A Critical Review of State-Of-The-Art Technologies for Cross-Language Speech Retrieval

Gareth J. F. Jones
Department of Computer Science
University of Exeter
Exeter EX4 4PT, U.K.
gareth@dcs.ex.ac.uk

David A. James
Ubilab
Union Bank of Switzerland
Zurich CH-8021, Switzerland
David.James@ubs.com

Abstract

Recent developments in monolingual speech retrieval, automatic speech recognition and cross-language text retrieval suggest that *cross-language speech retrieval* deserves some enquiry. To direct work in this new area, it is important to appreciate the possibilities and limitations of the component technologies. This paper gives a critical review of the state-of-the-art of technologies for speech retrieval in cross-language environments and outlines some possible experimental paradigms.

Introduction

Rapid developments in computing technology are bringing about new areas of research. One new research area has been the integration of information retrieval and spoken language processing to allow *monolingual* speech retrieval. This paper presents a critical review of the possible extension of this existing work to *cross-language* speech retrieval environments where queries[1] in one language can be used to retrieve documents in one or more others. Since this is a new area of research, we consider some possible definitions of cross-language speech retrieval (CLSR), and review the availability of technology to support these proposed systems. We then offer some thoughts about building systems and environments in which they might be deployed, and finally explore issues in the important area of system evaluation.

What do we mean by CLSR?

Before discussing possible systems, we need to establish what is meant by CLSR. This is not trivial, as even cross-language *text* retrieval (CLTR) can be interpreted to mean a number of different things. For example, if x and y represent known individual languages, and X and Y represent known groups of languages any of which may be observed at any time, the following CLTR scenarios may arise.

	Query Language	Document Language
(i)	x	x
(ii)	x	y
(iii)	x	Y
(iv)	X	y
(v)	X	Y

where (i) represents the trivial case (from a CLTR viewpoint) of monolingual text retrieval. The remaining scenarios shown are progressively less constrained and hence require increasingly complex language processing and retrieval technology.

The possible definitions of CLSR extend these existing paradigms. For each CLTR scenario the documents or queries may be spoken rather than textual items. Some examples of spoken documents in multiple languages might include: the output of multiple differing television, or simultaneously translated proceedings of international government groups like the United Nations or the European Union. Alternatively, a spoken query could be used to retrieve documents from a textual collection written in a different language.

It is not clear whether all possible permutations are likely. Certainly, as will hopefully become clear, the detailed experimental evaluation of each of these scenarios would be extremely costly. For simplicity here, we confine our discussion to the situation of monolingual textual queries to a cross-language spoken document archive and cross-language spoken queries to a monolingual text archive.

[1] For the purposes of this paper we do not distinguish between a *request*, the searcher's original expression of information need; and a *query*, the possibly transformed input to the retrieval engine. We assume both to be in the same language which we henceforth refer to as the *query language*.

Current Research Activities

The implementation of a CLSR system requires contributions from several related areas of research. For the purposes of this discussion we concentrate on automatic speech recognition, cross-language text retrieval and monolingual speech retrieval. Since these are all active research areas in their own right, this section reviews the state-of-the-art in these existing technologies and highlights pertinent features for research in CLSR. We shall not attempt to give an overview here of the textual cross-lingual retrieval world; instead, the reader is directed to Oard's excellent summary (OD96). Instead we concentrate here on existing work in monolingual speech retrieval and relevant spoken language technologies.

Monolingual Speech Retrieval

Monolingual speech retrieval combines speech recognition and monolingual information retrieval techniques. Research into the combination of these technologies has been going on for more than 5 years. An important distinction here is between work focusing on the retrieval of spoken documents, and spoken input for the retrieval of textual documents. The role of speech recognition systems in speech retrieval is to generate appropriate representations of the spoken data (whether it is the documents or the queries which are spoken).

Retrieval of Spoken Documents The central problem in retrieval of spoken documents is that document indexing cannot be carried out when the terms of a query are entered since recognition is on the order of real-time and often many times slower, depending on the complexity of the recogniser. Thus it is necessary to perform as much indexing as possible in advance of retrieval, thereby making the indexing process as query-independent as possible. The precise form of speech recognition to be used depends on the application, and the following review of existing work in retrieval of spoken documents gives examples of all methods investigated to date.

The earliest reported work in retrieval of spoken documents was by Rose (RCL91) in his work at the MIT Lincoln Laboratory. He used a speech recognition system to spot words from a relatively small predefined vocabulary (126 words). This speech recognition system, referred to as a *word spotter*, attempts to locate instances of these words in any spoken data presented to it. All other words present in the spoken material are outside the scope of the recognition process and are hence unavailable for indexing the content. Rose used the output of the word spotter to classify documents using pre-defined weights into one of 10 highly orthogonal static categories (termed *topics* in the speech literature). Later work by researchers at BBN (MNJ$^+$94) and Dragon (PCG$^+$96) reported related methods of achieving a similar objective using larger vocabularies.

The first work to consider the use of classical information retrieval (IR) methodology for spoken document retrieval was reported by Glavitsch and Schäuble of ETH Zürich in (GS92). In this work a selected set of 1000 sub-word features were used as the indexing vocabulary. The features were selected on the basis of frequency of occurrence within a text training corpus. The motivation for using these features instead of words was the presence of a very high number of compound nouns in German; the feature set would allow any German speech to be indexed independently of the actual vocabulary present. Thus provided the documents have been indexed in advance of retrieval, any terms can be entered as part of a query, the features extracted and then searched for in the document archive. Simulated results on existing text collections showed the potential of this method. In more recent work, Wechsler and Schäuble (WS95) have reported work on the retrieval of Swiss German radio news broadcasts using an alternative approach to subword indexing based on phone n–grams.

The first experimental study into the retrieval of spoken documents using classical IR and speech recognition on genuine spoken documents was performed at Cambridge University by James (Jam95). In this work a system for the retrieval of BBC radio news broadcasts was developed and evaluated. The spoken documents were indexed using a combination of a speaker and domain *dependent* medium vocabulary (3000 word vocabulary) speech recognition system and a subword *phone-lattice* to search for words outside of the recognition vocabulary. The medium vocabulary recognition system attempts to transcribe the complete contents of the spoken documents. In practice, many recognition errors arose due to the occurrence of words not contained in the recogniser vocabulary. The subword phone-lattice is a fast-searchable intermediate representation of the phonetic content of a spoken document. It takes the form of an acyclic graph where at any time multiple different phone hypotheses are represented. The output of the medium vocabulary speech recognition system can be treated as text at retrieval time, whereas the phone-lattice can be scanned several orders of magnitude faster than real-time for phone-strings corresponding to search terms. James used the output of these separate recognition systems in combination to realise better retrieval performance than either system in isolation. The best retrieval performance obtained on his test corpus of BBC radio news broadcasts using this system was 93% of standard TREC average precision relative to a perfect text transcription. James's work was extended in the Video Mail Retrieval Using Voice (VMR) project, also at Cambridge. The VMR project developed a domain and speaker *independent* system using a 20,000 word large vocabulary speech recognition (LVR) system and the phone-lattice method for the retrieval of voice and video mail. VMR demonstrated domain and speaker independent retrieval performance for an experimental collection of video mail messages and search request of 86% relative to manual text transcription of the messages (JFS$^+$96). Users of a deployed VMR demonstration system utilising only the phone-lattice indexing method found retrieval performance to be effective despite the errors in indexing (BFJ$^+$96).

We are not aware of research in retrieval of spoken documents in languages other than English and German.

Spoken Querying of Text Archives There has been similarly little work in this potentially important field. Kupiec (KKB94) reported a system which made use of so-called *N-best* speech recognition, which generates multiple word hypotheses for unknown speech in spoken queries, and the concept of *semantic cooccurrence filtering (SCF)* for the retrieval of text documents. He argued that it was possible to compensate for speech recogniser errors by generating multiple hypotheses for the spoken query terms. His argument was that recognition errors would be randomly distributed, whereas the correct recognition output would be semantically related and well-correlated to the contents of relevant documents, and hence retrieval performance would not suffer badly as a result of recognition errors. The principle of SCF is very powerful and is related to the standard information retrieval principle of redundancy.

A study at University of Massachusetts at Amherst (Cro95) found that when TREC-3 queries were spoken to a standard large vocabulary recogniser, retrieval performance was only slightly impaired compared to the original text queries. The speech recognition word error rate here was probably around 30%, but the redundancy in the long TREC queries ensured that SCF was able to compensate for the errors.

Spoken Querying of Spoken Documents Currently, the only study looking into spoken query input to retrieve spoken documents is the Informedia Digital Library Project at Carnegie Mellon University (HWC95) (HW97). The Informedia digital library system integrates technologies from the fields of natural language understanding, image processing, speech recognition, video compression and information retrieval to provide a system for multimedia information retrieval. This project has produced impressive working demonstrations of television news retrieval and recent experimental results indicate that (as observed in similar studies (Jam95) (BFJ$^+$96)) retrieval performance is fairly robust to significant errors in spoken document transcriptions generated by automatic speech recognition (HW97).

Speech Processing

In order to carry out CLSR several speech technologies may be required. For example, language identification may be necessary to ensure that a speech recogniser in the appropriate language is activated at any time. Once the language spoken has been identified, it is crucial that the speech recogniser generates a sufficiently accurate searchable representation of the spoken material. The reliable processing of spoken queries and documents is clearly dependent on the availability of suitable speech processing technology. In this section we review the state-of-the-art in potentially useful spoken language technologies for CLSR.

Language Identification If we assume that a CLSR document collection may contain multiple languages, then it is unlikely that the collection will be pre-processed to indicate which items are spoken in which languages. It would therefore be necessary to ensure that the identities of languages spoken throughout the collection are always known, so that the spoken data can be processed using the appropriate speech recogniser. A considerable number of papers in Language Identification (LID), as it is known, have appeared over the last couple of years; this boom can be ascribed, in part, to the release in 1994 of a public domain test corpus for LID experiments. A good overview of LID work up to this time is that of Muthusamy (MBC94).

One way of performing LID on some unknown speech is as follows: having obtained an acoustic-phonetic model (this could be based on a number of differing technologies, such as Hidden Markov Modelling, or neural networks) for each of n languages, the unknown speech is recognised n times using each model. Acoustic matching scores, a by-product of the recognition process, can be examined and the best-fitting model chosen. However, this method assumes that a sufficiency of model training speech is available for each language. Zissman *et al* recently experimented with using a set of American English phonetic models to identify differing dialects of Spanish, owing to a lack of appropriate training data (ZGRL96). Identification was performed by comparing histograms of n-gram sequences of output symbols from the recogniser. Incorporating lexical knowledge into the *n*-recognitions approach leads to the approach of Schultz *et al*, who discriminated between 3 languages using 3 large-vocabulary continuous speech recognisers (SRW96). Good performance was achieved, though at some considerable computational expense. Arguably the most economical approach is to train a small number of phonetic models for so-called *broad classes*, (vowels, nasals, fricatives, etc.), language-dependently or independently, and classify utterances on the basis of the output sequence (Zis93) (TCP94).

Once the contents of a spoken document have been automatically labelled with the languages spoken, speech recognition systems in the appropriate languages can be used to index the contents of the document.

Multilingual Speech Recognition For CLSR a speech recognition system must be available in any language in which spoken documents or queries appear. It has been noted previously that various approaches to speech recognition may be useful in CLSR. For example, we considered transcription using large vocabulary speech recognition and indexing of individual interesting content words using word spotting. Speech recognition research is currently mainly focused on LVR systems, however many of the techniques and the acoustic and linguistic data collected for this could be used in the other approaches to indexing of acoustic documents. This is important since much of the demonstrated progress in speech recognition research over recent years has been fostered by the availability of large commonly used corpora for system training and evaluation.

The majority of work in automatic speech recognition has focused on American English material. A major reason for this, and a major influence on research direction over the last five years, has been the annual ARPA system evaluation trials. Under this scheme speech researchers at laboratories in the USA and Europe have competed annually to achieve the lowest word recognition error rate on prescribed tasks. There are many similarities in operational method between the ARPA speech recognition evaluation programme and the NIST TREC text retrieval programme. Participants are provided with training, development and test data with which to develop and evaluate their systems. Over the lifetime of these trials, systems have advanced rapidly from recognition of individual sentences of spoken newspaper text with a vocabulary restricted to 5,000 words recorded in a quiet room (taken from the American *Wall Street Journal* (WSJ) newspaper), to transcription of a variety of US radio and television programmes with an unrestricted vocabulary and difficult features such as unmarked multiple speakers, varied quality of the speech channel, and background music. The transcription accuracy is naturally related to the difficulty of the task. For example, in the 1995 evaluation the test data was read material from newspaper articles recorded in a noisy environment using both a close talking microphone and a far-field microphone. In the latter case the speech signal was much more prone to corruption by the ambient noise. The best recognition error rates for the these tasks were achieved by the Cambridge University HTK recognition system at 6.63% for the close talking and 13.5% using the far-field microphones (WGP96). On this occasion, marginally worse results were achieved by the LIMSI group (GLA$^+$96). The 1996 evaluation task was concerned with the transcription U.S. television and radio broadcast news programmes. This time, the best results were achieved by researchers at LIMSI who achieved a word error rate of 27.1% (GAL$^+$97). This was slightly better than the Cambridge University HTK group which achieved a word error rate of 27.5% (WGP$^+$97). Many other groups regularly participate in this exercise including IBM (BCG$^+$97) and the Cambridge University connectionist speech recognition group (CKC$^+$97). The increased difficulty of the task is clearly reflected in the significantly worse results for the 1996 evaluation task. Breaking down the test data into the various acoustic channel types clearly shows that telephone interviews are much harder to recognise accurately than headlines read from a script by an announcer in the studio (GAL$^+$97) (WGP$^+$97). However advances in LVR systems in recent years would suggest that significant reductions in recognition errors for this harder task will be possible in the near future.

Although, by comparison limited, various work has also been carried out to develop LVR systems for languages other than American English. An evaluation programme called SQALE was carried out at three European speech research laboratories in 1995 (PWY95)(LADG95)(DAK95). This programme investigated recognition of American English, British English, French and German. The task here was recognition of read sentences and is some years behind the main ARPA evaluation for American English, however the results are interesting. Unsurprisingly, the best recognition performance was achieved for the most well-researched task, recognition of American English, followed by recognition of British English and then French. Each of these systems had a recognition vocabulary of 20,000 words. In order to achieve similar vocabulary coverage for German it was found necessary to build a recognition system with a vocabulary of

60,000 words. This problem arises because of the previously noted noun-compounding in German. SQALE provided a baseline condition for comparison of systems in each language from different sites using the same acoustic training data, vocabulary, and language model for each language. The average word recognition error rate was on the order of 15%. Clearly there is much work still to be done on speech recognition in these languages. The SQALE study was only a small effort for British English, French and German. One of the main constraints on further work in this area is the lack of acoustic and language model training data.

Other research in recognition of European languages is underway, for example in French (BDF+95) (LAG+96) and German (NHM+97) (AAL+96). Studies at Dragon Systems (BCG+96) and IBM (CDF+91) have reported comparative performance for various European languages including, as before, English, French and German, but also including Italian and Spanish.

Also less widely researched than American English, but of obvious importance for spoken document retrieval, is automatic speech recognition for Asian languages. Various groups have conducted research into large vocabulary Mandarin Chinese speech recognition (HL94) (CPC+97). Also under investigation is large vocabulary speech recognition in Japanese (MOM+97). In addition to their work on European languages, Dragon Systems are also researching Mandarin and Japanese recognition (BCG+96). The systems reported so far have vocabularies of less than 10,000 words, but word recognition error rates in dictation applications are encouragingly around 10%. These systems may appear much well less developed than the current ARPA evaluation systems; however, this type of system represented the state-of-the-art in earlier ARPA evaluations in 1992 and 1993 and hence rapid advances are likely if sufficient investment in spoken corpora and system development are made.

It has been observed that, in general, techniques and development strategies appear to port well from one language to another (LAG+96). However, unsurprisingly, to obtain optimal performance language-specific factors must be taken into account. Thus when porting to a new language certain system parameters or components have to be changed, i.e. those incorporating language-dependent knowledge sources such as the selection of the phone set, the recognition lexicon (including alternative pronunciations), and phonological rules.

The need for acoustic and linguistic training data to achieve continued advances in speech recognition research cannot be overstated. The systems used in the current ARPA evaluations make use of literally hundreds of hours of spoken acoustic training data and millions of words of training data for language models (as well as ever increasing computation power requirements). Such training data is very expensive to provide. For speaker-independent recognition acoustic training data must be collected from a group of speakers sufficiently acoustically diverse to represent the vocal characteristics of the anticipated speaker group. This data is usually collected by asking speakers to read aloud specially selected texts. The actual data spoken should be selected to give the widest possible coverage of the acoustic features of the language. For example, subword phone units should be featured in as many phonetic contexts as possible so that contextual variations can be incorporated into the models. Once collected, the spoken data must be manually verified to check that the speaker said what was required. This verification requires on the order of the time taken to speak the text. The spoken data must then be aligned with this accurate transcription before acoustic models can be trained. Significantly more costly is the collection of spontaneous speech data. This must first be manually transcribed by listening. We speak from our own experience when we say that manual transcription of spoken documents is a *very* time-consuming process. It takes many times ordinary speaking time (20 times is probably a realistic underestimate) to transcribe spontaneous speech accurately; in addition, it is also surprisingly tiring. After transcription, the acoustic data is aligned with the transcription before model training. Unfortunately for speech researchers it has been shown that use of spontaneous speech training data is of considerable benefit for the recognition of spontaneous speech (JEC+95).

Training data for linguistic models is similarly expensive to collect. Text data collected from different sources must be conditioned to standard formats and spellings. This is relatively straightforward for well-constructed material such as news archives, however it can be extremely time consuming for less well-controlled material such as transcriptions to interviews or conversations.

Work on CLSR will clearly be restricted only to those languages for which speech recognition resources are available.

Automatic Spoken Translation Systems One approach to cross-language text retrieval is to employ automatic machine translation. It would therefore seem appropriate to briefly review work in the area of automatic translation of spoken data.

Spoken language translation systems are built by combining expertise in speech recognition, natural language understanding and machine translation. In their most simple form these systems take the output from a standard LVR system and feed it to a rule-based natural language system which attempts to produce an output translation. Of course, this translation process is difficult even in text translation and the presence of speech recognition errors introduces further complications for spoken data. Even if these recognition errors could be eliminated, the informal linguistic structure of much spoken material makes the parsing process more difficult than for more formally structured text input. To this end researchers have had to develop novel robust parsing strategies to attempt to handle this informal structure and all forms of word recognition error: substitution, insertion and deletion. Robust parsing tools incorporate the ability to skip words in the output of LVR systems to attempt to produce a complete output parse and sometimes to produce such phrase level parsed output as can be found within the LVR output.

One method used to handle recognition errors is to offer multiple recognition output hypotheses to the translation system. These are either be in the form of *N-best* lists containing the N most likely output hypotheses as decoded by the speech recognition system, or word lattices representing the words the speech recognition system considers to be possible at anytime together with many possible preceding and following words. The underlying assumption here is that a completely correct, or at least more accurate, hypothesis will often be present among these N-best hypotheses, although it was not scored as the most likely output by the speech recognition system. The translation engine has to process these multiple hypotheses either individually or, particularly for the word lattice, in a single pass. Parsing a lattice introduces the need to simultaneously handle multiple word hypotheses as well as ambiguity. This involves non-trivial extensions to text parsing tools. The parsing system often incorporates special features to assist with the detection of recognition errors. The overall translated output is selected by combining scores from all evidence sources from the recognition and translation process.

There are currently a number of speech translation research systems in development. JANUS-II at Carnegie Mellon University (WFG+96) is a research system for the investigation of speech-to-speech translation systems. Currently components for English, German, Korean, Japanese and Spanish speech input and translation are being developed. Speech recognition performance varies depending on the difficulty of the task, but for a scheduling task recognition performance of around 70% is obtained for German, English and Spanish; and for the standard ATIS Switchboard task around 50%. The goal of the translation system is to preserve the content of an utterance. Thus end-to-end quality is assessed in terms of how well the meaning is preserved during the translation process. In a Spanish-English translation task, the JANUS-II system recorded over 80% acceptable translation for text input and over 70% with recognition performance of 70% to 80% (LWL+96).

The *Verbmobil* project is another major research project in the area of spoken language translation (BS96). Verbmobil is a large project with many partners. The system architecture is modular with separate groups often undertaking research for individual models. The objective of Verbmobil is to develop a portable simultaneous spoken language translation system for face-to-face dialogues. The task scenario for Verbmobil assumes a Japanese and a German manager discussing the date of their next meeting. As usual in similar situations both partners are required to speak English as a common second language. As is often the case in such situations the participants' understanding competence is higher than their speaking performance. At any point where their active knowledge of English proves insufficient the managers are allowed to speak in their native tongue. The Verbmobil system supports them by translating from their mother tongue, i.e. Japanese or German, into English. The Verbmobil functions using a mixture of keyword spotting for individual words to compute a dialogue history while the partners speak English, and deep linguistic analysis to perform translation when they are speaking their native language. The domain of operation is currently quite constrained. For example, the German recogniser has a vocabulary of only some 2461 words, which given the potential for noun compounding in German gives very limited coverage. Although there is consideration of the independent evaluation of individual modules and the whole system, results for such evaluations do not seem to be readily available.

Another example system is the Spoken Language Translator Project at SRI, Cambridge UK. The prototype system translates air travel (ATIS) queries from spoken English to spoken Swedish, French and Spanish (RC94) (RCB96). For a vocabulary of around 1,400 words, they report sentence recognition accuracy of over 70%. For sentences within the ATIS domain of up to 12 words in length, if the correct speech hypothesis is selected then translation is produced on about 75% of occasions and 90% of these translations are acceptable. The remaining 10% can be identified as errors because they are ungrammatical or unnatural. Divergence in meaning which might lead to serious forms of dialogue failure are extremely rare.

Rayner and Carter (RC95) suggest that practical systems for spoken language translation for limited domains of 1500 to 3000 words may be available within the next 5 to 10 years. However, although progress in this area, as in many areas of speech technology, has been rapid in recent years, these systems would not seem to have much utility for task-independent CLSR in the foreseeable future.

Specific Technology for CLSR The obvious strategies for implementing CLSR would seem to be to combine existing systems implementing the requisite component technologies. However further research may well inspire new technologies specifically applicable to CLSR in a similar manner to the index combination methods developed for monolingual speech retrieval (Jam95) (JFS+96).

Paradigms for Cross-Language Speech Retrieval

We can now discuss possible approaches to CLSR with reference to the preceding discussion of the current state-of-the-art in multilingual speech recognition and retrieval of speech recordings. We review two recent papers in the area of CLTR and examine the potential application of speech processing and speech retrieval techniques to each particular paradigm of CLTR. One paper takes a dictionary-based approach to CLTR and the other a parallel corpora method. We shall examine and explain the various issues and problems arising from each putative implementation of cross-lingual speech retrieval.

Appropriate Speech Technologies Before developing these possible CLSR paradigms it is worthwhile considering how these might best be explored using the currently available resources for speech technology. The ideal arrangement would clearly seem to be to build a LVR system in each language to be encountered and use the appropriate one to transcribe the spoken data. Further a phone-lattice approach could also be adopted for each language to supplement the LVR system. However, as an LVR system can only be built if sufficient acoustic and language model training data are available, the LVR approach may not be practical for many languages within current resources. As has been observed already, to create such resources would be very expensive, and a more economical method would be to adopt a phone-lattice only approach. The minimum requirement for this would be a sufficiently large corpus of training data for acoustic models and a phonetic dictionary for the words of the language. Ideally this arrangement could be augmented with a method of generating pronunciations for words outside the phonetic dictionary. Acoustic models for languages for which their is little available labelled training data could possibly be bootstrapped from existing acoustic models from other languages.

In the following putative CLSR approaches we assume the use of a phone-lattice indexing method, although where resources permitted this could be replaced or augmented with an LVR system.

Dictionary-Based Approach (Hull and Grefenstette, SIGIR96)

Hull and Grefenstette (HG96) recently proposed CLTR based on the use of multiple one-way dictionaries to translate a request in one language into the document language. Testing is performed using a conventional English document and request collection (in this case, the TIPSTER document collection, with TREC-3 queries), with queries translated by a human translator from English into a new query language, then automatically translated back into English. Retrieval then proceeds normally. Query translation is not performed using machine translation techniques owing to the lack of context in the queries and the fact that the query may not necessarily consist of well-formed sentences. In this approach, translation from some query language x to some document language y requires its own dictionary which cannot easily be inverted to provide translation in the other direction. While an interlingua might prove a more economical method of providing context-free translation between arbitrary languages, these typically prove to be too domain-dependent for use in modern retrieval tasks.

Spoken Documents This method could be adapted for CLSR with spoken documents as follows. If the document language is not known, it must be first be detected using language identification techniques. When a request is received an appropriate dictionary (assuming it exists) must be used to translate the query into the document language(s). Query terms can then be detected in the recogniser output either by scanning a phone-lattice (or directly for LVR output), and documents retrieved as with the monolingual speech retrieval techniques previously described. Thus, as in the case of CLTR, the actual retrieval process itself is relatively insulated from the language mismatch between the original query and the document collection.

One of the major problems in Hull and Grefenstette's attack on the problem is the production of large ambiguous queries by the translation process. For example, a sample 4-term French query *(troubles civils à caractère politique)* is translated into a 15-term English query. Hull and Grefenstette suggest, after the first pass of the retrieval system, the use of relevance feedback to refine the queries. However, this approach would not be easy to implement in a lattice-only CLSR system, since the identities of individual words are not known until explicitly searched for. However, some retrieval benefit might be realised from the re-weighting of the existing search terms using relevance feedback. It may be possible to derive indexing units automatically from phonetic transcriptions of the message collection, as Wechsler has done, but this method is still unproven in relevance feedback. James did achieve some success with relevance feedback in speech retrieval, but using transcriptions generated from a domain-dependent word recogniser. Even then, it would not necessarily make sense for the human user to review the suggested new terms and decide which should be incorporated into the revised query, since incorrectly-recognised terms may prove to be as discriminating as correct ones so long as the recogniser has made the same errors consistently. In this case, it would be far more appropriate to play back the section of audio corresponding to the fed-back term.

As mentioned earlier, in Kupiec's experiments, the semantic content of the textual document collection usefully disambiguated noisily-recognised spoken queries. James also observed this filtering, but in his case with textual queries and a noisily-recognised message collection. However, query mis-translation is unlikely to benefit from a similar filtering effect since the multiple variant translations of each query term are obviously semantically related. Returning to Hull and Grefenstette's example, several unwanted but perfectly valid-looking phrasal queries, such as "characters (of a) courteous nature", "political turmoil" or "diplomatic unrest" arise from the automatic translation of the original query. This does not compare favourably with Kupiec's "president kennedy" example, in which alternative queries generated by the system are almost all garbage such as "prescient remedy". This filtering can only work where term substitution is based on acoustic-phonetic rather than semantic criteria. It can be seen that CLSR would compound two error-prone processes; query translation and speech recognition. This is likely to manifest itself as a further reduction in retrieval performance, relative to some ideal monolingual test standard, beyond that observed for speech recognition or query translation in isolation.

Spoken Queries Another possible paradigm would be to have spoken queries in one language retrieving textual documents in another. A simple implementation would be straightforward to apply with Hull and Grefenstette's system. The French queries could be spoken into an appropriate speech recogniser and the output applied to the CLTR system. The output here would be error-prone, but semantic co-occurrence filtering and redundancy effects would come into play. This approach raises an interesting observation on Hull and Grefenstette's work. Reduced length TREC queries were used partially to compensate for a non-native speaker's inability to form expressive queries. However the obvious operating environment here would seem to be for spoken querying in the speaker's native language, in this case French. The speaker may well be able to read English to a level sufficient to make English texts useful, but their own command of English might mean that attempts to form English queries will lead to poorly formed queries lacking the richness and expressiveness of a native language. A system of this type, using Japanese text input to retrieve English text documents, is described in (KKSS96). Spoken queries in one's own language would seem likely to be more expressive and longer than attempts to form queries in an unfamiliar textual or spoken language.

Parallel Corpora Approach (Sheridan and Ballerini, SIGIR96)

Sheridan and Ballerini (SB96) have recently taken an approach to CLTR which exploits the availability of parallel, but not identical, corpora in differing languages. Pairs of files, which appear to address the same topic but which are written in 2 different languages, are "aligned" to form single cross-lingual documents; then, using a "dualist" model of information retrieval to cluster similar terms, a translation effect can be achieved through the expansion of a monolingual query, replacing each term with the corresponding cluster and then filtering to remove terms from the original query language. Since no dictionary is required, the approach is multi-directional between languages, although it depends heavily on the availability of a parallel corpus. If the corpora from which the cross-lingual thesaurus is compiled are also those from which documents are to be retrieved, then there is clearly nothing to be gained if document pairs are exact translations of each other, since there would be no need to perform CLTR at all. Once again, the system is dependent on relevance feedback, since the automatically translated queries will suffer from relatively low precision due to the translation ambiguity.

Since the core retrieval process need not know the exact provenance of the query, whether supplied by a native speaker of the language in question, or automatically translated using one or other appropriate method, a CLSR system could theoretically be built according to this paradigm. However, once again, the dependence on relevance feedback represents a major problem. In CLTR, it would not be unreasonable to assume that the user, who is not a speaker of the document language, will be able to make enough of a judgment on the relevance or otherwise of the top-ranked documents in order for relevance feedback to take place; however, this stage may be considerably more difficult when no text is available and such a judgment has to be made only by listening to audio or watching a video clip. A simple alternative would to be to assume the relevance of the top-ranked documents and perform feedback without user intervention; however, this assumes that units with some meaningful content, like words or word fragments, have been obtained, either from a word recogniser, or by derivation from some phonetic representation.

For a useful system to be built on this paradigm, the multilingual thesaurus would undoubtedly have to be generated without access to the retrieval collection, since the general availability of parallel corpora is unlikely, to say the least. The news domain would be a suitable one in which to build an experimental system, since it is obviously possible to build a domain-dependent parallel (or at least comparable) corpus, and subsequently a thesaurus, from textual news agency reports. This thesaurus could then be used to expand queries for a CLSR system. The thesaurus could be used either directly for retrieval of documents from the parallel corpora or (as would be more practical for CLSR) to translate queries for retrieval of documents from an unrelated spoken archive in another language. Results of an experimental study into CLSR using a comparable corpora approach with the ETH subword approach to spoken document indexing are due to appear shortly in (SWS97). In this work French textual queries are translated into German using a French-German thesaurus which is independent of the German spoken document collection. However, both the documents and the thesaurus are obtained from the domain of Swiss news reports (albeit from different sources and dates) and it is not clear how far this approach could be extended to material completely outside this domain.

Evaluation Issues

There are a number of issues which must be considered in the evaluation of CLSR systems. A common approach to CLTR evaluation (and that used by Hull and Grefenstette) has been to translate queries manually from existing collections into a new language and then use them to retrieve documents in the original language. There would appear to be only a minor, but nevertheless significant problem in extending this to spoken queries. The obvious way to enter these is to record a native speaker uttering the manual translation. However read speech data is considerably easier to recognise that spontaneous spoken material. Thus any conclusions drawn from experiments of this sort would refer only to queries read from text, therefore rather defeating the object of the exercise, since the text query might as well be used in the first place! Additionally, retrieval performance for such queries could only be taken as an upper bound on the effect of using spoken query input. It would thus be necessary to obtain new spontaneous queries in the new language, though this would require a new set of relevance assessments. This would be expensive but not impractical, and, so long as the experiment were carefully designed, such a collection could be used for long-term spoken query research.

As a simple test of the effectiveness of current LVR systems, we recorded a native French speaker reading aloud the queries used by Hull and Grefenstette. This data was then recognised using the current LIMSI 65,000 word vocabulary French LVR system. The overall word error rate of the recognised output was found to be 16% relative to the original text queries. Inspection of the recognised output showed that most of the content words had been recognised correctly. Only one word in the text queries, ***télévangélistes***, was outside the vocabulary of the LVR system [2]. Retrieval performance is likely to be somewhat degraded by these recognition errors, although overall we believe that existing lab-based speech technology, soon to be available in operational environments, is already good enough to make spoken querying a realistic proposition. However, it should be remembered with respect to this results that the queries were *read* from existing well-structured text. Spontaneous spoken queries are likely to be much more difficult to recognise reliably and hence recognition performance may be significantly worse. However, the ease with which spoken queries are generated relative to text may elicit longer queries in which redundancy effects may offset retrieval difficulties resulting from recognition errors.

The construction of collections to evaluate cross-language retrieval of spoken documents where the documents are in multiple languages would appear to be too expensive unless a genuine application could be established. For the construction of such a collection, spoken documents would have to be collected for each of the languages of interest. In addition, a query set in one or more languages would have to be collected and relevance assessment carried out. Queries could be

[2] It occurs to us that the out-of-vocabulary issue might have been more significant if the queries had been generated by a native speaker, who might be expected to use a broader vocabulary

manually translated into each document language and relevance assessments carried out for each document language. The cost of this would obviously be directly proportional to the number of query languages.

A cheaper alternative to exhaustive relevance assessment is to adopt so-called *known item searching.* A known-item search is a retrieval task that simulates a user seeking a particular, half-remembered document in a collection. In contrast to the exhaustive approach assumed so far, the goal in known-item searching is to retrieve one particular document deemed relevant. For a news archive this might be the full story corresponding to a particular headline. Construction of collections for CLSR evaluation with known-item searching would clearly be much cheaper. However, such tests would give no information about cross-language recall and little detail about precision.

To evaluate the effect of spoken recognition errors on performance a full correct manual transcription of the spoken material would be necessary. As was explained for the collection of acoustic training data, this is an extremely expensive process, particularly if the documents are of an unstructured spontaneous nature. This could then be used for control experiments to evaluate the effect of speech recognition errors on retrieval performance, as has been the practice for evaluation of monolingual spoken document retrieval.

Conclusions

Cross-lingual textual information retrieval, and monolingual speech retrieval, are both interesting and useful technologies, and their putative combination presents a number of challenges. In this paper we have aimed to give a flavour of how they might be combined. The proposed spoken document retrieval spoke of the TREC evaluations, combined with Hull and Grefenstette's CLTR approach, could give an opportunity for the first tests of cross-lingual speech retrieval. Alternatively, CLSR may be investigated in the CLTR TREC spoke if spoken documents were to be introduced.

There still remains the suspicion that CLSR is an interesting technical challenge without a very obvious real-world application. Certainly we would do well to consider which of the cross-lingual speech retrieval paradigms outlined at the start of this paper are likely to appear outside the laboratory, before starting extensive research efforts.

Acknowledgements

We are grateful to following for their assistance in this work; David Hull and Gregory Grefenstette of the Rank Xerox Research Centre, Grenoble for providing us with their French text queries; LIMSI, Paris and in particular Lori Lamel, for use of their French LVR system; Cambridge University Engineering Department for use of their speech recording facilities; and our willing French speaker Barbara Levienaise-Obiada.

References

M. Adda-Decker, G. Adda, L. Lamel, and J. L. Gauvain. Developments in Large Vocabulary, Continuous Speech Recognition of German. In *Proc. Int. Conf. Acoust., Speech., Sig. Processing*, volume I, pages 153–156. Atlanta, May 1996. IEEE.

R. Bakis, S. Chen, P. Gopalakrishan, et al. Transcription of Broadcast News – System Robustness Issues and Adaptation Techniques. In *Proc. Int. Conf. Acoust., Speech, Sig. Processing*, pages 711–714. Munich, April 1997. IEEE.

J. Barnett, A. Corrada, G. Gao, et al. Multilingual Speech Recognition at Dragon Systems. In *Proc. International Conference on Spoken Language Processing*, Philadelphia, October 1996.

M. G. Brown, J. T. Foote, G. J. F. Jones, K. Spärck Jones, and S. J. Young. Open-Vocabulary Speech Indexing for Voice and Video Mail Retrieval. In *Proc. ACM Multimedia 96*, Boston, November 1996. ACM.

J. Brousseau, C. Drouin, G. Foster, et al. French Speech Recognition in an Automatic Dictation System for Translators: the TransTalk Project. In *Proc. Eurospeech*, pages 193–196. Madrid, September 1995. ESCA.

T. Bub and J. Schwinn. VERBMOBIL: The Evolution of a Complex Large Speech-to-Speech Translation System. In *Proc. International Conference on Spoken Language Processing*, Philadelphia, October 1996.

H. Cerf-Danon, S. DeGennaro, M. Ferretti, et al. 1.0 Tangora - A Large Vocabulary Speech Recognition System for Five Languages. In *Proc. Eurospeech*, pages 183–192. Genoa, September 1991. ESCA.

T. Chiang, C.-M. Pengwu, S.-C. Chien, and C.-H.Chang. CCLMDS'96: Towards a Speaker-Independent Large-Vocabulary Mandarin Dictation System. In *Proc. Int. Conf. Acoust., Speech, Sig. Processing*, pages 1799–1802. Munich, April 1997. IEEE.

G. D. Cook, D. J. Kershaw, J. D. M. Christie, at al. Transcription of Broadcast Television and Radio News: The 1996 ABBOT System. In *Proc. Int. Conf. Acoust., Speech, Sig. Processing*, pages 723–726. Munich, April 1997. IEEE.

W. B. Croft. Personal Communication, 1995.

C. Dugast, X. Aubert, and R. Kneser. The Philips Large-Vocabulary Recognition System for American English, French and German. In *Proc. Eurospeech*, pages 197–200. Madrid, September 1995. ESCA.

J. L. Gauvain, L. Lamel, G. Adda, and D. Matrouf. Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task. In *Proc. Int. Conf. Acoust., Speech., Sig. Processing*, volume I, pages 73–76. Atlanta, May 1996. IEEE.

J. L. Gauvain, G. Adda, L. Lamel, and M. Adda-Decker. Transcribing Broadcast News Shows. In *Proc. Int. Conf. Acoust., Speech, Sig. Processing*, pages 715–718. Munich, April 1997. IEEE.

U. Glavitsch and P. Schäuble. A System for Retrieving Speech Documents. In *Proc. SIGIR*, pages 168–176. 1992. ACM.

D. A. Hull and G. Grefenstette. Querying Across Languages: A Dictionary-Based Approach To Multi-Lingual Information Retrieval. In *Proc. 19th Annual ACM SIGIR Conference*, pages 49–57. Zurich, August 1996. ACM.

A. G. Hauptmann, M. J. Witbrock, and M. G. Christel. News-on-Demand - An Application of Informedia Technology. *D-LIB Magazine*, September 1995.

A. G. Hauptmann and H. D. Wactlar. Indexing and Search of Multimodel Information. In *Proc. Int. Conf. Acoust., Speech, Sig. Processing*, pages 195–198. Munich, April 1997. IEEE.

A. G. Hauptmann and M. J. Witbrock. Informedia News on Demand: Multimedia Information Aquisition and Retrieval. In M. T. Maybury, Ed. *Intelligent Multimedia Information Retrieval*, AAAI Press/MIT Press, Menlo Park, CA, 1997.

H. W. Hon and K. F. Lee. Towards Large Vocabulary Mandarin Chinese Speech Recogntiion. In *Proc. Int. Conf. Acoust., Speech., Sig. Processing*, volume I, pages 545–548. Adelaide, April 1994. IEEE.

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

P. Jeanrenaud, E. Edie, U. Chaudhari et al. Reducing Word Error Rate on Conversational Speech from the Switchbaord Corpus. In *Proc. Int. Conf. Acoust., Speech., Sig. Processing*, volume I, pages 53–56. Detroit, April 1995. IEEE.

G. J. F. Jones, J. T. Foote, K. Spärck Jones, and S. J. Young. Retrieving Spoken Documents by Combining Multiple Index Sources. In *Proc. 19th Annual ACM SIGIR Conference*, pages 30–38. Zurich, August 1996. ACM.

J. Kupiec, D. Kimber, and V. Balasubramanian. Speech-based Retrieval using Semantic Co-Occurrence Filtering. In *Proceedings of the Human Language Technology (HLT) Conference*, pages 373–377. 1994. ARPA

Y. Kiyoshi, M. Kazunori, D. Shinichi, and K. Shinichiro. A Language Conversion Front-End for Cross-Linguistic Iformation Retrieval. In *Workshop on Cross-Linguistic Information Retrieval, 19th Annual ACM SIGIR Conference*, pages 34–39. Zurich, August 1996. ACM.

L. F. Lamel, M. Adda-Decker, and J. L. Gauvain. Issues in Large Vocabulary Multilingual Speech Recognition. In *Proc. Eurospeech*, pages 185–188. Madrid, September 1995. ESCA.

L. F. Lamel, M. Adda-Decker, J. L. Gauvain, and G. Adda. Spoken Language Processing in a Multilingual Context. In *Proc. International Conference on Spoken Language Processing*, Philadelphia, October 1996.

A. Lavie, A. Waibel, L. Levin, et al. Translation of Conversational Speech with JANUS-II. In *Proc. International Conference on Spoken Language Processing*, Philadelphia, October 1996.

T. Matsuoka, K. Ohtsuki, T. Mori, et al. Japanese Large-Vocabulary Continuous-Speech Recognition using a Business-Newspaper Corpus. In *Proc. Int. Conf. Acoust., Speech, Sig. Processing*, pages 1803–1806. Munich, April 1997. IEEE.

Y. K. Muthusamy, E. Barnard, and R. A. Cole. Automatic Language Identification: A Review/Tutorial. *IEEE Signal Processing Magazine*, October 1994. http://www.cse.ogi.edu/CSLU/publications/papers.html.

J. McDonough, K. Ng, P. Jeanrenaud, H. Gish, and J. R. Rohlicek. Approaches to Topic Identification on the Switchboard Corpus. In *Proc. Int. Conf. Acoust., Speech., Sig. Processing*, volume I, pages 385–388. Adelaide, April 1994. IEEE.

M. Niemöller, A. Hauenstein, E. Marschall, et al. A PC-Based Real-Time Large Vocabulary Continuous Speech Recognizer for German. In *Proc. Int. Conf. Acoust., Speech, Sig. Processing*, pages 1807–1810. Munich, April 1997. IEEE.

D. W. Oard and B. J. Dorr. A survey of Multilingual Text Retrieval. Technical Report UMIACS-TR-9619, University of Maryland, 1996. http://www.ee.umd.edu/ medlab/mlir/mlir.html.

B. Peskin, S. Connolly, L. Gillick, et al. Improvements in Switchboard Recognition and Topic Identification.

In *Proc. Int. Conf. Acoust., Speech., Sig. Processing*, volume II, pages 303–306. Atlanta, May 1996. IEEE.

D. Pye, P. C. Woodland, and S. J. Young. Large Vocabulary Multilingual Recognition Using HTK. In *Proc. Eurospeech*, pages 181–184. Madrid, September 1995. ESCA.

M. Rayner and D. Carter. The Speech-Language Interface in the Spoken Language Translator. In *Proceedings of TWLT-8, Twente Workshop on Language Technology, University of Twente, Holland*, 1994. http://www.cam.sri.com/tr/crc051/.

M. Rayner and D. Carter. The Spoken Language Translator Project. In *Proc. Language Engineering Convention*, London, October 1995.

M. Rayner, D. Carter and P. Bouillon. Adapting the Core Language Engine to French and Spanish In *Proc. NLP-IA*, Moncton, New Brunswick, 1996.

R. C. Rose, E. I. Chang, and R. P. Lippmann. Techniques for Information Retrieval from Voice Messages. In *Proc. Int. Conf. Acoust., Speech, Sig. Processing*, pages 317–320. Toronto, 1991. IEEE.

P. Sheridan and J. P. Ballerini. Experiments in Multilingual Information Retrieval Using the SPIDER System. In *Proc. 19th Annual ACM SIGIR Conference*, pages 58–65. Zurich, August 1996. ACM.

P. Sheridan, M. Wechsler, and P. Schäuble. Cross-Language Speech Retrieval: Establishing a Baseline Performance. In *Proc. 20th Annual ACM SIGIR Conference*, (to appear), Philadelphia, July 1997. ACM.

T. Schultz, I. Rogina, and A. Waibel. LVCSR-Based Language Identification. In *Proc. Int. Conf. Acoust., Speech., Sig. Processing*, volume II, pages 781–784. Atlanta, May 1996. IEEE.

R. C. F. Tucker, M. J. Carey, and E. S. Parris. Automatic Language Identification using Sub-Word Models. In *Proc. Int. Conf. Acoust., Speech., Sig. Processing*, volume I, pages 301–304. Adelaide, April 1994. IEEE.

A. Waibel, M. Finke, D. Gates, et al. Janus II – Translation of Spontaneous Conversational Speech. In *Proc. Int. Conf. Acoust., Speech., Sig. Processing*, volume I, pages 409–412. Atlanta, May 1996. IEEE.

M. Wechsler and P. Schäuble. Indexing Methods for a Speech Retrieval System. In *Proc. MIRO Workshop*, Glasgow, September 1995. http://www-ir.inf.ethz.ch/ISIR-Papers.html.

P. C. Woodland, M. J. F. Gales, and D. Pye. Improving Environmental Robustness in Large Vocabulary Speech Recognition. In *Proc. Int. Conf. Acoust., Speech., Sig. Processing*, volume I, pages 65-68. Atlanta, May 1996. IEEE.

P. C. Woodland, M. J. F. Gales, D. Pye, and S. J . Young. Broadcast News Transcription using HTK. In *Proc. Int. Conf. Acoust., Speech, Sig. Processing*, pages 719–722. Munich, April 1997. IEEE.

M. A. Zissman, T. P. Gleason, D. M. Rekart, and B. L. Losiewicz. Automatic Dialect Identification of Extemporaneous Conversational Latin American Spanish Speech. In *Proc. Int. Conf. Acoust., Speech., Sig. Processing*, volume II, pages 777–780. Atlanta, May 1996. IEEE.

M. A. Zissman. Automatic Language Identification Using Gaussian Mixture and Hidden Markov Models. In *Proc. Int. Conf. Acoust., Speech., Sig. Processing*, Minneapolis, 1993. IEEE.