A HIDDEN MARKOV MODEL BASED KEYWORD RECOGNITION SYSTEM[1]

Richard C. Rose and Douglas B. Paul

Lincoln Laboratory, MIT
Lexington, MA 02173–9108

ABSTRACT

A speaker independent hidden Markov model (HMM) keyword recognizer (KWR) based on a continuous speech recognition (CSR) model is presented. The paper describes the baseline keyword recognition system, and discusses techniques for dealing with non–keyword speech and linear channel effects. The training of acoustic models to provide an explicit representation of non–vocabulary speech is investigated [2]. A likelihood ratio scoring procedure is used to account for sources of variability affecting keyword likelihood scores [3]. Finally, an acoustic class dependent spectral normalization procedure is used to provide explicit compensation for linear channel effects. Keyword recognition results for a standard conversational speech task with a 20 keyword vocabulary reached 82% probability of detection at a false alarm rate of 12 false alarms per keyword per hour.

1 INTRODUCTION

The problem of detecting a given set of words in running speech has been approached in several different ways. Dynamic programming techniques for whole word template matching were originally introduced for keyword recognition by Bridle [1]. In these systems, a score is computed for every keyword template matched against every portion of the input. Each dynamic programming path taken to be a possible occurrence of a keyword is a putative hit. A second stage is required to remove overlapping putative hits, and normalize path likelihood scores so that thresholds can be established for separating true keyword hits from false alarms.

Following the reasoning of Higgins and Wohlford, we take a CSR approach to keyword recognition [2]. Higgins and Wohlford used a template based dynamic time warping connected speech recognition system to match a sequence of templates to an input utterance. They define "filler templates" to represent out of vocabulary or "non–keyword" speech. The output of such a system is a continuous stream of keyword and filler templates, and the occurrence of a keyword template in this output stream is taken as a putative hit. The advantage of such a system is that removal of overlapping putative hits is handled implicitly. However, the performance of the CSR based KWR relies heavily on the ability of the filler templates to match arbitrary speech.

The HMM KWR described in this paper also uses a filler model approach. A particular advantage of using a hidden Markov model representation of acoustic filler models, is that maximum likelihood training of statistical hidden Markov models allows acoustical filler models to assimilate information over many different speakers and word contexts. Therefore, it is reasonable to assume that a HMM based filler model system may do a better job of modeling arbitrary speech than a template based system. There have been previous implementations of HMM keyword recognition systems including that reported on by Rohlicek et. al. [5]. However, to our knowledge, the system described in this paper is the first HMM KWR based on a continuous speech recognition model.

The principle focus in this paper is on the problem of separating keyword speech from a background of non–keyword speech. While it is always assumed that robust estimation of keyword model parameters are important for this task, the emphasis in this paper is on investigating techniques for modeling background speech. The paper is organized as follows. Section 2 describes the baseline HMM KWR. The keyword recognition paradigm, including training and evaluation databases and the figure of merit used to represent system performance, is described in Section 3. Section 4 describes the experimental results, and, finally, discussion and summary is provided in Section 5.

[1]This work was sponsored by the Defense Advanced Research Projects Agency.

2 KEYWORD SPOTTING SYSTEM

2.1 Baseline System

The baseline keyword recognition system is intended as an initial version of a system capable of large vocabulary speaker independent keyword recognition in conversational speech messages over varied channel conditions. Originally derived from the speaker independent version of the Lincoln Continuous Speech Recognizer [4], continuous Gaussian observation subword hidden Markov models are used in the system. Observations are in the form of mel–frequency cepstra and difference cepstra. Keywords are represented by subword models to allow for very large keyword vocabularies, and also to allow for robust training of keyword variants that do not appear in training. The subword hidden Markov models are trained from orthographically transcribed speech, so hand labeling of keyword endpoints is not needed in training.

The recognizer itself is a null grammar time synchronous Viterbi beam search decoder consisting of a parallel network of keywords and fillers. Figure 1 shows a network consisting of N keywords and M fillers. Keywords are instantiated from three state, linear subword acoustic models. Fillers are represented in a number of different ways, and are described below. The operating point of the system can be adjusted by the settings of the interword transition weights, $W_{k,1},\ldots,W_{k,N}$ for keywords, and $W_{f,1},\ldots,W_{f,M}$ for fillers. Operating point in this context simply refers to the trade–off between the number of keyword misses and false alarms. The optimum choice of these weights is obtained by an empirically derived tradeoff, adjusting the operating point to maximize the performance criteria described in Section 3.

The score that is reported for each keyword decoded in the Viterbi network is a duration normalized word likelihood. The score, S_W^{KW}, for a keyword, KW, decoded for observations within the interval from time T_I to T_F with terminal state, s_F, is given as

$$S_{KW} = \frac{\log P(s_F, y_{T_I}, \ldots, y_{T_F})}{T_F - T_I}, \qquad (1)$$

where y_t is a cepstrum observation vector.

2.2 Partial Viterbi Backtrace

The input to the KWR is a continuous running utterance, and putative keyword hits must be reported with a minimum of delay even though there are generally many Viterbi paths active in the network at all times. Fortunately, a partial Viterbi backtrace can be used to locate

CH2847-2/90/0000-0129 $1.00 © 1990 IEEE

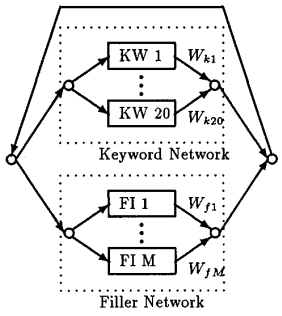

Figure 1: **The keyword–filler network**

states where all paths agree at a given time, and thereby locate an initial portion of the best scoring path [10]. This involves tracing back through active trellis nodes and identifying convergent nodes, or those trellis nodes from which extend only a single active path. The optimum sequence of fillers and keywords along this path can then be reported.

The advantage of the partial backtrace procedure is that putative hits are obtained from the optimum maximum likelihood path through the network, instead of from observed local maxima in likelihood scores. Since the partial backtrace procedure produces a continuous stream of keywords and fillers, putative hit overlap removal is performed implicitly. The potential disadvantage of the partial backtrace procedure is the delay from the instant when a keyword occurs to the instant when all active paths converge, and the keyword can be decoded. These network delays have been observed for typical utterances with a wide range of network pruning thresholds. In all examples observed, the network delays were less than three seconds.

3 KEYWORD SPOTTER TRAINING and EVALUATION

All keyword recognition results reported in this paper are for the same speaker independent keyword recognition paradigm. Acoustic model training is performed using read speech, and the KWR is evaluated using conversational speech oriented to the solution of an artificial task. There are a number of problems with this paradigm. First, there are fundamental differences in speaking style between read and conversational speech. Analysis of the data showed the variance of keyword duration for conversational speech to be roughly four times that for keywords occurring in read speech. Second, non–grammatical speech events are present in conversational speech that are not present in read speech. It is possible that the availability of labeled occurrences of these events for training background speech models would improve keyword recognition performance.

All keyword recognition results are reported for a continuous utterance consisting of excerpts from conversations for 8 male speakers. There is a twenty word keyword vocabulary. The total test utterance duration is 22 minutes, and there are a total of 353 keyword occurrences. The cepstrum observation vectors were obtained from the band energies of a mel–frequency spaced filter bank front–end [7]. All data was band limited to approximately 3200 Hz. The system was trained using orthographically transcribed read sentences and paragraphs from 14 male speakers, containing an average of 59 occurrences per keyword. The non–keyword speech in these training utterances is used for training filler acoustic models.

The system is evaluated using an automatic scoring procedure on a database containing hand–labeled keywords in the input utterance. A score is associated with each keyword occurrence, and these scores are used to compute a receiver operating curve relating the probability of detection to the number of false alarms per keyword per hour (fa/kw/hr). The results reported for this database are given as the probability of detection averaged over false alarm rates from 0 to 10 fa/kw/hr. The histogram in Figure 2 illustrates the keyword specific results for a typical system. The results are given as the total probability of detection for each keyword over the entire test utterance. The total number of false alarms that occurred for each keyword is indicated vertically along the histogram bars. The overall P_d for this system was 72% at a false alarm rate of 4.7 fa/kw/hr, while the average P_d over 0 to 10 fa/kw/hr was 64%.

4 EXPERIMENTAL RESULTS

4.1 Filler Models

Explicit modeling of non–vocabulary speech using filler acoustic models is motivated by the fact that better modeling of non–keywords will reduce the probability of false keyword detection. Of course, the robust estimation of keyword model parameters is just as important; however, in the following discussion it is assumed that keywords are represented as described in Section 2.1, and emphasis is placed on filler models. The use of many different types of fillers was investigated for the keyword recognition task described in Section 3. The different types of fillers, the motivation for using them, and the inherent drawbacks of each are outlined below. Keyword recognition performance for different sets of filler models is summarized in Table 1 using the the figure of merit described in Section 3. The relative computational complexity is indicated in column two of the table by the number of network states needed to represent the entire set of fillers for a particular system.

Acoustic word models It has been suggested that the best approach to keyword recognition is to design a very large vocabulary speech recognizer where the majority of the speech recognition vocabulary is chosen to model non–keyword speech [1]. Such a system would require a great deal of training data, and would also result in an extremely large recognition network. Despite these drawbacks, the use of acoustic word models was investigated here as a benchmark for comparing to lower complexity systems. In the training corpus used for this task, there were only 80 non–keyword vocabulary words (a total of 804 network states). The keyword recognition performance obtained using the network of Figure 1 with 80 word models trained from transcribed speech as fillers is given as the *Word Models* system in Table 1. The performance is actually quite remarkable when one considers how limited the set of 80 non–vocabulary models is with respect to the unconstrained vocabulary in conversational speech.

Acoustic subword models With a very large non–keyword vocabulary and a limited training corpus, the vast majority of words in the input utterance never appear in the training corpus. At a given level of computational complexity, keyword recognition performance should benefit from the use of filler models that can be shared across similar contexts. To illustrate this point, the triphone models (phone models defined in the context of adjacent left and right phones) forming the 80 non–vocabulary word models were used independently as fillers. In this system the filler model network of Figure 1 contains 268

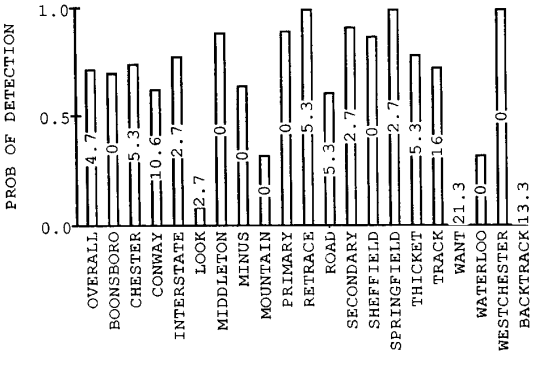

Figure 2: **Example of keyword specific performance**

Source of Fillers	Number of Filler States	Ave. P_d (%) 0–10 fa/kw/hr
Clustered Models	128	41.4
Word Models	804	59.2
Triphone Models	804	61.3
Monophone Models	135	60.6
VI Models (17 keywords)	135	44.0

Table 1: **Performance for different filler model networks**

triphones, and each triphone is represented by a 3 state, linear HMM. The performance of this *Triphone Models* system given in Table 1 represents a significant improvement over using word models as fillers. Not surprisingly, keyword recognition performance degraded when triphone models from the actual keyword vocabulary were used in the filler network. The overlapping contexts caused filler models to be decoded in place of keyword models in many instances, resulting in a large number of missed keyword occurrences.

Using a small number of general context monophones in the filler network of Figure 1 could significantly reduce the recognizer complexity. A system using monophone filler models trained from non-keyword speech is shown as the *Monophone Models* system in Table 1. While the performance of this system is slightly degraded from that of the triphone filler model system, its inherent simplicity makes it attractive from an implementational point of view.

Unsupervised clustering Training filler models from orthographically transcribed speech is labor intensive, making it difficult to reconfigure the system to a new keyword recognition task. A much easier way to train filler models would be to use unsupervised clustering of unlabeled observations from conversational speech utterances. A simple experiment was devised to investigate such an approach. The Gaussian observation means for 128 single state filler models were obtained as the cluster centroids of the Kmeans algorithm using a Malanobis distance measure. The clustering algorithm was initialized using the splitting procedure of Linde, Buzo, and Gray. HMM transition probabilities for the single state filler models were chosen so that each filler model would have an expected duration of 200 msec.

Keyword recognition performance using these clustered filler models is shown as the *Clustered Models* system in Table 1. The generally poor keyword recognition performance is due to the large number of false alarms decoded by the system, indicating that filler models trained from unsupervised clustering provide a relatively poor model of background speech. While it may be possible to obtain better performance using ergodic hidden Markov models or segment models trained from unlabeled observations, it is unlikely that such unsupervised training procedures will improve over those based on transcribed utterances.

Vocabulary independent models Vocabulary independent or task independent training of a KWR implies that there need be no retraining of subword models as the keyword vocabulary changes [5]. The existence of subword models trained from a large speech corpus associated with a completely separate task is assumed. In a preliminary experiment, the triphone subword models trained from the speaker independent DARPA Resource Management (RM) Database were used directly in the keyword recognizer. There are a total of 2430 triphones trained from a total of about 3 hours of speech with 72 speakers.

Roughly 60% of the keyword triphones were covered in this vocabulary independent data. For those triphones in the keyword vocabulary that were not not covered, the triphone was approximated by averaging over the monophone class. Monophone fillers were obtained by averaging the RM triphones over the monophone class. The detection and false alarm rate for nearly half of the keywords were only slightly degraded over using vocabulary specific models. The overall performance excluding 3 monosyllabic keywords, shown as the *VI Models* system in Table 1, resulted in a significant decrease in overall performance over the comparable system trained with vocabulary specific models. However, it was promising that the performance for many of the keywords in Figure 2 remained relatively unchanged.

4.2 Likelihood Ratio Scoring

In addition to using filler acoustic models, a modified scoring procedure can also be used to deal with non-keyword speech [6]. The keyword likelihood scores often exhibit variability in time, making it difficult to assign reliable decision regions for separating true keyword hits from false alarms. In order to account for these variabilities, a parallel "background" network of filler models is included as shown in Figure 3. The standard score of Equation 1 can be computed for the sequence of background fillers that overlaps a decoded keyword. A modified keyword log likelihood ratio score can then be reported as

$$S_{LR} = S_{KW} - S_{BA} \,, \tag{2}$$

where S_{BA} is the score for the overlapping string of background fillers.

Table 2 shows the KWR performance according to the standard figure of merit with different acoustic models populating both the filler network, shown in the first column, and the background network, shown in the second column. The interest here is in those systems where no spectral normalization is applied. There are several observations that can be drawn from these results. The first is that there is a significant improvement in performance obtained from likelihood ratio scoring, even for a task such as this one using clean speech. This is illustrated by the performance improvement obtained for the *Monophone Models* systems in Table 2 over the *Monophone Models* system in Table 1. The second observations is that the use of likelihood ratio scoring is not a replacement for a filler network containing acoustic models of speech, as opposed to a model of acoustic background. This is apparent from the relatively poor performance of the *Adapt. Bkg.* system in Table 1 where both the filler network and background network contain only a single acoustic background model. Finally, the results suggest that there is little to be gained in this clean channel keyword recognition paradigm by using detailed acoustic models of speech.

4.3 Spectral Normalization

By observing long-term spectra of training and test utterances, it is apparent that there is considerable inter-speaker and intersession speaker variability. It is assumed that the sources of variability in a message can be modeled as a fixed linear channel, and can therefore be represented by an constant additive bias term in the cepstrum domain. Hence, spectral normalization is accomplished by transforming the mean vectors of the Gaussian observation distributions for all hidden Markov

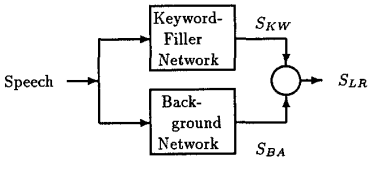

Figure 3: **Likelihood ratio scoring**

Filler Network Models	Background Network Models	Spectral Norm.	Ave. P_d (%) 0–10 fa/kw/hr
1 Adapt. Bkg.	1 Adapt. Bkg.	none	36.5
Monophone Models	Monophone Models	none	62.7
Monophone Models	1 Adapt. Bkg.	none	62.3
Monophone Models	1 Adapt. Bkg.	Blind Deconv.	64.2
Monophone Models	1 Adapt. Bkg.	Iter. Reest.	66.6

Table 2: **Performance for different filler and background networks.**

subword models. The cepstrum bias vector associated with this transformation can be estimated using a procedure described in [9], which is similar to a speaker adaptation technique due to Cox and Bridle [3]. In the procedure, a cepstrum bias vector is estimated from the unlabeled input observations so as to maximize the likelihood of the input utterance given the acoustic hidden Markov subword models.

It is assumed that the input utterance is divided into separate messages associated with each of the 8 evaluation speakers. The observations from each of these messages, ranging in length from 2 to 3 minutes, are used to estimate the parameters of the cepstrum bias vector. There are a total of 385 subword models in the KWR, and the Gaussian observation densities for each are treated as equally likely components of a Gaussian mixture. An iterative reestimation procedure is used to estimate the cepstrum bias vector as a fixed bias from the mean of each component Gaussian [9]. Table 2 shows the effect of this cepstrum transformation on KWR performance in comparison to a simple blind deconvolution procedure. This is the best average P_d obtained, and corresponds to $P_d = 82\%$ @ 12 fa/kw/hr.

5 SUMMARY

This paper has presented two techniques for separating occurrences of keywords in a running utterance from non–keyword speech. The difficulty of the task is characterized not only by the size of the keyword vocabulary, but also by the nature of the non–keyword speech. The particular focus this work is in large keyword vocabulary tasks involving unconstrained, non–stereotyped utterances arising from conversational speech. The problem has been approached in two ways. Filler models have provided explicit models for all speech, including both keywords and non–keywords, and has treated the keyword recognition problem from a continuous speech recognition point of view. Likelihood ratio scoring by itself treats keyword recognition as an open set word detection problem, where a limited set of keyword models are used to separate keyword speech from a background of non–keyword speech. The HMM based keyword recognition system described in this paper has combined the two approaches, and it was found that both approaches significantly contributed to the overall keyword recognition performance.

Good keyword recognition performance has been obtained on a standard task with a relatively low complexity system. An average P_d of 66% was obtained using monophone filler models, likelihood ratio scoring, and the spectral normalization procedure described in Section 4.3. Furthermore, it was found that good performance was only obtained for systems using filler models trained from transcribed speech. This suggests that keyword recognition performance may benefit most through the use of a more representative training corpus containing a large amount of transcribed non–keyword speech. It is also expected that more advanced general context models of background speech may further improve keyword recognition performance. Further work will emphasize techniques for improving the keyword recognition performance under a variety of conditions. These include techniques for noise compensation through pre–processing of input utterances in recognition, as well as secondary testing of putative keyword hits.

References

[1] James Baker. private communication.

[2] J. S. Bridle. An efficient elastic-template method for detecting given words in running speech. *Brit. Acoust. Soc. Meeting*, pages 1–4, April 1973.

[3] S. J. Cox and J. S. Bridle. Unsupervised speaker adaptation by probabilistic spectrum fitting. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, May 1989.

[4] A. L. Higgins and R. E. Wohlford. Keyword recognition using template concatenation. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pages 1233–1236, April 1985.

[5] H. W. Hon, K. F. Lee, and R. Weide. Towards speech recognition without vocabulary specific training. *European Conf. on Speech Comm. and Tech.*, September 1989.

[6] B. P. Landell, R. E. Wohlford, and L. G. Bahler. Improved speech recognition in noise. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pages 749–751, April 1986.

[7] D. B. Paul. The Lincoln robust continuous speech recognizer. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, May 1989.

[8] J. R. Rohlicek, W. Russel, S. Roucos, and H. Gish. Continuous HMM for speaker independent word spotting. *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, May 1989.

[9] R. C. Rose and D. A. Reynolds. Text independent speaker identification using automatic acoustic segmentation. *To be published in Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, May 1990.

[10] J. C. Spohrer, P. F. Brown, P. H. Hochschild, and J. K. Baker. Partial backtrace in continuous speech recognition. *Proc. Int. Conf. on Systems, Man, and Cybernetics*, pages 36–42, 1980.