Preliminary Study of a Learning Effect on Users to Develop a New Evaluation of the Spoken Dialogue System

Sunao Hara[†] **Ayako Shirose**[‡] **Chiyomi Miyajima**[†]
Katsunobu Ito[†] **Kazuya Takeda**[†]

[†]Graduate School of Information Science, Nagoya University
Furo-cho, Nagoya JAPAN
{hara, miyajima, k-itou, takeda}@sp.m.is.nagoya-u.ac.jp
[‡]Faculty of Human Sciences, Waseda University
Mikajima, Tokorozawa, Saitama JAPAN
shirose@aoni.waseda.jp

Abstract

In order to address a learning effect of the Spoken Dialogue System (SDS), we carried out field experiments using SDS of music retrieval for long periods and analyzed sequential shifts of the system performance and the users' utterances. Experimental results revealed that the user of SDS changed the strategy of dialogue with the system during the learning process. Analysis of the questionnaire suggested that the user changed the evaluation of SDS according to the learning experience. The leaning effect of SDS is known a priori, however, our results displayed a new form of the learning effect. We discussed that it may be taken into consideration to evaluate SDS users as well as the system.

1 Introduction

Although recent technology in speech recognition has shown high performance, the Spoken Dialogue System (SDS) has yet to become in popular use. The disadvantage associated with the usage of SDS has been assumed to be derived from the SDS performance itself, hence considerable research discussed the evaluation of SDS performance and users' impression of SDS to overcome the disadvantage. In contrast to SDS, it is rare to discuss the importance of user 's skill, but we can postulate that this quality represents the key to efficient usage.

It is empirically apparent that although experts, such as the SDS developer, can manage their system without difficulty, most users are confused when interacting with SDS. Their problem with SDS seems to be derived from limited experience and knowledge in the use of SDS. If they learn skills for using SDS, a more active uptake is expected. However, it is still unknown whether the user will learn the skill of using SDS or not and how to learn it and we should clarify whether and how the user learns the skill.

Thus, in order to address the SDS usage learning, we carried out field experiments using SDS of music retrieval and analyzed users' performance and the learning process of using SDS for long periods.

2 Music Retrieval System

2.1 Spoken Dialogue Interface for Music Retrieval

Our interface enables users to retrieve and listen to their favorite music through conversation with a spoken dialogue system. During the dialogue with the system, the user utters the title and artist name of the target song and the command words, e.g., "Next music" and "Stop playing". The system processes the user 's utterances, then subsequently retrieves and obtains a music list, including the target music, via the Internet. The music list is read out by a text-to-speech synthesis system. The users are able to enjoy listening to music without viewing the list or pushing any buttons.

2.2 Implementation of Modules

The system retrieves music from the Japanese commercial online music store "Mora" (http://mora.jp) on the Internet. It has several methods for retrieving music, e.g., by matching keywords, matching the first-letter of artist names, specifying a genre of music, and choosing music from a popularity ranking list. The method involving matching keywords was used in our interface.

The speech recognition module used an open-source Large Vocabulary Continuous Speech Recognition (LVCSR) engine, " Julius 3.1p2-sp4 "(Kawahara et al., 2004). We used a gender-independent acoustic model of Phonetically-Tied Mixtures (PTM) with 3,000 states (129 codebooks) and 64 Gaussians. Language models of a word bigram and a backward trigram were trained with the Pub-

licly Available Language Modeling Toolkit (Palmkit; http://parlmkit.sourceforge.net) using sentences generated from a grammar network.

The dictionary for speech recognition contains 7,710 words including 1,601 artist names, 6,071 music titles and command words used for retrieval. The artist names and titles were collected from the web site of Mora on Sept. 24, 2003 (1,404 artist names and 5,862 music titles) as well as from the top-ranked web site music lists ("Oricon"; http://www.oricon.co.jp) from October 2002 to September 2003.

3 Experiment

3.1 Procedure

Twelve subjects participated in our experiment. They were asked to use the music retrieval system for five days for about one hour every day. The one-hour session includes three experimental conditions. The detailed procedure of the experiment was as follows:

1. Speech training by reading newspaper articles (10 minutes)
2. Condition 1: Retrieval in a laboratory (20 minutes)
3. Condition 2: Retrieval while driving in a simulator (15 minutes)
4. Condition 3: Retrieval while driving on city streets (15 minutes)
5. Questionnaire about the interface

The text used for speech training included 1,714 sentences in 191 articles in a Japanese newspaper (Asahi Newspaper). Conditions 2 and 3 were carried out in consideration of the application to car navigation equipment. The subject's speech and action were recorded using PCs (Dell Inspiron 5150 and Sony VAIO PCFV505R/PB), audio interfaces (M-Audio Mobile Pre USB and EDIROL UA-5), a contact microphone (Sony ECM-77B) and DV Camera(Sony DCR-TRV900). The noise level of the room was about 21.2dB(A).

3.2 Questionnaire

The subjects were asked to evaluate the system's usability and impression via a questionnaire after using the system. The questionnaire was designed based on previous research (Möller and Skowronek, 2003). The criteria in the questionnaire are listed in Table 1.

Table 1: Questionnaire

Criterion	Question
EVERYDAY	
Input Quality	*How well did you feel understood by the system?*
Informativeness	*Was the information complete or incomplete?*
Task Success	*Did the system provide the desired information?*
User Satisfaction	*Are you satisfied with the dialogue?*
Ease of Use	*Did you comprehend the handling of the system?*
1st & 5th DAYS	
Dialogue Conciseness	*Did you perceive the dialogue as natural or unnatural?*
Dialogue Smoothness	*Was the course of the dialogue smooth or bumpy?*
User Dialogue	*How well did you feel your own dialogue?*
Pleasantness	*Did you enjoy the dialogues?*

4 Data Analysis

4.1 Summary of Data

We analyzed the utterances produced by the twelve users in Condition 1. The total number of utterances was 5,141. The rate of occurrence of Out Of Vocabularies (OOVs) was 15.7. The average Word Correct Rate (WCR) was examined, including OOVs and excluding them: the average WCR of whole words was 62.7 and that excluding OOVs was 76.9. Figure 1 shows a histogram of WCR of five days' data.

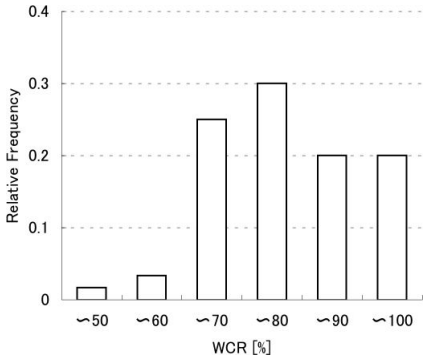

Figure 1: Histogram of Word Correct Rate.

In addition to WCR, the correspondence between the content of the user's utterance and the result of recognition was examined. It is assumed to indicate the system performance more

precisely than WCR. This index, Content Correct Rate (CCR), showed 77.4.

4.2 Sequential Shift of Performance

Sequential shifts of the system performance are considered to examine whether and how the users were changed according to successive use of the system.

Figure 2 shows the sequential shift of the average WCR. This was at a level of 70 on the first day and reached about 85 on the final day. Figure 3 shows the sequential shift in the average CCR, whereas it is also clear that CCR also rose over five days.

These data indicated that the system performance improved for five days, namely, according to the user learning experience. This result suggested that the user learned how to interact with the system and that the rate of recognition increased over the five-day session.

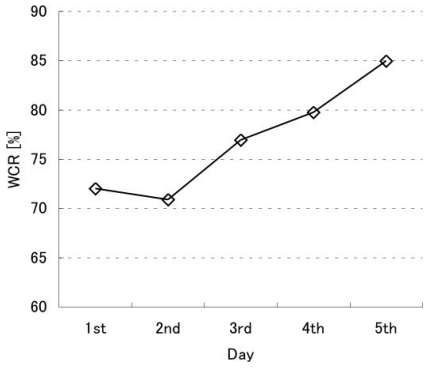

Figure 2: Sequential shift of WCR.

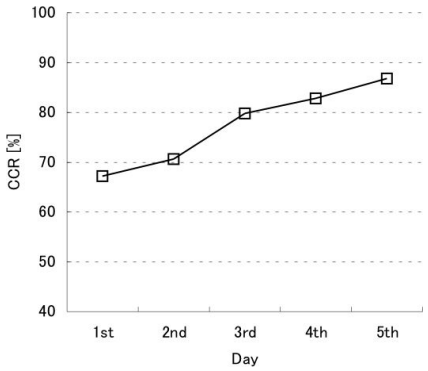

Figure 3: Sequential shift of Content Correct Rate.

4.3 Learning Effect

Since the above result shows potential for users learning how to use the system, it is also worthwhile considering what aspect they learn. In order to investigate the aspects of the learning effect, the users' utterances were analyzed in view of their phenotypes.

Figure 4 shows the occurrence of OOVs for five days. OOVs occurred around 15 with the fewest on the third day; seemingly likely to decrease during sessions. Figure 5 shows the sequential shift of perplexity, which increases till the 4th day. The increase in perplexity is caused by the user's production of complex grammar, in other words, utterance with many candidates. The user produces the music titles, artist names and commands to control the system during dialogue with our system. The music titles and artist names have more candidates than the commands, hence the increase in perplexity is presumably due to the expansion of the user's vocabulary of music titles and artist names.

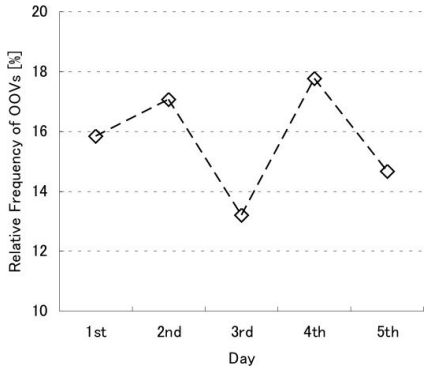

Figure 4: Sequential shift of occurrence of Out Of Vocabularies.

Figure 6 shows the occurrence rate of fillers over five days. The term 'fillers' here indicates expressions and exclamations such as "Umm", "Err" and so on. The occurrence of fillers decreases over five days, and specifically, bottoms out on the third day. Figure 7 shows the occurrence rate of utterances produced twice or more. The rate indicates how many times the user repeated the utterance of the same word. This figure shows that the rate of repeated words bottoms out on the third day as well as the forth and fifth days. It is speculated that the user repeats the utterance when it is not recognized by the system correctly. The reason why the rate of repeated words decreased was that the user learned how to produce utterances that were easily recognized at once and that they stopped producing the word again, with the possibility of OOVs in mind.

Our experimental results revealed that the user changed the strategy of dialogue with the system: the increase in perplexity indicating the

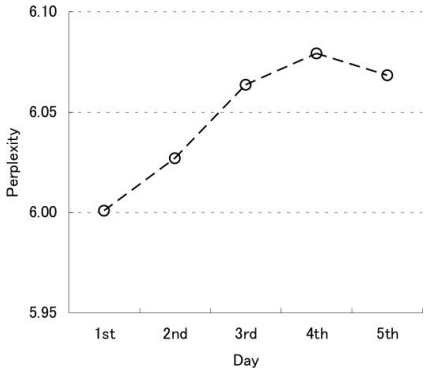

Figure 5: Sequential shift of perplexity.

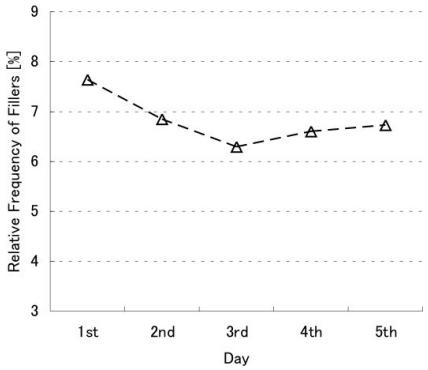

Figure 6: Sequential shift of occurrence of fillers.

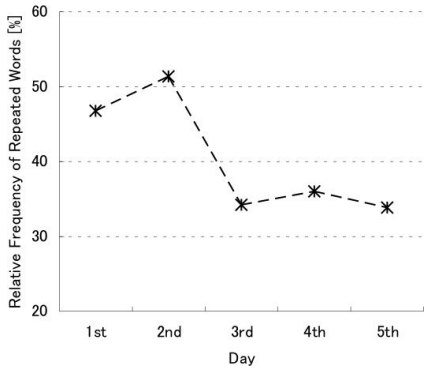

Figure 7: Sequential shift of occurrence of repeated word.

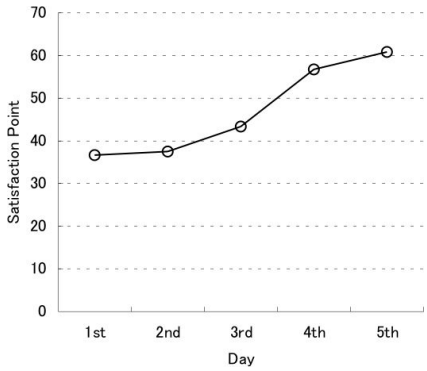

Figure 8: Sequential shift in the evaluation of "User Satisfaction".

expansion of the user's vocabulary, and a decreased use of unnecessary words such as fillers and repeated words. The users are postulated to learn how to interact with SDS.

4.4 Evaluation of Performance of SDS and User

We now discuss the shift in users' evaluation according to their learning experience. As shown in Table 1, the user answered several questions after using the system. Everyday questions are concerned with aspects of *speech input quality* and *dialogue cooperativity*, while the first and final-day questions are concerned with aspects of *communication efficiency* and *pleasantness* (Möller and Skowronek, 2003).

The following criteria improved over the five day period "Input Quality", "Informativeness", "Task Success", "User Satisfaction" and "Ease of Use", "User Dialogue" and "Pleasantness". In these criteria, all users obtained better scores on the final day as compared to the first. Examples of these results are show in Figure 8 . The improvements in "Input Quality", "Informativeness" and "Task Success" indicate that the user realizes the increase of WCR and CCR. The improvements in "User Satisfaction" and "Pleasantness" meanwhile, indicate that the user has begun to evaluate the system impression better according to the learning experience. Improvements in "Ease of Use" and "User Dialogue" indicate that the user correctly perceives the improvement in user performance. It is intriguing that the user is able to evaluate the learning effect.

Questions involving "Dialogue Conciseness", "Dialogue Smoothness" and "Partner Asymmetry" did not necessarily show improvement. The detailed results of the twelve users are shown in Table 2. "Good" indicates that the evaluation changed better on the final day as compared to the first.

Table 2: Results of evaluation in five days.

Criterion	Good	Bad	Same
Dialogue Conciseness	6	4	2
Dialogue Smoothness	7	1	4
Partner Asymmetry	7	3	2

These criteria are involved in aspects of *communication efficiency*. This aspect is supposed to decline in the evaluation during the user learning process.

Analysis of the questionnaire suggests that the user changes the evaluation of SDS according to the learning experience.

5 Path to a new evaluation of SDS

The experimental results illustrated how the usage and evaluation of SDS differed according to experience of the system: the user learned skills related to the use of SDS, especially, the phenotype in interaction. Subsequently, the user evaluated SDS better than pre-learning and learning effects appropriately. The leaning effect of SDS is known a priori, however, our results displayed a new form of the learning effect. We now propose that it may be taken into consideration to evaluate SDS users as well as the system.

Experimental results revealed that the system performance, WCR, CCR, improved and was evaluated better in terms of learning. This suggested the user's evaluation of SDS is a factor of the learning experience. With long-term usage in mind, user skills must be accessed and we should draw attention to the difference in the evaluation according to the user skills in the field test of usability.

With regard to the user, the phenotype of dialogue changed and the learning effect was evaluated precisely. Learning sessions may need to be implemented in SDS, during which the learning level of users can be evaluated by the users themselves. The learning of SDS may extend the potential for extensive usage, although further examination is needed to apply the learning effect of SDS to the system.

Acknowledgement

This research was supported in part by the MEXT e-Society leading project.

References

Asahi Newspaper: http://www.asahi.com/

Kawahara T. Lee A. Takeda K. Itou K. Shikano K. 2004. Recent Progress of Open-source LVCSR Engine Julius and Japanese Model Repository; Software of Continuous Speech Recognition Consortium. *Proceedings of ICSLP 2004*, 2:3069-3072.

Möller S. Skowronek J. 2003. Quantifying the Impact of System Characteristics on Perceived Quality Dimensions of a Spoken Dialogue Service. *Proceedings of EUROSPEECH 2003*, 1953-1956.

Mora music download site: http://mora.jp/

Oricon Style: http://www.oricon.co.jp/

Publicity Available Language Modeling Toolkit (Palmkit): http://palmkit.souceforge.net/