Proceedings of the Institute of Acoustics

IMPROVEMENTS TO A KEYWORD RECOGNITION ALGORITHM

Michael J Carey and Eluned S Parris

Ensigma Ltd., Chepstow, Gwent

1. INTRODUCTION

This paper describes a set of modifications made to a word spotting algorithm aimed at improving its performance. As a baseline we took as a starting point a system described by Rohlicek [1]. In this system a language or garbage model is included in addition to the models of the words to be detected. To improve this system we then implemented durational modelling by constructing models whose length was a function of the average length of the training utterances. As a second approach we used the alpha-nets paradigm first proposed by Bridle [2]. This technique provides for a neural network interpretation of a set of Hidden Markov Models allowing the parameters of the Markov Model to be adapted to maximize some classification criteria at their outputs. As a third approach we used a set of training tokens taken from continuous speech. It is important to note that a prime requirement of this work was the reduction of false alarms. Recognition rates are not, therefore, as high as some workers have reported.

2. DATABASE AND ANALYSIS

A test database used in a word spotting experiment must satisfy two criteria. Firstly it must have sufficient occurrences of the keywords to allow a valid statistical estimate of the keyword recognition rate to be made, and secondly it should contain enough speech from a wide range of speakers so that a good statistical estimate of the false alarm rate can be made. We therefore divided our database requirement into two parts. The first part consisted of a number of excerpts of small passages, rich in keywords, which were spoken by a variety of male and female speakers. The second part of the database was taken from three commercially published tapes entitled 'Accents of English' [3], 'English with a Dialect' [4] and 'English with an Accent' [5]. These contained a large selection of passages of approximately three minutes duration spoken by native and non-native male and female English speakers. False alarm testing was therefore carried out on a large number of different speaker types with the widest possible range of accents. A training vocabulary for a telephony application was available. This included such words as 'dial', 'expert', 'emergency', 'secretary', 'directory', 'telephone', 'operator' and 'number'. We considered that this vocabulary had a good mix of words of different phonemic complexity and was also available to use digitized and end point analysed. However, this vocabulary

was collected as isolated utterances whereas the words to be spotted were embedded in continuous speech where the phonetic contexts at the ends of words are more variable than in isolated words. We therefore collected a second training set consisting of frequently used words in BBC Radio Four weather forecasts.

The training and test data was bandlimited to 3.4 kHz and sampled at 8 ks/s prior to digitization. The speech was filtered using a bank of eleven filters and then the log power outputs of the filterbank were transformed to six cepstra at 20 ms intervals. Six differential cepstra were then derived from the cepstral coefficients which together with the differential energy comprised the feature vector. The pattern matching part of the algorithm was implemented using the Viterbi algorithm in the single pass form proposed for connected speech by Bridle et al [6]. Multiple mixture continuous density Hidden Markov Models were used.

3. BASELINE SYSTEM

As a baseline we took as a starting point a system described by Rohlicek [1]. In this system a language or garbage model is included in addition to the models of the words to be detected. The language model is trained on a wide variety of speech. For a valid word to be detected, the probability of that word model had to exceed the probability of the language model by some threshold. The justification for this approach is that speakers matching the language model of the speech badly would also match the individual word models badly and therefore the language model would give a reduced threshold against which the word models could be measured. Conversely, speakers fitting the language and specific word models well would operate against a higher threshold for acceptances. The language model was built from several hundred different English digits from a wide variety of speakers and had a left to right topology with ten states and two mixtures per state. To enable this model to fit a wide range of speech the transition matrix allowed for the skipping of an intervening state in addition to staying in the same state or making a transition from the preceding state. There were no penalties associated with any of these transitions. The performance of the baseline system is shown in Table 1.

Keyword	Recognition Rate (%)	False Alarms per hour
Secretary	45	1.14
Telephone	19	4.00
Office	9	0.29
Directory	18	0.57
Operator	32	0.00
Emergency	68	1.14
Dial	13	6.57
Number	31	6.00
Expert	19	3.71
Average	28	2.57

Table 1: Performance of Baseline System

4. MODELLING DURATION

One of the chief causes of false alarms in the baseline system was the matching of an inappropriate sequence of input frames of many frames of the input to a state in the model. We therefore adapted the model proposed by Picone [7] for this part of the study. This is a form of the Bakis model in which the number of states in this model is equal to the average number of frames in the training examples of the word. The minimum possible length of an utterance fitting this model with J number of states is $\frac{J}{2} + 1$. The durational modelling of the word is enforced by the transition probabilities. In the normal form of the stochastic model these have little effect. We therefore abandoned the stochastic constraint, in favour of experimentally determined penalties associated with the transitions $a_{ii}, a_{i,i+2}$. It is interesting to note that the overall form of this model is similar to that used in the symmetric template matching algorithms see for example [6].

A further refinement of the model can be achieved by making the penalty proportional to the number of time frames that the model stays in a particular state. We make the length of the model depend on the average length of the keyword. Typically, the number of states in the model is equal to 0.75 times the number of states in the average keyword. Since skips were also included in these models the shortest possible transit through the model would take 0.4 times the average length of the keyword. It is necessary to have this ability to make transits through the model since the keyword models were built from utterances taken from isolated speech, whereas the words to be spotted occur in continuous speech,

which can have a speaking rate over twice that observed for isolated words. A typical model now has approximately twenty states. Table 2 shows the results we obtained from this experiment over a variety of thresholds. Again for similar recognition rates we see a marked diminution in the false alarm rate and that with no threshold the recognition rate is above 50%.

Keyword	Recognition Rate (%)	False Alarms per hour
Secretary	45	0.57
Telephone	50	1.14
Office	36	0.00
Directory	55	0.57
Operator	27	0.00
Emergency	73	0.29
Dial	38	2.86
Number	56	2.86
Expert	31	2.00
Average	45	1.03

Table 2: System with Durational Modelling

5. ALPHA-NETS

An alpha-net is a discriminative neural-network interpretation of a set of Hidden Markov Models (HMMs). The parameters of the HMMs e.g. means, variances, are adapted to maximise some classification criteria at their outputs. The alpha-net is trained by back propagation of partial derivatives through time. The log probability score

$$J = -\log P_s$$

where P_s is the normalized final state likelihood of the correct model class s,

is minimized using a gradient algorithm. This is equivalent to maximizing the mutual information, which has been successfully used to improve the ability of HMMs to discriminate between words [11].

The probability b_{jt} in state j at time t is a function of the observations O_t and the parameters which are dependent on the state (e.g. means).

In alpha-nets these are adapted by the gradient algorithm

$$m_j(T) = m_j(T-1) - \xi\left(\frac{\partial J}{\partial m_j}\right)$$

where ξ is a coefficient which controls the rate of adaption. We therefore require $\frac{\partial J}{\partial m_j}$

The probability b_{jt} in state j at time t is a function of a multivariate mixture density Gaussian distribution and is given by

$$b_{jt} = \sum_{k=1}^{K} \frac{1}{\sqrt{2\pi}} c_{jk} \frac{1}{\sigma_{jk}} e^{-1/2 \left\| \frac{O_t - m_{jk}}{\sigma_{jk}} \right\|^2} \tag{1}$$

where c_{jk} is the mixture coefficient for state j mixture k, σ_{jk} is the vector of standard deviations for state j mixture k, m_{jk} is the vector of means for state j mixture k and K is the total number of mixture densities.

For a particular mixture density the probability b_{jkt} is given by

$$b_{jkt} = \frac{1}{\sqrt{2\pi}} c_{jk} \frac{1}{\sigma_{jk}} e^{-1/2 \left\| \frac{O_t - m_{jk}}{\sigma_{jk}} \right\|^2}$$

$$\begin{aligned} \frac{\partial J}{\partial m_{jk}} &= \frac{\partial J}{\partial L_{w_j}} \cdot \frac{\partial L_{w_j}}{\partial m_{jk}} \\ &= \frac{\left(P_{w_j} - \delta_{sw_j}\right)}{L_{w_j}} \cdot \sum_t \left(\frac{\partial L_{w_j}}{\partial b_{jt}} \cdot \frac{\partial b_{jt}}{\partial m_{jk}} \right) \end{aligned} \tag{2}$$

where

L_{w_j} is the likelihood of model w in state j,

$$\delta_{sw_j} \begin{cases} = 1 & \text{if } s \text{ is a member of class } w_j, \\ = 0 & \text{otherwise.} \end{cases}$$

Substituting

$$\frac{\partial L_{w_j}}{\partial b_{jt}} = \frac{\beta_{jt} . \alpha_{jt}}{b_{jt}} = \frac{\gamma_{jt}}{b_{jt}}$$

where β_{jt} is the backward probability at state j time t,
α_{jt} is the forward probability at state j time t,

and

$$\frac{\partial b_{jt}}{\partial m_{jk}} = b_{jkt}\left(\frac{O_t - m_{jk}}{\sigma_{jk}^2}\right) \tag{3}$$

in (2) gives

$$\frac{\partial J}{\partial m_{jk}} = \frac{\left(P_{w_j} - \delta_{sw_j}\right)}{L_{w_j}} \frac{1}{\sigma_{jk}^2}\left[\sum_t \gamma_{jt} O_t \frac{b_{jkt}}{b_{jt}} - m_{jk}\sum_t \gamma_{jt}\frac{b_{jkt}}{b_{jt}}\right]$$

Equations for the standard deviations and mixture coefficients can be similarly derived.

In word spotting, the keyword HMM and language HMM are trained using both valid occurrences of the keyword and false alarms. The keyword HMM is adapted to discriminate in favour of valid occurrences of the keyword and against false alarms. The language HMM is adapted to discriminate in favour of false alarms and against valid occurrences of the keyword. The use of different classification criteria can cause a reduction in false alarms and/or an improvement in recognition of valid occurrences of keywords.

All of the true occurrences and false alarms located in the experiment described in Section 4 were cut out of the original speech data into separate files. There were not many occurrences of false alarms for use in training the alpha-net. In particular, the keywords 'office', 'directory', 'operator' and 'emergency' all had less than three false alarms. For this reason only the means of the HMMs were adapted and the other state dependent parameters left static. The variances and mixture coefficients could also be adapted but a large number of occurrences of both false alarms and true keywords would be needed for training.

Each keyword model was taken as a pair with the language model and both models were adapted using the alpha-net to improve the false alarm rate. Half of the excised speech files were used for training the alpha-net.

Each adapted keyword and language HMM pair was tested with all of the original speech data. Figure 1 illustrates the effect of discrimination for the telephony vocabulary without the keyword 'dial'. Table 3 shows the recognition rates and number of false alarms for each keyword, before and after discrimination.

We can see from these results that substantial reductions in the false alarm rate of certain words, such as 'dial' and 'number', were achieved. However, this was at the expense of a correspondingly marked reduction in the recognition rate. It is worth noting that the type of false alarms we were getting with a word like 'dial' was confusion with the second half of 'missile'. Similarly, the word 'number' was confused with the word 'Humber', indicating the benefit of using phonetically more complex words like 'emergency', which had a small number of false alarms in the first place.

IMPROVEMENTS TO A KEYWORD RECOGNITION ALGORITHM

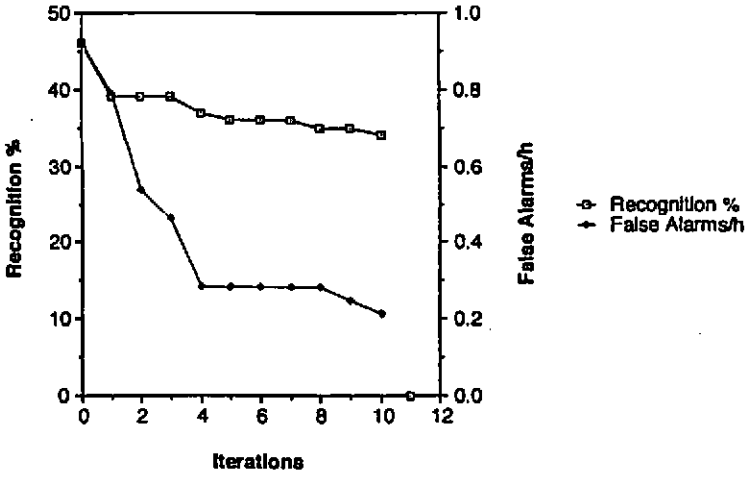

Figure 1: Reduction of False Alarm Rate by Means of Alpha-Nets

Keyword	Recognition Rate (%)	False Alarms per hour
Secretary	23	0.00
Telephone	44	0.57
Office	36	0.00
Directory	23	0.00
Operator	27	0.00
Emergency	59	0.00
Dial	13	0.00
Number	25	0.00
Expert	31	1.14
Average	32	0.17

Table 3: Performance after Adaptation

6. EMBEDDED TOKENS

The final experiment carried out was the replacement of the training set with a set of training utterances excised from continuous speech. Twenty-five men and twenty-five women were asked to read a passage in which each of the ten weather keywords occurred three times giving 150 training utterances. In this case the length of the models was set to be equal to the average length of the training set. The system was tested on fifty two minute broadcast weather forecasts and forty three hours of other broadcast speech. The results shown in Table 4 indicate that while there has been a considerable increase in the false alarm rate the recognition rate has also improved. This is consistent with the hypothesis that the presentation to the model building algorithm of speech with coarticulation effects would lead to models with broader distributions in the initial and final states.

Keyword	Recognition Rate (%)	False Alarms per hour
Temperature	21	1.0
Northern	42	59.0
Showers	74	3.1
Weather	64	47.1
England	88	53.4
Scotland	78	7.7
Ireland	67	20.6
Sunshine	71	1.7
Degrees	79	27.9
Tomorrow	92	6.6
Average	69.3	21.2

Table 4: Performance of Embedded Training Tokens

7. CONCLUSION

The most satisfactory result of our experiment was that the introduction of durational modelling both improved the recognition and the false alarm rates over those achieved in the baseline system. The alpha-net approach was capable of producing a level of false alarms hitherto unachievable. The use of embedded tokens in the training pass gave a large improvement in the recognition rate but at the expense of a disproportionately large number of false alarms. The form of wordspotter adopted will be dependent on the application which will determine the acceptable recognition rate and the level of false alarms.

8. REFERENCES

[1] J. R. Rohlicek et Al., 'Continuous Hidden Markov Modelling for Speaker-Independent Word Spotting', Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, Glasgow 1989, pp. 627-630.

[2] J. S. Bridle, 'Alpha-Nets: A Recurrent 'Neural' Network Architecture with a Hidden Markov Model Interpretation', Speech Communication, Vol. 9 No. 1 Feb. 1990.

[3] J. S. Wells, 'Accents of English', Cambridge University Press, 1982.

[4] 'English with a Dialect', BBC Publication, 1982.

[5] 'English with an Accent', BBC Publication, 1982.

[6] J. S. Bridle, M. D. Brown and R. M. Chamberlain, 'A One Pass Algorithm for Connected Word Recognition', Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, Paris 1982, pp. 899-902.

[7] J. Picone, 'On Modelling Duration in Context in Speech Recognition', Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, Glasgow 1989, pp. 421-424.

[8] L R Bahl, P F Brown, P V de Souza and R L Mercer, 'Maximum Mutual Information Estimation of HMM Parameters', Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, 1986, pp. 49-52.