CONTINUOUS HIDDEN MARKOV MODELING FOR SPEAKER-INDEPENDENT WORD SPOTTING

S12.8

J. Robin Rohlicek *William Russell* *Salim Roukos* *Herbert Gish*

BBN Systems and Technologies Corporation
Cambridge MA 02138

ABSTRACT

We present a word spotting system using Gaussian hidden Markov models. Several aspects of this problem are investigated. Specifically, we present results on the use of various signal processing and feature transformation techniques. We have observed that performance can be greatly affected by choice of features used, the covariance structure of the Gaussian models, as well as transformations based on energy and feature distributions. Also, due to the open-set nature of the problem, the specific techniques for modeling out-of-vocabulary speech and the choice of scoring metric can have a significant effect on performance.

1 INTRODUCTION

In this paper, we present a word spotting system using Gaussian likelihood, hidden Markov models (HMM). In the next section, the basic HMM model is presented as well as a discussion of novel aspects of the HMM techniques that deal with the open-set nature of word spotting. In the following section, we describe the various feature transformations considered. The experimental results are then presented, followed by an overall discussion.

2 HMM MODELINGS

Word spotting differs from continuous speech recognition in that the task involves locating a small vocabulary of words embedded in arbitrary conversation rather than determining an optimal word sequence taken from a fixed vocabulary. A corpus of speech with known locations of these words is used for training. Recognition is then performed using a single HMM network with unrestricted input speech. This single HMM network incorporates the keyword models and a model for non-keyword speech. The scores in the entire network are used to normalize the keyword likelihoods. We describe in the following two elements of the normalization. First, in the context of training, the construction of a non-keyword model is presented. Second, a probabilistic scoring method based on the computation of *a posteriori* probabilities is described.

2.1 Training

Whole-word models of each of the keywords are trained using the maximum-likelihood forward-backward algorithm [1]. Gaussian observation models are determined using either diagonal or full covariance structure. The word models are "left-to-right" linear sequences of states with approximately three states per phoneme. Two basic topologies have been considered; although the majority of the results use the simple structure shown in Figure 1a (Word-A), some further results use the structure shown in Figure 1b (Word-B). Word-B structures contain groups of three states each, where the states within each group share the same output distribution; for this structure, the duration of word segments is explicitly modeled instead of relying on the implicit geometric distributions of holding times for the states shown in Figure 1a. In addition, we have doubled the number of output distributions in the Word-B structures which leads to approximately 18 states per phoneme.

The second part of training involves creating a model which attempts to represent non-keyword speech (an "alternative" model). As we discuss in the next section, such a model has a significant effect on the scoring method we use. A low likelihood for a keyword can be caused by an overall low likelihood of the speech signal (due to noise, for example) rather than an unlikely realization of that keyword. To normalize for this effect, an alternative model is needed.

We have initially used two simple alternative model structures: first, a single-state model having a Gaussian mixture based on a uniform weighting of all the distributions in the keyword states (Alt-A); second, networks of the type shown in Figure 2b (Alt-B). The components of these models are composed of segments of the keyword models. Ultimately, an alternative model that is estimated from additional (non keyword) speech is desirable.

2.2 Scoring

A single HMM network, shown in Figure 3, is constructed; in this network, the keywords and alternative model components are replaced by the structures shown in Figures 1 and 2. The network is updated time-synchronously using a standard Baum-Welch scoring procedure. That is, given the sequence of observations $x_1, \ldots, x_t$, the score at state i at time t is

$$S_i(t) = \mathrm{p}(s_t = i, x_1, \ldots, x_t).$$

These scores can be used to obtain an *a posteriori* probability of occupying a particular state,

$$\Pr(s_t = i | x_1, \ldots, x_t) = \frac{\mathrm{p}(s_t = i, x_1, \ldots, x_t)}{\mathrm{p}(x_1, \ldots, x_t)} = \frac{S_i(t)}{\sum_j S_j(t)}, \quad (1)$$

where the sum is over all states of the HMM network. The significance of the alternative model is in the denominator of the last expression since the sum should typically be dominated by the states of the alternative model. Finally, the probability of a keyword ending can also be evaluated from these scores as

$$W_n(t) = \Pr(s_t = e_n | x_1, \ldots, x_t),$$

where e_n is the state index of the last state of word n. Possible keyword ending points are determined by local maxima in $W_n(t)$.

The structure of the alternative model affects the utility of the score in the real case where the HMM network is not an adequate model for all speech. Therefore, we present results using both alternative model structures shown in Figure 2.

Finally, we compared the use of *a posteriori* probability, $W(t)$, with a duration-normalized likelihood that has been previously used in word spotting,

$$\mathcal{L}_n(t) = \mathrm{p}(x_{t-d+1}, \ldots, x_t | \text{word } n)^{1/d},$$

where d is the estimated duration of the word. In a typical experiment, $\mathcal{L}(t)$ yielded a detection probability of only 0.2 times that of $W(t)$. All of our experimental results, described below, are based on the *a posteriori* score.

CH2673-2/89/0000-0627 $1.00 © 1989 IEEE

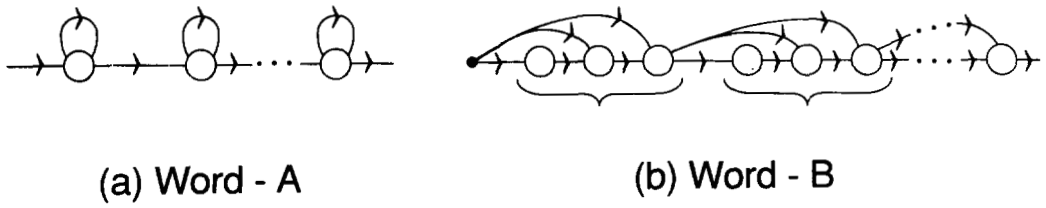

Figure 1: Word model structure

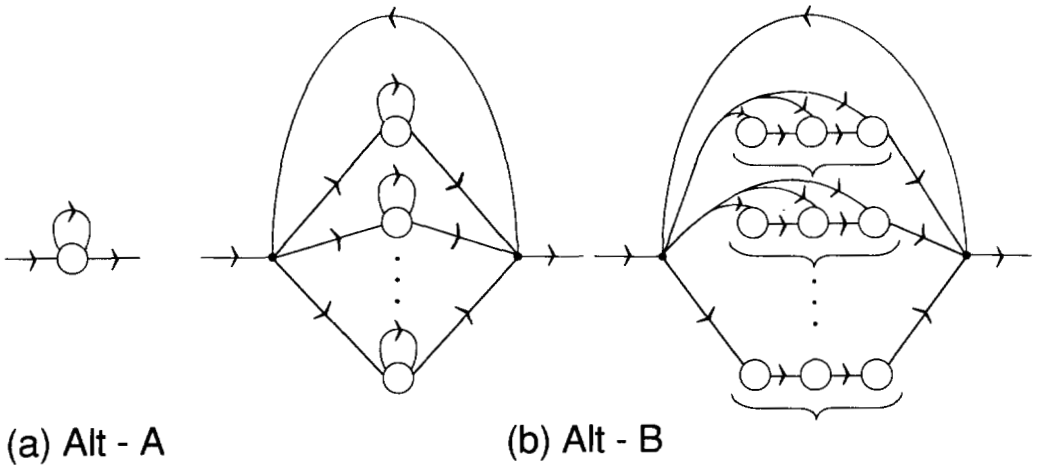

Figure 2: Alternative models

3 Feature selection

Several signal processing and feature transformation techniques are considered for telephone-bandwidth speech (sampled at 10 kHz.) The basic signal processing techniques used are:

- 18 Mel-spaced samples of a log LPC power spectrum over the full 5 kHz bandwidth (spectral bands, or SPBs), and
- cepstral coefficients of a Mel-warped spectrum in the range 300-3300 Hz ("selective" cepstra, c_0-c_{18}).

The feature transformations used are:

- shift normalization,
- range and distribution normalization,
- feature derivatives, and
- energy normalization

3.1 Shift normalization

To deal with the variability in speech from speaker to speaker as well as between the training and test databases, a speaker-dependent shift in each feature is introduced. Specifically, the distribution of each feature is determined from approximately two minutes of speech for each speaker. Then, to match the 80^{th} percentile points (computed from the speech regions), a speaker- and feature-dependent shift is introduced. This normalization attempts to remove the effect of an unknown, speaker-dependent, linear time-invariant channel. Furthermore, using a percentile rather than a mean should provide a more robust technique.

3.2 Range and distribution normalization

Two techniques are used to deal with feature distribution variations; they also obviate the need for the shift described above. Both methods map the entire dynamic range into the interval $[0, 1]$. These normalizations are again based on the distribution of each feature over a single speaker's speech. In range normalization, a speaker- and feature-dependent linear mapping (with hard limiting at 0 and 1) is computed for each feature. This map is determined by mapping the 10^{th} and 90^{th} percentiles to the values 0.1 and 1.0, respectively.

In distribution normalization, rather than using a linear map, each feature is mapped through a monotonic, nonlinear transformation so that the resulting distribution is uniform on $[0, 1]$. This is implemented by replacing the feature value by its image on the empirical cumulative distribution function.

3.3 Feature Derivatives

Derivatives of each feature are computed as the slope of a linear least-square fit of a straight line to a 5-frame window of data [2]. Note that the dimension of the feature space doubles when derivatives are used. Also, in the "full" covariance case, the covariance structure has two blocks, one for the features and one for their derivatives.

3.4 Energy normalization

Energy normalization is performed in a different manner for the SPB and cepstral cases. In the SPB case, this normalization is a memoryless transformation that is applied at each frame. Each component of the LPC power spectrum vector is scaled so that

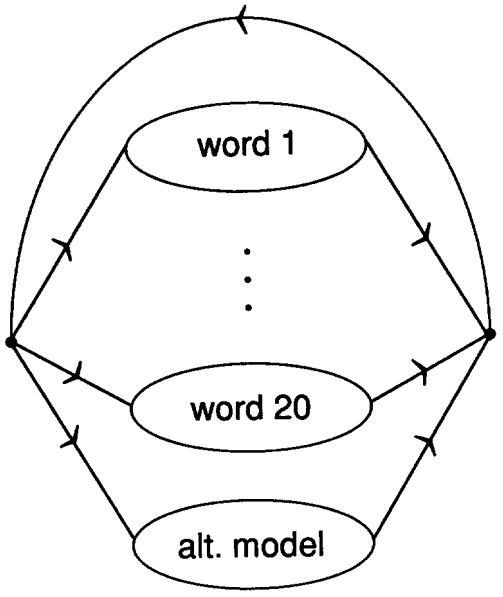

Figure 3: HMM network

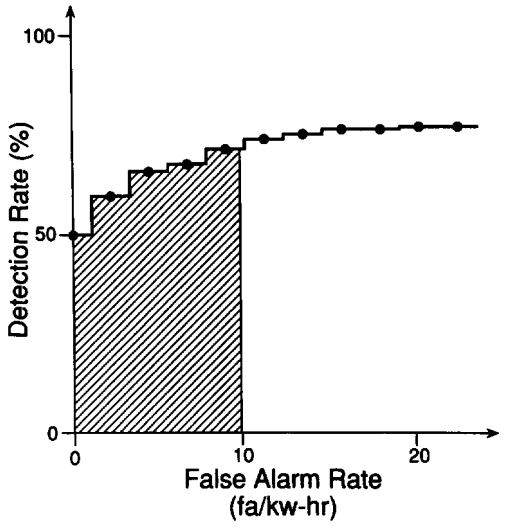

Figure 4: ROC curve: average value in shaded region is used

the sum of the components is one. In the cepstral case, experiments are conducted both with and without the "c_0" coefficient. Although these techniques are clearly not equivalent, they both have the effect of removing much of the short-term energy variation.

4 Experiments

In this section, the results of several experiments are presented. These experiments are aimed at examining the effects on performance of different basic feature sets (SPB vs. cepstra) and various feature transformations. Also, the sensitivity to the particular choice of an alternative speech model and the choice of word model structure is demonstrated. A set of comparative results are shown in Tables 1 through 5. The average detection probability (described below) is shown in the tables. Shift normalization is used in all the experiments.

4.1 Databases

Two separate databases with the same 20 keyword vocabulary were used. The first, used for training, consists of read paragraphs containing an average of 131 tokens per keyword from 28 male speakers. This data was recorded over local telephone lines. The second database, used for testing, consists of conversational speech from new speakers recorded with a dynamic microphone at 5 kHz speech bandwidth. Both databases were sampled at 10 kHz and bandpass filtered to 300-3300 Hz before further processing.

4.2 Performance Measure

In word spotting, one selects a threshold on the score $W(t)$ to determine a set of putative "hits" for a given amount of test speech. The fraction of false alarms in these putative hits is reported as a false-alarm rate per hour of speech. The fraction of keyword occurrences detected is an estimate of the detection probability. Clearly, the receiver operating characteristic (ROC) curve allows a trade-off between detection probability and false-alarm rates. Our estimate of the ROC curve for our word spotting system is determined by selecting a set of thresholds for the keywords for the score $W(t)$ such that various numbers of false alarms are observed. The fraction of true detections above these thresholds is defined as the detection probability estimate. The composite (over keywords) performance is evaluated as the total fraction of true detections for a given false-alarm rate. Figure 4 shows the ROC curve for a typical experiment. We are primarily interested in the false-alarm rates below 10 false alarms per keyword per hour (fa/kw-hr); therefore, we characterize the result of an experiment by the average value of the ROC curve over the range of 0 to 10 fa/kw-hr rates. This allows for a more stable statistic of performance and will be used as a figure of merit for comparing different systems.

In these experiments, we evaluate performance on a male, 10-speaker subset of the test database. The test set duration is 26 minutes of conversational speech (silence accounts for about 30%) for the 10 male speakers. A total of 405 tokens of the keyword vocabulary occurs in the test set. Assuming a binomial model on the detection rate, the standard deviation in the detection rate is approximately 2.5%.

4.3 Results

Table 1 shows the word spotting performance for SPBs with energy normalization using the Alt-A model. The table shows the effect of including the derivatives and range normalization described earlier. Note that the addition of derivatives increases detection rate by an average of 7.5% when range normalization is

Covariance	Diagonal			Full		
Normalization		Range	Dist.		Range	Dist.
SPB	25	23		25	27	
SPB w/deriv.	28	34	35	29	33	33

Table 1: SPB: Effect of range and distribution normalization, use of derivatives, and covariance structure

used, but by only 3% without range normalization. Furthermore, range normalization is insignificant when derivatives are not used. Also, distribution normalization appears equivalent to range normalization. Finally, the full covariance case is comparable to the diagonal case.

In another set of experiments, we compared the performance of c_1-c_{18}, c_0-c_{18}, and the 18 SPBs, all with derivatives and distribution normalization. We also used one of the two alternative models described earlier (Alt-A and Alt-B). Table 2 shows the

Covariance	Diagonal		Full	
Alternative Model	Alt-A	Alt-B	Alt-A	Alt-B
c_1-c_{18} w/deriv	14	49	16	52
c_0-c_{18} w/deriv	18	56	27	55
SPB w/deriv	35	38	33	41

Table 2: Comparison of SPB and cepstra: Effect of derivatives, alternative model, and covariance structure

performance of the diagonal and full covariance cases of these experiments. With the Alt-A alternative model, we found that c_0 seems to improve performance for the cepstrum; we also found that the SPB (which do not include energy) outperform the cepstral coefficients with Alt-A particularly in the diagonal case. The Alt-B alternative model gives a dramatic improvement in the detection rate for the cepstral features (it triples the detection rate for the diagonal case.) However, the Alt-B alternative model provides only a modest increase in performance for the SPBs. The reason for the large difference in performance for the cepstral coefficients requires further exploration. This result prompts us to be cautious about drawing strong conclusions about the significance of different performance results.

Next, we consider the effect of distribution normalization on cepstra and their derivatives. Distribution normalization affects the features in a way similar to the range normalization described above. Table 3 shows the average detection rate with and with-

Covariance	Diagonal	Full
c_0-c_{18} w/deriv	58	57
c_0-c_{18} w/deriv & dist.	56	55

Table 3: Cepstra with derivatives: Effect of distribution normalization and covariance structure

out distribution normalization. The results of Table 1 would predict an improvement through the use of distribution normalization with cepstra. However, no improvement is obtained; this suggests that distribution normalization is not required with cepstra. The remaining experiments use cepstra without distribution normalization (i.e. only shift normalization).

In Table 4 we consider cepstra without distribution normaliza-

	c_1-c_{18}	c_0-c_{18}
	44	39
deriv	48	58

Table 4: Cepstra: Effect of c_0 and derivatives, for diagonal covariance and Alt-B alternative speech model

tion and show the effect on word spotting performance of including c_0 and derivatives. The use of c_0 and derivatives improves performance by 14%. However, when derivatives are not used, performance actually decreases with the use of c_0.

Finally, Table 5 compares the performance of Word-A and

Word Model	Word-A	Word-B
c_0-c_{18} w/deriv	58	65
c_1-c_{18} w/deriv	48	61

Table 5: Cepstra: effect of word model structure and use of c_0.

Word-B word model structures (shown in Figure 1) for cepstra with derivatives. Performance with and without c_0 is shown. It is clear that the Word-B structure, where duration is explicitly modeled and where the number of output distributions is doubled, significantly improves performance.

5 Discussion of Results

Our experiments with feature transformations have shown that range and distribution normalization are beneficial for the SPB features but seem unimportant for cepstra. Furthermore, including derivatives improves performance in all situations. Finally, c_0 improves performance only when derivatives are used.

Using the full covariance structure does not yield significantly different performance from using the diagonal structure. This was not expected, particularly for the SPB case where successive samples of a smooth spectrum would be expected to be correlated. Given our training set size, we do not think that the lack of performance improvement in the full covariance case is due to insufficient training data; we suspect that the robustness of the model to new test conditions to have the more significant role.

Finally, compared to the Word-A word model structure, the Word-B structure provides substantially improved performance. The Word-B structure employs explicit duration modeling, and it contains more output distributions than the Word-A structure. More work is needed to determine the relative contribution of these two model characteristics to the recognition performance improvement.

6 Conclusion

We have described a word spotting system based on Gaussian hidden Markov models and evaluated the performance using two basic types of spectral envelopes: one based on linear prediction and the other based on the cepstrum. Various feature transformations and two alternative speech models were tested. The normalization of the likelihoods, which is implicit in our use of the *a posteriori* probability, is required for open-set problems such as word spotting. The normalization depends on the choice of alternate model and seems to affect performance to a great extent. Currently, our best performance is achieved with cepstra in combination with the appropriate alternate model. However, changing the alternate model can lead to a large decrease in performance. This sensitivity to which combination of features and alternate model requires further investigation and will be the focus of our future work.

Acknowledgement

This work was funded by the U.S. Department of Defense.

References

[1] L.R. Bahl, F. Jelinek, and R.L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190, March 1983.

[2] F.K. Soong and A.E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE ICASSP, Japan*, pp. 877-880, April 1986.