

Measurement of Workload for Voice User Interface Systems

Takuya NISHIMOTO *, Motoki TAKAYAMA **, Haruaki SAKURAI **,
and Masahiro ARAKI **

* Graduate School of Information Science and Technology, The University of Tokyo,
Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-8656 Japan

** Faculty of Engineering and Design, Kyoto Institute of Technology,
Matsugasaki, Sakyo-ku, Kyoto, 606-8585 Japan

1. Introduction

Recently, with the maturation of speech synthesis and speech recognition technology, voice interface enabling interactions between machine and human has become possible through the use of several spoken dialog systems. Two advantages of voice operations for information acquisition and machine operations are: (1) **Hands-free**: Keyboard operations or pointing gestures are no longer necessary, making the system suitable for the non-desktop environment, (2) **Eyes-free**: Visual attention to the display is no longer necessary.

When a voice-alone human-machine interface system is used, however, the user may feel fatigue from speaking for many hours. Moreover, the difficulties of understanding voice commands and the problem of assessing whether the user's utterances are accepted correctly may also exhaust the user. These kind of problems partly result from the quality of the synthesized voice (poor understandability, etc.), or the performance of speech recognition (low recognition rate, long response time, etc.). They may also result from the failure of the interaction design, which depends on the dialog patterns, vocabulary of speech input, and the choice of words in response sentences.

In this paper, various workloads using spoken dialog systems are grouped together as the dialog workload. Dialog workload includes both the physical process of speaking and so on, and the cognitive process of mental activities such as remembering, searching, reasoning, paying attentions, etc. Measuring the dialog workload involves not only comparing the physical processes such as the number of the utterances required to finish a task, but also involves the cognitive processes, such as the level of attention that the user must pay to the system. The measurement of dialog workload is

expected to show the ease of use, or the mental margins for using modalities other than the voice (visual, etc.). It can be used to know whether or not a voice interface system is appropriate for situations such as walking or driving a car.

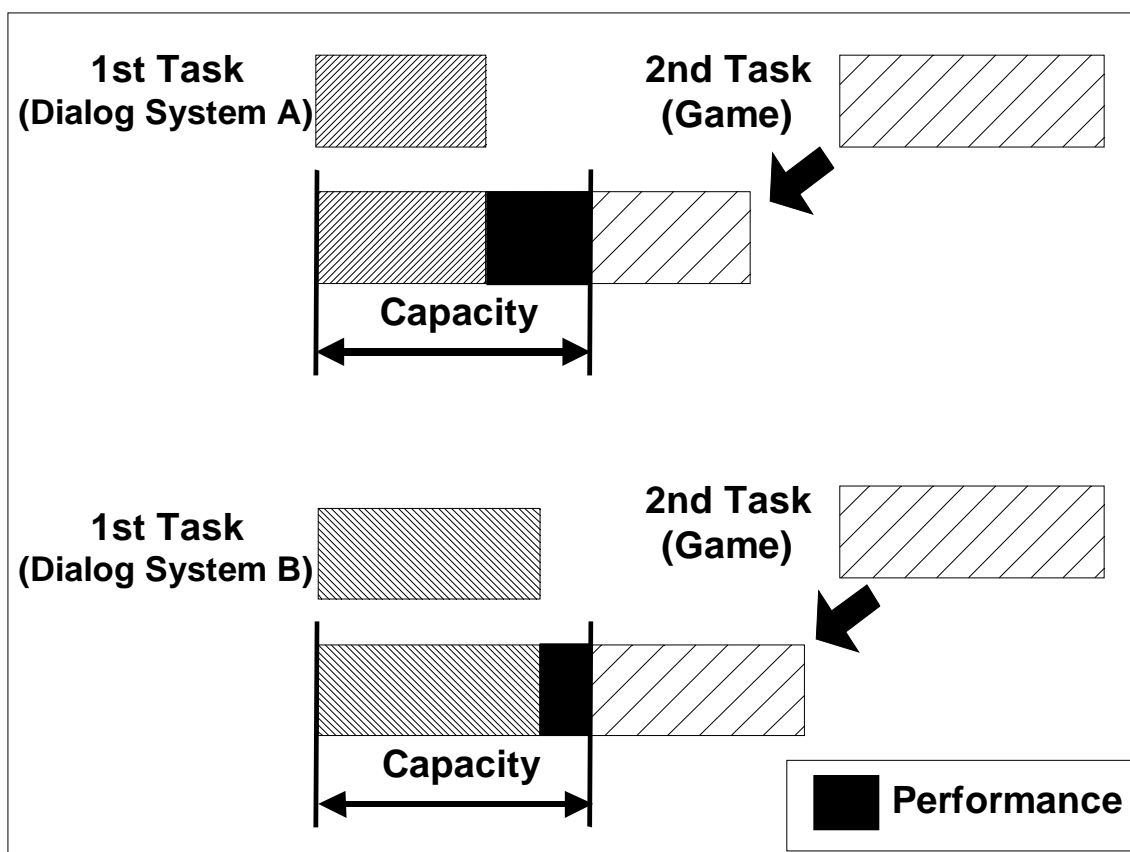
To make the spoken dialog system easy to use, it is important that the interaction design is planned in the early stages of development. Evaluations by users and improvements by the developer are on-going tasks that must be repeated [1, 2]. This can be performed efficiently without using the subjective opinions of users if the dialog workload is analyzed section by section to evaluate which parts need improvement. We propose a method of measurement for assessing the dialog workload in voice interface systems, followed by basic evaluations.

Dual-task methods are commonly used for the measurement of the level of attention to an object [3, 4, 5]. While performing a task, humans may have to divide their attention when other tasks are added. The Law of Fixed Capacity states that the human mental focus is limited, although to a certain degree it is possible to divide mental attention among different tasks. When the amount of the workload exceeds the upper limit of mental resources, a longer response time and/or an increase in errors may occur while doing the tasks.

The organization of this paper is as follows: The second section contains related works and unsolved problems in evaluating voice user interface systems using dual-task methods. The third section explains our method. The fourth section describes the preliminary experiments for testing the effectiveness of the proposed method. The fifth section discusses the experiments applying our method on a spoken dialog system. The sixth section states the summary and shows the problems that need to be solved.

2. Related Works

We propose a method for measuring the dialog workload of voice interface systems using the dual-task method (Figure 1).



(Figure 1: Measurement of workload for voice interface systems.)

The use of the voice interface system is regarded as the primary task, and a video game task or a driving simulator task in which the workload can be measured objectively is regarded as the secondary task. Subjects are asked to perform those tasks simultaneously. In situations where the performance of the secondary task is poor, the workload of the voice interface system is judged to be relatively high. Thus in Figure 1, the smaller black performance bar shown in dialog system B indicates that the task workload is heavier than that of the dialog system A.

To perform such experiments, the following conditions must be met: (1) The Law of Fixed Capacity applies, (2) The subjects can intentionally give priority to the primary task, and (3) The differences of performance in the secondary task are easily compared and reliable.

There are several related works which use multiple and simultaneous tasks that are specifically planned for the evaluation of voice user interface systems in the car driving environment.

Kojima et. al. [6, 7] performed subjective experiments with three tasks including car-driving, reactions to stimuli of LED displays, and shadowing of words.

Their results showed that the shadowing task affects the performance of the stimuli-reaction task (no significant difference in average reaction time was observed, but the ratio of trials in which reaction times were relatively long increased.) Their experiments have shown the weak points of using spoken dialog, but comparisons of the strength of influence among the dialog patterns have not been made.

Shimizu et. al [8] evaluated spoken dialog systems for car drivers from the viewpoint of safety. In their experiments, subjects were asked to perform three tasks including car-driving, reaction to the stimuli of LED displays, and spoken dialog task; however, the differences in influence among the dialog patterns were not shown. In their experiment, subjects drove a car on a circuit while performing other tasks including the stimuli reaction task and the spoken dialog task (traffic information retrieval). Their secondary task corresponds to both the driving task and the reaction task. It is difficult, however, to manage the priorities of each task, so the performance of the reaction task does not directly reflect the dialog system workload. Another difficulty is that unlike the reaction task, driving performance can not be objectively assessed.

Strayer et. al [9] investigated using the dual task method how conversation with mobile phone affects driving task. In their experiment, subjects perform a simulated driving task (steering on a course generated by adding three sinusoidal waves), a reaction task (when the red signal turns on, the subjects depress the brake pedal as soon as possible), and a spoken dialog task. Their results showed that conversation on a mobile phone had negative effects on both the simulated driving task and the reaction task. It was also shown that the difference in mobile phone shape (e.g. hands-free type, handheld type, etc.) does not have an effect on the task performances. They also reported that the word shadowing (repeating) task does not have affect on the driving task, but the word generation task (associating new words with given words) may affect driving.

Their study proved that the dual task method is useful between the primary task involving both visual attention and hand operation and the secondary task involving voice conversation. Their experiments, however, involve a triple task condition including steering and reaction to signals, not a dual task situation. Their method of workload measurement may not be appropriate for the evaluation of the word shadowing task because it can only measure a relatively large workload.

In a related experiment, Uchiyama et. al. [10] made an attempt to quantify the driver's capacity using a dual task method. They developed a voice interface system which adapts to driver's situation, estimating the driver's workload capacity using driving conditions. It displays information using spoken dialog when the driver is

expected to have adequate capacity.

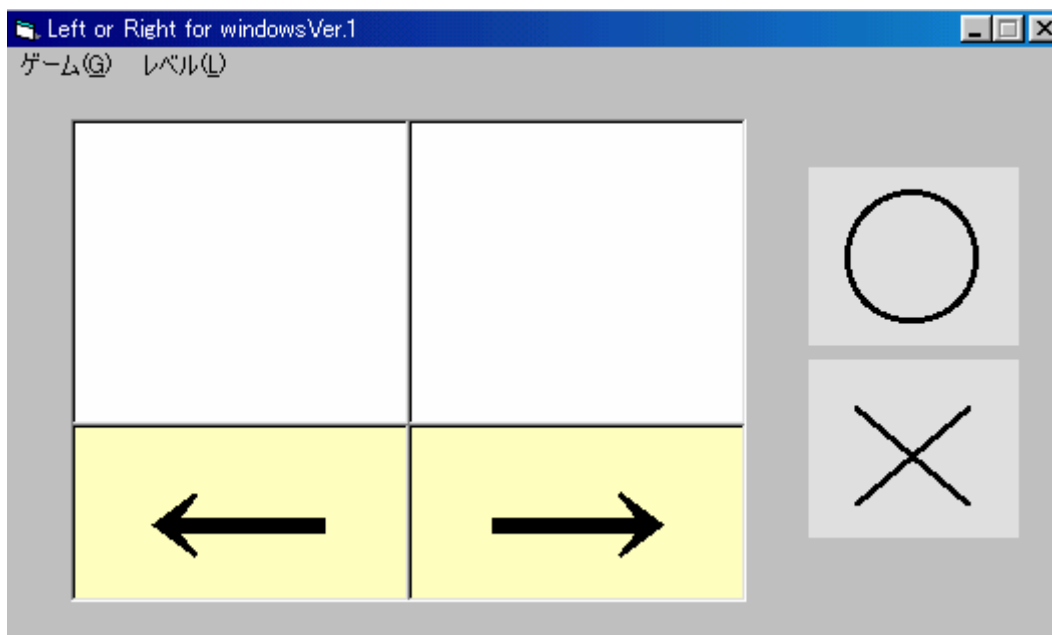
3. Proposed Method

From the viewpoint of comparing dialog workloads, priority is given to the following points in choosing the secondary task;

- (1) The results of the performance can be obtained periodically under the same conditions. In this situation, the similarity between the real task such as driving and the experimental task is not regarded as important.
- (2) The task has a certain level of difficulty so that practice will have little effect on the results.
- (3) The subjects can intentionally give priority to the primary task, i.e. the use of the spoken dialog system.

Using our experimental method, it is expected that we can specify the positions where the dialog workloads are relatively high in a dialog system. Even if the dialog system is designed for drivers and the secondary task in the real environment is car-driving, the magnitude relationship of the dialog workloads among the positions in the dialog may be the same as that under the conditions of our proposed method.

From such a viewpoint, we chose a PC game called 'the Left or Right' as the secondary task in our experiment [11, 12]. The screen image of our game is shown in Figure 2. The software is developed with Microsoft Visual Basic 6.0 and the target operating system is Microsoft Windows 2000.



(Figure 2: Screen image of the game used for workload measurement.)

There are two blank boxes with arrows beneath them on the left of the window. When a left pointing arrow appears in the left box, or a right pointing arrow appears in the right box, the player indicates “correct” by pressing the upper key. When the upper and lower arrows point in different directions, the player presses the lower key to indicate “incorrect.” A random arrow then appears one second after pressing the upper or lower arrow key, irrespective of success or failure. During the play, all events such as displayed arrows and pressed keys are recorded with the player’s response time measured in milliseconds.

The experimental procedure is as follows; at first, the subjects practice the game until they are fully accustomed to it. After this, the subjects perform the spoken dialog task as the primary task and the game as the secondary task. In advance of the experiment, the subjects are instructed to give priority to the spoken dialog task during the dual task condition, and respond as quickly and correctly as possible.

After the experiment, response times of each trial are statistically processed factoring in the level of difficulty of the spoken dialog task that was performed simultaneously. We judge the dialog workload during a period of time as significantly high, if the game response time during the period is significantly longer than those of the others.

4. Preliminary Experiment

To validate the proposed method, we performed a preliminary experiment to affirm that the workloads of the two spoken dialog tasks are different. The workload of the spoken dialog is determined taking into account both the cognitive and the physical workloads. Factors such as listening to system messages, remembering messages, judging the time to allot for response, and activation of speech organs to make utterances, etc. affect the difficulty of task and the response time.

As a simulated dialog task which produces such a workload, we use the listening and repeating of words task (word shadowing task). The number of words which a subject remembers simultaneously determines whether the workload difference is significant. This experiment was designed along the lines of the research of Uchiyama et. al. [10], although where Uchiyama's group used the word shadowing task as the secondary task, we used it for the primary task.

4.1 Procedure of Experiment

The examples of the word shadowing task are shown in Table 1 and 2. Two or four Japanese words are spoken to prompt the subject's answers. Followed by a beep sound, the subject answers remembered words one by one. The order of answered words is not restricted. The procedure mentioned above is counted as one trial of the word shadowing task. Approximate time required for one trial is 20 seconds. The trials are repeated followed by a time interval which is selected randomly between 10-20 seconds. The scheduled intervals between prompts and answers are selected so that the total duration of the different tasks (Table 1 and Table 2) is approximately same, although the timing of the individual steps within the tasks is not.

Table 1: Words shadowing task - 2 words.

Prompts (in Japanese)	User's responses
"Robotto" (robot) (followed by 3.5sec silence)	
"Taremake" (banner) (followed by 3.5 sec silence)	
(beep sound)	
(6 sec silence)	"Taremake"
(beep sound)	
(2 sec silence)	"Robotto"

Table 2: Word shadowing task - 4 words.

Prompts (in Japanese)	User's responses
"Robotto" (robot) (followed by 1 sec silence)	
"Taremake" (banner) (followed by 1 sec silence)	
"Keshigomu" (eraser) (followed by 1 sec silence)	
"Sukiyaki" (sukiyaki) (followed by 1 sec silence)	
(beep sound)	
(2 sec silence)	"Taremake"
(beep sound)	
(2 sec silence)	"Sukiyaki"
(beep sound)	
(2 sec silence)	"Robotto"
(beep sound)	
(2 sec silence)	"Keshigomu"

The vocabulary of the word shadowing task consists of 166 words. Japanese common noun words whose length is from two-moras to four-moras are randomly selected so that the frequency of appearances of the prefix syllabary characters is balanced. Both words start with dull sounds, and words which are difficult to hear clearly are removed from the vocabulary. They are randomly ordered and spoken using a commercial Text-to-Speech synthesis system (Toshiba LaLaVoice 2001).

The subjects include nine men and a woman between 20-25 years old. They are divided into two groups which are named Group-A and Group-B. The subjects of Group-A perform the 4-word tasks first, followed by the 2-word tasks. Those of Group-B perform the 2-word tasks first, followed by the 4-word tasks. The experimental procedure of Group-A is as follows (in Group-B, steps 3-4 and steps 6-7 are replaced);

1. The subjects listen to an explanation of the game, and practice the game.
2. They perform the game task with no other tasks (for eight minutes).
3. They practice the 4-word shadowing task until they are fully accustomed to it.
4. They perform the 4-word shadowing task and the game task simultaneously (for eight minutes).
5. They take a rest.
6. They practice the 2-word shadowing task until they are fully accustomed to it.
7. They perform the 2-word shadowing task and the game task simultaneously (for eight minutes).

4.2 Results and Discussion

In preparation, five male subjects, 20-25 years old, performed the first half of the procedure previously described (i.e. from step 1 to step 4). The response time of the game task increased significantly (with a risk rate of 1% for all subjects) during the performance of the 4-word shadowing task. In other words, the dual task condition required more time than the single task. Thus, we confirmed that the Law of Fixed Capacity can be applied to the proposed method. As the performance of the voice-related task during the performance of the game task may frequently exceed the capacity of human resources, the workload of the secondary task was considered appropriate.

Table 3 compares the 2-word and 4-word tasks. The ** mark indicates a significant difference with a risk rate of 1%. A significant difference was observed in nine out of ten subjects.

These results show that the proposed method can compare the dialog workloads of two different tasks.

Table 3 : Average response time (msec) in preliminary experiments.

Subject	2-word task	4-word task	Significance
A1	542	1191	**
A2	458	589	**
A3	486	675	**
A4	496	683	**
A5	426	562	**
B1	695	1000	**
B2	401	418	-
B3	523	868	**
B4	712	941	**
B5	478	519	**

5. Evaluation of the Spoken Dialog System

In the preliminary experiment, the subjects were asked to listen to the system prompts and to repeat the prompts. In this chapter, we evaluate the dialog workload of a spoken dialog system using our method.

5.1 Outline of the Dialog System

We created a restaurant database-search application to be used while driving. The Japanese version of the Nuance Voice Web Server (VWS) released by OMRON Corp. was used as the platform for operating the spoken dialog system. The subject uses a headset-microphone connected with the notebook-type personal computer. The VWS requires a client software that can input or output speech. We choose Pingtel Instant Xpressa as the client for our experiment.

The dialog application is written using VoiceXML 1.0 (<http://www.w3.org/Voice/>). The dialog is controlled with static documents which use the standard functions in VoiceXML specifications such as state-transitions and variables. The application uses a Japanese text-to-speech synthesizer to generate the responses from the system. We employed the VWS speech recognition and speech synthesis.

"Barge-in," or the user's interruption during the system utterances, is always enabled. Input is limited to voice responses. Other methods such as keyboards or pointing devices are not used. The speech recognition vocabulary consists mostly of isolated words. The recognition grammars are selected depending on the dialog states from 24 grammars, which contain an average of 8.3 words and a maximum of 30 words.

5.2 Dialog States

The subjects are asked to perform the following task: "Search for restaurants within 10 minutes of their current position, find the lowest budget Chinese restaurant and reserve seats." The expected dialog states to perform the task can be divided into six states S1-S6.

S1: [Main menu] The system shows various choices, and the user selects "Neighborhood Town Guide."

S2: [Neighborhood town guide] The system requests search conditions such as the category or approximate arrive time. In this task, the user selects "restaurant" for the category and "within 10 minutes" for the arrival time. The user is allowed to input both items in any order. At this dialog state, the user is also allowed to input two items in one utterance, such as "restaurants within 10 minutes."

S3: [Detailed category] The system shows the number of candidate restaurants. The user then adds a search condition to select only Chinese restaurants.

S4: [Price choice] The system shows a price choice: "higher or lower than 1500 yen a person." The user selects the latter option.

S5: [Candidates and details] The system shows the total number of restaurants which meet the conditions and lists their names. The user selects a restaurant by speaking its name. The system then shows detailed information about the restaurant, such as recommended menus and the average price level. The final choice is whether the user wants to reserve seats at the restaurant or not. If the user answers "No", the system repeats from the beginning of S5. If the user answers "Yes", the system moves to S6.

S6: [Reservation and exit] The system reserves the seats and exits.

5.3 Experimental Method

The subjects include four men and a woman between 20-25 years old. All the subjects have experience with other spoken dialog systems, although they have never used this one. The subjects first practice the game until they are fully accustomed to it.

Then they perform the restaurant database-search task only once. They are instructed to give priority to the spoken dialog task.

After the experiments, we analyzed the recorded response times and conversations. All states from S1 to S6 correspond to the dialog patterns written in VoiceXML. We analyzed average response times of the secondary task for each state.

5.4 Results

All subjects finished the given task. Average time to finish task was 343 seconds. The number of users' utterances ranged between 16 and 22. The average number of rejected users' utterances was 1.2 and the maximum was 5.

All game response times were used for analysis regardless of the correctness of the answers. Response times (R) showed deviations; R' represents the moving averages of 5 sequential samples from R. Figure 3 shows the R and R' from the beginning to the end of a dialog session. The peaks of the moving averages (R') are more significant than the original values (R).

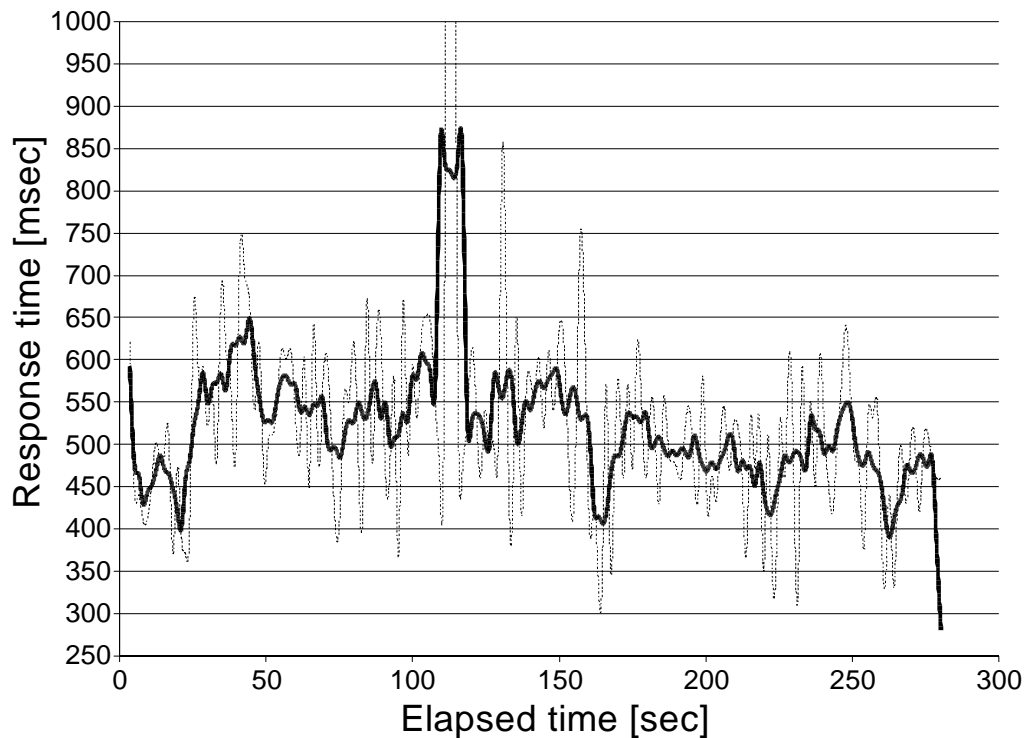


Figure 3: Example of response time while the subject is performing spoken dialog (thin

line: observed value R , bold line: moving average R)

Figure 4 shows the distributions of R' of all subjects with respect to each dialog state. The box shows the median value and the range between 25-75%. The whisker plots show 10% and 90%, and the outskirt samples are plotted at both ends. Significant response delays were observed in states S2 and S5, as judged from the outskirt samples.

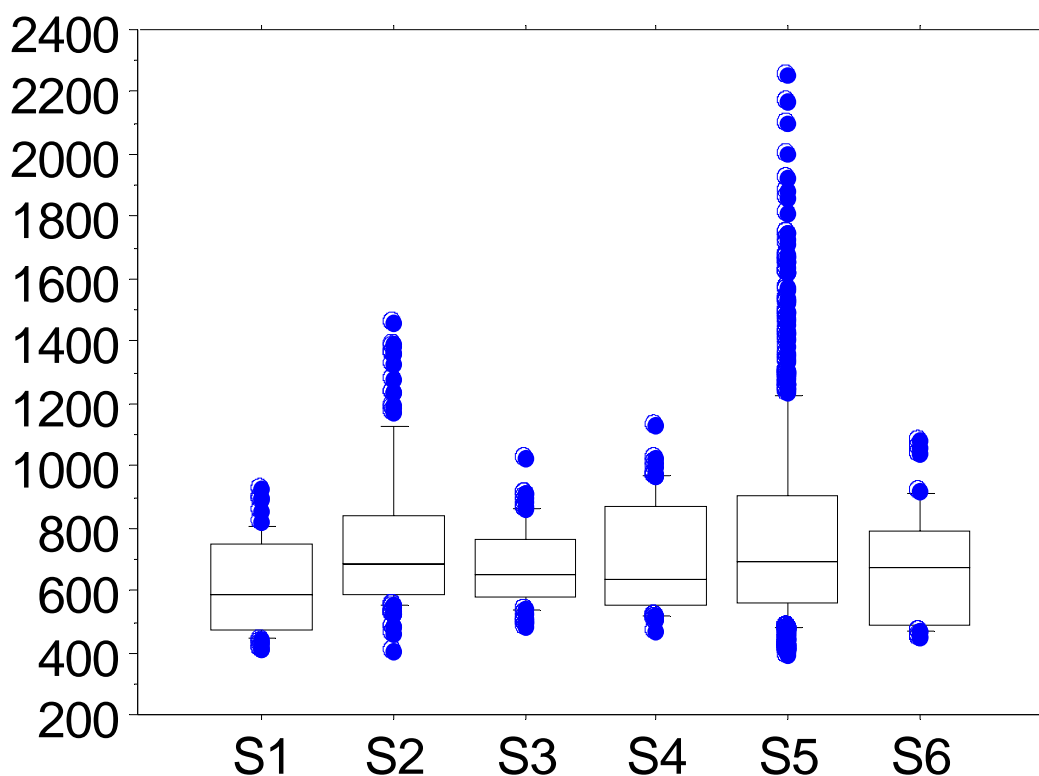


Figure 4: Distribution of response time (msec) in each dialog state.

The effect of interaction was significant ($F = 2.42$) according to the analysis of variance of R' with the subjects and the dialog states (S1-S6) as factors. This indicates the effects of the difference of subjects and dialog states on the response times are not independent [13]. Figure 5 shows the averages of R' with respect to each subject and dialog state. Table 4 shows the significance of the simple main effect of the different dialog states. F ratio indicates the ratio to the chance fluctuation. These results indicate that significant long average response times are observed in the following conditions: state S2 for subjects C1 and C2, and state S5 for subjects C4 and C5.

Table 4: Simple main effect of difference of dialog state.

Subject	F ratio	p	Significance
C1	$F(5, 177) = 3.380$	$p = .0061$	**
C2	$F(5, 244) = 7.222$	$p < .0001$	**
C3	$F(5, 151) = 1.591$	$p = .0658$	
C4	$F(5, 157) = 3.010$	$p = .0127$	*
C5	$F(5, 222) = 4.977$	$p = .0002$	**

5.5 Discussion

Kojima et. al. [6, 7] compared the delays of response times showing a difference in the single task and the dual task conditions. The results shown in Figure 4 indicate that the response times tend to increase at the dialog states S2 and S5, indicating that our method can be used to go one step further to compare dialog states or dialog patterns under the dual task conditions.

Figure 5 and Table 4 show the subject-ability effect. Our focus on the average response times was regarded as appropriate due to the significant differences observed among the different states, which are not obvious in the works of Kojima et. al.

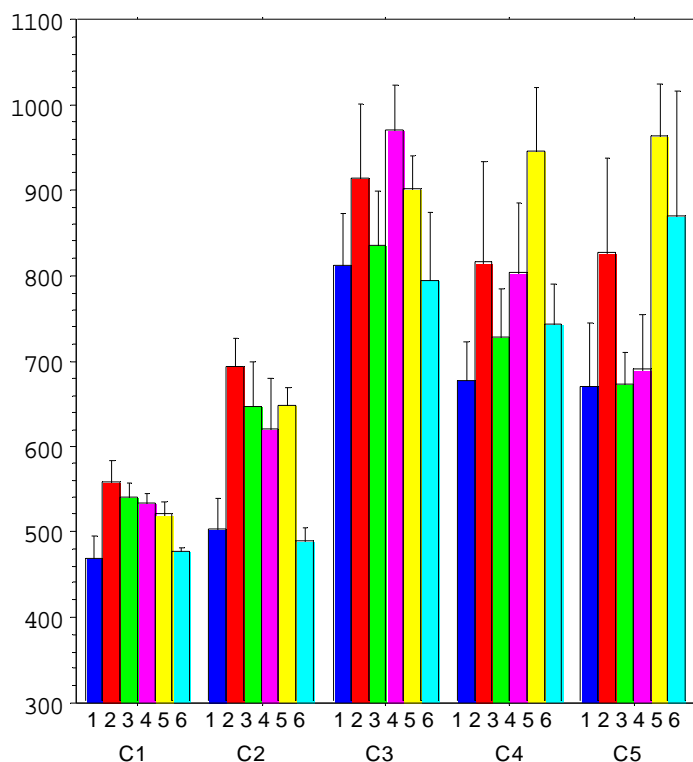


Figure 5: Average response time (msec) for each subjects C and dialog state S .

The difference between the main experiments and the preliminary experiments is that the performance of the main experiments depends on subject ability, but the tasks of the preliminary experiments were too simple to reflect subject ability. Preliminary experiments in Chapter 4 thus showed that there are no significant differences between the subjects' ability using our experimental model. Therefore, the experiments in Chapter 5 show that the dialog workloads of the real spoken dialog systems may change depending on subject ability such as the knowledge and the experience of individuals.

The appropriateness of our experiments can be proved if the dialogs in states S2 and S5 have adequate reasons of the delay of response time that was significant in the cases of two subjects respectively. The appropriateness of our experiments can be proved if the delays of response times seen in dialog states S2 and S5 are related to the difficulties of the tasks, as observed in the comparisons of two subjects.

In the dialog S2, the users fill two slots; the category and the arrival time. Figure 6 shows the prompt in dialog state S2. The system accepts the users' utterances such as "10 minutes," "Restaurant," or "Restaurant within 10 minutes." After the users'

utterances if they only specify one condition such as the time only or the category only, the system asks for the other condition. This increases the dialog workload because the users may have trouble finding the appropriate words to say, while in the other situation the users merely select a voice command from the list of the candidate words spoken as a prompt.

In dialog S5, the users listen to the names of restaurants, select a restaurant from the candidates, and say its name as a voice command. Figure 7 shows the prompt in dialog S5. If the user is not familiar with the name of the restaurant, or if the user cannot hear the name clearly, suitable voice commands cannot be chosen easily. For this reason the dialog workload is expected to increase.

Another factor that increases the dialog workload in state S5 is that the user must remember the name of the lowest priced restaurant and must also continue to compare price levels while listening to information about the next restaurant.

In summary, our method for obtaining the value R' can be measured objectively, and the value increases significantly when the dialog workload is increased. Therefore, we conclude that the proposed method is effective for searching for locations where the dialog workloads are relatively high.

Neighborhood town guide.
 The search conditions are category and arrival time.
 Category: restaurants, amusements, sports and advertisements.
 Arrival time: within 5 minutes, 10 minutes and 15 minutes.
 Please give a voice response for a category name and the arrival time.

(Figure 6: Prompt in dialog state S2.)

There are four candidate Chinese restaurants within 10 minutes with price levels under 1500 yen a person.
 The names are Ohsho, Miyoshi, Kowloon, and Tenka Ippin.
 Please give a voice response to ask for more information.

(Figure 7: Prompt in dialog state S5.)

6. Conclusion

This paper proposed a dual-task method to measure the workload of spoken

dialog tasks. In this method, subjects play a game using visual display and keyboard input. We evaluated the effectiveness of our method with a word shadowing task and a spoken dialog application.

One of the differences between our approach and that of others is that our evaluation is not based on the premise of use during car-driving. In contrast, our method can measure relatively small workload differences, such as in word shadowing tasks, which are difficult to measure by the method of Strayer's group [9] which uses a driving simulator task. Also, our method can identify positions in the dialogs which cause some users significant difficulty. Other groups [6-8] have not succeeded in comparing dialog workloads between dialog patterns or in finding positions where the workload is high.

Possible explanations of our success are as follows; (1) Our task is set up so that the subjects can intentionally give priority to the primary task, i.e. the use of the spoken dialog system, (2) The difficulty of the game task is appropriately large for this experiment, whereas the challenge of the workload in the other group's task [8], such as the LED stimuli reaction task, is too small, (3) The required time for one trial (i.e. the reaction time) is under 3 seconds, making it possible to obtain a large amount of data within a short period of time, (4) Subjects can be fully acclimated to our game quickly and easily.

Our method identifies positions where the dialog workload is high in short time intervals. It is especially useful for experiments in which the time required for a dialog task is long and when the experiment is limited to only one subject who performs the task repeatedly.

The method or tools we used to perform such experiments have yet to be fine-tuned. Moreover, the similarity of the game task and the dialog task may cause serial confusion. For example, in a route guide dialog which includes directions such as "turn left" or "turn right", our secondary task is similar to the primary task because both are related to the decision of left or right. We are also interested in factors such as physical and cognitive workloads in the spoken dialog, as discussed in Chapter 5. Guidelines or design templates need to be made for improving the spoken dialog interface based on such investigations. Finally, the discrepancy between the experimental workload and that in the real world during car-driving situations, and how this relates to the safety of voice telematics remains to be assessed.

Acknowledgements:

We thank the members of committee and the working group of the Network Voice

Telematics Project by the Association of Electronic Technology for Automobile Traffic and Driving (JSK: reorganized into JARI). Our special thanks are due to Professor Tetsunori Kobayashi of Waseda University for guiding our work. We also thank Dr. Hiromichi Hosoma of the University of Shiga Prefecture, Professor Emeritus Yasuhisa Niimi of Kyoto Institute of Technology, and the members of Pattern Information Processing Group, Kyoto Institute of Technology for useful advice and support.

References:

[1] Bruce Balentine, Devid P. Morgan: "How to build a speech recognition application," 2nd Ed., EIG Press, California, 2001.

[2] Hiroshi Daimoto, Hirohide Ushida, Hiroshi Nakajima, Tsutomu Ishida:
"A Practical Report about UI Design and Assessment for Telephony Voice Application,"
Correspondences on Human Interface, Vol.3, No.4, pp.35-40, 2001 (in Japanese).

[3] Byron Reeves, Clifford Nass:
"The Media Equation," CSLI, Cambridge, 1996.

[4] Michael W. Eysenck, Mark Keane:
"Cognitive psychology -- A student's handbook, "
4th Ed., Psychology Press, East Sussex, 2000.

[5] Nick Lund:
"Attention and pattern recognition,"
Routledge, Philadelphia, 2001.

[6] Shin'ichi Kojima, Takero Hongo, Hiroyuki Hoshino, Yuji Uchiyama:
"Development of an Evaluation Method for Verbal Interface in Driving,"
Proceedings of Annual Spring Congresses, Society of Automotive Engineers of Japan,
No. 91-99, pp. 17-20, 1999 (in Japanese).

[7] Shin'ichi Kojima, Takero Hongo, Hiroyuki Hoshino, Yuji Uchiyama:
"Development of an Evaluation Method for Verbal Interface in Driving,"
IPSJ SIG Technical Reports, 1999-MBL-010-010, 1999 (in Japanese).

[8] Tsukasa Shimizu, Shin'ichi Kojima, Toshihiro Wakita, Takero Hongo:

"Evaluation of Spoken Dialog Systems for a Vehicle,"
 IPSJ SIG Technical Reports, 2000-SLP-32-17, pp.87-92, 2000 (in Japanese).

[9] David L. Strayer, William A. Johnston:

"Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular phone,"

Psychological Science, 12, pp. 462-466, 2001.

[10] Yuji Uchiyama, Shin'ichi Kojima, Takero Hongo, Ryuta Terashima, Toshihiro Wakita:

"Voice Information System Adapted to Driving Situation,"

Proceedings of Symposium on Mobile Interactions and Navigations,
 pp.11-15, Nagoya, 2001 (in Japanese).

[11] Takuya Nishimoto, Motoki Takayama, Masahiro Araki:

"An Evaluation Method of Cognitive Load for the Voice Interface Systems,"

IPSJ SIG Technical Reports, 2002-SLP-45-5, pp.29-34, 2003 (in Japanese).

[12] Takuya Nishimoto, Motoki Takayama, Masahiro Araki:

"An Measuring Method of Cognitive Load for the Voice Interface Systems,"

Proc. of the 2003 Autumn Meeting of the Acoustical Society of Japan,
 2-4-13, pp.83-84, 2003 (in Japanese).

[13] Satoshi Tanaka:

Practical Psychological Data Analysis, Shin'yosha, Tokyo, 1996 (in Japanese).

Profiles:

Takuya Nishimoto:

He received B.E. and M.E. from Waseda University, Japan in 1993 and 1995, respectively. He is a Research Associate at the Graduate School of Information Science and Technology, The University of Tokyo, Japan. Research interests include spoken dialogue systems and human-machine interfaces. He is a member of Information Processing Society of Japan, Acoustical Society of Japan, Japanese Society for Artificial Intelligence, and Human Interface Society.

Motoki Takayama:

He received B.E. and M.E. from Kyoto Institute of Technology in 2001 and 2003, respectively. He joined

Ishida Co.,Ltd. in 2003.

Haruaki Sakurai:

He received B.E. and M.E. from Kyoto Institute of Technology in 2003 and 2005, respectively.

Masahiro Araki

He received M.E. in information science and D. Eng. from Kyoto University, Kyoto, Japan, in 1990 and 1998, respectively. He is an Associate Professor at Department of Electronics and Information Science, Kyoto Institute of Technology. His research interests include spoken dialogue systems and natural language understanding. He is a member of IPSJ, JSAI, ANLP, and ACL.