Processing Broadcast Audio for Information Access

Jean-Luc Gauvain, Lori Lamel, Gilles Adda, Martine Adda-Decker,
Claude Barras, Langzhou Chen, and Yannick de Kercadio

Spoken Language Processing Group
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France
(gauvain@limsi.fr http://www.limsi.fr/tlp)

Abstract

This paper addresses recent progress in speaker-independent, large vocabulary, continuous speech recognition, which has opened up a wide range of near and mid-term applications. One rapidly expanding application area is the processing of broadcast audio for information access. At LIMSI, broadcast news transcription systems have been developed for English, French, German, Mandarin and Portuguese, and systems for other languages are under development. Audio indexation must take into account the specificities of audio data, such as needing to deal with the continuous data stream and an imperfect word transcription. Some near-term applications areas are audio data mining, selective dissemination of information and media monitoring.

1 Introduction

A major advance in speech processing technology is the ability of todays systems to deal with non-homogeneous data as is exemplified by broadcast data. With the rapid expansion of different media sources, there is a pressing need for automatic processing of such audio streams. Broadcast audio is challenging as it contains segments of various acoustic and linguistic natures, which require appropriate modeling. A special section in the *Communications of the ACM* devoted to "News on Demand" (Maybury, 2000) includes contributions from many of the sites carrying out active research in this area.

Via speech recognition, spoken document retrieval (SDR) can support random access to relevant portions of audio documents, reducing the time needed to identify recordings in large multimedia databases. The TREC (Text REtrieval Conference) SDR evaluation showed that only small differences in information retrieval performance are observed for automatic and manual transcriptions (Garofolo et al., 2000).

Large vocabulary continuous speech recognition (LVCSR) is a key technology that can be used to enable content-based information access in audio and video documents. Since most of the linguistic information is encoded in the audio channel of video data, which once transcribed can be accessed using text-based tools. This research has been carried out in a multilingual environment in the context of several recent and ongoing European projects. We highlight recent progress in LVCSR and describe some of our work in developing a system for processing broadcast audio for information access. The system has two main components, the speech transcription component and the information retrieval component. Versions of the LIMSI broadcast news transcription system have been developed in American English, French, German, Mandarin and Portuguese.

2 Progress in LVCSR

Substantial advances in speech recognition technology have been achieved during the last decade. Only a few years ago speech recognition was primarily associated with small vocabulary isolated word recognition and with speaker-dependent (often also domain-specific) dictation systems. The same core technology serves as the basis for a range of applications such as voice-interactive database access or limited-domain dictation, as well as more demanding tasks such as the transcription of broadcast data. With the exception of the inherent variability of telephone channels, for most applications it is reasonable to assume that the speech is produced in relatively stable environmental and in some cases is spoken with the purpose of being recognized by the machine.

The ability of systems to deal with non-homogeneous data as is found in broadcast audio (changing speakers, languages, backgrounds, topics) has been enabled by advances in a variety of areas including techniques for robust signal processing and normalization; improved training techniques which can take advantage of very large audio and textual corpora; algorithms for audio segmentation; unsupervised acoustic model adaptation; efficient decoding with long span language models; ability to use much larger vocabularies than in the past - 64 k words or more is common to reduce errors due to out-of-vocabulary words.

With the rapid expansion of different media sources for information dissemination including via the internet, there is a pressing need for automatic processing of the audio data stream. The vast majority of audio and video documents that are produced and broadcast do not have associated annotations for indexation and retrieval purposes, and since most of today's annotation methods require substantial manual intervention, and the cost is too large to treat the ever increasing volume of documents. Broadcast audio is challenging to process as it contains segments of various acoustic and linguistic natures, which require appropriate modeling. Transcribing such data requires significantly higher processing power than what is needed to transcribe read speech data in a controlled environment, such as for speaker adapted dictation. Although it is usually assumed that processing time is not a major issue since computer power has been increasing continuously, it is also known that the amount of data appearing on information channels is increasing at a close rate. Therefore processing time is an important factor in making a speech transcription system viable for audio data mining and other related applications. Transcription word error rates of about 20% have been reported for unrestricted broadcast news data in several languages.

As shown in Figure 1 the LIMSI broadcast news transcription system for automatic indexation consists of an audio partitioner and a speech recognizer.

3 Audio partitioning

The goal of audio partitioning is to divide the acoustic signal into homogeneous segments, labeling and structuring the acoustic content of the data, and identifying and removing non-speech segments. The LIMSI BN audio partitioner relies on an audio stream mixture model (Gauvain et al., 1998). While it is possible to transcribe the continuous stream of audio data without any prior segmentation, partitioning offers several advantages over this straight-forward solution. First, in addition to the transcription of what was said, other interesting information can be extracted such as the division into speaker turns and the speaker identities, and background acoustic conditions. This information can be used both directly and indirectly for indexation and retrieval purposes. Second, by clustering segments from the same speaker, acoustic model adaptation can be carried out on a per cluster basis, as opposed to on a single segment basis, thus providing more adaptation data. Third, prior segmentation can avoid problems caused by linguistic discontinuity at speaker changes. Fourth, by using acoustic models trained on particular acoustic conditions (such as wide-band or telephone band), overall performance can be significantly improved. Finally, eliminating non-speech segments substantially reduces the computation time. The result of the partitioning process is a set of speech segments usually corresponding to speaker turns with speaker, gender and telephone/wide-band labels (see Figure 2).

4 Transcription of Broadcast News

For each speech segment, the word recognizer determines the sequence of words in the segment, associating start and end times and an optional confidence measure with each word. The LIMSI system, in common with most of today's state-of-the-art systems, makes use of statistical models of speech generation. From this point of view, message generation is represented by a language model which provides an estimate of the probability of any given word string, and the encoding of the message in the acoustic signal is represented by a probability density function. The speaker-independent 65k word, continuous speech recognizer makes use of 4-gram statistics for language modeling and of continuous density hidden Markov models (HMMs) with Gaussian mixtures for acoustic modeling. Each word is represented by one or more sequences of context-dependent phone models as determined by its pronunciation. The acoustic and language models are trained on large, representative corpora for each task and language.

Processing time is an important factor in making a speech transcription system viable for automatic indexation of radio and television broadcasts. For many applications there are limitations on the response time and the available computational resources, which in turn can significantly affect the design of the acoustic and language models. Word recognition is carried out in one or more decoding passes with more accurate acoustic and language models used in successive passes. A 4-gram single pass dynamic network decoder has been developed (Gauvain and Lamel, 2000) which can achieve faster than real-time decoding with a word error under 30%, running in less than 100 Mb of memory on widely available platforms such Pentium III or Alpha machines.

5 Multilinguality

A characteristic of the broadcast news domain is that, at least for what concerns major news events, similar topics are simultaneously covered in different emissions and in different countries and languages. Automatic processing carried out on contemporaneous data sources in different languages can serve for multi-lingual indexation and retrieval. Multilinguality is thus of particular interest for media watch applications, where news may first break in another country or language.

At LIMSI broadcast news transcription systems have been developed for the American English, French, German, Mandarin and Portuguese languages. The Mandarin language was chosen because it is quite different from the other languages (tone and syllable-based), and Mandarin resources are available via the LDC as well as reference performance results.

Our system and other state-of-the-art systems can transcribe unrestricted American English broadcast news data with word error rates under 20%. Our transcription systems for French and German have comparable error rates for news broadcasts (Adda-Decker et al., 2000). The character error rate for Mandarin is also about 20% (Chen et al., 2000). Based on our experience, it appears that with appropriately trained models, recognizer performance is more dependent upon the type and source of data, than on the language. For example, documentaries are particularly challenging to transcribe, as the audio quality is often not very high, and there is a large proportion of voice over.

6 Spoken Document Retrieval

The automatically generated partition and word transcription can be used for indexation and information retrieval purposes. Techniques commonly applied to automatic text indexation can be applied to the automatic transcriptions of the broadcast news radio and TV documents. These techniques are based on document term frequencies, where the terms are obtained after standard text processing, such as text normalization, tokenization, stopping and stemming. Most of these preprocessing steps are the same as those used to prepare the texts for training the speech recognizer language models. While this offers advantages for speech recognition, it can lead to IR errors. For better IR results, some words sequences corresponding to acronymns, multiword named-entities (e.g. Los Angeles), and words preceded by some particular prefixes (*anti*, *co*, *bi*, *counter*) are rewritten as a single word. Stemming is used to reduce the number of lexical items for a given word sense. The stemming lexicon contains about 32000 entries and was constructed using Porter's algorithm (Porter80, 1980) on the most frequent words in the collection, and then manually corrected.

The information retrieval system relies on a un-

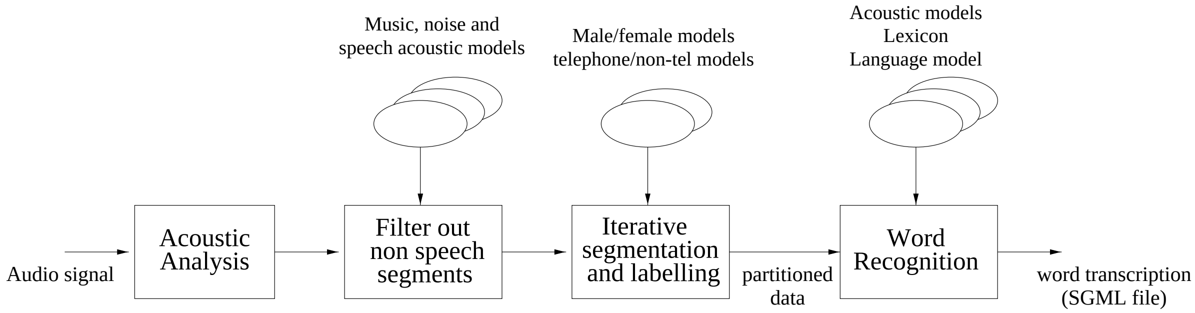

Figure 1: Overview of an audio transcription system. The audio partitioner divides the data stream into homogeneous acoustic segments, removing non-speech portions. The word recognizer identifies the words in each speech segment, associating time-markers with each word.

```
<audiofile filename=19980411_1600_1630_CNN_HDL language=english>
<segment type=wideband gender=female spkr=1 stime=50.25 etime=86.83>
<wtime stime=50.38 etime=50.77> c.n.n.
<wtime stime=50.77 etime=51.10> headline
<wtime stime=51.10 etime=51.44> news
<wtime stime=51.44 etime=51.63> i'm
<wtime stime=51.63 etime=51.92> robert
<wtime stime=51.92 etime=52.46> johnson
it is a day of final farewells in alabama the first funerals for victims of this week's tornadoes are being held today along
with causing massive property damage the twisters killed thirty three people in alabama five in georgia and one each
in mississippi and north carolina the national weather service says the tornado that hit jefferson county in alabama had
winds of more than two hundred sixty miles per hour authorities speculated was the most powerful tornado ever to hit the
southeast twisters destroyed two churches to fire stations and a school parishioners were in one church when the tornado
struck
</segment>
<segment type=wideband gender=female spkr=2 stime=88.37 etime=104.86>
at one point when the table came onto my back i thought yes this is it i'm ready ready protects protect the children because
the children screaming the children were screaming they were screaming in prayer that were screaming god help us
</segment>
<segment type=wideband gender=female spkr=1 stime=104.86 etime=132.37>
vice president al gore toured the area yesterday he called it the worst tornado devastation he's ever seen we will have a
complete look at the weather across the u. s. in our extended weather forecast in six minutes
</segment>
...
<segment type=wideband gender=male spkr=19 stime=1635.60 etime=1645.71>
so if their computing systems don't tackle this problem well we have a potential business disruption and either erroneous
deliveries or misdeliveries or whatever savvy businesses are preparing now so the january first two thousand would just be
another day on the town not a day when fast food and everything else slows down rick lockridge c.n.n.
</segment>
</audiofile>
```

Figure 2: Example system output obtained by automatic processing of the audio stream of a CNN show broadcasted on April 11, 1998 at 4pm. The output includes the partitioning and transcription results. To improve readability, word time stamps are given only for the first 6 words. Non speech segments have been removed and the following information is provided for each speech segment: signal bandwidth (telephone or wideband), speaker gender, and speaker identity (within the show).

Transcriptions	*Werr*	*Base*	*BRF*
Closed-captions	-	46.9%	54.3%
10xRT	20.5%	45.3%	53.9%
1.4xRT	32.6%	40.9%	49.4%

Table 1: Impact of the word error rate on the mean average precision using using a 1-gram document model. The document collection contains 557 hours of broadcast news from the period of February through June 1998. (21750 stories, 50 queries with the associated relevance judgments.)

igram model per story. The score of a story is obtained by summing the query term weights which are simply the log probabilities of the terms given the story model once interpolated with a general English model. This term weighting has been shown to perform as well as the popular TF∗IDF weighting scheme (Hiemstra and Wessel, 1998; Miller et al., 1998; Ng, 1999; Spärk Jones et al., 1998).

The text of the query may or may not include the index terms associated with relevant documents. One way to cope with this problem is to use query expansion (Blind Relevance Feedback, BRF (Walker and de Vere, 1990)) based on terms present in retrieved contemporary texts.

The system was evaluated in the TREC SDR track, with known story boundaries. The SDR data collection contains 557 hours of broadcast news from the period of February through June 1998. This data includes 21750 stories and a set of 50 queries with the associated relevance judgments (Garofolo et al., 2000).

In order to assess the effect of the recognition time on the information retrieval results we transcribed the 557 hours of broadcast news data using two decoder configurations: a single pass 1.4xRT system and a three pass 10xRT system. The word error rates are measured on a 10h test subset (Garofolo et al., 2000). The information retrieval results are given in terms of mean average precision (MAP), as is done for the TREC benchmarks in Table 1 with and without query expansion. For comparison, results are also given for manually produced closed captions. With query expansion comparable IR results are obtained using the closed captions and the 10xRT transcriptions, and a moderate degradation (4% absolute) is observed using the 1.4xRT transcriptions.

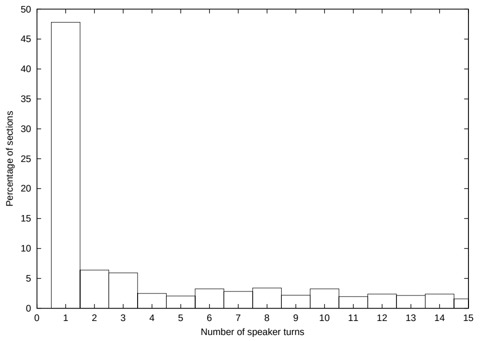

Figure 3: Histogram of the number of speaker turns per section in 100 hours of audio data from radio and TV sources (NPR, ABC, CNN, CSPAN) from May-June 1996.

7 Locating Story Boundaries

The broadcast news transcription system also provides non-lexical information along with the word transcription. This information is available in the partition of the audio track, which identifies speaker turns. It is interesting to see whether or not such information can be used to help locate story boundaries, since in the general case these are not known. Statistics were made on 100 hours of radio and television broadcast news with manual transcriptions including the speaker identities. Of the 2096 sections manually marked as reports (considered stories), 40% start without a manually annotated speaker change. This means that using only speaker change information for detecting document boundaries would miss 40% of the boundaries. With automatically detected speaker changes, the number of missed boundaries would certainly increase. At the same time, 11,160 of the 12,439 speaker turns occur in the middle of a document, resulting in a false alarm rate of almost 90%. A more detailed analysis shows that about 50% of the sections involve a single speaker, but that the distribution of the number of speaker turns per section falls off very gradually (see Figure 3). False alarms are not as harmful as missed detections, since it may be possible to merge adjacent turns into a single document in subsequent processing. These results show that even perfect

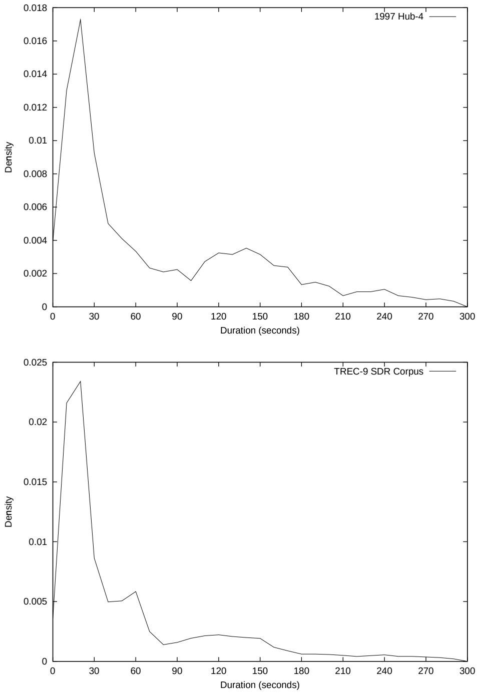

Figure 4: Distribution of document durations for 100 hours of data from May-June 1996 (top) and for 557 hours from February-June 1998 (bottom).

speaker turn boundaries cannot be used as the primary cue for locating document boundaries. They can, however, be used to refine the placement of a document boundary located near a speaker change.

We also investigated using simple statistics on the durations of the documents. A histogram of the 2096 sections is shown in Figure 4. One third of the sections are shorter than 30 seconds. The histogram has a bimodal distribution with a sharp peak around 20 seconds, and a smaller, flat peak around 2 minutes. Very short documents are typical of headlines which are uttered by single speaker, whereas longer documents are more likely to contain data from multiple talkers. This distribution led us to consider using a multi-scale segmentation of the audio stream into documents. Similar statistics were measured on the larger corpus (Figure 4 bottom).

As proposed in (Abberley et al., 1999; Johnson et al., 1999), we segment the audio stream into overlapping documents of a fixed duration. As a result of optimization, we chose a 30 second window duration with a 15 second overlap. Since there are many stories significantly shorter than 30s in broadcast shows (see Figure 4) we conjunctured that it may be of interest to use a double windowing system in order to better target short stories (Gauvain et al., 2000). The window size of the smaller window was selected to be 10 seconds. So for each query, we independently retrieved two sets of documents, one set for each window size. Then for each document set, document recombination is done by merging overlapping documents until no further merges are possible. The score of a combined document is set to maximum score of any one of the components. For each document derived from the 30s windows, we produce a time stamp located at the center point of the document. However, if any smaller documents are embedded in this document, we take the center of the best scoring document. This way we try to take advantage of both window sizes. The MAP using a single 30s window and the double windowing strategy are shown in Table 2. For comparison, the IR results using the manual story segmentation and the speaker turns located by the audio partitioner are also given. All conditions use the same word hypotheses obtained with a speech recognizer which had no knowledge about the story boundaries.

manual segmentation (NIST)	59.6%
audio partitioner	33.3%
single window (30s)	50.0%
double window	52.3%

Table 2: Mean average precision with manual and automatically determined story boundaries. The document collection contains 557 hours of broadcast news from the period of February through June 1998. (21750 stories, 50 queries with the associated relevance judgments.)

From these results we can clearly see the interest of using a search engine specifically designed to retrieve stories in the audio stream. Using an a priori acoustic segmentation, the mean average precision is significantly reduced compared to a “perfect” manual segmentation, whereas the window-based search engine results are much closer. Note that in the manual segmentation all non-story segments such as advertising have been removed. This reduces the risk of having out-of-topic hits and explains part of the difference between this condition and the other conditions.

The problem of locating story boundaries is being further pursued in the context of the ALERT project, where one of the goals is to identify "documents" given topic profiles. This project is investigating the combined use of audio and video segmentation to more accurately locate document boundaries in the continuous data stream.

8 Recent Research Projects

The work presented in this paper has benefited from a variety of research projects both at the European and National levels. These collaborative efforts have enabled access to real-world data allowing us to develop algorithms and models well-suited for near-term applications.

The European project LE-4 OLIVE: *A Multilingual Indexing Tool for Broadcast Material Based on Speech Recognition* (http://twentyone.tpd.tno.nl/ olive/) addressed methods to automate the disclosure of the information content of broadcast data thus allowing content-based indexation. Speech recognition was used to produce a time-linked transcript of the audio channel of a broadcast, which was then used to produce a concept index for retrieval. Broadcast news transcription systems for French and German were developed. The French data come from a variety of television news shows and radio stations. The German data consist of TV news and documentaries from ARTE. OLIVE also developed tools for users to query the database, as well as cross-lingual access based on off-line machine translation of the archived documents, and online query translation.

The European project IST ALERT: *Alert system for selective dissemination* (http://www.fb9-ti.uni-duisburg.de/alert) aims to associate state-of-the-art speech recognition with audio and video segmentation and automatic topic indexing to develop an automatic media monitoring demonstrator and evaluate it in the context of real world applications. The targeted languages are French, German and Portuguese. Major media-monitoring companies in Europe are participating in this project.

Two other related FP5 IST projects are: CORETEX: *Improving Core Speech Recognition Technology* and ECHO: *European CHronicles Online*. CORETEX (http://coretex.itc.it/), aims at improving core speech recognition technologies, which are central to most applications involving voice technology. In particular the project addresses the development of generic speech recognition technology and methods to rapidly port technology to new domains and languages with limited supervision, and to produce enriched symbolic speech transcriptions. The ECHO project (http://pc-erato2.iei.pi.cnr.it/echo) aims to develop an infrastructure for access to historical films belonging to large national audiovisual archives. The project will integrate state-of-the-art language technologies for indexing, searching and retrieval, cross-language retrieval capabilities and automatic film summary creation.

9 Conclusions

This paper has described some of the ongoing research activites at LIMSI in automatic transcription and indexation of broadcast data. Much of this research, which is at the forefront of todays technology, is carried out with partners with real needs for advanced audio processing technologies.

Automatic speech recognition is a key technology for audio and video indexing. Most of the linguistic information is encoded in the audio channel of video data, which once transcribed can be accessed using text-based tools. This is in contrast to the image data for which no common description language is widely adpoted. A variety of near-term applications are possible such as audio data mining, selective dissemination of information (News-on-Demand), media monitoring, content-based audio and video retrieval.

It appears that with word error rates on the order of 20%, comparable IR results to those obtained on text data can be achieved. Even with higher word error rates obtained by running a faster transcription system or by transcribing compressed audio data (Barras et al., 2000; J.M. Van Thong et al., 2000) (such as that can be loaded over the Internet), the IR performance remains quite good.

Acknowledgments

This work has been partially financed by the European Commission and the French Ministry of Defense. The authors thank Jean-Jacques Gangolf, Sylvia Hermier and Patrick Paroubek for their participation in the development of different aspects of the automatic indexation system described here.

References

Dave Abberley, Steve Renals, Dan Ellis and Tony Robinson, "The THISL SDR System at TREC-8", *Proc. of the 8th Text Retrieval Conference TREC-8*, Nov 1999.

Martine Adda-Decker, Gilles Adda, Lori Lamel, "Investigating text normalization and pronunciation variants for German broadcast transcription," *Proc. ICSLP'2000*, Beijing, China, October 2000.

Claude Barras, Lori Lamel, Jean-Luc Gauvain, "Automatic Transcription of Compressed Broadcast Audio *Proc. ICASSP'2001*, Salt Lake City, May 2001.

Langzhou Chen, Lori Lamel, Gilles Adda and Jean-Luc Gauvain, "Broadcast News Transcription in Mandarin," *Proc. ICSLP'2000*, Beijing, China, October 2000.

John S. Garofolo, Cedric G.P. Auzanne, and Ellen M. Voorhees, "The TREC Spoken Document Retrieval Track: A Success Story," *Proc. of the 6th RIAO Conference*, Paris, April 2000. Also John S. Garofolo et al., "1999 Trec-8 Spoken Document Retrieval Track Overview and Results," *Proc. 8th Text Retrieval Conference TREC-8*, Nov 1999. (http://trec.nist.gov).

Jean-Luc Gauvain, Lori Lamel, "Fast Decoding for Indexation of Broadcast Data," *Proc. ICSLP'2000*, **3**:794-798, Oct 2000.

Jean-Luc Gauvain, Lori Lamel, Gilles Adda, "Partitioning and Transcription of Broadcast News Data," *ICSLP'98*, **5**, pp. 1335-1338, Dec. 1998.

Jean-Luc Gauvain, Lori Lamel, Claude Barras, Gilles Adda, Yannick de Kercadio "The LIMSI SDR system for TREC-9," *Proc. of the 9th Text Retrieval Conference TREC-9*, Nov 2000.

Alexander G. Hauptmann and Michael J. Witbrock, "Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval," *Proc Intelligent Multimedia Information Retrieval*, M. Maybury, ed., AAAI Press, pp. 213-239, 1997.

Djoerd Hiemstra, Wessel Kraaij, "Twenty-One at TREC-7: Ad-hoc and Cross-language track," *Proc. of the 8th Text Retrieval Conference TREC-7*, Nov 1998.

Sue E. Johnson, Pierre Jourlin, Karen Spärck Jones, Phil C. Woodland, "Spoken Document Retrieval for TREC-8 at Cambridge University", *Proc. of the 8th Text Retrieval Conference TREC-8*, Nov 1999.

Mark Maybury, ed., Special Section on "News on Demand", *Communications of the ACM*, 43(2), Feb 2000.

David Miller, Tim Leek, Richard Schwartz, "Using Hidden Markov Models for Information Retrieval", *Proc. of the 8th Text Retrieval Conference TREC-7*, Nov 1998.

Kenney Ng, "A Maximum Likelihood Ratio Information Retrieval Model," *Proc. of the 8th Text Retrieval Conference TREC-8*, 413-435, Nov 1999.

M. F. Porter, "An algorithm for suffix stripping", *Program*, **14**, pp. 130–137, 1980.

Karen Spärk Jones, S. Walker, Stephen E. Robertson, "A probabilistic model of information retrieval: development and status," *Technical Report of the Computer Laboratory, University of Cambridge, U.K.*, 1998.

J.M. Van Thong, David Goddeau, Anna Litvinova, Beth Logan, Pedro Moreno, Michael Swain, "SpeechBot: a Speech Recognition based Audio Indexing System for the Web", *Proc. of the 6th RIAO Conference*, Paris, April 2000.

S. Walker, R. de Vere, "Improving subject retrieval in online catalogues: 2. Relevance feedback and query expansion", *British Library Research Paper 72*, British Library, London, U.K., 1990.