DEVELOPMENT OF A VERY FAST PREPROCESSOR

Young-Mok Ahn, Hoi-Rin Kim

Spoken Language Processing Section
Electronics and Telecommunications Research Institute

E-mail : aym@zenith.etri.re.kr

ABSTRACT — This paper proposes a very fast preprocessor for a large vocabulary isolated word recognition. This preprocessor extracts a few candidate words using the frequency and the time information for each word. For designing reference pattern, we use the order of amplitude of speech feature. So, the proposed preprocessor has a small computational load after the extraction of speech feature. In order to show the effectiveness of the proposed preprocessor, we compared it to a speech recognition system based on semi-continuous hidden Markov model and a VQ-based preprocessor by computing their recognition performances of a speaker independent isolated word recognition. In experiments, we use three types of speech database. The first, a speech database consists of 244 words including digits, English alphabets, etc. The second, a speech database consists of 22 words including section names. The third a speech database consists of 35 words. This preprocessor is composed of three major parts: the feature extraction, the feature sorting, and the reference pattern matching with reference templates. After sorting it requires only one vector addition per frame, namely, *vocabulary size x length of incoming frame*. In consequence, this approach is therefore much faster than our previous version which is the VQ-based preprocessor for isolated word recognition task (Ahn et al, 1994). In the experimental results, the accuracy of feature sorting based preprocessor is 99.86 % with 90 % reduction rate for the speech database of 244 words.

INTRODUCTION

In a large vocabulary task, the speech recognition system should have some efficient algorithms for reducing the computational load. To develop an efficient algorithm which is reducing the search space such as Viterbi beam search, it is necessary that the algorithm should not degrade the system performance. In real applications, the speech recognition system should have another efficient algorithm which can reduce the memory size. If the speech recognition system have a large memory space, it is not a emulous product. Considering the computational load and the memory size, the proposed approach shows the solvability on the isolated word recognition problem.

Although the proposed preprocessor is less exact compared to other speech recognition systems in the recognition performance of the first candidate word, it keep an appropriable preprocessor, because the preprocessor can extract a small candidate words including correct word from the large vocabulary within correct word on maintaining the system performance. And the preprocessor is easily constructed from the training speech database. On the other hand, use of the preprocessor for continuous speech recognition is not easy. This is because, to achieve the preprocessor with high accuracy performance, a large speech database with the precise labeling information should be prepared. So, the preprocessor combined in the continuous speech recognition system. To measure the performance of the proposed preprocessor, we performed the experiments on the speaker independent isolated word speech recognition. And we compared that with other speech recognition system performances.

In the rest of the paper, an overview of the proposed preprocessor and other speech recognition system is first introduced in the section of system overview, and than the performance of each system is described in the section of experiments. The experimental results and concluding remarks are given in the section of conclusions.

SYSTEM OVERVIEW

In order to show the effectiveness of the proposed preprocessor, we compared it to three types of speech recognition system for their recognition performances of a speaker independent isolated word recognition. The brief introduction is as below. Although the main purpose of each system is not same, but they provide a appraisal experimental result for isolated word recognition.

Feature sorting based preprocessor(FSP)

Recently, auditory system has begun to play a larger role in motivating the design of some speech recognition front-end systems. And spectral transitions play an important role in human auditory perception. In other words, the current speech signal is much influenced by the previous speech signals. And it is also meaningful to hypothesize that the speech signals have the relationship among orders of feature vectors. The idea of the feature sorting based preprocessor is derived from the hypothesis.

This preprocessor is composed of three major parts: the feature extraction, the feature sorting, and the reference pattern matching. After feature sorting it requires only one vector addition operation per frame, namely, *vocabulary size x length of incoming frames*. And the feature sorting requires very small computational load compared with vector quantization. The feature sorting based preprocessor is very similar to the VQ-based preprocessor except the vector quantization. Let $y_1, y_2, \ldots, y_n$ be a label sequence produced by the feature sorting based acoustic processor in response to an utterance of some unknown words. In the feature sorting based preprocessor of obtaining a short list of words, we seek a word scoring function of the form where S_w

$$S_w = \sum_{t=1}^{n} v(y_t, w) + i_w \qquad (w = 1, 2, \ldots, N)$$

denotes the score for word w, $v(y_t, w)$ denotes a real-valued vote cast by label y_t for word w, i_w denotes the initial value of S_w, and N denotes the number of words in the vocabulary. A short list can then be constructed from the highest scoring words.(Lalit et al, 1988) Further information for constructing reference patterns of the feature sorting based preprocessor can be found in (Ahn et al, 1994), (Lalit et al, 1988).

(Linde et al. 1980). In the feature sorting based preprocessor, the vector quantization is replaced sorting which is ordering of amplitude of speech features. So, In the aspect of computational load, the VQ-based preprocessor is not comparable to the feature sorting based preprocessor. The brief description of speech features is as bellows. Input speech was sampled by 16 kHz and quantized by 16 bit. For feature extraction, we use the perceptual linear prediction(PLP) feature. Each frame of speech is represented by a 16 dimensional feature vector that consists of only 16 PLP coefficients.

VQ-based preprocessor(VQP)

This system is a simple VQ-based preprocessor which can reduce the amount of computation in speaker independent isolated word recognition (Ahn et al, 1994), (Lalit et al, 1988), (Linde et al. 1980). This preprocessor is composed of three major parts: the feature extraction, the vector quantization, and the reference pattern matching. To evaluate this system, we use four feature vectors: LPC cepstrum with a bandpass lifter, delta cepstrum, delta-delta cepstrum, and energy(log energy, delta log energy, delta-delta log energy). The size of analysis window is 20 msec and the analysis interval is 10 msec.

Continuous speech recognition system(CSR)

This system is a Korean continuous speech recognition system using phone-based semi-continuous hidden Markov model(SCHMM) method for the hotel reservation domain (Kim et al, 1994), (Han et al, 1995). Task domain of the system is the query sentences of hotel reservation with 244 words. The vocabulary consists of room numbers from 10 to 99, 43 words for dates, 7 words for day of week, 26 words for English alphabet. We defined 339 context dependent phone models based on triphone model for pronunciation dictionary of each word. And we defined the model topology with 3 states and 8 transitions including skip transitions. In this system, we use four feature vectors: LPC cepstrum with a bandpass lifter, delta cepstrum, delta-delta cepstrum, and energy(log energy, delta log energy, delta-delta log energy). The size of analysis window is 20 msec and the analysis interval is 10 msec.

Spontaneous speech recognition system(SSR)

We implemented Korean spontaneous speech recognition system based on Janus system which was developed by Interactive Systems Laboratories(ISL) at Carnegie Mellon University(Lee et al, 1996), (Suhm et al, 1995). Janus system was aimed at the multi-lingual speech translation system, so that it was designed to accommodate easily a new language. The brief description of Korean spontaneous speech recognition system is as follows. Input speech was sampled by 16 kHz and quantized by 16 bit. For feature extraction, we use the perceptual linear prediction(PLP) feature. As an acoustic model unit, we used a set of 40 monophones. For the training of these units we used a speech database which is labeled 1,295 generalized triphones. Especially, the speech database for training the acoustic model have not involved the 244 words set.

EXPERIMENTS

The evaluation speech data consist of isolated word sets. In experiments, we use three types of speech database. In the comparison of performance with each system, we used a speech database which consists of 244 words set. This vocabulary consists of room numbers from 10 to 99, 43 words for date, 7 words for day of week, 26 words for English alphabets. The words uttered by 40 male speakers were used for the training data and the words uttered by another 11 male speakers were used for the test data. We also used this speech database for evaluating the performance of the proposed preprocessor concerning the reduction rate and the recognition rate of the top 4 candidates words. Another speech database which consist of 22 words including section names and consist of 35 words, are used for evaluating the performance of the small vocabulary task.

The comparison of performance with each system

The recognition results of each system using 244 words are shown in Table 1. In the CSR and the VQ-based preprocessor, we used four codebooks, with each 256 codewords, that use the LPC-based cepstrum of order 12, delta cepstrum, delta-delta cepstrum and normalized log power, delta power, delta-delta power. In the SSR and the feature sorting based preprocessor, we used the perceptual linear prediction(PLP) feature. We used only the PLP feature of order 16 for evaluating the feature sorting based preprocessor. On the other hand, the delta PLP feature is additionally used for evaluating the SSR. So the order of feature in the SSR is 26. The baseline experiment showed that the recognition rate of the CSR is 90.07 %, the recognition rate of the VQ-based preprocessor is 92.54 %, the recognition rate of the SSR is 71.86 %, the recognition rate of the feature sorting based preprocessor is 68.31 % on the 244 words set.

	CSR	VQP	SSR	FSP
Speech Feature	Cepstrum	Cepstrum	PLP	PLP
Order of Feature	39	39	26	16
Accuracy	90.07 %	92.54 %	71.86 %	68.31 %

Table 1. Recognition results of each system

The performance of the feature sorting based preprocessor

Concerning the reduction rate of candidate words, Table 2 summarizes the overall results of experiment. The condition of constraint such as Viterbi beam width in the Viterbi beam search for reducing the search space is the likelihood ratio of each candidate words in comparison with the likelihood of the first candidate word. The candidate words in the Table 2 denote the average words to be searched for choosing the first candidate word. In this experiments, the proposed preprocessor showed 91.80 % accuracy on the 5 % condition of constraint and the reduction rate is 98.5 % so that we can find the correct word only searching for the 3.6 words. And the proposed preprocessor showed 100.0 % accuracy on the 30 % condition of constraint and the reduction rate is 58.0 % so that we can find the correct word only searching for the 102.6 words.

To evaluate the performance of the feature sorting based preprocessor, we experimented on the fourth candidate words. The results are presented on Table 3. In this experiments, the performance of the feature sorting based preprocessor resulted in 90.98 % accuracy on the fourth candidate words.

Constraint	Accuracy	Candidate Words	Reduction Rate
5 %	91.80 %	3.6	98.5 %
10 %	98.91 %	10.6	95.7 %
15 %	99.86 %	24.3	90.0 %
20 %	99.86 %	45.5	86.4 %
25 %	99.86 %	72.7	70.2 %
30 %	100.00 %	102.6	58.0 %

Table 2. Recognition results of each constraint

Top N	Accuracy
Top1	68.31 %
Top2	81.69 %
Top3	87.71 %
Top4	90.98 %

Table 3. Recognition results of the fourth candidate words

On the small vocabulary task, the 35 words set and the 22 words set accuracy were 85.71 % and 97.00 % respectively. The similarity of inter words of the 35 words set is greater than the 22 words set. The 35 words set consists of the predicative words but the 22 words set consists of section name.

Speech DB	Accuracy
35 Words Set	85.71 %
22 Words Set	97.00 %

Table 4. small vocabulary performance

CONCLUSIONS

This paper described the feature sorting based preprocessor and introduced various systems for the comparison of performances. The experiments showed that use of the feature sorting based preprocessor is an effective way of reducing the candidate words for isolated word recognition. The advantages of the feature sorting based preprocessor are that it is easy to construct reference pattern and it is not required computational load after extracting of speech feature vector. This kind of isolated word recognition systems can be used for applications that need a very simple structure but accord a slightly accuracy such as toy. The future work is how to control the threshold for further reducing the candidate words and how to improve the recognition accuracy without increasing of computational load.

ACKNOWLEDGMENT

This work has been carried out in the research project sponsored by the ministry of information communication. The authors wish to thank Mr. J.W. Yang, Dr. Y.J. Lee, and other members in the Spoken Language Section for their continuous support.

REFERENCES

Suhm, B., et al. (1995) "JANUS: Towards Multilingual Spoken Language Translation", Proceedings of Spoken Language Systems Technology Workshop, 221−226.

Kim, H.R., Hwang, K.W., Han, N.Y., and Ahn, Y.M. (1994) "Korean Continuous Speech Recognition System Using Context−Dependent Phone SCHMMs", Proceedings of SST, Vol.2, pp. 694−699.

Lee, H.S., Park, J., Kim, H.R. (1996) "An Implementation of Korean Spontaneous Speech Recognition System", Proceedings of ICSPAT'96, Oct.

Lalit R. Bahl, Raimo Bakis, Peter V.de Souza and Robert L. Mercer (1988) "Obtaining candidate words by polling in a large vocabulary speech recognition system", Proceedings of ICASSP'98, pp. 489−493.

Han, N.Y., Kim, H.R., Hwang, K.W., Ahn, Y.M., Ryoo, J.H. (1995) "A Continuous Speech Recognition System Using Finite State Network and Viterbi Beam Search for the Automatic Interpretation", Proceedings of ICASSP'95, Vol. 1, pp. 117−120.

Linde Y., Buzo A., and Gray, R. M. (1980) "An algorithm for vector quantization", IEEE Trans. on Communication, Vol.28, pp. 84−95.

Ahn, Y.M., Kim, H.R., Hwang, K.W. (1994) "Development of a Simple VQ−Based Preprocessor", Proceedings of ICSPAT'94, Vol.2, pp. 1697−1699.