Two Ingredients for My Dinner with R2D2: Integration and Adjustable Autonomy

Dennis Perzanowski, Alan C. Schultz, Elaine Marsh, and William Adams

Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory
Codes 5512 and 5514
Washington, DC 20375-5337
< dennisp | schultz | marsh | adams > @aic.nrl.navy.mil

Abstract

While the tone of this paper is informal and tongue-in-cheek, we believe we raise two important issues in robotics and multi-modal interface research; namely, how crucial integration of multiple modes of communication are for adjustable autonomy, which in turn is crucial for having dinner with R2D2. Furthermore, we discuss how our multi-modal interface to autonomous robots addresses these issues by tracking goals, allowing for both natural and mechanical modes of input, and how our robotic system adjusts itself to ensure that goals are achieved, despite interruptions.

Introduction

The following situation should sound familiar to most, if not all, of us. You received your monthly credit card statement, and you have a question about something on the bill. So, you call Customer Service. Once connected, you are asked to press or say a number based on your request. After listening to your various options, you hear the appropriate number to press for Customer Service, and so you either press the telephone keypad corresponding to the number or you say it. This connection is made and you are asked to listen to another series of menu items and numeric choices. This may go on through several levels, but your menu-driven journey is still not over.

A pre-recorded voice asks you to punch in your credit card number. (Incidentally, the scenario is similar for utility bills, telephone shopping, etc.) You may be asked to verify your number, but once connected to Customer Service, a human asks for your credit card number. You may wonder why you have to repeat your number after you've mechanically punched it in. So much for interacting with an "intelligent" system..

You probably expected the Customer Service representative to have all that information and more on his/her screen, since you entered a very personal piece of information--your credit card number. But the representative did not have it. Where did it go? What did the system do with it? Surely, you were interacting with something when you punched in all those numbers previously, and hopefully, your credit card number did not go off into the ether when the system seemingly became dumber than dirt.

While this interaction may seem different from the kind of interactions one might have with autonomous robots, we would argue that the brief scenario outlined above involved a machine interface, and whether or not that machine is robot-like or simply a terminal, or even a telephone, on someone's desktop, you expect the interface to interact in certain intelligent ways.

One probably expects the interface to be easy to use, and for the system to be sufficiently developed that its various components are integrated in some intelligent fashion to produce intelligent results. From an architectural point of view, you probably would like one module to know what another module is doing, and not have to repeat yourself in order to get an appropriate reaction out of the system. Furthermore, one might even want the system to be a bit more intelligent than simply a slave to commands and actually interact with the human user on a more sophisticated level. This latter point projects us to a discussion of levels of independence, autonomy, and cooperation on the part of the system.

To achieve any level of independence, autonomy, and/or cooperation between humans and robots in completing a task, the system should allow either humans or robots to be the originators of goals and motivations. We refer to such systems as *mixed-initiative systems*.

In the context of mixed-initiative systems, *adjustable autonomy* is a critical requirement. Systems exhibiting this feature permit participants to interact with dynamically varying levels of independence, intelligence, and control. In these systems, human users and robots interact freely and cooperatively to achieve their goals. There is no master/slave relationship. Participants may adjust their level of autonomy as required by the current situation.

This requires that participants are aware of what the goal is and how each can contribute to achieve that goal effectively.

Our research addresses the case of human-robot interactions that require close interaction. To achieve success in such situations, we have been employing natural modes of communication, such as speech and gesture in our interface (Perzanowski, Schultz, and Adams 1998). As interface development progressed, we saw a need to integrate the language and gestural capabilities of the interface by tracking goals in human/robot interactions (Perzanowski et al 1999). We argued that tracking goals provided us with a means of achieving varying levels of autonomy. Recently, we included a mechanical means of communication with the robots via palm devices (Perzanowski et al. in review). We have expanded the kinds of interactions permitted in our interface and have argued that these added capabilities require the various components to be tightly integrated in order to achieve success. Our research has brought us to the conclusion that integration of multiple modes of communication affects adjustable autonomy.

Hey, what channel are you communicating on now?

Every mouse-click, every menu pulled down and item highlighted, every spoken command uttered to a natural language interface, is an act of communication of one sort or another. When we communicate with humans, we use certain channels of communication, and when we interact with machines, we may use the same or similar ones, or even different channels. For example, we don't mouse-click to a friend, or pull down a menu item to ask someone to get us a glass of water.

However, we do talk to people, gesture to them, point at items and locations in the real world, frown, laugh, accompany our speech with repetitive gestures to indicate underlying emotions--to name but a few of the channels of communication open to humans when they interact. Since we are constructing humanoid robots and even non-humanoid robots that must interact with humans in various ways, then we expect that at least some of these channels of communication are available to humans interacting with those robots. In other words, natural channels of communication must be available for humans interacting with both humanoid and non-humanoid robots, simply because we, humans, are trying to communicate with the robot. If you're trying to "communicate" with a machine that has some sort of anthropomorphic characteristics, shouldn't that machine "communicate" back at you in ways with which you are familiar and comfortable? Furthermore, even if the robot doesn't look at all human-like, but is just a screen with a speaker in front of it, humans will speak naturally to the system because a particular channel--a speech channel--has been provided for interaction.

One of the elements in Grice's maxims of communication theory (Grice 1957) is felicity. He includes it in one of his postulates attempting to explain how human communication is possible. Basically, human communication is effective when it embodies, among other characteristics, felicity, which he characterizes as the aptness and ease of expression incorporated in human communication.

It seems only logical, therefore, that if we are extending human communication to incorporate humanoid robots, then some of speech act theory must be incorporated in the interchange, which is embodied specifically in the interface. Of course, if you don't want to communicate with a machine as if it were almost human, that's another story. But if you provide a human-like channel of communication, let's say a speech channel, for example, people will probably wind up talking to the machine as if it were human or at least human-like. Furthermore, if you make this channel easy to communicate on, people will have to worry less about how to use the channel and simply interact with the system. Similar reasons prompted our work on incorporating natural gesture in our interface, contrary to using stylized gestures, as in (Kortenkamp, Huber, and Bonasso 1996).

We are not saying, however, that human-machine interaction must in some way mimic human-human communication. Providing a keyboard for human-machine interaction does not conjure up the same notions of interaction at all. We are simply saying that if you provide a human-like channel of communication, and indeed if you make the robot look like a human, then it better have some of the capabilities of interacting with a human like a human.

Finally, we believe that felicity in communication is achievable in a system whose communication capabilities are integrated. Our reasoning here is rather simple: if it's put together simply and all of its various components are fully integrated in a way that is compatible with human channels of communication, then communicating with it in whatever fashion we choose--in a way that is felicitous--should be simple. We believe those channels are things like pointing, gesturing, talking, etc.

Does R2D2's right hand know what its speech module is saying?

To achieve the kind of integration we are talking about, we have been designing and implementing a multi-modal interface to autonomous robots. In our research, we have been using Nomad 200s, XR-4000s, and an RWI ATRV-Jr. For a schematized overview of our system, see Figure 1 next page.

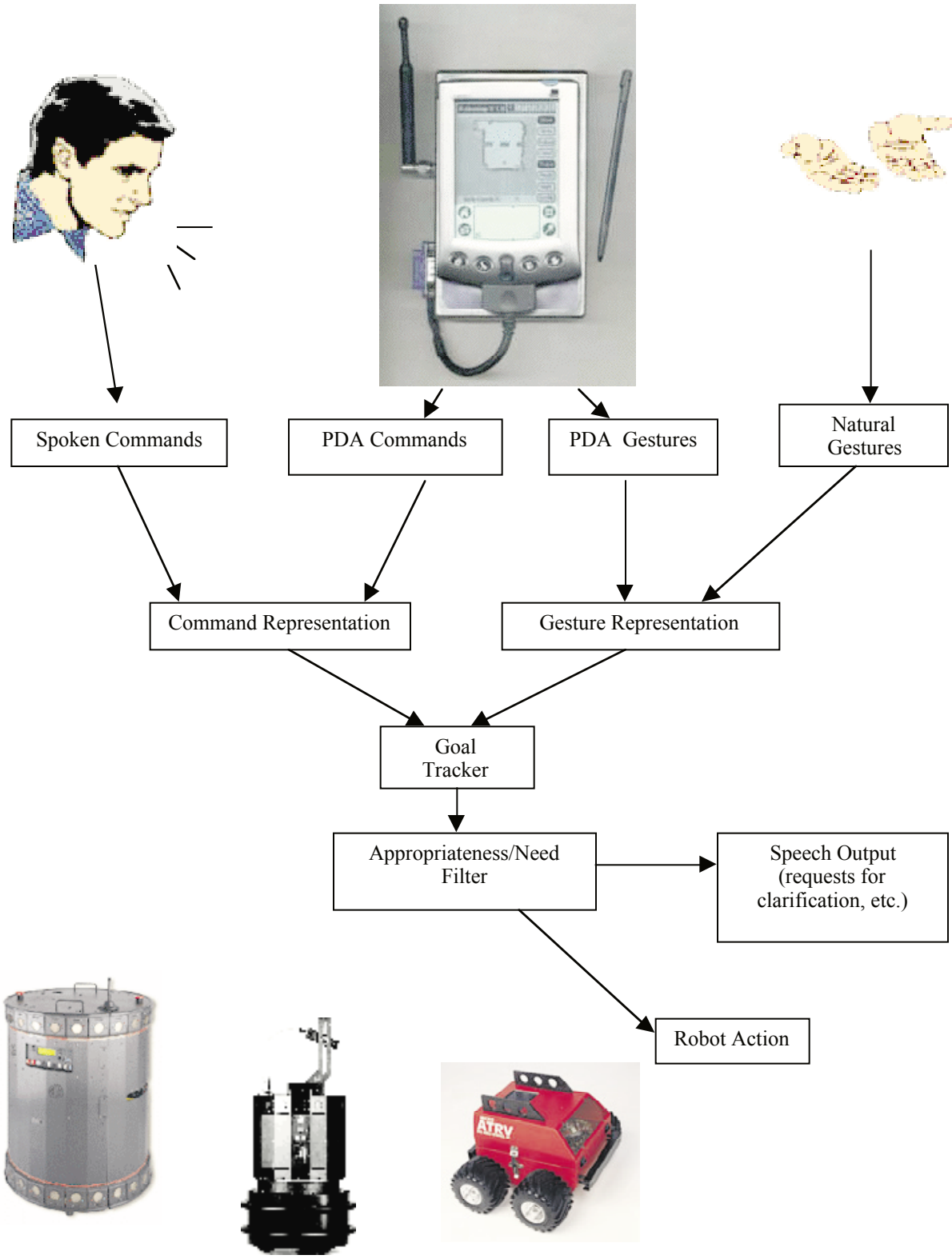

Figure 1: Schematic diagram of multi-modal interface

The robots understand speech, hand gestures, and input from a hand-held PDA, a personal digital assistant (in this case, any of the Palm devices). Speech is initially processed by a speech-to-text system (IBM's ViaVoice), and our natural language understanding system, Nautilus (Wauchope 1994), robustly parses the language input and translates it into a semantic representation which is then mapped to a command after gestural information is incorporated.

Gestures can be either distances, indicated by holding the hands apart to indicate a distance, or directions, indicated by tracing a line in the air.

Natural gestures are detected using a structured light rangefinder which emits a horizontal plane of laser light 30 inches above the floor. A camera fitted with a filter tuned to the laser wavelength is mounted on its side. Given that the laser and camera mount are at a right angle, and the camera is tilted a fixed amount, the distance to a laser-illuminated point can be easily triangulated. With the sensor, the robot is capable of tracking the user's hands and interpreting their motion as vectors or measured distances.

The palm device, which dynamically presents an adaptive map of the robots' environment, can be used to give certain commands to the robots. Users can tap on menu buttons on the device's touch screen, or gesture (by tapping or dragging across a map of the environment on the PDA screen) to indicate places or areas for the robots. The map on the display comes directly from the robot via a mapping and localization module (Schultz, Adams, Yamauchi 1999).

A system's level of integration can be seen as a function of how well the various modules share information in order to complete a task. Therefore, if the speech module needs information from the gesture module and can obtain that information readily, we would say that this system exhibits a greater degree of integration than one in which the modules do not have ready access to another module's information. The payoff for having information shareable is that the user can concentrate on communicating with the system, not with *how* to communicate with it. The system can get whatever information it needs, because it is available. Interacting with R2D2, therefore, should be easy: simply interact with it naturally and let the system do the work of putting all the pieces together.

We have been building a human-robot interface that allows for natural as well as mechanical interaction. We have constructed the interface so that it is responsible for integrating the information for the various input modalities. Users are free to interact with it as they see fit.

Our research on the natural language and gestural interface is based upon the premise that people communicate with other people easily and they use natural language and physical gesturing to do so, among other channels of communication, such as facial expression. We, therefore, assumed that people might readily interact with autonomous robots in much the same fashion; namely, by using speech and body gestures.

Of course, the number of channels of communication that are available in human communication are much more numerous than those we have elected to concentrate on here; we have limited our considerations to basically two, spoken natural language and body gestures of two types which we outline below, in order to determine the empirical consequences of using these two channels of communication in a human-robot interface.

Talking, Gesturing and Tracking

When people talk, they gesture. Some of those gestures are meaning-bearing, while others are superfluous, some redundant, and some indicate an emotional or intentional state of the speaker. We limit ourselves to the meaning-bearing gestures that disambiguate locative elements of spoken natural language. We do not consider other body movements, such as facial expression or head movements, at this time. We limit ourselves to hand and arm movements. Gestures can be made by pointing or gesturing to objects and locations in the real world, or by interacting with a PDA display that represents the same environment.

Furthermore, we have concentrated on two types of deictic gestures: natural and, what we call, synthetic ones (Perzanowski, et al. in review). Natural gestures we consider to be those made by natural movements of a person's arm and/or hand. Synthetic gestures are those made by pointing and clicking on a mechanical device, such as the touch screen of a PDA held in a user's hand.

In our interface, actions that the human user wishes to communicate to the robot are entered either verbally or by means of buttons on the PDA touch screen, and the system translates this input into domain predicates that are stored and the actions themselves are noted as either being completed or not. We are keeping a record of the actions or goals of the interactions so that we can address the issue of adjustable autonomy, to which we now turn.

Haven't you finished emptying the garbage yet?

Autonomous robots, equipped with sufficient knowledge, should be able to go off on their own and complete actions without the human having to intervene at every step.

For example, in attempting to deliver a letter to some office, a robot has to open a certain door in order to proceed. Once having gotten through the door on its own, the robot delivers the letter; however, for some reason the door was left open. If the human user should note this and tell the robot "You left the door open," the system should know what door is being referred to--not necessarily the one immediately in the current environment.

Granted, the goal of opening the door was achieved when that action was encountered, but the robot needs to know that "the door" being referred to is the one referred to in a former achieved goal. It should not attempt to close a door sensed in its immediate vicinity, or even query the

user about which door is being referred to. It should know what to do based upon its previous actions.

Recent work in planning (Grosz, Hungsberger, and Kraus 1999), indicates that a planning component is necessary for collaborative work between multiple agents. Collaboration entails team members adjusting their autonomy through cooperation. While we will not discuss the intricacies of a planning module here, we have started to incorporate some elements of planning by attempting to use natural language dialog and goal tracking as a planning activity. Thus, the natural language input, along with the gestural input, is incorporated into a history list, which keeps track of what goals have or have not been attained.

Thus, for example, if the user tells the robot to explore a particular area of a room but the robot is for some reason interrupted in the completion of the task, it will still be able to complete its prior task after the interruption, whenever it becomes feasible, based on nothing more than a simple command from the human to continue what it was doing. Likewise, with a team of robots, the task might be assigned to another available robot, simply by telling the second robot to pick up where the first robot left off. The human need not have to remember what the robot specifically was doing prior to the interruption, although the human could query the robot, if necessary, just to make sure that the robot was doing something worthy of continuation.

Robots in the team, using the list of goals and the record of which goals have been achieved or not, can use the information to determine exactly what is going on, and can adjust their own activities accordingly, based on the list of goals, their importance to the overall task, future re-directs, and to periodic interruptions.

The goals that we refer to here are actions, commands, directives, that are part of a dialog between humans and robots acting as a team, and they are incorporated in a planning component to achieve the kinds of cooperation and interaction needed for complex collaborative acts. Coupled with other dynamic factors, such as a changing environment (Pollack and Horty 1999), our dialog-driven planning component fosters the kind of adjustable autonomy we argue for, since it is based on immediate need--goals expressed in the dialog.

By generating plans from goals, and prioritizing them, almost on the fly as it were, the robotic system can achieve the kinds of coordination only obtainable by systems internally adjusting and cooperating with other systems that are themselves adapting to their role in a team and to a changing environment .

So how's your Aunt Sally?

When humans interact with each other on a daily basis, certain elements of their daily interactions become mutually understood. Teams are built. Co-operation develops between members of the group. Connections are made; for example, if a person is observant enough or cares (let's not get into inter-personal dynamics here), that person might know how another team member likes his or her coffee; remembers that one of the team member's Aunt Sally was sick the day before and as a result inquires about Aunt Sally's health and offers a cup of coffee with sugar and no cream for the co-worker at the team members' next meeting.

Also, people learn what their roles are in performing their daily work with other team members. People know that one team member's strength is to perform a certain function, such as debugging code. Another team member's strength might be to write research papers. Team members learn how to complement each other in their work. They do not necessarily ask what and how something should be done. They fit together and accomplish tasks based on the overall objectives given to them, and their individual strengths in performance. And humans seem somehow to learn this by working together in a group for a while. Much of this knowledge comes out of a tacit understanding based on the daily interactions of the team members as they become the unit.

If we are building robots that are going to interact with us in a similar way as other humans do, then co-operation and team interaction are going to be expected of our robot team members as well (Wilkes et al 1998). Our robot team members are going to have to be observant enough to construct the same kind of tacit understanding of their own roles and the roles of other team members and interact accordingly. They will have to adjust their own autonomy as needs arise and change.

How humans achieve this tacit understanding through observation and interaction is a complex phenomenon in itself. So, when we say that our robot team members must have the same capabilities, we are asking quite a bit. But knowing when to debug code solely, and when to offer a suggestion to a team member requires a skill which is achieved by knowing when to contribute and when to back off. We argue that these are all topics related to adjustable autonomy. Now that we've identified what we mean by the term, we now need to go back to our drawing boards and come up with ways to achieve this goal.

We have already started to consider how knowledge of the goals of a mission, for example, contribute to achieving one's individual autonomy or co-operation. Those tacit skills referred to earlier are much more complex than we characterize them here, but we believe knowledge about the individual goals and goals of the group can be used as constructs for achieving tacit knowledge. And with this knowledge robot team members can adjust their autonomy for easy interaction with their human and other robot team members.

Conclusions

Our work designing and implementing a multi-modal interface to autonomous robots is focussing on two broad research issues. Our research regarding the multi-modal interface is primarily concerned with the integration of the command and gesture modules of our interface. We believe considering integration as the impetus for constructing a multi-modal interface leads to a more

natural and easier-to-use interface. On the robot side of the house, we are focussing on autonomy, trying to construct a system that knows enough about itself, the world around it, and what it has been doing, so that it becomes more of a team player when interacting with humans and other robots. We are focussing our research on building a dialog-driven planner to track goals during human-robot interactions. When necessary, robots should be totally autonomous and carry out their duties and functions; however, they should be adaptable to any situation as it arises, becoming more dependent if necessary, acting closely with their team members as situations may demand.

A system that is firstly, integrated, and therefore more natural and easier to interact with, and secondly, capable of adjustable autonomy, is a much more fitting dinner partner than one that is not.

Acknowledgements

This work is funded in part by the Naval Research Laboratory and the Office of Naval Research.

References

Grice, H.P. 1957. In *The Philosophical Review,* 66: 377-388. Reprinted 1971 in *Semantics: An Interdisciplinary Readier in Philosophy, Linguistics, and Psychology,* Steinberg, D.A. and Jacobovits, L.A., eds. New York: Cambridge University Press.

Grosz, B., Hunsberger, L., and Kraus, S. 1999. Planning and Acting Together. *AI Magazine* 20(4): 23-34.

Grosz, B. and Sidner, C. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 12(3):175-204.

Kortenkamp, D., Huber, E., and Bonasso, R.P. 1996. Recognizing and Interpreting Gestures on a Mobile Robot. In Proceedings of the Thirteenth National AAAI Conference on Artificial Intelligence, 915-921. Menlo Park, CA: National AAAI Conference on Artificial Intelligence.

Perzanowski, D., Schultz, A.C., and Adams, W. 1998. Integrating Natural Language and Gesture in a Robotics Domain. In Proceedings of the IEEE International Symposium on Intelligent Control: ISIC/CIRA/ISAS Joint Conference, 247-252. Gaithersburg, MD: National Institute of Standards and Technology.

Perzanowski, D., Schultz, A., Marsh, E., and Adams, W. 1999. Goal Tracking in a Natural Language Interface: Towards Achieving Adjustable Autonomy. In Proceedings of the 1999 IEEE International Symposium on Computational Intelligence in Robotics and Automation: CIRA 1999, 144-149. Monterey, CA: IEEE Press.

Perzanowski, D., Schulz, A., Adams, W., and Marsh, E. in review. Towards Seamless Integration in a Multi-modal Interface.

Pollack, M. and Horty, J.F. 1999. There's More to Life Than Making Plans. *AI Magazine* 20(4): 71-83.

Pollack, M. and McCarthy, C. 1999. Towards Focused Plan Monitoring: A Technique and an Application to Mobile Robots. In Proceedings of the 1999 IEEE International Symposium on Computational Intelligence in Robotics and Automation: CIRA 1999, 144-149. Monterey, CA: IEEE Press.

Schultz, A.; Adams, W.; and Yamauchi, B. 1999. Integrating Exploration, Localization, Navigation and Planning With a Common Representation. *Autonomous Robots* 6(3*):* 293-308.

Wauchope, K. 1994. Eucalyptus: Integrating Natural Language Input with a Graphical User Interface, Technical Report, NRL/FR/5510--94-9711, Navy Center for Applied Research in Artificial Intelligence. Washington, DC: Naval Research Laboratory.

Wilkes, D.M., Alford, A., Pack, R.T., Rogers, T., Peters II, R.A., and Kawamura, K. 1998. Toward Socially Intelligent Service Robots. *Applied Artificial Intelligence* 12:729-766.