EURASIP Journal on Applied Signal Processing 2003:2, 140–150


# Structuring Broadcast Audio for Information Access

**Jean-Luc Gauvain**
*Spoken Language Processing Group, LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France*
*Email: gauvain@limsi.fr*

**Lori Lamel**
*Spoken Language Processing Group, LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France*
*Email: lamel@limsi.fr*

*Received 10 May 2002 and in revised form 3 November 2002*

One rapidly expanding application area for state-of-the-art speech recognition technology is the automatic processing of broadcast audiovisual data for information access. Since much of the linguistic information is found in the audio channel, speech recognition is a key enabling technology which, when combined with information retrieval techniques, can be used for searching large audiovisual document collections. Audio indexing must take into account the specificities of audio data such as needing to deal with the continuous data stream and an imperfect word transcription. Other important considerations are dealing with language specificities and facilitating language portability. At Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), broadcast news transcription systems have been developed for seven languages: English, French, German, Mandarin, Portuguese, Spanish, and Arabic. The transcription systems have been integrated into prototype demonstrators for several application areas such as audio data mining, structuring audiovisual archives, selective dissemination of information, and topic tracking for media monitoring. As examples, this paper addresses the spoken document retrieval and topic tracking tasks.

**Keywords and phrases:** audio indexing, structuring audio data, multilingual speech recognition, audio partitioning, spoken document retrieval, topic tracking.

## 1. INTRODUCTION

The amount of information accessible electronically is growing at a very fast rate. For what concerns the speech and audio processing domain, the information sources of primary interest are radio, television, and telephonic, a variety of which are available on the Internet. Extrapolating from Lesk (1997) [1], we can estimate that there are about 50,000 radio stations and 10,000 television stations worldwide. If each station transmits a few hours of unique broadcasts per day, there are well over 100,000 hours of potentially interesting data produced annually (excluding mainly music, films, and TV series). Although not the subject of this paper, evidently the largest amount of audio data produced consists of telephone conversations (estimated at over 30,000 petabytes annually). In contrast, the amount of textual data can be estimated as a few terabytes annually including newspapers and web texts. Despite the quantity and the rapid growth rate, it is possible to store all this data, should there be a reason to do so. What lacks is an efficient manner to access the content of the audio and audiovisual data.

As an example, the French National Institute of Audiovisual archives (INA) has over 1.5 million hours of audiovisual data. The vast majority of this data has only very limited associated metadata annotations (title, date, topic) which can be used to access the content. This is because today's indexing methods are largely manual, and consequently costly. They also have the drawback that modifications of the database structure or annotation scheme are likely to entail redoing the manual work. Another important application area is media watch where clients need to be aware of ongoing events as soon as they occur. Media watch companies offer services that require listening to and watching all main radio and television stations, and scanning major written news sources. However, given the cost of the services, many smaller and local radio and TV stations cannot be monitored since there is not a large enough client demand.

Automatic processing of audio streams [2, 3, 4] can reduce the need for manual intervention, allowing more information sources to be covered and significantly reducing processing costs while eliminating tedious work. Transcribing and annotating the broadcast audio data is the first step in providing access to its content, and large vocabulary continuous-speech recognition (LVCSR) is a key technology to automate audiovisual document processing [5, 6, 7, 8, 9, 10, 11, 12]. Once transcribed, the content can be accessed using text-based tools adapted to deal with the specificities of spoken language and automatic transcriptions.

The research reported here is carried out in a multilingual environment in the context of several recent and ongoing European projects. Versions of the LIMSI broadcast news transcription system have been developed for the American English, French, German, Mandarin, Portuguese, Spanish, and Arabic languages. The annotations can be used for indexing and retrieval purposes, as was explored in the EC HLT Olive project [13] and currently used in the EC IST Echo project [14] for the disclosure of audiovisual archives from the 20th century. 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 state-of-the-art automatic and manual transcriptions [15]. Another application area concerns detecting and tracking events on particular subjects of interest. The EC IST Alert [16] project and the French national RNRT Theoreme [17] project aim at combining state-of-the-art speech recognition with audio and video segmentation and automatic topic indexing to develop demonstrators for selective dissemination of information and to evaluate them within the context of real-world applications.

To the best of our knowledge, the earliest and longest ongoing work in this area is the Informedia project [18, 19] funded by the National Science Foundation (NSF) under the digital libraries news-on-demand action line. Some other notable activities for the preservation and access to oral heritage, also funded by the NSF, are Multilingual Access to Large Spoken Archives (MALACH) [20] and the National Gallery of the Spoken Word (NGSW) [21].

In this paper, we describe some of the work at LIMSI in developing LVCSR systems for processing broadcast audio for information access. An overview of the speech transcription system is given, which has two main components: an audio partitioner and a speech recognizer. Broadcast audio is challenging to process as it consists of a continuous flow of audio data made up of segments with various acoustic and linguistic natures. Processing such inhomogeneous data thus requires appropriate modeling at both levels. As discussed in Section 4, higher-level linguistic processing for information access also requires taking into account some of the specificities of spoken language.

## 2. AUDIO PARTITIONING

The goal of audio partitioning is to divide the acoustic signal into homogeneous segments, to label and structure the acoustic content of the data, and to identify and remove nonspeech segments. While it is possible to transcribe the continuous stream of audio data without any prior segmentation, partitioning offers several advantages over this straightforward solution. First, in addition to the transcription of what was said, other interesting information can be extracted such as the background acoustic conditions and the division into speaker turns and speaker identities. This information can be used both directly and indirectly for indexing 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 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 wideband or telephone band), overall performance can be significantly improved. Finally, eliminating nonspeech segments substantially reduces the computation time while avoiding potential insertion errors in these segments.

Various approaches have been proposed to partition the continuous stream of audio data. The segmentation procedures can be classified into three approaches: those based on phone decoding [22, 23, 24], distance-based segmentations [25, 26], and methods based on hypothesis testing [11, 27]. The LIMSI BN audio partitioner relies on an audio stream mixture model [28]. Each component audio source, representing a speaker in a particular background and channel condition, is in turn modeled by a mixture of Gaussians. The segment boundaries and labels are jointly identified using an iterative procedure.

The segmentation and labeling procedure introduced in [28, 29] first detects and rejects the nonspeech segments using Gaussian mixture models (GMMs). These GMMs, each with 64 Gaussians, serve to detect speech, pure music, and other (backgrounds). A maximum-likelihood segmentation/clustering iterative procedure is then applied to the speech segments using GMMs and an agglomerative clustering algorithm. Given the sequence of cepstral vectors for the given show, the goal is to find the number of sources of homogeneous data and the places of source changes. The result of the procedure is a sequence of nonoverlapping segments and their associated segment cluster labels, where each segment cluster is assumed to represent one speaker in a particular acoustic environment. More details about the partitioning procedure can be found in [7].

Speaker-independent GMMs corresponding to wideband speech and telephone speech (each with 64 Gaussians) are then used to label telephone segments. This is followed by segment-based gender identification, using 2 sets of GMMs: one for each bandwidth. The result of the partitioning process is a set of speech segments with cluster, gender, and telephone/wideband labels as illustrated in Figure 1.

The partitioner was evaluated to assess the segmentation frame error rate and quality of the resulting clusters [28]. Measured on 3 hours of BN data from 4 shows, the average speech/nonspeech detection frame error rate is 3.7% and the frame error in gender labeling is 1%. Another relevant factor is the cluster homogeneity. To this end, two measures were identified: the cluster purity, defined as the percentage of frames in the given cluster associated with the most represented speaker in the cluster; and the "best-cluster" coverage which is a measure of the dispersion of a given speaker's data across clusters. On average, 96% of the data in a cluster comes from a single speaker. When clusters are impure, they tend to include speakers with similar acoustic conditions. It was also found that on average, 80% of the speaker's data goes to the

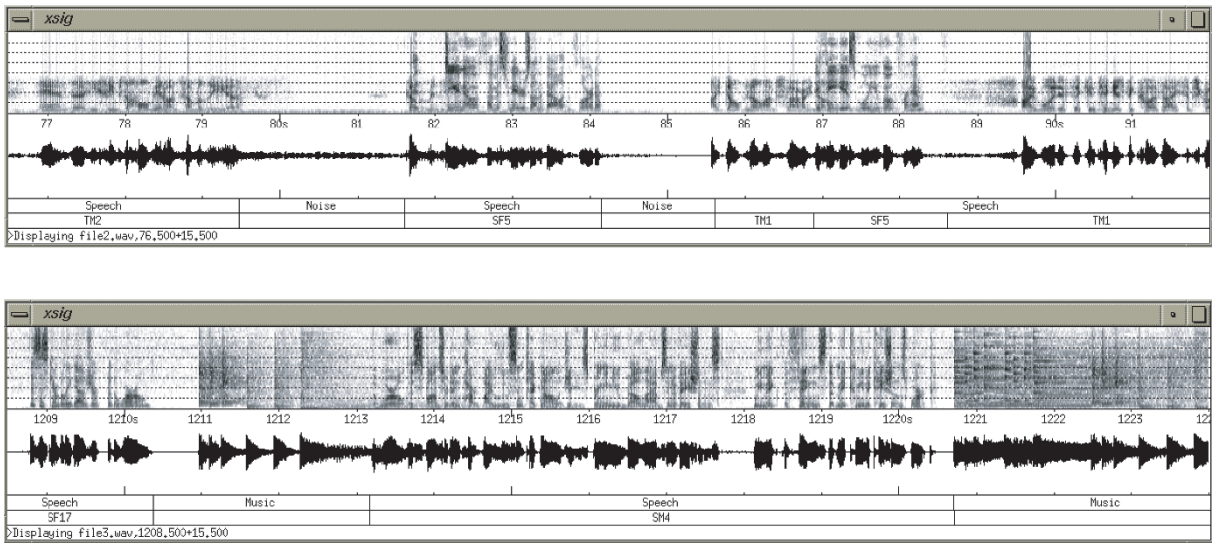


Figure 1: Spectrograms illustrating results of data partitioning on sequences extracted from broadcasts. The transcript gives automatically generated segment type: speech, music, or noise. For the speech segments, the cluster labels specify the identified bandwidth (T = telephone-band/S = wideband) and gender (M = male/F = female), as well as the number of the cluster.

same cluster. In fact, this average value is a bit misleading as there is a large variance in the best-cluster coverage across speakers. For most speakers, the cluster coverage is close to 100%, that is, a single cluster covers essentially all frames of their data. However, for a few speakers (for whom there is a lot of data), the speaker is covered by two or more clusters, each containing comparable amounts of data.

## 3. SPEECH RECOGNITION

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 such applications it is reasonable to assume that the speech is produced in a relatively stable environment and in some cases is spoken with the purpose of being recognized by the machine.

The ability of systems to deal with nonhomogeneous data as found in broadcast audios (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, and the 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 (OOV) words.

One of the criticisms of using LVCSR for audio indexing is the problem of how to deal with OOV words. A speech recognizer can only hypothesize words that it knows about, that is, those that are in the language model and for which there is a correct pronunciation in the system's pronunciation lexicon. In fact, there are really two types of transcription errors that need to be addressed: errors due to misrecognition and errors due to OOV words. The impact of the first type of error can be reduced by keeping alternative solutions for a given speech segment, whereas the second type of error can be solved by increasing the vocabulary size (it is now common to have vocabulary size larger than 100 k words) or by using a lattice of subword units instead of a word level one [30, 31, 32]. This latter alternative to LVCSR results in a considerably less compact representation and, as a result, the search effort during retrieval is more costly. In addition, since word transcripts are not available, the search results can only be browsed by listening, whereas LVCSR offers the possibility of browsing both the transcriptions or the audios. Another way to reduce the OOV problem for the transcription of contemporary data is to use text sources available on the Internet to keep the recognition vocabulary up-to-date [19]. Although keeping the recognition vocabulary up-to-date is quite important for certain tasks, such as media monitoring, when large recognition vocabularies (100 k words or larger) are used, the impact on the overall recognition performance is relatively small.

### 3.1. Recognizer overview

For each speech segment, the word recognizer has to determine the sequence of words in the segment, associating start

and end times and an optional confidence measure with each word.

Most state-of-the-art systems make use of hidden Markov models (HMM) for acoustic modeling which consists of modeling the probability density function of a sequence of acoustic feature vectors. These models are popular as they perform well and their parameters can be efficiently estimated using well-established techniques. A Markov model is described by the number of states and the transitions probabilities between states. The most widely used acoustic units in continuous-speech recognition systems are phone-based and typically have a small number of left-to-right states in order to capture the spectral change across time. Phone-based models offer the advantage that recognition lexicons can be described using the elementary units of the given language and thus benefit from many linguistic studies. Compared with larger units, small subword units reduce the number of parameters and, more importantly, can be associated with back-off mechanisms to model rare or unseen contexts and facilitate porting to new vocabularies.

A given HMM can represent a phone without consideration of its neighbors (context-independent or monophone model) or a phone in a particular context (context-dependent model). The context may or may not include the position of the phone within the word (word-position dependent), and word-internal and cross-word contexts may or may not be merged. Different approaches can be used to select the contextual units based on frequency or using clustering techniques or decision trees, and different types of contexts have been investigated. The model states are often clustered so as to reduce the model size, resulting in what are referred to as "tied-state" models. In the LIMSI system, states are clustered using a decision tree, where the questions concern the phone position, the distinctive features (and identities) of the phone, and the neighboring phones. For the languages under consideration except Mandarin, the recognition-word list contains 65 k entries where each word is associated with one or more pronunciations. For American English, the pronunciations are based on a 48-phone set (3 of them are used for silence, filler words, and breath noises).

The speech recognizer makes use of 4-gram statistics for language modeling and of continuous density 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 (20–200 hours) for each task and language.

Transcribing audiovisual 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.

Table 1: WERs after each decoding step with the LIMSI'99 BN system on the Nov98 evaluation data and the real-time decoding factors (xRT).

| System step | Decoding factor (xRT) | 3-hour test set (NIST Eval98) |
|---|---|---|
| Partitioner | 0.5 | — |
| Step 1 3-gram | 0.8 | 24.7% |
| Step 2 3-gram | 6.0 | 15.4% |
| Step 3 4-gram | 1.5 | 14.2% |

A single-pass, 4-gram dynamic network decoder has been developed [33]. Decoding can be carried out faster than real time on widely available platforms such as Pentium III or Alpha machines (using less than 100 Mb of memory) with a word error rate (WER) under 30%.

Prior to decoding, segments longer than 30s are chopped into smaller pieces so as to limit the memory required by the decoder. To do so, a bimodal distribution is estimated by fitting a mixture of 2 Gaussians to the log-RMS power for all frames of the segment. This distribution is used to determine locations which are likely to correspond to pauses, thus being reasonable places to cut the segment. Word recognition is performed in three steps: (1) initial hypothesis generation; (2) word graph generation; and (3) final hypothesis generation. The first step generates initial hypotheses which are used for cluster-based acoustic model adaptation. Unsupervised acoustic model adaptation of both the means and variances is performed for each segment cluster using the MLLR technique [34]. Acoustic model adaptation is quite important for reducing the WER with relative gains on the order of 20%. Experiments indicate that the WER of the first pass is not critical for adaptation. The second decoding step generates word graphs which are used in the third decoding pass to generate the final hypothesis using a 4-gram language model and adapted acoustic models. The word error rates and the real-time decoding factors after each decoding pass are given in Table 1 for the LIMSI'99 BN system on a 3-hour test set. The same decoding strategy has been successfully applied to the BN transcription in all the languages we have considered.

### 3.2. Language dependencies and portability

A characteristic of the broadcast news domain is that, at least for what concerns major news events, similar topics are simultaneously covered in different transmissions and in different countries and languages. Automatic processing carried out on contemporaneous data sources in different languages can serve for multilingual indexing and retrieval. Multilinguality is thus of particular interest for media watch applications, where news may first break in another country or language. The LIMSI American English broadcast news transcription system has been ported to six other languages.

Porting a recognizer to another language necessitates modifying those system components which incorporate language-dependent knowledge sources such as the phone set, the recognition lexicon, phonological rules, and the language model. Other considerations are the acoustic confusability of the words in the language (such as homophone,

Table 2: Some language characteristics. For each language are specified: the number of phones used to represent lexical pronunciations, the approximate vocabulary size in words (characters for Mandarin) and lexical coverage (of the test data), the test data perplexity with 4-gram language models *(3-gram for Portuguese and Arabic), and the word/character error rates. For Arabic, the vocabulary and language model are vowelized; however, the WER does not include vowel or gemination errors.

| Language | #Phones | Lexicon Size | Coverage | $N$-gram Perplexity | Test %WER |
|---|---|---|---|---|---|
| English | 48 | 65 k | 99.4% | 140 | 18 |
| French | 37 | 65 k | 98.8% | 98 | 23 |
| German | 51 | 65 k | 96.5% | 213 | 25 |
| Mandarin | 39 | 40 k+5 k chars | 99.7% | 190 | 20 |
| Spanish | 27 | 65 k | 94.3% | 159 | 20 |
| Portuguese | 39 | 65 k | 94.0% | 154* | 40 |
| Arabic | 40 | 65 k | 94.3% | 160* | 20 |

monophone, and compound word rates) and the word coverage of a given size recognition vocabulary. There are two predominant approaches taken to bootstrapping the acoustic models for another language. The first is to use acoustic models from an existing recognizer and a pronunciation dictionary to manually segment annotated training data for the target language. If recognizers for several languages are available, the seed models can be selected by taking the closest model in one of the available language-specific sets. An alternative approach is to use a set of global acoustic models, that cover a wide number of phonemes [35]. This approach offers the advantage of being able to use multilingual acoustic models to provide additional training data, which is particularly interesting when only very limited amounts of data (< 10 hours) for the target language are available.

There are some notable language specificities. The number of phones used in the recognition lexicon is seen to range from 25 for Spanish to 51 for German (see Table 2). The Mandarin phone set distinguishes 3 tones, which are associated with the vowels. If the tone distinctions are taken into account, the Mandarin phone set differentiates 61 units. For most of the languages, it is reasonably straightforward to generate pronunciations (and even some predictable variants) using grapheme-to-phoneme rules. These automatically generated pronunciations can optionally be verified manually. A notable exception is the English language, for which most of the pronunciations have been manually derived. Another important language characteristic is the lexical variety. The agglutination and case declension in German result in a large variety of lexical forms. French, Spanish, and Portuguese all have gender and number agreement which increases the lexical variety. Gender and number agreement in French also leads to a high homophone rate, particularly for verb forms. The Mandarin language poses the problem of word segmentation, but this is offset by the opportunity to eliminate OOVs by including all characters in the recognition word list [36]. The Arabic language also is agglutinative, but a larger challenge is to handle the lack of vowelization in written texts. This is compounded by a wide variety of Arabic dialects, many of which do not have a written form.

At LIMSI, broadcast news transcription systems have been developed for 7 languages. To give an idea of the resources used in developing these systems, there are roughly 200 hours of transcribed audio data for American English, about 50 hours for French and Arabic, 20 hours for German, Mandarin, and Spanish, with 3.5 hours for Portuguese. The data comes from a variety of radio and television sources. Obtaining appropriate language-model training data can be difficult. While newspaper and newswire texts are becoming widely available in many languages, these texts are quite different than transcriptions of spoken language. There are also significantly more language-model training texts available for American English (over 1 billion words including 240 million words corresponding to 10,000 hours of commercially produced transcripts). For the other languages, there are on the order of 200–300 million words of language-model training texts, with the exception of Portuguese where only 70 million words are available. It should be noted that French is the only language other than American English for which commercially produced transcripts were available for this work (20 million words).

Some of the system characteristics are shown in Table 2, along with indicative recognition performance rates. The WER on unrestricted American English broadcast news data is about 20% [33, 37]. The transcription systems for French and German have comparable error rates for news broadcasts [38]. The character error rate for Mandarin is also about 20% [36]. 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.

With today's technology, the adaptation of a recognition system to a different task or language requires the availability of sufficient amounts of transcribed training data. Obtaining such data is usually an expensive process in terms of manpower and time. Recent work has focused on reducing this development cost [39]. Standard HMM training requires an alignment between the audio signal and the phone models, which usually relies on an orthographic transcription of the speech data and a good phonemic lexicon. The orthographic transcription is usually considered as ground truth, that is, the word sequence should be hypothesized by the speech recognizer when confronted with the same speech segment. We can imagine training acoustic models in a less supervised manner. Any available related linguistic information about the audio sample can be used in place of the manual transcriptions required for alignment, by incorporating this information in a language model, which can be used to produce the most likely word transcription given the current models. An iterative procedure can successively refine the models and the transcription.

One approach is to use existing recognizer components (developed for other tasks or languages) to automatically transcribe task-specific training data. Although in the

beginning the error rate on new data is likely to be rather high, this speech data can be used to retrain a recognition system. If carried out in an iterative manner, the speech corpus can be cumulatively extended over time *without* direct manual transcription. This approach has been investigated in [40, 41, 42].

There are certain audio sources, such as radio and television news broadcasts, that can provide an essentially unlimited supply of acoustic training data. However, for the vast majority of audio data sources, there are no corresponding accurate word transcriptions. Some of these sources, particularly the main American television channels, also broadcast manually derived closed captions. The closed captions are a close, but inexact, transcriptions of what is spoken and are only coarsely time-aligned with the audio signal. Manual transcripts are also available for certain radio broadcasts. There also exist other sources of information with different levels of completeness such as approximate transcriptions or summaries, which can be used to provide some supervision.

Experiments exploring lightly supervised acoustic model training were carried out using unannotated audio data containing over 500 hours of BN audio data [43]. First, the recognition performance as a function of the available acoustic training data was assessed. With 200 hours of manually annotated acoustic training data (the standard Hub4 training data), a WER of 18.0% was obtained. Reducing the training data by a factor of two increases the WER to 19.1%, and by a factor of 4 to 20.7%. With only 1 hour of training data, the WER is 33.3%. A set of experiments investigated the impact of different levels of supervision via the language model training materials. Language models were estimated using various combinations of the text sources from the same epoch as the audio data or predating the period. Since newspaper and newswire sources have only an indirect correspondence with the audio data, they provide less supervision than the closed captions and commercially generated transcripts [41]. While all of the language models provided adequate supervision for the procedure to converge, those that included commercially produced transcripts in the training material performed slightly better. It was found that somewhat comparable acoustic models could be estimated at 400 hours of automatically annotated BN data and 200 hours of carefully annotated data.

This unsupervised approach was used to develop acoustic models for the Portuguese language for which substantially less manually transcribed data are available. Initial acoustic model trained on the 3.5 hours of available data were used to transcribe 30 hours of Portuguese TV broadcasts. These acoustic models had a WER of 42.6%. By training on the 30 hours of data using the automatic transcripts, the word error was reduced to 39.1%. This preliminary experiment supports the feasibility of lightly supervised and unsupervised acoustic model training.

## 4. ACCESSING CONTENT

The automatically generated partition and word transcription can be used for indexing and retrieval purposes. Techniques commonly applied in automatic text indexing can be applied to the automatic transcriptions of the broadcast news radio and TV documents. The two main application areas investigated are spoken document retrieval (SDR) [15, 37] and topic detection and tracking (TDT) [16, 44]. These applications have been the focus of several European and US projects [13, 14, 16, 17, 18, 20, 21].

There are some differences in processing automatic transcriptions and written texts that should be noted. Spoken language obeys different grammar rules than written language and is subject to disfluencies, fragments, false starts, and repetitions. In addition, there are no clear markings of stories or documents. Using automatic transcriptions is also complicated by recognition errors (substitutions, deletions, insertions). In the DARPA broadcast news evaluations [45], NIST found a strong correlation between WER and some IE metrics [15, 46, 47], and a slightly higher average WER on named-entity tagged words. This can in part be attributed to the limited system vocabulary, which was found to essentially affect only person names. On the positive side, the vocabulary of the system is known in advance and there are no typographical errors to deal with and no need for normalization. The transcripts are time-aligned with the audio data allowing precise access to the relevant portions. Word level confidence measures can also potentially reduce the impact of recognition errors on the information extraction task.

### 4.1. Spoken document retrieval

Via speech recognition, spoken document retrieval can support random access to relevant portions of audio documents, reducing the time needed to identify recordings in large multimedia databases. Commonly used text processing techniques based on document term frequencies can be applied to the automatic transcripts, 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. Some of the processing steps which aim at reducing the lexical variety (such as splitting of hyphenated words) for speech recognition can lead to IR errors. For better IR results, some word sequences corresponding to acronyms, multiword named-entities (e.g., Los Angeles), and words preceded by some particular prefixes (*anti*, *co*, *bi*, *counter*) are rewritten as single words. Stemming is used to reduce the number of lexical items for a given word sense [48].

In the LIMSI SDR system, a unigram model is estimated for each topic or query. The score of a story is obtained by summing the query term weights which are 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 [32, 49, 50, 51] but is more consistent with the modeling approaches used for speech recognition. Since the text of the query may not include the index terms associated with relevant documents, query expansion (blind relevance feedback, BRF [52]) based on terms present in retrieved contemporary texts is used. This is particularly important for

Table 3: Impact of the WER on the MAP 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.)

| Transcriptions | WER | Base | BRF |
|---|---|---|---|
| Closed captions | — | 46.9% | 54.3% |
| 10 xRT | 20.5% | 45.3% | 53.9% |
| 1.4 xRT | 32.6% | 40.9% | 49.4% |

indexing automatic transcripts as recognition errors, and missing vocabulary items can be partially compensated for since the parallel text corpus does not have the same limitations.

The system was evaluated using a data collection containing 557 hours of broadcast news from the period of February through June 1998, and a set of 50 queries with the associated relevance judgments [15]. This data includes 21750 stories with known boundaries. In order to assess the effect of the recognition time on the information retrieval results, the 557 hours of broadcast news data were transcribed using two decoder configurations of the LIMSI BN system: a three-pass 10 xRT system and a single-pass 1.4 xRT system [15]. The information retrieval results with and without query expansion are given in Table 3 in terms of mean average precision (MAP) as done for the TREC benchmarks. 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 10 xRT transcriptions, and a moderate degradation (4% absolute) is observed using the 1.4 xRT transcriptions. For WERs in the range of 20%, only small differences in the MAP were found with manually and automatically produced transcripts, when using query expansion on contemporaneous data.

The same basic technology was used in the European project LE-4 Olive: *A Multilingual Indexing Tool for Broadcast Material Based on Speech Recognition* which addressed methods to automate the disclosure of the information content of broadcast data, thus allowing content-based indexing. 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 comes from a variety of television news shows and radio stations. The German data consists 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 on-line query translation.

Automatic processing of audiovisual archives presents somewhat different challenges due to the linguistic content and wide stylistic variability of the data [53]. The objective of the European Community-funded Echo project is to provide technological support for digital film archives and to improve accessibility and searchability of large historical audio visual archives. Automatic transcription of the audio channel is the first step toward enabling automatic-content analysis. The Echo corpus consists of documents from national audiovisual archives in France, Italy, Holland, and Switzerland. The French data are provided by INA and cover the period from 1945 to 1995 on the construction of Europe. An analysis of the quality of the automatic transcriptions shows that in addition to the challenges of transcribing heterogeneous broadcast news (BN) data, the properties of the archive (audio quality, vocabulary items, and manner of expression) change over time. Several paths are being explored in an attempt to reduce the mismatch between contemporary statistical models and the archived data. New acoustic models were trained in order to match the bandwidth of the archive data and for speech/nonspeech detection. In order to deal with lexical and linguistic changes, sources of texts covering the data period were located to provide information from the older periods. An epoch corpus was created by extracting texts covering the period from 1945 to 1979 from a French video archive web site, which was used to adapt the contemporary language models. Due to a mismatch in acoustic conditions, the standard BN partitioner discarded some speech segments resulting in unrecoverable errors for the transcription system. Training new speech/nonspeech models on a subset of the data recovers about 80% of the partitioning errors at the frame level [53]. Interpolating language models trained on contemporary data with language models trained on data from older periods reduced the perplexity by about 9% but did not result in any significant reduction in the WER. Thus we can conclude that while the acoustic mismatch can be handled in a relatively straightforward manner, dealing with the linguistic mismatch is more challenging.

### 4.2. Locating story boundaries

While story boundaries are often marked or evident in many text sources, this is not the case for audio data. In fact, it is quite difficult to identify stories in a document without having some a priori knowledge about its nature. Story segmentation algorithms must take into account the specificity of each BN source in order to do a reasonable job [54]. The broadcast news transcription system also provides nonlexical information along with the word transcription. This information results from the automatic partitioning 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 carried out on 100 hours of radio and television broadcast news with manual transcriptions including the speaker identities showed that only 60% of annotated stories begin with a manually annotated speaker change. This means that using perfect speaker change information alone for detecting document boundaries would miss 40% of the boundaries. With automatically detected speaker changes, the number of missed boundaries would certainly increase. It was also observed that almost 90% of speaker turns occur in the middle of a document, which means that the vast majority of speaker turns do not signify story changes. Such false alarms are less harmful than missed detections since it may be possible to merge adjacent turns into a single document in subsequent processing. These results indicate, however, that

Table 4: MAP with manually 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.)

| | |
|---|---|
| Manual segmentation | 59.6% |
| Audio partitioner | 33.3% |
| Single window (30 s) | 50.0% |
| Double window | 52.3% |

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

The histogram of the duration of over 2000 American English BN document sections had a bimodal distribution [37], with a sharp peak around 20 seconds corresponding to headlines uttered by single speaker. A second smaller, flat peak was located around 2 minutes. This peak corresponds to longer documents which are likely to contain data from multiple talkers. This bimodal distribution suggested using a multiscale segmentation of the audio stream into documents.

We can also imagine performing story segmentation in conjunction with topic detection or identification, for instance, as in a topic tracking task; but for document retrieval tasks, since the topics of interest are not known at the time the document is processed, such an approach is not very viable. One way to address this problem is to use a sliding window-based approach with a window small enough to not include more than one story but large enough to get meaningful information about the story [55, 56]. For US BN data, the optimal configuration was found to be a 30-second window duration with a 15-second overlap. The 30-second window size is too large, however, to detect the short 20-second headlines. A second 10-second window can be used in order to better target short stories [37]. So for each query, two sets of documents, one set for each window size, are then independently retrieved. 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 30 s windows, a time stamp is located at the center point of the document. However, if any smaller documents are embedded in this document, time stamp is located at the center of the best scoring document taking advantage of both window sizes. This windowing scheme can be used for both information retrieval and on-line topic tracking applications.

The MAP using a single 30 s window and the double windowing strategy are shown in Table 4. 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. These results clearly show the interest in using a search engine specifically designed to retrieve stories in the audio stream. Using an a priori acoustic segmentation, the MAP 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 nonstory segments such as advertising have been removed. This reduces the risk of having out-of-topic hits and explains a part of the difference between this condition and the other conditions.

### 4.3. Topic tracking

Topic tracking consists of identifying and flagging on-topic segments in a data stream. A topic-tracking system was developed which relies on the same topic model as used for SDR, where a topic is defined by a set of keywords and/or topic-related audio and/or textual documents. This information is used to train a topic model, which is then used to locate on-topic documents in an incoming stream. The flow of documents is segmented into stories, and each story is compared to the topic model to decide if it is on- or off-topic. The similarity measure of the incoming document is the normalized likelihood ratio between the topic model and a general language model.

This technique can be applied in media-watch applications (IST Alert [16], RNRT Theoreme [17]) and to structure multimedia digital libraries (IST Echo [14]). Selective dissemination of information and media monitoring require identifying specific information in multimedia data such as TV or radio broadcasts, and alerting users whenever topics they are interested in are detected. Alerts concern the filtering of documents according to a known topic which may be defined by using explicit keywords or by a set of related documents.

A version of the LIMSI topic-tracking system was assessed in the NIST Topic Detection and Tracking (TDT2001) evaluation on the topic-tracking task [44]. For this task, a small set of on-topic news stories (one to four) is given for training and the system has to decide for each incoming story whether it is on- or off-topic. One of the difficulties of this task is that only a very limited amount of information about the topic may be available in the training data, in particular, when there is only one training story. The amount of information also varies across stories and topics: some stories contain fewer than 20 terms after stopping and stemming, whereas others may contain on the order of 300 terms. In order to compensate for the small amount of data available for estimating the on-topic model, document expansion techniques [37] relying on external information sources like past news were used, in conjunction, with unsupervised on-line adaptation techniques to update the on-topic model with information obtained from the test data itself. On-line adaptation consists of updating the topic model by adding incoming stories identified as on-topic by the system as long as the stories have a similarity score higher than an adaptation threshold. Compared with the baseline tracker, the combination of these two techniques reduced the tracking error by more than 50% [44].

A topic identification system has also been developed in conjunction with the LIMSI SDR system. This system

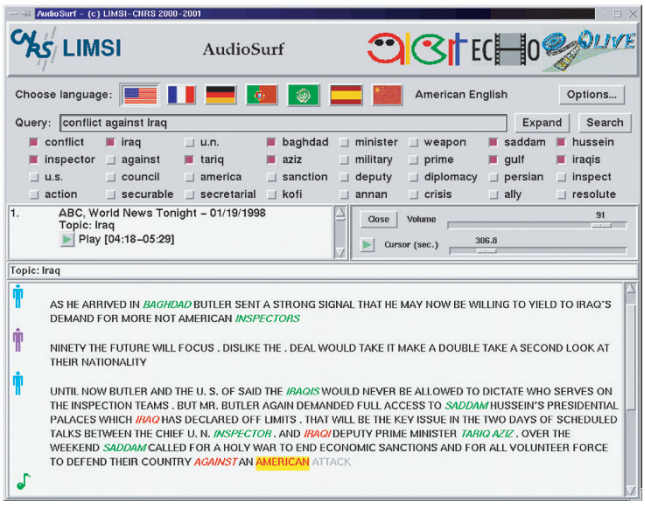


FIGURE 2: Example screen of the LIMSI SDR system able to process audio data in 7 languages. Shown is the sample query, the results of query expansion, the automatically detected topic, and the automatic transcription with segmentation into speaker turns and the document internal speaker identity. (Interface designed by Claude Barras.)

segments documents into *stories* dealing with only one topic, based on a set of 5000 predefined topics, each identified by one or more keywords. The topic model is trained on one or more on-topic stories, and segmentation and identification are simultaneously carried out using a Viterbi decoder. This approach has been tested on a corpus of one year of commercial transcriptions of American radio-TV broadcasts, with a correct topic identification rate of over 60%.

Figure 2 shows the user interface of the LIMSI BN audiovisual document retrieval system which is able to process data in 7 languages. This screen copy shows a sample query, the results of query expansion, the automatically identified topic, and the automatic transcription with segmentation into speaker turns as well as the document internal speaker identity.

In the framework of the Alert project, the capability of processing Internet audio has been added to the system. This capability was added to meet the needs of automatic processing of broadcast data over the web [57]. Given that the Internet audio is often highly compressed, the effect of such compression on transcription quality was investigated for a range of compression rates and compression codecs [58]. It was found that transmission rates at or above 32 kbps had no significant impact on the accuracy of the transcription system, thereby validating that the Internet audio can be automatically processed.

## 5. CONCLUSIONS

We have described some of the ongoing research activities at LIMSI in automatic transcription and indexing of broadcast data, demonstrating that automatic structuring of the audio data is feasible. Much of this research, which is at the forefront of today's technology in signal and natural language processing, 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 adopted. 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 WERs on the order of 20%, SDR results comparable to those obtained on manual transcriptions can be achieved. Even with somewhat higher word error rates (around 30%) obtained by running a faster transcription system or by transcribing compressed audio data (such as that can be loaded over the Internet), the SDR performance remains acceptable. However, to address a wider range of information extraction and retrieval tasks, we believe it is necessary to significantly reduce the recognition word error. One evident need is to provide a suitable transcript for browsing, and the quality output by a state-of-the-art transcriptions system is insufficient.

One outstanding challenge is automatically keeping the models up-to-date, in particular, by using sources of contemporary linguistic data. Another challenge is widening the type of audiovisual documents that can be automatically structured such as documentaries and teleconferences which have more varied linguistic and acoustic characteristics.

## ACKNOWLEDGMENTS

This work has been partially financed by the European Commission and the French Ministry of Defense. The authors thank their colleagues in the Spoken Language Processing Group at LIMSI for their participation in the development of different aspects of the automatic transcription and indexing systems reported here, from which results have been borrowed. They are also indebted to the anonymous reviewers for their valuable comments.

## REFERENCES

[1] http://www.lesk.com/mlesk/ksg97/ksg.html.
[2] C. Djeraba, Ed., "Special Issue on Content-Based Multimedia Indexing and Retrieval," *Multimedia Tools and Applications*, vol. 14, no. 2, 2001.
[3] M. Maybury, Ed., "Special Section on News on Demand," *Communications of the ACM*, vol. 43, no. 2, pp. 33–34, 35–79, 2000.
[4] M. Yuschik, Ed., "Special Issue on Multimedia Technologies, Applications and Performance," *International Journal of Speech Technology*, vol. 4, no. 3/4, 2001.
[5] P. Beyerlein, X. Aubert, R. Harb-Umbach, et al., "Large vocabulary continuous speech recognition of Broadcast News—The Philips/RWTH approach," *Speech Communication*, vol. 37, no. 1-2, pp. 109–131, 2002.

[6] S. S. Chen, E. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanevsky, and P. Olsen, "Automatic transcription of broadcast news," *Speech Communication*, vol. 37, no. 1-2, pp. 69–87, 2002.

[7] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[8] L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz, and J. Makhoul, "Progress in transcription of broadcast news using Byblos," *Speech Communication*, vol. 38, no. 1, pp. 213–230, 2002.

[9] A. J. Robinson, G. D. Cook, D. P. W. Ellis, E. Fosler-Lussier, S. J. Renals, and D. A. G. Williams, "Connectionist speech recognition of broadcast news," *Speech Communication*, vol. 37, no. 1-2, pp. 27–45, 2002.

[10] A. Sankar, V. Ramana Rao Gadde, A. Stolcke, and F. Weng, "Improved modeling and efficiency for automatic transcription of broadcast news," *Speech Communication*, vol. 37, no. 1-2, pp. 133–158, 2002.

[11] S. Wegmann, P. Zhan, and L. Gillick, "Progress in broadcast news transcription at dragon systems," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 33–36, Phoenix, Ariz, USA, March 1999.

[12] P. C. Woodland, "The development of the HTK broadcast news transcription system: An overview," *Speech Communication*, vol. 37, no. 1-2, pp. 47–67, 2002.

[13] A Multilingual Indexing Tool for Broadcast Material Based on Speech Recognition, Project No. LE4-8364, The European Commission, DG XIII, http://twentyone.tpd.tno.nl/olive/.

[14] European Chronicles On-line, http://pc-erato2.iei.pi.cnr.it/echo/.

[15] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "1999 TREC-8 spoken document retrieval track overview and results," in *Proc. 8th Text Retrieval Conference, TREC-8*, pp. 107–130, Gaithersburg, Md, USA, November 1999, http://trec.nist.gov.

[16] Alert System for Selective Dissemination of Multimedia Information, cost RTD project within the Human Language Technologies, Information Society Technologies (IST) http://alert.uni-duisburg.de/start.html.

[17] http://www-clips.imag.fr/mrim/projets/theoreme.html.

[18] http://www.informedia.cs.cmu.edu/.

[19] A. G. Hauptmann and M. J. Witbrock, "Informedia: News-on-demand multimedia information acquisition and retrieval," in *Proc. Intelligent Multimedia Information Retrieval*, M. Maybury, Ed., pp. 213–239, AAAI Press, Menlo Park, Calif, USA, 1997.

[20] Multilingual Access to Large Spoken Archives, National Science Foundation ITR Program, ITR Project # 0122466, http://www.clsp.jhu.edu/ research/malach/.

[21] The National Gallery of Spoken Word, National Science Foundation http://www.ngsw.org/.

[22] T. Hain, S. E. Johnson, A. Tuerk, P. C. Woodland, and S. J. Young, "Segment generation and clustering in the HTK broadcast news transcription system," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 133–137, Landsdowne, Va, USA, February 1998.

[23] D. Liu and F. Kubala, "Fast speaker change detection for broadcast news transcription and indexing," in *Proc. Eurospeech '99*, vol. 3, pp. 1031–1034, Budapest, Hungary, September 1999.

[24] S. Wegmann, F. Scattone, I. Carp, L. Gillick, R. Roth, and J. Yamron, "Dragon systems' 1997 broadcast news transcription system," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 60–65, Landsdowne, Va, USA, February 1998.

[25] F. Kubala, T. Anastasakos, H. Jin, et al., "Toward automatic recognition of broadcast news," in *Proc. DARPA Speech Recognition Workshop*, pp. 55–60, Arden House, NY, USA, February 1996.

[26] M. Siegler, U. Jain, B. Raj, and R. Stern, "Automatic segmentation and clustering of broadcast news audio," in *Proc. DARPA Speech Recognition Workshop*, pp. 97–99, Chantilly, Va, USA, February 1997.

[27] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127–132, Landsdowne, Va, USA, February 1998.

[28] J.-L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *Proc. International Conf. on Spoken Language Processing*, vol. 5, pp. 1335–1338, Sydney, Australia, December 1998.

[29] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI 1997 Hub-4E transcription system," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 75–79, Landsdowne, Va, USA, February 1998.

[30] M. Clements, P. Cardillo, and M. Miller, "Phonetic searching of digital audio," in *Proc. Broadcast Engineering Conference*, pp. 131–140, Washington, USA, 2001.

[31] D. A. James and S. J. Young, "A fast lattice-based approach to vocabulary independent word spotting," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 377–380, Adelaide, Australia, April 1994.

[32] K. Ng, "A maximum likelihood ratio information retrieval model," in *Proc. 8th Text Retrieval Conference, TREC-8*, pp. 413–435, Gaithersburg, Md, USA, November 1999.

[33] J.-L. Gauvain and L. Lamel, "Fast decoding for indexation of broadcast data," in *Proc. International Conf. on Spoken Language Processing*, vol. 3, pp. 794–798, Beijing, China, October 2000.

[34] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[35] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, 2001.

[36] L. Chen, L. Lamel, G. Adda, and J.-L. Gauvain, "Broadcast news transcription in Mandarin," in *Proc. International Conf. on Spoken Language Processing*, vol. II, pp. 1015–1018, Beijing, China, October 2000.

[37] J.-L. Gauvain, L. Lamel, C. Barras, G. Adda, and Y. de Kercadio, "The LIMSI SDR system for TREC-9," in *Proc. 9th Text Retrieval Conference, TREC-9*, pp. 335–341, Gaithersburg, Md, USA, November 2000.

[38] M. Adda-Decker, G. Adda, and L. Lamel, "Investigating text normalization and pronunciation variants for German broadcast transcription," in *Proc. International Conf. on Spoken Language Processing*, vol. I, pp. 266–269, Beijing, China, October 2000.

[39] Improving Core Speech Recognition Technology, EU shared-cost RTD project, Human Language RTD activities of the Information Society Technologies of the Fifth Framework Programme http://coretex.itc.it.

[40] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Proc. Eurospeech '99*, vol. 6, pp. 2725–2728, Budapest, Hungary, September 1999.

[41] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, no. 1, pp. 115–129, 2002.

[42] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve performance," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 301–305, Landsdowne, Va, USA, February 1998.
[43] C. Cieri, D. Graff, and M. Liberman, "The TDT-2 text and speech corpus," in *Proc. DARPA Broadcast News Workshop*, pp. 57–60, Herndon, Va, USA, February 1999.
[44] Y.-Y. Lo and J.-L. Gauvain, "The LIMSI topic tracking system for TDT2001," in *Proc. Topic Detection and Tracking Workshop*, Gaithersburg, Md, USA, November 2001.
[45] D. S. Pallett, "The role of the National Institute of Standards and Technology in DARPA's Broadcast News continuous speech recognition research program," *Speech Communication*, vol. 37, no. 1-2, pp. 3–14, 2002.
[46] *Proc. DARPA Broadcast News Workshop, (Herndon, Va, USA)*, Morgan Kaufmann Publishers, San Francisco, Calif, USA, February 1999, http://www.nist.gov/speech/publications/darpa99/index.htm.
[47] M. A. Przybocki, J. G. Fiscus, J. S. Garofolo, and D. S. Pallett, "1998 Hub-4 information extraction evaluation," in *Proc. DARPA Broadcast News Workshop, (Herndon, Va, USA)*, pp. 13–18, Morgan Kaufmann Publishers, San Francisco, Calif, USA, February 1999.
[48] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.
[49] D. Hiemstra and W. Kraaij, "Twenty-one at TREC-7: Ad-hoc and cross-language track," in *Proc. 7th Text Retrieval Conference, TREC-7*, pp. 227–238, Gaithersburg, Md, USA, November 1998.
[50] D. Miller, T. Leek, and R. Schwartz, "BBN at TREC-7: Using hidden Markov models for information retrieval," in *Proc. 7th Text Retrieval Conference, TREC-7*, pp. 133–142, Gaithersburg, Md, USA, November 1998.
[51] K. Spärck Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: Development and status," Tech. Rep. 446, University of Cambridge Computer Laboratory, London, UK, 1998.
[52] S. Walker and R. de Vere, "Improving subject retrieval in online catalogues. 2: Relevance feedback and query expansion," British Library Research Paper 72, British Library, London, UK, 1990.
[53] C. Barras, A. Allauzen, L. Lamel, and J.-L. Gauvain, "Transcribing audio-video archives," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. I, pp. 13–16, Orlando, Fla, USA, May 2002.
[54] M. Franz, J. S. McCarley, T. Ward, and W.-J. Zhu, "Segmentation and detection at IBM: Hybrid statistical models and two-tiered clustering," TDT 1999 workshop notebook, 1999.
[55] D. Abberley, S. Renals, D. Ellis, and T. Robinson, "The THISL SDR system at TREC-8," in *Proc. 8th Text Retrieval Conference, TREC-8*, pp. 699–706, Gaithersburg, Md, USA, November 1999.
[56] S. E. Johnson, P. Jourlin, K. Spärck Jones, and P. C. Woodland, "Spoken document retrieval for TREC-8 at Cambridge university," in *Proc. 8th Text Retrieval Conference, TREC-8*, pp. 197–206, Gaithersburg, Md, USA, November 1999.
[57] J.-M. Van Thong, D. Goddeau, A. Litvinova, B. Logan, P. Moreno, and M. Swain, "SpeechBot: a speech recognition based audio indexing system for the web," in *Proc. RIAO 2000 Content-Based Multimedia Information Access*, pp. 106–115, Paris, France, April 2000.
[58] C. Barras, L. Lamel, and J.-L. Gauvain, "Automatic transcription of compressed broadcast audio," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 13–16, Salt Lake City, Utah, USA, May 2001.

**Jean-Luc Gauvain** has been a permanent CNRS Researcher at LIMSI since 1983, where he is the Head of the Spoken Language Processing Group. He received a doctorate in electronics from the University of Paris XI in 1982. His primary research centers on large vocabulary continuous speech recognition and conversational interfaces. Other related research is on speaker identification, language identification, and audio indexing. He has participated in many speech-related projects both at the French and European levels and has led the LIMSI participation in the DARPA/NIST organized evaluations since 1992, in particular, for the transcription of broadcast news data. He has over 160 publications and received the 1996 IEEE SPS Best Paper Award in speech processing. He was a member of the IEEE Signal Processing Society's Speech Technical Committee from 1998 to 2001.

**Lori Lamel** joined LIMSI as a permanent CNRS Researcher in October 1991. She received her Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology in May 1988. Her primary research activities are speaker independent, large vocabulary continuous speech recognition, lexical and phonological modeling, speaker and language identification, and spoken language dialog systems. She has been involved in many European projects related to speech recognition and spoken-language dialog systems. She has over 150 publications and is a member of the Editorial Board of the Speech Communication Journal and the Permanent Council of ICSLP.