The Fourteenth International Symposium on Artificial Life and Robotics 2009 (AROB 14th '09),
B-Con Plaza, Beppu, Oita, Japan, February 5 - 7, 2009

Multi-modal Command Interface with Remote Home Robots

T. Abe[1], T. Oka[2], K. Sugita[2], and M. Yokota[2]

[1]*Graduate School of Fukuoka Institute of Technology, 3-30-1, Wajiro-higashi, Higashi-ku, Fukuoka, 811-0295, Japan*
[2]*Fukuoka Institute of Technology, 3-30-1, Wajiro-higashi, Higashi-ku, Fukuoka, 811-0295, Japan*
Tel : 81-92-606-4813; Fax : 81-92-606-0754
mdm07001@bene.fit.ac.jp, {oka, sugita, yokota}@fit.ac.jp

Abstract: This paper describes a multi-modal command interface with remote home robots based on a simple interaction model using a multi-modal command language and an animated humanoid communicator for multi-modal reactions. The interface can be realized using a mobile phone or a portable computer, and allows one to give multi-modal commands to a remote home robot through a microphone and buttons looking at pictures from home and the animated humanoid on the screen. It is designed to enable users unfamiliar with computers or robots to check rooms and operate appliances away from home. A test bed for user evaluation has been implemented on PCs based on a multi-agent platform, a speech recognition engine, and a robot simulator.

Keywords: Home robot, multi-modal interface, command language, animated humanoid, human-robot interaction

I. INTRODUCTION

In recent years, home robots for various purposes have been developed, some of which are already in the market. It is predicted that in near future they find their places to work as a new kind of interface with home appliances and achieve various tasks with or without their users at home. There are many kinds of potential tasks for those robots while their users are away from home. It would be desirable if tasks and commands could be given to them by remote users via a mobile phone or a small portable computer.

Some remote command interfaces with remote home robots have been recently developed. Typical examples of robots controlled by such an interface are Banryu and Roborior from Tmsuk (http://www.tmsuk.co.jp/), which can be operated pressing buttons on mobile phones. However, it is difficult to build a usable interface using only a small number of buttons. Another example is BlogAlpha developed by Toshiba [1] which allows users to type and send natural language commands and queries using a web browser. However, it is not appropriate for operating robots in real time and does not suit users unfamiliar with computers. Besides, such conventional interfaces have disadvantages on small mobile devices in general.

Voice-activated interfaces with robots [2, 3] have some advantages for our purpose. Even users with little training will be able to give spoken commands without trouble. However, speech recognition errors and background noises deteriorate command success rates and usability.

For the above reasons, the authors propose a new kind of input method combining spoken commands and button press actions, based on a multi-modal command language [4] in order for users to be able to intuitively operate robots without frequent errors in command recognition after a very short period of training.

Obviously, it is very difficult to operate a remote robot in real time if one cannot see the robot or its surroundings at all. Pictures from on-site cameras would help to give the robot tasks and see what is happening. Therefore, the proposed interface displays pictures from a camera on the operated robot on its screen. In addition, an animated humanoid robot appears on the screen and makes gestures to report what the real robot is doing.

The following part of this paper describes an overview of the multi-modal interface, an interaction model, the multi-modal command language, and the animated communicator designed for smooth and desirable multi-modal interaction.

II. MULTIMODAL INTERFACE

The multi-modal interface proposed in this paper can be realized using a mobile phone with a screen and a head set or a small portable computer. Fig. 1 illustrates a typical example of the interface, which displays pictures from a camera on a remote home robot and an animated humanoid "communicator" which speaks and makes gestures to communicate with users.

Using this multi-modal interface, one can command a home robot through a microphone and twelve buttons monitoring the screen and listening to the communicator. This interface is based on an interaction model described in the next section in order for users to command robots without confusion or communication problems.

Fig. 1 Multi-modal interface using a mobile phone

III. INTERACTION MODEL

1. Interaction states

Fig. 2 depicts the proposed interaction model for remote operation of home robots. This simple model has four interaction states, S1, S2, S3, and S4. In S1, the interface waits for a cue from a user without listening for a command. It listens for a new command in S2. If a valid command arrives, a transition to S3 will occur and the command will be interpreted. If executable, the command will be executed in S4. Otherwise, a transition back to S1 will occur.

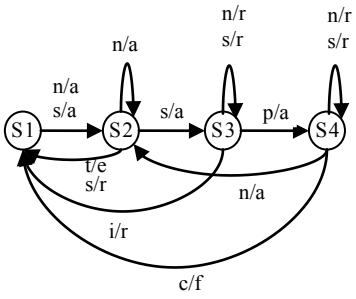

Fig. 2 Interaction model

2. User and system events

The interface is driven by user events and events generated by the interface itself. For instance, the transition from S1 to S2 in Fig. 2 is triggered by a non-verbal user event (n) or a speech event (s) from a user. A system event is either a possible action event (p), an impossible action event (i), a time out event (t), or a completion event (c).

3. Transition signals

In the interaction model, the system sends out transition signals whenever a state transition occurs. Transition signals include acceptance signals (a), rejection signals (r), finish signals (f), and empty signals (e). For example, in S3, a possible action event (p) will trigger a transition to S4 and an acceptance signal (a) sent to the user (hence, **p/a**, see the arrow from S3 to S4 in Fig. 2).

4. Multi-modal commands and communication c ues

In the model, users give their robots multi-modal commands and communication cues through a microphone and buttons, generating speech and non-verbal events. Here, a multi-modal command is a series of non-verbal events followed by a speech event arising in S2 of the model. A cue is any user event generated in S1, S3, or S4. Therefore, users can give communication cues only when the system is not listening, in which case user events are recognized as communication cues rather than components of a multi-modal command.

IV. MULTI-MODAL LANGUAGE

The multi-modal language, RUNA, comprises a set of grammar rules and a lexicon for spoken commands and communication cues, and a set of non-verbal events detected using buttons on a mobile phone, a keypad etc. The spoken language enables users to command home robots in Japanese utterances, completely specifying an action to be executed. Commands in the spoken language can be modified by non-verbal events. Speech and non-verbal events are also used as communication cues as described in the previous section. When the robot is unaware of the user, any button event can cue the robot to listen for a command.

In the version of RUNA for the remote interface, there are two types of commands, action commands and modifier commands. An action command consists of an action *type* such as *walk, turn, report,* and *lowertemp (for lowering the temperature setting)* and action *parameters* such as *speed, direction, angle, object* and *temperature*. Table 1 shows examples of action types and commands in RUNA.

The action types of RUNA are categorized into 24 classes based on the way action parameters are specified in Japanese. In other words, actions of different classes are commanded with different modifiers.

There are more than 300 generative rules for the latest full version of RUNA (Table 2). These rules allow Japanese speakers to command robots actions in a natural way by speech alone and to give communication cues. In RUNA, a spoken action command is an imperative utterance including a verb to determine the action type and other words to specify action parameters. For instance, a spoken command, "Yukkuri 2 metoru aruke! (Walk 2m slowly!)", indicates an action type *walk* and distance *2m* (Fig. 3). The third rule in Table 3 generates an action command of class 2 (AC2) which has *speed* and *distance* (SD) as parameters. The word category PE is for noise, silence or hesitation voice allowed between parameters. This helps speech recognition and command interpretation.

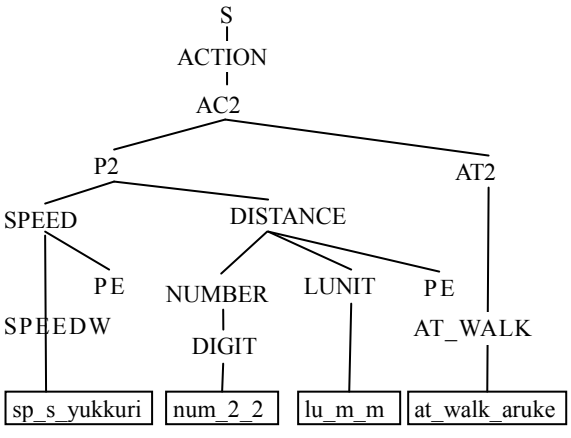

Fig. 3 An example parse tree for a spoken command

There are more than 250 words, categorized into about 100 groups identified by non-terminal symbols (Table 3). Because the language is simple, well-defined and based on the Japanese language, Japanese speakers would not need long training to learn it. Note that in

user test sessions, a reduced set of grammar rules and words can be employed to improve speech recognition performance.

Table 1 Examples of action commands

Type	Command	English Utterance
walk	*walk_s_3steps*	Take 3 steps slowly!
turn	*turn_f_l_30deg*	Turn 30°left quickly!
move	*move_m_r_2steps*	Move 2 steps right!
look	*look_f_l*	Look left quickly!
raisetemp	*raisetemp_room_2deg*	Raise the temperature of the room by 2 degrees!
settemp	*settemp_aircon_22deg*	Set the air-conditioner temperature around 22 degrees!
query	*query_aircon_all*	Report the status of the air-conditioner!

Table 2 Grammar rules of RUNA

Rule	Description
S → ACTION	action command
S → MODIFIER	modifier command
ACTION → SD AC2	class 2 command
AC2 → AT2	action type (class 2)
P2 → SPEED	speed (parameter)
P2 → DISTANCE SPEED	distance + speed
P2 → SPEED DISTANCE	speed + distance
SPEED → SPEEDW PE	one word for speed
DIST → NUMBER LUNIT PE	number + length unit
MODIFIER → REPEAT	repeat last action

Table 3 Part of RUNA's lexicon

Non-terminal	Terminal	Pronunciation
AT_WALK	at_walk_hokou	h o k o:
REPEAT	md_repeat_moikkai	m o: i q k a i
SPEED	sp_fast_isoide	i s o i d e
LUNIT	lu_cm_cm	s e N ch i
DIR_LR	dir_left_hidari	h i d a r i
TUNIT	tu_degree_do	d o
AIRCON	dev_aircon_eakon	e a k o N
AUNIT	au_degree_do	d o
PEND	mk_pe_q	q (pause)
	mk_pe_a:	a: (hesitation)
NI	joshi_ni_ni	n i

In RUNA, non-verbal events modify the meaning of spoken commands. They convey information about parameters of action commands. Table 4 shows examples of non-verbal events. For the remote interface, users can use keypad buttons to specify action parameters values instead of mentioning them. This reduces average number of words in a command and speech recognition errors. One can command a robot saying "Turn!" and pressing a button simultaneously instead of saying "Turn 33 degrees left slowly!" Furthermore, multi-modal commands are often more natural than spoken commands: e. g. pointing a glass and saying "Pick this up!" or saying "Lower the temperature!" pressing a button.

If a button event has been arrived within a short period of time, a spoken command will be modified as shown in Table 4. The twelve buttons are assigned to specific parameter values (Fig. 4). For example, the direction and speed of a turning action command are determined by the key pressed most recently by the user. If the key has been pressed once, the turning angle will be determined based on the duration of the key press event. If the key has been pressed more than once, a fixed angle value will be employed. Likewise, if a key has been pressed *twice* before a spoken command "Raise the room temperature!" the preset temperature will be *two* degrees higher.

Finally, the repeat button and query button allow users to command robots without speaking. The empty button helps to send a cue without specifying action parameters.

Table 4 Button event and action parameters

action type	duration	count	key
sidestep walk etc.	distance	distance	speed / direction
turn etc.	angle	angle	speed / direction
look etc.	-	-	speed/target
raisetemp settemp	-	temperature	-

← **left**	↑ **up**	→ **right**	Fast
← **left**		→ **right**	Moderate
← **left**	↓ **down**	→ **right**	Slow
empty	query	repeat	Cue

Fig. 4 Key assignment for action parameters

V. HUMANOID COMMUNICATOR

The animated humanoid communicator of the remote interface (hereinafter referred to as *communicator*) displays the current communication state of the interface changing his pose and sends out transition signals speaking and using gestures. When the system is not listening for a command, i.e. when it is in S3, the communicator looks the other way; he looks straight when the system is listening in S2 (Fig. 5). When the home robot is executing the action (S4), the communicator imitates the home robot's motion.

The communicator uses gestures to provide transition signals (Fig. 6) so that users give commands at the right moment. He repeats spoken commands like a parrot when he understands them (in S2 of Fig. 2). If the command is an executable one, the communicator

nods and says okay. He bows and says "I cannot do that!" if the command cannot be executed. Table 5 lists spoken messages and gestures on state transitions in Fig. 2.

Table 5 Gestures and spoken messages

transition	gesture	Message
S2 → S3	-	(repeat spoken command)
S2 → S1	shrug	"I don't understand!"
S3 → S4	nod	Okay!"
S4 → S1	salute	"I completed!"
S4 → S2	straighten up	"I stopped the action!"
S1 → S1	glimpse	"You cannot command
S3 → S3	shake hand	now!"
S4 → S4	cross arms	"I cannot stop now!"
S3 → S1	bow	"I cannot do that!"

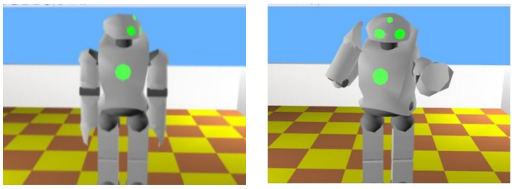

Fig. 5 Poses for S1 (left) and S2

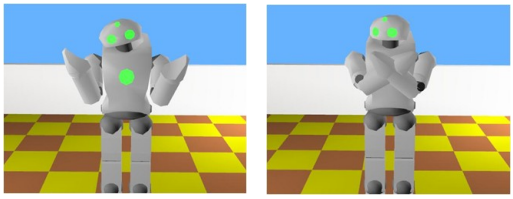

Fig. 6 Gestures for transition signals

Fig. 7 Simulated home environment on Webots5

VI. USER EVALUATION

1. Test bed

The remote interface has been implemented on top of the Open Agent Architecture (OAA) [5]. Speech events are detected by Julian, a grammar based version of a speech recognition engine [6], using a recognition grammar for RUNA. A command interpreter identifies the action type and parameter values mentioned in a spoken command; it determines unspecified parameter values using non-verbal events and default parameter values. Description of the multi-modal command interpreter in more detail can be found in some of the authors' previous papers [4, 7].

The communicator was built on Webots5.8.0 robot simulator (http://www.cyberbotics.com) and a free speech synthesizer. A simulated humanoid robot capable of execute actions in RUNA and a home environment were also created on the same robot simulator for user evaluation using a simulated home (Fig.7).

2. Methods

The remote interface can be evaluated in several ways. In user evaluation, potential users will be asked to achieve various tasks such as navigating the home robot, looking into rooms, operating home appliances and checking their status. Also, they will be asked to give one command at once from a command list for testing. It is planned that the simulated home robot system will be tested with more than 50 users.

VII. CONCLUSION

A multi-modal interface with remote home robots was presented. A test bed for user evaluation has been implemented based on a multi-agent architecture, a speech recognition engine, and a robot simulator. Future work includes full user evaluation using the test bed and implementation on portable devices.

AKNOWLEDGMENT

This work was supported by KAKENHI Grant-in-Aid for Scientific Research (C) (19500171).

REFERENCES

[1] Cho K, Kawamura T (2007) Home Security Robot Using Life Ontology and Blog Interface. Toshiba Review 62-12:50-53 (in Japanese)
[2] Prasad R, Saruwatari H, Shikano K (2004) Robots that can hear, understand and talk. Advanced Robotics 18-5:533-564
[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] 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 (to appear)
[5] Cheyer A & Martin D (2001), The open agent architecture. Journal of Autonomous Agents and Multi-Agent Systems, 4-1/2:143-148, March
[6] Lee A, Kawahara A, Shikano K (2001) Julius --- an open source real-time large vocabulary recognition engine. Proceedings of the 7th European Conference on Speech Communication and Technology, Aalborg, Denmark
[7] 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