Human-Computer Interaction -- INTERACT'03
M. Rauterberg et al. (Eds.)
Published by IOS Press, (c) IFIP, 2003, pp. 765-768

A Television Control System based on Spoken Natural Language Dialogue

Jun Goto Kazuteru Komine Yeun-Bae Kim Noriyoshi Uratani

NHK (Japan Broadcasting Corporation)
Science & Technical Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo, Japan

goto.j-fw@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 interaction using spoken natural language. At the current stage of our research, we are using this system to investigate the usefulness and problem areas of an interactive voice interface for television operation.

Keywords: spoken dialogue interface, Television operation, speech recognition

1 Introduction

The television reception environment in Japan has become quite diverse in recent years: Broadcast Satellite (BS) and Communication Satellite (CS) digital broadcasting is expanding and the launch of digital terrestrial broadcasting is scheduled for 2003. At the same time, however, reception operation is becoming all the more complicated. In addition, a wide variety of peripheral devices such as VTRs, DVD players, and game consoles are now being connected to televisions, and operating such equipment having different kinds of interfaces is becoming troublesome not only for the elderly but for general users as well.

Against the above background, we conducted a usability test last year targeting data broadcasts in BS digital broadcasting. It was found that a wide range of age groups in addition to the elderly had problems accessing hierarchically arranged data. This finding revealed the need for an easy means of accessing desired programs. One such means is a voice interface for operating reception equipment.

An interface that can understand user speech and then operate multi-functional televisions and peripheral devices accordingly while returning essential information by voice responses should be an extremely valuable interface in a multi-channel and multi-function viewing environment. With this in mind, we have set out to construct an experimental television control system based on natural language dialog to provide a form of television control that anybody can use. In this paper, we report on the results of an experiment that we performed to collect dialog data using the Wizard of OZ (WOZ) system, and describe a television control system using voice interaction based on the results of that experiment.

2 Experiment for Collecting Dialog Data in Television Operation

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 directions to that television? To put it another way, it is important that the words spoken by a user in such a situation be carefully studied when designing a television control system using voice interaction based on natural language. To this end, we decided to construct an experimental environment using the WOZ system that would allow users to make television control requests and to ask the television questions, and that would enable us to collect data on the dialog that occurred at that time.

2.1 Experimental conditions

In the experiment, 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.” In other words, they were made to believe that the television that they were to be using had a certain amount of intelligence. This computer-generated character that can converse with subjects is synthesized on the television screen, and two operators, one each for

character manipulation and television control, are placed in separate rooms.
The number of channels that could be selected in the experiment was 19, and screens displaying program information and interfaces for program searching were presented as needed. The subjects 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 thirty minutes under the conditions described above.

2.2 Experimental results

All dialog obtained by the experiment was recorded and analyzed for trends. 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 (program name, person's name, program genre, time, name of broadcast station, and control words). 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 requests could be made to a television possessing intelligence. In this system, program- and control-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 functions they saw as possible. It is also thought that the conventional image of television inside subjects' heads would serve to restrict user speech.

3 Television Control System using Voice Interaction

Based on the conditions of the WOZ experiment described above, it was found that users would voluntarily restrict their speech when interacting with the television, and that such speech would have a high probability of falling into specific categories. Taking these experimental results as preconditions, we developed a television control system using voice interaction based on natural dialog.
In terms of functions, users of this system are allowed to select programs on all the channels in terrestrial and BS broadcasts. The system, moreover, can present program information using data attached to digital broadcasts and data from the Internet, and based on this information, peripheral devices can be controlled to schedule program recordings or perform other functions. General Web browsing can also be performed, and all functions can be controlled in an interactive format based on natural language. At present, however, languages that can be used for voice interaction are limited to Japanese. Figure 1 shows the configuration of this system.

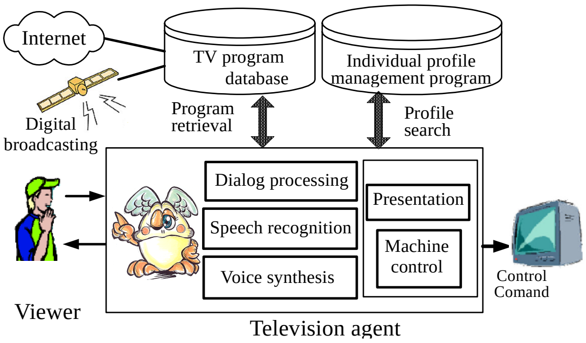

Figure 1: Structure of a television control system based on spoken natural language dialogue

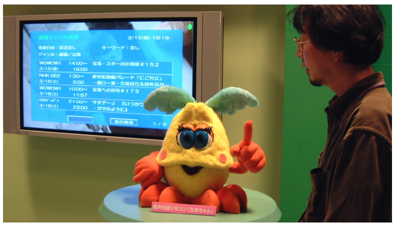

Figure 2: Television agent

The following overviews a television agent used as an input/output interface and main modules of the system.

3.1 Television agent

The user makes control requests using natural language to the television agent (TVA) shown in Fig. 2, and the TVA controls the television accordingly for the user. The TVA is equipped with a super-unidirectional microphone and a speaker, and communicates and activates the speech recognition, voice synthesis, and dialog processing modules of the system. The TVA has been given the appearance of a stuffed animal to facilitate ease of use.
On hearing a greeting or being called by its name, the TVA opens its eyes and enters a state that can perform various operations. For example, the TVA can help the user search for a program, can present program information on the television screen, and can return voice responses. Furthermore, if no voice input from the user has been received within a set time period, the TVA informs the user that it is entering a standby state. It then closes its eyes and waits for a startup word.

3.2 Dialog processing module

In dialog processing, it is generally difficult to understand intent by performing only a lexical analysis of speech. If, however, we limit tasks to dialog used in television control, 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 above WOZ experiment. As a consequence, user intent can be inferred from a combination of specific categories and, from the viewpoint of processing speed, processing can be performed in real time. This approach is also used in other dialog systems such as agent television in the FACTS (FIPA Agent Communication Technologies and Services) project [1].

When using a TVA, the dialog-processing module performs real-time morphological analysis of text input from the speech recognition module. This preprocessing is performed because Japanese is a language written with no space between words. A statement is then identified by pattern matching in units of morphemes and the meaning ascribed beforehand to that statement is obtained. Patterns are described as shown in Fig. 3 using the meta-characters listed in Table 1. In the pattern matching process, classes that are important to television control are stored as slots. Table 2 lists these classes and examples of their members. The words stored in these slots are then used as a basis for generating television control commands and search expressions for the television-program database. Response statements to input statements may take various forms depending on the pattern and current circumstances, and they are here generated by taking into account slot information, response history, results of searching for program information, etc.

3.3 Speech recognition module

The speech recognition module uses an algorithm [2] that can obtain recognition results in a sequential manner making for both real-time operation and a high speech-recognition rate. When applying this engine to a news program, a recognition rate of about 95% can be obtained. In speech that occurs during television control, words such as program names, broadcast-station names, and performer names have a high probability of occurring and are also updated frequently. For this reason, newly acquired word lists are automatically registered daily in a morpheme dictionary.

In addition, as program names often consist of multiple morphemes, it is necessary to register them as only one morpheme. This is done because speech-recognition rate improves by not dividing program names into multiple morphemes when creating the language model. It is also possible to create a language model whereby weights are added to program information close in time to the day of system operation. Using this model improves the recognition of speech that includes program information such as program names and performer names that is relatively new.

Additional forms of tuning are also performed, but perfect recognition is still difficult to achieve with current speech recognition technology. Erroneous results will therefore occur with a certain probability. To enable feedback to be given to the user at the time of erroneous recognition, results of recognition are always displayed in the lower-left corner of the television screen.

Meta-character	Function
*	any number of words
+	one word
!	non-matching word
{}	word within {} or non-existent
[]	word within []
()	any order OK
@	word class
|	‘or’ word delimiter
,	‘and’ word delimiter

Table 1: Meta-characters

slot	Example
@Program name	My fair lady, Hommamon, etc
@Performer’s	Chizuru Ikewaki , etc
@Genre	Drama, Animation, News, etc
@Time	10:20, Tomorrow, Tonight, etc
@Broadcast station	NHK, TBS, WOWOW, etc
@Direct operation	Volume, Channel, etc
@Action	Search, Watch, Turn up, etc

Table 2: Content of slots

Input statement

I’d like to watch my fair lady tonight.

Template pattern

*[watch|search] *[@Program name] *{@Time}

Figure 3: A simple example of pattern matching

3.4 Voice synthesis module

Nass et al. have reported on the psychological effects that a responding voice can have on users [3]. In particular, they have found that modifying the quality of the synthesized voice, the speed of speaking, and the expressions used in accordance with the personality of the user helps to make a response more trustworthy and friendly. From this, one means for reducing resistance to performing vocal control would be to make the character that makes the responses and the quality of its responding voice something that everyone could be comfortable with.

A questionnaire that was passed out during the WOZ experiment also revealed that a comparatively high number of subjects gave the responding voice as the reason for feeling awkward when performing voice control. This led us to conclude that incompatibility between the character’s external appearance and the quality of its synthesized voice as well as response speed and expressions can have a negative effect on users.

The responding voice of the TVA must therefore match its friendly appearance. We have therefore adopted a hybrid system consisting of (1) recordings made by a voice artists that provides the voice matching the characters appearance and (2) synthesized voice achieved by a learning process based on the same artist's voice. The recordings are used for phrases, program names, performer names, etc. that are frequently used in voice responses while synthesized voice is used and automatically generated for unregistered vocalizations corresponding to a new word list.

3.5 Information used by the system

Program information used by the system consists of information attached to digital broadcasts and information obtained from the Internet. The number of programs broadcast in one day is about 380 in BS broadcasting and about 260 in terrestrial broadcasting, and the names, summaries, broadcast times, and other data for these programs must be obtained. This data is stored in a relational database that returns appropriate information in response to a search request from the dialog-processing module. Programs can be searched for on the basis of program name, performer name, genre, time, or broadcast-station name or any combination of the above.

This system is also capable of switching viewer profiles so that the preferences of a particular user can be considered and programs recommended accordingly. Here, individual authentication for switching profiles is achieved by a user-recognition process that employs acoustical feature parameters. Profiles are automatically created from "stereotypes" prepared from a survey of viewers brought together at the NHK Broadcasting Culture Research Institute and from program-selection history stored in the interactive system.

3.6 Machine control module

Once an output statement has been determined by the dialog processing module, the machine control module specifies which device is to be controlled and the optimal operation to be performed. In addition to the television itself, this module can control peripheral devices such as recording equipment and DVD players. Television control includes not only volume adjustment and channel selection but also the display of electric program guides (EPG) and search results as an aid to vocal interaction.

4 System Operation Experiment

We performed an experiment to test the operation of our prototype television control system. In this experiment, five subjects in their 20s were asked to use the system to see whether they could perform basic television operations and obtain information on programs that they usually watch.

User:	What dramas are being shown today?
System:	Today, in addition to "Honmamon," there are 32 dramas. Any preferences?
User:	Well, show me what is on from four o'clock.
System:	There is "Mitokomon" and three other dramas.
User:	OK, I would like to see Mitokomon.
System:	Mitokomon starts in one hour and 20 minutes. Would you like to record it?
User:	Please.
System:	OK, it's scheduled to be recorded.

Figure 4: Example of system operation

The subjects were first shown how the system works by a person skilled in its operation. They were then asked to name 5 to 10 programs that they usually watch, and to select those programs, obtain information on them, and to control peripheral devices, all by controlling the system through voice interaction. Figure 4 shows an example of resulting dialog.

Experimental results revealed that speech recognition was not always performed well but that 90% of subject tasks could access the information of desired programs. A subsequent questionnaire, moreover, revealed that all five subjects expressed the opinion that "program selection was easy and desired information could be obtained."

5 Conclusion

We have constructed an interactive voice system based on the results of a WOZ experiment with the aim of achieving a television control interface easy enough for anybody to use. The results of a system operation experiment 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 control through interactive speech. We are currently conducting even more detailed experiments to demonstrate the usefulness of an interactive voice interface for television control and to examine problem areas.

References

1. FACTS (FIPA Agent Communication Technologies and Services). Available at http://www.labs.bt.com/profsoc/facts/

2. T. Imai: Progressive 2-pass Decoder for real-time Broadcast news captioning. In Proceedings of ICASSP-2000, Vol. 3 of 6 (2000) 1559-1562

3. C. Nass, K. M. Lee: Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction. In Proceedings of CHI (2000) 329-336