Analyzing Temporal Transition of Real User's Behaviors in a Spoken Dialogue System

Kazunori Komatani, Tatsuya Kawahara, Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University, Japan

komatani@i.kyoto-u.ac.jp

Abstract

Managing various behaviors of real users is indispensable for spoken dialogue systems to operate adequately in real environments. We have analyzed various users' behaviors using data collected over 34 months from the Kyoto City Bus Information System. We focused on "barge-in" and added barge-in rates to our analysis. Temporal transitions of users' behaviors, such as automatic speech recognition (ASR) accuracy, task success rates and barge-in rates, were initially investigated. We then examined the relationship between ASR accuracy and barge-in rates. Analysis revealed that the ASR accuracy of utterances inputted with barge-ins was lower because many novices, who were not accustomed to the timing when to utter, used the system. We also observed that the ASR accuracy of utterances with barge-ins differed based on the barge-in rates of individual users. The results indicate that the barge-in rate can be used as a novel user profile for detecting ASR errors.

Index Terms: spoken dialogue system, real user behavior, barge-in

1. Introduction

User's behaviors are one of significant elements that should be considered when designing a spoken dialogue system and improving its performance. Adaptation to various users [1] is indispensable for dialogue management modules when developing a system that can be used by real users. System performance will improve by predicting user's behaviors and by adapting speech recognition and dialogue management modules for each user. For example, if the system can predict that a user's automatic speech recognition (ASR) accuracy is low, dialogue management can be adaptively changed to system-initiated one, and a language model of ASR can also be narrowed down after providing adaptive help messages that guides user's utterances [2]. It is indispensable to know user's behaviors in real situations, to make the systems more robust for various users.

We constructed the Kyoto City Bus Information System, which is being available to the public now. Raux et al. developed a similar system at Pittsburgh, and reported its performance and issues [3]. In this paper, we analyzed various users' behaviors for individuals, which were based on callers' phone numbers, using real data collected on the system over 34 months. We focused on "barge-in", which is a specific feature of spoken dialogue systems, and added barge-in rate to our analysis. A barge-in is defined as the situation when a user starts speaking during a system prompt. When this occurs, the system stops its current prompt and starts recognizing the user's utterance. This barge-in function makes dialogues more efficient because users can interrupt lengthy system prompts when this function is installed into systems. We report on how the function works for general users, including novices, in real situations.

Sys:	What is your current bus stop, your destination, or specific bus route number?
User:	Shijo-Kawaramachi.
Sys:	Will you take a bus from Shijo-Kawaramachi?
User:	Yes.
Sys:	Where will you get off the bus?
User:	Arashiyama.
Sys:	Will you go from Shijo-Kawaramachi to Arashiyama?
User:	Yes.
Sys:	Bus number 11 bound for Arashiyama has departed Sanjo-Keihanmae, which is two bus stops away.

Figure 1: Example dialogue from bus system

As to diversity of user's behaviors, not only differences between individuals but also temporal transitions within one individual should be taken into consideration. In other words, users are expected to change their behaviors, such as how often they barge-in, until they get accustomed to the system[1]. First, we describe how user's behaviors change as they get used to the system. Then, we investigated the difference in barge-in rate among users and examined the relationship between the ASR accuracy and barge-in rates. The results show that the barge-in rate differed between individuals and suggest the rate can be used to predict a user's behaviors. We show that the barge-in rate is useful for detecting ASR errors in a user's utterances.

2. Target Data for Analysis

2.1. System Overview

We have developed the Kyoto City Bus Information System [1]. The system locates the bus the user wants to catch and tells them how long it will be before the bus arrives. The system can be accessed using a telephone, including cellular phones[2]. Users are required to input their boarding stop, the destination, or the bus route number by voice, and, as a result, obtain the appropriate bus information. The bus stops can be specified by using the names of famous landmarks or public facilities nearby. There is only one type of query: a request for information about specific buses. The system's ASR is grammar-based, and its vocabulary contains 652 bus stops and 756 famous landmarks and public facilities nearby.

The dialogue management is executed in a mixed-initiated manner. That is, when only one slot is filled by a user utterance

[1] We did not take a forgetting model [4] into consideration because of the simplicity of the system, which has only three slots.

[2] +81-75-326-3116

as shown in Figure 1, the system first confirms its content, and then the system ask a question to request information that has not been given. Users can also specify the required information in a single utterance. They can interrupt a system prompt while it is being generated, and this feature is called a "barge-in". If they already know the contents of the prompt, they can barge in.

2.2. Data Collection and Annotation

We analyzed data collected on the Kyoto City Bus Information System from May 2002 to February 2005. The data included 7,988 valid calls. The system logs the caller's phone numbers, whether all system prompts were presented, and the durations of each prompt, the times when calls are made, the ASR results for each utterance, and so on. If not all the system prompts are presented, we assumed that a barge-in had occurred. Caller's phone numbers, which are not recorded if the callers have dialed special numbers before the system's telephone number, were recorded for 5,927 of the 7,988 calls. We analyzed behaviors of individual users based on this data.

We manually assigned labels to each call and utterance. Each utterance was transcribed, and whether its ASR result was correct or not was given. We assumed that an ASR result was correct if the correct content words were contained in the transcription. The success of each task was determined manually. Based on the assessment of annotators, the task was considered successful when the required bus information matched the information outputted by the system.

3. Analyzing Temporal Transitions of User's Behavior

We analyzed temporal transitions of user's behaviors based on the following three measures:

- ASR accuracy
- task success rate
- barge-in rate

The barge-in rate was defined as the ratio of the number of calls when a user barges-in on system prompts and the number of total calls performed by the user. As a temporal axis, we calculated the ratios using the number of calls to a certain point and the number of total calls, and plotted them on the x-axis. Therefore, $0 < x \leq 1$. Averages of each measure , such as ASR accuracy, task success rate, and barge-in rate to a certain time x, were plotted on the y-axis.

We then approximated the plotted values by using the following function:

$$f(x) = c - a \cdot \exp(-bx)$$

Values a, b, c roughly correspond to values where the target measure had converged when the user got sufficiently accustomed to using the system, speeds of the convergence, and the amount of transitions during this interval $(0 < x \leq 1)$. These values were calculated by using the least-square method. We assumed $a \geq 0$. To describe rough shapes of the functions, we calculated x when the amount of changes of $f(x)$ became saturated. In this paper, we defined x_I as $\{x|\frac{df(x)}{dx} = 0.1\}$. This means that the change of $f(x)$ converges near x_I. We denote an average of the target measures in this interval as $f(1)$. We defined Δ as $f(1) - f(0)$, which represents the amount of changes in the measure for the target user in this interval[3]. Finally, x_I is not defined when the Δ is zero because there is no change on $f(x)$.

[3]We assumed $f(0) = 0$ if $f(0) < 0$ as a result of approximations of a function.

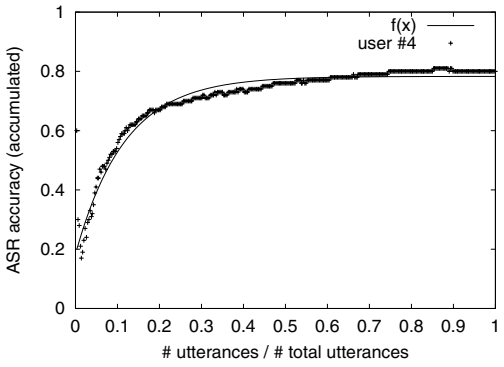

Figure 2: Temporal transition of ASR accuracy for user #4

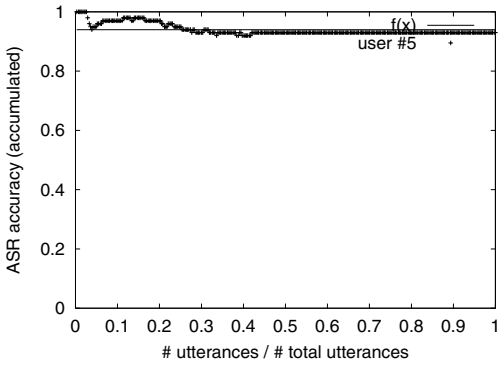

Figure 3: Temporal transition of ASR accuracy for user #5

3.1. Temporal Transition of ASR Accuracy and Task Success Rates

We analyzed temporal transitions of ASR accuracy and task success rate for each user. The ASR accuracy of users #4 and #5 is shown in Figure 2 and Figure 3. The accuracy of user #4 increased gradually and converged near on x of 0.6, whereas the accuracy of user #5 was kept high from the beginning and changed little when the x became larger.

Table 1 lists the temporal transitions of 12 users, who frequently used the system over 50 times. Mean square errors (MSEs) for the approximation results are also listed here in the exponential notation. The first and second columns in this table show the transition of the ASR accuracy and task success rate. The result indicates that the averages of the ASR accuracy and task success rate for the whole data, which are represented as $f(1)$, were generally high and that their variances were small. We can also find a correlation between the transitions of ASR accuracy and task success rate, which was also reported in [3]. However, the Δ values of some users, such as #4, were large. This indicates that ASR accuracy and task success rate gradually improved according as they became used to using the system. In other words, the results quantitatively showed that there were two types of users: ones who knew how to use the system from the beginning, and the others gradually became accustomed to the system while using it.

Table 1: Summary of temporal transitions for frequent users

User	ASR accuracy				Task success rate				Barge-in rate			
ID	$f(1)$	Δ	x_I	MSE	$f(1)$	Δ	x_I	MSE	$f(1)$	Δ	x_I	MSE
#1	.88	.20	.25	7.4E-5	.95	.28	.21	1.6E-4	.11	0	-	2.3E-4
#2	.89	.24	.47	2.5E-4	.94	.19	.25	2.8E-4	.19	0	-	1.9E-3
#3	.89	.03	< 0	6.8E-5	.96	.06	< 0	2.1E-4	.60	.60	> 1	6.4E-4
#4	.78	.60	.46	4.5E-4	.89	.89	.52	4.5E-4	.17	0	-	7.2E-4
#5	.94	0	-	3.3E-4	.98	0	-	1.3E-4	.74	.74	.58	4.6E-4
#6	.89	0	-	5.2E-4	.92	.40	.11	7.9E-4	.10	.06	< 0	1.1E-4
#7	.94	0	-	3.2E-4	.93	.09	.08	1.3E-3	.04	.04	.06	1.6E-4
#8	.89	0	-	1.7E-3	.87	.77	.37	1.0E-3	.71	0	-	1.0E-3
#9	.81	.27	.10	4.3E-4	.93	0	-	2.5E-3	.49	.47	.62	4.6E-4
#10	.90	0	-	1.1E-3	1	0	-	1.6E-4	.10	.10	.29	1.3E-4
#11	.72	.20	.17	1.5E-3	.79	.30	.19	2.2E-3	.15	.04	.13	9.8E-4
#12	.79	.37	.21	6.8E-4	.80	0	-	4.7E-3	.23	0	-	2.6E-3

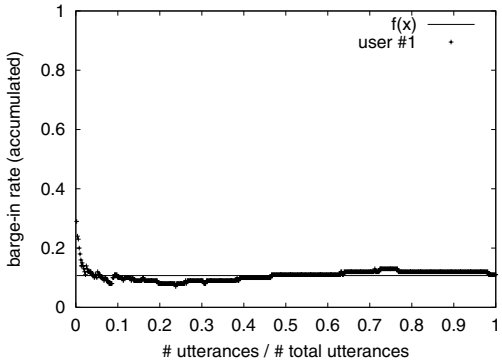

Figure 4: Temporal transition of barge-in rate for user #1

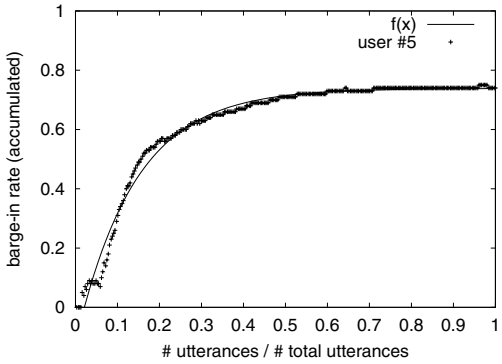

Figure 5: Temporal transition of barge-in rate for user #5

3.2. Temporal Transition of Barge-In Rates

Temporal transitions of the barge-in rates for users #1 and #5 are shown in Figures 4 and 5. The barge-in rate of user #1 was nearly static, whereas the barge-in rate of user #5 increased according as they became used to the system. As highlighted these by examples, variations in barge-in rates depended on individual users.

Variances in average barge-in rates ($f(1)$) were rather large, as shown in the third column in Table 1. These results show the diversity of user's behaviors while the system was being used. The result in the table also shows that the barge-in rates of some users, such as users #3, #5, and #9, increased steeply, whereas the rate of the other users did not change so much. This shows that the degree of behavioral transitions also differed among individuals.

3.3. Relationship of Behavioral Transitions among ASR accuracy, Task Success Rate and Barge-In Rate

The results in Table 1 indicate that the Δ in ASR accuracy and task success rate were small for users whose barge-in rates increased steeply, such as users #3, #5, and #9. This suggests that only users whose ASR accuracy was high enough may become accustomed to using the barge-in function.

On the other hand, barge-in rates were rather low for users whose Δ in ASR accuracy were rather large, such as users #1, #2, #4, #9, #11, and #12. This suggests that users who experienced many ASR errors when they started using the system tended to listen to every prompt of the system to its end.

Consequently, these results suggest that two phases exist while users become accustomed to the system: one phase is where ASR accuracy and task success rate improve, and the other phase is where users change their behaviors on how to complete tasks, such as barge-in. Considering these phases when designing dialogue management and help messages [2] would make them more useful: For example, system-initiated questions and instructing definite acceptable utterance patterns in the former phase; and suggestions about how to complete tasks, such as "You can barge in a system prompt while it is being generated" in the latter phase.

4. Predicting ASR Errors by using Barge-In Rate

In this section, we describe what a barge-in rate is useful for.

4.1. Relationship between Frequency of Barge-In and ASR accuracy

We analyzed the relationship between barge-in and ASR accuracy. Table 2 lists the ASR accuracy for all users per utterance, when the system prompts were played to their end (denoted as COMPLETE) and when the system prompts were barged in (BARGE_IN). The barge-in utterances amounted to 26.8% (7,940/29,580) of all utterances; however, half of those utterances contained ASR errors in their content words.

This result implies that many incorrect barge-in occurred despite the user's intention. Specifically, these included cases when background noises were incorrectly recognized as a

Table 2: ASR accuracy per barge-in

ASR results	Correct	Incorrect	Total	Accuracy
COMPLETE	17,921	3,719	21,640	(82.8%)
BARGE_IN	3,937	4,003	7,940	(49.6%)
Total	21,858	7,722	29,580	(73.9%)

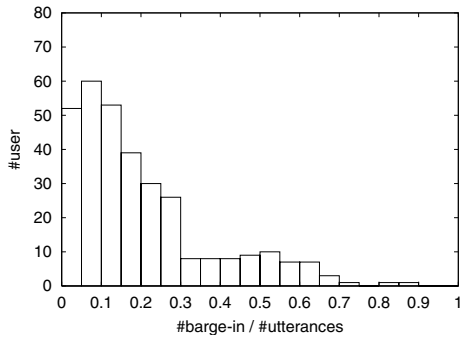

Figure 6: Number of users per barge-in rate

barge-in and the system's prompt stopped. This may often occur when the system is accessed using mobile phones in crowded places. Breathing and whispering were also prone to be incorrectly recognized as barge-ins. Moreover, disfluency in one utterance may be incorrectly divided into two portions, which causes further misrecognition and unexpected system's actions. The abovementioned phenomena, except background noises, are caused by a user's unfamiliarity with the system. That is, some novice users are not unaware of the timing when to utter and cause the system to misrecognize the utterance.

On the other hand, users who have already become accustomed to using the system often use the barge-in functions intentionally and, accordingly, make their dialogues more efficient. Average barge-in rates for the whole data were calculated for 323 users who used the system more than twice. The distribution is shown in Figure 6. The results in this figure show that the barge-in rates differed among users, which suggests that the barge-in rate may be used to profile user.

4.2. Predicting ASR Errors based on Barge-In Rates

We describe methods to detect ASR errors of utterances with barge-in based on each user's barge-in rate in this section. The results in Table 3 show the relationship between the barge-in rates per user and the corresponding ASR accuracies of utterances with barge-in. Since few users had barge-in rates greater than 0.8, which means almost all utterances were barge-ins, these users were removed from our analysis because almost all utterances were misrecognitions caused by background noises. Therefore, we focused on users whose barge-in rates were less than 0.8. For users whose barge-in rates were high, that is, they frequently barged-in, ASR accuracy with barge-in was high. This suggests that the barge-ins were intentionally performed. On the other hand, for users whose barge-in rates were less than 0.2, the ASR accuracies of their utterances with barge-ins was less than 20%. This suggests that the barge-ins of these users were unintentional.

Based on these results, for example, a strategy will be helpful in which the system does not accept utterances with a barge-in if the user's barge-in rate, which is accumulated per caller's telephone number, is lower than a threshold. This strategy will

Table 3: ASR accuracy of utterances with barge-in per each user's barge-in rates

Barge-in rate	Correct	Incorrect	ASR Acc. (%)
0.0 - 0.2	407	1,750	18.9
0.2 - 0.4	861	933	48.0
0.4 - 0.6	1,602	880	64.5
0.6 - 0.8	1,065	388	73.3
0.8 - 1.0	2	36	5.3
1.0	0	16	0.0
Total	3,937	4,003	49.6

enable utterances to be rejected that are likely to be ASR errors. Another strategy is where the system does not permit a barge-in for users whose barge-in rates are low. As indicated by these examples, we can use the barge-in rate as a new way to profile a user's characteristics.

5. Conclusion

We analyzed real user's behaviors using data collected from the Kyoto City Bus Information System. First, we analyzed temporal transitions of each user using three measures: ASR accuracy, task success rate, and barge-in rate. Consequently, a model was suggested to understand users' behaviors when they become accustomed to using a system. We also analyzed the relationship between the barge-in rates and ASR accuracy, and showed that the barge-in rate is helpful to predict ASR errors.

In our current analysis, ASR accuracy was calculated based on labels that were manually given. However, it is possible to estimate the accuracy after each dialogue has finished, by using the contents of user's responses for system's explicit confirmations. This estimation enables the adaptive dialogue management at runtime, based on our model explaining transitions while a user becomes accustomed to using the system.

Our future work includes developing a method in which barge-in rates are combined with other features and used for managing dialogue and detecting ASR errors. We will also verify the general trends reported in this paper after analyzing with more data collected using our system, which is now operating. Analysis of this data will enable us to develop a model to understand user behaviors and, ultimately, lead to the development of user-adapted dialogue strategies for use in spoken dialogue systems.

6. References

[1] K. Komatani, S. Ueno, T. Kawahara, and H. G. Okuno, "User modeling in spoken dialogue systems to generate flexible guidance," *User Modeling and User-Adapted Interaction*, vol. 15, no. 1, pp. 169–183, 2005.

[2] Y. Fukubayashi, K. Komatani, T. Ogata, and H. G. Okuno, "Dynamic help generation by estimating user's mental model in spoken dial ogue systems," in *Proc. INTERSPEECH*, 2006, pp. 1946–1949.

[3] A. Raux, D. Bohus, B. Langner, A.W. Black, and M. Eskenazi, "Doing research on a deployed spoken dialogue system: One year of let's go! experience," in *Proc. INTERSPEECH*, 2006, pp. 65–68.

[4] A. Hof, E. Hagen, and A. Huber, "Adaptive help for speech dialogue systems based on learning and forgetting of speech commands," in *Proc. of 7th SIGdial Workshop on Discourse and Dialogue*, 2006, pp. 1–8.