# Retrieving Spoken Documents by Combining Multiple Index Sources

G.J.F. Jones[†‡], J.T. Foote[‡], K. Spärck Jones[†] & S.J. Young[‡]

[†]Computer Laboratory, University of Cambridge,
New Museums Site, Pembroke Street
Cambridge CB2 3QG, England

[‡]Engineering Department, University of Cambridge,
Trumpington Street,
Cambridge CB2 1PZ, England

**Abstract**

This paper presents domain-independent methods of spoken document retrieval. Both a continuous-speech large vocabulary recognition system, and a phone-lattice word spotter, are used to locate index units within an experimental corpus of voice messages. Possible index terms are nearly unconstrained; terms not in a 20,000 word recognition system vocabulary can be identified by the word spotter at search time. Though either system alone can yield respectable retrieval performance, the two methods are complementary and work best in combination. Different ways of combining them are investigated, and it is shown that the best of these can increase retrieval average precision for a speaker-independent retrieval system to 85% of that achieved for full-text transcriptions of the test documents.

## 1 Introduction

Large archives of digitally stored audio and video data are becoming increasingly common, and pose difficult new problems in information retrieval. A fundamental question is how to index the contents of multimedia documents. This paper describes work on a retrieval system for spoken documents where correct prior indexing for all potential search terms cannot be guaranteed. The Video Mail Retrieval (VMR) project is investigating methods to retrieve spoken documents, using a combination of established document retrieval technology and state-of-the-art speech recognition.

In earlier VMR work, spoken document retrieval was based on word spotting (WS) techniques that used a fixed 35-word indexing vocabulary chosen *a priori* for a specific domain [Spärck Jones et al., 1996]. Because this is too limited for realistic tasks, the work presented here investigates large- and open-vocabulary indexing. The two index sources available for this are: a large-vocabulary speech recognition system (LVR), and a word identification system based on a phone-lattice scanner (PLS). Each index source has particular advantages, but also drawbacks: a main contribution of this paper is to contrast these advantages and disadvantages, and to show how the different indexing strategies may be combined for optimal retrieval performance.

Section 2 summarises the approaches to audio document retrieval presented in this paper, and compares them with previous work. Section 3 describes the principal retrieval problem to be solved, as well as the corpus of video mail messages used for experimental work. Section 4 describes the three speech-recognition techniques used to generate index terms in more detail, while Section 5 covers the indexing and retrieval methods investigated, including ways of combining the different information sources. Section 6 presents retrieval experiments using the different indexing sources. Finally, in Section 7, we assess the results and outline our future research.

## 2 Overview

It is by no means clear how best to locate index terms in acoustic data and how to derive the inverted files needed for searching. Our work so far assumes that words, as in text, are appropriate indexing terms (but it should be noted that the shorter the word, typically the poorer the spoken word recognition). This paper refers to three approaches to finding words in audio documents, as follows. If frequently used or important terms are known in advance of search time, then *word spotting* (WS) for these keywords may be used to good effect; however (especially with ad hoc queries) there will often be search terms not in the keyword list. *Large vocabulary recognition* (LVR) exploiting a reference word list supported by a language model (giving word sequence likelihoods) may be used to find any index terms within its vocabulary (which although large is not open). Unfortunately, this is difficult and computationally costly. For example, language models require training corpora of literally millions of words. Such corpora can be obtained for formal domains such as news material relatively easily, since suitable text is available; but obtaining them for informal conversational speech is far more costly and problematic, since explicit (and difficult) transcription to text is needed. Also, even with appropriate models LVR systems require significant computational resources for recognition. An alternative approach is *phone lattice scanning* (PLS) which can be used to find arbitrary terms, consisting of any sequence of phones, albeit with less accuracy. Thus PLS is particularly useful for locating search terms such as proper nouns which are unlikely to appear in the vocabulary of an LVR system.

To appear in the 19th Annual International ACM SIGIR 96 Conference, Zurich, Switzerland, August 1996.

### 2.1 Previous Work

Previously reported VMR project work has used WS techniques to perform retrieval for a small, carefully chosen, fixed keyword vocabulary, both for known speakers [Spärck Jones et al., 1996] and for unknown ones [Foote et al., 1995]. These experiments showed good retrieval performance when compared with text transcriptions of the test documents (about 90% of the text average precision), given that only the predefined keywords were used in both queries and messages.

Other researchers using WS techniques have developed systems to classify documents into a small number of categories[1] defined *a priori*. Examples of this approach are described in [Rose, 1991], [McDonough et al., 1994], [Nowell & Moore, 1995], and [Wright et al., 1995]. The word spotter vocabulary is selected in a post-hoc fashion specifically to optimise classification, where the parameters of the classifier are computed from training data. These systems are clearly limited in their application since they require not only that retrieval classification categories are specified in advance, but also that suitable data is available to train the system parameters.

### 2.2 Indexing problems for spoken documents

Speech retrieval using conventional WS is clearly limited both by the small number of practical search terms, and by the need to specify those terms in advance. But more flexible, open searching without a predetermined query vocabulary presents problems because of the difficulty of the word identification task. Word recognition models must first be trained on large amounts of example speech data; this can be computationally expensive given the typically large number of model parameters and the iterative training methods required. Even with trained models, the search necessary for finding the most likely model sequence (and thus the recognition results, given unknown speech data) is costly, and is impractical to do significantly faster than real time.

These restrictions place severe constraints on spoken document retrieval. Even if models exist for the terms in a given query, performing query-time recognition on a large archive is simply not practical: the computationally expensive recognition must be done in advance of the query. It is now becoming possible to use large-vocabulary speech recognition to provide *quasi-transcriptions* of spoken documents, even though recognition is still costly and prone to errors. These transcriptions can be treated as full text transcriptions and processed accordingly. For the work reported here we used an existing large vocabulary, speaker-independent continuous speech recogniser to generate quasi-transcriptions of the speech documents. The inverted files needed for actual searching were derived from these. The recogniser has a vocabulary of 20,000 words and was originally designed for the ARPA Wall Street Journal dictation task [Young et al., 1994]. It attempts to produce a complete transcription of the speech files, so the output of the recogniser maps all acoustic events to one of the 20K vocabulary words (or silence). But as a result, any spoken document words not in the vocabulary will be misrecognised, and so will not be available to match query terms correctly.

Though increasing the vocabulary size will reduce the number of out-of-vocabulary (OOV) query terms, there will inevitably be terms (such as proper names) not to be found in any vocabulary of practical size for speech recognition. One solution to this problem is to decompose terms into standard subword units, namely *phones*; for example the word "phone" is composed of an initial fricative /f/, a vowel /oh/, and a final nasal /n/. The phone sequence for any index term may be determined from a dictionary or by a rule-based algorithm. With a simple phone recogniser that attempts to transcribe all phones without taking into account the actual words, the result may be rapidly searched, at query time, for the particular phone sequence comprising a search term. (A technique of this type has been investigated at ETH Zürich [Wechsler & Schäuble, 1995].) The main drawback to this method that it is extremely difficult to produce a perfect phone transcription: the best automatic system to date is little better than 70% accurate [Robinson et al., 1994], which severely limits the ability to accurately find arbitrary phone sequences since even one phone recognition error will result in a search miss. Another method developed by [James, 1995] uses a more sophisticated recogniser to generate a phone lattice that contains multiple phone hypotheses. Thus when the lattice is searched for query terms several different phones will be considered at each point. Often many putative term occurrences will be found, but unfortunately a large number of these will be false alarms. The work presented here uses this lattice-based technique but seeks to improve retrieval performance.

[1] termed "topics" in the literature

## 3 The VMR Task

The VMR project data and tests so far have been described in [Spärck Jones et al., 1996, Foote et al., 1995]: we summarise only the parts relevant to the new work reported here. The primary goal of the VMR project is to design a video mail retrieval application for the Medusa multimedia environment developed at Olivetti Research Ltd. in Cambridge, UK [Wray et al., 1994]. Desktop microphones and cameras enable Medusa users to record, send, and archive video mail. However because most of the desired information is to be found in the audio record rather than the video, the VMR project is concentrating on extracting information from audio alone.

### 3.1 The VMR message corpus

Because there was no available video mail corpus and existing speech corpora were not suited to retrieval experiments, we had to create an archive of messages with known audio and information characteristics in order to evaluate both word recognition and message retrieval performance. This VMR1 corpus is described in detail in [Jones et al., 1994]. Ten broad subject categories were chosen to reflect the anticipated messages of Medusa users, including, for example, "management" and "equipment." The initial domain-dependent indexing used small vocabulary WS and a fixed set of 35 keywords was therefore provided for the ten categories; thus the keywords "staff," "time," and "meeting" refer to the "management" category (though keyword-category assignment is not exclusive). 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").

Fifteen speakers (11 men and 4 women) each provided about 45 minutes of speech data for a total of 5 hours of read training data and 5 hours of spontaneous speech messages. The acoustic training data consisted of isolated keywords, read sentences containing keywords in context, and

phonetically-rich sentences not containing keywords. For the message data, each speaker provided 20 spontaneous speech messages in response to 5 prompts chosen from 4 categories. The resulting 300 messages, along with their text transcriptions, served as a test corpus for the retrieval experiments presented later. The messages, though prompted, 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"). The messages were fully transcribed by hand, including non-speech events such as lip smacks, hesitations, and disfluencies. Basic punctuation was also added for ease of reading. These full transcriptions were used to evaluate both speech recognition and retrieval performance.

Data was recorded at a 16 kHz sampling rate, from a Sennheiser HMD 414 head-mounted microphone and the Medusa system desk-mounted microphone. For speech model training and recognition, the acoustic data was parameterized into a spectral representation at a 100 Hz frame rate. Experiments described here use only the data recorded using the Sennheiser head microphone. Experiments reported elsewhere indicate that, while somewhat degraded with respect to the head-mounted microphone, reasonable retrieval performance is achieved with the desk microphone [Jones et al., 1995b].

The VMR1 message set is very small by text retrieval standards, but as an experimental corpus for spoken document retrieval it compares respectably with [Wechsler & Schäuble, 1995, McDonough et al., 1994], and is also comparable with speech processing test data as used for ARPA experiments [Young et al., 1994].

### 3.2 Retrieval Collection VMR1b

For the retrieval experiments reported in this paper we used VMR Collection 1b. This consists of the 300-message corpus with a set of 50 requests and relevance assessments. VMR1b was obtained by asking users to generate natural requests as stimulated by a prompt for each message category. The users were asked to include at least one of the 35 keywords. A suitable relevance assessment subset was formed by combining the 30 messages in the category to which the original message prompt belonged, with the 5 messages from outside the category having the highest text retrieval scores. This gave 10.8 highly relevant documents on average per request. (The formation of this collection is fully described in [Jones et al., 1995a]: having to construct rather than select a test collection is regrettable but was unavoidable, and the test collection does share important properties with real test sets, eg variable topic overlap between documents.)

The requests average 12.0 words. After removing the standard van Rijsbergen stop words [van Rijsbergen, 1979], an average of 7.4 content words remain. On average 6.6 of the words (subsuming nearly all the keywords) are found in the 20K vocabulary, while keyword-only queries contain only 2.7 terms on average. We have already shown that performance improves when keywords are supplemented by large vocabulary terms [Jones et al., 1996]. Adding terms outside either the LVR 20K or keyword vocabulary, denoted out-of-vocabulary or OOV terms, should further improve retrieval performance. Even though for this collection there is only about 1 such term per query, OOV terms are likely to be domain specific and hence potentially useful.

## 4 Speech Recognition Techniques

All the speech recognition work described here exploits Hidden Markov Models (HMMs), which are a widely-used and successful method of speech recognition [Rabiner, 1989]. A hidden Markov model is a statistical representation of a speech event such as a word or phone. HMM parameters are typically trained on a large corpus of labelled speech data. Given a trained set of HMMs, there exists an efficient algorithm for finding the most likely model sequence (the recognised words), given unknown speech data. HMMs were used in WS, LVR and PLS systems, exploiting training data supplying *acoustic models* for individual phones, and also (for LVR) *language models* for words. The speech recognition systems for the work reported here used the HTK tool set developed at Cambridge University [Young et al., 1993]. This is a powerful and flexible set of software tools for developing HMM applications such as those used here. All the recognition techniques described below deliver an *acoustic score*, the log-likelihood that the observed sound or sound sequence is actually an instance of the matching phone or word model.

### 4.1 Word Spotting for Keywords (WS)

Although this paper concentrates on LVR and PLS systems, a short description of our WS system for fixed keywords is given here for reference. In the WS system, each keyword is modelled by concatenating the appropriate sequence of subword models (obtained from a phonetic dictionary). Phones vary depending on acoustic context and, as will be demonstrated, using context-dependent phone models can improve recognition performance. Hence biphones were used at the beginning and end of keywords, while triphones model their internal structure. For example, the keyword "find" is represented by the model sequence `f+ay f-ay+n ay-n+d n-d`. Keyword models were constructed from a set of 8-mixture word-internal tied-state triphone HMMs trained on the WSJCAM0 British English speech corpus [Robinson et al., 1995] using a tree-based state clustering technique [Young et al., 1994]. Non-keyword speech is modelled by an unconstrained parallel network of monophones (denoted "filler models"). Thus all speech is recognised as either a keyword or a phone from the filler network.

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 pass is done using a network of the keywords, silence, and filler models in parallel. In a manner similar to [Rose, 1991], keywords are rescored by normalising each hypothesis score by the the average filler model score over the keyword interval. This helps ensure that true keyword hits have scores greater than false alarms. Because low-scoring words are more likely to be false alarms, the operating point of the recognition system may be adjusted by ignoring words with a score below a given threshold.

The accuracy of a word spotter thus depends on its threshold and cannot be expressed as a single number if false alarms are taken into account. An accepted figure-of-merit (FOM) for word spotting is defined as the average percentage of correctly detected words as the threshold is varied from one to ten false alarms per word per hour. (This is quite similar to retrieval average precision, where

precision is averaged as output is varied.) The speaker-independent (SI) keyword spotter resulted in a 69.9% figure of merit (FOM) on the VMR1 data. The VMR corpus is realistic in that it contains speakers with varied backgrounds and accents, but the figure just given is derived from models trained exclusively on British English speakers. With speaker-adapted [Foote et al., 1995] or speaker-dependent (SD) models [Jones et al., 1995b] much better performance can be obtained (eg SD FOM 81.2%), though at the cost of having to adapt or train speaker-specific models, and suitable labelled adaptation data must be available.

### 4.2 Large Vocabulary Recognition (LVR)

Large vocabulary continuous recognition is only now becoming practical [Young et al., 1994]. For LVR experiments, a set of 8-mixture cross-word triphones was trained on the WSJCAM0 British English speech corpus [Robinson et al., 1995]. Ideally a suitable language model would be built using a large transcription archive of material typical of the application domain. Unfortunately since there was no available archive of this type, the standard WSJ 20K bigram language model from MIT Lincoln Labs was used. The WSJ triphone set and bigram language model when taken together yielded a 53% word recognition accuracy rate. This is low compared to read speech, where accuracy rates can exceed 90% in a limited domain, but is respectable given the difficulty of the spontaneous VMR task. Many factors impact recognition performance adversely: the VMR1 corpus has a significant out-of-vocabulary rate of 3.15%, including 4 of the 35 frequently-occurring fixed keywords. The WSJ North American business news language model is highly inappropriate for informal UK English monologues. Also problematic is the exclusively read training data, the spontaneous nature of our test speech, the lack of disfluency modelling for it, and its non-uniform accents (British, American, and Middle European) [Jeanrenaud et al., 1995]. Work is underway on developing a more appropriate language model, on adapting acoustic models to different accents, and on accounting for spontaneous speech phenomena. However, even the imperfect recognition of the existing system results in respectable retrieval performance.

### 4.3 Phone Lattice-based Word Spotting (PLS)

| | WS | | PLS | | |
|---|---|---|---|---|---|
| | SD | SI | SD mo | SI mo | SI bi |
| FOM | 81.2% | 69.9% | 73.6% | 48.0% | 60.4% |

Table 1: FOM summary for WS systems.

The PLS word spotting technique involves searching a phone lattice for the sequence of phones corresponding to a particular search term [James & Young, 1994]. A phone lattice is a directed acyclic graph whose nodes consist of start/end times, and whose arcs are putative phone occurrences, which are labelled with the phone's acoustic score. Phone lattices may be computed in advance, and rapidly scanned for an arbitrary phone sequence at search time. For the experiments reported here, separate phone lattices were generated using three model sets: 8-mixture SD monophones, 8-mixture SI monophones, and 8-mixture SI biphones. Bigram phone transition probabilities were enforced in a null-grammar network. Phone transition probabilities were trained using the transcriptions of the "z" data taken from VMR1. A phone lattice is generated from the $n$-best paths through a HMM network, given the network, models, and (unknown) acoustic data. The value of $n$ controls the number of simultaneous hypotheses that may end at a given time; thus $n$ controls the average lattice depth, and the number of possible paths through it.

At search time, the phone lattice is scanned for the phone sequence corresponding to the query search term. The phonetic composition of a search term is derived from a dictionary (British English pronunciations are taken from the Oxford Learner's Dictionary). Once a phone sequence is found (corresponding to a putative occurrence of the search term), an acoustic score for the term is estimated from the time-normalised scores of the component phones. In general, phone-lattice spotting accuracy is much poorer than the fixed-keyword spotting described earlier. For each model set the FOM for PLS was computed for the same 35 keywords as used for the standard WS system. FOM values for the three PLS systems are shown in Table 1; also shown for comparison are results for SD and SI fixed keyword WS systems. Whilst the PLS figures are lower than those for the WS system, it is important to remember that the phone lattices are completely general and that any set of words can be searched for without further speech recognition effort.

(Note that the PLS system works with exact word forms, not stems, so is not strictly comparable with our current implementation of WS and LVR; the effect in the tests reported here is likely to be negligible.)

### 4.4 Thresholding

Acoustic word spotting using either WS or PLS is prone both to missed words and to false alarms when seeking search terms, while LVR will wrongly transcribe a significant number of document words. All these errors will adversely affect retrieval performance. The degradation due to imperfect recognition can be measured by comparing retrieval performance on recognised speech with that for text transcriptions of the spoken documents. A particular problem with word spotting is that unrelated acoustic events will often resemble valid words. For example, the last part of "hello Kate" is acoustically quite similar to the keyword "locate." Moreover because even the most accurate acoustic models cannot discriminate between homophones, the output of an ideal word spotter that reports all word phone sequences - so-called *phonetic text* - provides a more legitimate standard of comparison than normal transcribed text. Previous checks have shown spoken document retrieval performance is relatively better when calibrated in this way, but the difference between using regular and phonetic transcriptions is not large and we have therefore used only the former here.

As mentioned earlier, when using WS output in an application a threshold is normally set on the acoustic score. Words with scores above the threshold are considered true hits, while those with scores below are considered false alarms and ignored. Choosing the appropriate threshold is a tradeoff between the number of Type I (missed words) and Type II (false alarm) errors, with the usual problem that reducing one increases the other. Retrieval performance varies with the choice of score threshold. At low threshold values, performance is somewhat impaired by a high proportion of false alarms (Type II errors); conversely, higher thresholds remove a significant number of true hits (Type I errors), also degrading performance. In our work thresholding was applied to both WS and PLS output hypotheses, but it is not

required for LVR because there are few false alarms.

## 5 Information Retrieval Techniques

Because there is always some uncertainty about whether a putative word is actually present in a spoken document, the sense and category ambiguities familiar in text retrieval are compounded, making speech retrieval more problematic than text retrieval. Some ambiguities may be resolved in the speech case by different pronunciations, but more ambiguities arise from homophones and misrecognised word boundaries. However as with text retrieval, redundancy can be exploited to reduce these uncertainties and ambiguities. Large vocabulary recognition is thus important for spoken document retrieval, compared with keyword spotting, as it significantly increases the number of potential matching keys. It can nevertheless still miss term occurrences. But phone lattice spotting can make more terms (such as proper names) available for matching. It is not immediately obvious, however, how to best combine the LVR and PLS index sources.

### 5.1 Indexing Methodology

For our experiments, standard indexing and matching techniques were applied both to the text transcription files and to the quasi-transcriptions generated by the speech recognition engines. Performance for the text transcriptions could then be used as a reference standard for the various speech retrieval strategies. The LVR output files were first processed to remove stop words. Several stop word lists were investigated but best results were obtained using the standard van Rijsbergen list [van Rijsbergen, 1979]. (The fixed keyword WS and PLS only return putative hits from the fixed vocabulary or the current query respectively and so contain no stop words.) Next all query terms and hypothesised document contents from all sources were suffix stripped using the Porter algorithm [Porter, 1980].

Retrieval tests compared *unweighted uw* matching performance with two forms of weighting. These were the standard *collection frequency weight cfw* (also called inverse document frequency weight), and the *combined weight cw* that incorporates within-document term frequencies and is normalised for document length (defined in [Robertson & Spärck Jones, 1994] and derived in [Robertson & Walker, 1994]; the *cw* scheme reflects the City University work for TREC [Robertson et al., 1995]). The *cw* weight for each term in each document is calculated as follows:

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

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. $ndl(j)$ is calculated as

$$ndl(j) = \frac{dl(j)}{\text{Average } dl \text{ for all documents}},$$

where $dl(j)$ is the total length of $j$. The combined weight constant $K$ has to be tuned empirically: after informal testing a value $K = 1$ was selected.

#### 5.1.1 Document Length $dl(j)$

The document length is ordinarily measured as the number of term occurrences in the document. This measure of $dl(j)$ is suitable for text and LVR where full transcriptions are available. However, for the PLS system the document is represented only by the search terms found for the current query, which may not be a good representation of the document length. But since $ndl(j)$ is the ratio between different document lengths, the absolute length is not important and alternative measures of $dl(j)$ can be considered. In our PLS tests we examined two alternative measures of $dl(j)$:

- the number of phones found in the most likely phone path, which is easily computed using the Viterbi algorithm during the speech recognition phase. We reason that on average the number of phones in a document is representative of the number of words.
- the total length of the document in seconds.

### 5.2 Index Combination Methods

Combining multiple information sources has been shown to improve text-based retrieval systems, for example in [Belkin et al., 1995]'s comparative TREC-2 study. Belkin et al. considered two approaches to information combination, referred to as *query combination* and *data fusion*. In query combination, multiple queries for the same information need are merged into a single query, from which a single ranked output list is generated. 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** In our data fusion work, matching scores for documents that have been computed independently by different indexing systems are added to form a final composite score. Since it is not clear whether scores for types of source are commensurable, we tried both normalising with respect to the highest scoring document in each list, and leaving scores as they were. With or without normalisation, the result is a new ranked list using the composite scores.

**Data Merging** In our data merging strategies, evidence from different indexing sources is combined in a way analogous to Belkin et al's query combination. Specifically, word hypotheses from the indexing systems are merged for a single document before computing the document's matching score. Hypotheses from the LVR output may be either augmented with all putative hits from a word spotter (PLS or WS), or only with those outside the WSJ 20K vocabulary. In the first approach, search keys are counted twice if hypothesised by both systems. This is not necessarily a drawback as it may help counteract acoustic stemming problems which may result in LVR misses when the term as instantiated is not in the LVR vocabulary. Because they are frequency based, *cfw* weights may be affected by spurious keys in other documents due to PLS false alarms. Combined weights (*cw* ), which take into account within-document term frequencies, may also be adversely influenced by multiple term counts in addition to these false alarms.

## 6 Retrieval Experiments

In this section we present our experiments in a series of comparisons as follows. First, for reference, text retrieval for all the query terms, versus only those terms in the 20K vocabulary. Second, with the spoken documents, comparisons between LVR alone and PLS alone. Third, between these

| Weight Scheme | Full Vocab. | | | 20K Vocab. | | |
|---|---|---|---|---|---|---|
| | *uw* | *cfw* | *cw* | *uw* | *cfw* | *cw* |
| Prec. | | | | | | |
| 5 docs | 0.392 | 0.375 | 0.371 | 0.346 | 0.313 | 0.321 |
| 10 docs | 0.313 | 0.308 | 0.344 | 0.281 | 0.277 | 0.294 |
| 15 docs | 0.279 | 0.292 | 0.308 | 0.257 | 0.257 | 0.272 |
| 20 docs | 0.250 | 0.271 | 0.290 | 0.227 | 0.242 | 0.258 |
| Av Prec. | 0.327 | 0.352 | 0.368 | 0.299 | 0.312 | 0.325 |

Table 2: Retrieval precision values for Full and WSJ 20K Vocabulary Text Transcriptions.

| Weight Scheme | | 20K LVR | | |
|---|---|---|---|---|
| | | *uw* | *cfw* | *cw* |
| Prec. | 5 docs | 0.254 | 0.271 | 0.300 |
| | 10 docs | 0.213 | 0.238 | 0.248 |
| | 15 docs | 0.204 | 0.219 | 0.242 |
| | 20 docs | 0.184 | 0.193 | 0.220 |
| Av Precision | | 0.225 | 0.246 | 0.263 |

Table 3: Retrieval precision values for WSJ 20K LVR.

individual methods and the different combination methods taking the two together. Detailed retrieval results for fixed-vocabulary WS are contained in [Jones et al., 1995b, Foote et al., 1995, Spärck Jones et al., 1996], and for the combination of WS and LVR in [Jones et al., 1996].

We recognise that with a small test collection specific figures are neither reliable nor significant: we concentrate therefore on the general picture that emerges from the results.

### 6.1 Text Retrieval

For reference, Table 2 shows text retrieval performance using both full open-vocabulary, and using only terms present in the WSJ 20K vocabulary. Results throughout this section show precision at ranked list cutoffs of 5, 10, 15 and 20 documents, and standard TREC average precision. The figures in Table 2 confirm that more sophisticated weighting schemes generally improve retrieval performance, but also show that restricting the vocabulary to 20K impairs performance.

### 6.2 Spoken Document Retrieval

#### 6.2.1 Large Vocabulary or Phone Lattice Retrieval: LVR vs PLS

As discussed earlier, retrieval performance for spoken documents is affected by varying the acoustic score threshold. This is illustrated in Figure 1 using search term hypotheses located in the SD monophone lattices. On the left of the figure, corresponding to low threshold values, retrieval performance 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 optimal threshold, in the central region, represents the best tradeoff between the numbers of true hits and false alarms.

Table 3 shows retrieval performance using the 20K LVR system. Retrieval results for the three sets of PLS phone lattices are shown in Tables 4, 5 and 6. These cover all, not just most, of the query terms as well as using a different

| Weight Scheme | | Speaker-Dependent Monophones | | | | |
|---|---|---|---|---|---|---|
| | | *uw* | *cfw* | *cw* | | |
| | | | | Terms | Time | Phones |
| Prec. | 5 docs | 0.329 | 0.288 | 0.313 | 0.342 | 0.338 |
| | 10 docs | 0.367 | 0.254 | 0.231 | 0.279 | 0.288 |
| | 15 docs | 0.213 | 0.225 | 0.218 | 0.243 | 0.246 |
| | 20 docs | 0.197 | 0.216 | 0.198 | 0.225 | 0.222 |
| Av Precision | | 0.262 | 0.285 | 0.284 | 0.311 | 0.315 |

Table 4: Retrieval precision values for SD monophones.

| Weight Scheme | | Speaker-Independent Monophones | | | | |
|---|---|---|---|---|---|---|
| | | *uw* | *cfw* | *cw* | | |
| | | | | Terms | Time | Phones |
| Prec. | 5 docs | 0.200 | 0.233 | 0.238 | 0.250 | 0.242 |
| | 10 docs | 0.160 | 0.183 | 0.190 | 0.190 | 0.200 |
| | 15 docs | 0.146 | 0.168 | 0.175 | 0.200 | 0.200 |
| | 20 docs | 0.148 | 0.157 | 0.170 | 0.183 | 0.185 |
| Av Precision | | 0.174 | 0.199 | 0.208 | 0.216 | 0.222 |

Table 5: Retrieval precision values for SI monophones.

| Weight Scheme | | Speaker-Independent Biphones | | | | |
|---|---|---|---|---|---|---|
| | | *uw* | *cfw* | *cw* | | |
| | | | | Terms | Time | Phones |
| Prec. | 5 docs | 0.296 | 0.279 | 0.233 | 0.313 | 0.317 |
| | 10 docs | 0.235 | 0.254 | 0.223 | 0.248 | 0.254 |
| | 15 docs | 0.199 | 0.226 | 0.206 | 0.226 | 0.236 |
| | 20 docs | 0.171 | 0.198 | 0.200 | 0.206 | 0.205 |
| Av Precision | | 0.224 | 0.262 | 0.248 | 0.269 | 0.277 |

Table 6: Retrieval precision values for SI biphones.

word identification technique. Lattice results are shown at the *a posteriori* best thresholds: this is clearly unrealistic, and in practice it would be necessary to choose an operating threshold on the basis of experience. However good performance could still be expected. Retrieval performance for PLS varies for the different model sets; performance (not surprisingly) is best with the SD monophone lattices, while for the SI models the more sophisticated biphones perform better than the monophones. The tables also show, for the *cw* scheme, that the estimation of document length is important in achieving effective retrieval. Using only the terms in the query to represent the document is clearly not suitable since performance is marginally worse than for *cfw* weighting alone. Using either the phone count or the duration time both give substantial improvement. However, the phone count is marginally better. This is intuitively reasonable since the phone count is independent of speaking rate and hence is a better estimate of the number of words actually spoken. All further *cw* weighted results were generated using phone count normalisation.

In comparing performance for LVR and PLS, the SD PLS performs better than (SI) LVR, but this is not a useful result for many practical purposes. The SI biphone performance on the other hand, is about the same as for LVR, perhaps because our data set is not large enough to discriminate between them. However, since performance for either is still below that for the text reference, the test results for combination of index sources described below have also to be

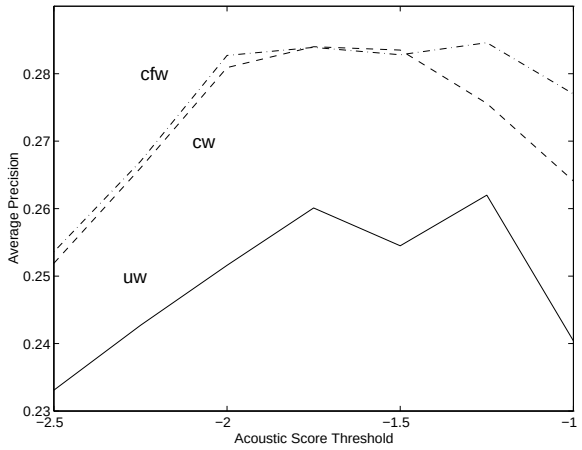


Figure 1: Retrieval performance for SD phone lattice vs. acoustic score threshold.

considered.

Additional experiments with an alternative (though artificial) query set with an average length of 19 terms indicate that LVR will give performance superior to PLS for longer queries. This is because individual search misses in the LVR may be less important than the proportionally high number of false alarms associated with PLS.

#### 6.2.2 Combination Methods: LVR + PLS

The combination experiments described here all used the SI biphone models, since these would normally be taken together with the SI LVR models in the fully SI systems most likely to be required in practice.

**Data Fusion** For the *data fusion* combination, document scores for the separate LVR and PLS sources were added and the document list was ordered by the resulting new scores. We tried two specific methods of data fusion for this, either simple addition (*simp. fuse*) of existing scores, or addition of scores normalised by the maximum score in the current list (*norm. fuse*). In the combination, lattice-derived hypotheses for all query terms were used regardless of whether they were also found using LVR. Ranked lists were taken at the best lattice thresholds, as in the PLS-alone test. Results for this experiment are contained in Table 7 and suggest that data fusion for LVR and PLS can improve retrieval performance by a small amount compared with either alone, but the choice of specific fusion method is immaterial.

**Data Merging** Table 8 illustrates performance for the alternative *data merging* strategy, combining document hypotheses from the LVR and PLS before computing a matching score. In this case we show both the use of lattices for all the query terms, as in the data fusion experiments, and the use of lattice hypotheses only for those query terms not in the 20K vocabulary. Not surprisingly, both approaches are helpful compared with either LVR or PLS alone, with the all-terms version somewhat preferable.

#### 6.2.3 Results Summary

Table 9 shows a summary of results from all the experiments just described. This gives average precision for all systems relative to the ideal open text standard. The conclusions to be drawn from such single-number performance indicators, especially when taken in the context of a small test collection, have to be treated with caution. In particular average precision, based on more information, may sharpen differences that are less apparent for the cutoff data which is likely to be more practically pertinent, However we can observe in the Table that the PLS approach performs slightly better in isolation than the 20K LVR system; that all LVR+PLS combination schemes give improved performance; and that data merging appears to perform better than data fusion as a specific combination technique. Perhaps the most important point is that the best combined performance for the SI biphone lattice is clearly better than performance for either method separately. Moreover compared with the text standards, the best levels of spoken document retrieval performance are quite respectable, reaching $80\% - 85\%$.

Finally we may also note, referring back to earlier VMR work using WS with a fixed keyword vocabulary, that keyword retrieval performance even for the SI case ([Jones et al., 1996]) is broadly comparable with our best results here. This is presumably because, though there are only 2.6 keywords on average per query, they are well chosen for the message set.

## 7 Conclusions and Further Work

The experiments reported here constitute only a first attack on open vocabulary retrieval for spoken documents. In particular, because our test collection is so small, we can only take the comments on individual test results made in the previous section as impressionistic and indicative. Nevertheless, we have shown that it is possible to obtain speech retrieval performance, using open search terms, approaching that obtainable for text. Further, it appears that the combination of two recognition techniques can perform better than either alone, and indeed achieve an average retrieval precision for a SI system degraded by only 15% from the best achievable text retrieval. Moreover this difference can be reduced by further improvements in speech recognition. Thus we have already found, for the SD case, that data combination retrieval performance using the current 20K LVR and PLS but with the SD monophones is only degraded by 7% compared to text.

There is little other work with which direct comparisons can be made. Thus the research on word spotting for categorisation mentioned earlier is not really pertinent. Schäuble [Wechsler & Schäuble, 1995]'s recent work, reported at the MIRO Workshop in September 1995, uses more documents, but they are equal in length and much shorter than ours, so the total amount of speech data is less than ours. There are also fewer requests. This work has explored various alternative methods of handling phone sequences. However, these tests and ours are so different in the data used, as well as the speech processing methods, that a straightforward comparison is impossible.

[James, 1995] reports experiments using all his test collection query terms as keywords along with a phone lattice. His results, for a news data set somewhat smaller than ours, show about the same relative performance for spoken document retrieval against text transcriptions as ours. He has also, like us, combined large vocabulary recognition with phone spotting, giving much improved performance. However, though performance was similar to ours, James's tests relied (unlike ours) on domain-dependent speech and lan-

| Weight Scheme | | Simp. Fuse | | | Norm. Fuse | | |
|---|---|---|---|---|---|---|---|
| | | *uw* | *cfw* | *cw* | *uw* | *cfw* | *cw* |
| Prec. | 5 docs | 0.317 | 0.325 | 0.317 | 0.321 | 0.308 | 0.308 |
| | 10 docs | 0.263 | 0.288 | 0.281 | 0.273 | 0.292 | 0.290 |
| | 15 docs | 0.233 | 0.251 | 0.260 | 0.232 | 0.247 | 0.263 |
| | 20 docs | 0.212 | 0.220 | 0.226 | 0.213 | 0.216 | 0.225 |
| Av Precision | | 0.255 | 0.286 | 0.301 | 0.258 | 0.281 | 0.300 |

Table 7: Retrieval precision values for data fusion combining 20K LVR and SI biphone PLS.

| Weight Scheme | | All Terms | | | OOV Terms | | |
|---|---|---|---|---|---|---|---|
| | | *uw* | *cfw* | *cw* | *uw* | *cfw* | *cw* |
| Prec. | 5 docs | 0.296 | 0.329 | 0.333 | 0.375 | 0.317 | 0.338 |
| | 10 docs | 0.250 | 0.283 | 0.292 | 0.242 | 0.277 | 0.288 |
| | 15 docs | 0.208 | 0.246 | 0.264 | 0.211 | 0.242 | 0.261 |
| | 20 docs | 0.184 | 0.214 | 0.239 | 0.199 | 0.216 | 0.237 |
| Av Precision | | 0.240 | 0.281 | 0.315 | 0.239 | 0.274 | 0.310 |

Table 8: Retrieval precision values for data merging combining 20K LVR and SI biphone PLS.

guage modelling. Spoken document retrieval in a video context is also being developed by the Informedia project at Carnegie Mellon University [Smith & Christel, 1995]. This is using established speech recognition and retrieval engines, much as we do. But with newscast data, for example, searching may be leveraged from accompanying text captions, and there is no reported formal performance evaluation yet.

The specific work reported in this paper needs to be followed up in several ways.

It is essential to conduct retrieval tests on a larger scale, and we have begun work on data capture and system development for television newscast retrieval.

At the same time, the approaches we have described must be developed to support a near real-time system. Even though expensive recognition is done offline, issues of storage and search efficiency must be addressed to yield a practical system. This is necessary both for larger-scale experiments and for operational use. Another problem to address is the use of desktop microphones rather than cumbersome head-mounted ones. We have already carried out work [Brown et al., 1994] on providing a suitable user interface with browsing facilities.

Fortunately for those concerned with spoken document retrieval, performance will continue to get better as the underlying speech recognition technology is improved. More sophisticated and efficient decoders mean that larger vocabularies may be used, which should reduce the OOV problem. We are developing more appropriate language models and vocabularies than those used for our experiments so far, using the British National Corpus [Burnard, 1995]. These will be used for further experiments in using LVR for retrieval. Improvements in speech recognition can only benefit spoken document retrieval.

## 8 Acknowledgements

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

## References

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

[Brown et al., 1994] Brown, M. G., Foote, J. T., Jones, G. J. F., Spärck Jones, K., & Young, S. J. (1994). Video Mail Retrieval using Voice: An overview of the Cambridge/Olivetti retrieval system. In *Proc. ACM Multimedia 94 Workshop on Multimedia Database Management Systems*, pp. 47–55, San Francisco, CA.

[Burnard, 1995] Burnard, L., editor (1995). *Users Reference Guide for the British National Corpus*. Oxford University Computing Services, Oxford.

[Foote et al., 1995] Foote, J. T., Jones, G. J. F., Spärck Jones, K., & Young, S. J. (1995). Talker-independent keyword spotting for information retrieval. In *Proc. Eurospeech 95*, volume 3, pp. 2145–2148, Madrid. ESCA.

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

[James & Young, 1994] James, D. A., & Young, S. J. (1994). A fast lattice-based approach to vocabulary independent wordspotting. In *Proc. ICASSP 94*, volume I, pp. 377–380, Adelaide. IEEE.

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

[Jones et al., 1995a] Jones, G. J. F., Foote, J. T., & Spärck Jones, K. (1995a). Video Mail Retrieval using voice: Report on collection of naturalistic requests and relevance assessments. VMR Project Working Document.

| Weight Scheme | | | Average Precision | | |
|---|---|---|---|---|---|
| | | | *uw* | *cfw* | *cw* |
| Text | Full Vocab | Avg. Prec. | 0.327 | 0.352 | 0.368 |
| | | (relative) | 100% | 100% | 100% |
| | 20K Vocab | | 91.5% | 88.8% | 88.3% |
| Spoken Documents | 20K LVR | | 68.8% | 69.8% | 71.5% |
| | PLS | SD Monophones | 80.1% | 81.0% | 85.6% |
| | | SI Monophones | 53.2% | 56.5% | 60.3% |
| | | SI Biphones | 68.5% | 74.4% | 75.3% |
| | LVR + PLS SI Biphone Data Fusion | Simp. Fuse | 78.0% | 81.3% | 81.8% |
| | | Norm. Fuse | 78.9% | 79.8% | 81.5% |
| | LVR + PLS SI Biphone Data Merging | 20K + all | 73.4% | 79.8% | 85.6% |
| | | 20K + OOV | 73.1% | 77.8% | 84.2% |

Table 9: Relative average precision from all experiments.


[Jones et al., 1994] Jones, G. J. F., Foote, J. T., Spärck Jones, K., & Young, S. J. (1994). VMR report on keyword definition and data collection. Technical Report 335, Cambridge University Computer Laboratory.

[Jones et al., 1995b] Jones, G. J. F., Foote, J. T., Spärck Jones, K., & Young, S. J. (1995b). Video Mail Retrieval: the effect of word spotting accuracy on precision. In *Proc. ICASSP 95*, volume I, pp. 309–312, Detroit. IEEE.

[Jones et al., 1996] Jones, G. J. F., Foote, J. T., Spärck Jones, K., & Young, S. J. (1996). Robust talker-independent audio document retrieval. In *Proc. ICASSP 96*, Atlanta, GA.

[McDonough et al., 1994] McDonough, J., Ng, K., Jeanrenaud, P., Gish, H., & Rohlicek, J. R. (1994). Approaches to topic identification on the switchboard corpus. In *Proc. ICASSP 94*, volume I, pp. 385–388, Adelaide. IEEE.

[Nowell & Moore, 1995] Nowell, P., & Moore, R. K. (1995). The application of dynamic programming techniques to non–word based topic spotting. In *Proc. Eurospeech 95*, volume 2, pp. 1355–1358, Madrid. ESCA.

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

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2),257–286.

[Robertson & Spärck Jones, 1994] Robertson, S. E., & Spärck Jones, K. (1994). Simple, proven approaches to text retrieval. Technical report, Cambridge University Computer Laboratory.

[Robertson & Walker, 1994] Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2–Poisson model for probabililtic weighted retrieval. In *Proc. SIGIR 94*, pp. 232–241, Dublin. ACM.

[Robertson et al., 1995] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). Okapi at TREC-3. In Harman, D. K., editor, *The Third Text REtrieval Conference (TREC-3)*, pp. 109–126.

[Robinson et al., 1995] Robinson, T., Fransen, J., Pye, D., Foote, J., & Renals, S. (1995). WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In *Proc. ICASSP 95*, pp. 81–84, Detroit. IEEE.

[Robinson et al., 1994] Robinson, T., Hochberg, M., & Renals, S. (1994). IPA: Improved phone modelling with recurrent neural networks. In *Proc. ICASSP 94*, volume 1, pp. 37–40, Adelaide, SA.

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

[Smith & Christel, 1995] Smith, M. A., & Christel, M. G. (1995). Automating the creation of a digital video library. In *Proc. ACM Multimedia 95*, pp. 357–358, San Francisco. ACM.

[Spärck Jones et al., 1996] Spärck Jones, K., Jones, G. J. F., Foote, J. T., & Young, S. J. (1996). Experiments in spoken document retrieval. *Information Processing and Management*. (In press).

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

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

[Wray et al., 1994] Wray, S., Glauert, T., & Hopper, A. (1994). The Medusa applications environment. In *Proceedings IEEE International Conference on Multimedia Computing and Systems*, pp. 265–273, Boston. IEEE.

[Wright et al., 1995] Wright, J. H., Carey, M. J., & Parris, E. S. (1995). Topic discrimination using higher–order statistical models of spotted keywords. *Computer Speech and Language*, 9(4),381–405.

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

[Young et al., 1993] Young, S. J., Woodland, P. C., & Byrne, W. J. (1993). *HTK: Hidden Markov Model Toolkit V1.5*. Entropic Research Laboratories, Inc., 600 Pennsylvania Ave. SE, Suite 202, Washington, DC 20003 USA.