An Effective Speech Understanding Method with a Multiple Speech Recognizer based on Output Selection using Edit Distance*

Kazutaka Shimada[a], Satomi Horiguchi[a], and Tsutomu Endo[a]

[a] Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Iizuka Fukuoka 820-8502, Japan
{shimada, s_horiguchi, endo}@pluto.ai.kyutech.ac.jp

Abstract. In this paper, we propose a simple and effective method for speech understanding. The method incorporates some speech recognizers. We use two recognizers, a large vocabulary continuous speech recognizer and a domain-specific speech recognizer. The integrated recognizer is a robust and flexible method for speech understanding. For the integration process, we use a simple edit distance measure of each output sentence from each recognizer. Our method has high scalability and accuracy. The experimental results show the effectiveness of the proposed method.

Keywords: Multiple speech recognizer, Integration, Output selection, Edit distance.

1. Introduction

Speech understanding and dialogue systems have been developed for practical use recently. These systems often recognize user utterances incorrectly. It is important to deal with speech recognition errors for speech understanding systems. Extracting keywords and understanding an utterance using them reduce speech recognition errors (Bouwman et al. 1999, Komatani and Kawahara 2000). Another approach is to use domain-specific grammars and linguistic models. However these methods can not handle out of domain and spontaneous utterances.

One approach for the improvement is to repair recognition errors by users. There are many studies on detection of recognition errors in a speech output. Goto et al. (2005) have proposed some systems with nonverbal speech information, such as “SPEECH STARTER” and “SPEECH SPOTTER”. Ogata and Goto (2005) have proposed a speech input interface with a speech-repair function. Although repairing recognition errors by humans is effective in terms of development of a speech understanding system with high recognition accuracy, it is costly for users.

Combining some recognizers is one of the best approaches to improve the accuracy of speech understanding systems (Isobe et al. 2007, Utsuro et al. 2004). Utsuro et al. (2004) have obtained high accuracy by using some speech recognizers' outputs. However they dealt with word error reduction only. Although Isobe et al. (2007) have proposed a multi-domain speech recognition system by using some domain-specific recognizers, their system cannot treat out-of-domain utterances such as a chat between users. However the chat utterances often include a significant role as the context of the dialogue.

In this paper we propose a simple and effective speech understanding method based on a large vocabulary continuous speech recognizer (LVCSR) and some domain-specific speech

 This research was supported by the New Energy and Industrial Technology Development Organization (NEDO), Intelligent RT Software Project, 2008.

recognizers (DSSR). The task of this system is speech understanding for a livelihood support robot. The DSSRs recognize particular utterances about orders; e.g., order utterances from elders who need care and order utterances from nurses. Figure 1 shows the outline of the proposed method. We construct the grammar-based DDSR for order utterances with small vocabulary and high accuracy for each order type. We use the LVCSR for recognition of utterances that the DDSR can not recognize, such as a chat between users. The information recognized by the LVCSR is of assistance for context construction of a dialogue. If we handle these different speech recognizers selectively and integratively, we realize a flexible and robust speech understanding method. Figure 2 shows the effectiveness of the proposed multiple recognizer. The DDSR achieves the order recognition with high accuracy and the LVCSR supplies lack of information in the order utterances.

In this paper we use two recognizers, a large vocabulary continuous speech recognizer and a domain-specific speech recognizer for user's order utterance understanding. In the experiment we focus on the selective usage of the multiple speech recognizer. In other words, it is to select outputs from each recognizer. For example, with respect to the utterance "Please pick up the remote" in Figure 2, it is important which result to select. For the selection we propose the *One Commoner and Some Specialists* (OCSS) model. In our system, the LVCSR is the commoner, namely domain-independent, and the DSSRs are specialists, namely domain-dependent. We focus on the difference between outputs generated from the commoner and specialists. For the method, we compare several features to judge whether an input utterance is an order to the robot or not.

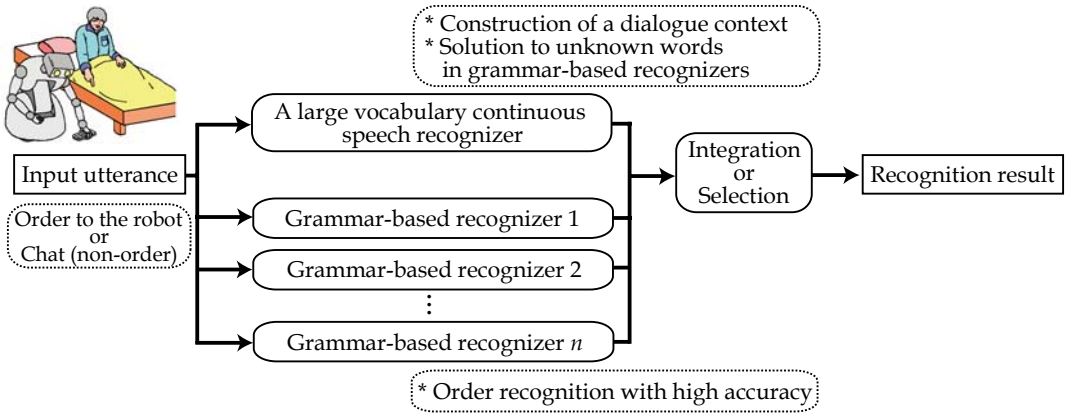

Figure 1: The outline of the proposed system.

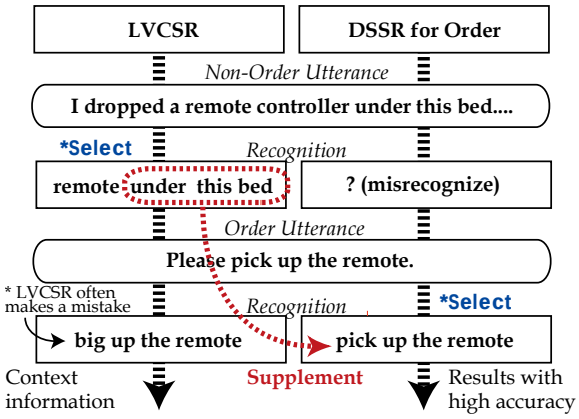

Figure 2: The effectiveness of a multiple recognizer.

2. Grammar-based recognizer

We use Julius as the LVCSR and Julian as the DSSR (Lee et al. 2001). The Julian consists of a vocabulary and a grammar file. For the grammar file we describe sentence structures in a BNF style, using word category names as terminate symbols. The vocabulary file defines words with its pronunciations (i.e., phoneme sequences) for each category.

Here we design grammar and vocabulary files of the Julian which accepts only order utterances from users. The acceptable utterances by our DSSR recognizer for order sentences are as follows:

- [Noun] wo [Verb] shi te kudasai or kureru ? [e.g., Please pick up the cellular telephone.]
- [Noun] wo [Verb] shitai [e.g., I want to eat the snack.]
- [Noun] wo [verb] kudasai or kureru? [e.g., Give me it.]
- [Location] ni aru [Noun] wo [Verb] shite kudasai or kureru? [e.g., Please bring the remote controller on the table.]
- [Noun] [e.g., The cellular telephone.]

We evaluated this grammar-based recognizer with 50 test utterances and 2 test subjects (male and female). The accuracy rate was 97% on the utterance-level. Here the utterance-level denotes that we judged that the output was correct if all words in an utterance were correct. This result shows that the DSSR used in our multiple recognizer is a robust and high accurate speech recognizer for the order utterances.

3. Output Selection

In our system, we need to judge whether an input utterance is an order to the robot or not. In this section we explain features and rules for the output selection.

3.1. Features

For the output selection, we compared each output in a preliminary exrperiment. As a result, we obtained 4 effective features for the selection; (1) confidence, (2) the number of candidates, (3) existance of a short pause mark and (4) a similarity between outputs.

- **Confidence**:

This is a confidence measure that is computed from the speech recgnizer Julius/Julian. This score is based on a posterior probability of each word (Lee et al. 2004). The range is from 0 to 1.

- **The number of candidates**:

The 2nd feature is the number of candidates of each DSSR. The number of candidates in a DSSR's output usually becomes small in the case that an input utterance is an order sentence. The reason is that the DSSR has high accuracy for target utterances because of small vocabulary. On the other hand, the number of candidates of a DSSR becomes large if the input is not an order utterance. In this situation, the DSSR generates many misunderstood results because the DSSR can not accept the input essentially.

- **Short paused mark**:

Julius/Julian can deal with a short pause in an utterance. If a short pause exists in an utterance, it output a short pause mark in the recognition result. The DSSR often contained the short pause mark in the result in the case that an input was not an order utterance[1]. Therefore the existance of the short pause mark in the output of the DSSR is effective to judge whether a input utterance is an order to the robot or not.

- **Similarity**:

The 4th feature is based on a similarity measure between outputs of the LVCSR and DSSR. Even human beings tend to misunderstand words which consist of similar pronunciations (Komatani et al. 2005). Here we focus on the output of the LVCSR. If an input is an order utterenace, the DSSR and the LVCSR generate similar outputs on phoneme-level because the

[1] I think that the reason is that the system replaced the part which it cannot analyze, with the mark.

LVCSR is domain independence. On the other hand, if the input is not an order utterenace, they often generate different outputs even on the phoneme-level because the DSSR never generates the correct result for non-order uttereances.

Komatani et al. (2007) have reported an utterance verification method based on difference of acoustic likelihood values computed from two recognizers. Kumar et al. (2005) have utilized Bhattacharyya distance to measure an acoustic similarity of different languages for multilingual speech recognition. In this paper, we use the edit distance as the similarity measure. The correspondence such as the edit distance is one of the most effective measures to identify high confidence words in outputs (Utsuro et al. 2004) and to extract similar word pairs (Komatani et al. 2005). In our method, if an input is an order utterance, the edit distance between the outputs from the DSSR and the LVCSR becomes small. However if the input is not an order utterance, that between the outputs from the DSSR and the LVCSR becomes large.

In our method, we compute the edit distance of utterance-level and word-level by using a DP matching algorithm. In the process, we compute the edit distance between phonemes of words for both levels.

3.2. Rules for the Selection

We apply the features to our selection process in the OCSS model. In the selection process, the rules to judge an utterance are applied in the following order:

1. If a short pause mark exists in the output of the DSSR or the output contains a word of which the confidence is 0, we select the output of the LVCSR as the final output.
2. If the number of candidates in the DSSR's output is less than 9, we select the output of the DSSR as the final output.
3. Compute the edit distance of the utterance-level (ED_{utter}) between the LVCSR and each DSSR. For the outputs of which the edit distance is less than $thresh_{utter}$, we select the output of the DSSR which contains the minimum ED_{utter}, as the final output.
4. Compute the edit distance of the word-level (ED_{word}) between the LVCSR and each DSSR. For the output of which the edit distance is less than $thresh_{word}$, we select the output of the DSSR which contains the minimum ED_{word} as the final output. Otherwise, the LVCSR as the final output.

The ED_{utter} is the edit distance value on the utterance-level. The ED_{word} is the average of the edit distance value computed on word-level. These values are normalized by the number of phonemes in the outputs. The $thresh_{utter}$ and $thresh_{word}$ are threshold values for the judgment. These values are decided experimentally.

In the computation of the word-level, we eliminate word pairs that are matched completely first. Next, we compute all the combinations of the other. Finally, we employ the minimum combinations as the word-level edit distance. Figure 3 shows an example of the calculation of the ED_{utter} and ED_{word}. In the figure, the dotted line denotes completely matched words. The numerals with arrows denote the original edit distance of the word pair. In the alignment process of word pairs, we select pairs which have the minimum value of the edit distance. In other words, we admit overlap of word pairs. For example, "noue vs. no" and "no vs. no" in Figure 3.

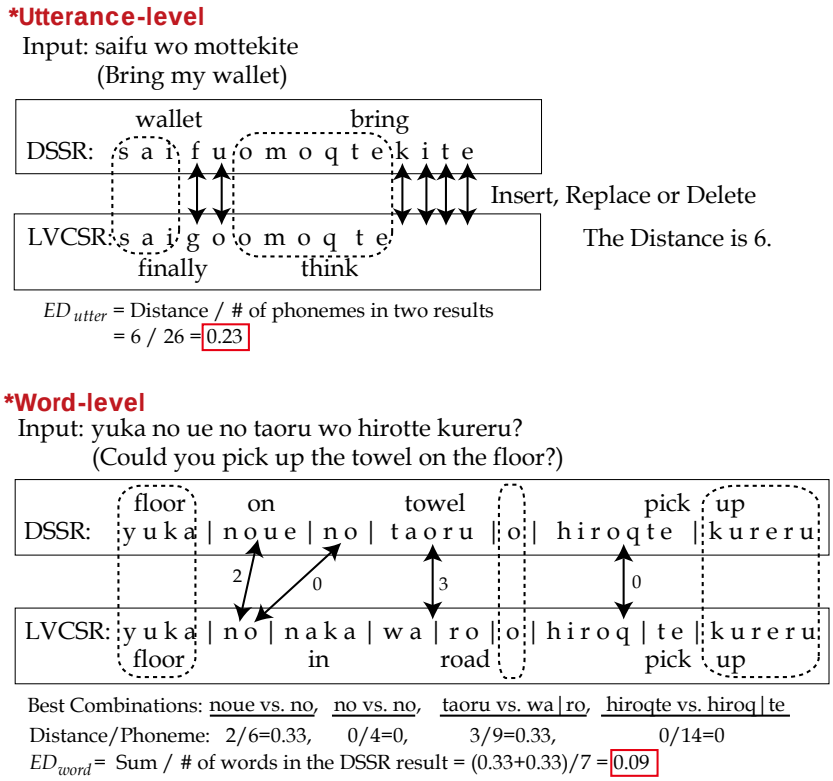

Figure 3: An example of the calculation of the word-level and utterance-level edit distance.

4. Experiment

We used 50 order utterances and 50 non-order utterances in this experiment. The DSSR accepts all the order utterances that used in this experiment. The non-order utterances consist of common greetings and daily conversation[2]. The number of test subjects was 4 persons (2 men and 2 women). We evaluated our method with cross-validation.

Table 1: The result of classification of the outputs.

Type	Precision	Recall	F-Value
Order	0.888	0.965	0.924
Non-Order	0.962	0.877	0.917

Table 1 shows the experimental result of the judgment process. Although our method was simple, we obtained high accuracy. In the experiment, the ranges of $thresh_{utter}$ and $thresh_{word}$ which were determined from training data were 0.24-0.26 and 0.08-0.13, respectively. In other words, the changes of these thresholds were little in the cross-validation. These results show the effectiveness and robustness of our method.

Next, we analyzed the correctness of each feature. Table 2 shows the distribution of each feature. As a result, we obtained the knowledge that the accuracy rates without the edit distance features were not always high.

On the basis of this result, we evaluated our method with the edit distance measures only. In other words, we used the rule 3 and 4 in Section 3.2. The thresholds, namely $thresh_{utter}$ and $thresh_{word}$, were the same as the previous experiment. Table 3 shows the experimental result. As a result, we obtained extremely high accuracy although the method was very simple.

[2] For example, “Today is busy.” and “I think that it rains in the afternoon of today.”

The standard deviation of the F-values in the cross-validation was approximately 0.007. Furthermore, the change of the F-value was at the most 0.01 even if the thresholds were fixed. These results show the effectiveness of our method.

Table 2: The correctness of each feature.

Feature	# of correct	# of incorrect
Confidence	48	15
Short pause	93	6
# of candidates	384	57
ED_{utter}	192	14
ED_{word}	388	3

Table 3: The result of classification of the outputs by using the edit distance only.

Type	Precision	Recall	F-Value
Order	0.963	0.985	0.974
Non-Order	0.985	0.963	0.974

5. Discussion

Our method is very simple and robust. In addition, our method has high scalability. This results from comparing the edit distance between the LVCSR and each DSSR. In general, a multiple recognizer consists of only DSSRs. Isobe et al. (2007) have proposed a multi-domain speech recognition system based on the model likelihoods of the different domain specific language models. Our method differs from it in use of the LVCSR. By using the LVCSR for a multiple recognizer, the system becomes simple and has high scalability. Figure 4 shows the advantage of our method as compared with previous studies. In general, systems in previous studies need to recalculate a model to select an output. In our method, when users add other grammar-based recognizers, what they need to change is 2 thresholds only ($thresh_{utter}$ and $thresh_{word}$)[3]. In fact, after the experiment, we added a new DSSR (the 3rd recognizer) to recognize some system commands for a robot such as “Stop”, “Move to the right by 50cm” and “Please go back a little”. As a result we also obtained high accuracy of output selection from 3 outputs without any changes of the thresholds; there was little decrease of the F-Value. Also we evaluated our method with 5 recognizers. The 4th recognizer was for question utterances such as “Where is my cellphone?”. The 5th recognizer was for order utterances from nurses such as “Carry these meals to patient's rooms”. In the additional experiment, we obtained approximately 0.95 on F-value with no change of the thresholds. Table 4 shows the details of the result. The test data consisted of 50 utterances (10 utterances for each category; daily conversation, orders from patients, system commands, orders from nurses and questions). The number of test subjects was 6 persons. These results show the effectiveness of the OCCS model with the edit distance.

Sako et al. (2006) have reported a method to discriminate a request to a system from a chat using AdaBoost. Tong et al. (2008) have reported a SVM-based method to separate a target language from other languages for spoken language recognition. Machine learning techniques generally need a large amount of training data to generate a classifier with high accuracy. However constructing training data by handwork is costly. As compared with them, our method can be realized with low cost. Besides, our method does not depend on particular speech recognizers although we used Julius/Julian in the experiment because it needs only phonemes of each output from the recognizer to select the final output.

Although we evaluated our method for output selection in this paper, we do not discuss integration of several outputs. To supplement information (e.g., “under the bed” in Figure 2)

[3] Note that if users want to select an output from DSSRs only, our method does not need any changes because the thresholds are used to distinguish between an order utterance and a chat. If the system does not need to recognize a chat in a dialogue, all it needs to do is just select the output that contains the highest similarity with the LVCSR as the final output.

and to construct a context of a dialogue by using the LVCSR are our future work. In the experiment, the word accuracy rates of the LVCSR were 65%[4] for order utterances and 32% for non-order utterances, respectively. This result shows the importance of improvement of the accuracy of the LVCSR. To apply the method in related work (Utsuro et al. 2004) to the LVCSR is also significant future work.

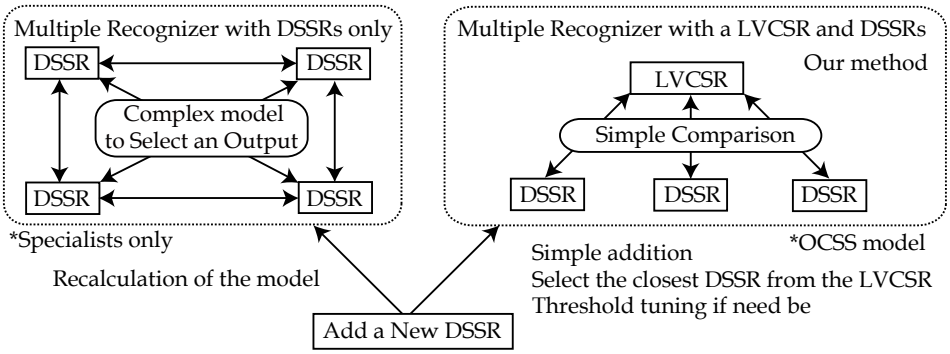

Figure 4: Comparison of a previous method and our method.

Table 4: The result of classification with 5 recognizers (edit distance only).

Type	Precision	Recall	F-Value
1st (LVCSR)	0.838	0.950	0.891
2nd (Patient's Order)	0.983	0.966	0.975
3rd (Command)	1.000	0.933	0.966
4th (Question)	0.948	0.917	0.932
5th (Nurse's Order)	1.000	0.983	0.992
Average	0.954	0.950	0.951

6. Conclusions

In this paper, we proposed a simple and effective method for speech understanding with some speech recognizers. Our method, OCSS model, does not need any complex computation and models. It uses an edit distance measure only. Furthermore, it can deal with out-of-domain utterances by using the LVCSR's output. We obtained the high precision and recall rates in the experiment. Future work includes (1) construction of the context using LVCSR and (2) development of multi modal interface with other sensors in the robot.

References

Bouwman, C., J. Sturm and L. Boves. 1999. Incorporating Confidence Measures in the Dutch Train Timetable Information System Developed in the ARICE project. *Proceedings of ICASSP*.

Goto, M., K. Itou and T. Kobayashi. 2005. Speech Interface Exploiting Intentionally-Controlled Nonverbal Speech Information. *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology* (UIST 2005), pp.35-36.

Isobe, T., K. Itou and K. Takeda. 2007. A Likelihood Normalization Method for the Domain Selection in the Multi-Decoder Speech Recognition System. *IEICE TRANSACTIONS on Information and Systems* (Japanese Edition) Vol.J90-D No.7 pp.1773-1780.

Komatani, K. and T. Kawahara. 2000. Flexible Mixed-initiative Dialogue Management Using Concept-level Confidence Measures of Speech Recognizer Output. *Proceedings of*

[4] Note that this accuracy rate is on word-level, i.e., word accuracy. On the other hand, the accuracy of the DSSR mentioned in Section 2 is utterance-level. In other words, even if the output utterance contains one mistake, it is incorrect. There is a significant difference between the LVCSR and the DSSR for the accuracy in this experiment.

COLING 2000, Vol. 1, pp. 467-473.

Komatani, K., R. Hamabe, T. Ogata, and H. G. Okuno. 2005. Generating Confirmation to Distinguish Phonologically Confusing Word Pairs in Spoken Dialogue Systems. *Proceedings of 4th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pp. 40-45

Komatani, K., Y. Fukubayashi, T. Ogata, and H. G. Okuno. 2007. Introducing Utterance Verification in Spoken Dialogue System to Improve Dynamic Help Generation for Novice Users. *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pp. 202--205.

Kumar, S. C., V. P. Mohandas and H. Li, 2005. Multilingual Speech Recognition: A Unified Approach. *Proceedings of InterSpeech 2005 - Eurospeech - 9th European Conference on Speech Communication and Technology*, pp. 3357-3360.

Lee, A., T. Kawahara and K. Shikano. 2001. Julius - an open source real-time large vocabulary recognition engine. *Proceedings of European Conference on Speech Communication and Technology* (EUROSPEECH), pp. 1691--1694, 2001.

Lee, A., and K. Shikano and Kawahara, T. 2004. Real-time word confidence scoring using local posterior probabilities on tree trellis search. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP2004), I, pp. 793-796.

Ogata, J. and M. Goto. 2005. Speech Repair: Quick Error Correction Just by Using Selection Operation for Speech Input Interfaces. *Proceedings of Interspeech 2005*, 2005, 133-136.

Sako, A., T. Takiguchi and Y. Ariki. 2006. System Request Discrimination Based on AdaBoost. *IEICE technical report. Natural language understanding and models of communication* (in Japanese), Vol. 106, No. 411, pp. 19-24.

Tong, R., B. Ma, H. Li, and E. S. Chng. 2008. Target-oriented phone tokenizers for spoken language recognition. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP 2008), pp. 4221-4224.

Utsuro, T., H. Nishizaki, Y. Kodama and S. Nakagawa. 2004. Estimating Highly Confident Portions Based on Agreement among Outputs of Multiple LVCSR Models. *Systems and Computers in Japan*, Vol.35, No.7, pp. 33-40.