VIDEO MAIL RETRIEVAL: THE EFFECT OF WORD SPOTTING ACCURACY ON PRECISION

G. J. F. Jones[1,2] *J. T. Foote*[1] *K. Sparck Jones*[2] *S. J. Young*[1]

[1]Cambridge University Engineering Department, Cambridge, CB2 1PZ, UK
[2]Cambridge University Computer Laboratory, Cambridge, CB2 3QG, UK

ABSTRACT

The goal of the Video Mail Retrieval project is to integrate state-of-the-art document retrieval methods with high accuracy word spotting to yield a robust and efficient retrieval system. This paper describes a preliminary study to determine the extent to which retrieval precision is affected by word spotting performance. It includes a description of the database design, the word spotting algorithm, and the information retrieval method used. Results are presented which show audio retrieval performance very close to that of text.

1. THE VIDEO MAIL RETRIEVAL TASK

The last few years has seen an increasing use of multimedia applications, including video conferencing and video and audio mail. Using these facilities can create large archives of video material, which poses a significant problem. Users are unable to find stored messages because, unlike text, there are no simple ways to search for a particular reference. The Video Mail Retrieval (VMR) project at Cambridge University is addressing this problem by developing a system to retrieve stored video messages by voice. A specific goal of the project is to develop a useful retrieval application for the MEDUSA multimedia environment installed at Olivetti Research Ltd. in Cambridge, UK.

In the simplest form of message retrieval, a user will specify a single search keyword and word spotting techniques will locate its occurrences in the audio soundtrack. A more robust system will use multiple search keys, both to minimize the effect of spotting errors and to refine the list of retrieved messages. Thus, the topic specification and search strategies developed for conventional text-based information retrieval (IR) must be adapted to this new environment [1]. Although later stages of the project will investigate open-keyword and open-user sets, the initial stage, and the study described here, focuses on a fixed, *a priori* known set of search keywords and users.

For the initial development of the VMR system, it was necessary to create a test archive of messages with known audio and information characteristics to evaluate word spotting and message retrieval performance. Unfortunately, existing corpora intended for word spotting research were not appropriate for this purpose for two reasons. The VMR system is intended to work in a multimedia system with high fidelity audio, while most existing corpora are only telephone-quality. In addition, the information content of existing corpora is not always appropriate for information retrieval experiments. A new corpus of audio mail messages, constituting a more natural document set, was recorded at the Cambridge University Engineering Department.

A key issue when designing a voice-based message retrieval system is the extent to which word spotting accuracy affects retrieval performance. One way to assess this is to compare the performance of a voice–based system with a standard IR system operating on orthographic transcriptions of the audio material. This paper describes some preliminary experiments using this approach on our message database designed specifically for this purpose.

1.1. The VMR message corpus

The VMR message corpus is a structured collection of audio training data and information–bearing audio messages. Ten "categories" were chosen to reflect the anticipated messages of actual users, including, for example, "management" and "equipment." A fixed set of 35 keywords was chosen to cover the ten categories; thus the "management" category includes the keywords "staff," "time," and "meeting." Keywords may be associated with more than one category; and the keyword set includes 11 difficult monosyllabic words (e.g "date" and "mail"), as well as overlapping words (e.g. "word" and "keyword") and word variants (e.g. "locate" and "location"). For the message data, talkers were asked to record a spontaneous response to a prompt (with five prompts per category), for a total of 50 unique prompts.

There were fifteen talkers, of which 11 were male and 4 female. Data was recorded at 16 kHz from both a Sennheiser HMD 414 close-talking microphone and the cardioid far-field desk microphone used in the MEDUSA system, in an acoustically isolated room. Each talker provided the following speech data:

- 77 read sentences ("r" data): sentences containing keywords, constructed such that each keyword occurred a minimum of five times.
- 170 isolated keywords ("i" data): 5 occurrences of each of the 35 keywords spoken in isolation.
- 150 read sentences ("z" data): phonetically-rich sentences from the TIMIT corpus.
- 20 natural speech messages ("p" data): the response to 20 unique prompts from 4 categories.
- 20 "tags" ("t" data): natural speech responses to a prompt requesting a summary for each of their "p" messages.

The "i," "r," and "z" data are intended for use as training data; the "p" and "t" data, along with their transcriptions, serve as a test corpus for both keyword spotting and preliminary IR experiments. The tag data was not used for the presented work, so the test corpus consists of the 300

Section	Training Data			Test Data	
	"i"	"r"	"z"	"p"	"t"
Amount	62.6	91.6	145.8	261.2	37.5
Total	Train: 300.0			Test: 298.7	

Table 1. VMR Speech Corpus: minutes of recorded speech

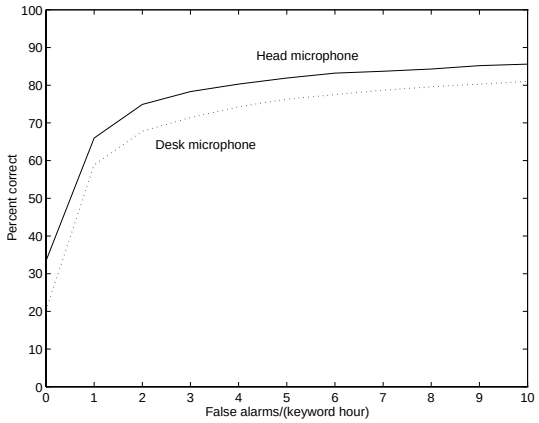

Figure 1. VMR Task Keyword Spotting ROC

spontaneous messages. There were 6 messages (from 6 different users) for each of the 50 prompts. Additional data about the corpus is summarized in Table 1.1..

All files were verified and transcribed at the word level; non-speech events and disfluencies such as partially spoken words, pauses, and hesitations were transcribed in accordance with the Wall Street Journal data collection procedures. Phonetic transcriptions were automatically generated from a machine-readable version of the Oxford Learners Dictionary. The standard reduced TIMIT phone set was augmented with additional vowels specific to British English pronunciation. A full description of the VMR corpus may be found in [2].

2. KEYWORD SPOTTING

For HMM training and recognition, the acoustic data was parameterized into 12 mel-cepstral coefficients at a 100 Hz frame rate, and difference and acceleration coefficients were appended. The HTK tool set was used to construct whole-word talker-dependent keyword models and monophone filler models for each of the 15 talkers [3]. 3-state monophone models are used for both keywords and filler; separate phone models are used for filler phones and each keyword phone instance. Word-dependent keyword models are then constructed by concatenating the word-specific phone models at the network level.

2.1. Model Training

For every training utterance, a phone sequence was generated from the text transcription and a dictionary. These sequences were used to estimate both filler and keyword phone models. Keyword models were trained on approximately 10 instances of each keyword, of which half were spoken in isolation ("i" data) and half in the context of read messages ("r" data). The filler monophone models were trained on sentences from the TIMIT database ("z" data) and the non-keyword parts of the read messages ("r" data). Once single-mixture monophone models had been initialized, the number of mixture components was increased, and the parameters re-estimated in the usual way. Re-estimation was halted after a small number of mixture components, as additional components did not improve spotting performance. This is almost certainly because some of the word-specific phones had limited training data. All experiments reported here are based on two-component mixture models.

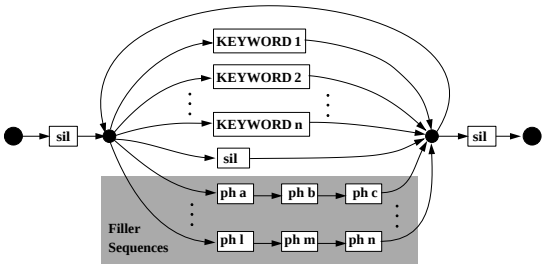

Figure 2. Keyword recognition network.

Data set:	head	desk
Talker Avg.	81.2%	76.4%

Table 2. Average Figures of Merit

Separate models were trained for both the head-microphone and desk-microphone data; the head-microphone data had an average SNR of 35-45 dB. Though the SNR of the desk-microphone data was substantially less (about 20-25 dB), the recordings are subjectively crisp, with little of the audible reverberation characteristic of omnidirectional microphones.

2.2. Keyword Recognition

Keyword spotting is done with a two-pass recognition procedure. First, Viterbi decoding is performed on a network of just the filler models, yielding a time-aligned sequence of the maximum-likelihood filler monophones and their associated log-likelihood scores. Secondly, another Viterbi decoding is done using a network of the keywords, silence, and filler models in parallel. In a manner similar to Rose & Paul [4], keywords are rescored by normalizing each hypothesis score by the the average filler model score over the keyword interval. Unlike [4], however, the average log likelihoods are divided rather than subtracted, which results in somewhat better performance [5].

Because of the limited training data, the monophone filler models are better-trained than the keyword models, and it was necessary to tune the filler models so that they did not match an undue number of keywords. A satisfactory solution was to introduce filler models of common 3-phone sequences by concatenating three monophone models, and adjusting the word transition penalty to penalize the filler sequences (which must be traversed in groups of three). To minimize keyword misses (filler models recognized for valid keywords), an attempt was made to to construct filler sequences as "orthogonal" to the keywords as possible. The 3-phone filler sequences were obtained from a list of the 100 most common 3-phone sequences in the Wall Street Journal training corpus by eliminating those identical or similar to keyword sequences, leaving a final set of 43 filler sequences. Figure 2 shows the network topology for the keyword recognition Viterbi pass.

2.3. Recognition Results

An accepted figure-of-merit (FOM) for word spotting is

defined as the average percentage of correctly detected keywords as the threshold is varied from one to ten false alarms per keyword per hour. Keyword spotting results were scored against aligned text transcriptions containing the keywords. The FOM for the two audio channels are shown in Table 2, averaged across both the 15 talkers and the 35 keywords. The receiver operating characteristic (ROC) curve for both the close-talking and far-field microphone data is shown in Figure 1. Keyword spotting results for both the head and desk microphone data were used for the retrieval experiments of Section 3.

3. INFORMATION RETRIEVAL

Information retrieval (IR) techniques are used to satisfy an operator's information need by the retrieval of potentially relevant messages from an archive. The intent of the VMR system is to select specific documents from the archive, rather than perform broad subject classification as in related research on "topic" identification [4, 6]. The contents of a video mail archive are dynamic over time, and hence there is no opportunity to pre-determine keyword weights or thresholds. Fortunately, there exist methods from text-based IR research that enable messages to be scored relative to a user's request with a minimum of *a-priori* knowledge.

3.1. Methodology

Information retrieval experiments require message *queries*, expressing a user's information requirements; and assessments of message *relevance* to the queries. Since real user queries and assessments were not available for the message corpus, they were simulated for our first tests as follows. Queries were constructed from the message prompts used in the database recording. To reduce variations in word form that inhibit retrieval matching, query words were suffix-stripped to stems using a standard algorithm [7]. Queries were formed from the prompts by selecting those stems also found in the keyword stem list. For example, given the prompt

```
Your current project is lagging behind schedule.
Send a message pointing this out to the other
project management staff. Suggest some days and
times over the next week when you would be
willing to hold a meeting to discuss the
situation.
```

the following query was obtained:

```
project messag project manag staff time meet
```

Word fragments such as "`messag`" are the suffix-stripped keyword roots. The 6 recorded messages generated in response to each prompt were assumed relevant to the query constructed from that prompt. Note that the 24 other messages in the same category, which are quite likely to contain similar keywords, are assumed to be not relevant; retrieval of one of these messages is construed as an error. Thus, the retrieval task is comparatively difficult.

In retrieval, a score is computed for each query-document pair which may then be used to rank documents [8]. Considering keyword presence/absence only, the score is the number of keywords in common, often called the *coordination level* (cl) score. Keywords may also be usefully weighted, for instance by the *inverse document frequency* (idf) weight,

$$idf_i = \log \frac{N}{n[i]}$$

	Text		Phonetic	
Weight scheme	cl	idf	cl	idf
Ave. precision	0.293	0.332	0.279	0.317

Table 3. Text and phonetic message retrieval performance.

where N is the total number of documents and $n[i]$ is the number of documents that contain keyword i. Thus, keywords occurring in a small number of documents are favoured. For this weighting scheme the query-document score is the sum of keyword weights for each keyword which occurs in both the query and the document.

Retrieval performance is often measured by *precision*, the proportion of retrieved messages that are relevant to a particular query. One conventional single-number performance figure, *average precision*, is derived as follows: the precision values obtained at each new relevant document in the ranked output for an individual query are averaged, and the results are then averaged across the query set. Other retrieval evaluation metrics are available and generally preferable, but this single-number performance measure is useful for comparing text and word spotting results.

3.2. Calibration via Text Retrieval

Acoustic word spotting is prone to false alarms and missed keywords, so retrieval performance can be expected to suffer degradation relative to text documents. The extent of the degradation can be measured by comparing retrieval performance on word spotting results with that on text. We used our transcribed corpus to provide us with this text performance standard, applying suffix-stripping, matching and scoring as described in the previous section.

An additional problem of word spotting is that unrelated acoustic events will often resemble valid keywords. For example, the last part of "hello Kate" is acoustically quite similar to the keyword "locate." Because even the most accurate acoustic models cannot discriminate between homophones, the output of an ideal word spotter that reports all keyword phone sequences provides a more legitimate standard of comparison than text. This was simulated by scanning the message phonetic transcriptions for sequences that match keyword phone sequences. Table 3. shows a comparison of *text* and *phonetic* standard retrieval performances for both cl and idf weighting.

3.3. Acoustic Message Retrieval Performance

The word spotter outputs a list of putative keyword hits and associated acoustic scores. Because the message retrieval scheme uses only the presence/absence of a keyword in a message, the acoustic scores are thresholded such that only hits with a score above the threshold are counted. The effect of thresholding is shown in Figure 3; because false alarms typically score worse than true hits, high thresholding values will remove a greater proportion of the false alarms. In practice, it is necessary to select an operating threshold. The FOM is not useful for this as it is an average across multiple threshold values. A good measure of spotting performance at a given threshold is the *accuracy*, defined as the number of keyword hits minus the number of false alarms divided by the number of actual keywords.

Accuracy, and IR performance, depend strongly on the threshold value. The top two curves in Figure 4 show plots of average idf retrieval precision against score threshold for both the head and desk microphones (precision is shown relative to the phonetic standard). On the left of the figure, corresponding to low threshold values, retrieval per-

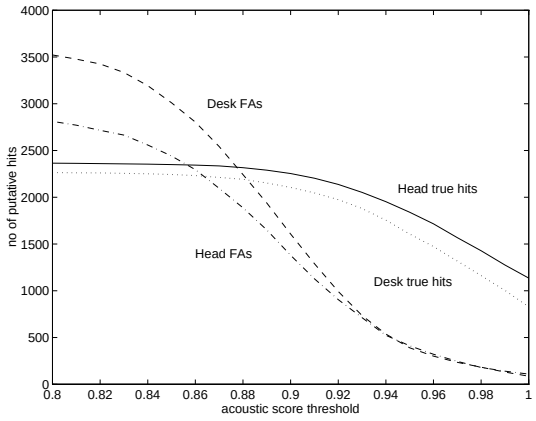

Figure 3. Nos of putative hits versus threshold.

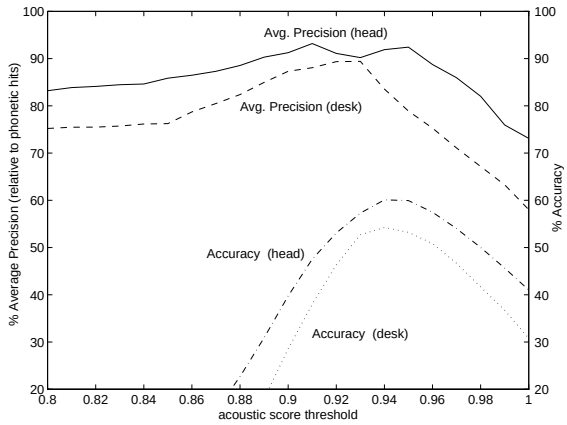

Figure 4. IR results versus threshold.

formance is impaired by a high proportion of false alarms; conversely, high thresholds (towards the right) remove a significant number of true hits, also degrading performance. The lower curves in Figure 4 are the accuracies versus the threshold. The optimal threshold, in the central region, represents the best tradeoff between the numbers of true hits and false alarms. In both cases, the threshold giving the highest *accuracy* also provides near optimal retrieval performance.

Table 4 compares acoustic retrieval performance with ideal text and phonetic standards, at the *a posteriori* best thresholds. These results show that ideal phonetic retrieval performance is degraded by about 5% relative to that of the standard text transcription, because of homophones; thus retrieval using even an ideal word spotter will not perform as well as retrieval from a full text transcription. However, even for an imperfect word spotting system considering both head and desk microphones, retrieval performance is encouragingly around 90% of the ideal phonetic figure. As anticipated from the lower FOM and the behaviour shown in Figure 3, retrieval performance for the desk microphone is slightly lower, although probably still sufficient for successful incorporation into the MEDUSA system.

One reason for the good performance of the retrieval system is the inherent robustness of idf weighting with respect to false alarms. Idf weighting penalises both frequently occurring, and hence undiscriminating, keywords (as in the text case), and also keywords having high numbers of false alarms across the document set. This illustrates the advantage of using text-based IR methods on acoustic tasks.

	Absolute	Text Relative	Phon. Relative
Text	0.332	100%	—
Phonetic	0.317	95.5%	100%
Head	0.295	88.8%	93.2%
Desk	0.283	85.2%	89.4%

Table 4. Relative idf retrieval performance.

4. CONCLUSIONS AND FURTHER WORK

Our tests so far have been very limited in scale, and in an artificial laboratory retrieval environment. Work is underway on enhancing the retrieval system to accommodate open keyword and user sets, allowing a search for arbitrary words spoken by anyone [9]. Using the system in a real-world office environment will undoubtedly raise other issues which must be addressed, such as noise robustness and the actual information content of video mail messages. These should not be insurmountable, however, and the first steps presented here suggest that state-of-the-art information retrieval and word spotting techniques can be combined successfully to provide a useful retrieval environment.

5. ACKNOWLEDGEMENTS

This project is supported by the UK DTI Grant IED4/1/5804 and SERC Grant GR/H87629. Olivetti Research Limited is an industrial partner of the VMR project. The VMR corpus used for this work will be available for public distribution in the near future.

REFERENCES

[1] U. Glavitsch and P. Schäuble. A system for retrieving speech documents. In *Proceedings SIGIR '92*, pages 168–176. ACM, 1992.

[2] G. J. F. Jones, J. T. Foote, K. Sparck Jones, and S. J. Young. Video Mail Retrieval Using Voice: Report on keyword definition and data collection. Technical Report 335, Cambridge University Computer Laboratory, May 1994.

[3] S. J. Young, P. C. Woodland, and W. J. Byrne. *HTK: Hidden Markov Model Toolkit V1.5*. Entropic Research Laboratories, Inc., 1993.

[4] R. C. Rose. Techniques for information retrieval from speech messages. *Lincoln Laboratory Journal*, 4(1):45–60, 1991.

[5] K. M. Knill and S. J. Young. Speaker dependent keyword spotting for hand-held devices. Technical Report 193, Cambridge University Engineering Department, July 1994.

[6] J. McDonough, K. Ng, P. Jeanrenaud, H. Gish, and J. R. Rohlicek. Approaches to topic identification on the switchboard corpus. In *Proceedings of ICASSP 94*, pages I–(385–390), Adelaide, 1994. IEEE.

[7] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.

[8] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.

[9] D. A. James and S. J. Young. A fast lattice-based approach to vocabulary independent wordspotting. In *Proceedings of ICASSP 94*, pages I–(377–380), Adelaide, 1994. IEEE.