New Postprocessing Methods for Rejecting Out-of-Vocabulary Words

*Myung-Gyu Song, **Kab-Jong Shim, and *Hyung-Soon Kim

Abstract

The goal of postprocessing in automatic speech recognition is to improve recognition performance by utterance verification at the output of recognition stage. It is focused on the effective rejection of out-of vocabulary words based on the confidence score of hypothesized candidate word. We present two methods for computing confidence scores. Both methods are based on the distance between each observation vector and the representative code vector, which is defined by the most likely code vector at each state. While the first method employs a simple time normalization, the second one uses a normalization technique based on the concept of on-line garbage model[1]. According to the speaker independent isolated words recognition experiment with discrete density HMM, the second method outperforms both the first one and conventional likelihood ratio scoring method[2].

Ⅰ. Introduction

In practical speech recognition applications, the capability to reject out-of-vocabulary(OOV) words is very important. When an input speech lies outside pre-defined vocabulary words, the recognition system without an appropriate postprocessing determines the best candidate word in the active vocabulary (pre-defined vocabulary of a recognition system) as the recognition output, which results in the false acceptance error. To deal with this problem, a reliable postprocessing method for rejecting less confident words is needed. A few approaches have been considered for rejecting OOV words. Those include the method using the likelihood ratio score[2], the method using artificial neural network[3], and the method using linear discriminators[4], etc.

In this paper, we present two measures of confidence score of the hypothesized candidate word. Both are based on the distance between each observation vector and the representative code vector. While the first method uses a simple time normalization technique, the second one uses a normalization technique based on the concept of on-line garbage model. In both methods, the postprocessor will reject the first candidate word if its confidence score is lower than the pre-defined rejection threshold. The performance of the rejection capability was tested for the discrete-HMM-based isolated word recognition system.

*Department of Electronics Engineering, Pusan National University, Korea
** Passenger Car E&R Center II, Hyundai Motor Company, Korea

Manuscript Received: September 23, 1997.

Ⅱ. Postprocessing Methods for Rejection

One of the most widely used techniques for rejecting OOV words is the method using first/second likelihood ratio score as a confidence measure of the first candidate word. If the difference between the accumulated likelihood of the first candidate and that of the second candidate is small, the method considers the confidence of the first candidate word low and then rejects the first candidate word. The confidence score for this method is given by:

$$D_{LRS} = WL_1 - WL_2 \tag{1}$$

where WL_1 is the accumulated log likelihood of the first candidate word and WL_2 is that of the second candidate word. Since large D_{LRS} value indicates high confidence, the decision rule is given as follows:

If $D_{LRS} <$ RTH Reject the first candidate word
If $D_{LRS} >$ RTH Accept the first candidate word

where **RTH** is predefined rejection threshold. In this method, the rejection capability is highly dependent on the contents of the active vocabulary set. Indeed, if the active vocabulary set consists of similar words and on input speech is one of them, it will be likely to reject the first candidate word, since both the first and second candidate words may have about the same accumulated likelihood value.

To avoid the vocabulary dependency of the confidence measure, one may employ only the highest accumulated likelihood value or its normalized version by the length of the frame as a measure of confidence. But these measures

also lack consistency in separating the active vocabulary words. Fig. 1 shows a few examples of the local likelihood contours along the Viterbi decoded state sequence for the discrete-HMM-based isolated word recognition. In these examples, the solid line indicates the local likelihood contour for the correct word model and the dotted line indicates that of the best candidate word model except the correct one. Fig. 1(a) shows the accumulated likelihood value represents the confidence of the model properly. But (b) and (c) of the same figure show some local likelihood values fall into the floored minimum probability even in a correctly recognized case, which is mainly caused by the insufficient training data. There are several approaches to handle this problem without increasing the training data[5], but fundamental cure is impossible due to quantization error effect of discrete HMM. As a result, the accumulated likelihood value does not represent the confidence of the model appropriately.

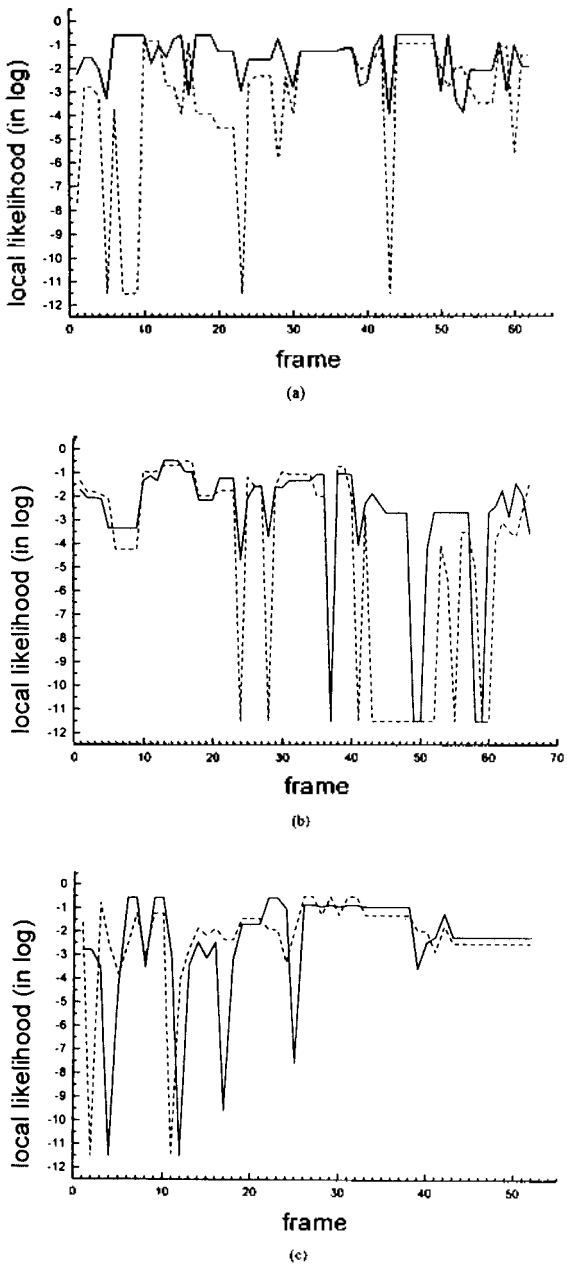

Figure 1. Examples of the local likelihood contours. Solid line for the correct word model, dotted line for the best candidate word model except the correct word model. (a) a correctly recognized example in training data. (b) a correctly recognized example in test data. (c) an incorrectly recognized example in test data.

To alleviate this problem, we introduce a new confidence measure, which is based on the distance between each observation vector and the representative code vector. We define the representative code vector at each state as the code vector which has the maximum probability at the corresponding state. In this method, the confidence score of the first candidate word is computed as:

$$D_1 = \frac{\sum_{t=1}^{T} d(x_t, C_{q_t})}{T} \tag{2}$$

where x_t is t-th frame observation vector, C_{q_t} is the representative code vector of the Viterbi decoded state at t-th frame, T is the frame length of input speech (the length of observation sequence), and $d(\cdot,\cdot)$ is the Euclidean distance between two vectors. We decide if the first candidate word is rejected or not by comparing the confidence score defined in (2) to the rejection threshold(RTH) as follows:

If $D_1 >$ RTH Reject the first candidate word
If $D_1 <$ RTH Accept the first candidate word

Notice that large D_1 value indicates low confidence. Increasing the rejection threshold leads to increase the number of false acceptance of OOV words. Conversely, decreasing the rejection threshold leads to increase the number of false rejection (incorrect rejection of acitve vocabulary). We call this method as the proposed method A.

In (2), we use a simple time normalization technique based on the length of observation sequence. We propose another normalization technique using the concept of online garbage model[1]. This normalization technique was introduced to emphasis the superiority of the nearest code vector over the N nearest code vector candidates corresponding to an observation vector. We use the accumulated average distance between an observation vector and N nearest code vectors corresponding to it as a normalization factor. The confidence score adopted for this new

normalization technique is computed as:

$$D_2 = \frac{\sum_{t=1}^{T} d(x_t, C_{q_t})}{\sum_{t=1}^{T} \left| \frac{1}{N} \sum_{i=2}^{N+1} d(x_t, C_{i_t}) \right|} \quad (3)$$

where x_t, C_{q_t}, and T are the same as in (2), and C_{i_t} is the i-th nearest code vector of x_t, N is the number of code vector used for normalization, which is to be optimized by the experimental results. The same decision rule used for D_1 is used for D_2. We call this method as the proposed method B.

Ⅲ. Experiments and Results

The speech corpus used in the experiments contains 49 isolated words. Speech was collected from 59 male speakers and sampled at 8 kHz. The whole corpus has been split into two sets: a 25 active vocabulary words set and the rest 24 OOV words set. Training has been performed over 49 speakers and test to assess the rejection capability over 10 speakers.

Discrete HMM for each active vocabulary word was modeled using HTK V2.0[6]. The number of states of each word model is determined as three times the number of phonemes of the word. The size of VQ codebook used is 64. The observation vectors are 12 MFCCs computerd every 10ms with an analysis window of 20ms. The number of filter banks used for computing MFCC is 26. The error rate of the recognition system (with no rejection) is 2.73%.

To assess the performance of speech recognition system in terms of rejection capability, we use two curves. One is the OOV rejection rate as a function of the false rejection rate, the other is the recognition accuracy of non rejected active vocabulary words as a function of that. Each point of the curve corresponds to a particular rejection threshold (RTH). At first, we evaluated the rejection performance of the proposed method B over several N values. Fig. 2 shows that the rejection performance is not sensitive to N. We set $N = 5$ to maximize the rejection on OOV when the rejection rate on the vocabulary varies from 2% to 5%.

Then we compared the performance of proposed rejection methods to that of a conventional rejection method using first/second likelihood ratio score and the results are shown in Figs. 3 and 4. From Fig.3, we can see that the proposed method B outperforms the others. For example, if the rejection rate of active vocabulary words is allowed to 4%, the proposed method B can reject 70% of OOV

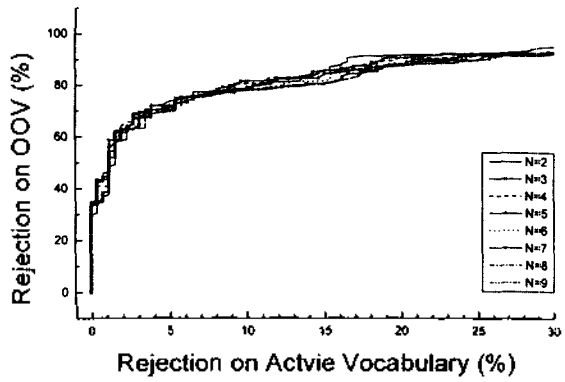

Figure 2. Rejection performance of the proposed method B over several N values

while the proposed method A 52% and the conventional method only 30%. As a result, the performance of the proposed method B is much better than that of proposed method A as well as that of the conventional rejection method using first/second likelihood ratio score. On the other hand, it can be seen from Fig.4 that recognition accuracy with proposed rejection methods is slightly lower than that with conventional method. For example, when the rejection rate of active vocabulary word is allowed to 4%, the difference in the recognition accuracy is about

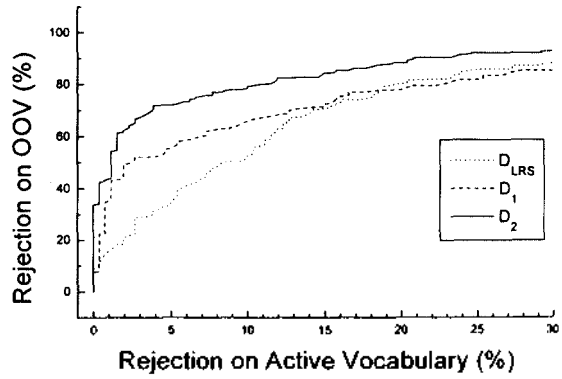

Figure 3. OOV rejection versus active vocabulary rejection

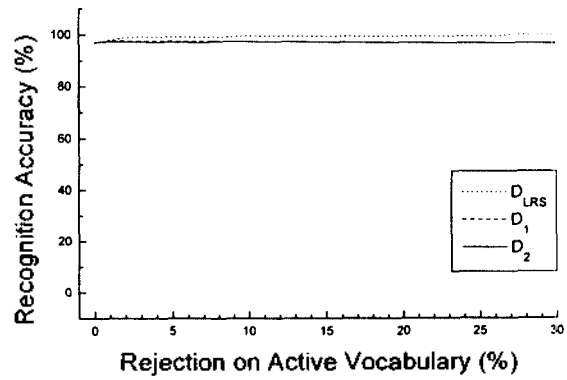

Figure 4. Accuracy on non-rejected active vocabulary versus active vocabulary rejection

2%. However, it should be noted that, in real situations where considerable amount of OOV words is involved, the powerful OOV rejection capability of the proposed method B compensate for its slightly low recognition accuracy.

Finally, we evaluated the vocabulary dependency of the rejection performance by changing the active vocabulary set. Fig. 5 is the result of the proposed method B and Fig. 6 is that of the conventional first/second likelihood ratio scoring method. These figures indicate that the proposed method B is less dependent on active vocabulary set than the conventional rejection method.

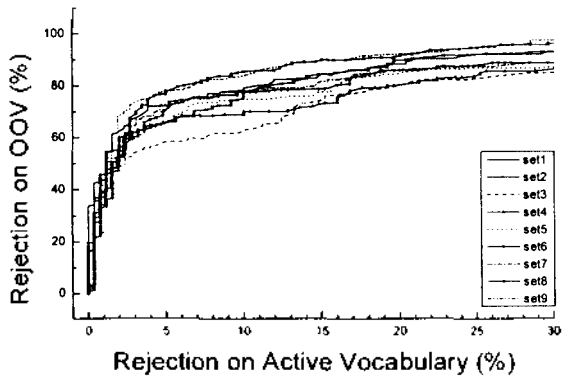

Figure 5. Vocabulary dependency of the rejection performance for the proposed method B

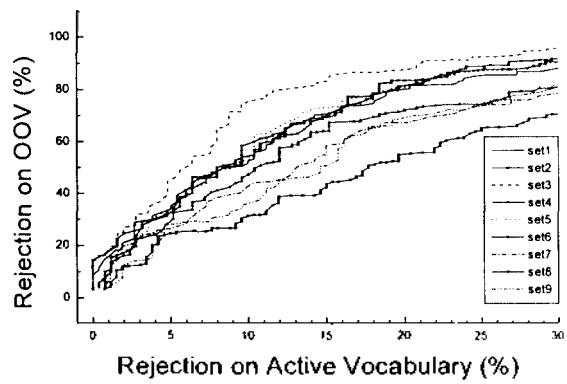

Figure 6. Vocabulary dependency of the rejection performance for the conventional method

Ⅳ. Conclusions

In this paper, we proposed two postprocessing methods for rejecting out-of-vocabulary words. Both methods are based on the distance between each observation vector and the representative code vector of each state corresponding to it. While the first method employs a simple time normalization, and the second one uses a normalization technique based on the concept of the on-line garbage model. According to the speaker-independent isolated word recognition experiment, the latter is superior to both the former and the conventional rejection method using first/second likelihood ratio score. Moreover, we can show that the rejection performance of the proposed method B is less dependent on the active vocabulary set than that of the conventional method.

In the proposed methods, we defined a representative code vector as the code vector which has the maximum probability at each state. But the validity of the representative code vector still has problem, especially when several code vectors are competing for the representative one. Our current research includes the introduction of multiple representative code vectors and combining the proposed normalization technique with the conventional confidence measure, for example, the likelihood ratio score.

References

1. J. M. Boite, H. Bourland, B. D'hoore, and M. Haesen, "A new approach towards keyword spotting," *Eurospeech'93*, pp. 1273-1276, 1993.
2. R. C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system," *ICASSP'90*, pp. 129-132, 1990.
3. D. P. Morgan, et al., "A keyword spotter which incorporates neural networks for secondary processing," *ICASSP'90*, pp. 113-116, 1990.
4. R. A. Sukkar, et al., "A two pass classifier for utterance rejection in keyword spotting," *ICASSP'93* pp. 451-454, 1993.
5. K. F. Lee and H. W. Hon, "Speaker independent phone recognition using hidden markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, pp. 1641-1648, Nov. 1989.
6. S. J. Young, et al., *The HTK Book*, Cambridge University and Entropic Research Laboratories Ltd., 1996.

▲Myung Gyu Song

Myung Gyu Song was born in Pusan, Korea, on October 14, 1971. He received the B.S. degree in Electronics Engineering at Pusan National University, Pusan, Korea, in 1996. From March 1996, he has been a student in M.S. course in Electronics Engineering at Pusan National University. His current research interests include speech recognition, speech synthesis and speech coding.

▲Kab Jong Shim

Kab Jong Shim received his B.S. and M.S. degree in electronics engineering from KangWon National University, Korea, in 1989 and 1991, respectively. He has been working for Research & Development Division in Hyundai

Motor Company since 1991 and was engaged in the research and development of automotive electronic systems. His current research interests include speech recognition, synthesis and intelligent human interface systems in vehicle.

▲Hyung Soon Kim

Hyung Soon Kim received the B.S. degree in electronic engineering from Seoul National University in 1983, and the Ph.D. degree in electrical and electronic engineering from the Korea Advanced Institute of Science and Technology(KAIST) in 1989.

From 1987 to 1992, he was with Digicom Institute of Telematics, where he was Technical manager of the Speech Communication Division. Since 1992, he has been with the faculty of the Department of Electronics Engineering at Pusan National University, and is an Assistant Professor. His research interests include digital signal processing, speech recognition and speech synthesis.