ROBUST TALKER-INDEPENDENT AUDIO DOCUMENT RETRIEVAL

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 (VMR) project is to integrate state-of-the-art document retrieval methods with speech recognition to yield a robust and efficient retrieval system. The work presented here extends VMR towards an open-vocabulary, talker-independent system for retrieving spontaneously-spoken audio and video messages. We present results showing successful retrieval using a standard large-vocabulary (LV) recogniser, despite the lack of a matched language model and vocabulary. We further show that integrating a LV recogniser with conventional word spotting (WS) gives more robust retrieval performance than either method alone. This paper gives details of the message archive used, the speech recognition methodologies, the information retrieval methods, and experimental results.

1. INTRODUCTION

The last few years has seen an increasing use of multimedia applications, including video conferencing and video and audio mail. Using these facilities can result in 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.

In this system speech recognition techniques are used to locate potential search keys in the audio soundtrack. To retrieve messages, a user enters a search request and the recogniser output is examined for occurrences of these search keys. A robust system uses multiple search keys, both to minimize the effect of recognition 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. This is a challenging problem as indicated in related work [1] [2].

Earlier work in the VMR project successfully demonstrated retrieval of spontaneously spoken messages using a small, *a-priori* known set of 35 search keys, for both talker-dependent [3], and talker-independent [4] word spotting. This paper describes work using large-vocabulary (LV) recognition to extend the set of possible search keys to the size of the recogniser lexicon (here 20,000 words). For the retrieval task, small-vocabulary word spotting complements large-vocabulary recognition in that word spotting may be quicker, more robust, or better able to find terms outside the LV recognition lexicon. We show that we can get good retrieval performance when using the Wall Street Journal (WSJ) 20K lexicon and language model, even though these are not well matched to spontaneously-spoken video mail messages. To recognise task-specific keywords not in the LV recognition lexicon, LV results may be augmented with small-vocabulary word spotting (WS) results. We present experiments showing that combining LV and WS recognition improves retrieval performance over our previous WS-only systems, and that the combination can give better retrieval performance than either source alone.

2. METHODOLOGY

The LV recognition and WS systems are used to identify occurrences of the search keys in the messages. In this initial study we used the output of the HTK 20K WSJ large-vocabulary continuous-word recognition system [5]. LV performance suffers from low word accuracy due to the spontaneous nature of the data, the lack of a suitable language model, and a large number of out-of-vocabulary (OOV) terms. Nevertheless the LV system can substantially improve retrieval performance compared to our earlier fixed keyword system when additional search keys are available to give better coverage of the user's information request. Furthermore, word accuracy and OOV problems can be somewhat ameliorated by combining LV recognition results with small-vocabulary WS results. This is particularly true because the WS system is more robust than the LV one, though it is anticipated that these advantages will diminish as the lexicon and language models of LV recognisers are improved.

It is important to determine how retrieval performance is affected by recognition and spotting accuracy. 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 experiments using this approach on a message database designed specifically for this purpose.

3. THE VMR MESSAGE CORPUS

For the initial development of the VMR system, it was necessary to create an archive of messages (VMR1) with known audio and information characteristics to evaluate both word spotting and message retrieval performance [6]. A fixed set of 35 keywords were chosen to reflect the anticipated messages of actual users; for example, keywords included "staff," "time," and "meeting." 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"). This keyword set included four keywords not in the 20K LV recognition lexicon: "indigo," "keyword," "pandora," and "spotting".

Fifteen talkers (11 men and 4 women) each provided 20 spontaneous speech messages in response to 5 prompts from 4 out of 10 available categories. Data was recorded at 16 kHz from a Sennheiser HMD 414 close-talking microphone. The resulting 300 messages (5 hours of data), along with their text transcriptions, serve as a test corpus for both the keyword spotting and large-vocabulary IR experiments. The messages are fully spontaneous, and contain a large number of disfluencies such as "um" and "ah," partially uttered words and false starts, laughter, sentence fragments, and informalities and slang ("'fraid" and "whizzo"). This corpus provides a challenging spontaneous speech recognition task, particularly since there is not sufficient training data to estimate a task-specific language model.

4. LARGE VOCABULARY RECOGNITION AND WORD SPOTTING

For both LV and WS HMM training and recognition, the acoustic data was parametrized into 12 mel-cepstral coefficients (plus energy) at a 100 Hz frame rate, and difference and acceleration coefficients were appended.

In the word spotting system, the keywords were constructed from a set of 8-mixture word-internal tied-state triphone HMMs trained on the WSJCAM0 British English speech corpus [7] using a tree-based state clustering technique [8]. Each keyword is modelled by concatenating the appropriate sequence of subword models (obtained from a phonetic dictionary). Biphones are used at the beginning and end of keywords, while triphones model the internal structure. For example, the keyword "find" is represented by the model sequence `f+ay f-ay+n ay-n+d n-d`. Non-keyword speech is modelled by an unconstrained parallel network of monophones (denoted "filler models"). This strategy resulted in a 69.9% figure of merit (FOM) on the VMR1 data (which can be improved using talker-adaptation) [4].

For large-vocabulary recognition experiments, a set of 8-mixture cross-word triphones was trained on the same WSJCAM0 corpus of British English speech. These were used with the standard WSJ 20K large-vocabulary bigram language model from MIT Lincoln Labs to yield a 53% word accuracy rate. Though this is relatively low, it is to be expected for several reasons. The VMR1 corpus has a significant out-of-vocabulary rate of 3.15%, including 4 of the 35 frequently-occurring fixed keywords. The North American business news language model is highly inappropriate for informal UK English monologues, as demonstrated by an estimated perplexity of 356 on the VMR1 data. Also problematic is the exclusively read training data, the spontaneous nature of the test speech, the current lack of disfluency modelling, and the non-uniform accents (British, American, and Middle European) in the corpus [9]. Work is underway in developing a more appropriate language model, adapting acoustic models to different accents, and accounting for spontaneous speech phenomena. However, even the relatively poor recognition of the existing system results in respectable retrieval performance.

When LV and WS are compared for the 35 fixed keywords, the LV system is less robust, yielding only a 53% FOM versus the 69.9% FOM of the WS system. Because only about 1% of the LV system results are false alarms, the comparison would be fairer if the word spotter operating point was adjusted to give a similar number of false alarms. This would reduce the number of true hits and thus the FOM. Though the language model should theoretically help the LV recogniser distinguish between homophones, it may hinder word spotting performance because of word form variations. For example, for WS the keyword "retrieve" will be spotted for "retrieval" which occurs frequently in the test set. However, since "retrieval" is not in the 20K lexicon and the LV recogniser is constrained by the language model, many similar word form variations may be mis-recognised.

5. INFORMATION RETRIEVAL

Information retrieval (IR) techniques are used to satisfy a user's information need by retrieving relevant messages from an archive. In practice, the user composes a search *request* by typing in a sentence or set of words; from this a group of messages is returned, ranked by a matching score on the request content words. The user can then browse the high-scoring messages to find the desired information.

5.1. IR Experiments

Requests and Relevance Assessment Evaluating an IR system requires a set of message requests, together with assessments of the *relevance* of each message to each of these requests. Previous experiments [3] used a simulated request and assessment set (VMR1a); however a more realistic set (VMR1b) has since been collected.

For VMR1b a total of 50 requests were collected, 5 for each of the 10 categories used in message collection. These were gathered from 10 users, each of whom generated 5 requests and corresponding relevance assessments. This was achieved by forming 10 unique sets of 5 categories, and assigning each to a user knowledgeable about the categories. These users were asked to compose a natural language request from the information given in a text prompt, and to include at least one of the fixed keywords, One such prompt was formed for each of the message categories by combining information given in the 5 message prompts associated with the category. Ideally, the relevance of all archived messages should be assessed; however this is not practical even for our 300-message archive (which is considered relatively small). A suitable assessment subset was formed by combining the 30 messages in the category to which the original message prompt belonged, plus 5 messages from outside the category having the highest retrieval scores.

Data Preprocessing. IR benchmarks are established using text. Speech recognition outputs are considered to be message "pseudo-transcriptions". These, as well as written requests and the true transcriptions, are processed before search by removing function words using a standard *stop list* [10], and reducing the remaining words to stems. Function words are of no value to retrieval when using this approach, and stemming (using a standard method such as the Porter algorithm [11]) suppresses variations in word form that inhibit term matching. Once processed, a request is referred to as a search *query*. For example, given the request

```
In what ways can the windows interface of a
workstation be personalised?
```

the following query is obtained:

```
wai window interfac workstat personalis
```

Message Scoring Given a query, a matching score for each message can be computed, and the messages in the archive ranked by this query-document matching score [10]. Considering search key presence/absence only, the simplest

scoring method is just to count the number of keys in common, often called the *coordination level* (*cl*) score. However, it is more useful to weight keys, for instance by the *inverse document frequency* (*idf*) weight,

$$idf(i) = \log \frac{N}{n[i]}$$

where N is the total number of documents and $n[i]$ is the number of documents that contain search key i. This scheme favours rarer (and hence more selective) keys. The query-document matching score is then the sum of the matching query term weights. A more sophisticated weighting scheme takes into account the number of times each term occurs in each document and normalises with respect to the document length. This latter factor is important since a document's relevance does not depend on its length and hence neither should the score. A well tested *combined weight* (*cw*), described further in [12] is

$$cw(i,j) = \frac{idf(i) \times tf(i,j) \times (K+1)}{K \times ndl(j) + tf(i,j)}$$

where $cw(i,j)$ represents the cw weight of term i in document j, $tf(i,j)$ is the document term frequency and $ndl(j)$ the normalised document length. The combined weight constant K has to be determined empirically.

5.2. Integrating Large-Vocabulary Recognition and Word Spotting

Information retrieval using the LV and WS systems generally yields different results. In this paper we investigate methods of combining the two to produce better retrieval performance than that of either system alone. Combining multiple information sources has been shown to improve text-based IR systems; see [13] for a comparative study. Two approaches to information combination were considered in this study, referred to as *query combination* and *data fusion*. In query combination, multiple queries for the same information need are combined into a single query which is used to form a single ranked output list for a document set. In data fusion, multiple ranked lists (from different data representations) are combined to form a single overall ranked list. The methods described below use elements of both these techniques.

Data Fusion The data fusion method used here is to form a weighted sum of the matching scores computed independently from the LV and WS hypotheses. Since the scores from the two systems may be incommensurable, they may be normalised with respect to the highest scoring document in each list. (The most effective weighting for each component must be determined empirically.) The result is a new ranked list of the combined scores.

Data Combination In data combination we combine evidence from different sources in a way analogous to query combination. Specifically, word hypotheses from both systems are combined into a single document before computing the matching score. Hypotheses from the LV output may be either augmented with putative WS hits for keywords not in the LV lexicon, or pooled with all WS hypotheses regardless of LV lexicon and keyword overlap. The latter may help counteract acoustic stemming problems, but search keys are counted twice if hypothesised by both systems. So because they are frequency-based, *idf* weights may be affected by spurious keys in other documents due to WS false alarms. The *cw*, which takes into account within-document term frequency, may also be adversely influenced by this multiple counting of terms as well as the WS false alarms.

Weighting			*cl*	*idf*	*cw*
VMR1a	Avg. precision	fixed	0.293	0.332	0.358
		open	0.600	0.671	0.718
VMR1b	Avg. precision	fixed	0.296	0.332	0.346
		open	0.327	0.352	0.368

Table 1. Text retrieval performance for VMR1a and VMR1b.

5.3. Measuring IR performance

Retrieval performance is often measured by *precision*, the proportion of retrieved messages that are relevant to a particular query at a given position in the ranked list. A convenient (if crude) single-number performance measure, *average precision*, is derived as follows. For each query, the ranked-list precision values are averaged, and the results are then averaged across the query set. To assess spoken document retrieval, and specifically the effect of imperfect word recognition, performance for recognition can be compared with that for text transcriptions of the documents.

5.4. Information Retrieval Results

5.4.1. Large Vocabulary vs Word Spotting

Table 1 shows retrieval performance for text transcriptions using the full *open* vocabulary and for occurrences of only the 35 *fixed* keywords. Using the open vocabulary average precision is doubled for VMR1a compared to the fixed keywords, but increased by only around 10% for VMR1b. This can be explained by the relative differences in average query length between the fixed keyword and open vocabularies; VMR1a increases from 5.7 to 18.7 words, but VMR1b only changes from 2.7 to 5.7 words. An open vocabulary will naturally be more helpful to the longer requests, since these are more likely to contain keys matching message contents. In addition, the chance of missing all query terms (and thus the message) goes down dramatically as the number of terms increases (this effect is rather similar to "semantic co-occurrence filtering" described by Kupiec *et al.* [14]). For both query sets, retrieval performance improves with increased term weighting scheme sophistication. Results are shown for the *cw* scheme with $K = 1$, which gave good performance for both query sets.

5.4.2. Combining Word Hypothesis Sources

Table 2 shows retrieval performance using various speech recognition configurations. The 20K WSJ system performs better than WS system for VMR1a, but not as well for the shorter queries in VMR1b. The effect of OOV keywords and recognition errors is clearly greater for the shorter VMR1b queries. Again, term weighting improves performance overall. The WS performance depends on applying a threshold to putative hit scores [3] (results here are at the *a posteriori* best threshold). It is interesting to note that the *cw* scheme not only improves absolute performance but also makes it much less sensitive to the choice of threshold [12].

Data Fusion Directly adding query-document matching scores from the two lists results in the scores labelled *simp. merge* in Table 2; for *norm. merge*, scores in each list are normalised by the maximum score before summation. Normalisation prevents one of the lists dominating the overall score. For VMR1b, the best overall result is produced

Weighting			*cl*	*idf*	*cw*
VMR 1a	Avg. prec.	35 kw	0.220	0.249	0.287
		20K vocab	0.475	0.523	0.576
		simp. merge	0.468	0.538	0.591
		norm. merge	0.426	0.482	0.521
		+ all KW's	0.490	0.540	0.607
		+ OOV KW's	0.496	0.543	0.588
VMR 1b	Avg. prec.	35 kw	0.241	0.284	0.309
		20K vocab	0.225	0.246	0.263
		simp. merge	0.285	0.312	0.335
		norm. merge	0.289	0.319	0.342
		+ all KW's	0.248	0.265	0.306
		+ OOV KW's	0.250	0.272	0.290

Table 2. Speech retrieval performance

by this data fusion approach (here both components are weighted equally). Though not shown in the table, VMR1a results were improved by weighting in favour of the LV component, and VMR1b by favouring the WS component. This is expected since the additional terms in the LV system disproportionately aided the longer VMR1a queries. This also explains the relative performance of the *simp.* and *norm.* strategies for the two query sets. The general value of data fusion was also observed in [13].

Data Combination Two combination strategies were investigated; combining all putative term occurrences from the WS with the LV or adding only the OOV terms from the WS. The latter was more effective for VMR1a, since this enables the LV output hypotheses to dominate. Conversely this approach is less useful for the VMR1b query set where the WS output is more important.

6. CONCLUSIONS AND FURTHER WORK

The work reported here differs from previous topic spotting for spoken documents using LV [15] or WS [16] recognition. In these cases the choice of keywords for topics was based on their occurrence in manually transcribed training corpora and on recognition reliability, and was geared to optimimising retrieval for standing topics. But such training material is not always available, and information needs may change over time. The audio document information retrieval system described here handles previously unseen *ad hoc* topic specifications in the form of user generated textual requests. Using this method we have shown retrieval performance to be robust to an unmatched lexicon and language model. Furthermore, it has been shown that better retrieval performance can be obtained by combining LV and WS systems, though the best combination method depends somewhat on the nature of the queries. Current work is in progress to augment these two methods with a phone-lattice scanning approach [2] to cover search keys not in the recognition lexicon or a fixed keyword set, yielding a task independent, truly open-vocabulary audio retrieval system.

7. ACKNOWLEDGEMENTS

This project is supported by the UK DTI Grant IED4/1/5804 and SERC Grant GR/H87629. The authors would like to thank David James for useful discussions, David Pye and Phil Woodland for word-external acoustic models, and Julian Odell for the word-internal and language models used in this work.

REFERENCES

[1] M. Wechsler and P. Schäuble. Indexing methods for a speech retrieval system. In C. J. van Rijsbergen, editor, *Proceedings of the MIRO Workshop*, University of Glasgow, September 1995.

[2] D. A. James. *The Application of Classical Information Retrieval Techniques to Spoken Documents.* PhD thesis, Cambridge University, February 1995.

[3] G. J. F. Jones, J. T. Foote, K. Sparck Jones, and S. J. Young. Video Mail Retrieval: the effect of word spotting accuracy on precision. In *Proc. ICASSP 95*, pages 309–312, Detroit, May 1995. IEEE.

[4] J. T. Foote, G. J. F. Jones, K. Sparck Jones, and S. J. Young. Talker-independent keyword spotting for information retrieval. In *Proc. Eurospeech 95*, pages 2145–2148, Madrid, 1995. ESCA.

[5] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young. Large vocabulary continuous speech recognition using HTK. In *Proc. ICASSP 94*, volume II, pages 125–128, Adelaide, SA, 1994.

[6] G. J. F. Jones, J. T. Foote, K. Sparck Jones, and S. J. Young. VMR report on keyword definition and data collection. Technical Report 335, Cambridge University Computer Laboratory, May 1994.

[7] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals. WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In *Proc. ICASSP 95*, pages 81–84, Detroit, May 1995. IEEE.

[8] S. J. Young, J. J. Odell, and P. C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proc. ARPA Spoken Language Technology Workshop*, Plainsboro, NJ, 1994.

[9] P. Jeanrenaud, E. Eide, U. Chaudhari, J. McDonough, K. Ng, M. Siu, and H. Gish. Reducing word error rate on conversational speech from the Switchboard corpus. In *Proc. ICASSP 95*, pages 53–56, Detroit, May 1995. IEEE.

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

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

[12] K. Sparck Jones, G. J. F. Jones, J. T. Foote, and S. J. Young. Experiments in spoken document retrieval. *Information Processing and Management.* In Press.

[13] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. Combining the evidence of multiple query representations for information retrieval. *Information Processing and Management*, 31(3):431–448, 1995.

[14] J. Kupiec, D. Kimber, and V. Balasubramanian. Speech-based retrieval using semantic co-occurrence filtering. In *Proc. HLT 94*, pages 350–354. ARPA, 1994.

[15] L. Gillick, J. Baker, and J. Bridle et al. Application of large vocabulary continuous speech recognition to topic and speaker identification using telephone speech. In *Proc. ICASSP 93*, pages II–(471–474), San Francisco, May 1993. IEEE.

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