A spoken dialogue interface for TV operations based on data collected by using WOZ method

Jun Goto	Yeun-Bae Kim	Masaru Miyazaki	Kazuteru Komine	Noriyoshi Uratani
NHK STRL Human Science	NHK STRL Human Science	NHK STRL Human Science	NHK STRL Human Science	NHK STRL Human Science
Tokyo 157-8510 Japan	Tokyo 157-8510 Japan	Tokyo 157-8510 Japan	Tokyo 157-8510 Japan	Tokyo 157-8510 Japan
goto.j-fw @nhk.or.jp	kimu.y-go @nhk.or.jp	miyazaki.m-fk @nhk.or.jp	komine.k-cy @nhk.or.jp	uratani.n-fc @nhk.or.jp

Abstract

The development of multi-channel digital broadcasting has generated a demand not only for new services but also for smart and highly functional capabilities in all broadcast-related devices. This is especially true of the television receivers on the viewer's side. With the aim of achieving a friendly interface that anybody can use with ease, we built a prototype interface system that operates a television through voice interactions using natural language. At the current stage of our research, we are using this system to investigate the usefulness and problem areas of the spoken dialogue interface for television operations.

1 Introduction

In Japan, the television reception environment has become quite diverse in recent years. In addition to analog broadcasts, BS (Broadcast Satellite) digital television and data broadcasts have been operating since 2000. At the same time, TV operations for receiving such broadcasts are becoming increasingly complex, and an ever increasing variety of peripheral devices such as video tape recorders, disk recorders, DVD players, and game consoles are now being connected to televisions, and operating such devices with different kinds of interfaces is becoming troublesome not only for the elderly but for general users as well (Komine et al., 2000).

Recently we conducted a usability test targeting data broadcasts in BS digital broadcasting. The results of the test revealed that many subjects had trouble accessing hierarchically arranged data.

This finding revealed the need for an easy means of accessing desired programs. One such means is a spoken natural language dialogue (hereafter spoken dialogue) interface for TV operations. If spoken dialogue could be used to select and search for programs, to operate peripheral devices, and to give information in reply to system queries, we can envisage such an interface as being extremely valuable in a multi-channel and multi-service function viewing environment. With this in mind, we have set out to build an interface system that could operate a television via spoken dialogue in place of manual operations.

2 Collecting dialogue data for TV operations

Assuming that a television is intelligent enough to understand the words spoken by a human, what kind of language expressions would a user use to give commands to that television? In other words, it is important that the words spoken by a user in such a situation be carefully examined when designing a television interface using spoken dialogues. Therefore first we built an experimental environment that would enable us to collect dialogue data based on WOZ (Wizard of OZ) method.

2.1 Wizard of OZ

We set up a television-operation environment according to the WOZ framework in which the subjects were instructed that “the character appearing on the television screen can understand anything

you say, and that the character will operate the television for you."

The number of channels that could be selected was 19, and screens displaying Electronic Program Guide (EPG) and user interface for program searching were presented as needed (Komine et al., 2002).

This WOZ environment required two operators, one in charge of voice responses and the other of user interface operations. The voice-response operator returns a voice response to the subject by a speech synthesizer after selecting a reply from about 50 previously prepared statements or inputting replies directly from a keyboard. If the subject happens to be silent, the operator returns a response that introduces new services or prompts the subject to say something. The user interface operator first determines what the subject wants, and then manipulates user interface or EPG and performs basic television operations such as changing channels.

The subjects selected for data collection consisted of 10 men and 10 women ranging in age from 24 to 31 (average age: 28.7), and each was allowed to speak freely with the television for 5 minutes under an assumption that the "television has a certain amount of intelligence."

2.2 Results of data analysis

Figure 1 shows an example of dialogue data recorded during a WOZ session. On analyzing collected utterances made by the subjects (1,268 utterances in total), it was found that 83% of user utterances concerned requests made to the television, and that 89% of those requests included words belonging to specific categories such as *program title*, *genre*, *performer*, *station*, *time*, and *TV operation commands*. The remaining 17% of utterances did not concern the system but were rather a result of subjects talking or muttering to themselves for self-confirmation and the like.

Here, we consider the following reason why most utterances belonged to specific categories despite the fact that a variety of request could be made. In this system, TV program- and operation-related information is displayed on the television screen, and based on this information, subjects tended to underestimate television capability and to omit utterances not dealing with service functions they saw as possible. It is also thought that the conventional image of television inside subjects' minds served to restrict user utterances.

00:27:08	Subject	Well, I'm looking for a program.
00:30:23	WOZ	You can also choose by genre. Would you like to see the list of programs by genre?
00:36:25	Subject	Yes.
00:38:00	WOZ	All right.
00:47:02	Subject	Ah!
00:47:02	WOZ	Please select a genre.
00:50:04	Subject	Well, let's see. How about "Variety?"
00:55:11	WOZ	OK!
01:02:06	Subject	I see.
01:03:29	WOZ	Please select the program you would like to see.
01:08:27	Subject	Well, I would like see more at the bottom of the screen.
01:12:09	WOZ	OK, I will do it.
01:15:23	Subject	Um, Just a little bit more.
01:17:27	WOZ	OK, how's that?

Figure 1: Example of dialogue data

As a part of this WOZ experiment, we also had the subjects fill out a questionnaire with regards to television operations by using spoken dialogue interface. When asked to give an opinion on operating a television by voice, more than half replied "*Yes, I would like to*" therefore apparently indicating a high demand for the spoken dialogue interface. On the other hand, most subjects that replied "*No, I would not like to*" gave simple embarrassment at speaking out loud as one reason and a reluctance to vocalize commands when watching television together with their families as another. In this regard, we think that embarrassment could probably be reduced through user experience and appropriate environment configuration.

3 Spoken dialogue interface system for TV operations

Based on the results of the data analysis, we built a prototype system that enables television operations via spoken dialogue. Figure 2 shows the configuration of this system. The system allows users to select real-time broadcast programs from 19 channels. It also enables the presentation of program in-

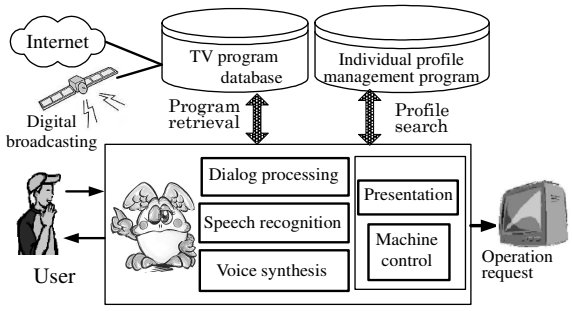

Figure 2: Configuration of interface system

Figure 3: Interface robot and an operation scene

formation obtained from the Internet or overlaid data in digital broadcasts; the scheduling of program recording; and the browsing of program-related information from Internet. All of these functions can be operated through spoken natural language interactions. The main processing modules of the system are described below.

3.1 Robot interface

The user makes operation requests to interface robot (IFR) as shown in Figure 3, and the IFR operates the television accordingly for the user. The IFR is equipped with a super-unidirectional microphone and a speaker, and communicates and activates the speech recognition and voice synthesis, and dialogue processing of the system. The IFR has been given the appearance of a stuffed animal. One advantage of this IFR is that it can be directly touched and manipulated to create a feeling of warmth and closeness.

On hearing a greeting or being called by its name, the IFR opens its eyes and enters a state that can perform various operations. For example, the IFR can assist the user search for a program, can present information about any program on the television screen, and can return voice responses.

3.2 Speech recognition

The speech recognition module uses an algorithm that can finalize recognition results in a sequential manner for a real-time operation and a high speech recognition rate. When applying this module to a news program, a speech recognition rate of about 95% can be obtained (Imai, 2000).

In speech that occurs during television operations, the words such as program titles, names of broadcast stations, names of entertainers and etc. have a high probability of occurring and are also updated frequently. For this reason, newly acquired word-lists are automatically registered in a dictionary on a daily basis. In addition, as program titles often consist of multiple words, it is necessary to register them as a single word in order to improve the recognition rate.

Despite several additional forms of tuning, it is still difficult to achieve perfect results with current speech recognition technology. To enable feedback to be given to the user at the time of erroneous recognition, results of recognition are always displayed on the lower left corner of the television screen.

3.3 Dialogue processing

In dialogue processing, it is generally difficult to understand intent by performing only a lexical analysis of speech. If we limit tasks to dialogue used in television operation, the words spoken by a user have a high probability of falling into specific categories such as program name, as indicated by the results of the data analysis described in 2.2. As a consequence, user intent can be inferred from a combination of specific categories and predicates. From the viewpoint of processing speed, processing can be performed in real time if we use pattern-base approach. This approach is also used in other dialogue systems such as PC-based agent television systems in the (FACTS) project and (Sumiyoshi et al., 2002).

The dialogue processing module performs real-time morphological analysis of input statements from the speech recognition module. A statement is then identified by pattern matching in units of morphemes and the meaning ascribed beforehand to that statement is obtained. An example of such pattern is shown in Figure 4 using the meta-characters listed in Table 1:

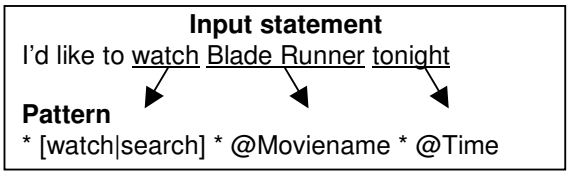

Figure 4: Example of pattern matching

Table 1: Meta-characters used in pattern

Meta-character	Description
*	any number of any words
+	one word
!	non-matching word
{}	optional
[]	mandatory
()	any order
@	slots
|	or
,	delimiter

In the pattern matching process, categories important to television operations are stored as slots. Table 2 lists these category-slots and examples of their members. The words stored in these slots are then used as a basis for generating television operation commands and search expressions to access the TV program database. Response statements to input statements may take various forms depending on the patterns and current circumstances, and they are here generated by taking into account slot information, response history, results of searching for program information.

Table 2: Content of category-slots

Slot	Examples
@Moviename	Blade Runner, My Fair Lady etc
@Performer's name	Harrison Ford, Chizuru Ikewaki Norika Fujiwara, etc
@Genre	Drama, Animation, News, etc
@Time	10:20, Tomorrow, Tonight, etc
@Broadcast station name	NHK, TBS, WOWOW, etc
@Direct operation	Volume, Channel, etc
@Action	Search, Watch, Turn up, etc

4 Conclusion

We have built a spoken dialogue system based on the results of a WOZ experiment with the aim of achieving a television operation interface easy enough for anybody to use.

In the preliminary system operation test, 5 subjects were asked to give some examples of TV programs that they watch at home, and to use this system to see whether they could obtain information in relation to those programs. Results of this test showed that all subjects could access information on desired programs. In a subsequent questionnaire, moreover, all subjects stated that "program selection was easy, and particularly there was no need to know about hierarchical structure of program information."

On the other hand, the test also revealed that some issues remain to be addressed in speech recognition but that a favorable evaluation could be obtained from all subjects with regard to television operations via spoken dialogue. We are currently conducting even more detailed experiments to demonstrate the usefulness of a spoken dialogue interface for television control and to examine problem areas.

References

FACTS (FIPA Agent Communication Technologies and Services) A1 Work Package. Available at http://sharon.cselt.it/projects/facts-a1/.

Hideki Sumiyoshi, Ichiro Yamada, and Nobuyuki Yagi. 2002. Multimedia Education System for Interactive Educational Services. *Proceedings of IEEE International Conference on Multimedia and Expo*, CD-ROM.

Kazuteru Komine, Nobuyuki Hiruma, Tatsuya Ishihara, Eiji Makino, Takao Tsuda, Takayuki Ito, and Haruo Isono. 2000. Usability Evaluation of Remote Controllers for Digital Television receivers. *Proceedings of SPIE, Human Vision and Electronic Imaging 5*, Vol. 3959:458-467.

Kazuteru Komine, Toshiya Morita, Jun Goto, and Noriyoshi Uratani. 2002. Analysis of Speech Utterances in TV Program Selection Operations using a Spoken Dialogue Interface. *Proceeding of Human Interface Symposium*, No.3231:631-634. (in Japanese).

Toru Imai. 2000. Progressive 2-pass Decoder for real-time Broadcast news captioning. *Proceedings of ICASSP-2000*, Vol.3:1559-1562.