INTERSPEECH 2009 BRIGHTON

Enabling a User to Specify an Item at Any Time During System Enumeration – Item Identification for Barge-In-Able Conversational Dialogue Systems –

Kyoko Matsuyama, Kazunori Komatani, Tetsuya Ogata, Hiroshi G. Okuno

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

Abstract

In conversational dialogue systems, users prefer to speak at any time and to use natural expressions. We have developed an Independent Component Analysis (ICA) based semi-blind source separation method, which allows users to barge-in over system utterances at any time. We created a novel method from timing information derived from barge-in utterances to identify one item that a user indicates during system enumeration. First, we determine the timing distribution of user utterances containing referential expressions and then approximate it using a gamma distribution. Second, we represent both the utterance timing and automatic speech recognition (ASR) results as probabilities of the desired selection from the system's enumeration. We then integrate these two probabilities to identify the item having the maximum likelihood of selection. Experimental results using 400 utterances indicated that our method outperformed two methods used as a baseline (one of ASR results only and one of utterance timing only) in identification accuracy.

Index Terms: spoken dialogue system, conversational interaction, barge-in, utterance timing

1. Introduction

Natural conversational dialogue systems should allow users to freely express their utterances anytime. Of particular importance is that the user should be able to interrupt the system's utterances. For example, the user should be able to occasionally interrupt the system by specifying an item when the system is listing items. This ability to **barge-in** is difficult to attain with normal input devices such as a stereo microphone. Takeda *et al.* developed a sound source separation method by suppressing self-generating sound and its reflections [1]. This method is based on Independent Component Analysis (ICA) based semi-blind source separation and provides conversational dialogue systems with the capability to accept users' barge-in utterances. From such an utterance, the system can obtain not only its automatic speech recognition (ASR) result but also the timing when the user starts speaking. Thus, the spoken dialogue system can utilize the separated utterance as well as its barge-in timing to achieve a Barge-In-Able Conversational Dialogue System (BIACDS).

With the BIACDS, 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···"

User That one.

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

In this case, the user interrupts the system while it reads out "Ginkaku-ji temple." The semi-blind ICA technology first separates the user utterance "That one" with its barge-in timing and then provides both pieces of information to the **item identification** sub-system. This sub-system of BIACDS 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.

In this paper, we present a method for identifying the user's referent by focusing on barge-in utterances in the form of a pronoun, object, or abbreviation during the system's listing of choices. This sub-system that reads out each item in a list is important for two reasons. First, the user can indicate the referent by timing information, which is detected robustly. Barge-in timing is more reliable than ASR results in many cases. Second, this dialogue often appears when a system displays a retrieval result in the information retrieval task, which is a promising task in conversational dialogue systems that is being developed at several companies such as Microsoft [2] and Google [1].

Users' language expressions should not be restricted. Therefore, we handle both cases of interpreting utterances: using its barge-in timing and using its content. We integrate the two different information sources such as numerical barge-in timing and symbolic ASR results. The problems with the item identification sub-system are summarized below:

(1) Modeling barge-in timing to identify the user's referent

(2) Integrating timing information with ASR results

For the former, we determine the relationships between the timing and content of a user utterance. For the latter, we construct a framework in which both timing information and ASR results are represented as probabilities. Using these probabilistic representations, we can obtain the most relevant interpretation as the one having the maximum likelihood.

Barge-in has attracted the attention of researchers concerned with spoken dialogue systems, specifically, the issue of barge-in detection [3, 4]. Their purposes are to detect users' barge-in occurrences quickly and accurately. McTear [5] focused on how to stop a system utterance in order to recognize a user's barge-in. Ström [6] discussed a system's behavior when barge-ins were incorrectly detected. In this paper, we report a new interpretation by utilizing the locutionary act of barge-in, assuming that the barge-in detection is correct.

[1] http://www.google.com/goog411/

2. Modeling of user's utterance timing

We investigate the relationships between the content of user utterances and utterance timing to utilize barge-in timing. Here, we define **utterance timing** as the temporal subtraction of when a system utterance starts and when a user utterance starts (see Figure 1). While a system enumerates choices for selection, the user utters **referential expressions** or **content expressions** to select one item. The former indicates an utterance that contains a reference term, such as "that one" or a pronoun. The latter indicates an utterance containing content words, such as "Kinkaku-ji Temple." If the user utters a content expression, the user conveys his intention not by the timing but instead by the content. On the other hand, there must be a characteristic distribution of the utterance timing in the referential expression to convey a user's intention.

We determine a distribution of utterance timing of referential expressions. We collected user utterances under two different conditions: one was based on 35 user utterances when a system enumerated items with an average length of 0.73 seconds. The pause length between items is approximately 1.0 second (Cond. #1). The other condition was based on 69 user utterances when a system enumerated items with an average length of 5.27 seconds. The pause length was 2.0 seconds (Cond. #2). Utterance timing was detected by using the Voice Activity Detection of an ASR engine, Julius [7]. The distributions of utterance timing of both conditions are shown in Figures 2 and 3 as histograms. The bars in the histograms denote the relative frequencies of utterances in their timing, multiplied by the bar's width to represent the probabilistic density. The widths are set to 0.5 seconds. We can see clear peaks in both figures, although their peak positions and attenuation are different.

We model the histograms representing utterance timing of referential expressions by a gamma distribution:

$$f(t) = \frac{1}{(\rho-1)!\sigma^{\rho}}(t-\mu)^{\rho-1}e^{-(t-\mu)\frac{1}{\sigma}} \quad (1)$$

Zhou *et al.* also claimed that the time required for human perception follows a gamma distribution [7]. Equation (1) has three parameters: μ, ρ, and σ. We fix ρ, representing the rough shape of the distribution, to 2.0. We then assume that the remaining two parameters depend on a set of enumerated items and pause lengths between the items, and can therefore be determined beforehand on the basis of such criteria. We first set μ as the average length of time required to speak a noun. Parameter μ represents the time lag between when a system starts reading an item and when a user starts his utterance. Because a user needs to listen to at least a certain portion of an item before he decides to select it, we set the average length to μ. We then assume σ is proportional to the sum of an average length of enumerated items and the pause length. Parameter σ represents the attenuation speed of a gamma distribution. We thus set $\sigma = \beta \times$(average length of enumerated items + pause length between the items). The coefficient β is empirically set to 0.2. The gamma distributions are also illustrated in Figures 2 and 3. Their parameters are as follows: $\mu = 1.2$ and $\sigma = 0.3$ in Figure 2; $\mu = 2.2$ and $\sigma = 1.5$ in Figure 3.

3. Identifying a user's referent using barge-in timing and ASR results

We construct a framework in which both utterance timing and ASR results are uniformly represented as probabilities. This enable us to identify a user's referent as an item having the maximum likelihood.

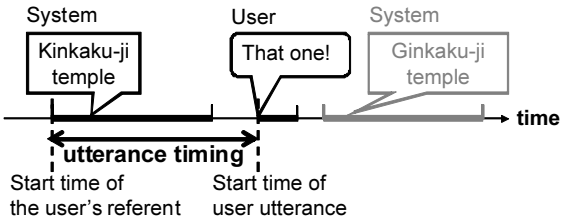

Figure 1: *Definition of utterance timing*

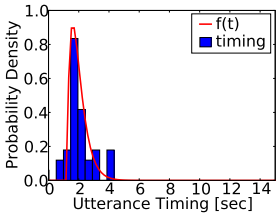

Figure 2: *Timing distribution in Cond. #1.*

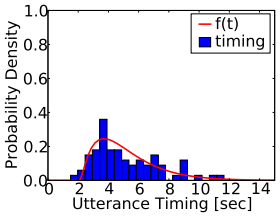

Figure 3: *Timing distribution in Cond. #2.*

3.1. Basic formulation

We formulate the problem by detecting T_i such that the probability $P(T_i|U)$ is maximized. Here, T_i denotes the i-th item enumerated by a system, and U denotes a user utterance. That is, $P(T_i|U)$ represents how probable it is that U indicates T_i corresponding to each item in the system's enumeration. We calculate the probability for each T_i and then determine the user's intention, T.

$$\begin{aligned} T &= \operatorname*{argmax}_{T_i} P(T_i|U) = \operatorname*{argmax}_{T_i} \frac{P(T_i,U)}{P(U)} \\ &= \operatorname*{argmax}_{T_i} P(T_i,U) \end{aligned} \quad (2)$$

We calculate $P(T_i,U)$ according to Equation (2) by considering the possibilities of two cases: when the user indicates his intention by either the utterance timing *or* the content of the utterance. We denote these cases as C_1 and C_2, respectively. Then, $P(T_i,U)$ can be represented as the following sum:

$$P(T_i,U) = P(T_i,U,C_1) + P(T_i,U,C_2) \quad (3)$$

Equation (3) denotes that the two cases are considered for all user utterances. $P(T_i,U,C_k)$ denotes the joint probability of an occurrence of user utterance U, which should be interpreted as case C_k and indicates the item T_i. We assume that U contains two elements: $U = \{X, t_b\}$. X indicates an ASR result and t_b denotes the time at which the user barge in the system's utterances. Both $P(T_i,U,C_1)$ and $P(T_i,U,C_2)$ are defined in the following subsections.

3.2. Probability defined by using barge-in timing

We define $P(T_i,U,C_1)$ by using utterance timing since C_1 is defined as the case when a user expresses his intention using utterance timing. Therefore, we assume probability $P(T_i,U,C_1)$ depends not on an ASR result X but on barge-in time t_b only.

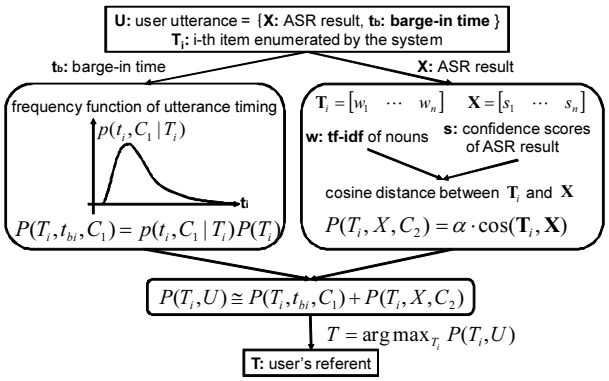

Figure 4: *Flow of identifying a user's referent*

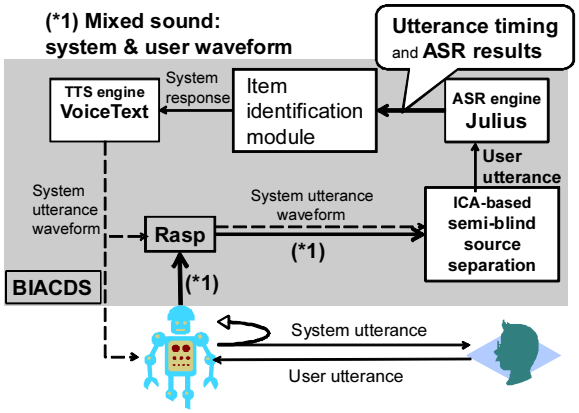

Figure 5: *System architecture*

Here, t_i denotes the utterance timing after the system starts enumerating item T_i; that is,

$$t_i = t_b - start(T_i) \quad (4)$$

Then, $P(T_i, U, C_1)$ is calculated as follows:

$$\begin{aligned} P(T_i, U, C_1) &\approx P(T_i, t_b, C_1) \\ &= P(t_i, C_1 | T_i) P(T_i) \end{aligned} \quad (5)$$

Note that the probability $P(t_i, C_1|T_i)$ represents a case when a user indicates a specific item, T_i, in timing t_i. Therefore, the probability corresponds to the gamma distribution we found in Section 2. As a result, we use the distribution $f(t_i)$ as $P(t_i, C_1|T_i)$. We also make all the prior probabilities $P(T_i)$ equal and set $P(T_i) = 1/N$. This N denotes the number of enumerated items.

3.3. Probability defined by using ASR results

We define $P(T_i, U, C_2)$ by using an ASR result according to the definition of C_2. In this case, we assume that the probability $P(T_i, U, C_2)$ do not depend on the utterance timing. This probability is a measure of the similarity between a user utterance U, that is an ASR result X, and each item T_i. To calculate this similarity, two M-dimensional vectors $\mathbf{X}$ and $\mathbf{T_i}$ are defined. The vector $\mathbf{X}$ corresponds to an ASR result of the user utterance U. M is the number of nouns contained in all items enumerated by the system. The elements of $\mathbf{T_i}$ are TF-IDF values [8] of all nouns in the enumerated items in order to account for the word importance. The vector $\mathbf{X}$ consists of ASR confidence scores for the M nouns. By considering ASR confidence scores when calculating the probability, damage caused by ASR errors is alleviated. We define the closeness using the cosine distance; that is,

$$\begin{aligned} P(T_i, U, C_2) &\approx P(T_i, X, C_2) \\ &= \alpha \cdot cos(\mathbf{T_i}, \mathbf{X}) \end{aligned} \quad (6)$$

The coefficient α adjusts the score ranges between $P(T_i, U, C_1)$ and $P(T_i, U, C_2)$. We empirically set $\alpha = 0.01$. The flow of our method of identifying a user's referent is shown in Figure 4.

4. Experimental evaluation

4.1. Implementation of BIACDS

We implemented our BIACDS, whose components are shown in Figure 5. The process flow is as follows: the sound of a user utterance is mixed with a system utterance, as the system uses a microphone embedded in a robot under a real environment. The mixed sound is synchronized with the sound waveform of the system utterance in RASP[2] and separated into the user utterance and the system utterance by the semi-blind ICA component [1]. An ASR engine, Julius [9], then recognizes the separated user utterance and records when the utterance starts. The item-identification sub-system identifies the user's referent on the basis of ASR results and the barge-in timing and generates a system response. We used VoiceText[3] developed by PENTAX Inc. as a Text-to-Speech (TTS) engine.

Our system enumerates updated news titles that are automatically obtained from RSS feeds. After the system identifies the user's referent from barge-in utterances, it reads out the details of the news the user indicates. Figure 6 represents an example in which both ASR results and utterance timing should be considered. From the user utterance "Tell me about the article on foreign students," the system cannot identify the user's referent only by the ASR result because the word "foreign student" is included in the two enumerated items. This dialogue example shows that the system should use both pieces of information even when a user utterance contains content words. Actually, our system can identify the second item "An event for foreign students *KIZUNA* ⋯" as the user's referent by using utterance timing together.

4.2. Conditions of experimental evaluation

We collected evaluation data from 20 subjects. The system listed news titles in RSS feeds, and the subjects were told they could interrupt the system utterance and speak their free expressions. We set three pause lengths between enumerated items: 1.5, 2.0, and 3.0 seconds. The parameter of a gamma distribution used in our method was determined beforehand as follows: $\mu = 0.73$.

We evaluated the identification accuracy of our method, that is, how the system correctly identified the user's referents. We set the following two baseline methods for comparisons:

Baseline #1: Using ASR results only

A user's referent was identified only by the cosine distance between an ASR result and each news title. If all cosine distances were 0, no referent was identified, and the identification was considered to have failed.

[2]Realtime Array Signal Processor (RASP) was developed by JEOL System Technology co., Ltd.

[3]http://voice.pentax.jp/

System: "Accepting foreign students at Kyoto Univ.",
"An event for foreign students, KIZUNA···"
User: Tell me about the article on foreign students.

Figure 6: *Dialogue example*

Baseline #2: Using barge-in timing only

A user's referent was the item that had just been read out or presented when a user started speaking.

We made a statistical language model based on the CIAIR corpus [10] and news articles obtained from RSS feeds. The vocabulary size was 6,831. The vector size M and the number of items N varied depending on the enumerated news articles. On average, M was 104.5, and N was 15.8. The parameter ρ of a gamma distribution was determined according to the pause lengths between items and the contents of enumerated items.

4.3. Experimental results

We collected 400 utterances from the 20 subjects. The utterances consisted of 263 reference expressions, 107 content expressions, and 30 utterances not classified as either. The ASR word accuracy for all utterances was 35.8%. One of the reasons for the low accuracy was sound reflections or distortions during the sound source separation since we used a microphone embedded in a robot instead of using a normal close-talk microphone.

The identification accuracies of the two baselines and of our method are listed in Table 1. The identification accuracy was 4.2% when only ASR results were used (Baseline #1). Most significantly, the accuracy for referential expressions was very low: 4.2%. This is because it was impossible to identify referential expressions using only the ASR results because no content words were contained in such utterances. The accuracy for content expressions was also low at 5.6%. This was directly due to the low ASR accuracy under the severe ASR condition where no close-talk microphone was used.

The identification accuracy was 64.5% when only utterance timing was used (Baseline #2). The accuracy for referential expressions was much higher that of Baseline #1. This fact indicates that timing information is useful in this situation, as expected. Furthermore, the accuracy for content expressions also improved by 24.4 points compared with Baseline #1. The results showed that the timing information was also effective for interpreting content expressions.

The identification accuracy of our method was 69.5% for all of the utterances, which outperformed the accuracies of the two baseline methods. The differences between Baseline #2 and our method for referential expressions, content expressions, and total utterances were statistically significant ($p < 0.01$) by t-tests. It is noteworthy that the accuracies of our method were better than Baseline #2 for all kinds of utterances including referential expressions. This fact indicates that considering ASR results together was effective even when the user tried to convey his intention by utterance timing.

Our method could not correctly handle 30 utterances that were neither referential expressions nor content expressions and thus were categorized as "other" in Table 1. These utterances, for instance, included "The second item please" and "I want to get the results of the game." In these cases, the user tried to convey his intention by his ASR results, but these utterances contained no content words that were included in the enumerated items. Therefore, we could not measure the closeness between the utterances and each item on the basis of the simple cosine distance. Enabling our system to handle these utterances is a task we plan to address in the future. The former example utterance will be resolved by considering anaphora expressions with numbers. We will also utilize Latent Semantic Mapping [11] to measure the latent relationship between items and ASR results for the latter example.

Table 1: *Identification accuracy [%] for user utterances*

	Referential (#:263)	Content (#:107)	Others (#:30)	Total (#:400)
(1) only ASR	4.2	5.6	0	4.2
(2) only timing	84.8	30.0	10.0	64.5
Our method	**88.2**	**39.3**	**13.3**	**69.5**

5. Conclusion

We created a novel model of users' barge-in timing and developed an identification method by integrating the timing model with ASR results as a probabilistic representation. We implemented a barge-in-able conversational dialogue system that reads out news articles obtained from RSS feeds.

Our method covers only a sub-dialogue where a user selects one item when a system lists choices. 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. Nevertheless, this study is a first step towards achieving such an intuitive interaction in conversational dialogue systems. We developed a new interaction exploiting barge-in timing and showed that it can improve the accuracy of identifying a user's referent especially in barge-in-able conversational dialogue systems.

6. References

[1] 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.

[2] Y.-Y. Wang, D. Yu, Y.-C. Ju, and A. Acero, "An introduction to voice search," *IEEE Signal Processing Magazine*, May 2008.

[3] 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.

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

[5] M. F. McTear, "Spoken Dialogue Technology: Enabling the Conversational User Interface." *ACM Computing Surveys*, pp. 90–169, 2002.

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

[7] Y. Zhou, J. Gao, K. White, I. Merk, and K. Yao, "Perceptual Dominance Time Distributions in Multistable Visual Perception," *Biological Cybernetics*, vol. 90, no. 4, pp. 256–263, 2004.

[8] G. Salton., *Automatic Text Processing*. Addison-Wesley, 1988.

[9] 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.

[10] N. Kawaguchi, S. Matsubara, K. Takeda, and F. Itakura, "CIAIR In-Car Speech Corpus -Influence of Driving Status-," in *IEICE Transactions on Information and Systems*, vol. E88-D, no. 3, 2005, pp. 578–582.

[11] J.Bellegarda, "Latent semantic mapping," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 70–80, 2005.