The Sixteenth International Symposium on Artificial Life and Robotics 2011 (AROB 16th '11),
B-Con Plaza, Beppu,Oita, Japan, January 27-29, 2011

A multimodal language to communicate with life supporting robots through a touch screen and a speech interface

T. Oka, H. Matsumoto, and R. Kibayashi

College of Industrial Technology, Nihon University, 1-2-1 Izumicho, Narashino, Chiba, 275-8575 JAPAN
Tel : 81-47-474-9693; Fax : 81-47-474-2669
oka.tetsushi@nihon-u.ac.jp

Abstract: This paper proposes a multimodal language to communicate with life supporting robots through a touch screen and a speech interface. The language is designed for untrained users who need support in daily lives from cost effective robots. In this language, the users can combine spoken and pointing messages in an interactive manner in order to convey their intensions to robots. Spoken messages include verb and noun phrases which describe intensions. Pointing messages are given when users finger-touch a camera image, a picture containing a robot body, or a button on a touch screen at hand, which convey a location in their environment, a direction, a body part of the robot, a cue, a reply to a query, or other information to help the robot. This work presents the philosophy and structure of the language.

Keywords: Life-supporting robot, multi-modal language, speech, touch screen, human-robot interaction

I. INTRODUCTION

It is expected that life supporting robots will help people in their daily lives in near future. Such robots must be easy to communicate with those untrained including elderly and disabled, and particularly those people must be able to learn how to communicate with their robot in a short period of time without much effort. On the one hand, although a conventional user interface using pull-down menus and buttons is a cost effective choice, it is difficult to build an easy-to-learn natural interface with life supporting robots that are given many kinds of tasks. Unless users learn many short-cuts, they have to go through a long sequence of choosing among menu items or buttons. On the other hand, it will be a long journey to realize robots that can speak and communicate like humans. Such robots will need sophisticated sensors to collect information and powerful computers to recognize different situations, perceive verbal and nonverbal messages from their counterpart, and disambiguate them based on various knowledge sources[1-3]. Therefore, we predict that it will take decades to develop affordable robots one can convey intentions to in the same way as to humans. Thus, in the near future we will need an *artificial language* for natural communication with robots.

In human face-to-face communications, nonverbal messages play a great role [4]. They can fill in gaps in verbal messages and convey some kind of information more efficiently than explicit words. Since robots are often given physical tasks, nonverbal messages are particularly important for smooth and efficient communication between humans and robots. For this reason, nonverbal and multimodal communications between humans and robots are subject of current research [5-7]. However, there are few arguments on multimodal or nonverbal *languages* between humans and robots.

RUNA [8-12] is the first multimodal language designed for communication between untrained people and cost effective robots. In the language, one can command a robot by combining a spoken message in Japanese and nonverbal message such as a hand gesture, body touch, and button press action. In some recent user studies, novice users were able to successfully command a robot and achieve some tasks without a long training. Many of them preferred multimodal commands to single modal spoken commands. The results of these studies imply that more experienced users would be able to command robots combining verbal and nonverbal messages more efficiently and successfully. In addition, since the language is designed so that one can give a command on a context free basis without ambiguity, one can build a responsive command interpreter at low cost.

Although most of the commands by the novice users were successful, there were some human errors as well as system errors. Many of them failed to give spoken commands with many words which specify a robot action in detail. It looks also difficult for beginners to convey multiple values in a single nonverbal message. For example, those users who commanded a robot simultaneously conveying a speed, direction, and angle in a gesture or button press action thought that it was harder to communicate in the language than the other users. Therefore, the efficiency of RUNA seems to be an obstacle for beginners.

Another problem of the current version of RUNA is that it does not have a means to point at an object or a location. For this reason, users had to direct a robot to a location or an object in an indirect manner using spoken words, buttons and gestures. Direct pointing at locations and objects would make it much easier to convey intentions to life supporting robots.

Based on the advantages and disadvantages of RUNA, this paper discusses a multimodal language in which one can communicate with life supporting robots

through a speech interface and a touch screen, both of which are common and not expensive in recent years. The language allows users to communicate with a robot more slowly and in a more interactive manner than RUNA. Using a touch screen at hand, they can point at a location or an object in a camera image and send cues, pieces of advice and parts of intentions to their robot by tapping on a button. In the following sections, we describe the principles and structures of the language.

II. COMMUNICATION

The multimodal language proposed in this paper will be developed so that untrained novice users should be able to communicate with a life supporting robot without taking much time to learn about the language and their robot. Based on the results of studies of RUNA, we predict that an affordable user friendly life supporting robot will be realized if we develop a good multimodal language. Above all, a touch screen and a speech interface should provide a good means for users of such a robot. For this reason, the language is strategically-designed to utilize such devices. Figure 1 depicts a touch screen showing windows for interactions in the language. The camera image window shows real images from a camera on the robot in real time. Some buttons appear in the button window, which guide users to convey intentions. Users may point at a location or a body part of the robot and use gestures on the screen. The language is designed based on the following *laws*:

1. Users must be able to convey their intensions in an interactive manner.
2. Users must be able to send messages at any time.
3. Users must be allowed to speak in as natural a way as possible.
4. A touch screen must help users as much as possible.
5. Users must be able to point at 3D positions or locations in their environment without difficulty.
6. Users must be able to point body parts of their life supporting robot.
7. Whenever necessary, buttons must appear on the touch screen.
8. Users must be able to use gestures to convey their intensions.
9. The touch screen must display messages to help users and realize smooth communications.
10. Robots must send spoken messages to help users.

III. MULTIMODAL LANGUAGE

A. Semantic Representation

Table 1 exemplifies intensions users can convey to robots in the language. Intensions are identified by their type and parameter values. For instance, an intension to start a robot moving forward slowly is formally represented by a type *move forward* and a speed value *slow*. The robot must identify the type and parameter value by interpreting verbal and nonverbal messages from the user. In our language, users can leave out optional parameter values if they do not care about them; if one does not care about the speed, one does not need mention it. As seen in Table 1, all the types of intension belong to one of the classes, such as *action*, *cue*, and *advice*.

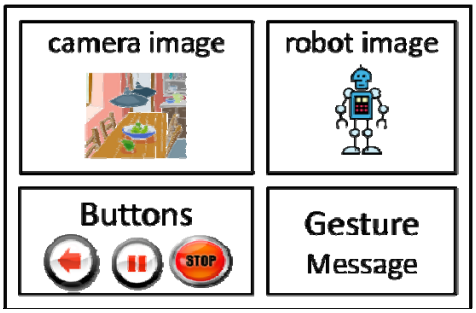

Fig. 1 Touch Screen Windows

Table 1 Intension Representation

Class	Type	Parameters
action	*move forward*	speed(**o**ptional)
	turn/move aside	direction(**m**andatory) speed(o)
	go to	destination(m) speed(o)
	come to	destination(o) speed(o)
	turn to	target position(m)
	pick up	position(m) size(o) hand(o)
	bring	position(m) size(o) hand(o)
	place	position(m)
	push / pull	position(m) size(o) hand(o)
	look	target position(m)
	grasp	hand(m) size(o) speed(o)
	release	hand(m) speed(o)
	rotate wrist	hand(m) direction(m) speed(o)
	move hand	direction(m) hand (o)
	move object	position(m) destination(m) size(o)
	throw away	position(m) destination(m) type(o) size(o)
	vacuum- clean	room(m) location(o) area(o) mode(o)
	device op	device (m) operation(m) value(m)
	report	content(m)
	sing / dance	genre(o)
cue	*start stop*	
	pause	
	resume	
	cancel ok	
advice	*subgoal*	position(m)
	obstacle	position(m)
	landmark	position(m) label(m)
	bail out	direction(o)

B. Interpretation

Since inexperienced users may not be able to express their intensions without effort, our language allows them to send verbal and nonverbal messages that contain partial information. In other words, one can inform a robot of the type and mandatory and optional parameters of their intension bit by bit in separate

messages (Fig. 2). For instance, a spoken message "turn" brings a *left button* and a *right button* on the touch screen to help the user to convey a direction. The robot also sends spoken messages and screen messages to inform the user of what has been received and what information is needed. Unmentioned parameter values are inferred by the robot based on the default values and context. In addition, one can restate a type or parameter value at any time before the robot starts its action.

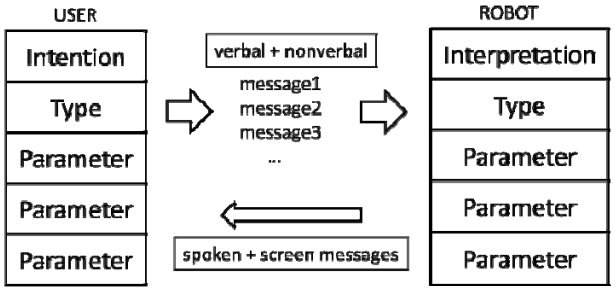

Fig. 2 Communication of Intentions

After conveying a complete intension, namely, the type and mandatory parameter values, one can cue the robot to start by sending a spoken or pointing message. A cancel cue removes received information and stops the current action if being executed. While the robot is executing an action, the user can send a cue, an overriding parameter value, or advice to the robot.

Table 2 Grammar rules

Rule	Description
S → INTENTION	intention
S → INTETION_PARAMS	parameters
S → CUE	cue
S → ADVICE	advice
INTENTION → IP1 IT11	intention type and parameters
INTETION_PARAMS → IP1	parameters(type1)
IP1 → SPEED SHORTSLIENCE	speed + short silence
IP3 → TO_LOCATION	location
TO_LOCATION → HERE TO	to here
TO_LOCATION → KITCHEN	the kitchen
TO_LOCATION → THIS DOOR TO	to this door
IT1 → MOVEFORWARD	move forward
CUE → START	cue to start action

IV. SPOKEN MESSAGES

A spoken message of our language is a sentence, phrase, or word in Japanese to partially convey an intension that can be represented as in Table 1. A sentence includes a word or phrase denoting an intension type, so one can specify the type and one or more parameter values in a sentence like "turn to the left slowly." It is also possible to send a single phrase message that contains only a parameter value or type. Thus, a beginner can convey an intension bit by bit slowly. It is also easy to modify parameter values that have been already sent.

The lexicon of the language includes deictic words such as *kono*/*kore* (this) and *koko* (here). For example, one may give commands like "place the bottle *here*," "turn to *this* quickly," "pick up *this can*," or "move *this here*." Although such words certainly do not help robots to determine parameter values, one can command a robot in a natural manner by combining a pointing message and such a word.

Table 2 shows some examples of grammar rules of our new spoken language. The grammar is similar to RUNA's grammar, but it includes more deictic expressions and less numerical expressions.

V. POINTING MESSAGES

Pointing messages are such messages that are sent through a touch screen. A touch on a camera image conveys a parameter value, e. g. the position of an object or a location in the robot's view. One can touch a button on the screen to send a cue, a parameter value, or a piece of advice. For instance, a touch on a picture of the robot's body creates a nonverbal message containing a location or a body part.

A cue can be sent by a touch on a cancel, start, stop, pause, or resume button. Users can send a cancel cue and restart commanding at any time, terminate or interrupt an ongoing action, and resume an interrupted action by touching a button on the screen. After the type of an intension is identified in a spoken message, buttons appear on the screen to allow the user to choose among parameter values. The user can modify parameter values at any time before the action terminates.

VI. 2D GESTURES

Finger gestures on a touch screen can convey contours, trajectories, directions, speeds, lengths, heights, angles, locations, and symbols. First, using gestures on a camera image and deictic expressions in spoken messages, one can inform a robot a safe route to a location, a hand trajectory to an object, the contour and size of an object, a direction, etc. Next, 2D gestures on a robot picture help users to intuitively communicate directions, speeds, lengths, angles, body motions, heights, etc. Moving a finger across the body of the robot with a spoken message "put the cup up" is quite a natural way to command a robot. Finally, symbolic gestures substitute for or emphasize spoken words. We expect that using these gestures users will be able to communicate more naturally.

VII. INTERACTIONS

This section illustrates some example of interactions between a robot and its user.

Interaction A (moving to a location)

U: Turn…
R: Left or right? (Displays buttons)
U: (Touches on the "right button")
R: OK! (Starts moving and displays new buttons)
U: (Touches on the "slow" button)
R: (Slows down)

U: (Touches on the "stop" button)
R: OK!
U: (Points at the camera image.)
R: (Displays a mark on the point.)
U: Go here!
R: Shall I go here? (Displays a "start" button)
U: (Touches on the "start" button.)
R: OK! (Starts moving and displays buttons.)

Interaction B (moving an object to a location)

U: (Points at an object)
R: (Displays a mark on the object)
U; (Points at a location)
R: (Displays a mark at the point)
U: Move this here!
R: Shall I move this here? (Displays a start button)
…

Interaction C (throwing away a bottle)

U: (Points at a bottle on the floor)
U: Uh, this bottle.
R: This bottle? (Displays a mark)
U: Throw it away!
R: Shall I throw this away? (Displays a "start" button)
…

Interaction D (changing the temperature setting)

U: The, uh, air conditioner…
R: Shall I operate the air conditioner? (Displays buttons)
U: (Touches on the "temperature" button)
R: (Displays the current temperature setting and buttons to change the temperature)
U: (Changes the temperature and cues OK)

VIII. DISCUSSION

In general, direct pointing and speech are among the easiest input methods for most users. Our language allows beginners, who cannot efficiently combine pointing gestures and speech, to convey their intentions slowly and interactively. Using key words and phrases in Japanese, one can inform a robot of the type of an intention without using a keyboard shortcut or selecting among menu items, and then convey parameter values guided by the robot. Experienced users of our language will be able to achieve their goals more quickly by sending spoken and nonverbal messages.

Needless to say, it is difficult to specify locations, objects, and body parts without pointing at or touching them. Spoken commands to life supporting robots tend to be wordy and tongue-twisting [12, 13]. In addition, the use of nonverbal messages instead of verbal messages will often decrease cognitive loads.

IX. SUMMARY AND FUTURE WORK

This paper proposed a multimodal language and a user interface for untrained users of life supporting robots. The language is designed based on experiences and results of some earlier studies on RUNA, and allows beginners to convey their intentions on a more interactive basis. A touch screen will help them to select among options and interact with a robot in an intuitive manner. At this moment, we are developing a test bed for user studies of the new multimodal language. Our future work include user studies of life supporting robots based on the language, design of a multimodal language integrating speech, 3D gesture, body and screen touch, and other modalities.

AKNOWLEDGMENT

This work was supported by KAKENHI Grant-in-Aid for Scientific Research (C) (22500179).

REFERENCES

[1] Prasad R, Saruwatari H, Shikano K (2004) Robots that can hear, understand and talk. Advanced Robotics 18-5:533-564
[2] Jurafsky D, Martin JH (2000) Speech and Language Processing. Prentice Hall
[3] Bos J, Oka T (2007) A spoken language interface with a mobile robot. Journal of Artificial Life and Robotics 11-1:42-47
[4] Knapp ML, Hall JA (2010) Nonverbal Communication in Human Interaction. Wadsworth
[5] Perzanowski D, et. al. (2001) Building a multimodal human-robot interface. IEEE Intelligent Systems, 16-1, pp. 16-21
[6] Iba S, Paredis CJJ, Adams W, Khosla PK (2004) Interactive multi-modal robot programming, The 9th International Symposium on Experimental Robotics (ISER '04), pp. 503-512
[7] Igarashi T (2008) User Interface for Robots, Journal of Robotics Society of Japan, Vol. 28, No. 3, pp.246-248 (in Japanese)
[8] Oka T, Abe T, Shimoji M, Nakamura T, Sugita K, Yokota M (2008) Directing humanoids in a multi-modal command language. The 17th International Symposium on Robot and Human Interactive Communication
[9] Oka T, Abe T, Sugita K, Yokota M (2009) RUNA: a multi-modal command language for home robot users. Journal on Artificial Life and Robotics 13-2: 455-459
[10] Oka T, Abe T, Sugita K, Yokota M (2009) Success rates in a multimodal command language for home robot users. Journal on Artificial Life and Robotics 14-2:219-223
[11] Oka T, Sugita K, Yokota M (2010) Commanding a humanoid to move objects in a multimodal language. Journal on Artificial Life and Robotics 15-1:17-20
[12] Oka T, Abe T, Sugita K, Yokota M (2011) User study of a life supporting humanoid directed in a multimodal language. The 16th International Symposium on Artificial Life and Robotics (AROB11)
[13] Oka T, Sugita K, Yokota M (2009) Spoken command language to direct a robot cleaner. FAN2009