SELECTED PAPER

Selection and Rejection of Recognition Results Obtained with Multiple Recognizers of Various Recognition Units Using Support Vector Machine

Nobuhisa Suzuki, Tomoyuki Tanaka, Masafumi Nishida, and Seiichi Yamamoto

Graduate School of Engineering, Doshisha University
1-3 Tataramiyakodani, Kyoutanabe-shi, Kyoto 610-0394, Japan
dtl0761@mail4.doshisha.ac.jp, {mnishida, seyamamo}@mail.doshisha.ac.jp

Abstract

We propose a novel speech recognition method using multiple recognizers of various recognition units. Our method can improve the word recognition accuracy and reject incorrect recognition results when the correct recognition results are obtained. We use a Support Vector Machine (SVM) to select a single recognition result or reject recognition results from multiple recognition results. Experimental results show that the proposed method achieves higher word recognition accuracy and classification accuracy in comparison with a single recognizer and ROVER.

1. Introduction

It is impossible to recognize every utterance perfectly even with speech recognition technology. The confidence measure is the key technology of deciding whether recognition results are correct or false.

Kemp and Schaaf [1] and Wessel et al. [2] used confidence measures obtained with a single recognizer. Although these methods using a single recognizer can classify recognition results into correct or false, they do not improve the word accuracy itself. Various methods have been proposed to select a single recognition result out of multiple recognition results. Fiscus [3] proposed ROVER, a method of discerning multiple recognition results by majority vote. As a method based on ROVER, Schwenk and Gauvain [4] proposed a method of incorporating the language model information in ROVER and Hahn et al. [5] proposed a method of selecting the best result from multiple recognizers using ROVER in spoken language understanding. Nakagawa et al. [6] proposed a method of using an intersection set of multiple acoustic models and decoders. Katsumaru et al. [7] proposed a selection method using confidence measures obtained from multiple language models and language understanding models. In those methods, a recognition result is selected even if the correct result is not included in the candidates.

In this paper, we propose a novel speech recognition method that uses a Support Vector Machine (SVM) to classify multiple recognition results into correct or false. The proposed method can also improve word accuracy by selecting the best result from among multiple recognizers.

The method uses confidence measures obtained from a combination of recognizers of various recognition units. We use a “partial word unit”, as well as a morpheme unit and word unit, as a recognition unit. The partial word unit has certain combinations of morphemes obtained by computing the connection probabilities between consecutive morphemes.

The rest of this paper is organized as follows. Section 2 introduces the recognition units and language models using the multiple recognition units. Section 3 describes the method of selecting recognition results based on confidence measures. Section 4 describes the experimental setup and results. Section 5 covers the conclusion and future work.

2. Recognition Units and Language Models

Our research target is the recognition of destination names in utterances input to a car navigation system, e.g., “The destination is *Tokyo Disney Resort*”. The destination names are usually compound words and combinations of commonly used morphemes or words. Morphemes are commonly used as recognition units; however, they are so short that a number of acoustically similar recognition candidates appear. Longer recognition units are desirable from the viewpoint of word accuracy, but the number of lexicon items is huge. Thus, we explored the use of a new recognition unit, i.e., a “partial word unit”, which is selected on the basis of the connection probability between consecutive morphemes.

Figure 1 shows the scheme of making the “partial word unit”. When the destination phrase “*Tokyo Disney Resort*” is given, the recognition target is separated into “*Tokyo*”, “*Disney*”, and “*Resort*” by morphological analysis. The connection probability between morphemes is defined as follows.

$$P(w_i|w_{i-1}) > \theta \quad (1)$$

Here, w is a morpheme and θ is a threshold. Morphemes are combined when the connection probability is greater than θ . The result of a preliminary experiment led us to set θ as 0.03 because word recognition accuracies are highest at this value.

We used two different language models, e.g., N-grams and a regular grammar.

Figure 1: Scheme of making “partial word unit” based on connection probability between morphemes

Bi-gram and tri-gram language models with morpheme units and partial word units were trained with 150k destination names using a corpus containing 72 types of expressions per destination name (see Figure 2). We formulated 72 grammar rules corresponding to the 72 expression types. We implemented the following recognizers.

- (1) Bi-gram and tri-gram with morpheme unit
- (2) Bi-gram and tri-gram with partial word unit
- (3) Regular grammar with word unit

We selected two recognizers from the above (recognizer-A and recognizer-B).

Figure 2: Example of utterance-included expressions

3. Selection of Recognition Results Based on Confidence Measures

Table 1 shows the features for confidence measures. Two recognition results obtained from the two recognizers are classified into the following four patterns.

- (1) Both recognition results are correct.
- (2) Recognizer-A’s result is correct and recognizer-B’s result is false.
- (3) Recognizer-A’s result is false and recognizer-B’s result is correct.
- (4) Both recognition results are false.

Our method uses SVM to classify two recognition results into the above four patterns according to confidence measures shown in Table 1, and it selects a single recognition result or no recognition result. This method can improve word accuracy by itself. Figure 3 shows the classification and scheme for selecting recognition results. Recognizer-A and recognizer-B provide speech recognitions and the confidence measures shown in Table 1.

Table 1: Confidence measures

<i>CMscore</i>	Posterior probability for each recognition unit [8]
<i>Appearance rate</i>	Rate that the 1 st recognition candidate appears in 2 nd to 10 th recognition candidates
<i>Acoustic likelihood</i>	Acoustic score for each recognizer
<i>Language likelihood</i>	Language score for each recognizer (N-gram only)
<i>Correspondence</i>	Recognition results of two recognizers correspond or not correspond
<i>Phoneme difference</i>	Difference in number of phonemes of two recognizers

Figure 3: Proposed method

We selected the hyperparameters and Radial Basis Function (RBF) as the kernel function of the SVM because classification accuracies were highest in the preliminary experiment. We conducted experiments using the LIBSVM [8].

4. Experimental Evaluation

4.1 Experimental conditions

Speech data was sampled at 16 kHz and segmented into 25 ms frames every 10 ms with a 15 ms overlap. The acoustic features consisted of 25 components derived from 12 Mel Frequency Cepstral Coefficients (MFCCs), 12 delta MFCC, and delta power. A Phonetic Tied-Mixture (PTM) triphone of 3000 states and 64 mixture distributions model was used as the acoustic model.

As language models, we used forward bi-gram language models in the first pass and backward tri-gram language models in the second pass, and finite-state automaton models. We used Julius as a decoder [9]. We registered 150k destination names in the word dictionary. There were about 18k recognition units for the morpheme unit, about 85k for the partial word unit, and about 505k for the word unit. 400 speech datums, in which 100 destination names were expressed in different patterns, were prepared as the evaluation test set and were uttered by 10 males. 1200 utterances of the 4000 evaluation data contain Out-of-Vocabulary (OOV) words as the destination names. Finally, recognition results were classified by 10-fold cross-validation.

Word accuracies were 75.3% for N-gram with morpheme unit, 81.7% for N-gram with partial word unit, and 85.9% for regular grammar with word unit, in the case that evaluation test sets did not contain OOV words.

4.2 Experimental results

Classification accuracy is calculated as follows:

$$Acc = \frac{\text{number of utterances correctly classified into four patterns}}{\text{number of utterances}} \times 100 \quad (2)$$

Figure 4: Word accuracies of various combinations of recognizers with SVM and ROVER

Classification accuracies for the single recognizer and ROVER were calculated as the ratio of patterns correctly classified into correct or false to total number of patterns. Table 2 shows classification accuracies of recognition results by each single recognizer. OOV words were not included in the 2800 speech data for evaluation. Table 3 shows the classification accuracy given by ROVER, which selects a single recognition result from the above three recognizers by majority vote.

Table 2: Classification accuracy of each single recognizer

recognizer	accuracy [%]
N-gram (morpheme unit)	89.1
N-gram (partial word unit)	91.0
Grammar (word unit)	90.0

Table 3: Classification accuracy of ROVER

recognizer	accuracy [%]
N-gram (morpheme unit)	87.9
N-gram (partial word unit)	
Grammar (word unit)	

Table 4: Classification accuracy of proposed method

recognizer	accuracy [%]
N-gram (morpheme unit)	90.8
N-gram (partial word unit)	
N-gram (partial word unit)	95.6
Grammar (word unit)	
Grammar (word unit)	95.5
N-gram (morpheme unit)	

Table 4 shows the classification accuracies of proposed recognizers. Our method achieved the highest classification accuracy. All combinations of recognizers based on N-gram models and regular grammar were effective. Figure 4 shows word accuracies based on the classifications using confidence measures from the single recognizer, ROVER, and proposed method. The proposed method achieved the highest word accuracy.

4.3 Evaluation of OOV utterances

We also evaluated speech data containing OOV words. The evaluation used 1200 of the 4000 speech datums containing OOV words as the destination name. Table 5 shows the detection rate of OOV for each single recognizer, Table 6 shows the result obtained by ROVER, and Table 7 shows the result obtained by the proposed method. These tables show that the proposed method can classify multiple recognition results with high accuracy even if OOV words are included in the utterances.

Table 5: OOV detection rate of single recognizer

recognizer	accuracy [%]
N-gram (morpheme unit)	81.4
N-gram (partial word unit)	79.7
Grammar (word unit)	73.2

Table 6: OOV detection rate of ROVER

recognizer	accuracy [%]
N-gram (morpheme unit)	69.8
N-gram (partial word unit)	
Grammar (word unit)	

Table 7: OOV detection rate of proposed method

recognizer	accuracy [%]
N-gram (morpheme unit)	79.0
N-gram (partial word unit)	
N-gram (partial word unit)	83.1
Grammar (word unit)	
Grammar (word unit)	81.5
N-gram (morpheme unit)	

5. Conclusion

We proposed a novel speech recognition method in which recognition results are selected or rejected by using SVM on the basis of confidence measures obtained from multiple recognizers of various recognition units. Experimental results showed that the proposed method achieves higher word accuracy and classification accuracy in comparison with conventional methods. In particular, the combination of

the N-gram-based recognizer and grammar-based recognizer achieved the highest performance in our experiment.

As future work, we will evaluate the method that judges whether the correct recognition result is included under the 2nd candidate if the recognition result is rejected. We will also study new confidence measures by combining recognizers.

Acknowledgment

This work was supported by a Grant-in-Aid for Young Scientists (B), 21700184.

References

- [1] T. Kemp and T. Schaaf: Estimating confidence using word lattices, Proc. Eurospeech, pp. 827-830, 1997.
- [2] F. Wessel, K. Macherey and H. Ney: A comparison of word graph and N-best list based confidence measures, Proc. Eurospeech, pp. 315-318, 1999.
- [3] J. G. Fiscus: A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER), Proc. ASRU, pp. 347-354, 1997.
- [4] H. Schwenk and J. -L. Gauvain: Combining multiple speech recognizers using voting and language model information, Proc. ICSLP, pp. 915-918, 2000.
- [5] S. Hahn, P. Lehnen and H. Ney: System combination for spoken language understanding, Proc. Interspeech, pp. 236-239, 2008.
- [6] S. Nakagawa, T. Watanabe, H. Nishizaki and T. Utsuro: An unsupervised speaker adaptation method for lecture-style spontaneous speech recognition using multiple recognition systems, IEICE, Vol. E88-D, No. 3, pp. 463-471, 2005.
- [7] M. Katsumaru, M. Nakano, K. Komatani, K. Funakoshi, T. Ogata and H. G. Okuno: Improving speech understanding accuracy with limited training data using multiple language models and multiple understanding models, Proc. Interspeech, pp. 2735-2738, 2009.
- [8] C.-C. Chang and C.-J. Lin: LIBSVM: a library for support vector machines, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [9] A. Lee and T. Kawahara: Recent development of open-source speech recognition engine Julius, Proc. APSIPA, pp. 131-137, 2009.