Analyzing User Utterances in Barge-in-able Spoken Dialogue System for Improving Identification Accuracy

Kyoko Matsuyama, Kazunori Komatani, Ryu Takeda, Toru Takahashi, Tetsuya Ogata, Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University
Yoshida-Hommachi, Sakyo, Kyoto 606-8501, Japan.
{matuyama, komatani, rtakeda, tall, ogata, okuno}@kuis.kyoto-u.ac.jp

Abstract

In our barge-in-able spoken dialogue system, the user's behaviors such as barge-in timing and utterance expressions vary according to his/her characteristics and situations. The system adapts to the behaviors by modeling them. We analyzed 1584 utterances collected by our systems of quiz and news-listing tasks and showed that ratio of using referential expressions depends on individual users and average lengths of listed items. This tendency was incorporated as a prior probability into our method and improved the identification accuracy of the user's intended items.

Index Terms: barge-in, spoken dialogue systems, utterance timing, user characteristics

1. Introduction

Since barge-in-able spoken dialogue systems allow users to freely express their utterances anytime, it is expected to improve the quality of user interfaces. The user can interrupt the system's utterances to convey his/her intention. This interruption is called barge-in. Barge-in has attracted the attention of researchers concerned with spoken dialogue systems, specifically, the issue of barge-in detection [1, 2]. Their purpose has been to detect users' barge-in occurrences quickly and accurately. Ström [3] discussed a system's behavior when barge-ins were incorrectly detected. Since the user's utterance is mixed with the system's in the case of barge-in, it was difficult to exploit utterance timing under environments where no close-talk microphone is used.

We have developed a dialogue strategy that enables robust interaction under noisy environments where automatic speech recognition (ASR) results are not necessarily reliable [4]. Our barge-in-able spoken dialogue system can detect the user's barge-in utterance thanks to an ICA-based sound source separation method [5] that extracts the user's utterance from a mixture of the user's and system's utterances. This system exploits utterance timing together with ASR results to interpret user intention: that is, to identify the item that a user wants to indicate from items the system enumerates one by one. For example, the system and the user can interact as follows:

User Tell me which temple you suggest visiting.

System There are ten temples that I would suggest. "Kinkaku-ji Temple", "Ginkaku-ji Temple$\cdots$"

User That one.

System OK, you mean "Ginkaku-ji temple." It is the most famous one $\cdots$

In this case, the user interrupts the system while it reads out "Ginkaku-ji temple." This system identifies the user's referent, that is, what the user indicates by "That one." By using the barge-in timing of the user utterance, it determines that "Ginkaku-ji Temple" is specified by the user.

We investigate the user's behaviors such as barge-in timing and utterance expressions when he/she specifies his/her referent on the basis of data collected by two barge-in-able systems. More specifically, we analyze how often each user barges in and uses referential expressions. Some users prefer to barge in and to use referential expressions, e.g., "that one", while others prefer to specify items by words or phrases in enumerated items. In the former case, the system should make much use of the timing; in the latter case, it should give weight to interpretations based on the ASR result.

We furthermore show that the user's preference and the lengths of enumerated items are helpful to adapt the weight between the utterance timing and ASR results. Characteristics of users, enumerated items, and ASR results are used as features of the logistic regression to adapt the weight.

The rest of the paper is organized as follows: Section 2 explains the framework for identifying the user's referent. The analysis of users' behavior is explained in Section 3. Section 4 presents the estimation of the weight by using logistic regression and shows its experimental results. Section 5 concludes this paper with future works.

2. Use of barge-in timing and ASR results to identify user's referent

2.1. Maximum likelihood estimation of user's referent

This section describes a probabilistic framework in which we integrate utterance timing and ASR results, both of which are represented as probabilities [4]. This enables us to identify a user's referent as the item with the maximum likelihood. Here, we define utterance timing as the temporal difference between when a system utterance starts and when a user utterance starts (see Figure 1).

We formulate the problem of identifying a user's referent T_i that maximizes the probability $P(T_i|U)$. Here, T_i denotes the i-th item enumerated by the system, and U denotes a user utterance. We calculate the probability for each T_i and then determine the user's intention, T.

$$\begin{aligned} T &= \underset{T_i}{\operatorname{argmax}}\, P(T_i|U) = \underset{T_i}{\operatorname{argmax}}\, \frac{P(U|T_i)P(T_i)}{P(U)} \\ &= \underset{T_i}{\operatorname{argmax}}\, P(U|T_i). \end{aligned} \quad (1)$$

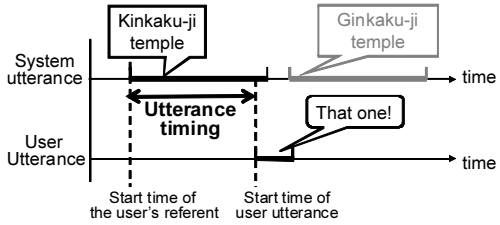

Figure 1: *Definition of utterance timing*

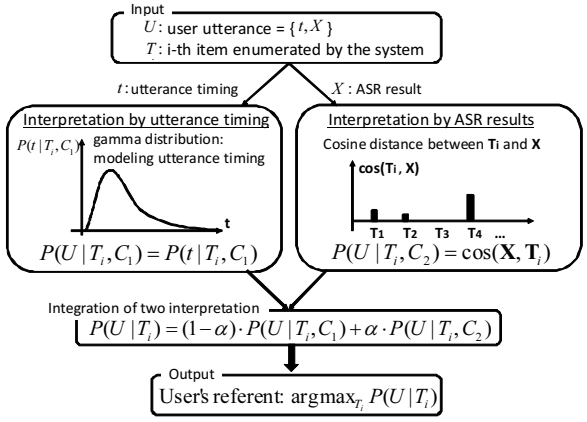

Figure 2: *Flow of identifying user's referent*

Here, we assume all the prior probabilities $P(T_i)$ are equal and $P(U)$ is not dependent on i.

We calculate $P(U|T_i)$ in accordance with Equation (1) by considering the possibilities of two cases. Here, Case C_1 is when the user conveys his/her intention by utterance timing, and Case C_2 is when he/she does by content of the utterance. Thus, $P(U|T_i)$ can be represented as the following sum:

$$P(U|T_i) = \sum_{k=1}^{2} P(U|T_i, C_k)P(C_k|T_i). \quad (2)$$

Here we set the coefficient α as the prior probabilities $P(C_k|T_i)$ as shown in Equation (3).

$$P(U|T_i) = (1-\alpha)P(U|T_i, C_1) + \alpha P(U|T_i, C_2). \quad (3)$$

The parameter α gives weight to interpretations based on either utterance timing or ASR results. $P(U|T_i, C_k)$ denotes the probability of an occurrence of user utterance U in the case of C_k for each item T_i. $P(U|T_i, C_1)$ is calculated by a gamma distribution assumed for utterance timing. $P(U|T_i, C_2)$ is calculated as a cosine distance between an ASR result and each item T_i in the vector space model. Figure 2 shows the flow of identifying a user's referent. The details how to calculate $P(U|T_i, C_k)$ are given in [4].

3. Tendency of user in uttering the referential expression

The parameter α should be adequately determined for correct interpretations of the user's utterances. Understanding when users convey their intention by utterance timing would help us determine how to set α properly. This section presents how often each subject uses referential expressions and reveals the correlation between the ratio of referential expressions and the average length of listed items.

Table 1: *Number of user utterances*

Tasks	Referential	Content	Total
News-listing (20 subjects)	263 (65.7%)	137 (34.3%)	400
Quiz (31 subjects)	434 (36.7%)	750 (63.3%)	1184

Setting α properly corresponds to changing the prior probability $P(C_k|T_i)$ adaptively to the following situations. When a user specifies his referent, he uses referential expressions such as "That one" or a pronoun, or content expressions such as "Kinkaku-ji Temple" or "The second." If a user utters a content expression, the user conveys his intention not by the timing but by the content. In this case, the ASR result is important to interpret his intention. On the other hand, a user's referential expression should be interpreted not by the content but by the timing. If a user frequently utters referential expressions, α in Equation (3) should be smaller to give weight to an interpretation based on utterance timing.

3.1. Data for analysis

We designed two different tasks to collect the user's barge-in utterances, and we analyze the difference of user expressions between two systems. In the first task, the system enumerates updated news titles that are automatically obtained from 10 RSS feeds [4]. User can specify the news title after he/she consider whether each news title seems interesting. In the other task, the system enumerates 8 choices for each of 40 quizzes and subjects choose one as an answer. User can specify the item immediately when it is enumerated by the system. The average length of news titles is 5.65 seconds and the titles often contain unknown words for users. Whereas the average length of items enumerated in quizzes is 1.59 seconds and the items consist of well-known words.

We collected 400 utterances in the former system from 20 subjects and 1184 utterances in the latter from 31 subjects. The number of referential or content expressions in two tasks is shown in Table 1. This table indicates the difference of the ratio of using referential expressions between the two tasks. Their utterances from a news-listing task consist of more referential expressions than content expressions and those from the quiz task have the opposite composition. It is because a subject often uses a referential expression when the enumerated item is long in seconds and when the item contains of unknown words. Another reason of this is that a subject tends to use a content expression when the enumerated item is short in seconds, because there is less time for him/her to say "That one".

3.2. Ratio of using referential expressions and hypothesis

How often each user used referential expressions is shown in Figures 3 and 5. The horizontal axis denotes the numbers of subjects and the vertical axis denotes the ratio of referential expressions. Eight subjects in Figure 3 and two subjects in Figure 5 used referential expressions more frequently than 80% of their utterances. Some subjects prefer the referential expressions because the barge-in timing is often more reliable than ASR results. In this case, the interpretation by the utterance timing is a good strategy to identify the user's referent. Their utterances are interpreted robustly with regard to their utterance timing rather than their ASR results. Determining α to give weight to inter-

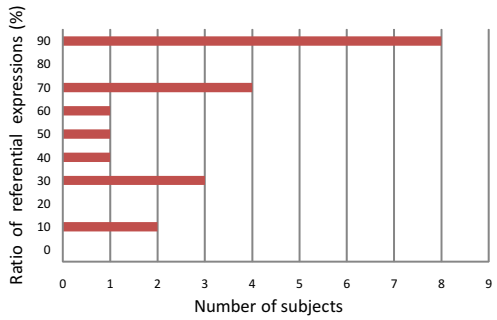

Figure 3: *Ratio of referential expression for subjects in news-listing task*

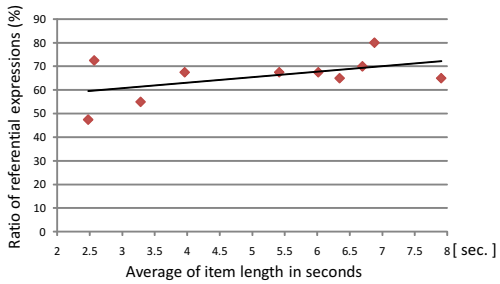

Figure 4: *Ratio of referential expression for item lengths in news-listing task*

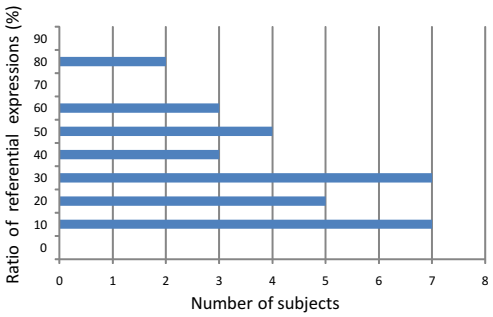

Figure 5: *Ratio of referential expression for subjects in quiz task*

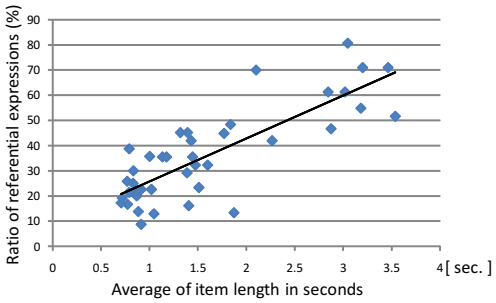

Figure 6: *Ratio of referential expression for item lengths in quiz task*

pretations obtained by utterance timing is important to decline identification error caused by only using ASR results.

Figures 4 and 6 show the ratio of the referential expressions increases as average item lengths become longer in seconds for each news-listing and quiz. The horizontal axis denotes the average lengths of items and the vertical axis denotes the ratio of referential expression. Each point corresponds to the one news-listing or quiz. The coefficient of correlation in Figure 4 is 0.51 and that in Figure 6 is 0.81. This result suggests that it is an effective strategy to rely on the utterance timing interpretation when a system lists up longer items. This result indicate that the system should give weight to an utterance timing interpretation according to the lengths of the items.

4. Automatic estimation of parameter α and experimental evaluation

We verify the relevance of changing α in accordance with each user's characteristics and situations in which the system enumerates. First, we show the oracle of the identification accuracy. We then describe the automatic estimation of parameter α using logistic regression. The evaluation of our method is presented subsequently.

4.1. Oracle of identification accuracy in changing α for each user utterance

We show that there is still room for improvement for identification accuracy by optimizing α. This oracle is calculated as follows:

1. Calculating $P(U|T_i)$ when α is changed from 0.0 to 1.0 in increments of 0.1.
2. If there exists an α that correctly identifies a user's referent, we regard the user's utterance as identifiable.

Table 2 shows the identification accuracy when α is 0.0, 0.1 and 1.0 and the oracle accuracy obtained by the steps above. The identification accuracy when α is 0.0 corresponds to using only the utterance timing, and that when α is 1.0 corresponds to using only the ASR results. The case $\alpha = 0.1$ was when the highest accuracy was obtained among all α. The accuracy in the oracle, in which α was determined by hand for each utterance, was higher by 4.9 points than the best one when α was fixed ($\alpha = 0.1$). This result indicates the effectiveness of changing α for each user's characteristics and item length to improve the identification accuracy.

4.2. Determining α on basis of features of user characteristics and enumerated items

We estimate α using logistic regression, which uses several features obtained from the user characteristics and the average length of items that the system enumerates, as shown in Equation (4).

$$\alpha = \frac{1}{1 + \exp(-(a_1 F_1 + a_2 F_2 + \cdots + a_6 F_6 + b))}. \quad (4)$$

The advantage of using logistic regression is it can calculate α automatically. Calculating α by simple rule is difficult because not all rules can be defined. For example, it is not always appropriate to set α to large value whenever a user utters content expressions because we found a case where the user's referent is estimated correctly using utterance timing. The coefficients $a_1, \cdots, a_6$ and b are fitted using training data. The independent variables $F_1, \cdots, F_6$ are the features shown in Table 3.

Table 2: *Identification accuracy [%] for user utterances*

α	Referential (#:434)	Content (#:750)	Total (#:1184)
0.0	98.4 (#:427)	84.3 (#:632)	89.4 (#:1059)
0.1	98.2 (#:426)	86.8 (#:651)	91.0 (#:1077)
1.0	3.23 (#:14)	58.9 (#:442)	38.3 (#:456)
Oracle	98.6 (#:428)	94.3 (#:707)	95.9 (#:1135)
Our method	98.2 (#:426)	88.9 (#:667)	92.3 (#:1093)

Table 3: *Features of user's and system's characteristics*

F_1:	ASR word correctness
F_2:	maximum of ASR confidence scores
F_3:	ratio of referential expression
F_4:	whether user barges in user's referent or not
F_5:	utterance timing [sec.]
F_6:	average length of listed items [sec.]

Features F_1 and F_2 represent the characteristics of the ASR results. F_1 corresponds to the user's ASR word correctness and it is calculated after the user's all utterances are collected. F_2 is the maximum ASR confidence score among words contained in the user utterance. Features F_3 to F_5 represent the characteristics about when and what the user speaks. The ratio of referential expression, which is updated everytime the user utters, is used as F_3. Whether the utterance is a referential expression or not is decided on the basis of transcriptions. F_4 is whether the user barges in the user's referent or not, and F_5 is the utterance timing defined in Section 2. Utterance timing is detected using the voice activity detection module of the ASR engine, Julius [6]. Feature F_6 represents the characteristics of items the system enumerates, and F_6 is the average lengths of listed items. Features F_1 and F_3 are given by hand in our experiments. Calculating these features online and automatically is one of our future works.

4.3. Target data for evaluation

We used 1184 barge-in utterances collected by our system in the quiz task. We used Julius as the speech recognizer and a 3000-state phonetic tied-mixture triphone model as the acoustic model. We made a statistical language model by using transcriptions of user utterances. The vocabulary size was 409. The ASR word accuracy for all utterances was 51.9%. Reasons for the low accuracy include reverberation of the room and distortions caused by the sound source separation since we used a microphone embedded in a robot instead of using a normal close-talk microphone. Moreover, it was because they often speak quickly or too softly.

4.4. Experimental results

We fitted the coefficients of the logistic regression with a 10-fold cross-validation. The identification accuracy when α was calculated by logistic regression for each utterance is also shown in Table 2. This accuracy was 92.3% for all utterances, which outperformed the accuracy when $\alpha = 0.1$. The differences between this method and $\alpha = 0.1$ for content expressions and total utterances were statistically significant ($p < 0.01$) by t-tests.

In fact, 16 utterances that could not be identified correctly when $\alpha = 0.1$ became identified correctly. For example, the utterance "the forth item" were interpreted that it specified the fifth item based on utterance timing. By applying α estimated from logistic regression, this utterance became identified correctly because α was automatically set larger; that is, the system gave more weight to interpretations based on the ASR results.

Next, we examine the coefficients $a_1, \cdots, a_6$ obtained from logistic regression. Coefficients a_3 and a_4 had positive values and the rest had negative one. These results were reasonable as explained below: When F_1 and F_2 are large, the ASR results of user utterances are reliable. In particular, when the user uttered content expressions, α should be large. Thus, coefficients a_1 and a_2 should be negative as obtained. On the other hand, if F_3 and F_4 are large, it is expected that α should be small because the utterances seem to be identified by an interpretation based on utterance timing Thus coefficients a_3 and a_4 should be positive, as also obtained. If F_5 is large, the user tends to use content expressions and α should be large. Thus a_5 should be negative, as also obtained. If F_6 is large, the user tends to use referential expressions and a_6 should be positive to set α large. However, a_6 had a negative value, which was against our expectation. It is because some content expressions can be interpreted by not only ASR results but also utterance timing and then α does not matter.

5. Conclusion

We have experimentally demonstrated that the identification accuracy improves by exploiting information about how a user barges in and what a system lists. First, we collected 1584 barge-in utterances by our systems of quiz and news-listing tasks and showed the ratio of using referential expressions depends on individual users and average lengths of listed items. We then incorporated this observation as a prior probability in identifying the user's referent by logistic regression.

Future works include the online estimation of features F_1 and F_3 that were given by hand in this experiments. Another future work is to enable spoken dialogue systems to accept various kinds of barge-in utterances. In a natural conversational interaction, users can make a variety of barge-in utterances; for example, to conclude the conversation quickly, to correct misunderstandings, or to assert themselves strongly - not only to indicate their referent.

6. References

[1] R. C. Rose and H. K. Kim, "A hybrid barge-in procedure for more reliable turn-taking in human-machine dialogue systems," in *Proc. ASRU*, 2003, pp. 198–203.

[2] A. Ljolje and V. Goffin, "Discriminative training of multi-state barge-in models," in *Proc. ASRU*, 2007, pp. 353–358.

[3] N. Ström and S. Seneff, "Intelligent Barge-in in Conversational Systems," in *Proc. ICSLP*, 2000.

[4] K. Matsuyama, K. Komatani, T. Ogata, and H. G. Okuno, "Enabling a User to Specify an Item at Any Time During System Enumeration – Item Identification for Barge-In-Able Conversational Dialogue Systems –," in *Interspeech*, 2009, pp. 252–255.

[5] R. Takeda, K. Nakadai, K. Komatani, T. Ogata, and H. G. Okuno, "Barge-in-able Robot Audition Based on ICA and Missing Feature Theory under Semi-Blind Situation," in *Proc. IEEE/RSJ IROS*, 2008, pp. 1718–1723.

[6] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, "Recent progress of open-source LVCSR Engine Julius and Japanese model repository," in *Proc. ICSLP*, 2004, pp. 3069–3072.