AN EFFICIENT KEYWORD SPOTTING TECHNIQUE USING A COMPLEMENTARY LANGUAGE FOR FILLER MODELS TRAINING

Panikos Heracleous, Tohru Shimizu

KDDI R&D Laboratories, Inc, Japan
Email:{panikos, shimizu}@kddilabs.jp

ABSTRACT

The task of keyword spotting is to detect a set of keywords in the input continuous speech. In a keyword spotter, not only the keywords, but also the non-keyword intervals must be modeled. For this purpose, filler (or garbage) models are used. To date, most of the keyword spotters have been based on hidden Markov models (HMM). More specifically, a set of HMM is used as garbage models. In this paper, a two-pass keyword spotting technique based on bilingual hidden Markov models is presented. In the first pass, our technique uses phonemic garbage models to represent the non-keyword intervals, and in the second stage the putative hits are verified using normalized scores. The main difference from similar approaches lies in the way the non-keyword intervals are modeled. In this work, the target language is Japanese, and English was chosen as the 'garbage' language for training the phonemic garbage models. Experimental results on both clean and noisy telephone speech data showed higher performance compared with using a common set of acoustic models. Moreover, parameter tuning (e.g. word insertion penalty tuning) does not have a serious effect on the performance. For a vocabulary of 100 keywords and using clean telephone speech test data we achieved a 92.04% recognition rate with only a 7.96% false alarm rate, and without word insertion penalty tuning. Using noisy telephone speech test data we achieved a 87.29% recognition rate with only a 12.71% false alarm rate.

1. INTRODUCTION

The task of keyword spotting is to detect a set of keywords (single or multiple keywords) in the input continuous speech. In some applications the keyword spotting may have an important role. Especially, in applications based on telephone speech recognition the performance can be increased by implementing a keyword spotting technique, compared to when the syntax contains only keywords without garbage models. However, usually in such applications, the input speech includes out-of-task words, noise components, mobile phone speech or spontaneous speech characteristics. By introducing a keyword spotting technique these phenomena can be rejected, and therefore the system allows users the flexibility to speak naturally. The final goal of this study is to develop a keyword spotting method for telephone speech.

The problem of keyword spotting has been approached in several ways. Bridle approaches the problem by introducing dynamic programming techniques for whole word templates [1]. Higgins et al., introduces a continuous speech recognition approach to keyword spotting, and they also define filler templates to represent the non-keyword portions [2]. Finally, Rose et al. introduces the HMM based keyword spotting [3].

In a keyword spotter not only the keywords, but also the non-keywords or noise components must be explicitly modeled. Although several approaches exist for modeling these intervals, the most common ones are based on HMM. In such approaches, a set of HMM (garbage or filler models) is chosen to represent the non-keyword intervals [3, 4, 5, 6].

The performance of an HMM based keyword spotter heavily depends on the ability of the garbage models to represent non-speech intervals, without rejecting the correct keyword hits (false rejections). Therefore, the choice of an appropriate garbage model set is a critical issue. The most common approaches are as follows:

- The training corpus for a specific task is split into keyword and non-keyword (extraneous) data. [3, 4].

  The keywords are represented by HMM trained using the keyword speech, and the garbage models are trained using the extraneous speech. The main disadvantage of such approaches is the task-dependency. Model retraining is required when the vocabulary changes. Moreover, the training data must include a large number of keyword occurrences for robust training.

- The garbage models are selected from a set of common acoustic models [3].

  In this case, a speech corpus for a separate task is used to train only one common acoustic set. A subset of this set is used as garbage models. A typical case is to represent the keywords by context-dependent HMM, and the non-keyword portions by context-independent HMM. In some works, a subset of context-dependent models is also used as garbage models [3, 4].

  The main disadvantage of such methods is the high rate of false rejections (percentage of true keywords rejected). However, since the syntax allows any garbage models sequence, the keywords are also included in these sequences. The overlapping of contexts causes garbage models to be decoded instead of keywords.

Most keyword spotters use two sets of acoustic models trained using keyword and non-keyword speech, respectively. Rose and Paul suggest that the performance of a keyword spotter may be increased by training phonemic garbage models using a large corpus of non-keyword speech [3].

In this paper, we propose a novel method for modeling the non-keyword intervals based on the use of bilingual hidden Markov models. In our method, we use garbage models trained using a speech corpus of a language other than the target language. Our goal is to develop a task-independent keyword spotter, and to overcome the problem of the overlapping of contexts.
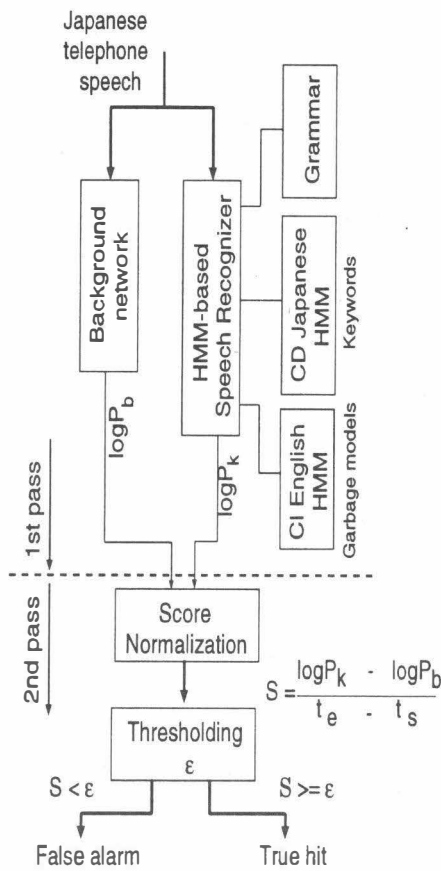
Fig. 1. Block diagram of the system

2. PROPOSED KEYWORD SPOTTING APPROACH

The main requirement in our approach is acoustic similarity between the target and 'garbage' languages. Instead of analytical comparison, we compared the two languages based on the International Phonetic Alphabet (IPA) [7]. The IPA has been developed by the International Phonetic Association, and is a set of symbols which represents the sounds of language in written form. Previously, in some studies [9] the IPA was also used for selecting a common phoneme set for multilingual speech recognition. In those studies, the obtained results showed the efficiency of using the IPA. Based on IPA, American English acoustically covers the Japanese language efficiently. Therefore, the choice of English is reasonable, and the English HMM garbage models - trained from a large speech corpus of guaranteed non-keyword speech - are expected to represent the non-keyword intervals without rejecting the true keyword hits. However, HMM trained using a database of American English efficiently covers the non-keyword intervals, but not the keyword portions.

Figure 1 shows the block diagram of the keyword spotting system based on bilingual HMM. As can be seen, our approach is a two-pass keyword spotting technique. In the first pass, English garbage models are connected with Japanese keywords. In the
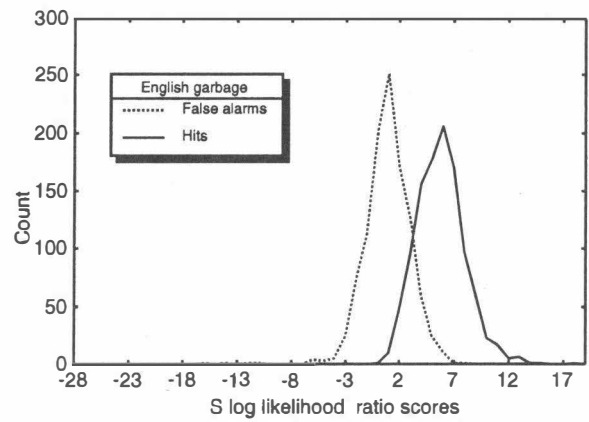

Fig. 2. Histogram of log likelihood ratio scores

second pass, the putative hits are verified using log likelihood ratio scores [3, 8] as follows:

$$S = \frac{logP_k - logP_b}{t_e - t_s} \qquad (1)$$

$$S < \epsilon, \text{ putative hit is rejected} \qquad (2)$$
$$S \geq \epsilon, \text{ putative hit is accepted},$$

where $logP_k$ is the log likelihood of the putative hit, $logP_b$ is the log likelihood of the alternate hypothesis given by the background network, t_s is the starting time of the putative hit, t_e is the ending time of the putative hit, and ϵ is the threshold that must be adjusted. The background network is composed of garbage models connected to form syllables as in Japanese language. The garbage models which are used in the recognizer are used in the background network, too. The sequence of the decoded garbage models, which overlaps the decoded putative keyword hit is used to provide the alternate hypothesis. Using background network, we can account for the variabilities in time of the keyword scores, and the decision for separating true keyword hits from false alarms is more reliable.

Figure 2 shows the histogram of the log likelihood ratio normalized scores. As can be seen, the true hits and the false alarms are very well separated. Therefore, a threshold can be established to provide a trade-off between the recognition rate and rejection rate.
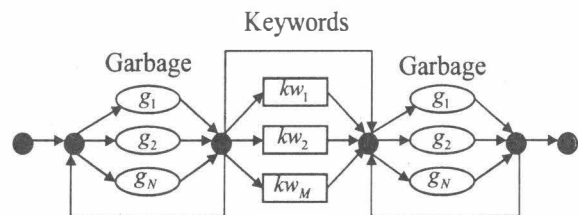

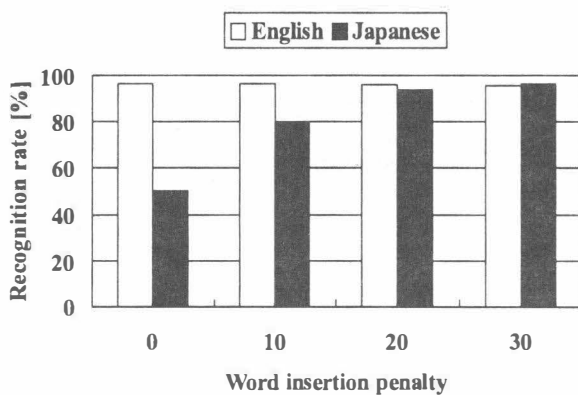**Fig. 3**. Recognition network for keyword spotting system
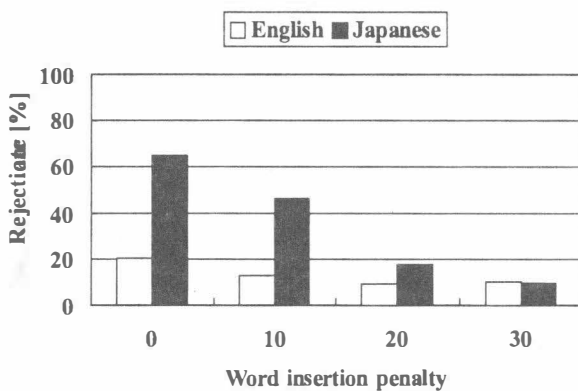
Fig. 4. Recognition rates using clean test data

Fig. 5. Rejections rates using clean test data

Fig. 6. Performance using English garbage models (clean test data)
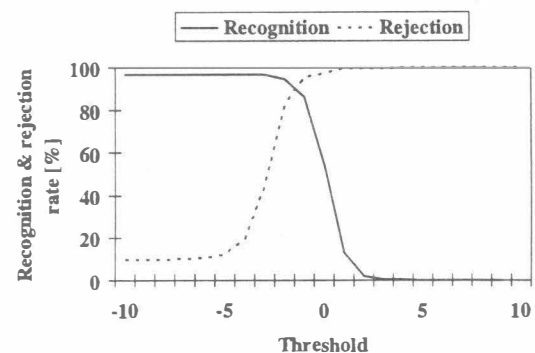
Fig. 7. Performance using Japanese garbage models (clean test data)

3. ACOUSTIC MODELS - SYNTAX

The Japanese keywords are represented by gender-dependent, context-dependent HMM. The models are trained using a large corpus of telephone speech (fixed line speech and mobile telephone speech). The feature vectors are of size 38 (12 MFCC + 12 ΔMFCC + 12 $\Delta\Delta$MFCC + ΔEnergy + $\Delta\Delta$Energy). A set of 28 context-independent, 3-state single Gaussian HMM trained using the same speech corpus is chosen as the Japanese garbage models for comparison purposes. The English garbage models are represented by context-independent, 3-state single Gaussian HMM. Twenty-eight models trained using the MACROPHONE American English telephone speech corpus are used. Figure 3 shows the syntax, which allows at most one keyword per utterance.

4. DEFINITIONS OF EVALUATION MEASURES

For the evaluation, the following three measures are used:

- Recognition Rate (RCR) - The percentage of keywords detected.
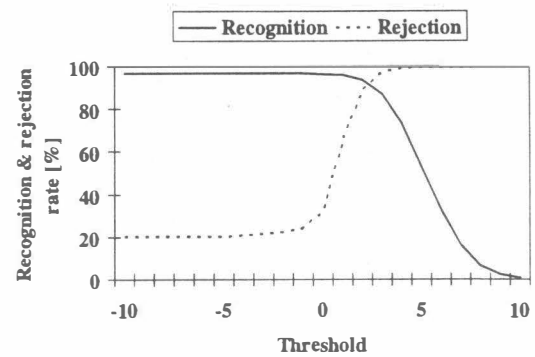- Rejection Rate (RJR) - The percentage of non-keywords rejected.
- Equal Rate (ER) - It shows equal RCR and RJR.

5. EXPERIMENTS AND RESULTS

In these experiments, the test set consists of Japanese telephone speech recorded in a clean environment, and it contains 1,548 short utterances (of which only 1,133 contain a keyword). In total, the test set contains 7740 non-keywords. The vocabulary consists of 100 keywords (country names), and the grammar allows at most one keyword per utterance.

Figures 4 and 5 show the results after the first pass. As can be seen, the Japanese garbage models cause a high rejection rate and low recognition rate. The recognition rate is increased by tuning the word insertion penalty. However, the rejection rate is drastically decreased.

The figures show also that by using English garbage models, the word insertion penalty appears not to have a significant effect on the recognition rate. Although in both cases the RCR is high, the number of false alarms is still large. To reduce the false alarms, a posteriori thresholding based on the normalized log likelihood ratio scores is used in the second pass. Figures 6 and 7 show the performance of the system. Using Japanese garbage models,
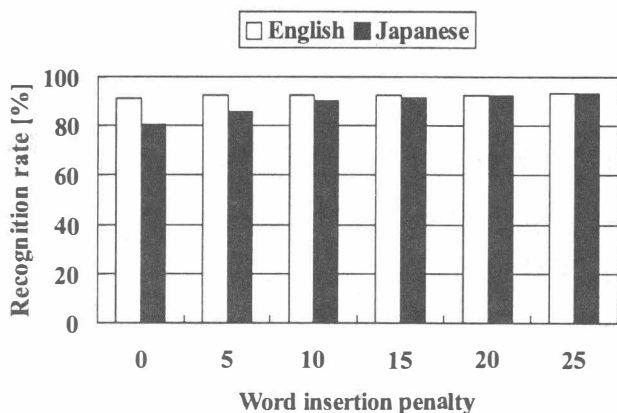
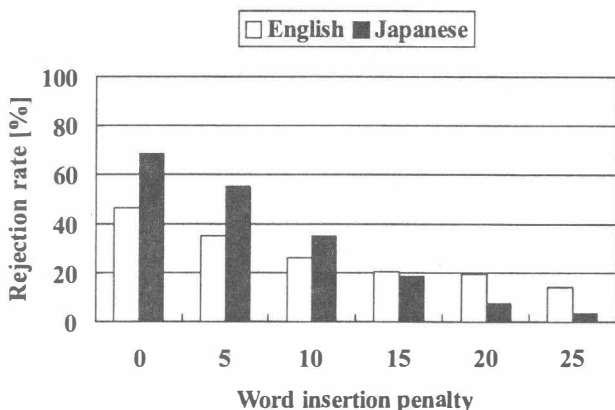**Fig. 8**. Recognition rates using noisy test data

Fig. 9. Rejections rates using noisy test data

the word insertion penalty is tuned in the first pass to achieve the maximum recognition rate (word insertion penalty = 30). As we can see, by adjusting the threshold, the rejection rate is drastically increased (the false alarms are decreased) without significant decrease in the recognition rate. With English garbage models, we achieved a 92.04% equal rate, with Japanese garbage models, 91.57%. The results show the benefits of using English garbage models. Using English garbage models, we achieved higher performance, without word insertion penalty tuning.

Figures 8 and 9 show the results obtained using noisy data. In these experiments, the test set consists of mobile phone speech collected under noisy conditions, and contains 686 short utterances (of which only 564 contain a keyword). In total, the test set contains 2164 non-keywords. The vocabulary consists of 100 keywords, and the grammar allows at most one keyword per utterance. The results show that using English garbage models, the improvement in the recognition rate is not significant, even after tuning the word insertion penalty. With Japanese garbage models, word insertion penalty tuning is necessary. In both cases, however, by tuning the word insertion penalty, the rejection rate is drastically decreased. Using English garbage models, we achieved after thresh-

olding and without word insertion penalty tuning a 85.64% equal rate. After word insertion penalty tuning (maximum recognition rate), we achieved a 87.29% equal rate. Using English garbage models, our system performed better than when using Japanese garbage models (85.57% after word insertion penalty tuning), even without word insertion penalty tuning.

6. CONCLUSION AND FUTURE WORK

This paper presents a novel two-pass keyword spotting technique. The proposed technique is based on hidden Markov models, and uses as garbage models phonemic HMM trained using a speech corpus of a language other than the target language. In our study, the target language is Japanese, and English was chosen as the 'garbage' language. The main advantage of our method is the task-independency, and also parameter tuning (e.g. word insertion penalty) does not have a serious effect on the performance. The method was evaluated through experiments on clean and noisy telephone speech data. The results showed the effectiveness of our method compared to a conventional method. For a vocabulary of 100 keywords and using clean test data we achieved a 92.04% equal rate. Using noisy telephone speech data and a vocabulary of 100 keywords we achieved a 87.29% equal rate. In both cases, the achieved results are very promising and show the effectiveness of our proposed method. In a future study, we plan to evaluate our method using larger vocabularies.

7. REFERENCES

[1] J. S. Bridle, "An Efficient Elastic-template Method for Detecting Given Words in Running Speech," *Brit. Acoust. Soc. meeting*, pages 1–4, 1973.

[2] A. L. Higgins and R. E. Wohlford, "Keyword Recognition Using Template Concatenation," *Proc. ICASSP*, pages 1233-1236, 1985.

[3] R. C. Rose and Douglas B. Paul, "A Hidden Markov Model Based Keyword Recognition System," *Proc. ICASSP*, pages 129–132, 1990.

[4] H. Bourlard, B. D'hoore and J. Boite, "Optimizing Recognition and Rejection Performance in Wordspotting Systems," *Proc. ICASSP*, pages 373–376, 1994.

[5] K. M. Knill and S. J. Young, "Fast Implementation Methods for Viterbi-based Word-spotting," *Proc. ICASSP*, pages 522–525, 1996.

[6] A. S. Manos and V.W. Zue, "Segment-based Wordspotter Using Phonetic Filler Models," *Proc. ICASSP*, pages 899–902, 1997.

[7] "Handbook of the International Phonetic Association," *Book*, pages 41–44 and 117–119, Cambridge University Press, 1999.

[8] T. Kawahara, C. H. Lee, and B.H. Juang, "Combining Key-phrase Detection and Subword-based Verification for Flexible Speech Understanding," *Proc. ICASSP*, pages 1159–1162, 1997.

[9] T. Schultz and A. Waibel, "Multilingual and Crosslingual Speech Recognition," *Proc. DARPA Broadcast News Workshop*, 1998.