Predicting ASR Errors by Exploiting Barge-In Rate of Individual Users for Spoken Dialogue Systems

Kazunori Komatani, Tatsuya Kawahara, Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University, Japan
komatani@kuis.kyoto-u.ac.jp

Abstract

We exploit the barge-in rate of individual users to predict automatic speech recognition (ASR) errors. A barge-in is a situation in which a user starts speaking during a system prompt, and it can be detected even when ASR results are not reliable. Such features not using ASR results can be a clue for managing a situation in which user utterances cannot be successfully recognized. Since individual users in our system can be identified by their phone numbers, we accumulate how often each user barges in and use this rate as a user profile for determining whether a current "barge-in" utterance should be accepted or not. We furthermore set a window that reflects the temporal transition of the user's behavior as they get accustomed to the system. Experimental results show that setting the window improves the prediction accuracy of whether the utterance should be accepted or not. The experiments also clarify the minimum window width for improving accuracy.

Index Terms: spoken dialogue system, user modeling, barge-in

1. Introduction

User behaviors are important factors affecting the performance of spoken dialogue systems. User satisfaction and task success rate can be improved by generating adaptive responses [1]. We focused on the "*barge-in rate*" as a new user profile. A barge-in is a situation in which a user starts speaking during a system prompt. When this occurs, the system stops its current prompt and starts recognizing the user's utterance. The barge-in rate seems to correspond to how well a user is accustomed to the system [1] and how he or she regards the system as a social being. These are promising features for adapting dialogue management to individual users. Furthermore, the barge-in rate can be obtained relatively easily because it is defined not from automatic speech recognition (ASR) results but from only the timing information of the utterances.

We use the barge-in rate for predicting whether a "barge-in utterance" (an utterance at which a user barges in during the system prompt) is correctly recognized or not. Such utterances often cause ASR errors [2]. ASR errors often occur in fragments of utterances especially when novices use the system [3]. An example case is one in which users are not accustomed to the timing at which to speak and interrupt their utterances when they notice the system prompt continues. These ASR errors cannot be correctly rejected by a classifier distinguishing speech from noises on the basis of the Gaussian mixture model (GMM) [4], because the fragments are parts of human speech. We detect ASR errors by exploiting the barge-in rate as a new profile at the dialogue level. There were some studies in which dialogue-level features were used for detecting ASR errors [5, 6, 7]. We exploit new dialogue features that can be robustly obtained when the system is used repeatedly.

We calculate barge-in rates by setting a window. That is, barge-in rates are calculated by using only N utterances before the current utterance. This is because we cannot ignore the temporal transition of the user's behavior as he or she gains experience on the system. The temporal transitions of user behaviors in relation to ASR accuracy, task success rate, and barge-in rate was shown by using dialogue data obtained from a publicly deployed spoken dialogue system [2]. The window reflects temporal transitions in user behaviors by discarding their old histories before they were accustomed to the system, and it accordingly improves the prediction accuracy of ASR errors. We also clarify the window width required as a history for obtaining reliable barge-in rates and empirically show the minimum width for improving prediction accuracy.

2. Analyses of behaviors of real users in deployed spoken dialogue system

2.1. System overview

We have developed the Kyoto City Bus Information System [1]. The system locates a bus the user wants to catch and tells him/her how long it will be before the bus arrives. The system can be accessed by telephone, including cellular phone. Users are required to input their boarding stop, destination, or bus route number by voice, and, as a result, obtain 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 operated on the Voice Web Server, a product of Nuance Communications, Inc.[1], by dynamically generating VoiceXML scripts [8]. 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, the system first confirms its content and then requests 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 content of the prompt, they can barge in.

2.2. Target data for analysis

We analyzed data collected on the Kyoto City Bus Information System for 34 months (from May 2002 to February 2005) [2]. The data contained 7,988 valid calls from 671 users. The system logs the caller's phone numbers, whether all system prompts

[1] http://www.nuance.com/

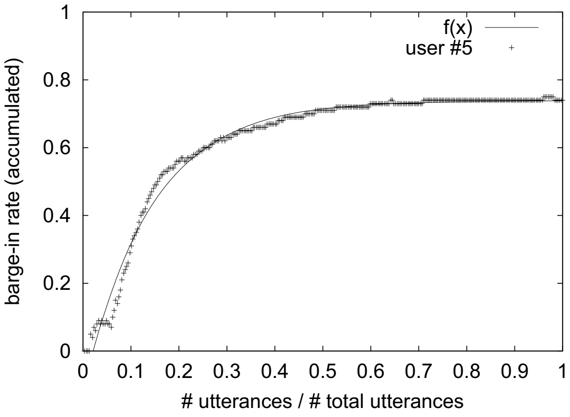

Figure 1: Temporal transition of a user's barge-in rate

Table 1: 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%)

were presented, 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 were presented, we assumed that a barge-in had occurred. Caller's phone numbers, which are not recorded if the callers dialed special numbers before the system's telephone number, were recorded for 5,927 of the 7,988 calls. We analyzed the behaviors of individual users on the basis of this data. Because they were real users, whose behaviors were different from recruited users [9], the number of calls per user as well as their behaviors was various. Several users repeatedly used the system.

We manually assigned labels to each call and utterance. We transcribed each utterance and indicated whether its ASR result was correct or not. An ASR result was assumed to be correct if the correct content words were contained in the transcription.

2.3. Temporal transition of users' behaviors

We previously analyzed temporal transitions of user behaviors for the following three measures [2]: ASR accuracy, task success rate, and barge-in rate. Here, we only mention the 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 to the number of total calls performed by the user.

A temporal transition of the barge-in rate for a certain user is shown in Figure 1. This user called the system 273 times, and the total number of utterances was 1,010. As the temporal axis x, we calculated the ratios using the number of calls to a certain point and the number of total calls. Therefore, $0 < x \leq 1$. Average barge-in rates at a certain time x are plotted on the y-axis. The line $f(x)$ shows the result of the approximation for the plotted values by using the following function: $f(x) = c - a \cdot \exp(-bx)$. This figure indicates that users' behaviors change as they get accustomed to the system. This result therefore suggests a user model considering temporal transitions caused by user experience as well as representing differences between users.

Table 2: ASR accuracy of utterances with barge-ins per average barge-in rate of each user

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

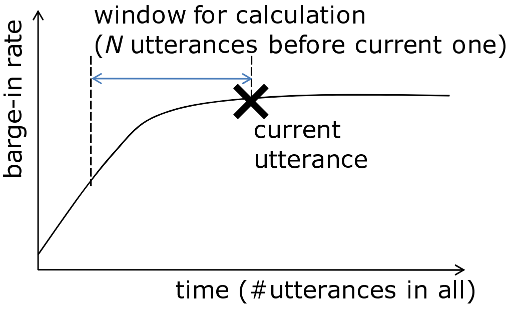

Figure 2: Calculating average and variance of barge-in rates within window

3. Predicting ASR errors by using barge-in rate

Barge-in utterances are prone to contain more ASR errors than those without barge-ins. Table 1 lists the ASR accuracy for the cases when the system prompts were played to their end (denoted as COMPLETE) and when the system prompts were barged in upon (BARGE_IN). This table shows that the barge-in utterances amounted to 26.8% (7,940/29,580) of all utterances; however, half of those utterances contained ASR errors. These were caused by background noise or the user's unfamiliarity with the system.

We took into consideration how often individual users barge in the system prompts for detecting ASR errors of barge-in utterances. That is, we calculated the average barge-in rates per user for their all utterances. The results in Table 2 show the relationship between the barge-in rate per user and the corresponding ASR accuracies of utterances with barge-ins. For users whose barge-in rates were high, that is, they frequently barged-in, the ASR accuracy of barged-in utterances was high[2]. 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 barge-in utterances were less than 20%. This suggests that the barge-ins of these users were unintentional and they caused ASR errors.

3.1. Setting window for calculating barge-in rate

In order to take the temporal transition of a user's behavior into consideration, we set a window for calculating barge-in rates at each point of the dialogue. That is, barge-in rates are calculated by using N utterances before the current one, as shown in Figure 2. We call this N the window width.

[2] We ignored users whose barge-in rates were higher than 0.8 because they were few and thus can be regarded as noises.

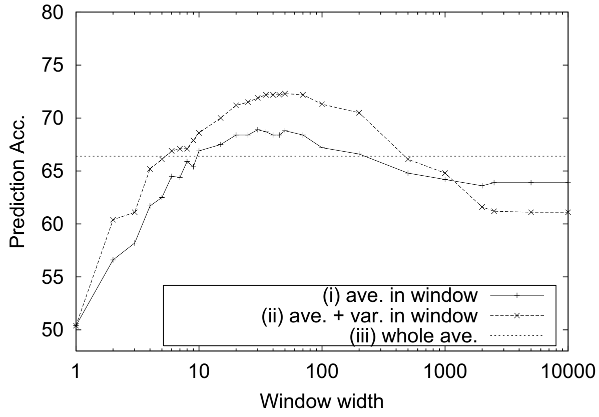

Figure 3: Prediction accuracies of ASR errors when window width varies for all barge-in utterances

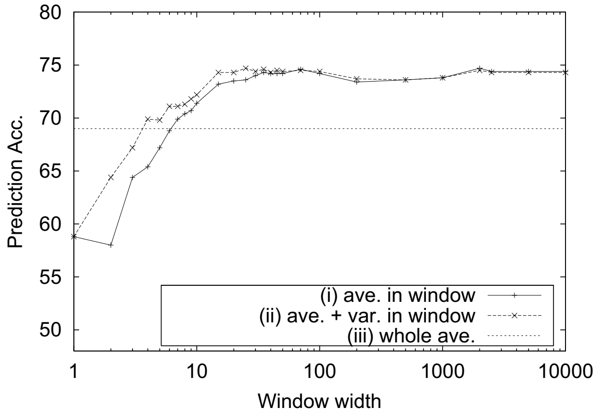

Figure 4: Prediction accuracies of ASR errors when window width varies for barge-in utterances after excluding users who used the system for less than 10 times

We also calculate the variances of the barge-in rates within this window width, as well as their averages. When the variance is small, the barge-in rates have converged and do not change much. This means the average barge-in rate can be used as a reliable profile. On the other hand, a large variance means that the user's behaviors are not stable and the average barge-in rate is not reliable as a profile.

3.2. Experimental verification

The experiments sought to answer the following questions:

1. Does the prediction accuracy of ASR errors improve by using a window to calculate the barge-in rate?
2. How wide a window is appropriate as a history for estimating reliable barge-in rates?

Accuracies given various window widths were compared with a baseline, where barge-in rates were calculated by using all utterances. When too wide a window is used, the barge-in rates become equivalent to the baseline and the average barge-in rate cannot reflect the temporal transitions in the user's behavior. When the window is too narrow, the average barge-in rate is not reliable as a user profile.

We assessed the prediction results for each situation as to whether or not the user profile gave correct ASR results for barge-in utterances. The prediction accuracy was calculated by comparing the prediction results and manually annotated reference labels.

To predict ASR errors, we constructed a logistic regression model that relates barge-in rates to the probability of ASR correctness. Denoting the probability that an ASR result of a barge-in utterance is correct as P_{ASR}, the regression function can be written as:

$$P_{ASR} = \frac{1}{1 + \exp(-(a_1 x_1 + a_2 x_2 + b))}.$$

The independent variables are the average and variance of barge-in rates within the window width, x_1 and x_2, respectively. The dependent variable is a binary value indicating whether the ASR is correct or not. The coefficients a_1, a_2, b are obtained after fitting by using training data. The fitting and prediction processes were performed with a 10-fold cross validation.

We set three conditions on calculating the average prediction accuracies.

(i) ave. in window: using only the average barge-in rate (x_1) within each window width as the input to the regression function.

(ii) ave. + var. in window: using both average and variance of barge-in rate (x_1, x_2) within each window width as the input to the regression function.

(iii) whole ave.: using average barge-in rate for whole barge-in utterances as the input at every point in the dialogue. This condition does not take into consideration temporal transition of user behaviors, and it corresponds to the performance shown in Table 2. We chose this as the baseline method.

First, we calculated the prediction accuracy by using all 7,940 barge-in utterances as training and evaluation data. Figure 3 shows the prediction accuracies when the window width varies. We then limited target users to those who called the system over 10 times[3]. The number of such users was 74. This is to verify accuracies after excluding utterances for which sufficient data was not available as a history. Each call contained approximately 2-5 utterances. Their utterances amounted to 6,216, and prediction accuracies in these cases are shown in Figure 4.

Accuracies and window widths for the best performance in Figures 3 and 4 are listed in Table 3. "Maj." in this table means the majority baseline; that is, when all utterances were classified to either binary value. The parameters for the cases listed in the table are as follows:

- All utterances / Cond. (i) / window = 30
  $a_1 = 3.08$, $b = -1.60$
- All utterances / Cond. (ii) / window = 50
  $a_1 = 3.05$, $a_2 = -7.54$, $b = -1.04$
- Limited to over-10-call users / Cond. (i) / window = 70
  $a_1 = 4.16$, $b = -1.65$
- Limited to over-10-call users / Cond. (ii) / window = 25
  $a_1 = 4.13$, $a_2 = -3.66$, $b = -1.50$

[3] We tried several values as the number of minimum calls, but no steep change was observed. Therefore, 10 times is used as a representative value.

Table 3: Best prediction accuracy and corresponding window width

Conds.	(i)	(ii)	(iii)	Maj.
For all utterances	68.9% (w=30)	72.3% (w=50)	66.4% (-)	50.4% (-)
Limited to over-10-calls users	74.6% (w=70)	74.7% (w=25)	69.1% (-)	58.8% (-)

Figure in () is window width.

3.3. Discussion

Necessity of the temporal transition model: In both Figures 3 and 4, the prediction accuracies for appropriately set windows (around 30-80) were better for Conds. (i) and (ii) than for Cond. (iii) (i.e., when the average of whole utterances were used). The window discarded the users' old histories and thus reflected temporal transitions of their behaviors. The barge-in rates were not constant but varied as the users got accustomed to the system, as shown in Figure 1. Therefore, consideration of temporal transitions resulted in the better performance of Conds. (i) and (ii).

Required window width: As we can see from Figures 3 and 4, prediction accuracies leveled off for window widths larger than around 30. This means that 30 utterances at least need to be used to calculate the average barge-in rate.

Effect of variance: The prediction accuracy when all barge-in utterances were used was lower than the result when the target users were limited to those who called the system over 10 times, as shown in Table 3. This was because many users were contained whose utterances were fewer than the window widths when all barge-in utterances were used. In this case, the number of utterances used to calculate the average was not sufficient to be regarded as reliable history information.

Figure 4 shows that using the variance together improved prediction accuracy when the window width was narrow. When the window width was wide enough, the variance caused no significant change in prediction accuracy. This result means that the variance can be used in combination with the average barge-in rate when the average barge-in rate is not reliable.

4. Conclusion and future work

We used the barge-in rate of individual users as a new user profile to predict ASR errors. This rate can be obtained relatively reliably because it is defined not with ASR results but by using only the timing information of the utterances. We calculated the average and variance of barge-in rates by setting a window of N utterances and used them as a user profile at each point in the dialogue. Setting the window width to around 30-80 improved the prediction accuracy of ASR errors relative to a baseline in which an average of whole barge-in utterances was used as the user profile. This result quantitatively showed that a temporal transition model in the form of a window is required.

The prediction of ASR errors described in this paper is based on the positive correlation between the average barge-in rate of individual users and their ASR accuracy, both of which vary as they get accustomed to the system. However, after observing the temporal behaviors of real users, we also found that ASR accuracy for a user tends to improve earlier than her barge-in rate does [2]. That is, there is a phase where the user's barge-in rate is still low but her ASR accuracy is high, as the user gets accustomed to the system, as shown in Figure 5. Therefore, if we can estimate a user's ASR accuracy during interactions and take the degree of its change into account, the preciseness of our model will be enhanced. We will investigate an online estimation of ASR accuracy by utilizing users' responses for the system's explicit confirmation [10, 11], and we will try to use the result to improve our model.

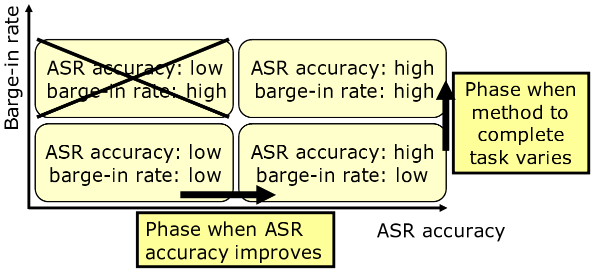

Figure 5: Two phases in getting accustomed to the system

5. 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] K. Komatani, T. Kawahara, and H. G. Okuno, "Analyzing temporal transition of real user's behaviors in a spoken dialogue system," in *Proc. INTERSPEECH*, 2007, pp. 142–145.

[3] A. Raux, D. Bohus, B. Langner, A. 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. Lee, K. Nakamura, R. Nisimura, H. Saruwatari, and K. Shikano, "Noice robust real world spoken dialogue system using gmm based rejection of unintended inputs," in *Proc. ICSLP*, 2004, pp. 173–176.

[5] D. J. Litman, M. A. Walker, and M. S. Kearns, "Automatic detection of poor speech recognition at the dialogue level," in *Proc. ACL*, 1999, pp. 309–316.

[6] M. Gabsdil and O. Lemon, "Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems," in *Proc. ACL*, 2004, pp. 343–350.

[7] D. Bohus and A. Rudnicky, "A "k hypotheses + other" belief updating model," in *Proc. AAAI Workshop on Statistical and Empirical Approaches to Spoken Dialogue Systems*, 2006.

[8] K. Komatani, F. Adachi, S. Ueno, T. Kawahara, and H. G. Okuno, "Flexible spoken dialogue system based on user models and dynamic generation of VoiceXML scripts," in *Proc. 4th SIGdial Workshop on Discourse and Dialogue*, 2003, pp. 87–96.

[9] H. Ai, A. Raux, D. Bohus, M. Eskenazi, and D. Litman, "Comparing spoken dialog corpora collected with recruited subjects versus real users," in *Proc. 8th SIGdial Workshop on Discourse and Dialogue*, 2007, pp. 124–131.

[10] K. Sudoh and M. Nanano, "Post-dialogue confidence scoring for unsupervised statistical language model training," *Speech Communication*, vol. 45, pp. 387–400, 2005.

[11] D. Bohus and A. Rudnicky, "Implicitly-supervised learning in spoken language interfaces: an application to the confidence annotation problem," in *Proc. 8th SIGdial Workshop on Discourse and Dialogue*, 2007, pp. 256–264.