Recognition of Dialogue Speech

Shozo MAKINO, Akinori ITO, Takashi OTSUKI, Yukihiro OSAKA, Yoshiyuki OKIMOTO and Motoyuki SUZUKI
Computer Center, Tohoku University
Aoba-ku, Sendai 980-77, JAPAN
e-mail:makino@cc.tohoku.ac.jp

1. Introduction

In dialogue speech, variations of coarticulation effect and speaking rate are larger comparing to those in isolated spoken word and continuous speech. Furthermore, language models needed in dialogue speech recognition are much different from those used in continuous speech recognition. In order to solve these problems, we have carried out several researches for building up a spoken dialogue system. The researches are separated into three parts.

(1) robust phoneme recognition with high accuracy,

(2) language processing for dialogue speech, and

(3) performance evaluation method when using various language models.

2. Robust phoneme recognition

2.1 Phoneme recognition using HM-net

Many methods have been proposed to construct context-dependent phoneme models using HMMs to get better performance. These conventional methods require the previously given contextual factors. If the given factors are not sufficient, the constructed models do not give good recognition performance. We proposed a new HMnet construction algorithm requiring no contextual factors. The new algorithm is very similar to the Successive State Splitting(SSS) algorithm(called SSS-original) proposed by J. Takami and S. Sagayama. The difference of the new algorithm(called SSS-free) from the SSS-original is the way of splitting the state on the contextual domain. Traning samples passing through a single path in the HMnet made by the SSS-free have similar acoustical characteristics.

In order to show the effectiveness of the SSS-free, we carried out a speaker-dependent phoneme recognition for 24 Japanese phonemes in the ATR Japanese sentense speech uttered by two male and two female speakers. Training data was 400 sentences in the ATR Japanese sentence speech Database, and test data was other 103 sentences. The maximum number of states per phoneme was set to 30 and the toral number of states was set to 500. Average recognition score of the SSS-free was 93.1% and that of the SSS-original was 91.3%. The average recognition score increased by 1.8%. We found that, in the SSS-original, there were several phonemes that the state splitting of the phonemes could not be done before the number of states reached the pre-defined number.

2.2 Phoneme recognition using discriminative training and reference patterns with variable length

We proposed the modified LVQ2 algorithm for phoneme recognition. The length of phoneme reference pattern made by the algorithm was fixed to 70 ms. It showed high phoneme recognition performances for continuous speech. Hoewever, it did not show high performances for spontaneous speech because the phoneme reference patterns with fixed length could not deal with big change in phoneme duration. Then we developed a new construction algorithm of the reference patterns with variable length based on discriminative training.

At first, we construct the initial reference patterns for each phoneme from a large amount of phoneme training samples with variable length using the clustering method. If the input traing sample is correctly recognized, the modification of the reference patterns is not carried out. Otherwise, the modification is carried out.

(1) The input sample is matched with the nearest reference patterns in each of the incorrect and correct phoneme classes using the DP-based time-warping method.

(2) The correspondence between the frame of the input sample and the frame of the nearest reference pattern in each of the incorrect and correct phoneme classes is calculated.

(3) The corresponding frame vector of the reference pattern of the incorrect phoneme class is moved away from the frame vector of the input sample based on discriminative training algorithm. On the other hand, the corresponding frame vector of the reference pattern of the correct phoneme class is moved nearer to the frame vector of the input sample.

Furthermore, we proposed a new segmentation algorithm taking into account the distribution of phoneme duration.

The recognition and segmentation experiments were carried out for the 212 word vocabulary uttered by 2 male and 3 female speakers, where the reference patterns of phonemes were made by training samples in the 212 word vocabulary uttered by another 8 male and 7 female speakers. We obtained 73.1% of phoneme accuracy.

2.3 Phoneme recognition with adaptability function to speaking rate of input speech

We proposed a spoken word recognition system based on phoneme duration information estimated from the speaking rate of an input speech. In general, the speaking rate is expressed by number of morae per second. We found that normalization of the duration of each mora with the average mora duration in input speech and with the average duration of each mora class was very effective to reduce the variation. Because recognition of mora is very difficult, we examined utilization of the average vowel duration in input speech instead of the average mora duration. The average vowel duration was also very effective for the normalization. Furthermore, the variation of duration of each phoneme class became small using the normalization with the average vowel duration in input speech and with the average duration of each phoneme class. By taking into account the above results, we proposed the first-order linear regressive equation as a function of the average vowel duration for predicting the duration of each phoneme in input speech. We estimated the errors between the predicted duration and the manually obtained phoneme duration. The errors became small comparing to that without the prediction.

We applied this method to isolated spoken word recognition. A phoneme sequence is at first recognized and then the average vowel duration is computed from the recognition result. The duration of each phoneme in each word template is predicted using the linear regressive equation based on the average vowel duration by taking into account the phoneme context in the phoneme sequence of the word template. The matching between the input sequence of similarity vectors and the phoneme sequence of each word temlate is carried out using the dynamic time-warping method by taking into consideration the predicted phoneme duration constraints.

The word recognition score was 97.3% for the 212 word vocabulary uttered by 10 male and 10 female speakers using the equation taking into account phoneme context. The score increased by 1.6% comparing to that without the information of speaking rate.

3. Language processing for dialogue speech

3.1 Bunsetsu-automaton model for daialogue speech

Our language model for spoken dialogue is constructed based on Bunsetsu-units and a dependency grammar. We constructed the Bunsetsu-automaton so as to accept various Bunsetsu-units appeared in spontaneous transcriptions in the ASJ spontaneous speech corpus. We analyzed 1134 utterances and 7113 uttered Bunsentsu in the utterances. We showed some differences between the automaton for spontaneous speech and that for sentense speech. The perplexity of the model is discussed based on the frequency of each word in the model.

3.2 Word pre-selection method based on redundant hash addressing for continuous speech recognition

In order to reduce the recognition time and to realize an on-line dialogue system with a large size vocabulary, it is important to develop an efficient and accurate bottom-up word pre-selection method.

We proposed a new word-preselection method called "extended redundant hash addressing method". This method is based on isolated word recognition method using the redundant hash addressing priciple proposed by T. Kohonen et. al. The redundant hash addressing method can not be used for continuous speech recognition because the word boundary in continuous speech is not known. We have extended Kohonen's method to continuous speech recognition. The proposed method spots words from the input speech using the information about locations of features in reference patterns. Combination of this method and an ordinary parser makes a speech recognition system faster. The proposed method gave comparable performance to that with traditional continuous-DP matching.

3.3 Acquisition of language model based on HMnet

Statistic language models obtained from a large number of training samples play an important role in speech recognition. In order to obtain higher recognition performance, we should introduce long distance correlations between words. However, conventional statistic language models such as Word n-grams and ergodic HMMs are insufficient for expressing long distance correlations between words. Then we proposed an acquisition method for a language model based on HMnet taking into consideration long distance correlations and word location.

The effectiveness of this method is supported by experiments involving comparison to conventional methods. A source language model expressed by a finite state automaton represents control commands for a text editor, where the number of states is 23 and the vocabulary size is 36 words. Transition probabilities were assumed to be equal. Training and test samples were randomly generated from the model. The number of training samples was 30,000, the number of test samples was 20,000. The HMnets showed lower perplexities than traditional models such as word n-grams and ergodic HMMs.

4. Performance evaluation method when using various language models

We proposed the performance evaluation methods on sentense recognition using a transition information between words and using a finite state automaton. For the both cases, we showed the algorithms to calculate the number of word sequences at a short distance from a sentense in polynomial time for predicting the relationships between word recognition scores and sentense recognition scores. The comparison between the prediction and the simulation showed that our methods can be applied to getting the lower limit of the sentense recognition scores when the quality of the word lattice is good.

5. Conclusion

We have proposed three phoneme recognition methods and three language processing methods for recognition of dialogue speech. Furthermore, we have proposed performance evaluation methods when using various language models. We showed the effectiveness of these methods by experiments.

Now, we have been developing a spoken dialogue system based on the Bunsetsu-automaton model and the phonem recognition method described in the section 2.2.

List of papers presented by the authors

[1] T. OTSUKI, A. ITO, S. MSAKINO and T. SONE:"Performance Prediction of Word Recognition Using the Transition Information between Phonemes or between Characters", IEICE Trans.(D-II), J76-D-II, 1090-1096 (1993) (in Japanese)

[2] T. OTSUKI, A. ITO, S. MAKINO and T. OTOMO:"Performance Prediction of Word Recognition Using the Probability of Word Occurence", IEICE Trans.(D-II), J77-A, 274-281 (1994)(in Japanese)

[3] S. MAKINO, A. ITO, M. ENDO, K. KIDO:"A CONTINUOUS SPEECH RECOGNITION SYSTEM USING A MODIFIED LVQ2 METHOD AND A DEPENDENCY GRAMMAR WITH SEMANTIC CONSTRAINTS", Int. Jour. Pattern Recognition & Artificail Intelligence., 8, 197-213 (1994)

[4] H. Koga, S. Makino and K. Kido:"Effect of time window and lifter of speech analysis on isolated vowel recognition using local spectral peaks", Jour. Acoust. Soc. Jpn., 51, 130-132 (1995)(in Japanese)

[5] A. ITO and S. MAKINO:"Word Pre-Selection Using Extended Redundant Hash Addressing Method for Continuous Speech Recognition", IEICE Trans.(D-II), J-78-D-II, 400-408 (1995)(in Japanese)

[6] M. SUZUKI, S. MAKINO, A. ITO, H. ASO, H. SHIMODAIRA:"A New HMnet Construction Algorithm Requiring No Contextual Factors", IEICE Trans. INF.&SYST., E-78-D, 662-668 (1995)

[7] A. ITO, S. MAKINO:"A NEW WORD PRE-SELECTION METHOD BASED ON AN EXTENDED REDUNDANT HASH ADDRESSING FOR CONTINUOUS SPEECH RECOGNITION", Proc. ICASSP'93, II-299-II-302 (1993)

[8] S. Makino, Y. Osaka, T. Sone:“A Spoken Word Recognition System with Adaptability to Speaking Rate of Input Speech”, Proc. IWSP’93, 19-24 (1993)

[9] T. Otsuki, A. Ito, S. Makino, T. Otomo:“THE PERFORMANCE EVALUATION METHOD ON SENTENCE RECOGNITION SYSTEM USING A FINITE STATE AUTOMATON”, Proc. ICASSP’94, I-397-I-401 (1994)

[10] Y. OSAKA, S. MAKINO, T. SONE:“Spoken Word Recognition Using Phoneme Duration Information Estimated from Speaking Rate of Input Speech”, Proc. ICSLP’94, 191-194 (1994)

[11] S. Makino, Y. Okimoto, K. Kido, H. R. Kim, Y. J. Lee:“KOREAN CONSONANT RECOGNITION USING A MODIFIED LVQ2 METHOD”, Proc. WESTPRAC’94, 1033-1038 (1994)

Keywords:HMnet, Adaptation to input speech, Discriminative training, Acquisition of language model, Performance evaluation