Audio Partitioning and Transcription for Broadcast Data Indexation

J.L. Gauvain (gauvain@limsi.fr), L. Lamel (lamel@limsi.fr) and
G. Adda (gadda@limsi.fr)
Spoken Language Processing Group
LIMSI-CNRS, BP 133, 91403 Orsay, France
http://www.limsi.fr/tlp

February 2000

Abstract. This work addresses automatic transcription of television and radio broadcasts in multiple languages. Transcription of such types of data is a major step in developing automatic tools for indexation and retrieval of the vast amounts of information generated on a daily basis. Radio and television broadcasts consist of a continuous data stream made up of segments of different linguistic and acoustic natures, which poses challenges for transcription. Prior to word recognition, the data is partitioned into homogeneous acoustic segments. Non-speech segments are identified and removed, and the speech segments are clustered and labeled according to bandwidth and gender. Word recognition is carried out with a speaker-independent large vocabulary, continuous speech recognizer which makes use of n-gram statistics for language modeling and of continuous density HMMs with Gaussian mixtures for acoustic modeling. This system has consistently obtained top-level performance in DARPA evaluations. Over 500 hours of unpartitioned unrestricted American English broadcast data have been partitioned, transcribed and indexed, with an average word error of about 20%. With current IR technology there is essentially no degradation in information retrieval performance for automatic and manual transcriptions on this data set.

Keywords: audio segmentation, speech recognition, audio indexation

1. Introduction

With the rapid expansion of different media sources for information dissemination, there is a need for automatic processing of the data. For the most part todays methods for transcription and indexation are manual, with humans reading, listening and watching, annotating topics and selecting items of interest for the user. Automation of some of these activities can allow more information sources to be covered and significantly reduce processing costs while eliminating tedious work. Radio and television broadcast shows are challenging to transcribe as they contain signal segments with various acoustic and linguistic natures. The signal may be of studio quality or have been transmitted over a telephone or other noisy channel (i.e., corrupted by additive noise and nonlinear distortions), it can contain speech in the presence of background music and pure music segments. Gradual transitions between segments occur when there is background music or noise with changing volume, whereas abrupt changes are common when there is switching between speakers in different locations. The speech is produced by a wide variety of speakers: news anchors and talk show hosts, reporters in remote locations, interviews with politicians and common people, unknown speakers, new dialects, non-native speakers, etc. Speech from the same speaker may occur in different parts of the broadcast, and with different channel conditions. The linguistic style ranges from prepared speech to spontaneous speech.

Two principle types of problems are encountered in transcribing broadcast news data: those relating to the varied acoustic properties of the signal, and those related to the linguistic properties of the speech. Problems associated with the acoustic signal properties are handled using appropriate signal analyses, by classifying the signal according to segment type and by training specific acoustic models for the different acoustic conditions. This process, known as audio partitioning is described in the next section. Section 3 describes the process for automatically transcribing the speech data. Section 4 presents an evaluation of the word transcription quality and the performance of an information retrieval system using the automatic transcriptions of the data from the 1999 TREC-8 Spoken Document Retrieval task [5].

2. Data Partitioning

While it is evidently 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

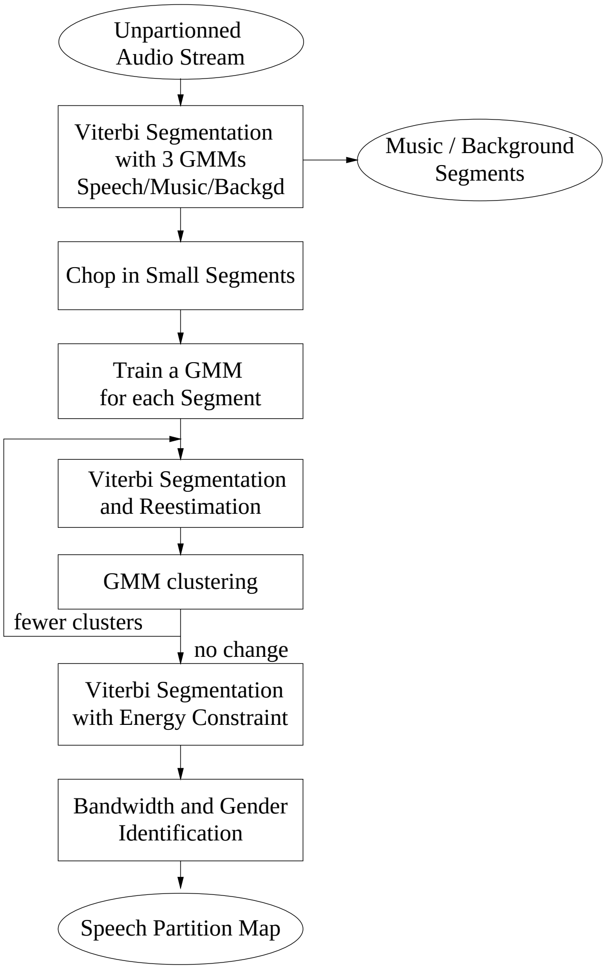

Unpartionned Audio Stream
Viterbi Segmentation with 3 GMMs Speech/Music/Backgd
Music / Background Segments
Chop in Small Segments
Train a GMM for each Segment
Viterbi Segmentation and Reestimation
GMM clustering
fewer clusters
no change
Viterbi Segmentation with Energy Constraint
Bandwidth and Gender Identification
Speech Partition Map

Figure 1. Partitioning algorithm.

the transcription of what was said, other interesting information can be extracted such as the division into speaker turns and the speaker identities. Prior segmentation can avoid problems caused by linguistic discontinuity at speaker changes. By using acoustic models trained on particular acoustic conditions, overall performance can be significantly improved, particularly when cluster-based adaptation is performed. Finally by eliminating non-speech segments and dividing the data into shorter segments (which can still be several minutes long), reduces the computation time and simplifies decoding.

The segmentation and labeling procedure introduced in [9] is shown in Figure 1. First, the non-speech segments are detected (and rejected) using Gaussian mixture models. The GMMs, each with 64 Gaussians, serve to detect speech, pure-music and other (background). The acoustic feature vector used for segmentation contains 38 parameters. It is the same as the recognition feature vector described in the next section, except that it does not include the energy, although the delta energy parameters are included. The GMMs were each trained on about 1h of acoustic data, extracted from the training data after segmentation with the transcriptions. The speech model was trained on data of all types, with the exception of pure music segments and the silence portions of segments transcribed as speech over music. In order to detect speech in noisy conditions a second speech GMM was trained only on noisy speech segments. These model are expected to match all speech segments. The music model was trained only on portions of the data that were labeled as pure music, so as to avoid mistakenly detecting speech over music segments. The silence model was trained on the segments labeled as silence during forced alignment, after excluding silences in segments labeled as containing speech in the presence of background music. All test segments labeled as music or silence are removed prior to further processing.

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 corresponding to a show $(x_1, \ldots, x_T)$, the goal is to find the number of sources of homogeneous data (modeled by the p.d.f. $f(\cdot|\lambda_k)$ with a known number of parameters) and the places of source changes. The result of the procedure is a sequence of non-overlapping segments $(s_1, \ldots, s_N)$ with their associated segment cluster labels $(c_1, \ldots, c_N)$, where $c_i \in [1, K]$ and $K \leq N$. Each segment cluster is assumed to represent one speaker in a particular acoustic environment. In absence of any prior knowledge about the stochastic process governing (K, N) and the segment lengths, we use as objective function a penalized log-likelihood of the form

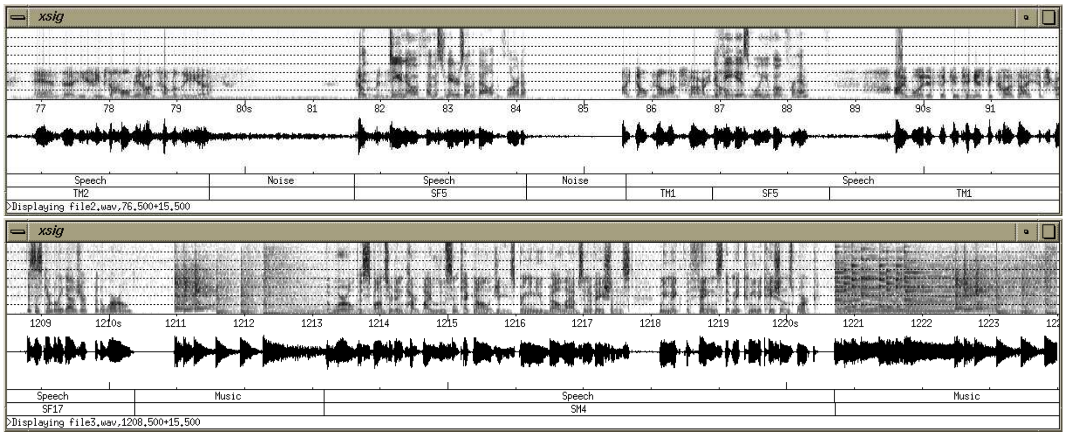
$$\sum_{i=1}^{N} \log f(s_i|\lambda_{c_i}) - \alpha N - \beta K$$

where $\alpha > 0$ and $\beta > 0$. The terms αN and βK, which can be seen as segment and cluster penalties, correspond to the parameters of exponential prior distributions for N and K. It is easy to prove that starting with overestimates of N and K, alternate Viterbi re-estimation and agglomerative clustering gives a sequence of estimates of (K, N, λ_k) with non decreasing values of the objective function. In the Viterbi step we reestimate (N, λ_k) so as to increase $\sum_i \log f(s_i|\lambda_{c_i}) - \alpha N$ (i.e. adding a segment penalty α in the Viterbi search) whereas in the clustering step two or more clusters can be merged as long as the resulting log-likelihood loss per merge is less than β.[1] Since merging two models can reduce the number of segments, the change in segment penalty is taken into account during clustering.

The process is initialized using a simple segmentation algorithm based on the detection of spectral change (similar to the first step used in [17]). The threshold is set so as to over-generate segments, roughly 5 times as many segments as true speaker turns. Initially, the cluster set consists of a cluster per segment. This is followed by Viterbi training of the set of GMMs (one 8-component GMM per cluster). This procedure is controlled by 3 parameters: the minimum cluster size (10s), the maximum log-likelihood loss for a merge (α), and the segment boundary penalty (β). When no more merges are possible, the segment boundaries are refined using the last set of GMMs and an additional relative energy-based boundary penalty, within a 1s interval. This is done to locate the segment boundaries at silence portions, attempting to avoid cutting words (but sometimes this still occurs).

Speaker-independent GMMs corresponding to wide-band speech and telephone speech (each with 64 Gaussians) are then used to identify telephone segments. This is followed by segment-based gender identification, using 2 sets of GMMs with 64 Gaussians (one for each bandwidth). The result of the partitioning process is a set of speech segments

[1] This clustering criterion is closely related to the MDL or BIC criterion.



>Displaying file2.wav,76.500+15.500

>Displaying file3.wav,1208.500+15.500

Figure 2. Spectrograms illustrating results of data partitioning on sequences extracted from broadcasts. The upper transcript is the automatically generated segment type: Speech, Music, or Noise. The lower transcript shows the clustering results for the speech segments, after bandwidth (T=telephone-band/S=wide-band) and gender (M=male/F=female) identification. The number identifies the cluster.

with cluster, gender and telephone/wide-band labels, as illustrated in Figure 2.

We evaluated the frame level segmentation error (similar to [11]) on the 4 half-hour shows in the DARPA Hub-4E eval96 test data [4] using the manual segmentation found in the reference transcriptions. The NIST transcriptions of the test data contain segments that are not scored, since they contain overlapping or foreign speech, and occasionally there are small gaps between consecutive transcribed segments. Since we consider that the partitioner should also work correctly on these portions, we relabeled all excluded segments as speech, music or other background.

Table I(top) shows the segmentation frame error rate and speech/non-speech errors for the 4 shows. The average frame error is 3.7%, but is much higher for show 1 than for the others. This is due to a long and very noisy segment that was deleted. Averaged across shows the gender labeling has a 1% frame error. The bottom of Table I shows measures of the cluster homogeneity. The first entry gives the total number of speakers and identified clusters per file. In general there are more clusters than speakers, as a cluster can represent a speaker in a

Table I. Top: Speech/non-speech frame segmentation error (%), using NIST labels, where missing and excluded segments were manually labeled as speech or non-speech. Bottom: Cluster purity and best cluster coverage (%).

Show	*1*	*2*	*3*	*4*	*Avg.*
Frame Error	7.9	2.3	3.3	2.3	3.7
M/F Error	0.4	0.6	0.6	2.2	1.0
#spkrs/#clusters	7/10	13/17	15/21	20/21	-
ClusterPurity	99.5	93.2	96.9	94.9	95.9
Coverage	87.6	71.0	78.0	81.1	78.7

given acoustic environment. The second measure is the cluster purity, defined as the percentage of frames in the given cluster associated with the most represented speaker in the cluster. (A similar measure was proposed in [3], but at the segment level.) The table shows the weighted average cluster purities for the 4 shows. 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. The "best cluster" coverage is a measure of the dispersion of a given speaker's data across clusters. We averaged the percentage of data for each speaker in the cluster which has most of his/her data. On average, 80% of the speaker's data goes to the 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%, i.e., 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. Transcribing Partitioned Broadcast Data

The speech recognizer uses continuous density hidden Markov models (CD-HMMs) with Gaussian mixture for acoustic modeling and n-gram statistics estimated on large text corpora for language modeling [8]. For acoustic modeling, 39 cepstral parameters are derived from a Mel frequency spectrum estimated on the 0-8kHz band (0-3.5kHz for telephone speech models) every 10 ms. The LPC-based cepstrum coefficients are normalized on a segment cluster basis using cepstral mean removal and variance normalization. Each resulting cepstral coefficient for each cluster has a zero mean and unity variance. Each context-dependent phone model is a tied-state left-to-right CD-HMM with Gaussian mixture observation densities (about 32 components) where the tied states are obtained by means of a phonemic decision tree. Gender-dependent acoustic models were built using MAP adaptation of speaker-independent seed models for wide-band and telephone band speech [6]. The acoustic models for American English were trained on about 150 hours of Broadcast News data.

Language models (LMs) were obtained by interpolation of back-off n-gram language models trained on different data sets: Broadcast news transcriptions, North American Business newspapers and Associated Press Wordstream texts, and transcriptions of the broadcast news acoustic training data. The interpolation coefficients of these 4 LMs were chosen so as to minimize the perplexity on a set of development texts. The recognition vocabulary contains 65122 words and has a lexical coverage of about 99% on the development and test data. The pronunciations are based on a 48 phone set (3 of them are used for silence, filler words, and breath noises). A pronunciation graph is associated with each word so as to allow for alternate pronunciations, including optional phones. Compound words for about 300 frequent word sequences subject to reduced pronunciations were included in the lexicon as well as the representation of frequent acronyms as words.

In order to address variability observed in the linguistic properties, we analyzed differences in read and spontaneous speech, with regard to lexical items, word and word sequence pronunciations, and the frequencies and distribution of hesitations, filler words, and respiration noises. As a result of this analysis, these phenomena were explicitly modeled in both the acoustic and language models as described in [8].

The word decoding procedure is shown in Figure 3. Prior to decoding, segments longer than 30s are chopped into smaller pieces so as to

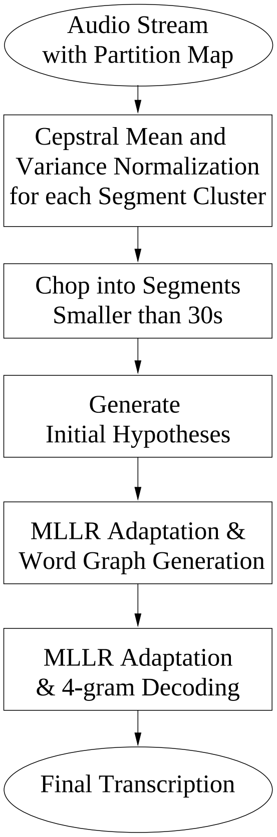


Figure 3. Word decoding.

limit the memory required for the 4-gram decoding pass [8]. 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. Cuts are made at the most probable pause 15s to 30s from the previous cut. Word recognition is performed in three steps: 1) initial hypothesis generation, 2) word graph generation, 3) final hypothesis generation. The initial hypothesis are used for cluster-based acoustic model adaptation using the MLLR technique [14] prior to the 2nd and 3rd decoding passes. The final hypothesis is generated using a 4-gram language model.

The first step generates initial hypotheses which are used for cluster-based acoustic model adaptation. This single pass decoding makes use of a trigram backoff language model (about 8M trigrams and 17M bigrams) and gender-specific sets of 5416 position-dependent, cross word triphones with about 11500 tied states. Band-limited acoustic models are used for the telephone speech segments.

The second decoding step generates accurate word graphs. Unsupervised acoustic model adaptation (both means and variances) is performed for each segment cluster using the MLLR technique [14]. The mean vectors are adapted using a single block-diagonal regression matrix, and a diagonal matrix is used to adapt the variances. Each segment is decoded with the trigram language model and an adapted version of the larger set of acoustic models 28000 position-dependent, cross word triphones with about 11500 tied states (350K Gaussians).

The final hypothesis is generated using a 4-gram language model, and the large set of acoustic models adapted with the hypothesis from the second decoding step.

Broadcast news transcription systems have been also developed for the French and German languages, partially supported by the LE4 Olive project. The same partitioning and recognition algorithms have been successfully applied in conjunction with language-specific lexicons. and acoustic and language models. The French and German lexicons are represented with 37 and 51 phones respectively, including specific phones for silence, breath and fillers. The acoustic models for each language were trained on about 20 hours of audio data from radio and television broadcasts. Trigram backoff language models are formed by interpolation of individual LMs estimated on the transcriptions of the acoustic training data and on texts from newspapers and newswires. The out-of-vocabulary (OOV) rate is 1.15% for the French 65k lexicon, and 4.5% for the German 65k lexicon. The lower lexical coverages than for English are due to the large number of verb forms, and number and gender agreement for French and German and for case declension and compounding in German.

Table II. Summary of broadcast news transcription word error rates for 3 test sets. *Only the 1996 system used a manual partition. All other results are with an automatic partition.

		Test set	
	Eval'96	*Eval'97*	*Eval'98*
System	*1.8 hours*	*3 hours*	*3 hours*
1996 system	**27.1***		
1997 system	25.3	**18.3**	
1998 system	19.8	13.9	**13.6**

4. Evaluation

This section presents an evaluation of the broadcast news transcription system both in terms of transcription accuracy, and the potential for using the automatic transcription for information indexing and retrieval.

4.1. Speech recognizer word accuracy

In Table II reports the word recognition results on DARPA evaluation test sets from the last three years. Each data set contains a few hours of broadcast audio data selected by NIST [4]. The commonly used error metric is the "word error" rate defined as: *%word error = %substitutions + %insertions + %deletions*. The results shown in bold are the official NIST scores obtained by the different systems. For the 1997 system our main development effort was devoted to moving from a manual to an automatic partitioning process. This system nevertheless achieved a performance improvement of 6% on the eval'96 test data. The 1998 system [10] has more accurate acoustic and language models, and achieves a relative word error reduction of over 20% compared to the 1997 system. These tests were carried out without any restriction on the computation time and required over 100 hours to process each hour of data.

Even though it is usually assumed that processing time is not a major issue since computer processing power increase continuously,[2] 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. Transcribing "found" data requires significantly higher processing power than what is needed to transcribed read speech data by speaker adapted dictation systems. This is due to the lack of control of the recordings and linguistic content, which on average results in lower SNR ratios, a poorer fit of the acoustic and language models to the data, and as a consequence a need for larger models. Processing time constraints significantly change the way we select our models. For each operating point, the right balance between model complexity and search pruning level must be found. Two fast systems were optimized for decoding at 10 and 1.4 times real-time (RT), including audio partitioning. On the eval'98 data set the word error rates are 14.2% for the 10xRT system and 24.7% for the 1.4xRT on a Compaq XP1000 500MHz machine.

Figure 4 shows an example portion of an SGML file created from the automatically generated word transcription, taking into account the information available from the partitioning process. Each audio segment starts with a <segment> tag with its start and end times as well as labels for the signal type, gender and speaker. The word transcription is given, with an illustration of the word time codes. Although not shown, a word level confidence score can optionally be associated with each word.

The French and German transcription systems have been evaluated on about 1.5 hours of data. The French data come from television news shows (ARTE) and radio station (France Inter). The German data consist of TV news and documentaries from ARTE. The average word error on the French data is under 20%. The average word error on the German news data is about 20%, and lower than the error on documentaries which is closer to 35%. This difference can be partially

[2] It is common practice to develop systems that run in 100 times real-time or more, especially to evaluate the absolute quality of the acoustic and language models.

```
<audiofile  filename=CSPAN-WJ-960917  language=English>
   <segment type=wideband gender=female spkr=5 stime=81.6 etime=84.2>
     do you know if that mr. nader's on the ballot in florida
   </segment>
   <segment type=telephone gender=male spkr=1 stime=84.72 etime=86.09>
    <wtime stime=84.72 etime=84.97> i
     <wtime stime=84.97 etime=85.22> don't
     <wtime stime=85.22 etime=85.47> know
     <wtime stime=85.47 etime=85.63> i'm
     <wtime stime=85.63 etime=86.09> sorry
   </segment>
   <segment type=wideband gender=female spkr=5 stime=86.09 etime=87.59>
    <wtime stime=86.09 etime=86.21> if
     <wtime stime=86.21 etime=86.41> he
     <wtime stime=86.41 etime=86.67> is
     <wtime stime=86.67 etime=86.79> will
     <wtime stime=86.79 etime=86.94> you
     <wtime stime=86.94 etime=87.16> vote
     <wtime stime=87.16 etime=87.32> for
     <wtime stime=87.32 etime=87.59> him
   </segment>
   <segment type=telephone gender=male spkr=1 stime=87.59 etime=106.22>
     i would if it ...
   </segment>
</audiofile>
```

Figure 4. Example SGML format for the system output. For each segment the signal type, gender and speaker labels, and start and end times are given, as well as the word transcription. For simplicity not all time codes are shown.

attributed to the better language model representivity for the news data.

4.2. Experiments with Spoken Document Retrieval

One of the main motivations for automatic processing of the audio channels of broadcast data is to serve as a basis for automatic disclosure and indexation for information retrieval (IR) purposes. The aim of the

OLIVE project[3] was to develop an archiving and retrieval system for broadcast data to enable efficient access to large multimedia libraries, such as the French INA audio-visual archive [13]. Disclosure of video material plays an important role for the user organizations, but is too costly to carry out manually for all broadcast data. As a result, the vast majority of data is archived with only minimal annotations. The audio stream is automatically partitioned and the speech segments transcribed and time-coded using the methods described above. The transcription is used to generate an index which is linked to the appropriate portions of the audio or video data. 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.

We have assessed the performance in spoken document retrieval (SDR) on 600 hours of audio data (100 hours from TREC-7 SDR'98 and 500 hours from TREC-8 SDR'98). Although for IR purposes the story boundaries are assumed to be known, this information is not used by the speech recognizer. Most of the development work was carried out using the SDR'98 test data (100h), consisting of about 2800 documents with the associated 23 queries. The SDR'99 test data (500h) consists of 21750 documents with an associated set of 50 queries. It should be noted that the reference transcripts of the SDR'98 data are detailed manual transcriptions, whereas for the SDR'99 data these are closed captions.

In order for the same IR system to be applied to different text data types (automatic transcriptions, closed captions, additional texts from newspapers or newswires), all of the documents were preprocessed in a homogeneous manner. This preprocessing, or tokenization, is the same as the text source preparation for training the speech recognizer language models [7], and attempts to transform them to be closer to the observed American speaking style. The basic operations include translating numbers and sums into words, removing all the punctuation symbols, removing case distinctions and detecting acronyms and

[3] The LE4-8364 OLIVE project (http://twentyone.tpd.tno.nl/olive) was funded by the European Commission under the Telematics Application Programme in the sector Language Engineering.

spelled names. However removing all punctuations implies that certain hyphenated words such as *anti-communist*, *non-profit* are rewritten as *anti communist* and *non profit*. While this offers advantages for speech recognition, it can lead to IR errors. To avoid IR problems due to this transformation, the output of the tokenizer (and recognizer) is checked for common prefixes, in order to rewrite a sequence of words as a single word. The prefixes that are handled include *anti*, *co*, *bi*, *counter*. A set of rewrite rules covering compound words formed with these prefixes and a limited number of named entities (such as *Los-Angeles*) is used to transform the texts. Similarly all numbers less than one hundred are treated as a single entity (such as *twenty-seven*).

In order to reduce the number of lexical items for a given word sense, each word is mapped to its stem (as defined in [2, 16]) or, more generally, into a form that is chosen as being representative of its semantic family.

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 do query expansion based on terms present in retrieved documents on the same (Blind Relevance Feedback) or other (Parallel Blind Relevance Feedback) data collections [19]. We combined the two approaches in our system. For the latter 6 months of commercially available broadcast news transcripts from the period of June through December 1997 [1] were used. This corpus contains 50 000 stories and 49.5 M words. For a given query, the terms found in the top 15 documents from the baseline search are ranked by their offer weight [18], and the top 10 terms are added to the query. Since only the terms with best offer weights are kept, the terms are filtered using a stop list of 144 common words, in order to increase the likelihood that the resulting terms are relevant.

The information retrieval system relies on a unigram model per story. 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 [12, 15].

Table III gives the IR results in terms of mean average precision (MAP), as is done for the TREC benchmarks. Four experimental configurations are reported: baseline search (*base*), query expansion using blind relevance feedback (BRF), query expansion with parallel blind relevance feedback (PBRF) and query expansion using both BRF and PBRF. The results clearly demonstrate the interest of using both BRF and PBRF expansion techniques, as consistent improvements are obtained over the baseline system for the two conditions (R1 and S1). Average precisions of 57% and 54% respectively were obtained on the SDR'98 and SDR'99 test sets using the automatic transcriptions. These values are quite close to the average precisions obtained on manual transcripts, even though the 10xRT recognizer transcripts have an estimated 20.5% word error rate. Using transcriptions generated with the 1.4xRT system (word error rate of about 32%), the baseline MAP is 41% and the MAP with query expansion is 49% for the SDR'99 test conditions.

Table III. Mean average precision (%) for the SDR'98 and SDR'99 data sets using unigram term weightings. R1: reference transcript. S1: automatic speech transcription obtained with a 10xRT system.

dataset	*base*	*BRF*	*PBRF*	*BRF+PBRF*
98-R1	46.95	59.36	55.74	58.89
98-S1	45.58	51.21	58.84	57.45
99-R1	46.91	53.54	50.98	54.30
99-S1	44.12	53.02	49.43	53.98

5. Conclusions

In this paper we have presented our recent research in partitioning and transcribing television and radio broadcasts. These are necessary steps to enable automated processing of the vast amounts of audio and video data produced on a daily basis. The data partitioning algorithm makes use of Gaussian mixture models and an iterative segmentation and clustering procedure. The resulting segments are labeled according to gender and bandwidth using 64-component GMMs. The speech detection frame error is less than 4%, and gender identification has a frame error of 1%. Many of the errors occur at the boundary between segments, and can involve silence segments which can be considered as with speech or non-speech without influencing transcription performance.

Word recognition is carried out in multiple passes for each speech segment progressively using more accurate models. The generation of word graphs with adapted acoustic models is essential for obtaining word graphs with low word error rates, particularly in light of the variety of talkers and acoustic conditions. On unrestricted American English broadcast news shows the word error rate is about 20%. Due to the availability of large, transcribed corpora available through the LDC our initial work focused on American English, however, in the context of the LE4 Olive project the transcription system system has been sucessfully ported to the French and German languages with word error rates under 20% for news shows.

Our experience is that radio news shows are usually easier to transcribe than television news shows, probably due to the fact that only the audio channel is used to transmit the information, whereas for television the audio stream is supported by visual data. Broadcast news data is also easier to transcribe than documentaries.

A complete indexing system has been built by applying text IR techniques on the output of our broadcast news speech recognizer. Quite comparable average precisions were obtained on manual and reference transcriptions (which for the SDR'99 data were closed captions), indicating that the transcription quality is not the limiting factor on IR performance for current IR techniques.

Some existing applications that could greatly benefit from this technology are the creation and access to digital multimedia libraries (disclosure of the information content and content-based indexation), media monitoring services (selective dissemination of information based on automatic detection of topics of interest) as well as new emerging applications such as news-on-demand and Internet watch services.

Acknowledgements

This work has been partially financed by the European Commission and the French Ministry of Defense. The authors gratefully acknowledge the participation Martine Adda-Decker in the development of the recognition systems and of Michèle Jardino, Yannick de Kercadio, Rémi Lejeune and Patrick Paroubek to the SDR work.

References

1. PSMedia. http://www.thomson.com/psmedia/bnews.html
2. UMass. ftp://ciir-ftp.cs.umass.edu/pub/stemming/
3. S.S. Chen and P.S. Gopalakrishnan. Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion. *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, Virginia, 127–132, February 1998.
4. J.S. Garofolo, E.M. Voorhees, C.G.P. Auzanne, V.M. Stanford and B.A. Lund. Design and Preparation of the 1996 Hub-4 Broadcast News Benchmark Test Corpora. *Proc. of the DARPA Speech Recognition Workshop*, Chantilly, Virginia, 15–21, February 1997. (see also http://www.nist.gov/speech/tests/).
5. J.S. Garofolo, C.G.P. Auzanne, E.M. Voorhees, B. Fisher. The TREC Spoken Document Retrieval Track: A Success Story. *Proc. 8th Text Retrieval Conference TREC-8*, 107–130, Gaithersburg, Maryland, November 1998.
6. J.L. Gauvain and C.H. Lee. Maximum *a Posteriori* Estimation for Multivariate Gaussain Mixture Observation of Markov Chains. *IEEE Trans. on SAP*, **2**(2), 291–298, April 1994.
7. J.L. Gauvain, L. Lamel, G. Adda and M. Adda-Decker, The LIMSI Nov93 WSJ System. *Proc. ARPA Spoken Language Technologies Workshop*, Plainsboro, New Jersey, 125–128, March 1994.
8. J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, Transcribing Broadcast News: The LIMSI Nov96 Hub4 System. *Proc. ARPA Speech Recognition Workshop*, Chantilly, Virginia, 56–63, February 1997.
9. J.L. Gauvain, Y. de Kercadio, L. Lamel and G. Adda The LIMSI SDR System for TREC-8. *Proc. 8th Text Retrieval Conference TREC-8*, 475–482, Gaithersburg, Maryland, November 1999.
10. J.L. Gauvain, L. Lamel, G. Adda and M. Jardino. The LIMSI 1998 Hub-4E Transcription System. *Proc. DARPA Broadcast News Workshop.* Herndon, Virginia, 99–104, February 1999.
11. T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland, and S.J. Young. Segment Generation and Clustering in the HTK Broadcast News Transcription Sys-

tem. *DARPA Broadcast News Transcription and Understanding Workshop.* Landsdowne, Virginia, 133–137, February 1998.

12. D. Hiemstra and K. Wessel. Twenty-One at TREC-7: Ad-hoc and Cross-language track. *Proc. 7th Text Retrieval Conference TREC-7*, 227–238, Gaithersburg, Maryland, November 1999.
13. F.M.G. de Jong, J.L. Gauvain, J. den Hartog, K. Netter, OLIVE: Speech Based Video Retrieval. *Proc. CBMI'99*, Toulouse, France, October 1999.
14. C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, **9**(2), 171–185, 1995.
15. D.R.H. Miller, T. Leek and R.M. Schwartz. BBN at TREC7: Using Hidden Markov Models for Information Retrieval *Proc. 7th Text Retrieval Conference TREC-7*, 133–142, Gaithersburg, Maryland, November 1999.
16. M.F. Porter. An Algorithm for Suffix Stripping. *Program* **14**(3), 130–137, 1980.
17. M. Siegler, U. Jain, B. Raj, R. Stern. Automatic Segmentation, Classification and Clustering of Broadcast News Audio. *Proc. DARPA Speech Recognition Workshop.* Chantilly, Virginia, 97–99, February 1997.
18. K. Spärk Jones, S. Walker and S. E. Robertson. A probabilistic model of information retrieval: development and status. *a Technical Report of the Computer Laboratory*, University of Cambridge, U.K., 1998.
19. 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, U.K., 1990.