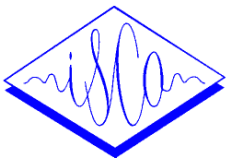
ISCA Archive
http://www.isca-speech.org/archive



4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

TOPIC SPOTTING WITH TASK INDEPENDENT MODELS

Michael J. Carey and Eluned S. Parris.

Ensigma Ltd, Turing House, Station Road, Chepstow, Gwent, NP6 5PB, U.K.

E-mail: michael@ensigma.com, eluned @ensigma.com

ABSTRACT

Topic spotting with whole-word models has been shown to give high detection rate with low false alarms. However the system must be capable of the generation of keyword models without repeated data collection sessions to be flexible in use. Since the required vocabulary is unknown a priori the models must be task independent. This however degrades the system performance which then must be restored. This can be achieved by using linear discriminant analysis in the feature extractor and the generation of context dependent subword models using decision trees. The system uses concatenations of the context dependent models to form the keyword models. Keywords are selected according to their usefulness. Non-keyword speech is modelled by a set of monophone models. During topic spotting the significance of the occurrence of keywords is weighted according to the discrimination they provide between topic and non-topic material. The system was tested on the BBC database spotting two minute weather forecasts. It detected 95% of the weather forecasts at a rate of one false alarm per hour. It was also tested on three other topics where its performance was not as good but still useful.

Keywords: topic spotting, word-spotting, context dependent models, linear discriminant analysis.

10.21437/Eurospeech.1995-510

1. INTRODUCTION

Two approaches to the problem of topic identification have been pursued. In the larger vocabulary approach [1] an attempt is made to transcribe the whole of the message prior to the operation of the topic spotting algorithm. In the small vocabulary approach [2] the system searches for only a small set of key words and uses occurrences of these for the identification of the presence of the topic. This is computationally simpler than the large vocabulary approach and does not require knowledge of, and the modelling of, the whole vocabulary.

We previously implemented a small vocabulary topic spotting system. This system achieved a 93% topic identification rate at one false topic per hour searching for the occurrence of weather forecasts in news broadcasts from a database of material collected from BBC Radio 4. In this system the word spotting algorithm used whole word continuous density Hidden Markov Models(HMMs). The production of the whole word models required the collection and hand editing of continuous speech from a minimum of fifty speakers. This was a major obstacle to the extension of the system to the identification of other topics.

In this paper we describe work we have carried out to produce task independent models for topic identification. Whole word models are constructed by concatenating phonetically appropriate subword models. The generation of task independent models has been achieved by the use of a number of techniques, including Linear Discriminant Analysis(LDA), the incorporation of explicit duration modelling and the generation of context sensitive models using decision trees. These are described in Section 2. The technique we use for word spotting with these models is described in Section 3, and the application of the word spotting system to topic identification, with the results achieved, is described in Section 4.

2. SUBWORD MODELLING

2.1 Acoustic Processing and Database

The database used for testing comprised 44 hours of news and current affairs programs collected from BBC FM Radio transmissions. The material was sampled and digitised at a rate of 8kHz. Fifteen hours of the database was transcribed and the vocabulary of this section was found to be about 7000 words. The transcriptions contain 150,000 words giving a speaking rate of 100 words per minute.

The SCRIBE and SCRIBE SRU databases were combined for use as training data. These are phonetically balanced databases of British English, which together contain utterances from over 160 speakers each saying five sentences. In each case the data was down sampled to a sampling rate of 8 kHz. The data was then filtered using the SRU Bank which contains nineteen filters. The filterbank outputs were transformed into twelve cepstral coefficients and twelve delta cepstral coefficients at a frame rate of 10 ms.

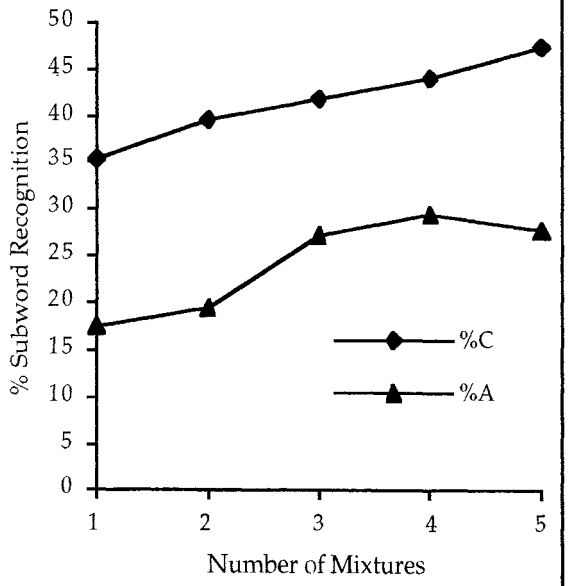


Figure 1. Subword Recognition Performance with Decision Tree Models.

These coefficients were augmented by energy and delta energy parameters to give a twenty six element feature vector. The subword models were three state HMMs with continuous mixture distributions and a left to right topology. Duration was modelled explicitly by a gamma distribution.

2.2 Linear Discriminant Analysis

It has been previously shown [3] that linear discriminant analysis can improve recogniser performance. An LDA transform was therefore applied to the feature vectors to improve the subword recognition rate, where we observed that increasing the number of mixtures in the subword models did not give the expected improvement in performance. We therefore modified the LDA algorithm to pool data over mixtures rather than states, which, as Table 1 shows, gave a marked improvement in performance. This is described more fully in [4].

Mixtures	State LDA %C	State LDA %A	Mixture LDA %C	Mixture LDA %A
1	28.8	18.5	32.4	22.8
5	30.5	18.4	37.9	24.2

Table 1. LDA Performance with State and Mixture Pools

2.3 Decision Tree Models

Context sensitive models [5] have been shown to give better subword recognition accuracy than monophone models. We therefore generated a set of decision trees based on the combined SCRIBE and SCRIBE SRU database. Other approaches used in building decision trees have used node splitting algorithms based on a dissimilarity measure or Poisson models. We however built new HMMs at each node for each question. We split the node using the likelihood of the model given the data as the metric used to determine the most most appropriate question. This approach is more computationally intensive than the previously mentioned techniques but is advantageous because the decisions are made directly using the models which will be used for recognition. A set of approximately 400 context sensitive models were built using this technique. The improvement in recognition performance as the number of mixtures in the models increases is shown in Figure 1. Comparing Table 1 and Figure 1 demonstrates how the context sensitive models gave an overall improvement in recognition performance.

3. WORD SPOTTING

3.1 Model Generation

The task independent whole-word models for keywords were generated as follows. Each keyword was transcribed into a sequence of phonemes taken from a pronunciation dictionary. When the keyword was based on monophones the transcription was used to select the appropriate subword models, which were then concatenated together to produce a whole word model. The same method was applied when using context sensitive models. However, in the latter case the left and right context of each phoneme was used together with that phoneme's decision tree to select the most appropriate context sensitive model. For word boundaries general left and right context sensitive models were used. Variations in pronunciation were accommodated by generating a whole word model for each major variation.

3.2 The Word Spotting Algorithm

Word spotting was carried out with a set of models comprising the whole word models and the set of monophone models. A bigram grammar was used to constrain the transitions between the monophones to linguistically plausible sequences. The transition probability between successive subword units of the whole word model was set to unity. If the phonetic sequence corresponding to a whole word model occurred, then the whole word model would be matched in preference to the corresponding series of monophones, since the monophones would be penalised by their lower inter-model transition penalties. The degree to which this occurred could be controlled by introducing a penalty, gamma, which exponentiated the transition probabilities to modify the penalty incurred by the background models.

Increasing the value of gamma allows the keyword models to match a wider range of input sequences since

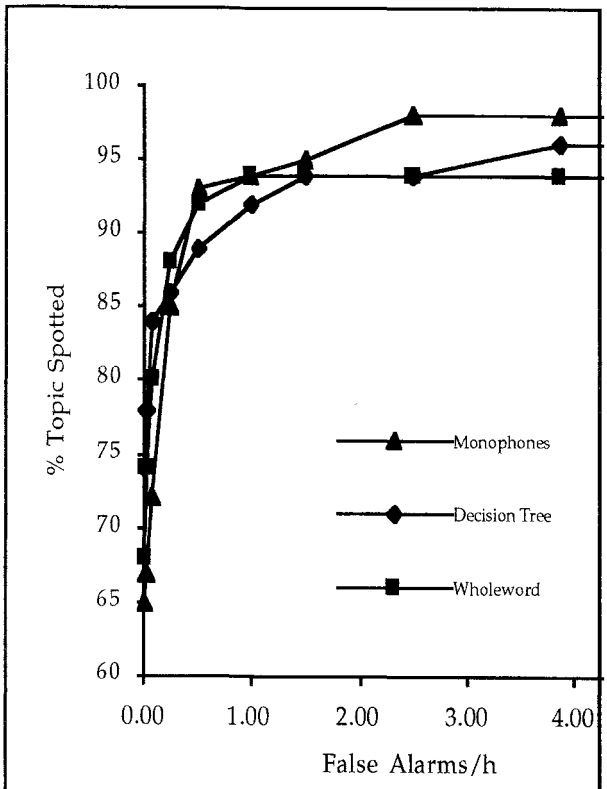


Figure 2. Operating Characteristic on the Weather Forecast Topic.

decreased spectral penalty in the background models is offset by the increased transition penalties, causing an increase in recognition rate at the expense of more false alarms. Using the decision tree models further increased the probability of keyword recognition as these models, unlike the background monophone models, had improved modelling of transitions between the keywords due to their context sensitivity which is advantageous when the true word occurs. The overall performance of the system was a 46% keyword recognition accuracy at a false alarm rate of ten false alarms per keyword per hour.

4. TOPIC SPOTTING

We have shown that usefulness of a keyword in the identification of a topic [6] is given by

$$U_{w_j} = p(w_j|C) \log \frac{p(w_j|C)}{p(w_j|\overline{C})}$$

where $p(w_j|C)$ and $p(w_j|\overline{C})$ are the probabilities of detecting the keyword in the topic and in unwanted material.

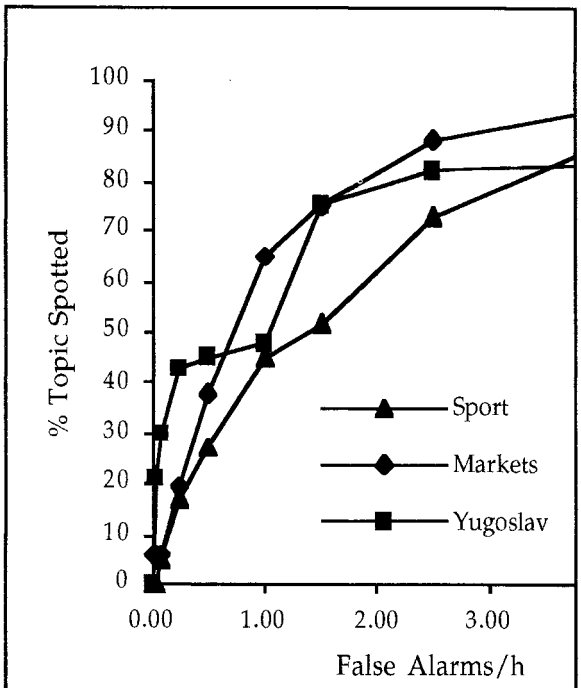


Figure 3. Operating Characteristic on the Markets, Sport and Yugoslavia Topics.

The keywords are ranked according to the usefulness, as shown in Table 2, selecting those which maximise the discrimination between topic and non-topic speech. The topic identification algorithm then operates by accumulating the n_j occurrences of each keyword output by the word spotter over the expected length of the topic, and weighting these occurrences by $\log \frac{p(w_j|C)}{p(w_j|\overline{C})}$, the log likelihood ratio of seeing the keyword occurrence in wanted and unwanted material, to give the score for the window

$$S_m = \sum_j n_j \log \frac{p(w_j|C)}{p(w_j|\overline{C})}$$

When this score exceeded a threshold the topic is deemed to be detected. The system was tested on fifty weather forecast broadcasts of two minutes duration each, and gave the results shown in Figure 2. As can be seen, the context independent models based on LDA monophone subwords ard decision trees performed as well as context dependent whole word models in this application. Since it was now possible to generate a new set of models in a few minutes we were able to apply the system to several other topics which occurred regularly in the BBC database. These were sports, Yugoslavia and the financial markets, whose keywords are shown in Table 3. These gave the results shown in Figure 3. While the system did not perform as well on these topics as on the weather forecast, it nonetheless is able to identify a high proportion of the topics occurring at an acceptably low false alarm rate.

Word	Transcription	Usefulness
weather	w e D @ @	24.6
tomorrow	t @ m Q r @U	31.3
sunshine	s V n S aI n	100.3
showers	S aU @ z	145.8
England	I N l @ n d	50.4
Ireland	aI @ l @ n d	18.7
Scotland	s k Q t l @ n d	75.0
northern	n O D @ n	107.5
temperature	t e m p @ r I tS @	81.2
degrees	d I g r I z	20.4

Table 2 Weather Forecast Keywords

Markets	Sport	Yugoslavia
market	player	Croatia
hundred	England	Yugoslavia
exchange	racing	Serbia
pound	against	Zagreb
nikkai	football	fighting
sterling	played	republic
Wall Street	number	federal
thousand	beating	president
index	winner	Vukovar
Dow Jones	finals	barracks

Table 3 Keywords for Other Topics

5. DISCUSSION AND CONCLUSIONS

The difference in results between the weather forecast and other topics shown above is worthy of further examination, since there are specific reasons which account for these differences and point to the future utility of a system such as the one we have described.

In the weather forecasts the keywords occurred regularly. This also applied to the financial markets topic but not to the others. However the length of items on the financial markets was typically one minute or less, which is probably the major factor degrading the performance of the system on this topic. The sport topic was not a homogenous topic, but a heterogeneous mixture of several sub-topics; football, rugby, cricket and horse racing being the main ones. The occurrence of keywords in sport was dependent on whether the sub-topic was discussed within a particular item. The unreliability of the occurrences of the keywords is thought to be a major contributor to the poor performance on this topic.

The final topic, Yugoslavia, seems to suffer from two problems, the first of which was that the context sensitive models of the keywords performed very badly in this topic. The results shown are for monophone models. We believe this is caused by most of the keywords not being English. The sequences of phonemes in these words were infrequently observed in the training data and hence had contexts which were poorly represented in the decision tree models. Also the broadcasts on Yugoslavia mapped a fluid political situation in which the focus of attention moved from one area to another, so that the relative usefulness of the keywords used in the system changed over time contributing to the poor performance on this topic.

The topic spotting system we have described in this paper assumes that the occurrences of the keywords are independent. The performance of the system can be improved if the dependencies between the occurrences of keywords are accurately modelled. This aspect of the system is addressed in [7].

6. ACKNOWLEDGEMENT

The authors would like to thank the Speech Research Unit of the Defence Research Agency, Malvern, UK for the Scribe SRU database and the phonological rules used in the decision tree modelling.

REFERENCES

[1]J. Baker et al. "Application of Large Vocabulary Continuous Speech Recognition to Topic and Speaker Identification", Proc ICASSP 93

[2]R. Rose et al. "Technique for Information Retrieval from Voice Messages", Proc ICASSP 91

[3] M. Hunt et al. "An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination", Proc ICASSP 91

[4]E. Parris and M. Carey, "Estimating Linear Discriminant Parameters for Continuous Density Hidden Markov Models" Proc ICSLP 94

[5]L Bahl et al. "Decision Trees for Phonological Rules in Continuous Speech", Proc ICASSP 91

[6]E. Parris and M. Carey, "Discriminative Phonemes for speaker Identification" Proc ICSLP 94

[7]J. Wright, M. Carey, and E Parris, 'Topic Spotting using Higher Order Statistics', Proc. ICASSP 95