

Spoken Document Retrieval, Automatic

A Hauptmann, Carnegie Mellon University,
Pittsburgh, PA, USA

© 2006 Elsevier Ltd. All rights reserved.

What Is Automatic Spoken Document Retrieval?

Spoken document retrieval (SDR) aims to provide content-based retrieval of passages from archives of recordings of speech. Thus, SDR refers to the retrieval of spoken material in digital audio files, ordered by relevance to some textual query. In the most common realization, the query consists of a typed sequence of words, a question, or a statement of information need. The spoken documents are previously indexed audio recordings of human speech, which were automatically transcribed by a speech recognition system. The goal of SDR, therefore, is a form of information retrieval—that is, automatically finding those complete documents or excerpts that either contain the words and expressions in the query or are semantically relevant to the information need. Thus, the speech transcription has to be automatic and the retrieval has to be based on a user-supplied query. Retrieval of music, songs, or nonlanguage sounds does not fall into the category of spoken document retrieval. Concise question-answering, where a user would expect to hear a short answer to a question as opposed to being presented with a complete audio document, is also not considered part of the definition of automatic SDR.

The purpose for developing mechanisms that provide access to spoken information is fairly obvious. The availability of inexpensive computers, storage devices, and high-bandwidth transmission capacity has resulted in many large multimedia collections. People have become accustomed to full access to virtually all textual information available on the Internet. Without SDR, access to audio archives, or at least spoken audio collections, would be restricted to those limited documents that have been manually transcribed or indexed with key words. Although a significant fraction of current television and radio broadcasts have manually created transcripts or at least approximate script outlines, a much larger amount of spoken audio recordings are untranscribed because the cost of human transcription is too high relative to the potential useful value or because they consist of older radio and television ‘legacy’ productions, for which transcripts were lost or never created.

SDR creates a full search and retrieval capability in a way that is already widely available for text content. This capability is useful in applications such as

- Searching the transcripts of video conferences
- Accessing portions of taped educational lectures
- Finding specific content in training audio and video
- Organizing archived voice mail by spoken content
- Accessing news of interest from television or radio
- Archiving meetings as a form of corporate memory and documentation.
- Retrieving audio transcripts from sports and entertainment broadcasts, including practically any talk shows, films, quiz programs, etc. with a significant spoken component.

Problems arise in SDR from insufficient fidelity of the speech recognition transcript, the accuracy of the retrieval matching between the query and the transcript, and the lack of segmentation (i.e., the absence of clear document, story, or passage boundaries in continuously recorded audio streams).

The Basic Approach to SDR

As the first step, an SDR system generates a text transcript from the audio recordings to enable text-based retrieval over these audio documents. The idea is to first apply automatic speech recognition to the audio documents to obtain a text transcript. The transcript recognition units are usually words; however, syllables and phonemes have also been used. Each word in the transcript is marked with the time at which it occurred. In general, SDR is designed to search vast archives of audio documents that have been previously indexed. In the case of a stable collection of audio recordings, speech recognition occurs once before the archive is indexed. In archives that are actively accumulating spoken documents, speech recognition is performed on a continuous basis as the audio stream is captured and saved.

The next phase in the SDR process is information retrieval index creation. To facilitate matching root words with different inflections and morphologies, a stemming process converts all speech-recognized words into their canonical root stems. After removing very frequently occurring words, the so-called ‘stop-words,’ the remaining word stems are compiled into an inverted index, which allows fast access based on target query words. Each word stem constitutes an entry in the index, listing all documents in which it was found, the number of times it occurred in the documents, and the time of each individual occurrence. Statistics are kept on the distribution of the

AU:1

p0020

s0010

p0025

p0015

p0030

AU:2

2 Spoken Document Retrieval, Automatic

words across the documents in the collection for retrieval similarity ranking.

When the indexed archive is accessed, a query indicating the user's information need is expressed as a sequence of words or syntactically complete sentences. Although queries are typically typed, in some situations the query may also be spoken. In a query preprocessing step, the query orthography is normalized with all punctuation and capitalization removed. Thus, for example, after normalization, stemming, and stopword removal, the query *C.I.A.'s use of spy planes in 1985?* might appear as *CIA SPY PLANE NINETEEN EIGHTY FIVE*. The query is then transformed into a sequence (vector) of units of the same type as the indexed transcript recognition units (i.e., words, syllables, or phonemes).

To retrieve spoken documents relevant to a user query, each document in the archive is automatically compared to the query vector to determine how well the document 'matches' the query. The matching function $R(Q, D)$ computes the relevance (R) of a query (Q) to a document (D). A list of the query-relevant audio documents is then returned to the user, where the locations of the matching words are indicated. The list of audio documents or passages is ordered by decreasing similarity between the query and the content (transcription) of the document (Robertson *et al.*, 1996). A typical SDR process is shown in Figure 1.

Evaluation of SDR

Evaluation of SDR systems can be done along different dimensions. The complete system performance can be measured in terms of how well the SDR system retrieves spoken documents, but there are often useful aspects in measuring the performance of the components, notably the speech recognition, segmentation, and text information retrieval modules.

The complete SDR system performance is usually measured in terms of mean average precision (MAP), a measure also widely used for text information retrieval. 'Precision,' in general, is defined as the number of relevant documents retrieved divided by the number of items retrieved. Average precision is a single valued measure that reflects performance over all relevant documents for a given query, rewarding systems that retrieve relevant documents at high ranks. It is defined as the average of the precision value obtained after each relevant document retrieved. When a relevant document is not retrieved, its precision is set to zero.

Formally, for a given query, there are a total of N_r items in the collection that are relevant to this query. Assume that the system retrieves k relevant items and

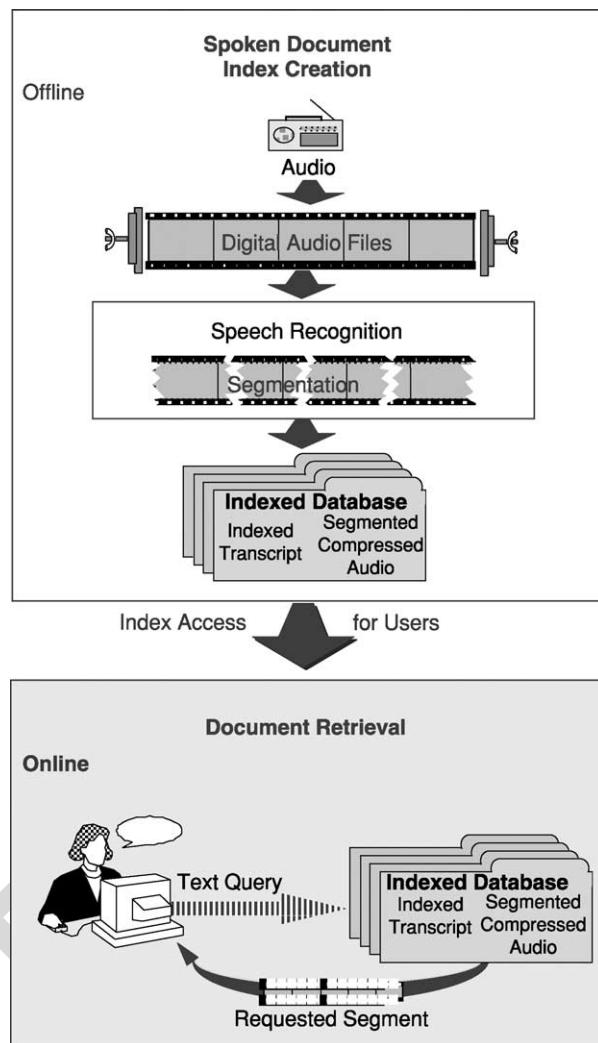


Figure 1 The basic architecture of a spoken document retrieval system.

they are ranked as r_1, r_2, \dots, r_k . Then, the average precision is computed as

$$AP = \left\{ \sum_{i=1}^k i/r_i \right\} / N_r \quad (1)$$

MAP is the average of the precision values over a set of queries.

Evaluations use a standardized evaluation format, a common corpus (frequently broadcast news), and test data to ensure different systems and their approaches can be compared fairly. The largest SDR evaluations to date were carried out by the National Institute of Standards and Technology, eventually comprising more than 20,000 broadcast news stories (which translated to approximately 500 h of audio), with 50 queries in a typical evaluation 'task' (Garofolo *et al.*, 2000).

Speech recognition is usually evaluated in terms of the word error rate (WER) in the transcript produced by the automatic speech recognizer. It is defined as the sum of the inserted words, deleted words, and substituted words, divided by the total number of words in a 'perfect' human-transcribed reference transcript.

Publications on SDR will occasionally refer to the term error rate (TER) in speech transcripts, which is similar to WER but ignores all information retrieval 'stop words' (usually a list of very common and function words, which do not contribute to the relevance ranking of a document) in the speech transcripts. A variation sometimes referred to as the STER (stemmed term error rate) also collapses all morphological and inflectional variants of a word into the same stem. For example, in the sentence, 'The large apple was very rotten,' if the recognizer produced an output of 'That large apples were very rotten,' the WER would be 50% (3/6), the TER would be 33.3% (1/3, since 'that,' 'were,' and 'very' are stop words and thus ignored), and STER would be 0 (0/3) because the roots for the words 'large,' 'apple,' and 'rotten' were all correctly preserved. Although it can be argued that TER and STER (or even QTER, which only considers actual query words) are more appropriate measures than WER for the assessment of how good the speech recognition is for the purposes of SDR, in practice all these measures are approximately equivalent (Garofolo *et al.*, 2000).

SDR systems performance evaluation is usually done not only in terms of absolute MAP but also in terms of what degradation is due to the errorful nature of the automatic speech transcription process. To this end, identical systems are evaluated using speech-recognized transcripts vs. perfectly transcribed documents, and the decrease in MAP is reported. Note that some evaluations speak of a 'relative' decrease in MAP, whereas others report absolute decrease in MAP, often leading to confusion in the reader.

Speech Recognition for SDR

The major challenge of automatic speech recognition for SDR is to transcribe real-world data that have not been recorded for SDR purposes. In such material, the audio may be of varying quality, from many speakers, recorded with unknown microphones under widely different environmental conditions, and cover a wide range of topical content. Each of these factors contributes to making the speech recognition problem very challenging.

A speech recognizer is applied to an audio stream and generates a time-marked transcription of the speech. Time alignment allows pinpointing exactly where each word is spoken. Thus, all query terms or

known relevant words can be marked on a timeline of the playback viewer. It is also possible to skip to the relevant portions of a longer document to facilitate efficient playback that skips irrelevant portions within a single audio document (Arons, 1994).

This transcript usually consists of words, but syllable or phoneme transcripts have also been used. The transcript may be phone- or word-based in either a lattice (probability network), n -best list (multiple individual transcriptions), or, more typically, a 1 -best transcript (the most probable transcription as determined by the recognizer).

Depending on the speech recognition technology and computer hardware used, the automatic speech recognition can take less than real time, such that a 1 -h document is transcribed in 60 min or less, or the speech recognition can involve multiple passes and take up to several hundred times real time, where many days are needed to process a single hour of speech with high accuracy. The redeeming factor is that this process only needs to occur once, at the time the archive of spoken documents is created.

The main speech recognition issues that affect the performance of SDR are related to (1) the vocabulary size of the automatic speech recognition system, (2) the word error rate of the automatic speech recognition system, and (3) removing nonspeech audio sections from the data.

Vocabulary Size in Automatic Speech Recognition for SDR

Automatic speech recognition systems tend to have fixed vocabularies of words that they can recognize. If the recognizer's vocabulary is small, many words will not be recognized, resulting in an inability to find these words later in the retrieval process. This is known as the out-of-vocabulary (OOV) problem. The OOV rate is the percentage of test set words that are not included in the recognizer's vocabulary and that therefore can never be correctly recognized. The problem is actually compounded by the fact that speech recognition systems tend not to know when they encounter an unknown word that is outside their vocabulary. Instead, they will fill in another word as a substitute, and to preserve continuity in the acoustic and language sequence, often an additional error is made afterwards. Studies have shown that each OOV results in almost one and a half word errors in the transcript (Rosenfeld, 1991). For vocabularies of less than 10,000 words, the OOV rate can be expected to be in the double digits, resulting in significant reductions in transcript accuracy and, consequently, lower retrieval effectiveness.

Although one would like the recognizer to distinguish between as many words as possible, each

additional word in the vocabulary also results in additional potential confusions during recognition. As a practical matter, current recognizers are limited to a maximum of 64,000 words, which can be indexed in 16 bits or 2 bytes. A larger vocabulary quickly becomes prohibitively unwieldy, when one considers that a typical trigram language model allows $64,000^3$ word sequences, even though not all combinations are likely to exist. A larger index, in addition to increasing the storage to 4 bytes per word, also implies a cubic increase in the space necessary for the language model. Fortunately, 64,000 words have been shown to be sufficient for many tasks due to Zipf's power law. In fact, there is only a minimal improvement in the OOV rate when recognizers increase their vocabulary from 51,000 to 64,000 words, where the OOV rate is approximately 1 or 2% (Woodland *et al.*, 2000). Clever selection of the right 64,000 word vocabulary, based on the expected thematic content of the audio, can further reduce the OOV problem (Garofolo *et al.*, 2000).

Even when the speech recognizer vocabulary is too small and actual query terms are not included in it, some of the degradation in the speech transcripts can be mitigated using phonetic search. In general, phonetic search is considered inferior to search on full words, but it can recover errors due to OOV words. The standard approach is to replace each word by an n-gram of phonemes (typically three or more phonemes). Thus, the name missing from the recognizer vocabulary such as 'Albright,' phonetically transcribed as 'ax l b r ay t,' would translate into a search

for the phoneme trigrams 'ax l b,' 'l b r,' 'b r ay,' and 'r ay t' as if they were words. Of course, an appropriate phone retrieval index would initially need to be constructed from the archived documents (Ng *et al.*, 2000).

Relationship of Word Error Rate to Retrieval Effectiveness

In general, WER is a good predictor of information. It also predicts how well named entities can be extracted (Kubala *et al.*, 1998)—that is, whether names, organizations, places, dates, and numbers can be correctly tagged as such in the speech transcript.

Although it is interesting to see how much information retrieval degrades with respect to a particular recognition word error rate, Figure 2 shows sample experimental data estimating spoken document retrieval effectiveness over a range of transcripts with different error rates. The figure shows the relationship between information retrieval precision and speech recognition accuracy plotted as relative degradation to retrieval from manually transcribed text documents. The performance of a 'perfect' system is defined by the relevance judgments for documents and queries of a human judge of document relevance. This is equivalent to the WER of 0%. The quality of the information retrieval decreases as the speech recognition word error rate increases. For word error rates less than approximately 25%, the graph shows that retrieval performance degrades very little for transcripts with increasing word error rates and that retrieval is fairly robust to recognition errors. At

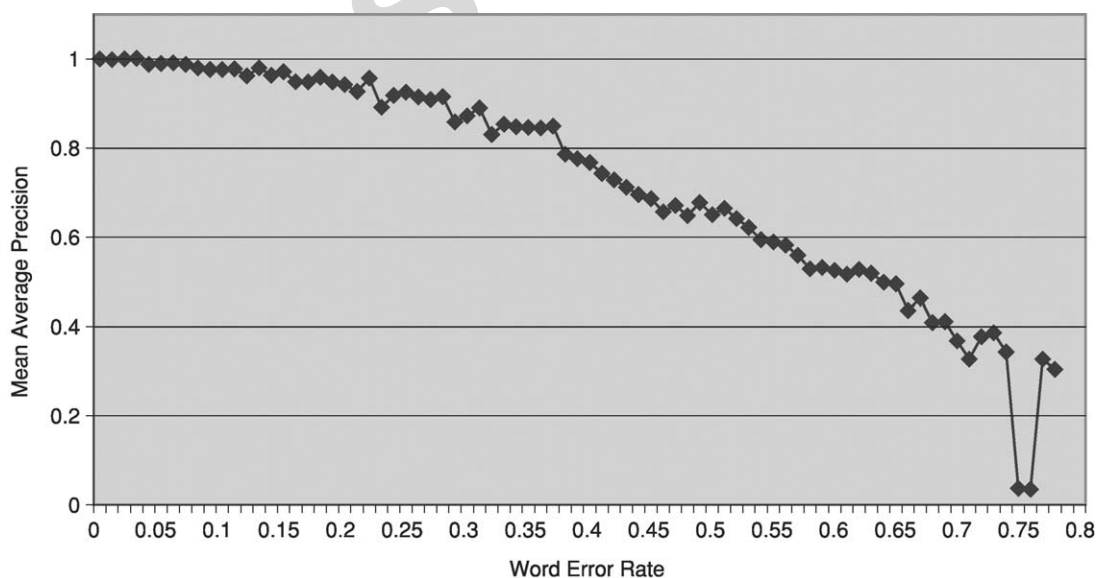


Figure 2 Degradation of retrieval effectiveness as a function of word error rate in a basic SDR system without query or document expansion. Data from Hauptmann and Wactlar (1997).

speech error rates higher than 35%, the information retrieval effectiveness starts to decline noticeably. Similar results were obtained by Woodland *et al.* (2000).

p0130 Fortunately, the most accurate recognizers to date on the type of broadcast news data shown in **Figure 2** produced a word error rate of approximately 20% for English broadcast news (Johnson *et al.*, 2001). One possible explanation is that the repetition and redundancy of key words in spoken language allow the relevant documents to be retrieved even when a substantial fraction of the words are incorrectly recognized. Once an essential set of correct key words are removed, however, the whole information retrieval paradigm quickly falls apart.

p0135 Other data show that modifying the speech recognizer to directly produce a transcript does not work because the speech recognition process depends strongly on the syntactic constraints provided by morphological variants and function words (Hauptmann *et al.*, 1998). There is evidence that allowing a speech recognizer to produce multiple alternate word candidates in the form of a confidence annotation lattice can reduce the effect of words missing from the single ‘best guess’ word candidate provided by the speech recognizer and thus improve information retrieval effectiveness (Siegler, 1999).

p0140 The consensus from a number of published experiments in this area is that as long as speech recognition has a word error rate less than 35%, information retrieval from the transcripts of spoken documents is only 2–10% worse than information retrieval on perfect text transcriptions of the same documents. Speech recognition within a reasonable error range is sufficient to perform information retrieval tasks with only a small decrease in information retrieval effectiveness (Garofolo *et al.*, 2000; Sparck-Jones *et al.*, 2001). Effective retrieval performance in SDR, especially from broadcast news, is thus a perfectly practical proposition.

s0035 **Removing Nonspeech Audio Sections from the Data**

p0145 A final contribution of speech recognition is to identify regions in the audio that do not contain speech. Nonlexical information derived directly from the audio, which would not normally be transcribed, can be used to improve SDR systems. Eliminating difficult to recognize sections of the audio track helps retrieval accuracy. Thus, when a system removes commercial advertisements, music, and poor quality audio regions, word transcription accuracy improves because there are fewer misrecognized words that spuriously match query words during retrieval. Speech recognition approaches can aggregate acoustically

similar passages; identify gender of the speaker; detect speech portions recorded by telephone; and identify silence, noise or music, repeated passages, and commercials. Audio repeats can accurately predict the presence of commercials, which can be filtered out before retrieval, and some broadcast structure information can be recovered by analyzing bandwidth, signal energy, and the presence of music. Segmentation of audio streams into coherent passages and browsing can also be improved by including automatically inserting markers for sentence boundaries and speaker turns. Removing music, commercials with speech, song and music, and other irrelevant material can further benefit the retrieval effectiveness for SDR (Johnson and Woodland, 2000). This automatically derived information also assists in the detection of ‘document’ boundaries in continuous audio streams.

Information Retrieval in SDR

This section provides an overview of the approaches used from the information retrieval side of SDR to improve retrieval efficiency and lessen the effects of incorrect speech transcriptions. The improvements on retrieval effectiveness concentrate on query expansion. If a collection of errorless text documents is available that is comparable to the audio archive, then document expansion and parallel corpus query expansion are also able to improve retrieval.

Query Expansion through Blind Relevance Feedback

p0155 It has long been known that longer queries result in higher precision retrieval. Blind (or pseudo-) relevance feedback, common for text retrieval, attempts to augment the query with terms from other relevant documents. For example, the system might take the top 10 documents returned after an initial retrieval cycle and pick the five most ‘valuable’ terms (according to a defined measure of term value) found in these documents. The assumption is that the top documents are likely to be highly relevant, and important terms in these documents not already in the query should help find other relevant documents. Since the retrieval is automatic, no user decides on the actual relevance of the selected documents or terms; hence, the selection is considered ‘blind’ or pseudo-relevance feedback. The top valuable terms are then added to the query before it is issued again against the archive. The effect is to make the query longer and richer in the words used to describe the query topic. The optimal number of terms and

the optimal number of initial documents to use are dependent on the collection and subject to empirical determination.

Query Expansion through Blind Relevance Feedback from a Parallel Corpus

A variation on this theme is possible if a closely matched parallel corpus of similar documents is available. Similar to blind relevance feedback, the query is first issued to the parallel collection to retrieve highly relevant documents, from which valuable additional query terms are selected before the query is reissued against the spoken document archive. Of course, one can also combine both blind relevance feedback and parallel blind relevance feedback for query expansion.

Document Expansion through Blind Relevance Feedback from a Parallel Corpus

Another improvement in information retrieval for audio documents comes through document expansion, provided a parallel corpus is available. Instead of adding terms to the query, terms from the parallel collection are added to the actual audio document transcripts. This is accomplished by treating each audio document (or a passage from the audio document) as a query into the parallel collection. Again ('blindly') the most similar parallel documents retrieved are assumed to be relevant, and the most valuable terms are extracted from these documents and added to the original audio transcript.

Combined, as shown in Figure 3, these document and query expansion techniques can improve mean

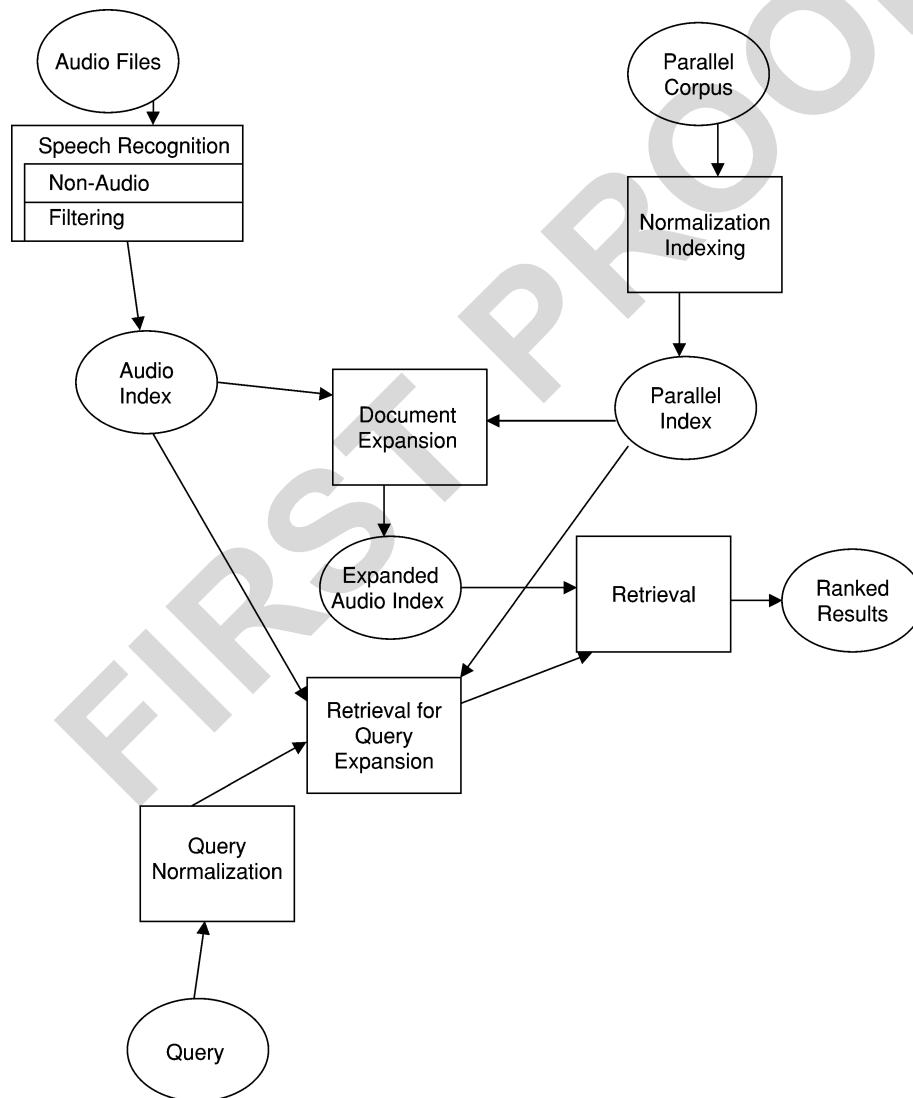


Figure 3 Architecture of a full SDR system with parallel corpus document expansion, blind relevance feedback, and parallel blind relevance feedback.

average precision by 17% or more. However, depending on the speech transcription error rate, the initial mean average precision, and the type of content, not all techniques work equally well all the time. For example, document expansion does not improve precision when the documents are already almost perfectly transcribed. In the most definitive studies, the degradation in mean average precision due to a WER of 25% was reduced to approximately 2% when all methods were combined.

Other variants of the cross-lingual SDR have attempted to use automatically transcribed documents in one language to find matching audio documents in a comparable, parallel collection in another language (Meng *et al.*, 2000). To date, combining SDR and machine translation into cross-lingual SDR still results in substantial degradation of retrieval effectiveness, which varies according to language pair and retrieval approaches used (Federico and Jones, 2003).

p0195

Video Retrieval

s0075

Video is a rich source of information, with aspects of content available both visually and acoustically. One prerequisite for achieving the goal of video retrieval is the automated information extraction and metadata creation from digitized video. The video retrieval research challenge is to automatically analyze both the visual and the audio component of video and make them retrievable by a user. Once the metadata have been extracted, queries can search the combined metadata of spoken language and visual documents. Since the video and image analysis does not have a well-defined vocabulary like speech recognition, a variety of ‘semantic concepts’ have been proposed to aid in automatically labeling the content of video shots. A successful approach for video retrieval requires the integration of speech retrieval, semantic concept analysis, and image retrieval (Kraaij *et al.*, 2004; Yan *et al.*, 2004).

p0200

SDR with Unsegmented Audio Data

s0060

Although one can think of many different ways to break a continuous audio stream into individual passages, stories, or documents, a simple approach of declaring a story boundary every 30 s has proved to be a remarkably good approximation, given that the WER and the lack of punctuation and formatting in speech recognition transcripts prevent many approaches based on linguistic features. To avoid splitting a passage at some particularly critical junction, the story boundary windows are overlapped by 15 s. After retrieval, multiple overlapping windows are merged together for the final result list.

p0175

Compared to manually segmented audio documents, the retrieval effectiveness suffers an 8–10% absolute reduction in MAP, despite the use of various query and document expansion techniques described previously (Abberley *et al.*, 1999; Johnson *et al.*, 2001).

p0180

Research Trends in SDR

s0065

Evaluations of SDR showed that retrieval of audio documents can be extremely effective despite significant error rates in the speech transcripts. This insight resulted in SDR being considered a ‘solved’ problem, and research interest faded. Although some researchers have examined spoken queries, or heterogeneous collections of spoken, OCR, and text documents, there are two main directions of active research in SDR: cross-lingual SDR and video retrieval.

p0185

Cross-Lingual SDR

s0070

Cross-lingual SDR attempts to find audio documents spoken (and automatically transcribed) in one language using queries formulated in a different language. The problem becomes one of adequately translating the query with consideration of the of spoken document retrieval. In particular, bad translations and misrecognition of proper nouns can become very problematic so that a multiscale approach seems promising, which combines word translation with syllable or phoneme translation and retrieval (Meng *et al.*, 2004).

p0190

Applications Using SDR

s0080

One of the fundamental lessons from SDR experiments is that achievable WERs in the speech recognition permit effective information retrieval. Using fairly standard information retrieval techniques, the current state of the art in speech recognition is adequate for information retrieval from audio archives. It is not too surprising that some interesting applications have transitioned from the research environments into full daily use.

p0205

There are a number of smaller companies that market meeting and lecture browsers that include speech recognition and SDR. The fundamental concept is to record audio and video of a presentation or meeting, analyze the audio, and allow retrieval, browsing, or summarization based on the extracted metadata. However, poor quality audio recording and high error rate transcriptions often make SDR difficult in these environments.

p0210

The HP SpeechBot is a successful attempt to index every audio file on the Web and make it available for search on a much larger scale (Thong *et al.*, 2002). It incorporates the basic SDR paradigm, extending it to cover a much richer variety of content and formats

p0215

than has been studied in research labs. A recent addition to this is the NewsTuner, which is focused on broadcast audio streams available on the Web (Logan *et al.*, 2004).

p0220 One of the more ambitious SDR projects is associated with the Survivors of the SHOAH Visual History Foundation, which has videotaped interviews with 52,000 speakers in 32 different languages. These audio tracks are filled with verbal disfluencies, heavy accents by nonnative speakers, age-related pronunciation difficulties, and uncued speaker and language switching. A limited amount of metadata have been manually entered describing the speakers, the locations mentioned, and some content key words. A major effort is under way to convert these video interviews into a fully searchable SDR archive (Byrne *et al.*, 2004).

See also:

Bibliography

Abberley D, Kirby D, Renals S & Robinson T (1999). 'The THISL broadcast news retrieval system.' Proceedings of the ESCA ETRW workshop on accessing information in spoken audio, Cambridge, UK April.

Arons B (1993). 'Speech Skimmer: interactively skimming recorded speech.' In *Proceedings of UIST'93*. New York: ACM Press.

Byrne W, Doermann D, Franz M, Gustman S, Hajic J, Oard D, Picheny M, Psutka J, Ramabhadran B, Soergel D, Ward T & Zhu W-J (2004). 'Automatic recognition of spontaneous speech for access to multilingual oral history archives.' *IEEE Transactions on Speech and Audio Processing* (Special issue on spontaneous speech processing).

Federico M & Jones G J F (2003). 'The CLEF 2003 cross-language spoken document retrieval track.' Proceedings of the CLEF 2003: Workshop on cross-language information retrieval and evaluation, Trondheim, Norway.

Garofolo J, Auzanne G P & Voorhees E M (2000). 'The TREC spoken document retrieval task: a success story.' Proceedings of the Recherche d'Informations Assistée par Ordinateur: content-based multimedia information access conference, www.nist.gov/speech/tests/sdr/sdr2000/.

Hauptmann A G & Wactlar H D (1997). 'Indexing and search of multimodal information.' International conference on acoustics, speech and signal processing (ICASSP-97), Munich, Germany, April.

Hauptmann A G, Jones R E, Seymore K, Siegler M A, Slattery S T & Witbrock M J (1998). 'Experiments in information retrieval from spoken documents.' BNTUW-98 proceedings of the DARPA workshop on broadcast news understanding systems, Lansdowne, VA, February.

Johnson S E & Woodland P C (2000). 'A method for direct audio search with applications to indexing and retrieval.' Proceedings of ICASSP 2000, Istanbul, Turkey, June.

Johnson S E, Jourlin P, Spärck Jones K & Woodland P C (2001). 'Spoken document retrieval for TREC-9 at Cambridge University.' In *Proceedings of TREC-9*. Gaithersburg, MD: National Institute of Standards and Technology.

Kraaij W, Smeaton A F, Over P & Arlandis J (2004). 'TRECVID—an introduction.' In *Proceedings of TRECVID 2004*. Gaithersburg, MD: National Institute of Standards and Technology.

Kubala F, Schwartz R, Stone R & Weischedel R (1998). 'Named entity extraction from speech.' Proceedings of DARPA broadcast news workshop, Lansdowne, VA, February.

Logan B, Moreno P, Thong J M V, Marston J & MacCarthy G (2004). 'NewsTuner: a simple interface for searching and browsing radio archives.' IEEE International Conference on Multimedia and Expo (ICME), June.

Meng H, Lo W K, Li Y C & Ching P C (2000). 'Multi-scale audio indexing for Chinese spoken document retrieval.' *Proceedings of the Sixth International Conference on Spoken Language Processing* 4, 101–104.

Meng H M, Chen B, Khudanpur S *et al.* (2004). 'Mandarin-English Information (MEI): investigating translingual speech retrieval.' *Computer Speech & Language* 18, 163–179.

Ng C, Wilkinson R & Zobel J (2000). 'Experiments in spoken document retrieval using phoneme n-grams.' *Speech Communication* 32(1–2), 61–77.

Robertson S E, Walker S, Hancock-Beaulieu M M, Gatford M & Payne A (1996). 'Okapi at TREC-4.' In Harman D K (ed.) *The fourth text retrieval conference (TREC-4)* (special publication No. 500–236). Gaithersburg, MD: National Institute of Standards and Technology. 73–96.

Rosenfeld R (1995). 'Optimizing lexical and Ngram coverage via judicious use of linguistic data.' Proceedings of Eurospeech '95, Madrid, Spain, September.

Siegler M (1999). Integration of continuous speech recognition and information retrieval for mutually optimal performance. Ph.D. diss., Carnegie Mellon University.

Spärck Jones K, Jourlin P, Johnson S E & Woodland P C (2001, July). 'The Cambridge multimedia document retrieval (MDR) project: summary of experiments' (technical report No. 517). Cambridge, UK: Cambridge University Computer Laboratory.

Thong V, Moreno P J, Logan B, Fidler B, Maffey K & Moores M (2002). 'Speechbot: an experimental speech-based search engine for multimedia content on the Web.' *IEEE Transactions on Multimedia* 4(1), 88–96.

Woodland P C, Johnson S E, Jourlin P & Spärck Jones K (2000). 'Effects of out of vocabulary words in spoken document retrieval' (poster session). Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, Athens, Greece, July 24–28.

Yan R, Yang J & Hauptmann A (2004). 'Learning query-class dependent weights in automatic video retrieval.' Proceedings of ACM Multimedia 2004, New York, October 10–16.

AU:4

AU:5

Relevant Websites

<http://www.newstuner.org>—NewsTuner.

<http://www.speechbot.research.compaq.com>—HP
SpeechBot.

<http://www.vhf.org>—Survivors of the SHOAH Visual
History Foundation.

FIRST PROOF