Online Error Detection of Barge-In Utterances by Using Individual Users' Utterance Histories in Spoken Dialogue System

Kazunori Komatani* **Hiroshi G. Okuno**

Kyoto University
Yoshida-Hommachi, Sakyo, Kyoto 606-8501, Japan
{komatani,okuno}@kuis.kyoto-u.ac.jp

Abstract

We develop a method to detect erroneous interpretation results of user utterances by exploiting utterance histories of individual users in spoken dialogue systems that were deployed for the general public and repeatedly utilized. More specifically, we classify barge-in utterances into correctly and erroneously interpreted ones by using features of individual users' utterance histories such as their barge-in rates and estimated automatic speech recognition (ASR) accuracies. Online detection is enabled by making these features obtainable without any manual annotation or labeling. We experimentally compare classification accuracies for several cases when an ASR confidence measure is used alone or in combination with the features based on the user's utterance history. The error reduction rate was 15% when the utterance history was used.

1 Introduction

Many researchers have tackled the problem of automatic speech recognition (ASR) errors by developing ASR confidence measures based on utterance-level (Komatani and Kawahara, 2000) or dialogue-level information (Litman et al., 1999; Walker et al., 2000; Hazen et al., 2000). Especially in systems deployed for the general public such as those of (Komatani et al., 2005; Raux et al., 2006), the systems need to correctly detect interpretation errors caused by various utterances made by various users, including novices. Error detection using individual user models would be a promising way of improving performance in such systems because users often access them repeatedly (Komatani et al., 2007).

We choose to detect interpretation errors of barge-in utterances, mostly caused by ASR errors, as a task for showing the effectiveness of the user's utterance histories. We try to improve the accuracy of classifying barge-in utterances into correctly and erroneously interpreted ones without any manual labeling. By classifying utterances accurately, the system can reduce erroneous responses caused by the errors and unnecessary confirmations. Here, a "barge-in utterance" is a user utterance that interrupts the system's prompt. In this situation, the system stops its prompt and starts recognizing the user utterance.

In this study, we combine the ASR confidence measure with features obtained from the user's utterance history, i.e., the estimated ASR accuracy and the barge-in rate, to detect interpretation errors of barge-in utterances. We show that the features are still effective when they are used together with the ASR confidence measure, which is usually used to detect erroneous ASR results. The characteristics of our method are summarized as follows:

1. The user's utterance history used as his/her profile: The user's current barge-in rate and ASR accuracy are used for error detection.

2. Online user modeling: We try to obtain the user profiles listed above without any manual labeling after the dialogue has been completed. This means that the system can improve its performance while it is deployed.

In our earlier report (Komatani and Rudnicky, 2009), we defined the estimated ASR accuracy and showed that it is helpful in improving the accuracy of classifying barge-in utterances into correctly and erroneously interpreted ones, by using it in conjunction with the user's barge-in rate. In this

*Currently with Graduate School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan. komatani@nuee.nagoya-u.ac.jp

Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 289–296,
The University of Tokyo, September 24-25, 2010.

Table 1: ASR accuracy per barge-in

	Correct	Incorrect	Total	Accuracy
w/o barge-in	16,694	3,612	20,306	(82.2%)
w/ barge-in	3,281	3,912	7,193	(45.6%)
Total	19,975	7,524	27,499	(72.6%)

report, we verify our approach when the ASR confidence measure is also incorporated into it. Thus, we show the individual user's utterance history is helpful as a user profile and works as prior information for the ASR confidence.

2 Barge-in Utterance and its Errors

Barge-in utterances were often incorrectly interpreted mainly because of ASR errors in our data as shown in Table 1. The table lists the ASR accuracy per utterance for two cases: when the system prompts were played to the end (denoted as "w/o barge-in") and when the system prompts were barged in ("w/ barge-in"). Here, an utterance is assumed to be correct only when all content words in the utterance are correctly recognized; one is counted as an error if any word in it is misrecognized. Table 1 shows that barge-in utterances amounted to 26.2% (7,193/27,499) of all utterances, and half of those utterances contained ASR errors in their content words.

This result implies that many false barge-ins occurred despite the user's intention. Specifically, the false barge-ins included instances when background noises were incorrectly regarded as barge-ins and the system's prompt stopped. Such instances often occur when the user accesses the system using mobile phones in crowded places. Breathing and whispering were also prone to be incorrectly regarded as barge-ins. Moreover, disfluency in one utterance may be unintentionally divided into two portions, which causes further misrecognitions and unexpected system actions. The abovementioned phenomena, except background noises, are caused by the user's unfamiliarity with the system. That is, some novice users are not unaware of the timing at which to utter, and this causes the system to misrecognize the utterance. On the other hand, users who have already become accustomed to the system often use the barge-in functions intentionally and, accordingly, make their dialogues more efficient.

The results in Table 2 show the relationship between barge-in rate per user and the corresponding ASR accuracies of barge-in utterances. We

Table 2: ASR accuracy of barge-in utterances for different barge-in rates

Barge-in rate	Correct	Incorrect	Acc. (%)
0.0 - 0.2	407	1,750	18.9
0.2 - 0.4	205	842	19.6
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	3,281	3,912	45.6

here ignore a small number of users whose barge-in rates were greater than 0.8, which means almost all utterances were barge-ins, because most of their utterances were misrecognized because of severe background noises and accordingly they gave up using the system. We thus focus on users whose barge-in rates were less than 0.8. The ASR accuracy of barge-in utterances was high for users who frequently barged-in. This suggests that the barge-ins were intentional. On the other hand, the ASR accuracies of barge-in utterances were less than 20% for users whose barge-in rates were less than 0.4. This suggests that the barge-ins of these users were unintentional.

A user study conducted by Rose and Kim (2003) revealed that there are many more disfluencies when users barge in compared with when users wait until the system prompt ends. Because such disfluencies and resulting utterance fragments are parts of human speech, it is difficult to select erroneous utterances to be rejected by using a classifier that distinguishes speech from noise on the basis of the Gaussian Mixture Model (Lee et al., 2004). These errors cannot be detected by using only bottom-up information obtained from single utterances such as acoustic features and ASR results.

To cope with the problem, we use individual users' utterance histories as their profiles. More specifically, we use each user's average barge-in rate and ASR accuracy from the time the user started using the system until the current utterance. The barge-in rate intuitively corresponds to the degree to which the user is accustomed to using the system, especially to using its barge-in function. That is, this reflects the tendency shown in Table 2; that is, the ASR accuracy of barge-in utterances is higher for users whose barge-in rates are higher. Each user's ASR accuracy also indicates the user's habituation. This corresponds to an empirical tendency that ASR accuracies of more accustomed

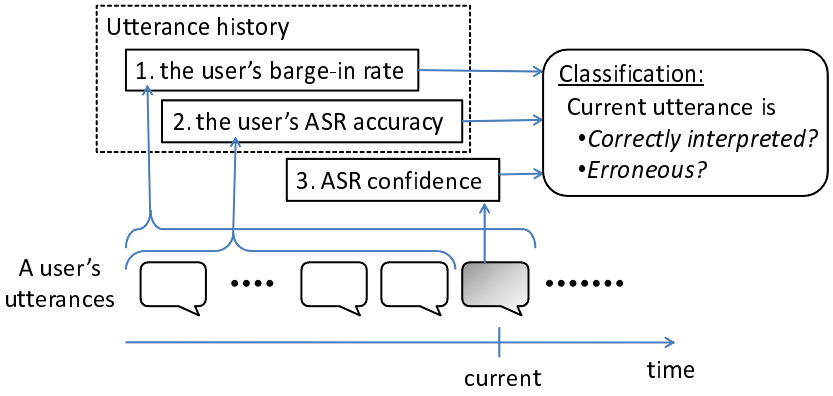

Figure 1: Overview of detecting interpretation errors

users are higher (Komatani et al., 2007; Levow, 2003). To account for another fact that some expert users have low barge-in rates, and, accordingly, not all expert users barge in frequently (Komatani et al., 2007), we use both the user's barge-in rate and ASR accuracy to represent degree of habituation, and verify their effectiveness as prior information for detecting erroneous interpretation results when they are used together with an ASR confidence measure.

To obtain the user's ASR accuracy without any manual labeling, we exploit certain dialogue patterns indicating that ASR results at certain positions are reliable. For example, Sudoh and Nakano (2005) proposed a "post-dialogue confidence scoring" in which ASR results corresponding to the user's intention upon dialogue completion are assumed to be correct and are used for confidence scoring. Bohus and Rudnicky (2007) proposed "implicitly supervised learning" in which user responses following the system's explicit confirmations are used for confidence scoring. If the ASR results can be regarded as reliable after the dialogue, machine learning algorithms can use them as teacher signals. This approach does not need any manual labeling or transcription, a task which requires much time and labor when spoken dialogue systems are being developed. We focus on users' affirmative and negative responses to the system's explicit confirmations, and estimated the user's ASR accuracy on the basis of his or her history of responses (Komatani and Rudnicky, 2009). This estimated ASR accuracy can be also used as an online feature representing a user's utterance history.

3 Detecting Errors by using the User's Utterance History

We detect interpretation errors of barge-in utterances by using the following three information sources:

1. the current user's barge-in rate,
2. the current user's ASR accuracy, and
3. ASR confidence of the current utterance.

The error detection method is depicted in Figure 1. Barge-in rate and ASR accuracy are accumulated and averaged from the beginning until the current utterance and are used as each user's utterance history. Then, at every point a user makes an utterance, the barge-in utterances are classified into correctly or erroneously interpreted ones by using a logistic regression function:

$$P = \frac{1}{1 + \exp(-(a_1 x_1 + a_2 x_2 + a_3 x_3 + b))}, \quad (1)$$

where x_1, x_2 and x_3 denote the barge-in rate, the ASR accuracy until the current utterance, and the ASR confidence measure of the current utterance, respectively. Coefficients a_i and b are determined by 10-fold cross validation on evaluation data. In the following subsections, we describe how to obtain these features.

3.1 Barge-In Rate

The barge-in rate is defined as the ratio of the number of barge-in utterances to all the user's utterances until the current utterance. Note that the current utterance itself is included in this calculation. We confirmed that the barge-in rate changes as the user becomes accustomed to the system

U1:	205. (Number 100)
S1:	Will you use bus number 100?
U2:	No. (No)
S2:	Please tell me your bus stop or bus route number.
U3:	Nishioji Matsu... *[disfluency]* (*Rejected*)
S3:	Please tell me your bus stop or bus route number.
U4:	From Nishioji Matsubara. (From Nishioji Matsubara)
S4:	Do you get on a bus at Nishioji Matsubara?
U5:	Yes. (Yes)

Initial characters 'U' and 'S' denote the user and system utterance.
A string in parentheses denotes the ASR result of the utterance.

Figure 2: Example dialogue

(Komatani et al., 2007). To take these temporal changes into consideration, we set a window when calculating the rate (Komatani et al., 2008). That is, when the window width is N, the rate is calculated on the basis of only the last N utterances, and utterances before those ones are discarded. When the window width exceeds the total number of utterances by the user, the barge-in rate is calculated on the basis of all the user's utterances. Thus, when the width exceeds 2,838, the maximum number of utterances made by one user in our data, the barge-in rates equal the average rates of all utterances by the user.

3.2 ASR Accuracy

ASR accuracy is calculated per utterance. It is defined as the ratio of the number of correctly recognized utterances to all the user's utterances until the previous utterance. Note that the current utterance is not included in this calculation. The "correctly recognized" utterance denotes a case when every content word in the ASR result of the utterance was correctly recognized and no content word was incorrectly inserted. The ASR accuracy of the user's initial utterance is regarded as 0, because there is no utterance before it. We do not set any window when calculating the ASR accuracies, because classification accuracy did not improve as a result of setting one (Komatani and Rudnicky, 2009). This is because each users' ASR accuracies tend to converge faster than the barge-in rates do (Komatani et al., 2007), and the changes in the ASR accuracies are relatively small in comparison with those of the barge-in rates.

We use two kinds of ASR accuracies:

1. actual ASR accuracy and
2. estimated ASR accuracy (Komatani and Rudnicky, 2009).

The actual ASR accuracy is calculated from manual transcriptions for investigating the upper limit of improvement of the classification accuracy when ASR accuracy is used. Thus, it cannot be obtained online because manual transcriptions are required.

The estimated ASR accuracy is calculated on the basis of the user's utterance history. This is obtainable online, that is, without the need for manual transcriptions after collecting the utterances. We focus on users' affirmative or negative responses following the system's explicit confirmations, such as "Leaving from Kyoto Station. Is that correct?" To estimate the accuracy, we make three assumptions as follows:

1. The ASR results of the users' affirmative or negative responses are correctly recognized. This assumption will be verified in Section 4.2.

2. A user utterance corresponding to the content of the affirmative responses is also correctly recognized, because the user affirms the system's explicit confirmation for it.

3. The remaining utterances are not correctly recognized. This corresponds to when users do not just say "no" in response to explicit confirmations with incorrect content and instead use other expressions.

To summarize the above, we assume that the ASR results of the following utterances are correct: an affirmative response, its corresponding utterance which is immediately preceded by it, and

Table 3: Distribution of ASR confidence measures for barge-in utterances

Confidence measure	Correct	Incorrect	(%)
0.0 - 0.1	0	1491	0.0
0.1 - 0.2	0	69	0.0
0.2 - 0.3	0	265	0.0
0.3 - 0.4	0	708	0.0
0.4 - 0.5	241	958	20.1
0.5 - 0.6	639	333	65.7
0.6 - 0.7	1038	68	93.9
0.7 - 0.8	1079	20	98.2
0.8 - 0.9	284	0	100.0
0.9 - 1.0	0	0	–
Total	3281	3912	45.6

a negative response. All other utterances are assumed to be incorrect. We thus calculate the user's estimated ASR accuracy as follows:

$$(\text{Estimated ASR accuracy}) = \frac{2 \times (\#\text{affirmatives}) + (\#\text{negatives})}{(\#\text{all utterances})} \quad (2)$$

Here is an example of the calculation for the example dialogue shown in Figure 2. U2 is a negative response, and U5 is an affirmative response. When the dialogue reaches the point of U5, U2 and U5 are regarded as correctly recognized on the basis of the first assumption. Next, U4 is regarded as correct on the basis of the second assumption, because the explicit confirmation for it (S4) was affirmed by the user as U5. Then, the remaining U1 and U3 are regarded as misrecognized on the basis of the third assumption. As a result, the estimated ASR accuracy at U5 is 60%.

The estimated ASR accuracy is updated for every affirmative or negative response by the user. For a neither affirmative nor negative response, the latest estimated accuracy before it was used instead.

3.3 ASR Confidence Measure

We use an ASR confidence measure calculated per utterance. Specifically, we use the one derived from the ASR engine in the Voice Web Server, a product of Nuance Communications, Inc.[1]

Table 3 shows the distribution of ASR confidence measures for barge-in utterances. By using this ASR confidence, even a naive method can have high classification accuracy (90.8%) in which just one threshold ($\theta = 0.516$) is set and utterances whose confidence measure is greater than the threshold are accepted. This accuracy is regarded as the baseline.

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

Table 4: ASR accuracy by user response type

	Correct	Incorrect	Total	(Acc.)
Affirmative	9,055	243	9,298	(97.4%)
Negative	2,006	286	2,292	(87.5%)
Other	8,914	6,995	15,909	(56.0%)
Total	19,975	7,524	27,499	(72.6%)

4 Experimental Evaluation

4.1 Data

We used data collected by the Kyoto City Bus Information System (Komatani et al., 2005). This system locates a bus that a user wants to ride and tells the user how long it will be before the bus arrives. The system was accessible to the public by telephone. It adopted the safest strategy to prevent erroneous responses; that is, it makes explicit confirmations for every user utterance except for affirmative or negative responses such as "Yes" or "No".

We used 27,499 utterances that did not involve calls whose phone numbers were not recorded or those the system developer used for debugging. The data contained 7,988 valid calls from 671 users. Out of these, there were 7,193 barge-in utterances (Table 1). All the utterances were manually transcribed for evaluation; human annotators decided whether every content word in the ASR results was correctly recognized or not.

The phone numbers of most of the calls were recorded, and we assumed that each number corresponded to one individual. Most of the numbers were those of mobile phones, which are usually not shared; thus, the assumption seems reasonable.

4.2 Verifying Assumption in Calculating Estimated ASR Accuracy

We confirmed our assumption that the ASR results of affirmative or negative responses following explicit confirmations are correct. We classified the user utterances into affirmatives, negatives, and other, and calculated the ASR accuracies (precision rates) per utterance as shown in Table 4. Affirmatives include *hai* ('yes'), *soudesu* ('that's right'), OK, etc; and negatives include *iie* ('no'), *chigaimasu* ('I don't agree'), *dame* ('No good'), etc. The table indicates that the ASR accuracies of affirmatives and negatives were high. One of the reasons for the high accuracy was that

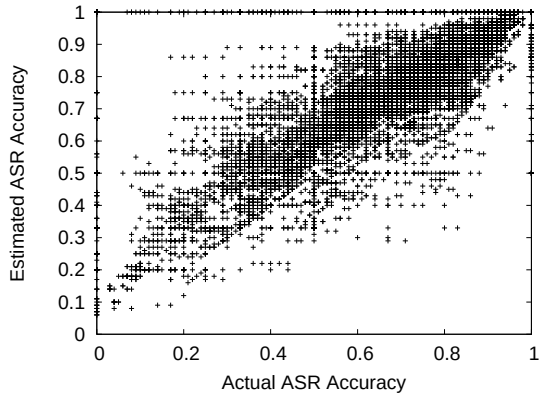

Figure 3: Correlation between actual and estimated ASR accuracy

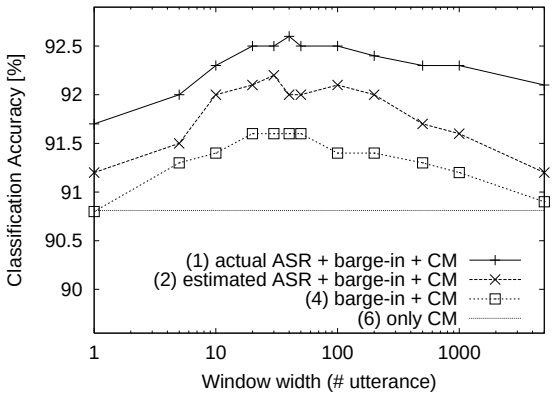

Figure 4: Classification accuracy when window width varies used to calculate barge-in rate

these utterances are much shorter than other content words, so they were less confused with other content words. Another reason was that the system often gave help messages such as "Please answer *yes* or *no*."

We then analyzed the correlation between the actual ASR accuracy and the estimated ASR accuracy based on Equation 2. We plotted the two ASR accuracies (Figure 3) for 26,231 utterances made after at least one affirmative/negative response by the user. The correlation coefficient between them was 0.806. Although the assumption that all ASR results of affirmative/negative responses are correct might be rather strong, the estimated ASR accuracy had a high correlation with the actual ASR accuracy.

4.3 Comparing Classification Accuracies When the Used Features Vary

We investigated the classification accuracy of the 7,193 barge-in utterances. The classification accuracies are shown in Table 5 in descending order for various sets of features x_i used as input into Equation 1. The conditions for when barge-in rates are used also show the window width w for the highest classification accuracy. The mean average error (MAE) is also listed, which is the average of the differences between an output of the logistic regression function X_j and a reference label manually given $\hat{X}_j$ (0 or 1):

$$MAE = \frac{1}{m}\sum_{j}^{m} |\hat{X}_j - X_j|, \qquad (3)$$

where m denotes the total number of barge-in utterances. This indicates how well the output of the logistic regression function (Equation 1) distributes. Regarding Condition (12) in Table 5 (majority baseline), the MAE was calculated by assuming $X_j = 0.456$, which is the average ASR accuracy, for all j. Its classification accuracy is the majority baseline; that is, all interpretation results are regarded as incorrect.

4.4 Experimental Results

The results are shown in Table 5. First, we can see that the classification accuracies for Conditions (1) to (6) are high because the ASR confidence measure (CM) works well (Table 3). The MAEs are also small, which means the outputs of the logistic regression functions are good indicators of the reliability of the interpretation result.

Upon comparing Condition (6) with Conditions (1) to (5), we can see that the classification accuracies improve as a result of incorporating the user's utterance histories such as barge-in rates and ASR accuracies. Table 6 lists p-values of the differences when the barge-in rate and the estimated ASR accuracy were used in addition to the CM. The significance test was based on the McNemar test. As shown in the table, all the differences were statistically significant ($p < 0.01$). That is, it was experimentally shown that these utterance histories of users are different information sources from those of single utterances and that they contribute to improving the classification accuracy even when used together with ASR confidence measures. The relative improvement in the error reduction rate was 15.2% between Conditions (2) and (6), that is, by adding the barge-in rate and the estimated ASR accuracy, both of which can be obtained without manual labeling.

Table 5: Best classification accuracy for each condition and optimal window width

Conditions (features used)	Window width	Classification accuracy (%)	MAE
(1) CM + barge-in rate + actual ASR acc.	w=40	92.6	0.112
(2) CM + barge-in rate + estimated ASR acc	w=30	92.2	0.119
(3) CM + actual ASR acc.	-	91.7	0.121
(4) CM + barge-in rate	w=30	91.6	0.126
(5) CM + estimated ASR acc.	-	91.2	0.128
(6) CM	-	90.8	0.134
(7) barge-in rate + actual ASR acc.	w=50	80.0	0.312
(8) barge-in rate + estimated ASR acc.	w=50	77.7	0.338
(9) actual ASR acc.	-	72.8	0.402
(10) barge-in rate	w=30	71.8	0.404
(11) estimated ASR acc.	-	57.6	0.431
(12) majority baseline	-	54.4	0.496

CM: confidence measure
MAE: mean absolute error

Table 6: Results of significance test

Condition pair	p-value
(2) vs (4)	0.00066
(2) vs (5)	0.00003
(4) vs (6)	0.00017
(5) vs (6)	0.00876

Figure 4 shows the results in more detail; the classification accuracies for Conditions (1), (2), (4), and (6) are shown for various window widths. Under Condition (6), the classification accuracy does not depend on the window width because the barge-in rate is not used. Under Conditions (1), (2), and (4), the accuracies depend on the window width for the barge-in rate and are highest when the width is 30 or 40. These results show the effectiveness of the window, which indicates that temporal changes in user behaviors should be taken into consideration, and match those of our earlier reports (Komatani et al., 2008; Komatani and Rudnicky, 2009): the user's utterance history becomes effective after he/she uses the system about ten times because the average number of utterances per dialogue is around five.

By comparing Conditions (2) and (4), we can see that the classification accuracy improves after adding the estimated ASR accuracy to Condition (4). This shows that the estimated ASR accuracy also contributes to improving the classification accuracy. By comparing Conditions (1) and (2), we can see that Condition (1), in which the actual ASR accuracy is used, outperforms Condition (2), in which the estimated one is used. This suggests that the classification accuracy, whose upper limit is Condition (1), can be improved by making the ASR accuracy estimation shown in Section 3.2 more accurate.

5 Conclusion

We described a method of detecting interpretation errors of barge-in utterances by exploiting the utterance histories of individual users, such as their barge-in rate and ASR accuracy. The estimated ASR accuracy as well as the barge-in rate and the ASR confidence measure is obtainable online. Thus, the detection method does not require manual labeling. We showed through experiments that the utterance history of each user is helpful for detecting interpretation errors even when the ASR confidence measure is used.

The proposed method is effective in systems that are repeatedly used by the same user over 10 times, as indicated by the results of Figure 4. It is also assumed that the user's ID is known (we used their telephone number). The part of our method that estimates the user's ASR accuracy assumes that the system's dialogue strategy is to make explicit confirmations about every utterance by the user and that all affirmative and negative responses followed by explicit confirmations are correctly recognized. Our future work will attempt to reduce or remove these assumptions and to enhance the generality of our method. The experimental

result was shown only in the Kyoto City Bus domain, in which dialogues were rather well structured. Experimental evaluations in other domains will assure the generality.

Acknowledgments

We are grateful to Prof. Tatsuya Kawahara of Kyoto University who led the Kyoto City Bus Information System project. The evaluation data used in this study was collected during the project. This research was partly supported by Grants-in-Aid for Scientific Research (KAKENHI).

References

Dan Bohus and Alexander Rudnicky. 2007. Implicitly-supervised learning in spoken language interfaces: an application to the confidence annotation problem. In *Proc. SIGdial Workshop on Discourse and Dialogue*, pages 256–264.

Timothy J. Hazen, Theresa Burianek, Joseph Polifroni, and Stephanie Seneff. 2000. Integrating recognition confidence scoring with language understanding and dialogue modeling. In *Proc. Int'l Conf. Spoken Language Processing (ICSLP)*, pages 1042–1045, Beijing, China.

Kazunori Komatani and Tatsuya Kawahara. 2000. Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output. In *Proc. Int'l Conf. Computational Linguistics (COLING)*, pages 467–473.

Kazunori Komatani and Alexander I. Rudnicky. 2009. Predicting barge-in utterance errors by using impricitly-supervised asr accuracy and barge-in rate per user. In *Proc. ACL-IJCNLP*, pages 89–92.

Kazunori Komatani, Shinichi Ueno, Tatsuya Kawahara, and Hiroshi G. Okuno. 2005. User modeling in spoken dialogue systems to generate flexible guidance. *User Modeling and User-Adapted Interaction*, 15(1):169–183.

Kazunori Komatani, Tatuya Kawahara, and Hiroshi G. Okuno. 2007. Analyzing temporal transition of real user's behaviors in a spoken dialogue system. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 142–145.

Kazunori Komatani, Tatuya Kawahara, and Hiroshi G. Okuno. 2008. Predicting asr errors by exploiting barge-in rate of individual users for spoken dialogue systems. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 183–186.

Akinobu Lee, Keisuke Nakamura, Ryuichi Nisimura, Hiroshi Saruwatari, and Kiyohiro Shikano. 2004. Noice robust real world spoken dialogue system using GMM based rejection of unintended inputs. In *Proc. Int'l Conf. Spoken Language Processing (ICSLP)*, pages 173–176.

Gina-Anne Levow. 2003. Learning to speak to a spoken language system: Vocabulary convergence in novice users. In *Proc. 4th SIGdial Workshop on Discourse and Dialogue*, pages 149–153.

Diane J. Litman, Marilyn A. Walker, and Michael S. Kearns. 1999. Automatic detection of poor speech recognition at the dialogue level. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 309–316.

Antoine Raux, Dan Bohus, Brian Langner, Alan W. Black, and Maxine Eskenazi. 2006. Doing research on a deployed spoken dialogue system: One year of Let's Go! experience. In *Proc. Int'l Conf. Spoken Language Processing (INTERSPEECH)*.

Richard C. Rose and Hong Kook Kim. 2003. A hybrid barge-in procedure for more reliable turn-taking in human-machine dialog systems. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 198–203.

Katsuhito Sudoh and Mikio Nakano. 2005. Post-dialogue confidence scoring for unsupervised statistical language model training. *Speech Communication*, 45:387–400.

Marilyn Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane Litman. 2000. Learning to predict problematic situations in a spoken dialogue system: Experiments with How May I Help You? In *Proc. North American Chapter of Association for Computational Linguistics (NAACL)*, pages 210–217.