KEYWORD TRAINING USING A SINGLE SPOKEN EXAMPLE FOR APPLICATIONS IN AUDIO DOCUMENT RETRIEVAL

K.M.Knill and S.J.Young

Speech, Vision and Robotics Group,
Department of Engineering,
University of Cambridge

ABSTRACT - An open keyword vocabulary word-spotting system is described for audio document retrieval. To model each keyword, an N-Best recogniser is used to hypothesise the keyword's phonetic transcription based on a single spoken example. The system is evaluated on a database of spoken messages and performance is found to be comparable with that obtained using pronunciations taken from a dictionary.

INTRODUCTION

Speech is an attractive input medium as it can be a faster and more expressive means of capturing information than handwritten or typed records. This ability is being increasingly exploited in applications such as voice and video mail, and hand-held organisers. In the latter case, voice input also has the advantage of negating the need for a keyboard, enabling smaller devices to be designed. However unlike visual documents, stored speech is slow to access and cannot be scanned easily.

Word-spotting is generally employed to retrieve audio documents of interest. Since personal computing devices are not intended to be restricted to a particular activity or profession, a task specific keyword vocabulary cannot be pre-defined. To enable a fully open keyword set to be used, a keyword model should be built using one or more spoken examples of the keyword. Ideally, the user should be able to give a single search request, such as 'Find contract', to retrieve messages containing the keyword, 'contract'. The difficulty with this, of course, is that a single spoken example is insufficient by itself to build an acoustic model which is accurate enough to give reasonable word-spotting performance.

Wilcox and Bush (1991) have proposed using pre-trained sub-word HMMs, representing general acoustic units, to construct the required keyword. Kupiec et al (1994) used phonetic sub-word HMMs in a recogniser to hypothesise the keyword, then applied a dictionary and knowledge of likely recognition errors to produce an n-best list of keywords. In this paper we present an alternative approach in which the required keyword is constructed from pre-trained sub-word phone models using one or more pronunciations determined by applying a N-Best phone recogniser to the single spoken example. This removes the need for a dictionary, hence, reducing the storage requirements, of particular concern in hand-held devices. Additionally it avoids the problems caused by out-of-vocabulary keywords.

This approach has been evaluated within an HMM-based speaker-dependent word-spotting system.

SYSTEM

The overall architecture of the system is shown in Figure 1. An isolated example of each keyword is presented to the system, and a model derived using the N-Best recogniser. This model is then used to search the set of stored speech messages for occurrences of the keyword.

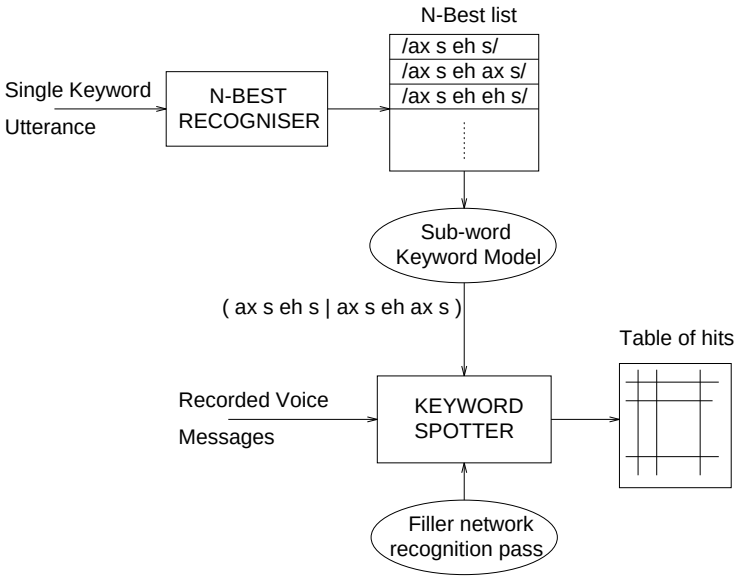

Figure 1. Overall system.

HMMs

Speaker dependent, continuous density multiple component Gaussian distribution, sub-word HMMs were used. Each model had 3 emitting states, with a left-to-right topology, no skips. The parameters used in the system were: 12 Mel frequency coefficients, normalised log energy and first and second derivatives of all parameters giving a vector size of 39. Two sets of models were used in the tests: 43, 8 mixture component, monophone models; 705, 2 mixture component, word-internal biphone models.

Database

The Video Mail Retrieval (VMR) DATABASE 1 (Jones et al 1994) developed at Cambridge University Engineering Department was used for evaluating the system. This database has been designed for word-spotting and information retrieval in spontaneous speech. The HMMs were trained on all read speech, the N-Best keyword models were recognised from the isolated word examples, and word-spotting was performed on the 20 spontaneous messages, for speaker 3 in the database. Only those keywords for which there are examples in the test data were used, this gave a subset of 26 from the 35 predefined VMR keyword set.

Baseline word-spotting system

A two-pass word-spotting system was implemented. Both passes use the monophone HMMs in a null grammar, time synchronous Viterbi beam search decoder. A parallel network of the keyword and monophone models is used in the main word-spotting pass, with silence enforced at the start and end of each sentence. The monophone models act as background or filler models to match non-keyword speech.

Rose and Paul (1990) showed that significant performance gains can be achieved by re-scoring the putative keyword hits against a recognition pass consisting of the filler models only. In this pass, a parallel network of the monophone models is used. Since it is independent of the choice of keywords, this pass only needs to be carried out once, and can be performed as the message is recorded. The

keyword hits are re-scored in this work by dividing the maximum log likelihood keyword score by the average filler model score over the same time frames. This has been shown to improve performance more than other proposed re-scoring schemes (Knill & Young, 1994).

In the baseline system, the keyword model is built by linking the monophone models according to the keyword's pronunciation rules given by a phonetic dictionary (based on the Advanced Oxford Learner's Dictionary (1)). Table 1 shows the performance of this baseline system, where for the ith false alarm,

$$\% \text{ hits / } i\text{th FA} = \frac{\sum_{j=1}^{m} H_{i,j}}{H} \times 100 \quad (1)$$

where m is the number of keywords, $H_{i,j}$ the number of true hits scored for the jth keyword before the ith false alarm, and H is the total number of true hits.

System	No. of False Alarms			
	1	2	3	10
baseline	63.5	65.5	69.5	80.7

Table 1. Baseline word-spotting performance; % hits per false alarm for 8 mixture component models.

N-Best keyword recognition

An N-Best recogniser was used to automatically derive pronunciations for a keyword based on a single isolated utterance of the keyword. The recogniser had a null grammar, with sub-word models connected in a parallel network allowing any sequence of phones. Silence was enforced at the start and end of the utterance. Unlike a standard Viterbi decoder, the recogniser produced the N most likely phone string hypotheses. The resulting phone sequences were treated with equal weight and combined in parallel to form a sub-word model for the keyword, as shown in Figure 1 where the vertical bar in the example pronunciation denotes an alternative.

Sets of monophone and word-internal biphone models were tested in the N-Best recogniser. In the former case, some initial phones were made optional in the keyword string using sight-derived rules e.g. /f t/ became /[f] t/. This took account of errors caused by breath effects owing to the isolated word context which were poorly accounted for by the models since they were trained on read speech. In the biphone case, some strings were repeated due a context mismatch between the final biphone and word-end monophone model in the recogniser output. When this occurred, only one example of the repeated string was used in the keyword model.

RESULTS

Monophone N-Best

The N-Best recognition was performed with 8 mixture component monophone models. Table 2 shows the word-spotting performance achieved when the number of N-Best phone strings used in the keyword model was varied between 1 and 9.

From Table 2 it can be seen that introducing the N-Best phone strings gives better word-spotting performance than if the optimal Viterbi phone sequence was taken as the keyword model. Performance improves as N is increased, up to $N = 7$. Taking the N-Best phone sequences increases the likelihood of one of the keyword sequences matching the dictionary definition. It also allows some

variations in the keyword pronunciation to be modelled.

N-Best	No. of False Alarms			
	1	2	3	10
1	44.2	53.3	57.4	63.5
3	48.7	59.9	62.9	72.1
5	48.2	59.9	65.0	71.6
7	49.2	59.9	64.5	73.1
9	49.2	57.9	64.5	72.6

Table 2. % Hits per false alarm using N-Best derived keyword models; 1 to 9 N-Best.

The best system, N=7, has a lower overall performance than the baseline system, with between 5 and 14% difference in the number of hits per false alarm. However, for 21/26 keywords, the 7 N-Best system achieves at least the same number of hits at the 3rd false alarm level as the baseline system. In the cases where the N-Best system performed least well compared to the baseline system, the cause was found to be due to poor modelling of the keyword pronunciation by the N-Best recogniser.

Biphone N-Best

Word-internal biphone models were used in the N-Best recogniser to see if improving the acoustic models would improve the accuracy of the N-Best strings, and, hence, improve the overall word-spotting performance.

The accuracy of the keyword pronunciations can be judged by considering the number of N-Best hypotheses in which the phonetic dictionary definition is included. For N=7, the dictionary definition occurred in 15 keyword hypotheses when the biphone models were used in the N-Best recogniser, compared to 9 hypotheses when the monophone models were used. In the remaining keywords, at least one hypothesis with only a single phone difference from the dictionary definition was observed for 9 and 10 keywords using the biphone and monophone models, respectively. Table 3 shows that this increase in keyword model accuracy results in better word-spotting performance, particularly at the 1st and 2nd false alarm level.

System	No. of False Alarms			
	1	2	3	10
baseline	63.5	65.5	69.5	80.7
monophone	49.2	59.9	64.5	73.1
biphone	55.8	62.4	64.5	74.6

Table 3. % Hits per false alarm for; baseline, 7 N-Best monophone and biphone systems.

The biphone system does not perform as well as the baseline system. However, as with the monophone system, the number of true hits achieved at the 3rd false alarm level by the biphone system equals or exceeds the baseline system for 21/26 keywords. Errors in the acoustic phone model was again found to be the cause of poor performance for the keywords where a noticable difference in hit rate was observed.

As the number of syllables in a keyword increased, the performance of all the systems improved and the gap between the baseline and N-Best systems narrowed. This is shown in Table 4 for the biphone system.

No. of Syllables	System	No. of False Alarms			
		1	2	3	10
1	baseline	40.9	43.2	50.0	71.6
	biphone	30.7	38.6	43.2	62.5
2	baseline	77.3	80.0	82.7	86.7
	biphone	73.3	78.7	78.7	82.7
3	baseline	91.2	91.2	91.2	91.2
	biphone	82.4	88.2	88.2	88.2

Table 4. % hits per false alarm by keyword syllable length for; baseline, and 7 N-Best biphone systems.

CONCLUSION

We have presented a preliminary study into the use of an N-Best recogniser to enable an open keyword vocabulary for audio document retrieval. The technique allows names, places, abbreviated terms etc to be recognised. Since the technique does not require a pronunciation dictionary, it reduces the computational complexity and, hence, cost of voice activated systems. At present, the performance is not as good as a dictionary based system. However, only simple means of combining the N-Best phone sequences have been tried. Alternative approaches to handling the prior information to be gained from the N-Best output, for example from the likelihood scores and phone pattern matching, should help reduce the gap in performance.

ACKNOWLEDGEMENTS

This work is funded by Hewlett-Packard Laboratories, Bristol, U.K.

NOTES

(1) The original dictionary was obtained from ftp: black.ox.ac.uk.

REFERENCES

Jones, G. J. F., Foote, J. T., Sparck Jones, K., & Young, S. J. (1994) *Video Mail Retrieval Using Voice: Report on Keyword Definition and Data Collection (Deliverable Report on VMR Task No 1)*, University of Cambridge Computer Laboratory, Tech. Report No. 335, May.

Knill, K.M. & Young, S.J. (1994) *Speaker Dependent Keyword Spotting for Accessing Stored Speech*, Cambridge University Engineering Dept., Tech. Report No. CUED/F-INFENG/TR 193.

Kupiec, J., Kimber, D., & Balabsubramanian, V. (1994) *Speech-Based Retrieval Using Semantic Co-Occurence Filtering*, Proc. ARPA Human Language Technology Workshop.

Rose, R. C. & Paul, D. B. (1990) *A Hidden Markov Model Based Keyword Recognition System.*, Proc. ICASSP, Albuquerque, S2.24, 129-132.