Success Rates in a Multi-modal Command Language for Home Robot Users

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 discusses success rates in a multi-modal command language for home robot users. In the command language, one specifies action types and action parameter values to direct robots in multiple modes such as speech, touch and gesture. Success rates of commands in the language can be estimated by user evaluation in several ways. This paper presents some user evaluation methods and results from recent studies on command success rates. The results show that the language enables users without much training to command home robots at success rates of as high as 88-100%. It is also shown that multi-modal commands combining speech and button press actions included fewer words and were significantly more successful than single-modal spoken commands.

Keywords: Home robot, multi-modal, command language, success rate, 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 help people in homes in many ways achieving various tasks with or without their users at home. There is no doubt that such multi-purpose home robots should be easily directed by a wide range of untrained non-experts.

The authors proposed to design and use an artificial command language for home robot users, in order to enable non-experts to direct home robots without much training [1,2]. RUNA [2], a multi-modal command language for home robot users is designed to be learned and used by a wide range of users without effort. This language allows users to direct robots in multiple modes such as speech, touch and gesture. It is so carefully designed that one can realize a command interpreter which can interpret multi-modal command without much computational cost for speech and language understanding or gesture detection.

The authors have developed some real and simulated robots that can be directed in RUNA and conducted user studies. Some of the studies showed that novice users who had never directed robots were able to give valid commands in RUNA [2, 3]. However, some users had communication problems and were confused during commanding a robot in RUNA. Some commands given by them were too early and ignored; some were misrecognized due to speech recognition errors and the robot responded with "I don't understand!" or "I cannot do that!" Success rates of those users were low especially in the beginning. However, even those users might be able to improve their success rates in a short period of time with some training if the communication problems can be removed or reduced.

Based on the above discussion, the user evaluation systems were modified in order to reduce all sorts of communication problems including speech recognition errors and improve overall command success rates. This paper presents some user evaluation methods to estimate command success rates and results from recent studies on command success rates. The results show that the language enables users without much training to command home robots at success rates of as high as 88-100%. It is also shown that multi-modal commands combining speech and button press actions included fewer words and were significantly more successful than single-modal spoken commands.

II. MULTIMODAL COMMANDS

In RUNA, users specify an action type and action parameter values for each action command. For instance, to command a robot to turn, one should convey an action type, *turn,* and action parameter values (direction, angle, and speed). In the language, there are 24 action classes; each action type belongs to one of the action classes and has its own parameters. Table 1 shows some of the action classes of RUNA and parameters required.

In the command language, one can direct a robot in a spoken command, verbally specifying an action type and parameters: e. g. "Turn left by 45 degrees very slowly!" Parameter values can be left out: e. g. just saying "Turn left!" or "Turn slowly!" because each action parameter has a default value and this value will be set if not specified. Therefore, users need not mention every parameter and can reduce speech recognition errors.

The multi-modal language also allows users to give a parameter value using a gesture, pressing a button, or touching the robot they are commanding. For example, they can touch the robot's left shoulder for a short while and then say "Turn quickly!" to make the robot turn left (direction) by about 10 degrees (angle).

The authors are interested in how quickly users can learn the command language, how well they can command robots, and the success rate of their multi-modal commands.

Table 1 Action classes, types, and parameters

Class	Type	Parameters
3	*look ,turnto, lookAround*	*speed, target*
4	*turn*	*speed, directionlrni, angle*
5	*sidestep*	*speed, directionlrni, distance*
6	*move*	*speed, directionni, distance*
7	*handshake, highfive*	*speed, handde*
8	*punch*	*speed, handed, directionni*
9	*kich*	*speed, footde, directionni*

III. COMMAND SUCCESS RATES

To command robots in RUNA, users should have a clear purpose in mind. They should decide what action type and action parameter values to specify before giving a command to a robot. The command is successful if and only if the robot executes it correctly. Note that user intention is not observable although it is possible to guess action types and parameter values in spoken commands, gestures, touches, button press actions, etc. One should also note that novice users may not have some of action parameters clearly in mind.

Multi-modal commands must be valid to be correctly interpreted by robots. More specifically, spoken commands must be grammatical, and gestures, touches on robots, and button press actions must be presented to robots so that action types and action parameter values should be precisely recognized. Therefore, novice users must learn the multi-modal language to be successful in directing robots.

Users will make errors and give invalid commands, which cause a lower success rate, even after learning the language. This means that in order to realize a higher success rate with little learning effort, the command language should be carefully designed.

Even valid and reasonable commands can fail, since robots cannot always execute them as users give. First, robots may fail to understand spoken commands due to speech recognition problems or noises. Secondly, non-verbal messages can be misunderstood or ignored due to gesture recognition errors etc. Thirdly, robots may fail to execute even correctly recognized commands for physical reasons; they may fail to pick up a glass or stumble taking three steps forward.

False alarms by speech recognizers and gesture detectors can confuse users, making it harder for them to learn the language, and result in lower success rates. False non-verbal events can set a wrong parameter value. For example, a false alarm of a gesture to indicate “a very low speed” will cause an unwanted very slow action.

As stated above, the command is successful if and only if the robot executes it correctly. Thus, a success rate is determined by the number of successful commands in a set of multi-modal commands. However, as user intention is not observable, one must watch a user giving a command and guess what action type and parameter values were in mind. The problem, however, is that the user might have made a slip of the tongue or an error in non-verbal parameter specifications. Therefore, asking what a user intended and asking users to give particular commands are essential to discuss success rates in user evaluation.

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 various sensors on robots and buttons on computers, mobile phones, controllers, 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.

In RUNA, there are two types of commands, action commands and modifier commands. An action command consists of an action *type* such as *walk, turn, pickup,* and *lowertemp (for lowering the temperature setting)* and action *parameters* such as *speed, direction, angle, object* and *temperature*. Table 2 shows examples of action types and commands in RUNA.

Table 2 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!

The action types of RUNA are categorized into 24 classes based on the way action parameters are specified in Japanese (Table 1). 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 3). These rules allow Japanese speakers to command robots actions in a natural way by speech alone. 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. 1). 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 category was introduced lately to solve some problems and would help speech recognition and command interpretation.

There are more than 250 words, categorized into about 100 groups identified by non-terminal symbols (Table 4). 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 success rate.

In RUNA, non-verbal events modify the meaning of spoken commands. They convey information about parameters of action commands. For instance, users can use keypad buttons to give action parameters values instead of mentioning them. This will reduce average number of words in a command and speech recognition errors. If a non-verbal event has been arrived within a short period of time, a spoken command will be modified (see Table 5 for examples of mappings of button event parameters to action parameters).

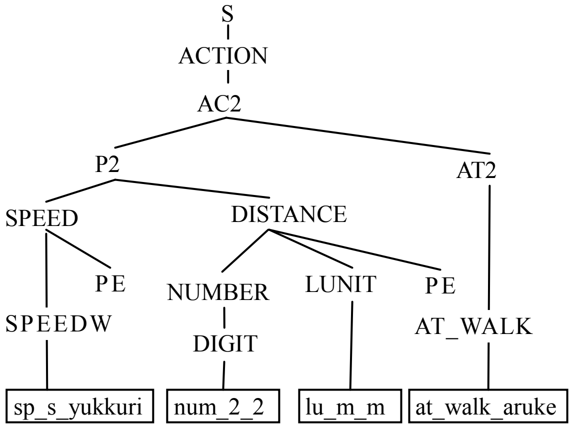

Fig. 1 An example parse tree for a spoken command

Table 3 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 4 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
AUNIT	au_degree_do	d o
PEND	mk_pe_q mk_pe_a:	q (pause) a: (hesitation)
NI	joshi_ni_ni	n i

Table 5 Button event and action parameters

action type	duration	count	Key
move	distance	distance	speed / direction
turn	angle	angle	speed / direction
walk	distance	distance	Speed
look	-	-	speed/direction

V. METHODS

There are several ways to estimate command success rates of novice and trained users. One can teach users exactly how to give commands and then let them give the same commands. For spoken commands, one can show them a printed list of utterances presenting every word. Users can be tested given some general instructions and a set of actions, with a type and parameters for each, to be commanded in RUNA or a set of goals to be achieved by commanding a robot in the language.

The authors have developed a command recognition system on top of a multi-agent architecture [4] which interprets multi-modal commands in RUNA, integrating a grammar-based speech recognition engine [5], a gesture detector using the OpenCV library for computer vision (http://www.intel.com), a button press event detector, a tactile event detector, and a command interpreter which utilizes an action database [2, 3]. Recently, a speech synthesizer was added to the system to repeat valid spoken commands to help untrained users. This system has been applied to some test beds to direct real small humanoids and simulated robots on Webots5 simulator (http://www.cyberbotics.com) [6].

Each of 14 novice users, mostly high school students who visited Fukuoka Institute of Technology, gave 25 spoken commands to a small humanoid robot on a table in a noisy environment with many people (Test A). Each verbal command was displayed on a computer screen and the users were given an opportunity of practice before giving each command. We videotaped the users and logged all the system events including non-verbal events, speech recognition results and command interpretations. These users were tested using a reduced grammar with 148 rules and 133 words.

Another 14 novice users of a wider range who visited one of the authors' offices were given a three-page note which explains how to operate robots in RUNA (Test B). After a five minute practice of giving commands in the language, they were asked to remotely operate a humanoid in a simulated environment on Webots5 through a microphone and a keypad to explore a room monitoring images from the camera until they can answer three questions about the room: "Is there a note on the refrigerator?", "Is one of the drawers near the sink open?" and "Is there anything on the floor in front of the sink?" We recorded time to complete the task, multi-modal commands given by the users, and all the system events. We also video-recorded the users while they were giving commands to the simulated robot. The users were tested with a smaller grammar including 110 rules and 92 words.

VI. RESULTS

Some results of the former test (Test A) are shown in Table 6: each user's success rate of spoken commands on the screen (SR), speech recognition rate (RR), and word error rate (WER). Some of the commands were misrecognized by the speech recognizer but successful without errors in action types and parameter values for some reasons. There were one ungrammatical command (0.29%) and two grammatical commands which were slightly different from the command displayed on the screen (0.57%).

Table 6 Success rates of listed spoken commands

User	SR(%)	RR(%)	WER(%)
A1-A5	100	100	0
A6,A7	96.0	96.0	2.5
A8	96.0	92.0	3.7
A9	96.0	92.0	4.9
A10	96.0	88.0	3.7
A11,A12	92.0	92.0	4.9
A13	92.0	88.0	6.2
A14	92.0	84.0	6.2

Table 7 Success rates when achieving a goal

user	SRA(%)	RR(%)	WER(%)	SRM(%)
B1-B5	100	100	0	100
B6	100	97.2	1.7	100
B7	96.9	96.9	3.8	100
B8	95.4	95.4	2.2	97.5
B9	94.4	94.4	4.2	100
B10	93.7	93.7	6.9	94.7
B11	91.9	91.9	6.1	-
B12	90.7	88.9	11.1	100
B13	90.5	90.5	7.2	100
B14	88.0	88.0	14.8	75.0

Table 7 summarizes the results of the latter test (Test B): each user's success rate of valid commands (SRA), speech recognition rate (RR), word error rate (WER), and success rate of valid multi-modal commands comprising speech and button press actions (SRM). The first five users in Table 7 succeeded in every command they gave with no speech recognition errors although there were a few false alarms per user. For several users, the success rate of multi-modal commands was higher than that of single-modal spoken commands. User B11 gave only spoken commands and user B14 was more successful in giving spoken commands. The users gave 17 to 79 (38.9 on average) valid commands in RUNA and it took about five to thirteen minutes for them to complete their task. Eight of them gave one or two invalid (ungrammatical) commands, which were 2% of all the user commands.

On average, 98.1 % of multi-modal commands and 91.3 % of single-modal spoken commands were successful. The difference is proved to be statistically significant (p=0.0012). The average success rate of the 14 users was 95.1%. The average command length of the multi-modal commands was 1.14 words which was significantly shorter than 2.28 words for the spoken commands.

VII. DISCUSSION AND SUMMARY

The facts that in the both user tests more than 95 percent of commands were successful and that there were little invalid commands imply that the users learned how to command robots in RUNA in a very short period of time. In Test B, multi-modal commands were more successful, presumably because multi-modal commands included fewer spoken words and numerical phrases such as "25 cm" or "45 degrees." The user were more successful, even in a noisier place, than the users described in the authors' previous work (78%) [3], partly because of the repetitions of spoken commands using the speech synthesizer, short practice, and the new grammar rules for hesitations and pauses. The users learned the language, how to effectively use the microphone, and how loud they should speak very quickly thanks to the speech feedback. In fact, there were much less frequent communication problems in Test B than in user in earlier studies [2, 3]. Using grammars for speech recognition for a specific purpose, the word error rates in most of the users' commands were as low as 0-8%.

The users in Test B commanded the robot spontaneously without words on paper or the screen which might have slightly lowered their success rates although the grammar was a smaller one. Another hypothesis implied by the data is that young users can adapt to the system more quickly than older users including B11 and B14.

In summary, the results of two user tests show that the language enables users without much training to command home robots at high success rates. It is also shown by the results that multi-modal commands combining speech and button press actions included fewer words and were significantly more successful than single-modal spoken commands.

ACKNOWLEDGMENT

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

REFERENCES

[1] Oka T, Yokota M (2007) Designing a multi-modal language for directing multipurpose home robots. Proceedings of the 12th International Symposium on Artificial Life and Robotics (AROB '07), Beppu, Japan

[2] 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)

[3] 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

[4] Cheyer A & Martin D (2001), The open agent architecture. Journal of Autonomous Agents and Multi-Agent Systems, 4-1/2:143-148, March

[5] 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

[6] Michael O (2004) Cyberbotics Ltd – Webots^{TM}: Professional Mobile Robot Simulation. International Journal of Advanced Robotic Systems 1-1:39-42