Systems and Computers in Japan, Vol. 38, No. 9, 2007
Translated from Denshi Joho Tsushin Gakkai Ronbunshi, Vol. J89-D, No. 7, July 2006, pp. 1493–1503

Spoken Language Understanding Method Using Confidence Measure and Dialogue History

Noriki Fujiwara,[1] Toshihiko Itoh,[1] Kenji Araki,[1] Atsuhiko Kai,[2] Tatsuhiro Konishi,[3] and Yukihiro Itoh[3]

[1]Graduate School of Information Science and Technology, Hokkaido University, Sapporo, 060-0814 Japan

[2]Faculty of Engineering, Shizuoka University, Hamamatsu, 432-8561 Japan

[3]Faculty of Informatics, Shizuoka University, Hamamatsu, 432-8011 Japan

SUMMARY

In the real environment, it is hard for a speech recognizer to avoid misrecognitions completely. However, if misrecognitions occur, user's intentions are usually misunderstood by a conventional language understanding technique, which simply gives priority to the higher rank hypothesis of a speech recognition result (N-best). The utterances in a dialogue are coherent and correct user's intentions might appear in the lower rank hypothesis of N-best. To understand user's speech intentions in the real environment, we propose the language understanding technique that utilizes the dialogue context and confidence measure, which is the word posterior probability. The experimental results show that proposed technique is more efficient (about 15%) than the conventional technique. © 2007 Wiley Periodicals, Inc. Syst Comp Jpn, 38(9): 21–31, 2007; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/scj.20758

Key words: speech language understanding; speech dialogue system; speech recognition error; confidence measure; dialogue history.

1. Introduction

Spoken dialogue systems have recently attracted attention as practical applications [1–4]. Speech interface has some advantages, which are eyes-free and hands-free functions, and do not need any extra practice. However, they have some problems in dealing with spontaneous speech. An example is speech misrecognition, which prevents smooth dialogue and leads to repetitive correcting input, which upsets users.

There are many studies which aim to reduce the misrecognition. Some of them research utilizing directional microphones for reducing surrounding noise, and learning of a speech recognizer with speech data overlapped noise [5]. Moreover, many spoken dialogue systems utilize confidence measures as criteria which indicate reliability for a recognition result [6–8]. However, the current technology can reduce speech misrecognition but cannot avoid it completely.

We consider that a spoken dialogue system needs to take into account misrecognition in a real environment, because speech recognizers have a difficult time completely recognizing speech due to the surrounding noises and speech ambiguity. If the systems have misrecognized or misunderstood, they have to detect and modify the errors as soon as possible. Otherwise they cannot create a smooth dialogue and then the degree of user satisfaction drops.

Some studies aim to detect a system's misrecognitions or misunderstanding. Some utilize acoustic features for user utterances [9, 10]. Other studies detect the repetition of user utterances [11] and the misunderstanding from the question–answer dialogue [12]. Other studies aim to understand a user's intentions correctly with more information. They utilize the dialogue context for the language understanding [13–17] because a user's utterances in a dialogue are coherent. Still other studies aim to increase the degree of user satisfaction. In Ref. 18, the dialogue manager does not include doubtful information in system responses, because the information which might be misunderstood decreases the degree of user satisfaction.

We aim to develop a dialogue system that creates a smooth dialogue with a high degree of user satisfaction. For that purpose, our research utilizes the following concepts.

- When user's utterance is not recognized correctly, the correct sentence or a part of the correct sentence is included in N-best in a lot of cases [19].
- User almost always corrects misrecognition immediately [20].
- User's utterances in a task-oriented dialogue are coherent and have semantic relationship [21].

We propose a language understanding method which can estimate correct user's intention even if user's utterance is not recognized correctly. In the proposed method, we give each keyword the criterion which is based on the confidence measure and the context in the dialogue. The criteria are updated whenever a new recognition result is received. In this way, our proposed method estimates user's intention in consideration of the context.

In many of the studies mentioned above, the systems rescore a speech recognition result N-best to estimate user's intention. Therefore, they never estimate correct user's intention when correct words are not included in N-best. In our proposed method, the system has all keywords that are likely to appear in recognition results and considers semantic relationship of each keyword. Therefore, the system can output keywords not included in N-best.

In this paper, we propose a language understanding method using confidence measure and dialogue history, which includes all recognition results and system responses, for estimating user's intention even if user's utterance is recognized correctly. Moreover we describe experimental results with the proposed method.

2. System Task

2.1. Dialogue task

In this paper, we selected a landmark setting in a car navigation system as a dialogue task. When drivers use a car navigation system, they should not use their hands or eyes to direct the system while driving a car. Therefore, speech input is much safer than handling a menu on the display or a remote controller. In addition, engine noise and voices of fellow passengers create more misrecognition.

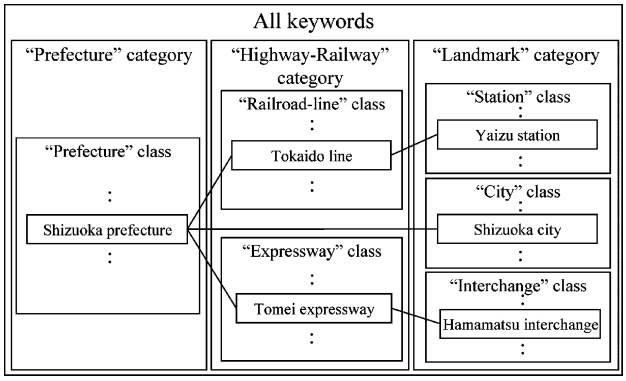

Fig. 1. Relation of categories, classes, and keywords.

Table 1. Speech type

Speech types	Expositions and dialogue examples
Narrowing	The utterance for adding new information to system response. U1: In Shizuoka prefecture. S1: Shizuoka prefecture. U2: Hamamatsunishi interchange of Tomei expressway.
Correction	The utterance for correcting the system response. U1: Hamamatsunishi interchange of Tomei expressway. S1: Are you going to the Hamamatsu interchange of Tomei expressway? U2: No, Hamamatsunishi interchange.
Answer	The utterance for answering system question. U1: I'm going to Hamamatsunishi interchange of Tomei expressway. S1: Which interchange of Tomei expressway are you going to? U2: Hamamatsunishi interchange.
Re-input	The utterance after system re-input request. U1: I'm going to the Hamamatsunishi interchange. S1: Please say that once again. U2: Hamamatsunishi interchange.

U: user utterance, S: system response

Users must input by voice landmark(s) name(s) along driving routes. The landmarks names include the names of interchanges, train stations, and cities. The users input landmarks names (Landmarks) to the system. Users can supplement Landmarks with prefectures (Prefectures) and routes (Highway-Railways). Highway-Railways include the names of railroad lines and expressways. In this paper, Prefectures, Highway-Railways, and Landmarks are called "categories." The subcategories included in each category are called "classes" (interchanges, train stations, railroad lines, and so on). The relation of categories and classes is shown in Fig. 1. The maximum number of keywords which a user can input is three, for example, "Shizuoka prefecture," "Tomei expressway," and "Hamamatsu interchange." The user can input three keywords at once or can input them separately. The following utterance patterns were used.

2.2. Speech type

In our task, all user's utterances are classified into four speech types, which are narrowing, correction, answer, and re-input. In narrowing, the user inputs new or additional words which narrow the relationship with previous input. In correction, the user corrects the system response. In answer, the user answers a question from the system. In re-input, the user re-inputs in formation in accordance with a request from the system. Examples of each speech type are shown in Table 1.

3. System Architecture

An outline of our spoken language system is shown in Fig. 2. The system consists of a speech recognizer, a generator of the confidence measure, a language understanding component, a response generator, a speech synthesizer, and a GUI component [22].

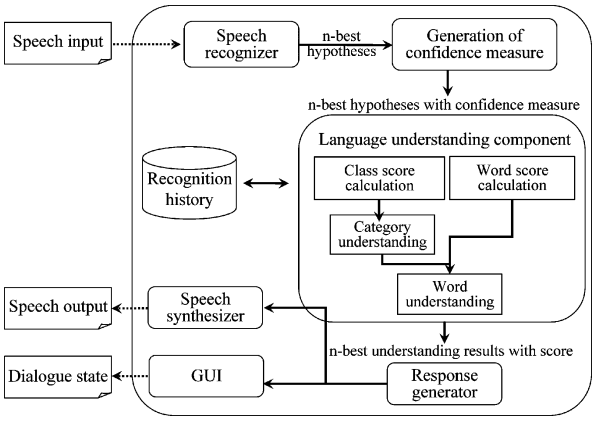

Fig. 2. Outline of speech dialogue system.

The speech recognizer outputs the n-best hypotheses ordered according to their acoustic probability. In this paper, we use SPOJUS [23] as the speech recognizer.

The generator of the confidence measure outputs the confidence measure of a word w for a given speech x is estimated as a posteriori probability $P(w|x)$, which is calculated from the likelihood scores of the n-best sentence hypotheses, $P(x|w)$ [6]. By the same process, the confidence measure of a class c is also estimated.

For the language understanding component, the first step is to calculate word scores and class scores, which are criteria based on the confidence measures and the recognition history. The second step is to calculate the category scores, which are criteria based on class scores, and furthermore to predict categories which the user uttered using the criteria. This process is termed "category understanding." For example, if the user spoke keywords of a prefecture and an interchange, the correct prediction in this process is that Prefecture category and Landmark category are uttered. The final step is to determine the word sequence that has the highest score of all the word sequences of predicted category combinations. The detailed process is described in the next section.

The response generator outputs system response for a confirmation, a question, or a request for next user utterance utilizing word scores, class scores, and the language understanding results. This process generates optimal sentences based on various dialogue strategies. For example, when the language understanding component outputs certain understanding result which is the first ranking candidate in the system belief but whose score is low, the system pretends to have understood and requests next user utterance. By this strategy the system does not output the information which might be misunderstood and can draw out more information. This strategy is the same idea as in Ref. 18, where asking casually or requesting other information causes a higher degree of user satisfaction than does frequently requesting for the next user utterance or confirming system understanding.

A speech synthesizer conveys system responses by synthesized speech, and a GUI component displays current states of system understanding.

4. Language Understanding Component

This section describes the detailed process in the language understanding component. The first step is to calculate word scores and class scores, which are criteria based on the confidence measures and the recognition history. The second step is to calculate the category scores and category understanding is performed on the scores. The

final step is to determine the word sequence that has the highest score of all the word sequences of predicted category combinations. These word scores and class scores are utilized by not only language understanding but also system response generating.

When a user adds new information, the system has to add it into current system belief. On the other hand, when the user corrects the system response, the system has to modify current system belief. In our proposed method, those processes are realized by the addition and the subtraction of word scores and class scores. Calculated word scores and class scores are saved as integrated recognition results based on the history.

4.1. Class score calculation

In class score calculation, whenever a new recognition result is input (turn t), the system judges speech types (Table 2) with the recognition history and the latest recognition result. According to the judgment, all classes are calculated by each equation.

Class score calculation is performed by two different patterns. In one, the speech types are narrowing or answer. When these utterance patterns are uttered, the system has to add new information to current system belief. In the other, the speech types are correction or re-input. When these utterance patterns are uttered, the system has to modify current system belief. Utterance pattern judgment is shown in Table 2. This judgment is based on our heuristics, therefore there is a possibility that better judgment exists.

4.1.1. Class score calculation for narrowing or answer

When the user utters a narrowing utterance or a positive answer, the system can assume that the previous recognition is successful. Therefore, the class scores related to the current recognition results are added to the confidence measures of that class. The class score of class c is calculated according to the following equation. In this paper, class scores of all classes picked up from the word dictionary of the speech recognizer are initially set to 0, and do not have the upper and lower limit.

$$Score_t(c) = Score_{t-1}(c) * weight_{na} + Conf_t(c) \quad (1)$$

$Score$: class scores of recognition history
$Conf_t$: confidence measures of the latest
$weight_{na}$: coefficient ($0.0 < weight_{na} < 1.0$) recognition results
c: class to be processed

For considering previous user utterances, the class score in the latest recognition result is added to the class score in the recognition history. The class score in the recognition history is reduced with the coefficient $weight_{na}$ by the concept that "as information becomes older, its reliability decreases." This coefficient $weight_{na}$ is optimized by the dialogue data described in Section 5. Calculated class scores are saved as integrated recognition results based on the history.

Table 2. Utterance pattern judgment

Condition	Judgment
utterance after re-input request	correction or re-input
negation involved in recognition result	correction or re-input
new category introduced into the dialogue	narrowing or answer
otherwise	correction or re-input

4.1.2. Class score calculation for correction or re-input

The calculation for the correction and re-input utterances is fundamentally the same as that for the narrowing and answer utterances, except that the confidence measures of the different class in the same category are reduced from the whole scores. This reduction contributes to recover from misunderstanding.

$$Score_t(c_a) = Score_{t-1}(c_a) * weight_{cr} - Conf_t(c_b) + Conf_t(c_a) \quad (2)$$

$Score_t$: class scores of recognition history
$Conf_t$: confidence measures of the latest recognition results
$weight_{cr}$: coefficient ($0.0 < weight_{cr} < 1.0$)
c_a: class to be processed
c_b: class different from c_a in the same category

4.2. Category understanding

A category understanding process is performed to understand user utterances roughly at category level. An example of a category understanding process is shown in Fig. 3. The category understanding process calculates category scores from the class scores of the current recognition results and the recognition history. Each category score is the sum of all class scores belonging to each category, and is judged by each threshold. If the score is over the threshold, the judgment result is "1." Otherwise it is "0." The logical addition of the judgment result for each category is the category understanding result.

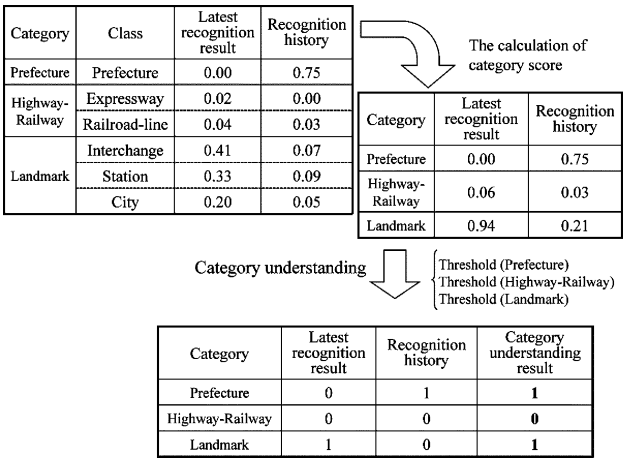

Fig. 3. Category understanding process.

4.3. Word score calculation

The word scores are calculated using confidence measures of the latest recognition results and previous word scores of the recognition history. Word scores of all keywords picked up from the word dictionary of the speech recognizer are initially set to 0, and changes are calculated by the following 10 strategies whenever a new recognition result is input. By preparing word scores for all keywords and using the following strategies (e.g., using the relationship of each word in the latest recognition results and recognition history), a word that does not exist in recognition results can be included in the understanding result. In this paper, we do not set up upper and lower limits for word scores.

Strategy (1): Strategy (1) is based on the concept that if information becomes older, its reliability decreases. Whenever a new recognition result is received, all word scores in the history are lowered. The word score of "w_A" from the recognition history is calculated according to the following equation:

$$Score_t(w_A) = Score_{t-1}(w_A) - weight_1 \quad (3)$$

$Score_t$: word scores of recognition history
$weight_1$: coefficient ($0.0 < weight_1 < 1.0$)
w_A: word A

Strategy (2): When a word "w_A" in the history and a word "w_B" in the latest results have a semantic relation (in this paper, local relation: e.g., they belong to the same prefecture), the word score of "w_A" increases according to the equation

$$Score_t(w_A) = Score_{t-1}(w_A) + weight_2 * Conf_t(w_B) \quad (4)$$

$Score_t$: word scores of recognition history
$Conf_t$: confidence measures of the latest recognition results
$weight_2$: coefficient ($0.0 < weight_2 < 1.0$)
w_A: word A
w_B: word B

In this strategy (2), the strength of the semantic relation is different, whether a word "w_A" and a word "w_B" are the same word or not. Therefore, when a word "w_A" and a word "w_B" are the same words, the coefficient $weight_2$ is replaced by another coefficient $weight_2'$.

Strategy (3): When a word "w_A" in the history and a word "w_B" in the latest results do not have any semantic relation (local relation), the word score of "w_A" decreases according to the equation

$$Score_t(w_A) = Score_{t-1}(w_A) + weight_3 * Conf_t(w_B) \quad (5)$$

$weight_3$: coefficient ($0.0 < weight_3 < 1.0$)

Strategy (4): When an affirmation word "w_{yes}" (e.g., "yes") is contained in the latest recognition results, the word score of "w_A" contained in the previous system increases according to the equation

$$Score_t(w_A) = Score_{t-1}(w_A) + weight_4 * Conf_t(w_{yes}) \quad (6)$$

$weight_4$: coefficient ($0.0 < weight_4 < 1.0$)
w_{yes}: an affirmation word

Strategy (5): When a negation word "w_{no}" (e.g., "no") is contained in the latest recognition results, the word score of "w_A" contained in the previous system response decreases according to the equation

$$Score_t(w_A) = Score_{t-1}(w_A) - weight_5 * Conf_t(w_{no}) \quad (7)$$

$weight_5$: coefficient ($0.0 < weight_5 < 1.0$)
w_{no}: a negation word

Strategy (6): When a word "w_A" in the latest recognition results and a word "w_B" in the system response have a semantic relation (local relation), the score of the word "w_B" increases according to the equation

$$Score_t(w_B) = Score_{t-1}(w_B) + weight_6 * Conf_t(w_A) \quad (8)$$

$weight_6$: coefficient ($0.0 < weight_6 < 1.0$)

Strategy (7): When a word "w_A" in the latest recognition results and a word "w_B" in the system response do not have a semantic relation (local relation), the word score of "w_B" decreases according to the equation

$$Score_t(w_B) = Score_{t-1}(w_B) + weight_7 * Conf_t(w_A) \quad (9)$$

$weight_7$: coefficient ($0.0 < weight_7 < 1.0$)

Strategy (8): If the system asks and the subsequent user response contains possible answers for the question, the word score of "w_A" in the possible answers increases. For example, the system asks "Which interchange?" and the user responds "Hamamatsu interchange." The word score

of "Hamamatsu interchange" increases according to the equation

$$Score_t(w_A) = Score_{t-1}(w_A) + weight_8 * Conf_t(w_A) \quad (10)$$

$weight_8$: coefficient ($0.0 < weight_8 < 1.0$)

Strategy (9): As the ranking in n-best hypotheses increases, the words possess higher reliability. Therefore, the bonus scores based on the ranking are given to words ranked higher in the recognition results (n-best hypotheses).

$$Score_t(w_A) = Score_{t-1}(w_A) + weight_9 * Conf_t(w_A) \quad (11)$$

$weight_9$: coefficient ($0.0 < weight_9 < 1.0$)

The coefficient "$weight_9$" takes different value for ranking in recognition result.

Strategy (10): The speech recognizer in our system has the behavior that longer utterances are easier to recognize correctly. In other words, user's utterance of only one category is easier to recognize correctly than that of two or three categories. For this reason, the bonus scores are given based on the length of recognized word sequences. The sentence including "w_A" in a recognition result is longer, and the bonus scores thus become larger.

$$Score_t(w_A) = Score_{t-1}(w_A) + weight_{10} * Conf_t(w_A) \quad (12)$$

$weight_{10}$: coefficient ($0.0 < weight_{10} < 1.0$)

The coefficient "$weight_{10}$" takes different value for the sentence length including "w_A" in a recognition result.

The coefficients in each strategy are optimized by training speech data mentioned in Section 5.1.

Word scores are calculated using the above strategies. The calculated word scores are updated and saved as new recognition history. These processes can increase the word scores of the keywords not in the recognition results using the relationship to the keywords in the recognition results. The repetitive calculation of word scores with new recognition results makes the scores of reliable words increase and the scores of unreliable words decrease. As a result, regardless of the ranking of the present recognition results (n-best hypotheses), priority is given to words with the highest possibility to appear in a given dialogue.

4.4. Word understanding

Word understanding is used to search most likely word sequence in the dialogue for a certain goal (in this paper, one landmark setting). An example of the word understanding process is shown in Fig. 4. In this process, the system selects the word combination which has the maximum sum of word scores with category understanding result. In Fig. 4, category understanding result is that Prefecture and Landmark category are uttered. Therefore, the sum of word scores belonging to Prefecture and Landmark category is calculated. In this process the system utilizes semantic restriction of words. In other words, this process never outputs word sequences that do not exist.

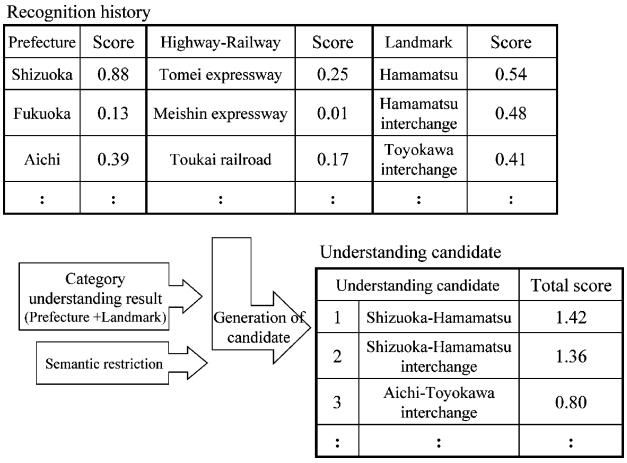

Fig. 4. Word understanding process.

5. Experiment

5.1. Speech data

Our system uses many weights and thresholds in a language understanding component. Many utterance patterns are needed to optimize these parameters. Therefore, we collect utterances in the following way. First, "U1-S1-U2" pattern is assumed as a dialogue and it is not considered whether to have completed the landmark setting. Second, we record various possible utterance patterns by a sentence when a user sets a landmark name. These utterances collected in this way are called "speech data." We input a part of speech data which are appropriate for U1 to our system and get system responses (S1). Then we select utterances of speech data which are appropriate utterances following S1. In this way we construct U1-S1-U2 dialogue, which is called "imitation dialogue."

We collected speech data for training our system. First, we prepared 21 sentences (Table 3) about the Hamamatsu-nishi interchange, which is one of the interchanges along the Tomei expressway in Shizuoka prefecture, Japan. We asked five speakers to record each prepared sentence three times in a sound booth. In consideration of the real environment, car noise was mixed in with the speech, and the recognition rate was lowered. We collected 315 sentences. Their recognition rate is 56.8%. In this

Table 3. Speech pattern

1	Shizuoka ken (Shizuoka prefecture.)
2	Tomei jidousyadou (Tomei expressway.)
3	Hamamatsu-nishi intah (Hamamatsu-nishi interchange.)
4	Shizuoka ken no Hamamatsu-nishi intah (Hamamatsu-nishi interchange in Shizuoka prefecture.)
5	Shizuoka ken no Tomei jidousyadou (Tomei expressway in Shizuoka prefecture.)
6	Tomei jidousyadou no Hamamatsu-nishi intah (Hamamatsu-nishi interchange of Tomei expressway.)
7	Shizuoka ken no Tomei jidousyadou no Hamamatsu-nishi intah (Hamamatsu-nishi interchange of Tomei expressway in Shizuoka prefecture.)
8	Hai, Shizuoka ken (Yes, Shizuoka prefecture.)
9	Hai, Tomei jidousyadou (Yes, Tomei expressway.)
10	Hai, Hamamatsu-nishi intah (Yes, Hamamatsu-nishi interchange.)
11	Hai, Shizuoka ken no Hamamatsu-nishi intah (Yes, Hamamatsu-nishi interchange in Shizuoka prefecture.)
12	Hai, Shizuoka ken no Tomei jidousyadou (Yes, Tomei expressway in Shizuoka prefecture.)
13	Hai, Tomei jidousyadou no Hamamatsu-nishi intah (Yes, Hamamatsu-nishi interchange of Tomei expressway.)
14	Hai, Shizuoka ken no Tomei jidousyadou no Hamamatsu-nishi intah (Yes, Hamamatsu-nishi interchange of Tomei expressway in Shizuoka prefecture.)
15	Iie, Shizuoka ken (No, Shizuoka prefecture.)
16	Iie, Tomei jidousyadou (No, Tomei expressway.)
17	Iie, Hamamatsu-nishi intah (No, Hamamatsu-nishi interchange.)
18	Iie, Shizuoka ken no Hamamatsu-nishi intah (No, Hamamatsu-nishi interchange in Shizuoka prefecture.)
19	Iie, Shizuoka ken no Tomei jidousyadou (No, Tomei expressway in Shizuoka prefecture.)
20	Iie, Tomei jidousyadou no Hamamatsu-nishi intah (No, Hamamatsu-nishi interchange of Tomei expressway.)
21	Iie, Shizuoka ken no Tomei jidousyadou no Hamamatsu-nishi intah (No, Hamamatsu-nishi interchange of Tomei expressway in Shizuoka prefecture.)

paper, the maximum number of candidates output by a speech recognizer is 100, in other words N-best is 100-best. Note that when the acoustic likelihood for a candidate is lower than the thresholds, the speech recognizer never outputs the candidate. Therefore, the speech recognizer does not always output 100 candidates. The average number of candidates for training speech data mentioned above is about 30. Moreover, we investigated how "Shizuoka," "Tomei," and "Hamamatsu-nishi" are misrecognized as other words. As a result, the misrecognition from "Shizuoka" to "Mie" (Mie Prefecture) is the strongest tendency. The number of misrecognitions for 180 utterances including "Shizuoka" is 73 utterances. The rate of misrecognition from "Shizuoka" to "Mie" in 73 utterances is 20.5% (15/73). As a training set, we made 7530 imitation dialogues from the training speech data. Using this training set, we optimized the weights and thresholds in the language understanding component. Here we define two criteria for the evaluation. They are called *dialogue accuracy* and *word accuracy*. The dialogue accuracy denotes the rate which the keywords output by the system correspond to the keywords in user's utterance for all categories (Prefecture, Highway-Railway, and Landmark category). For example, when a user utters w_A belonging to Prefecture category and w_B belonging to Landmark category and the system outputs w_A and w_B, the dialogue accuracy is 100.0% (1/1). In this case, the keywords output by the system correspond to the keywords in user's utterance in just proportion. If the system outputs only w_A, or w_A, w_A, and w_A belonging to Highway-Railway category, the dialogue accuracy is 0.0% (0/1). The word accuracy denotes the rate which the keyword output by the system corresponds to the keyword in user's utterance at a word level. For example, when user utters w_A belonging to Prefecture category and w_B belonging to Landmark category and the system outputs w_A, w_B, and w_C belonging to Highway-Railway category, the word accuracy is 66.7% (2/3). The parameters are optimized so that the dialogue accuracy might increase the most. The dialogue accuracy with the optimized parameters is 71.5% (5386/7530 dialogues) and the word accuracy is 87.0% (19,655/22,590 words). Moreover, we investigate the rate at which the categories estimated by the system correspond to the categories uttered by user at category level, which is called *category understanding accuracy*. The category understanding accuracy with the optimized parameters is 74.7% (5632/7530 dialogues). The denominator of the word accuracy is three times the denominator of the dialogue accuracy, because the system has to estimate three keywords by a dialogue. Even if user's utterance includes only two categories, the system has to specify the category being uttered and the category not being uttered.

In the same way as for training speech data and training imitation dialogues, we collected speech data and made imitation dialogue for testing. We asked 10 speakers

to record each prepared sentence three times in a sound booth, and mixed car noise in with the speech. We collected 630 sentences. Their recognition rate is 67.7%. We made 29,670 imitation dialogues; the number of dialogues which do not include correct words is 3305 dialogues.

5.2. Experiential method

For our evaluation, we prepared three spoken-language understanding systems: SYS-A, SYS-B, and SYS-C. SYS-A gives top priority to the highest-rank candidate of the latest recognition results (N-best hypotheses), searches the recognition history for words related to the candidate, and outputs an understanding result. SYS-B uses the language understanding method mentioned in this paper. SYS-C is identical to SYS-B except that it uses the correct category understanding result. In SYS-B, if a category understanding fails, the system cannot output a correct understanding result. To investigate the performance of word score calculation independently, we prepared SYS-C. These systems have the same performance except the language understanding component.

5.3. Results and discussion

The dialogue accuracies of the systems, SYS-A, SYS-B, and SYS-C, were 57.9, 72.2, and 89.2%, respectively. SYS-B was about 15 points more accurate than SYS-A. These results mean that our system is effective. SYS-C is about 17 points more accurate than SYS-B. Therefore, the understanding performance can be increased by improving the accuracy of the category understanding. The category understanding accuracy of SYS-B was 78.8% (23,408/29,670 dialogues).

We separated a test set by four speech types. The dialogue accuracy of each system is shown in Fig. 5. In this figure, "U1:OK" means the first user's utterance (U1) is recognized correctly. "U1:NG" means that U1 is recognized incorrectly. "SB:OK" means that the system belief corresponds to keywords included in user's utterances, which are U1 and U2. The numerical value above the graph means the number of dialogues of each speech type, and the numerical value in the graph means the understanding rate. This figure shows that the understanding rate is higher in order of SYS-C, SYS-B, and SYS-A except correction dialogues and a part of re-input dialogues. This result means that the proposed method is effective except correction dialogues even if user's utterances are misrecognized at any turns.

The word accuracies of SYS-A, SYS-B, and SYS-C were 75.4, 87.1, and 95.5%, respectively. This means that our system is effective at word level similarly. Figure 6 shows that the understanding rate is higher in order of SYS-C, SYS-B, and SYS-A except correction dialogues and a part of re-input dialogues. In Fig. 6, the numerical value above the graph means the number of words of each speech type. The word accuracies of SYS-B and SYS-C are better than their dialogue accuracies. This means that the systems can understand user's intention partly even if they cannot understand completely. The systems could understand completely by continuing the dialogue. This evaluation demonstrated our proposed method is effective for spoken language understanding.

In Figs. 5 and 6, the correction dialogue accuracies of the proposed method are lower. The reason for this is that the number of correction dialogues in a training set is less than other speech types and our system learned the tendency which was not robust for correction dialogues. At the same time, the response generator in our system is based on the concept that the system response including misunderstanding information decreases the degree of user satisfaction. Therefore, the keywords with lower reliability are not included in system responses, and the training set has less correction dialogues. In addition, the speech data used for training and test are recorded and then mixed in car noise, therefore it is not necessarily the case that the speech data correspond to utterances in a real environment. For these reasons, we have to record utterances of a speaker who actually drives a car and to evaluate with the speech data.

We investigated the effects of each strategy in the word score calculation. We compared the largeness of optimized $weight_n$ (n = 1, . . . , 10) and the number of applications in each strategy. We found that the weights in strategies (1), (2), and (3) are smaller than the others, but they are applied very frequently. Therefore, those strategies have leverage over the word score calculation. The weights in strategies (4), (7), and (8) are sometimes applied, but they

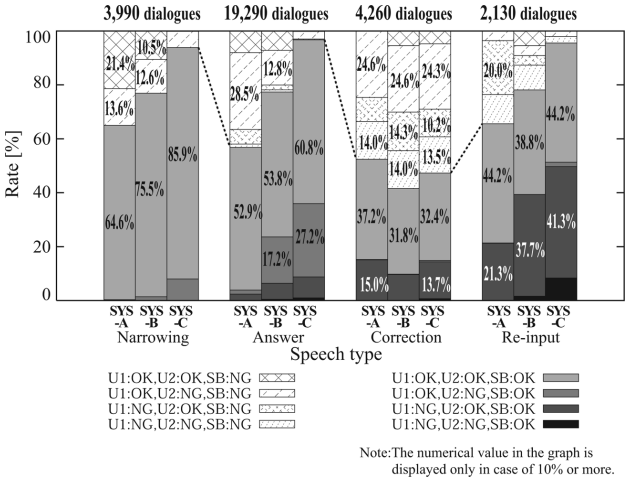

Fig. 5. Understanding results (dialogue accuracy).

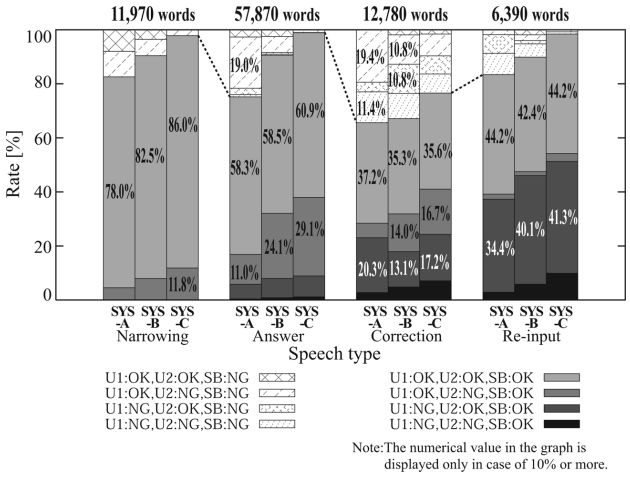

Fig. 6. Understanding results (word accuracy).

are larger than the others, therefore those strategies also have leverage over the word score calculation. As to the other strategies not mentioned above, our system often understood correctly by the mutual influence of some strategies included in the other strategies. In addition, the weights get larger or smaller by training speech data or tasks. Therefore, we do not consider that the strategies which have less effect in this investigation are needless. It is a future work to investigate qualitative effects of each strategy.

As to the domain which can be used in the proposed method, it is hard to use for any spoken dialogue tasks. However, the keywords almost have hierarchic and dependency relations as a tree structure in the slot-filling tasks or information inputting tasks, which are the mainstream of many spoken dialogue systems of the present time. In these tasks, the proposed method is effective and can be used widely. Some tasks might require some modifications and additions regarding the word score calculation.

6. Conclusion

In this paper, we proposed a method for improving spoken language understanding in car navigation systems using confidence measures and a dialogue history. Experimental results showed the understanding accuracy of our method is more than 15% higher than the language understanding method that gives top priority to the higher-rank candidate of the n-best hypotheses. Even when misunderstanding has occurred, the system understands the user's intention partially. The system could understand completely by continuing the dialogue. It is also possible to improve the understanding rate by improving the category understanding accuracy.

Future work should consider speech data. Because the speech data used for a training and test are recorded and then mixed in car noise, it is not necessarily the case that the speech data correspond to the utterances in a real environment. We have to record utterances of a speaker who actually drives a car. Other future work concerns the length of a dialogue. Because the dialogues in this experiment are U1-S1-U2, landmark setting in some dialogues are finished and in others are on the way. Thus, we have to evaluate our system in longer dialogues. Next, it is necessary to investigate the understanding rate and the degree of user satisfaction [24] when the system is used in conditions similar to an actual driving situation (by using a drive simulator and so on). Moreover, we have to investigate qualitative effects of each strategy in word score calculation, because the effects of the strategies are changed by training speech data or tasks. Finally, we would extend the system to be able to handle dialogues in complex situations. The situations are that user searches the landmark and then sets, or does not know exact landmark names.

REFERENCES

1. Hara S, Shirose A, Miyajima C, Itou K, Takeda K. A music searching system by spoken dialogue. IPSJ SIG Technical Reports, SLP-53, p 31–36, 2004. (in Japanese)
2. Watanabe Y, Sekiguchi Y, Suzuki Y. On the speech dialogue function of household electric appliances goods as an example of video tape recoder. IPSJ J 2003;44:2690–2698. (in Japanese)
3. Kawaguchi N, Ushikubo S, Matsubara S, Iwa H, Kajita S, Takeda K. In-car spoken dialogue rescoring system. IEICE Trans Inf Syst (Japan Ed) 2001;J84-D-II:909–917. (in Japanese)
4. Itoh T, Kai A, Konishi T, Itoh Y. An understanding strategy based on plausibility score in recognition history using CSR confidence measure. Proc ICSLP '04, p 2133–2136, Jeju Island, Korea.
5. Kokubo H, Amano A, Hataoka N. Robust speech recognition for car environment noise. IEICE Trans Inf Syst (Japan Ed) 2000;J83-D-II:2190–2197. (in Japanese)
6. Komatani K, Kawahara T. Flexible dialogue management for generating efficient confirmation and guidance using confidence measures of speech recognition result. IPSJ J 2002;43:3078–3086. (in Japanese)
7. Raymond C, Esteve Y, Bechet F, De Mori R, Damnati G. Belief confirmation in spoken dialog systems us-

ing confidence measures. Proc ASRU 2003, p 150–155, St. Thomas, U.S. Virgin Islands.
8. Tsutsumi S, Isobe T, Morishima M. Confidence measure using multi-normalized likelihoods. IPSJ SIG Technical Reports, SLP-57, p 31–36, 2005. (in Japanese)
9. Litman DJ, Hirschberg JB, Swerts M. Predicting automatic speech recognition performance using prosodic cues. Proc 6th Applied Natural Language Processing Conference (NALP-NAACL00), p 218–225, Seattle, USA, 2000.
10. Kai A, Ishimaru A, Itoh T, Konishi T, Itoh Y. Analysis and detection of spoken corrections in spoken dialog between human and car navigation system. 2001 Autumn Meeting of the Acoustical Society of Japan, 2-1-8, p 63–64. (in Japanese)
11. Kitaoka N, Kakutani N, Nakagawa S. Detection and recognition of correction utterances on misrecognition of spoken dialog system. IEICE Trans Inf Syst (Japan Ed) 2004;J87-D-II:1441–1450. (in Japanese)
12. Hirasawa J, Miyazaki N, Aikawa K. Detection of misunderstandings in spoken dialogue system using system-user utterance sequence. IPSJ SIG Technical Reports, SLP-34, p 239–244, 2000. (in Japanese)
13. Kanda N, Komatani K, Ogata T, Okuno H. Experimental evaluation of spoken dialogue system using contextual constraint in database retrieval task. IPSJ SIG Technical Reports, SLP-55, p 107–112, 2005. (in Japanese)
14. Yamamoto H, Tanigaki K, Sagisaka Y. A statistical language model for conversational speech reflecting the previous utterance of the other participant. IEICE Trans Inf Syst (Japan Ed) 2001;J84-D-II:2507–2514. (in Japanese)
15. Bousquet-Vernhettes C, Vigouroux N. Context use to improve the speech understanding processing. Proc SPECOM 2001, p 89–92, Moscow.
16. Wutiwiwatchai C, Furui S. Nonlinear rescoring based on a dialogue context in discourse understanding. IPSJ SIG Technical Reports, SLP-51, p 37–42, 2004.
17. Higashinaka R, Sudoh K, Nakano M. Incorporating discourse features into confidence scoring of intention recognition results in spoken dialogue systems. ICASSP2005, Vol. 1, p 25–28, Philadelphia.
18. Ohmori K, Higashida M. New dialogue control method that efficiently ascertains customer intent through speech recognition. IPSJ SIG Technical Reports, SLP-32, p 45–50, 2000. (in Japanese)
19. Cho K, Miyayama A, Yamashita Y. Determination of the number of candidates using recognition scores for N-best based speech interface. IEICE Trans Inf Syst (Japan Ed) 2005;J88-D-II:1003–1011. (in Japanese)
20. Hirasawa J, Miyazaki N, Nakano M, Aikawa K. Studies on how users respond to misunderstandings of spoken dialogue systems. 2000 Spring Meeting of the Acoustical Society of Japan, 3-8-10, p 85–86. (in Japanese)
21. Grice HP. Logic and conversation. In Cole P, Morgan J (editors). Speech acts, syntax and semantics, Vol. 3. Academic Press; 1975. p 41–58.
22. Mizutani M, Itoh T, Kai A, Konishi T, Itoh Y. Maximum-likelihood spoken language understanding using CSR confidence measure and dialogue history. IPSJ SIG Technical Reports, SLP-45-19, p 113–118, 2003. (in Japanese)
23. Nakagawa S, Kai A. A context-free grammar driven, one pass HMM-based continuous speech recognition method. IEICE Trans Inf Syst (Japan Ed) 1993;J76-D-II:1337–1345. (in Japanese)
24. Ishikawa Y, Sawada K, Kido E. The evaluation of the speech interface. J Acoust Soc Japan 2005;61:79–84. (in Japanese)

AUTHORS

Noriki Fujiwara received his M.E. degree from Hokkaido University in 2004 and enrolled in the Ph.D. program. His scientific interests are speech language processing and spoken dialogue systems. He is a member of ASJ, IPSJ, and IEICE.

AUTHORS (continued) (from left to right)

Toshihiko Itoh received his D.Eng. degree from Toyohashi University of Technology in 1999 and became an assistant professor on the Faculty of Informatics at Shizuoka University. He is currently an associate professor in the Graduate School of Information Science and Technology at Hokkaido University. His research interests are speech language processing and spoken dialogue systems. He is a member of IPSJ, JSAI, ASJ, HIS, and IEICE.

Kenji Araki received his D.Eng. degree from Hokkaido University in 1988 and joined the Faculty of Informatics at Hokkai-Gakuen University. He moved to Hokkaido University in 1998, and is currently a professor in the Graduate School of Information Science and Technology, Hokkaido University. His research interests are natural language processing, especially machine translation and spoken dialogue processing. He is a member of IPSJ, NLP, JSAI, JCSS, ACL, IEEE, and AAAI.

Atsuhiko Kai received his D.Eng. degree from Toyohashi University of Technology in 1996 and joined the faculty as an assistant professor. He moved to Shizuoka University in 1999, and is currently an associate professor. His scientific interests are spoken language processing and dialogue processing with a focus on speech recognition. He is a member of ASJ, IPSJ, and JSAI.

Tatsuhiro Konishi received his D.Eng. degree from Waseda University in 1992 and became an assistant professor at Shizuoka University. He is currently an associate professor on the Faculty of Informatics. His scientific interests are intelligent education systems and intelligent dialogue systems. He is a member of IPSJ, JSAI, JSiSE, and JCSS.

Yukihiro Itoh received his D.Eng. degree from Waseda University in 1987 and became an assistant professor on the Faculty of Science and Engineering. He is currently a professor on the Faculty of Informatics at Shizuoka University. His scientific interests are natural language processing, dialogue systems, and intelligent education systems. He is a member of IPSJ, JSAI, NLP, JSiSE, and JCSS.