Improving Identification Accuracy by Extending Acceptable Utterances in Spoken Dialogue System Using Barge-in Timing

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

Graduate School of Informatics, Kyoto University, Kyoto, Japan
{matuyama,komatani,tall,ogata,okuno}@kuis.kyoto-u.ac.jp

Abstract. We describe a novel dialogue strategy enabling robust interaction under noisy environments where automatic speech recognition (ASR) results are not necessarily reliable. We have developed a method that exploits utterance timing together with ASR results to interpret user intention, that is, to identify one item that a user wants to indicate from system enumeration. The timing of utterances containing referential expressions is approximated by Gamma distribution, which is integrated with ASR results by expressing both of them as probabilities. In this paper, we improve the identification accuracy by extending the method. First, we enable interpretation of utterances including ordinal numbers, which appear several times in our data collected from users. Then we use proper acoustic models and parameters, improving the identification accuracy by 4.0% in total. We also show that Latent Semantic Mapping (LSM) enables more expressions to be handled in our framework.

Index Terms: spoken dialogue systems, 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. This ability to **barge in** is useful to convey the user's intention. The user should be able to occasionally interrupt the system by specifying an item when the system is listing items. 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." This system identifies the user's referent, that is, what the user indicates by

N. García-Pedrajas et al. (Eds.): IEA/AIE 2010, Part II, LNAI 6097, pp. 585–594, 2010.

"That one." By using the barge-in timing of the user utterance, it determines that "Ginkaku-ji Temple" is specified by the user. This kind of dialogue in which items are read out 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. Therefore, this new dialogue strategy enables the system to obtain the user intention by reading out each item even in noisy environments. Second, this dialogue often appears when a system displays a retrieval result in the information retrieval task. This task is a promising one for conversational dialogue systems and is being developed at several companies such as Microsoft [1] and Google[1].

We have developed a method for identifying the user's *referent* during system enumeration by focusing on barge-in utterances while the system lists choices [2]. Our purpose is to identify the user's referent with a high degree of accuracy. We exploit utterance timing together with ASR results to identify the user's referent as follows. First, we determine the relationships between the timing and content of a user utterance in order to use timing information. Then we construct a framework in which both timing information and ASR results are represented as probabilities. By using these probabilistic representations, we can obtain the most relevant interpretation as the one having the maximum likelihood [2]. We furthermore improve the interpretation obtained from ASR results in order to handle user utterances that include no content words in each item. Specifically, we introduce the interpretation of utterances with numbers. We also propose interpreting utterances that include words related to the items. We collect documents from the Web and use Latent Semantic Mapping (LSM) [3] to measure the closeness between the utterance and each item.

Interpretation using utterance timing has not been investigated although barge-in has attracted the attention of researchers concerned with spoken dialogue systems, specifically, the issue of barge-in detection [4,5]. Their purpose has been to detect users' barge-in occurrences quickly and accurately. McTear [6] focused on how to stop a system utterance in order to recognize a user's barge-in. Strӧm [7] discussed a system's behavior when barge-ins were incorrectly detected. We report a new interpretation that utilizes the locutionary act of barge-in, on the assumption that the barge-in detection is correct.

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 a 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,

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

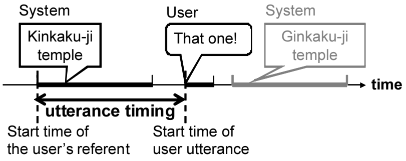

Fig. 1. Definition of utterance timing

Table 1. Two different conditions

Condition	Cond. A	Cond. B
# user utterances	35	69
PAUSE (sec.)	1.0	2.0
AVERAGE (sec.)	0.73	5.27

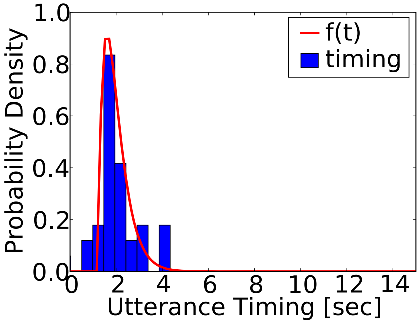

Fig. 2. Timing distribution in Cond. A

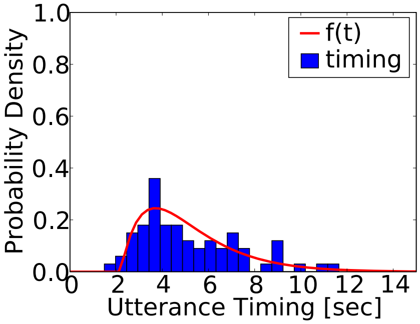

Fig. 3. Timing distribution in Cond. B

the user conveys his intention not by the timing but by the content. On the other hand, a characteristic distribution of the utterance timing must be in the referential expression to convey a user's intention.

We determine how utterance timing of referential expressions is distributed. We collected user utterances under two different conditions (see Table 1). PAUSE represents the interval of time between items and AVERAGE represents an average length of enumerated items. Utterance timing is detected by using the voice activity detection of an ASR engine, Julius [8]. 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 Gamma distribution:

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

Zhou *et al.* also claimed that the time required for human perception follows Gamma distribution [9]. Equation (1) has three parameters: μ, ρ, and σ. The details of how these parameters are set was explained in our previous paper [2]. The Gamma distributions are also illustrated in Figures 2 and 3. Their parameters are as follows: $\mu = 1.2$, $\rho = 0.3$ and $\sigma = 2.0$ in Figure 2; $\mu = 2.2$, $\rho = 1.5$and $\sigma = 2.0$ in Figure 3.

3 Identifying User's Referent Using Barge-in Timing and ASR Results

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

3.1 Basic Formulation

We formulate the problem of identifying a user's referent by calculating 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.

$$T = \underset{T_i}{\operatorname{argmax}} P(T_i|U) = \underset{T_i}{\operatorname{argmax}} \frac{P(U|T_i)P(T_i)}{P(U)} = \underset{T_i}{\operatorname{argmax}} P(U|T_i) \tag{2}$$

We assume all the prior probabilities $P(T_i)$ are equal. $P(U)$ is not dependent on i.

We calculate $P(U|T_i)$ in accordance with Equation (2) by considering the possibilities of two cases: interpreting user's intention by either the utterance timing, C_1 *or* the content of the utterance, C_2. Thus, $P(U|T_i)$ can be represented as the following sum:

$$P(U|T_i) = \Sigma_{k=1,2} P(U|T_i, C_k) P(C_k|T_i) \tag{3}$$

$$= \frac{1}{2} \Sigma_{k=1,2} P(U|T_i, C_k) \tag{4}$$

Here we assume that these prior probabilities $P(C_k|T_i)$ are even. We set the coefficient α as the score ranges between $P(U|T_i, C_1)$ and $P(U|T_i, C_2)$ by setting a parameter α , as shown in Equation (5).

$$P(U|T_i) = (1 - \alpha) P(U|T_i, C_1) + \alpha P(U|T_i, C_2) \tag{5}$$

Equation (5) denotes that the two cases are considered for all user utterances. $P(U|T_i, C_k)$ denotes the probability of an occurrence of user utterance U in the case of C_k for each item T_i. We assume that U contains two elements: $U = \{X, t_b\}$. Here, X indicates an ASR result and t_b denotes the time at which the user barges in during the system's utterance. Both $P(U|T_i, C_1)$ and $P(U|T_i, C_2)$ are defined in the following subsections. The flow of our method of identifying a user's referent is shown in Figure 4.

3.2 Probability Defined by Using Barge-in Timing

We define $P(U|T_i, C_1)$ by using utterance timing since C_1 is defined as the case when a user expresses his intention by using utterance timing. Therefore, we

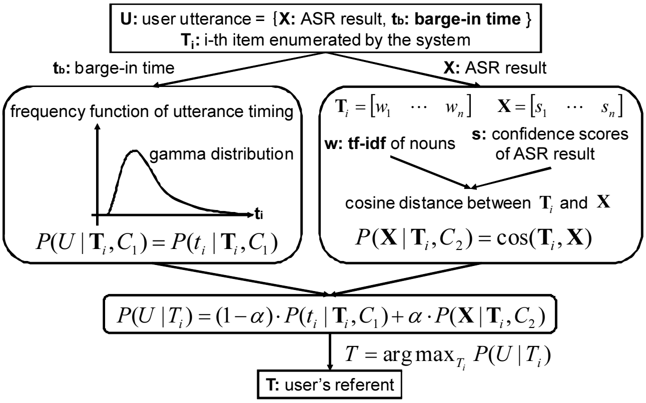
U: user utterance = {X: ASR result, tb: barge-in time}
Ti: i-th item enumerated by the system
tb: barge-in time
X: ASR result
frequency function of utterance timing
gamma distribution
ti
$P(U|\mathbf{T}_i,C_1)=P(t_i|\mathbf{T}_i,C_1)$
$\mathbf{T}_i=[w_1 \cdots w_n]$ $\mathbf{X}=[s_1 \cdots s_n]$
w: tf-idf of nouns
s: confidence scores of ASR result
cosine distance between $\mathbf{T}_i$ and $\mathbf{X}$
$P(\mathbf{X}|\mathbf{T}_i,C_2)=\cos(\mathbf{T}_i,\mathbf{X})$
$P(U|T_i)=(1-\alpha)\cdot P(t_i|\mathbf{T}_i,C_1)+\alpha\cdot P(\mathbf{X}|\mathbf{T}_i,C_2)$
$T=\arg\max_{T_i} P(U|T_i)$
T: user's referent

Fig. 4. Flow of identifying user's referent

assume probability $P(U|T_i, C_1)$ depends not on an ASR result X but on barge-in time t_b only. Here, t_i denotes the utterance timing after the system starts enumerating item T_i (see Figure 1); that is,

$$t_i = t_b - start(T_i) \tag{6}$$

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

$$P(U|T_i, C_1) = P(t_i|T_i, C_1) \tag{7}$$

Note that the probability $P(t_i|T_i, C_1)$ 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. We use the distribution $f(t_i)$ as $P(t_i|T_i, C_1)$.

3.3 Probability Defined by Using ASR Results

The probability $P(U|T_i, C_2)$ represents how close a user utterance U (ASR result X) and each item T_i are. We define $P(U|T_i, C_2)$ by using an ASR result in accordance with the definition of C_2 [2], except for some utterances for which we also need to use barge-in timing t_b. One example utterance is "The item before last." This example needs to be interpreted by using both the user's barge-in timing and the ASR result. That is, we need to know what a user said, and when.

The closeness is defined by cosine distance:

$$P(U|T_i, C_2) = cos(\mathbf{T_i}, \mathbf{X}) \tag{8}$$

where $\mathbf{X}$ and $\mathbf{T_i}$ are M-dimensional vectors. M is the vocabulary size of the system. The elements of $\mathbf{T_i}$ are TF-IDF values [10] of all nouns in the enumerated items in order to account for the word importance. The vector $\mathbf{X}$ corresponds to

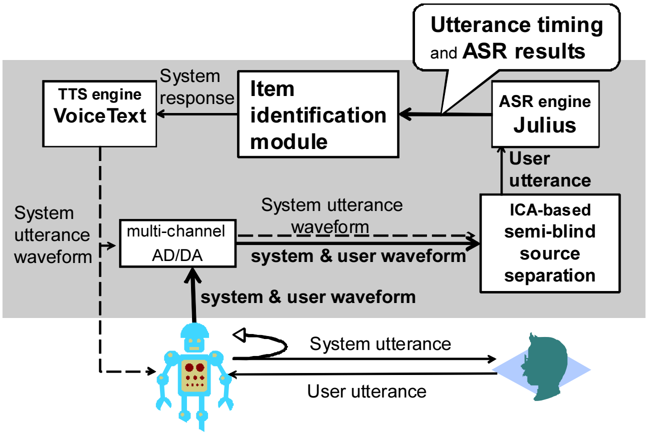

Fig. 5. System architecture

the ASR result for the user utterance U. This vector 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.

To interpret utterances that include numbers such as “The second one”, we add such number words into the vocabulary. For example, “first” is added to the vector T_1 corresponding to the first item. The size of X also increases accordingly. After adding these words, “the second” can be interpreted to indicate the second item in the system's enumeration, for example. When “The item before last” is recognized, we first estimate the interrupted item by using barge-in timing t_b and calculate the most likely user selection by the number of the items and ASR result. Then we assign the average confidence scores for words in ASR results to the corresponding element of vector X.

4 Experimental Evaluation

4.1 Implementation of Barge-in-able Dialogue System

The overview of the architecture of our barge-in-able dialogue system is depicted in Figure 5. The process flow is summarized as follows: The multi-channel AD/DA, RASP of JOEL System Technology captures a mixed sound and the wave file of the system utterance into a 2-channel wave stream. The ICA-based semi-blind source separation [11] obtains this wave stream and separates the user utterance incrementally. The ASR engine, Julius [8], then recognizes the separated user utterance and begins to record 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 Voice-Text[2] developed by PENTAX Inc. as a Text-to-Speech (TTS) engine.

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

4.2 Conditions of Experimental Evaluation

We collected 400 utterances from 20 subjects. The utterances consisted of 263 referential expressions and 137 content expressions. The system listed news titles in 10 RSS feeds, and the subjects were told they could interrupt the system utterance and say whatever they liked. The number and length of titles are different for each RSS feed. We set three pause lengths between enumerated items: 1.5, 2.0, and 3.0 seconds. The parameters of the Gamma distribution used in our method were determined beforehand as follows: $\mu = 0.73$ and $\sigma = 2.0$. The parameter ρ of Gamma distribution was determined in accordance with the pause lengths between items and the contents of enumerated items. We set α in Equation (5) to 0.6 empirically. Accuracies when α is changed are shown in Section 4.3. We used an acoustic model containing pink noise, which reflects the actual acoustic environment. We made a statistical language model by using the CIAIR corpus [12] and news articles obtained from each RSS feed. On average, the vocabulary size was 5835.

We evaluated several methods by identification accuracies, that is, how well the system correctly identified the user's referents. Each method is listed below:

Cond. 1: Use of barge-in timing only
A user's referent was the item that had just been read out or presented when a user started speaking.

Cond. 2: Use of barge-in timing model only
A user's referent was identified by the using the timing model of Gamma distribution.

Cond. 3: Our method (not extended to interpret numbers)
A user's referent was identified by the identical method to [2].

Cond. 4: Our method (explained in Section 3)
A user's referent was identified by our method extended to interpret numbers.

Conds. 1 and 2 correspond to simpler methods in which no ASR results are used. We set these to verify how well the timing model works and whether ASR results are necessary or not. In Cond. 3, the vector size M and the number of items N varied with the number of enumerated news articles. On average, M was 104.5, and N was 15.8. In Cond. 4, M was 173.5. The ASR word accuracy for all utterances was 38.3%. Reasons for the low accuracy include 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. Also, correctly recognizing a user's utterances is difficult because these users often speak quickly or quietly.

4.3 Experimental Results

The identification accuracies of the user's referent for 263 utterances with referential expressions, 137 utterances with content expressions, and all 400 utterances are shown in Table 2. Accuracy of Cond. 2 was better than that of Cond. 1.

Table 2. Identification accuracy [%] for user utterances

Condition	Referential expression (#:263)	Content expression (#:137)	Total (#:400)
1: only barge-in timing	84.8	25.5	64.5
2: only barge-in timing model	87.8	32.1	68.8
3: our method	81.4	53.3	71.8
4: + **numbers**	**85.2**	**57.7**	**75.8**

Table 3. Identification accuracy [%] for α in Cond. 4

α value	Referential expression (#:263)	Content expression (#:137)	Total (#:400)
0.0	87.8	32.1	68.8
0.2	86.7	42.3	71.5
0.4	85.9	54.7	75.3
0.6	85.2	57.7	75.8
0.8	84.8	56.9	75.3
1.0	0.76	43.1	15.3

This result shows the utterance timing model formulated as Gamma distribution works effectively. Moreover, the timing information is also effectively used for interpreting content expressions, because some content utterances were identified correctly even though users conveyed their referent by content words.

The identification accuracy of Cond. 3 was 71.8% for all utterances, outperforming the accuracies of Conds. 1 and 2. In particular, the accuracy for content expressions also improved by 21.2 points compared with that of Cond. 2. The result suggests using the ASR results is effective although its accuracy is not high. The identification accuracy of Cond. 4 was 75.8% for all utterances, which outperformed the accuracy of Cond. 3. In fact, the identification accuracy of content expressions including numbers improved by 27 points more than that of Cond. 3. The differences between Cond. 3 and 4 for referential expressions and total utterances were statistically significant ($p < 0.01$) by t-tests. Most significantly, the accuracy for referential expressions of Cond. 4 also improved by 3.8 points more than that of Cond. 3. These utterances can be identified after scores of incorrect ASR results decreased due to the number being considered.

The highest accuracy of referential expressions was obtained by Cond. 2. This case corresponds to $\alpha = 0.0$. Table 3 lists identification accuracies in Cond. 4 when α is changed from 0.0 to 1.0. When we set α to 1.0, a user's referent is identified by only $P(U|T_i, C_2)$. In this case, the identification accuracy of referential expressions is very low because ASR results of referential expressions such as "That one" contain no information associated with any items. When we set α smaller, $P(U|T_i, C_1)$ was emphasized and more referential expressions were correctly identified. This result indicates the trade-off between $P(U|T_i, C_1)$ and $P(U|T_i, C_2)$. To improve the accuracy for referential expressions in our methods, we should dynamically determine α in Equation (5) for each user's utterance.

Table 4. Identification accuracy [%] by using LSM

Condition	Referential expression (#: 263)	Content expression (#: 137)	Total (#: 400)
Using LSM	85.2	58.3	76.0

5 Extending Acceptable Utterances by LSM

The user often tries to convey his or her intention using related words, that is, content words that were not included in the enumerated items. This utterance, for instance, includes "The Beckham's result" corresponding to the item "Soccer." To deal with this utterance, we collect the documents obtained by copying sentences from Wikipedia[3] pages related to each item. Here, $P(U|T_i, C_2)$ represents how close a user utterance U (ASR result X) and the documents from the Web corresponding to each item T_i are, and it is calculated by using LSM [3]. We decompose the co-occurrence matrix to obtain the k-dimensional vectors of all the documents. We construct a $M \times N$ co-occurrence matrix between the items and the documents, where M is the vocabulary size and N is the total number of the documents. We applied singular value decomposition (SVD) to the matrix and compressed its rank to k. Here, k corresponds to $N - 2$. The k-dimensional vectors were calculated on the basis of the matrix obtained from the SVD.

We estimate $P(U|T_i, C_2)$ by calculating the cosine distance between the k-dimensional vectors of the user's utterance and those of the documents. The user's utterance was recognized using a statistical language model that was based on the documents for each RSS feed. The documents consist of the data from Wikipedia and the 115 command utterances such as "Let me hear the news". On average, the size of the vocabulary was 17253. The ASR word accuracy was 37.5%. The size of the co-occurrence matrix M corresponds to the size of vocabulary. The k-dimensional vector of the user's utterance was calculated from its ASR confidence scores and the matrix obtained from the SVD.

We apply LSM only when a user specifies the item by using related words to avoid misinterpretation by applying LSM to all utterances. We compare two acoustic likelihoods to select utterances to apply LSM. One is calculated by using a language model for LSM and the other by using language model used in Cond. 4. We obtain the difference between them by subtracting the latter from the former. We use LSM only when the difference is more than 90. This value is empirically determined.

We evaluated the effectiveness of using LSM to identify the user's referent. The identification accuracy by using LSM is shown in Table 4. Here we set α in Equation (5) to 0.6. Table 4 shows that the identification accuracy outperformed that of Cond. 4. In fact, the one utterance that only has a content expression with related words in the data became identified correctly.

[3] http://ja.wikipedia.org/

6 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. As a result, we made 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 work is the first step towards achieving such an intuitive interaction in conversational dialogue systems. We developed a new interaction exploiting barge-in timing model and showed that it can improve the accuracy of identifying a user's referent, especially in barge-in-able conversational dialogue systems.

References

1. Wang, Y.Y., Yu, D., Ju, Y.C., Acero, A.: An introduction to voice search. IEEE Signal Processing Magazine (May 2008)
2. Matsuyama, K., Komatani, K., Ogata, T., Okuno, H.G.: 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 (2009)
3. Bellegarda, J.R.: Latent semantic mapping. IEEE Signal Processing Magazine 22(5), 70–80 (2005)
4. Rose, R.C., Kim, H.K.: A hybrid barge-in procedure for more reliable turn-taking in human-machine dialogue systems. In: Proceeding of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 198–203 (2003)
5. Ljolje, A., Goffin, V.: Discriminative training of multi-state barge-in models. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 353–358 (2007)
6. McTear, M.F.: Spoken Dialogue Technology: Enabling the Conversational User Interface. ACM Computing Surveys, 90–169 (2002)
7. Ström, N., Seneff, S.: Intelligent Barge-in in Conversational Systems. In: Proceeding of International Conference on Spoken Language Processing (2000)
8. Kawahara, T., Lee, A., Takeda, K., Itou, K., Shikano, K.: Recent progress of open-source LVCSR Engine Julius and Japanese model repository. In: Proceeding of International Conference on Spoken Language Processing, pp. 3069–3072 (2004)
9. Zhou, Y., Gao, J., White, K., Merk, I., Yao, K.: Perceptual Dominance Time Distributions in Multistable Visual Perception. Biological Cybernetics 90(4), 256–263 (2004)
10. Salton., G.: Automatic Text Processing. Addison-Wesley, Reading (1988)
11. Takeda, R., Nakadai, K., Komatani, K., Ogata, T., Okuno, H.G.: Barge-in-able Robot Audition Based on ICA and Missing Feature Theory under Semi-Blind Situation. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1718–1723 (2008)
12. Kawaguchi, N., Matsubara, S., Takeda, K., Itakura, F.: CIAIR In-Car Speech Corpus -Influence of Driving Status-. IEICE Transactions on Information and Systems, 578–582 (2005)