants were asked which one of the two videos was more natural, and then to give a few comments describing why it was more natural. Note that participants did not know they were going to be asked to choose which system was more natural when they performed the rating.

5.2 Results and Discussion

First, we examined what people thought of the "attentive" and "distracted" models in an absolute sense using Likert-scale ratings. Consistent with our hypothesis, participants rated the attentive model (Mean = 4.5, SD=1.5) as much more natural than the distracted model (Mean = 3.0, SD=1.2), $F(1, 13) = 9$, $MS_e = 1.75$, $p < 0.05$. This result shows that people not only thought that the attentive model was more natural than the distracted model, but that, in an absolute sense, the attentive model was perceived as very natural and above the mid-point of 4 on a 7-point scale. In contrast, the distracted model was seen as very un-natural and below the mid-point.

Second, we examined which of the two models was seen as more natural. 10 of 14 participants (71%) thought that the attentive model was more natural than the distracted model; this difference was significant via Wilcox signed rank test, $V = 55, p < 0.05$.

Fortunately, both the absolute and relative analyses in this experiment showed consistent results: people thought that the attentive model was more natural than the distracted model. Probably the most important finding in this study was that people thought that, in an absolute sense, the attentive robot was quite natural.

6. CONCLUSION

Previous research has used gaze tracking to track conversations. We suggested that, as good a cue as gaze tracking is for conversation tracking, a better cue for who is talking is the physical (spatial) location of the sound source (i.e., the speaker).

We believe that the best way to utilize this insight is to build computational cognitive models that think and act in the same manner that people do. We have modified ACT-R [1, 2] and created a novel cognitive robotic architecture which we call ACT-R/E.

ACT-R/E enhances the traditional ACT-R visual and aural modules to allow the system to find and localize both people and sounds using our on-board robotic sensors.

We used ACT-R/E to create a model that integrated both aural location information with visual location information to find the person that is talking. The model uses primarily bottom up (perceptual) rather than goal-related cues to control which productions will fire. As long as the same person is talking, the model continues to look at the speaker. However, once a 500ms break in the conversation occurs and another person talks, the model switches visual attention. This decision of when and who to switch to is a critical component of the model and comes directly from psychological research. The results of the model's visual attention are shown on a computerized face. The overall model looks at the person who is talking and switches when someone else speaks, as long as the original speaker is silent. The model itself can be characterized as quite attentive.

We compared our attentive model to a model that appears much more distracted. Our empirical results showed that people not only preferred the more attentive system in a relative sense, but actually thought that the more attentive system was highly natural in an absolute sense as well.

The research reported here suggests that the auditory location of a speaker can, in fact, be used to influence (though not control) visual attention. While we did not use gaze direction at all (and in fact it is not that useful as a conversation cue in dyadic conversations), it would be interesting to incorporate gaze direction into our system to make it even more robust than our current system.

Our conversation tracking system shares some similarities with other researchers who have integrated visual and auditory information [9, 17, 27]. All of these projects attempt to automatically integrate vision and auditory sound localization information and do an excellent job. Our system uses similar computational techniques for audition and vision, but we do our integration at the cognitive level within ACT-R rather than at or just after the sensor level, as other systems do. Our system also correlates 360° auditory tracking with 360° person tracking. While people clearly can not see 360° around them, we use the co-occurrence of sound and person to reduce the number of errors, making the assumption that only people can talk. Other researchers have only used front-back auditory information [27]. Our system also has a principled method of deciding when to change visual focus of attention (i.e., when the current speaker has stopped talking for approximately 500ms). Finally, our system works in a dynamic environment, with speakers moving around.

ACT-R/E is also a substantial improvement over our previous effort that used ACT-R for hide and seek [43]. In that report, we took sensor information, converted it into symbols, and put it directly into declarative memory. By putting it into declarative memory, we created several issues that were difficult to overcome. First, from a cognitive point of view, it is not plausible that sensor information goes directly into memory. Second, once something enters into declarative memory, it is subject to decay, forgetting, interference, and other memory issues. Also, the fact that only one object at a time can enter any particular buffer makes it difficult or impossible to perform human-level reasoning.

In the current system, we are true to the architecture because we take the information from our visual and auditory sensors and put it into the appropriate ACT-R module. We are then are able to use and reason with that information within ACT-R's architectural constraints. This is one of the ways that allows our models to maintain cognitive plausibility, though further testing is needed to confirm that our model is plausible.

One of the strengths of our current method is that we took both visual and auditory streams separately and then combined them at the cognitive/perceptual level within ACT-R rather than at the sensor level. This integration was what allowed us to find the person that was speaking and form a unified percept of the speaking person, which we then used to direct the system's visual attention. This type of principled (not ad-hoc) integration is one of the many strengths that using a computational cognitive architecture gives us. Another advantage is that as we build additional cognitive models, they can be integrated into a coherent whole, providing cognitively plausible, multi-capability HRI.

7. ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research under work order number N0001406WRZ0156 and N0001406-WXZ0001. The authors would like to thank Bill Adams for his care and feeding of the robots as well as help with filming for the experiment as well as Derek Brock for consultation on the spatial localization software. We would also like to thank Alan Schultz, Bill Kennedy, and Sarah Kriz for comments on a previous draft. The views and conclusions contained in this document should not be interpreted as necessarily representing the official policies of the U. S. Navy.

8. REFERENCES

[1] J. Anderson. *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press, 2007.

[2] J. Anderson, D. Bothell, M. Byrne, S. Douglass, C. Lebiere, and Y. Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060, 2004.

[3] M. Argyle and M. Cook. *Gaze and mutual gaze.* Cambridge University press, Cambridge, MA, 1976.

[4] M. Argyle and J. Graham. The Central Europe Experiment-looking at persons and looking at things. *Journal of Environmental Psychology and Nonverbal Behaviour*, 1(1977):6–16, 1977.

[5] B. Best and C. Lebiere. Cognitive agents interacting in real and virtual worlds. In R. Sun, editor, *Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation*, pages 186–218, New York, NY, 2006. Cambridge University press.

[6] C. Breazeal, G. Hoffman, and A. Lockerd. Teaching and working with robots as a collaboration. *Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004. Proceedings of the Third International Joint Conference on*, pages 1030–1037, 2004.

[7] M. Bull and M. Aylett. An analysis of the timing of turn-taking in a corpus of goal-oriented dialog. In *Proceedings of the IC-SLP'98*, volume volumne 4, pages 1175–1178, Sydney, Australia, 1998.

[8] L. Cepstral. Swift TM: Small Footprint Text-to-Speech Synthesizer, 2005.

[9] C. Choi, D. Kong, S. Lee, K. Park, S. Hong, H. Lee, S. Bang, Y. Lee, and S. Kim. Real-time audio-visual localization of user using microphone array and vision camera. *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 1935–1940, 2005.

[10] H. Clark. *Using Language*. Cambridge University Press, 1996.

[11] N. Fay, S. Garrod, and J. Carletta. Group discussion as interactive dialogue or as serial monologue: the influence of group size. *Psychol Sci*, 11(6):481–6, 2000.

[12] B. R. Fransen, V. I. Morariu, E. Martinson, S. Blisard, M. Marge, S. Thomas, A. C. Schultz, and D. Perzanowski. Using vision, acoustics, and natural language for disambiguation. In *HRI*, pages 73–80. ACM, 2007.

[13] T. Kanda, H. Ishiguro, and T. Ono. Development and evaluation of an interactive humanoid robot robovie. *Proc. Intl Conf Robotics and Automation*, 2002.

[14] M. Katzenmaier, R. Stiefelhagen, and T. Schultz. Identifying the addressee in human-human-robot interactions based on head pose and speech. *Proceedings of the 6th international conference on Multimodal interfaces*, pages 144–151, 2004.

[15] W. G. Kennedy, M. Bugajska, M. Marge, W. Adams, B. R. Fransen, D. Perzanowski, A. C. Schultz, and J. G. Trafton. Spatial representation and reasoning for human-robot collaboration. In *The Proceedings of the AAAI national Conference on Artificial Intelligence*, 2007.

[16] D. Kieras and D. E. Meyer. An overview of the epic architecture for cognition and performance with application to human-computer interaction. *Human Computer Interaction*, 12:391–438, 1997.

[17] H. Kim, J. Choi, and M. Kim. Speaker Localization among multi-faces in noisy environment by audio-visual Integration. *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 1305–1310, 2006.

[18] H. Kozima, C. Nakagawa, and H. Yano. Can a robot empathize with people? *Artificial Life and Robotics*, 8(1):83–88, 2004.

[19] J. E. Laird, A. Newell, and P. S. Rosenbloom. SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33:1–64, 1987.

[20] C. Lalanne and J. Lorenceau. Crossmodal integration for perception and action. *Journal of Physiology-Paris*, 98(1-3):265–279, 2004.

[21] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Proc. IEEE International Conference on Image Processing ICIP 2002*, volume 1, pages 900–903, 2002.

[22] E. Martinson and D. P. Brock. Improving human-robot interaction through adaptation to the auditory scene. In *Proceedings of the 2007 conference on Human Robot Interaction*, pages 113–120. ACM Press New York, NY, USA, 2007.

[23] E. Martinson and A. C. Schultz. Auditory evidence grids. In *Proc. of the International Conference on Intelligent Robots and Systems (IROS)*, pages 1139–1144, 2006.

[24] L. Meltzer. Interruption Outcomes and Vocal Amplitude: Explorations in Social Psychophysics. *Journal of Personality and Social Psychology*, 18(3):392–402, 1971.

[25] A. Newell. *Unified theories of cognition*. Harvard University Press, Cambridge, MA, 1990.

[26] G. Nielsen. *Studies in self confrontation*. Monksgaard, Copenhagen, 1962.

[27] H. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi, and H. Kitano. Human-robot interaction through real-time auditory and visualmultiple-talker tracking. *Intelligent Robots and Systems, 2001. Proceedings. 2001 IEEE/RSJ International Conference on*, 3, 2001.

[28] J. Ou, L. Oh, S. Fussell, T. Blum, and J. Yang. Analyzing and predicting focus of attention in remote collaborative tasks. *Proc. of the 7th international conference on Multimodal interfaces*, pages 116–123, 2005.

[29] J. Ou, Y. Shi, J. Wong, S. Fussell, and J. Yang. Combining audio and video to predict helpers' focus of attention in multiparty remote collaboration on physical tasks. *Proc. of the 8th int. conference on Multimodal interfaces*, pages 217–224, 2006.

[30] F. Parke and K. Waters. *Computer facial animation*. AK Peters, Ltd. Natick, MA, USA, 1996.

[31] P. Qvarfordt and S. Zhai. Conversing with the user based on eye-gaze patterns. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221–230, 2005.

[32] S. Raidt, G. Bailly, and F. Elisei. Basic components of a face-to-face interaction with a conversational agent: mutual attention and deixis. *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, pages 247–252, 2005.

[33] F. Ritter, D. Van Rooy, R. Amant, and K. Simpson. Providing user models direct access to interfaces: an exploratory study of a simple interface with implications for HRI and HCI. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 36(3):592–601, 2006.

[34] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn taking for conversation. *Language*, 50(4):696–735, 1974.

[35] B. Scassellati. Theory of Mind for a Humanoid Robot. *Autonomous Robots*, 12(1):13–24, 2002.

[36] A. Schultz, W. Adams, and B. Yamauchi. Integrating Exploration, Localization, Navigation and Planning with a Common Representation. *Autonomous Robots*, 6(3):293–308, 1999.

[37] C. Sidner, C. Kidd, C. Lee, and N. Lesh. Where to look: a study of human-robot engagement. *Proceedings of the 9th international conference on Intelligent user interfaces*, pages 78–84, 2004.

[38] R. Simmons, W. Adams, A. Atrash, M. Bugajska, M. Coblenz, M. MacMahon, D. Perzanowski, I. Horswill, R. Zubek, D. Kortenkamp, et al. GRACE: an autonomous robot for the AAAI Robot challenge. *AI Magazine*, 24(2):51–72, 2003.

[39] N. Taatgen, H. van Rijn, and J. Anderson. An Integrated Theory of Prospective Time Interval Estimation: The Role of Cognition, Attention, and Learning. *Psychological Review*, 114:577–598, 2007.

[40] J. Trafton, A. Schultz, M. Bugajska, and F. Mintz. Perspective-taking with robots: experiments and models. *ROMAN 2005*, pages 580–584, 2005.

[41] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz. Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man, and Cybernetics*, 35(4):460–470, 2005.

[42] J. G. Trafton, A. C. Schultz, N. L. Cassimatis, L. M. Hiatt, D. Perzanowski, D. P. Brock, M. D. Bugajska, and W. Adams. Cognition and multi-agent interaction: From cognitive modeling to social simulation. In R. Sun, editor, *Communicating and collaborating with robotic agents*, in press.

[43] J. G. Trafton, A. C. Schultz, D. Perzanowski, W. Adams, M. D. Bugajska, N. L. Cassimatis, and D. P. Brock. Children and robots learning to play hide and seek. In A. C. Schultz and M. Goodrich, editors, *Proceedings of the 2006 ACM conference on HRI*. ACM Press, 2006.

[44] R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 301–308, 2001.

[45] Y. Yoshikawa, K. Shinozawa, H. Ishiguro, N. Hagita, and T. Miyamoto. Responsive robot gaze to interaction partner. *Proceedings of robotics: Science and systems*, 2006.