Project ref. no.	*IST-1999 10354*
Project title	**ALERT - Alert System for selective dissemination of multimedia information**

Deliverable status	*Public*
Contractual date of delivery	t30 (delayed until t34)
Actual date of delivery	t36
Deliverable number	*D4.1*
Deliverable title	*Report on Topic Detection on the output of a speech recognizer*
Type	*RE-Report*
Status and version	*Final version*
Number of pages	*44*
WP contributing to the deliverable	*WP4*
WP / Task responsible	*LIMSI*
Author(s)	*R. Amaral, I. Trancoso (INESC), C. Barras, E. Bilinski, J.L. Gauvain, L. Lamel, Y.Y. Lo, (LIMSI), U. Iurgel, A. Kosmala (UniDu)*
EC Project Officier	*Mats Ljungqvist*
Keywords	*topic detection, story segmentation, topic spotting, topic tracking*
Abstract (for dissemination)	This document overviews the work carried out in the ALERT project concerning the detection of topics in audiovisual data using the output of a speech recognizer. The objectives of this workpackage are two-fold: to automatically divide the audio stream into topically homogeneous segments and associate one or more topics with each audio segment. The emphasis is to develop statistical methods for topic detection, comparing the performance of different techniques in the presence of speech recognition errors.

Contents

1 Introduction

This document provides an overview of the work carried out in the Workpackage 4 of the ALERT project addressing Automatic Topic Detection for multimedia data using the output of a speech recognizer. The objectives of this workpackage are two-fold: to automatically divide the audio stream into topically homogeneous segments and associate one or more topics with each audio segment. Due to the availability of topic-labeled data in American English and benchmark tests (SDR, TDT), some of the investigations were first developed for English and then applied and adapted for use with the French, German and Portuguese languages. The main advances in this workpackage are highlighted, leaving the more detailed descriptions to the included list of related articles. These publications are all available on the project website [1] or at the individual partner websites.

The next section describes the general topic detection problem, highlighting some of the specificities of detecting topics in automatically generated transcriptions. ALERT demonstrators have been developed for the French, German and Portuguese languages in order to address the needs of the industrial partners in the consortium. The topic detection aspects of these demonstrators are described in the following three sections written by INESC, LIMSI and UNIDU respectively.

2 Automatic Topic Detection

Keeping aware of information is of strategic importance for many businesses and governmental agencies. With the rapid expansion of different media sources (newspapers, newswire, radio, television, internet) for information dissemination, there is a growing demand for automatic processing for monitoring the different data sources. In the ALERT project we have explored the combined use of state-of-the-art speech recognition with audio and video segmentation and automatic topic detection in order to develop an automatic media monitoring demonstration system. One of the project aims has been to reduce the manual intervention required to detect topics of interest for users.

This workpackage has been concerned with the development of statistical methods for topic detection. Statistical methods for topic detection are particularly attractive when applied to automatically generated transcriptions, since they are less sensitive to the characteristics of spoken language and to transcription errors than approaches relying on a linguistic analysis. Since standard text-based techniques make use of punctuation markers often combined with syntactic analyses, they cannot be directly applied to the output of automatic speech recognizers for which these markers are generally lacking. The topic detection algorithms also have to deal with multiple topics per story, an unknown number of topics and with stories of differing sizes.

[1] http://alert.uni-duisburg.de

Statistical methods are also less dependent upon the a priori definition of topic-specific keywords, since they can be trained on sample on-topic stories. There can be a variety of ways to speak about a given topic, and in general the presence of particular keywords is not mandatory for the topic to be present. Statistical methods exploit the probability distribution of words in the transcription for each topic of interest.

The topic detectors provide an ordered list of topics with associated probabilities. The decision module (which is used to decide whether or not a segment is on-topic) minimizes a detection cost function which takes into account the relative costs of two types of errors: false alarm (incorrectly detecting a topic in an off-topic segment) and miss (not detecting an on-topic segment).

3 Topic Detection for Portuguese (INESC)

The huge amount of information we can access nowadays in very different formats (audio, video, text) and through distinct channels revealed the necessity to build systems that can efficiently store and retrieve this data in order to satisfy future information needs. This is the framework for the ALERT European Project (Alert System for Selective Dissemination of Multimedia Information), whose goal was to build a system capable of continuously monitoring a TV channel, and searching inside their news programs for the stories that match the profile of a given client. The system may be tuned to a particular TV channel in order to automatically detect the start and end of a broadcast news program. Once the start is detected, the system automatically records, transcribes, indexes and stores the program. Each of the segments or stories that have been identified is indexed according to a thematic thesaurus. The system then searches in all the client profiles for the ones that fit into the detected categories. If any topic story matches the client preferences, an email is send to that client indicating the occurrence and location of one or more stories about the selected topics. This alert message enables a client to find in the System Website the video clips referring to the selected stories. This report concerns only the segmentation and indexation modules of the Alert system.

The broadcast news corpus and thesaurus used in this work are described in Section 3.1. The following two sections present our segmentation and indexation algorithms, respectively. Section 3.4 presents the story segmentation results, using as input stream data that was automatically segmented into sentences together with information about background acoustical environment and speaker identification for each sentence. Section 3.5 shows the results of an indexation task where the descriptors of the thematic thesaurus were used as indexing keys in stories whose boundaries were manually identified. The report concludes with a discussion of these results and our plans for future research in this area.

3.1 Topic Detection Corpus Description

This section presents the Topic Detection Corpus (TDC) that was used to develop and test our indexation algorithm. In this work, the “topic” concept is formally defined by a set of descriptors included in a thematic thesaurus. According to this topic definition the TV Broadcast News Corpus in European Portuguese was manually segmented and topic indexed, in cooperation with the national public broadcasting company - RTP (*Rádio Televisão Portuguesa*).

3.1.1 Broadcast News Corpus

Collected in the scope of the ALERT project over a period of 9 months (February - October 2001), the BN corpus contains around 300 hours of audio data from 133 TV broadcast evening news programs. The corresponding orthographic transcriptions were automati-

cally generated by our speech recognition engine [20]. All the programs were manually segmented into stories or fillers, and each story was also manually indexed according to a thematic, geographic and onomastic thesaurus. The manual segmentation also identified the so called filler segments, containing either headlines or short story descriptions presented to draw the audience attention to stories that will be presented later in the program. Filler segments were not indexed.

3.1.2 Thematic, Geographic and Onomastic Thesaurus

Manual indexation was done using a thematic, geographic and onomastic (names of persons and organizations) thesaurus by RTP trained annotators.

The Thematic Thesaurus contains 21 hierarchical concept trees whose domains are:

Justice and Rights (JR)
Defense and Security (DS)
Society and Sociology (SS)
Political Organisation (PO)
Sports and Leisure (SL)
Transportation (TR)
Science and Technology (ST)
Communication and Documentation (CD)
Work and Employment (WE)
Economy and Finance (EF)
Health and Feeding (HF)
Religion and Ethics (RE)
Arts and Culture (AC)
House and Living (HL)
Industry (IN)
Environment and Energy (EE)
Agriculture (AG)
European Union (EU)
History (HI)
Weather Forecast (WF)
Events (EV)
Education (ED)

The Thematic Thesaurus contains 7781 descriptors and 1615 non-descriptors whose relations are established via:

- hierarchical relations of the type **general term (GT)** and **specific term (ST)**;
- associative relations such as **related term (RT)**

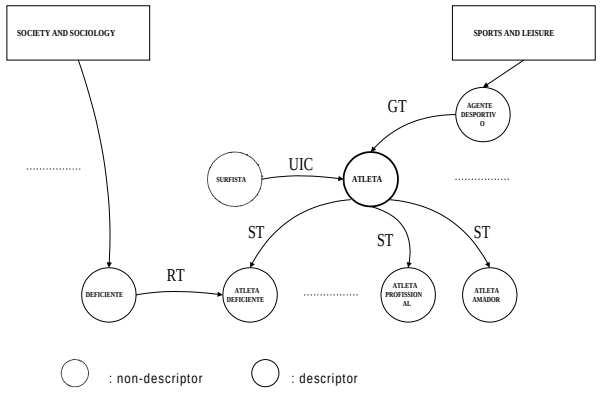

Figure 1: Hierarchical concept tree - Thematic Thesaurus

- equivalence relations such as **use for (UF)** or **use in case (UIC)**

Besides the above relations, there is a tag called **explanation node (NE)**, which gives some contextual indications for the use of the corresponding descriptor.

Figure 1 illustrates the thesaurus structure. Let us focus our attention in the thesaurus **descriptor** *atleta* (athlete). This descriptor has a **general term** called *agente desportivo* (sports agent), which is its immediate upper node of the tree, and belongs to the Sports and Pleasure thematic domain. The descriptor *atleta* has at least three **specific terms**, namely: *atleta deficiente*, *atleta amador* and *atleta profissional* (handicapped, amateur and professional). That means that any of these three terms have a higher degree of specification than the upper node for the current domain. In the example, we can also see the **non-descriptor** term *surfista* (surfer) that when present should be replaced by the thesaurus descriptor *atleta*. This descriptor has an **explanation note** that indicates the situations where the descriptor should be used. The descriptor *atleta deficiente* has at least one **related term** called *deficiente* (handicapped) meaning that there is at least one related descriptor in another thematic domain tree (Society and Sociology).

The distribution of the descriptors among the several levels is represented in table 1.

The onomastic and geographic thesauri have 1765 and 1890 entries, respectively. The first ones include institution names, as well as person names. These entries are used to identify the story speakers, and not the persons who are the subject of the story.

Thesaurus level	Descriptors %
1st-level	0.21%
2nd-level	7.62%
3rd-level	48.32%
4th-level	26.47%
5th-level	11.83%
6th-level	3.84%
7th-level	0.86%
8th-level	0.63%
9th-level	0.19%
10th-level	0.03%

Table 1: Descriptors distribution among thesaurus levels.

3.1.3 Training, Development and Evaluation Subsets

The topic detection corpus was divided into three subsets for training, development and evaluation purposes.

The training corpus was collected from March to mid August 2001. It includes 85 programs, corresponding to 2451 report segments and 530 fillers. The report segments involve 6073 stories with 312 words each, on average. Very frequently, a report segment is classified into more than one topic (for instance, Sports and Leisure and Society and Sociology, as in the example above). Such report segments will originate multiple stories, which justifies the difference between the number of report segments and stories. In the above case, a single report segment will originate one story used for building the SL language model and another for building the SS language model. The distribution of the thematic domains among the story programs is shown in Figure 2.

In the next figure (fig. 3) we can see the relation between one topic stories and multiple topic stories for each of the 22 domains.

The development corpus was collected in September 2001 and includes 21 programs, corresponding to 699 report segments and 144 fillers. The report segments involve 1172 stories, whose thematic distribution is shown in figure 4.

The evaluation corpus was collected in October 2001. It includes 27 programs corresponding to 871 report segments, and 134 fillers. The segments involve 1528 stories, whose thematic distribution is shown in Figure 5.

3.2 Story Segmentation

The input to the segmentation algorithm is a stream of audio data, which was automatically segmented into sentences (or rather “transcript segments”, defined by pauses), and

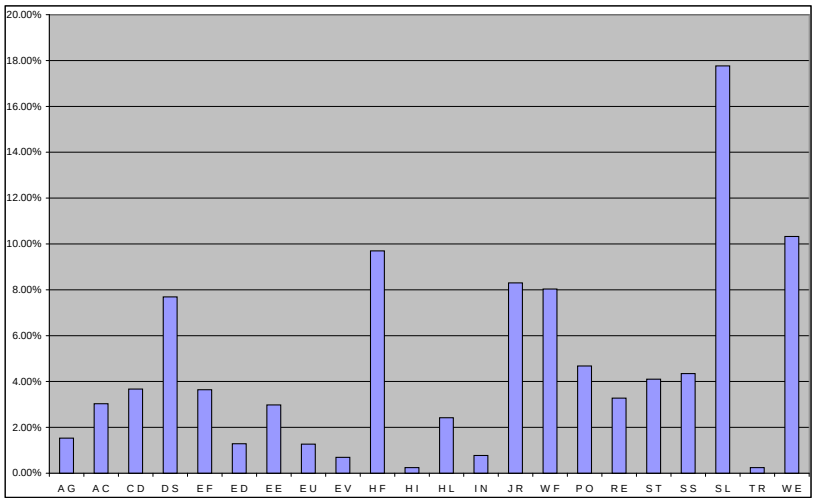

Figure 2: Topic representation in the training corpus

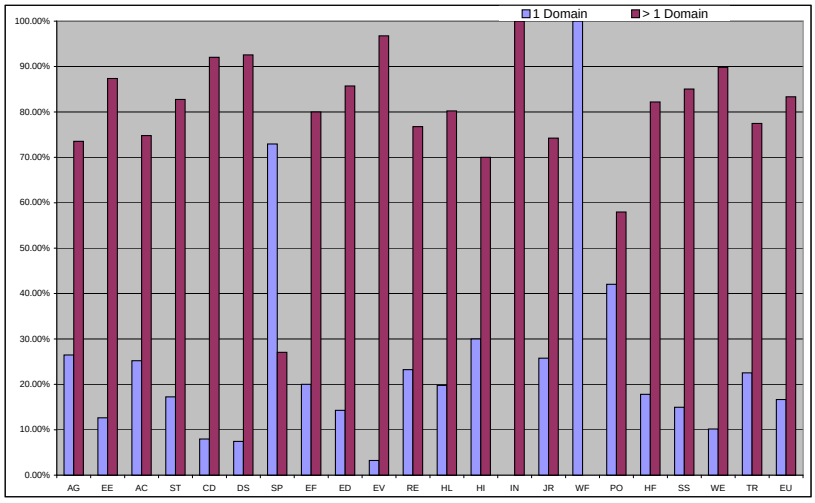

Figure 3: Percentage of segments with one and multiple topics, for each of the 22 domains.

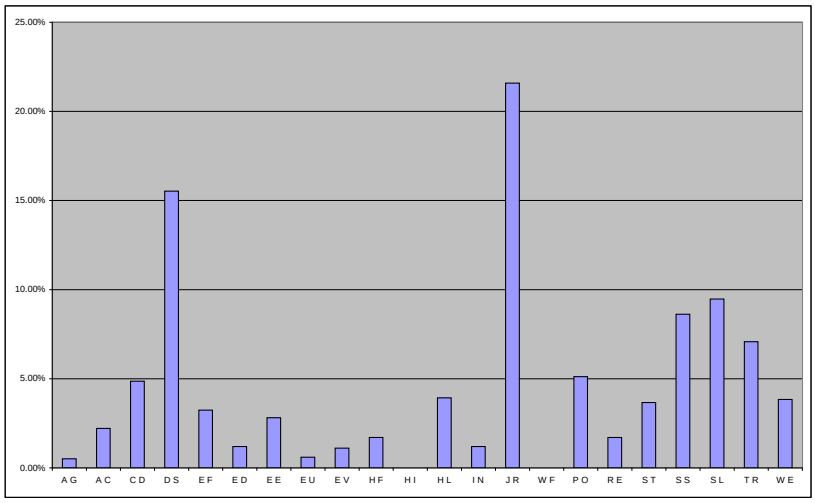

Figure 4: Topic representation in the development corpus.

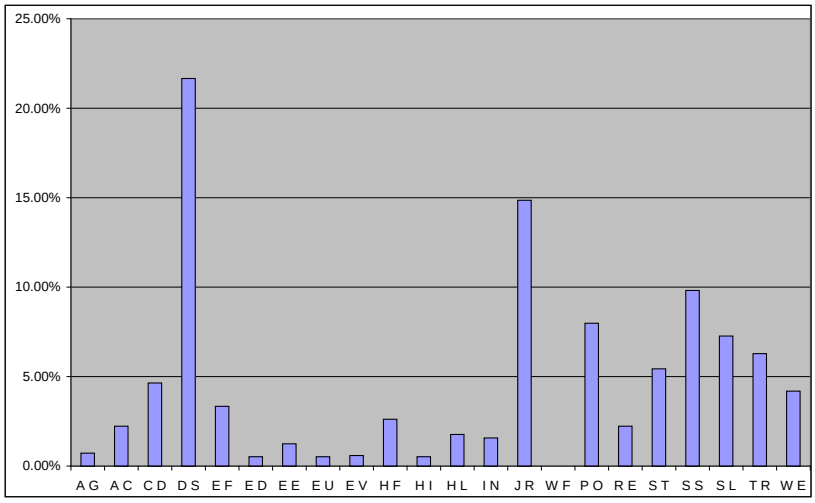

Figure 5: Topic representation in the evaluation corpus.

later transcribed by our automatic speech recognition (ASR) system. Each transcript segment contains as well some information related to the background acoustic environment, the speaker gender, and the speaker identification. All this metadata is also automatically extracted from the speech signal.

The speaker identification is of particular importance to the segmentation algorithm, namely, the classification as anchor or non-anchor. In fact, in our broadcast news programs, the anchors are responsible for introducing most stories. They are also the speakers whose id-numbers appear most often during the whole program, independent of the duration of each talk.

Our segmentation algorithm is based on a very simple heuristic derived from the above assumptions. It identifies the transcript segments belonging to the most frequent speaker-id number (the anchor), and defines potential story boundaries in every transition "non-anchor transcript segment/anchor transcript segment".

In the next step, we try to eliminate stories that are too short (containing less than 3 spoken transcript segments), because of the difficulty of assigning a topic with so little transcribed material. In fact, the classification of shorts stories is very "noisy". This type of situation occurs every time a sequence of transcript segments spoken by the anchor is interrupted by one or more transcript segments spoken by someone unknown. In these cases, the short story segment is merged with the following one.

The next stage, following this two-step algorithm, is indexation, as described in the next section. After this classification stage, a post-processing segmentation step may be performed, in order to merge all the adjacent segments classified with the same topic.

3.3 Story Indexation

Story indexation is performed in two steps. We start by detecting the most probable story topic, using the automatically transcribed text for each story. Our decoder is based on the HMM (Hidden Markov Model) methodology and the search for the best hypothesis is accomplished with the Viterbi algorithm. The topology used to model each of the 22 thematic domains is single-state HMMs with self-loops, transition probabilities, and bigram language models. For each of the 22 domains, a smoothed bigram model was built with an absolute discount strategy and a cutoff of 8, meaning that bigrams occurring 8 or fewer times are discarded. The referred models built from the training corpus, give the state observation probabilities. The statistics for each domain were computed from automatically transcribed stories with manually placed boundaries. The corresponding text was post-processed in order to remove all function words (527) and lemmatizing the remaining ones. Lemmatization was performed using a subset of the SMORPH dictionary with 97524 entries [11]. Smoothed bigram statistics were then extracted from this processed corpus using the CMU-Cambridge Statistical Language Modeling Toolkit v2 [4].

In the second step, we find for the detected domain all the second and third level descriptors that are relevant for the indexation of the story. To accomplish that, we count the

Judgment	Situation
Correct	There is a computed and a reference boundary inside the evaluation window.
Correct	Neither a computed nor a reference boundary is inside the evaluation window.
Miss	No computed boundary is inside the evaluation window that contains a reference boundary.
False Alarm	A computed boundary is inside the evaluation window that does not contain a reference boundary.

Table 2: Segmentation judgment for each window translation.

number of occurrences of the words corresponding to the domain tree leafs and normalize these values with the number of words in the story text. Once the tree leaf occurrences are counted, we go up the tree accumulating in each node all the normalized occurrences from the nodes below [10]. The decision of whether a node concept is relevant for the story is made only at the second and third upper node levels, by comparing the accumulated occurrences with a pre-defined threshold. The decision to restrict indexation to the second and third node levels was made taking into account the ALERT project goals and the data sparseness at the thesaurus lower levels.

3.4 Segmentation Results

For the evaluation of our simple segmentation algorithm, we adopted the metric used in the 2001 Topic Detection and Tracking (TDT 2001) benchmark NIST evaluation [22]. In this Evaluation Plan, the evaluation performance is defined in terms of probability of miss and false alarm errors (P_{Miss} and P_{FA}). A miss is considered when the algorithm fails to identify an existing boundary. A false alarm occurs when the algorithm outputs a non-existing boundary, according to the reference boundaries. To evaluate the outputted boundaries produced by the algorithm, an evaluation window of 50 words was used, and for each window translation, a judgment was done according to Table 2.

The cost segmentation function is defined as:

$$(C_{\text{Seg}})_{\text{Norm}} = C_{\text{Seg}} / \min(C_{\text{Miss}} \times P_{\text{Target}}, C_{\text{FA}} \times P_{\text{Non-Target}})$$

where

$$C_{\text{Seg}} = C_{\text{Miss}} \times P_{\text{Miss}} \times P_{\text{Target}} + C_{\text{FA}} \times P_{\text{FA}} \times P_{\text{Non-Target}}$$

and

C_{Miss} : cost of a miss

P_{Miss} : conditional probability of a miss

P_{Target} : a priori target probability

C_{FA} : cost of a false alarm

P_{FA} : conditional probability of a false alarm

$P_{\text{Non-Target}}$: a priori non-target probability $(1 - P_{\text{Target}})$

Using the values of C_{Miss} and C_{FA} adopted in TDT2001 [22] (1 and 0.3, respectively), we achieved a normalized value for the segmentation cost of 0.835 for a P_{Target} of 0.8.

The segmentation cost value did not reach 0.9, which was state-of-the-art in TDT2001 for this task. One potential reason for this low value is the post-processing step, in which adjacent story segments are merged if their topic classification is equal. Our next segmentation experiments were hence aimed at studying the influence of the merging criterion in this post-processing stage. In fact, we compared this post-processing stage that was based on merging adjacent stories with the same domain classification (*1st-level*), with 3 others: a post-processing stage based on merging adjacent stories with the same second level descriptors (*2nd-level*), a post-processing stage based on merging adjacent stories with the same third level descriptors (*3rd-level*), and no post-processing stage at all (Non). The results are shown in Figure 6.

We note that the correction and accuracy values are very close indicating that there is no real advantage in using any merging stage.

Even without this stage, however, the ”miss” rate is still very high, which motivated a closer look at the segmentation results. Several critical problems were detected: one of the reasons for boundary deletion is related to anchor detection in filler segments. Filler segments are very short segments spoken by the anchor and usually followed by a new story introduced by the anchor. In this scenario, and since all potential story boundaries are located in transitions ”non-anchor transcript segment/anchor transcript segment”, the boundary mark will be placed at the beginning of the filler region and no more boundary marks will be placed. To make the problem even more complex, filler segments are often partially corrupted by music, which makes them difficult to transcribe correctly.

Another reason for boundary deletion is the presence of multiple anchors in a broadcast news program. Some of the broadcast news programs in our corpus had in fact two anchors, one of which was responsible only for the sports stories. Our simple heuristic was based on defining a single anchor as the speaker that appeared more often, independently of the talk duration. Using this criterion, we got only the main anchor and not the sports anchor. The story boundaries introduced by the latter will all be missing. This obviously calls for a more refined anchor detection procedure.

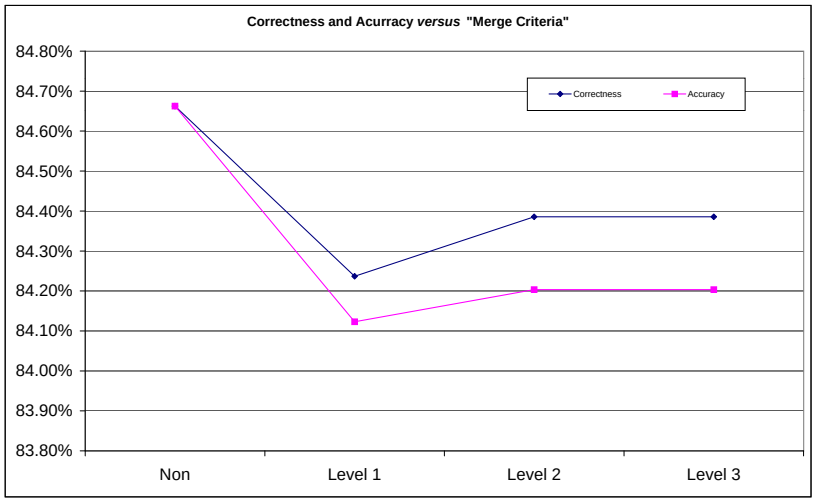

Figure 6: Experiment results using different merge criteria in the segmentation process.

3.5 Indexation Results

To measure the performance of the indexation algorithm, an experiment was done using the stories of the evaluation corpus and ignoring all the filler segments. In order to discard the influence of segmentation errors, this experiment was done using manually placed story boundaries and automatically transcribed texts.

In the evaluation of the indexation algorithm, we had to take into account the fact that there are stories that were manually indexed with more than one thematic domain (39% of the stories). We considered a hit every time the topic decoded is present in the topics manually identified in the story by the human annotators.

Our first set of experiments considered only the classification into the 22 hierarchical domains. The correctness achieved in the evaluation corpus using our bigram model was 73.80%. Figure 6 shows the confusion matrix that can be obtained using only the subset of the evaluation corpus corresponding to stories that were manually topic annotated with a single topic.

The rightmost column of the matrix indicates the number of stories accounted for. By observing this matrix, we see that the least confusable topic is ”weather forecast” which is never confused in a one-to-one classification. Some of the substitution errors are easily understood, given the topic proximity. Examples are: ”defense and security” which is confused in 22% of the cases with ”society and sociology” (32% of ”defense and security”

	AC	AG	CDI	DS	ED	EE	EF	EU	EV	HF	HL	HI	IN	JR	PO	RE	SL	SS	ST	TR	WE	WF	TOTAL
AC			25%								25%				25%			25%					4
AG																		100%					1
CDI			**33%**	67%																			3
DS				**70%**										3%	3%			22%		2%			63
ED					**25%**					25%					25%			25%					4
EE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
EF							**67%**							8%	17%						8%		12
EU	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
EV	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
HF										**100%**													1
HL																	100%						1
HI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
IN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
JR				8%										**65%**	15%			8%			4%		26
PO			3%	7%			10%			3%	3%				**69%**			3%					29
RE	100%																						1
SL			1%														**98%**	1%					93
SS			10%	10%										20%				**60%**					10
ST											14%							14%	**43%**	29%			7
TR																		14%		**71%**	14%		7
WE							33%							33%	33%								3
WF																						**100%**	27

Figure 7: Confusion matrix for a subset of the Evaluation Corpus.

stories were also classified as ”society and sociology” stories in the training corpus), and ”economy and finance” which is confused in 17% of the cases with ”political organization” (16% of ”economy and finance” stories were also classified as ” political organization ” stories in the training corpus).

This experiment was also repeated using unigram topic models, yielding a correctness value of 73.53%. The proximity of the results indicates that the amount of training data is not enough to build robust bigram models.

In terms of the second and third level descriptors, the results achieved a precision of 76.39% and 61.76%, respectively, but the accuracy is rather low given the high insertion rate (order of 20%). It is important to notice that this evaluation was performed only on those stories whose top level domain was correctly identified. Given the nature of the algorithm, the descriptor search is restricted to a specific domain identified in an earlier stage of the decoding process.

3.6 Summary

This report presented a topic segmentation and detection system for performing the indexation of broadcast new stories that have been automatically transcribed. Despite the limitations described in the report, the complete system is already in its test phase at RTP.

Our current work is aimed at improving the story segmentation method. We intend to

explore some information related to the speaker role inside the news programs. The anchors usually introduce stories and conduct the news program. Journalists usually develop the introduced stories where guests can appear in an interview sequence. We believe that the knowledge of the news program structure will enhance the precision of the segmentation task.

As future work in terms of indexation, we also intend to collect more data in order to build better bigram language models, because the ones used in this work were built using a high cutoff value. In addition, we also plan to allow the decoder to output all the domain scores associated with confidence values. This procedure will enable us to allocate more than one topic per story. This is indeed the situation of 39% of the Topic Detection Corpus.

4 Topic Detection for French (LIMSI)

In this section we describe the work carried out at LIMSI in developing topic detection and tracking methods. Due to the lack of ALERT-specific data in the early period of the project and in order to compare our methods to the work done by other sites, we have used labeled corpora in American English for part of our system development and evaluation. We participated in the TREC SDR'00 and TDT'01 and TDT'02 evaluations coordinated by NIST. The following subsections describe the ALERT topic detection corpus, the methods developed for SDR and TDT, and then apply these to the ALERT data.

4.1 Topic Detection Corpus

The topic detection corpus was provided by Secodip and contains recordings of radio transmissions from the station "France-Inter" dating from April 16-21, 2002. The recordings were made continuously for all the broadcast time and cut into hour-long segments. The content varies from hour to hour but in general contains a variety of types: news, interviews, weather reports, music, interviews, special reports, etc. The 76 hours of data were automatically processed by the French transcription system. There are no manually annotated story boundaries.

The date, time and channel are specified in the file names which are of the form:

20020416_0000_0100_finter
20020416_0100_0200_finter
. . .
20020421_0100_0200_finter
20020421_0200_0250_finter

Secodip also provided the list of "badges" which specify the topics of interest for their clients. For the period of interest, Secodip also provided the alerts (produced using their standard processing methods) sent to the various clients.

4.1.1 Client Badges

For each client, there is an associated badge which specifies the list of topics (and any particular restrictions or explanations) which the client is interested in. Each badge consists of the client_name (client), title, and a list of keywords (topics) which define the topic along with any comments. Two example badges are shown in Figure 8.

The corpus contains badges for each of 22 clients, shown in Table 3. The names of non-public institutions have been replaced with generic identifiers to respect the client's privacy. The first column shows the number of topics for each client, which is seen to range from 1 to 13. In total there are 96 topics.

#Topics	*Client*
4	ADP - Aeroport de Paris
3	Regional agency
5	European Community
1	Regional agency2
2	Political party
2	CSA - Conseil Superieur de l'Audiovisuel
6	Governmental agency
5	European Company
1	News Magazine
5	FFF Federation Football Francaise
3	TV Station
5	Telephone company
5	Company
1	Paris newspaper
5	Ministry of Agriculture
7	Ministry of Culture
13	Minisity of Employment
5	Ministry of Public Services
4	Minister
10	Company2
2	Political party2
2	International Company

Table 3: List of clients and the number of topics for each.

* **Client:** AEROPORTS DE PARIS
* **Topcis:**
ADP (tous sujets sur l'Aéroport De Paris)
AEROPORTS (exclusivement les aéroports parisiens)
BRUIT (exclusivement les problèmes de nuisance sonore autour des aéroports parisiens)
ROISSY (exclusivement l'aéroport de Roissy Charles de Gaulle)

* **Client:** GOVERNMENT AGENCY
* **Topics:**
ARRONDISSEMENT (exclusivement les arrondissements parisiens)
LA TOUR EIFFEL(tous sujets sur la Tour Eiffel)
MAIRIE DE PARIS (tous sujets sur la Mairie de Paris)
PARIS (tous sujets sur Paris)

Figure 8: Two example badges.

Some clients have a wide range of topics of interest which tends to result in a large number of alerts. For example, the European Company in Table 3 has 5 wide ranging topics such as *digital TV, medical radiology, pain, ...*; which are quite different in nature. Other clients have very broad topics - a Government Agency that wants to know all about *Paris*, or the ADP that wants to be informed about any stories concerning any of the Parisian airports.

4.1.2 Manual alerts

The manual alerts for each client were provided by Secodip. The alerts did not specify the particular topic which triggered the alert. For some clients there are only a few topics of interest, but for others the set of topics is relatively large and can be diverse. For each alert the following information is available:

Date
Start time of alert
Duration of the segment of interest
End time of alert
Title of the show
Name of speaker or host
Resume of the segment

There were a total of 330 manual alerts during the period, with a total duration of 53,888 seconds (about 900 minutes) of corresponding audio data. The number of alerts is shown

#Alerts	*Client*
2	ADP - Aeroport de Paris
0	Regional agency
21	European Community
8	Regional agency2
11	Political party
16	CSA - Conseil Superieur de l'Audiovisuel
19	Government agency
1	European Company
11	News Magazine
16	FFF Federation Football Francaise
14	TV Station
1	Telephone company
20	Company
12	Paris newspaper
9	Ministry of Agriculture
21	Ministry of Culture
29	Ministy of Employment
10	Ministry of Public Services
11	Minister
15	Company2
18	Political party2
65	International Company

Table 4: Number of manual alerts for each client occurring in the evaluation corpus.

17/04/02
Heure début 07:49:14
Durée 00:01:55
Heure fin 07:51:09
EST&OUEST / NORD&SUD
STEPHANE PAOLI

LA POLEMIQUE SE POURSUIT AUTOUR DE L'IMPLANTATION DU TROISIEME AEROPORT PARISIEN EN AUSTRALIE ; CAR SUR LE SITE SE TROUVERAIENT DES CIMETIERES MILITAIRES DE LA PREMIERE GUERRE MONDIALE .. LE VICE PREMIER MINISTRE AUSTRALIEN DE VISITE A PARIS IL SERA RECU PAR LE SECRETAIRE D'ETAT FRANCAIS AUX ANCIENS COMBATTANTS JACQUES FLOCH .. 07:49:34 COMMENTAIRE DE STEFANYS COCK .." UNE PETITION NATIONALE AUSTRALIENNE A ETE DEPOSEE A L'AMBASSADE DE FRANCE A CAMBERA .." .. (DEFENSE TRANSPORTS)

Figure 9: Example alert for the client ADP.

for each client in Table 4. Figure 9 shows an example alert for the client ADP (Aeroport de Paris).

The timestamps in the manual alerts are approximative and offsets ranging from 0 to 20 seconds were observed with respect to the true time in the audio signal. These offsets, which are particularly bad for short alerts that are only a few seconds long, were readjusted.

4.2 Spoken Document Retrieval (TREC-SDR)

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 audiovisual 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 words 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 a single word. Stemming is used to reduce the number of lexical items for a given word sense [24].

In the LIMSI SDR system, a unigram model is estimated for each topic or query. The

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

Table 5: Impact of the word error rate on the mean average precision using using a 1-gram document model. The document collection contains 557 hours of broadcast news from the period of February through June 1998. (21750 stories, 50 queries with the associated relevance judgments.)

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 [13, 21, 23, 28] 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 [29]) based on terms present in retrieved contemporary texts is used. This is particularly important for 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 [7]. 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 10xRT system and a single pass 1.4xRT system [7]. The information retrieval results with and without query expansion are given in Table 5 in terms of mean average precision (MAP), as is 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 10xRT transcriptions, and a moderate degradation (4% absolute) is observed using the 1.4xRT transcriptions. For word error rates in the range of 20%, only small differences in the mean average precision were found with manually and automatically produced transcripts, when using query expansion on contemporaneous data.

4.3 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 [6]. The broadcast news transcription system also provides non-lexical 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 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 [9], 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 multi-scale segmentation of the audio stream into documents.

One 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 [1, 16]. 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 [9]. 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 30s 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 online topic tracking applications.

The mean average precision (MAP) using a single 30s window and the double win-

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

Table 6: Mean average precision with manual and automatically determined story boundaries. The document collection contains 557 hours of broadcast news from the period of February through June 1998. (21750 stories, 50 queries with the associated relevance judgments.)

dowing strategy are shown in Table 6. 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 of using a search engine specifically designed to retrieve stories in the audio stream. Using an a priori acoustic segmentation, the mean average precision is significantly reduced compared to a "perfect" manual segmentation, whereas the window-based search engine results are much closer. Note that in the manual segmentation all non-story segments such as advertising have been removed. This reduces the risk of having out-of-topic hits and explains part of the difference between this condition and the other conditions.

4.4 Topic Tracking (TDT)

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 is 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 to the ALERT media-watch application as well as to the problem of structuring multimedia digital libraries.

A version of the LIMSI topic tracking system was assessed in the NIST Topic Detection and Tracking (TDT'01,TDT'02) evaluations on the topic tracking task [17, 18]. For this task, a small set of on-topic news stories (one to four) are 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 the 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 [9] relying on external information sources like past news were used, in conjunction with unsupervised online adaptation techniques to update the on-topic model with information obtained from the test data itself. Online 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% [17, 18]. In order to deal with recognition errors in the automatic transcripts the use of word-level confidence measures and word lattices was investigated. However, these were not found to improve the topic detection results.

A topic identification system has also been developed in conjunction with the LIMSI SDR system. This system 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 10 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.

4.5 Topic Detection for Alert

The methods developed and tested for the SDR [9] and TDT [17, 18] evaluations were compared and combined for use in ALERT. In this section we describe the main specificities of detecting topics in French using the ALERT topic detection corpus.

The TDT method uses a unigram topic tracker as described in [17, 18]. The similarity score $S(d,T)$ for the incoming document d and the topic T is the normalized log-likelihood ratio between the topic model and the general background French language model:

$$S(d,T) = \frac{1}{L_d} \sum_{w \in d} tf(w,d) \log \frac{\lambda P(w|T) + (1-\lambda) P(w)}{P(w)}$$

where $P(w|T)$ is the ML estimate of the probability of word w given the topic model, $P(w)$ is the general French background language model probability of w, $tf(w,d)$ is the term frequency in the incoming story d, and L_d is the story length. If the similarity score

Figure 10: Example screen of the LIMSI spoken document retrieval 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.

	Background Model_1		Background Model_2	
Lexicon	59012 words		34866 words	
Stoplist	139 words		228 words	
Training texts	*period*	*size*	*period*	*size*
Le Monde	jan01-sep01	1984862 words	may01-sep01	877976 words
F2	aug01-jul02	592736 words	aug01-jul02	592736 words
TF1	aug01-jul02	665660 words	aug01-jul02	665660 words

Table 7: Background models used for TDT based system.

of a segment $S(d,T)$ is higher than the threshold th_A, the segment is ontopic. Since we do not have development data, results are reported for different thresholds.

Since there are no story boundaries available, the overlapping sliding window approach described above was used to determine the extent of the stories. For the SDR approach the double-window method was used and for the TDT approach, a fixed window of 50 words with an offset of 25 words was used.

4.5.1 Background model and Lexicon

The background (or general French) model is trained on texts from the French newspaper *Le Monde* and automatic transcriptions of broadcast news data from the *F2, TF1* television shows. Two background models were estimated using different size lexicons and stoplists, and different amounts of newspaper texts for training (see Table 7). The texts were processed according to our standard language modeling normalization, followed by stopping, stemming, and compounding. The lexicon is the word list of the background model after normalization, stopping, stemming, and compounding.

4.5.2 Query Expansion

The available information (title and keywords) can be used directly as a query or can be expanded. Query expansion is carried out using a parallel text corpus from the French newspaper *Le Monde*, dating from April 2001 to March 2002. (Although the newspaper texts are purchased on CDROM, the most recent texts were taken from the Internet since the CDs are produced quarterly.) Applying the query expansion technique to the terms in the badge results in a larger set of terms related to the topic of interest. The result of query expansion applied to the title and keywords was manually verified, and unrelated terms were removed.

In total, three text sources are used for querying the system: the topic titles; the relevant terms in the topic definition (obtained after stopping and stemming); and the terms found by query expansion.

System	*Training data*	*Threshold*	*#detected*	*#correct*	*%detected*
SDR	-	-	3332	277	84
Background Model_1					
System	*Training data*	*Threshold*	*#detected*	*#correct*	*%detected*
TDT	title & qexp	0.0	2933	211	64
TDT	title & keywords & qexp	0.0	3640	249	75
		0.1	2921	214	65
SDR+TDT	title & keywords & qexp	0.0	5273	293	89
		0.1	4783	286	88
Background Model_2					
system	*training*	*threshold*	*#detected*	*#correct*	*%detected*
TDT	title & keywords & qexp	0.1	4807	270	82
		0.15	3421	247	75
		0.2	2566	221	67
SDR+TDT	title & keywords & qexp	0.1	6272	306	93

Table 8: Topic detection results.

4.5.3 Training the on-topic models

The texts available for training the topic models are the title, the keywords (as described in Section 4.1) and the expanded query qexp (as described in the previous section). The number of words for training is very small, ranging from 1 to 24. For example, the topic for one of the clients contains only the single word *Paris*. This is an extreme case. On average, there are 10 words for training a topic. The ontopic model is interpolated with the background model with interpolation coefficient(0.25).

4.5.4 Experimental Results

The SDR-based system and the TDT tracker were tested individually and in combination. In both cases query expansion with manual correction was used. Detection results are given for the two background general French models, with a few different thresholds for two training conditions. In the first training condition the training data consist of the title and qexp, whereas in the second condition the keywords were also used.

The top part of Table 8 gives the results of the SDR, TDT and combined SDR-TDT systems for Background Model_1. The SDR system has a detection rate of 84%, correctly detecting 277 of the 330 alerts. With the Background Model_1, using all available texts (title & keywords & qexp) outperforms using just the title & qexp.The tracking performance is seen to depend on the decision threshold. Combining the SDR and TDT based

systems gives the highest detection rates.

The lower part of the table gives results using Background Model_2, trained on fewer newspaper texts and having a smaller lexicon size. The stoplist is also larger than that used for Background Model_1. Both TDT alone and SDR+TDT have better performance with the second background language model. The best detection rate for TDT alone is 82%, using a threshold of 0.1. With this same threshold, combining SDR with TDT gives a correct detection rate of 93% (306 of 330 alerts being correction detected. This high correct detection rate also results in a large total number of detections (6272).

4.5.5 Improving the performance

A frame size of 50 words (or approximately 15s) is used for the TDT tracker. The large number of detected alerts, and their locations, suggests that multiple alerts are being detected for the same segment. Therefore we decided to investigate merging adjacent segments that are within the same 1 minute interval if they have the same detected topic. This reduces the total number of alerts, but can increase the duration covered by the alerts.

Online unsupervised adaptation was found to be successful for the American English TDT task [17, 18]. The topic model is adapted by adding incoming stories identified as on-topic by the system to the training data, as long as the stories have a similarity score $S(d, T)$ that is higher than an adaptation threshold th_A, where $th_A \geq th_D$. The topic model term frequencies are updated by adding the story term frequencies of the incoming story weighted with an adaptation weight (fixed weight). Since no data were available to tune the adaptation parameters, the threshold with the best detection rate was chosen. Online adaptation is carried out by adding a maximum 5 stories (we found that adding more stories increased the number of false alarms). The following parameters were used for adaptation:

Frame size: 50 words, offset: 25 words
Decision threshold th_D= 0.1
Adaptation threshold th_A = 0.2 (0.3)
Adaptation weight adp_w = 0.05
Maximum number of stories for adaptation: 5

Table 10 compares the results with merging of adjacent segments for different systems: SDR, SDR+TDT, and SDR+TDT with online unsupervised adaptation. Merging can be seen to decrease the total number of detected alerts, but increases the duration of the detected alerts. The total time of the detected alerts is an important measure since it gives an idea of the percentage of the corpus the client will need to look at. In the case of manual verification of the automatically detected alerts, it gives an idea of the verification time needed. Merging segments within 60s is quite reasonable, if two ontopic segments occur within 60s, they are properly from same story. As the merging window is widened (120s to 300s) it becomes less likely that the detected topics are really from the same story.

	TDT		*SDR*		*SDR+TDT*		
Classes	*detected*	*corr.*	*detected*	*corr.*	*detected*	*corr.*	*manual*
ADP	61	2	25	2	73	2	2
Regional agency	47	0	78	0	92	0	0
European Community	113	10	195	15	237	17	19
Regional agency2	64	3	91	7	135	7	8
Political party	236	9	25	7	243	10	11
CSA	145	16	116	16	188	16	16
Government agency	326	15	641	19	730	19	19
European Company	212	1	181	1	326	1	1
News Magazine	42	5	26	8	54	8	11
FFF	158	15	117	15	190	15	16
TV Station	298	8	468	14	608	14	14
Telephone company	120	1	251	1	298	1	1
Company	108	15	85	19	140	19	20
Paris newspaper	81	7	247	10	292	10	12
Ministry of Agriculture	239	7	62	7	238	8	9
Ministry of Culture	496	21	151	16	555	21	21
Ministy of Employment	642	23	164	21	638	25	29
Ministry of Public Services	374	9	93	8	360	10	10
Minister	220	11	62	9	222	11	11
Company2	346	13	96	11	376	13	15
Political party2	115	18	44	16	114	18	18
International Company	186	61	114	55	163	61	65
Total	4629	270	3332	277	6272	306	330

Table 9: Number of automatically detected alerts per client for the three approaches: TDT, SDR and SDR+TDT. The total detected time is 274581s.

System	*merging time(s)*	*correct*	*#detected*	*detected time*
SDR	0 (no-merge)	277	3332	199845
	60	277	2990	209940
	120	278	2772	231255
	180	278	2512	260835
	300	278	2213	330150
SDR+TDT	0 (no-merge)	306	6272	274581
	60	306	5744	281549
	120	306	4872	350490
	180	308	4478	410172
	300	308	4031	509182
SDR+TDT with	0 (no-merge)	311	8315	342376
adaptation	60	310	5970	375698
	120	311	5355	500845
	180	312	4869	627725
	300	313	4319	658750

Table 10: Results with merging of stories for SDR, SDR+TDT and SDR+TDT with online unsupervised adaptation.

With online adaptation, the number of correct detections increases from 306 to 311, but the total number of detections is quite high: 8315 without merging and 5970 with 60s merging. Adaptation can add noise to the ontopic model especially the topic is wide.

4.6 Keeping the Vocabulary Up-to-Date

Most broadcast news speech recognizers use static LMs trained on very large text corpora (hundreds of million words) from a variety of sources and spanning several years. There is often a substantial gap (several months or longer) between the epoch of the LM training corpus and the audio data to process. This is due to difficulties in obtaining and processing training texts. In contrast, the news domain is characterized by sudden changes in topics, with resulting changes in vocabulary. The longer the gap between the LM training data epoch and the processed data, the higher the expected proportion of out-of-vocabulary (OOV) words. These words are unknown to the system since they did not appear often enough or even at all in the training corpus.

In order to reduce the LM aging problem, HTML files are collected daily from 5 French websites. These texts are cleaned and normalized using appropriate scripts adapted to each website in order to extract the useful information (date, title and text) and convert the texts to the form used for language model training.Four of the data sources are used for training,

Source	Mean	Minimum	Maximum
Site1	37.2k	11.7k	68.9k
Site2	1.2k	0.0k	4.6k
Site3	51.3k	1.3k	95.8k
Site4	34.9k	0.0k	59.1k
Dev	5.6k	0.2k	10.9k

Table 11: Mean, minimum and maximum of number of words per day and per source for the months of December 2001 and January 2002.

and the fifth is used to test the adapted models. Table 11 gives the mean, minimum and maximum of number of words per day and per source for the months of December 2001 and January 2002.

In order to adapt the recognition vocabulary, each day new candidate lexical entries are identified by analyzing the frequency of out-of-vocabulary words in the adaptation corpus. Words that occur at least twice in the given day and at least 5 times during the preceding four weeks are retained. In order to keep a fixed vocabulary size (65k words) the least likely words in the original vocabulary are removed. This approach allows us to reduce the out-of-vocabulary rate by 32% and the language model perplexity by 13%. It should be noted that proper names account for a large portion of the new entries [2, 3].

4.7 Error analysis

With the combined SDR+TDT systems, 24 of the manual alerts were missed. Eight of the alerts were missed because of recognition errors that were unable to be recovered. An additional 8 errors are at the semantic level. The scores were too low for four of the alerts, so they were eliminated. Two of the missed alerts can be attributed to incomplete user profiles, and two due to the normalization procedures.

Figure 11 shows the percentage of correct alerts, the percent coverage and the precision as a function of the number of retained alerts. The coverage is seen to increase rapidly and then to taper off, however still increases, indicating that some relevant alerts appear quite far down on the list and must have weak scores.

4.8 Discussion

This study has highlighted several criteria which can be related to difficulties in automatically detecting alerts.

- Wording not in the recognition lexicon are often unrecoverable. Missing words are

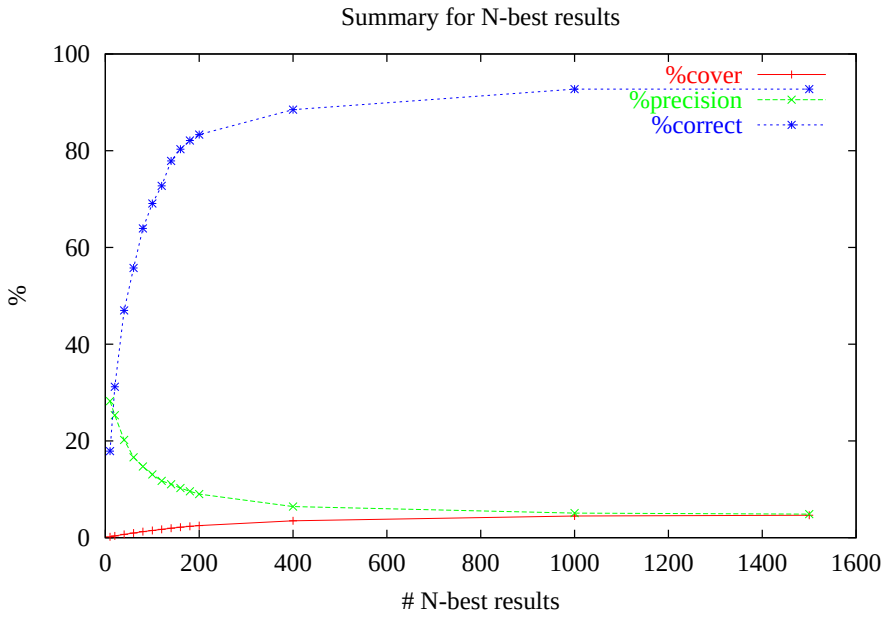

Figure 11: Coverage, false alerts and percent correct as a function of the number of alerts retained.

often proper names or names of organizations, products, etc. It is quite difficult to detect alerts if these are not present.

- Specific terms should be used for searching. The more specific the terms used for searching are, the fewer the number of false alerts will be. If the subject is vague or is defined exclusively using frequent words of the language, it requires interpretation and there are likely to be false detections.
- The frequency of the relevant terms in the language should be taken into account. Rare words may be more closely related to a topic, but are not often observed.
- The additional information in the comments of the badges is important as it links the terms to other related terms. Other considerations are the importance of the order of terms and whether or not they are required or just pertinent. One way of dealing with such constraints would be to include boolean operators.

The combined use of the SDR and TDT search engines enables 93% of the manual alerts in French news to be automatically detected. Although the number of false alerts is quite high, if the automatic system is seen as aid to prefilter the audio data, it allows most of the audio documents to be ignored, greatly reducing the listening time.

The definition of a badge and the current state of events are strongly linked, so if a user is able to modify the profile in an interactive manner based on the alerts detected by the system, the precision of the system can greatly improved. Although this method requires human intervention, the profile can be expected to remain valid for a certain period of time. However since the nature of news is that it is constantly changing, the ability to revise profiles is crucial for an service. Improving the underlying statistical models (acoustic and language), and methods for adaptation should also improve the detection precision.

5 Topic Detection for German (UniDu)

The aim of the topic detection module is to classify text automatically into pre-defined categories. Automatically transcribed speech data is considered as the major text source within this project. However, the processing of further text sources, like electronic news texts or manually generated summaries is of economic interest for the users as well. Accordingly, the topic detection module should be capable to cope with such kind of text sources too.

In contrast to electronic texts, the speech recognizer output does not provide any clues regarding the structure of the text. In this case, a document appears as a plain unstructured text stream. In addition to the actual classification problem, the system has to find boundaries between topically homogeneous segments. Another challenge results from the specific requirements of the user partners. We know from further investigations within the project, that search profiles, which were requested and ordered by the end users, appear quite sparsely in common news shows. From this requirement, a third key feature can be derived, which is the rejection of off-topic segments.

The following sections present the developed methods for topic detection and the achieved results. The terms used within this report correspond to the definitions in D6.1 ”Plans for multilingual evaluation and demonstration”.

5.1 The German topic detection module

We have implemented and investigated two approaches to topic detection that are presented in the following sections. Part of the results presented here were published at [15], [30] and [14]. Note that the German topic detection module does not perform any topic segmentation and therefore assumes that the story boundaries it has got from the topic segmentation module separate the text correctly into homogeneous topics.

5.2 Standard Approach

The standard approach works on a word feature level. A unique index number is assigned to every word in the vocabulary list of the ASR module. For feature extraction, each word in the text corpus (may be the output of the speech recognizer module or, for training, manually created summaries) is represented by its index number. We model each topic with a discrete single state HMM and use the word index numbers as observations.

As text-pre-processing, the words are reduced to their stem, and words are removed that appear in a manually created stop word list with 150 entries. The vocabulary list has been limited to the 32k most frequent words.

5.2.1 Novel approach

Overview Our novel approach works on a character basis. The system scans the text with a sliding window and extracts binary feature vectors from the windows. They are quantized and then used to model each topic with a discrete Hidden Markov Model (HMM). Thus, in contrast to standard approaches, where discrete HMMs emit VQ labels representing entire words, our approach uses VQ labels to represent feature vectors generated from character sequences and is therefore independent of a word lexicon.

Feature extraction In a pre-processing step, numbers and punctuation marks are removed and all characters are converted to lower case. Then, a sliding window W of size w (typically $w = 3$) characters scans the text. From each character *C* in the window, we extract a 32-dimensional binary feature vector $\vec{O}_C$. Exactly one component of $\vec{O}_C$ gets a value of $+1$, the others are assigned a value of 0 (in unipolar case) or -1 (in bipolar case). The vector representing an a "+1" value at its first component, a b gets a "+1" at the second component, and so on. The feature vector of each text window $\vec{O}_W$, thus has a size of $w*32$ with w components being "+1". We call these vectors *frames*. The resulting concatenated feature vector is made up of f adjacent overlapping frames: $\vec{O}_R = [\vec{O}_{W_1} \cdots \vec{O}_{W_f}]$. This not only considers the context of a window, but also allows us to duplicate characters that are in the center of the window. Several experiments using different numbers of frames have shown that 3 frames show the best recognition result. For numerical reasons, it turned out that it is better to add noise to the vectors in order to avoid singularities in the feature distribution. Noisy features improved the recognition result from 48.4% to 50.3%.

In German, words tend to change their stem vowel or umlaut when changing their grammatical function; this change may not be detected by the speech recognizer. The idea behind using these big feature vectors is that if characters are wrongly recognized by the speech recognition module, the distance between the vector of the correct spelling and the vector with one wrong character is the same whatever the wrong character may be. Scanning with a window, provided its size is properly chosen, emphasizes the morphemes of the text, and thus the semantic information carriers. Our experiments show best results when five characters are covered, which may well be the average size of German morphemes.

Vector quantization To reduce the dimensionality (eg. 288 components) of the vectors $\vec{O}_R$, they are quantised using B prototype vectors. Each of the prototype vectors $\vec{p}_b$ gets an index number I. Every vector $\vec{O}_R$ is represented by the number of its nearest prototype vector. We have investigated two methods for creating the prototypes. The first one clusters the feature vectors of the training set using the k-means clustering algorithm. The second one trains a single-layer neural net (no hidden layers) to produce optimal prototypes. Its optimization criterion is to maximize the mutual information (MI) $I(P, T)$ between the prototype vectors $\vec{p}$ and the topics t. The vector quantization of the feature vectors $\vec{O}_R$ results in a loss of information. If we use prototypes with a high mutual information between

$\vec{p}$ and t, this loss is reduced.

Topic Modeling Every topic is modeled with a discrete single-state HMM by using the indices of the prototype vectors as observations.

Detecting topics of no interest Our first experiments were made with no out-of-topic model, because all tests topics were restricted to the trained topics. Then we tested an out-of-topic-model that was created using all training texts. Finally, a confidence measure was applied to identify topics that are of no interest to the user. For every story, we computed a confidence measure C as presented in [12] :

$$C = \frac{S_1 - \frac{1}{N-2}\sum_{i=2}^{N-1} S_i}{S_1 - S_N} \quad (1)$$

where N is the number of N-best results of topic recognition for the story, and S_i is the log-likelihood of the i-th best result. Comparing the ROC-curves using (1) and using an out-of-topic model, the superiority of the confidence measure was shown.

Experiments and Results on summaries test sets In all cases, the training material consisted of topic-assigned summaries manually created by Observer. They summarize reports of radio and TV programmes that are of interest to the customers of Observer. As Hidden Markov Models need a lot of training material, different topics labels that belong to the same company (i.e. "Daimler" and "Daimler Unternehmen") were combined.

The test sets were both summaries and output of the speech recogniser module.

For our experiments we have defined the following test sets:

- **A**: no stop word removal, no stemming, no text optimization.
  - **A1**: 22 topics in 3037 stories for training and 1319 stories for testing
  - **A2**: 173 topics in 6039 stories for training and 2700 stories for testing
- **B**: deletion of 150 stop words, text optimization, 22 topics in 2956 stories for training and 1284 stories for testing.
- **C**: Training on summaries of 898 topics, testing on 48 manually segmented, automatically transcribed radio news episodes.

In the summaries of test set **A**, some words are separated into two single words because they were entered in two different lines. Besides, there are some special abbreviations. Thus, the text basis is not optimal, but we decided not to change the text in order to simulate in a more or less rough way errors which are made by automatic speech recognition. For test set **B**, which is made up of different texts, we did compensate for these errors (called

text optimization in the listing above). To get the correct parameter configuration of the topic detection module, many experiments were conducted with test sets **A1**, **A2** and **B**, some of which are presented here.

Duplicating Characters When using feature vectors that are made up of more than one frame $\vec{O}_W$, the characters in the center of the window are duplicated. Consider for example a window size of w=3, a frame number of f=3 and a text window abcde. The feature vector then represents the characters abcbcdcde, or once a and e, twice b and d and three times c. In the table, the results for f=3 (duplicated characters) and f=1 (single characters) are compared. In both cases, the same 5 characters are covered, but the repetition of the center characters leads to an improved recognition result.

	f=3, w=3	f=5, w=1	Set
1-best	48.4%	45.5%	A1
2-best	67.6%	64.6%	A1
1-best	32.1%	30.4%	A2
2-best	46.8%	45.1%	A2

Varying the number of iterations It would be ideal to improve the quantization prototypes with every iteration of the k-means algorithm and of the MMI net training. We have investigated the performance of both quantizers for different numbers of iteration steps. The recognition result of the k-means system increases until 20 iterations and then remains constant. The MMI system tends to get slightly worse (from 47.8% to 45.9%) with a growing number of iterations, except for a peak at 20 iterations. Interestingly, this is contrary to the increase of the mutual information $I(P,T)$. However, $I(P,T)$ does not always grow monotonically, which indicates that the global optimum won't necessarily be reached by the training algorithm applied to the neural net. Figure 12 shows how $I(P,T)$ grows with the number of iterations of the neural net training.

5.3 Using more prototype vectors

As the prototype vectors have to represent the information that is in the texts in the best possible way, the right choice of the number of prototypes is important. We have made experiments with several numbers of prototypes on test set **A1** whose results are listed in Table 12.

The k-means system shows its best recognition results for 500 prototypes and decreases significantly with more prototypes. The MMI system's peak is at 1000 prototypes. This indicates that the number of important lexical morphemes in our training set is somewhere in the range between 500 and 1000. The decrease in performance with a high number of prototypes might be due to over-fitting to the training set and thus losing the ability to generalize.

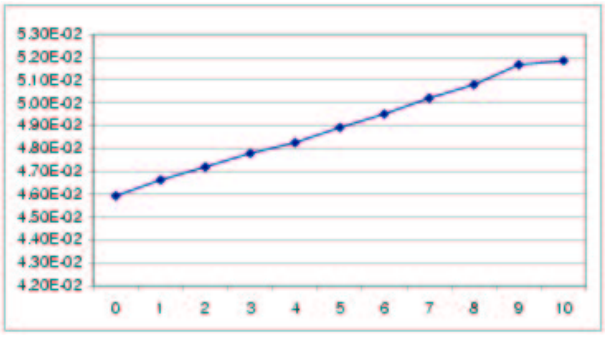

Figure 12: The MMI $I(P,T)$ in bit as a function of the number of iterations

K-means	*MMI*	*Number of prototypes*
48.4%	47.9%	200
50.1%	49.7%	500
46.9%	50.4%	1000
46.7%	47.3%	2000

Table 12: The influence of the number of quantization prototypes. w=f=3,5 HMM states, set A1

Other experiments conducted were the variation of number of HMM states, using unipolar and bipolar feature vectors, and deletion of keywords (i.e. the words that appear in the topic name) for the k-means and MMI approach.

Results on non-optimal summaries and on ASR transcription; Comparison to Standard approach Several tests were performed with non-optimal texts. Besides using the test sets **A1** and **A2**, which are already not optimal, we removed all spaces from the test set **A1**. This is to simulate wrong speech recognizer output, i.e. combining separated words to compound words (recall that the German language makes extensive use of compound words). The standard system will fail on this test set, because it works on a word basis, while the new approach decreases only slightly.

Test set **C** consists of 48 reports from radio news that were transcribed by means of our ASR module. 898 topics were used to train the system for this test set. The recognition results using both the new system and the standard system are listed in Table 13. The results for the new system are significantly lower than those of the standard system. Thus, for a high number of topics, the standard system is still better, though it has not been designed with respect to erroneous texts of a speech recognizer.

New system	*Standard system*	*Sets*
50.3%	50.4%	A1
48.0%	-	A1, no spaces
32.1%	35.3%	A2
66.6%	78.0%	B
22.9%	35.4%	C

Table 13: Comparison of the best recognition rates of the new approach to the standard approach

The results on the output of the ASR module for the development and evaluation set are given in the Evaluation Report.

It must be emphasized that the training and the recognition operate on two different text corpora: the topic models are trained on manually created summaries, while the recognition is made on erroneous output of the text recognizer module. It was necessary to use the summaries because statistical approaches (like Hidden- Markov-Models) must be trained with a large amount of data, and this is the only way to get the necessary amount of data.

6 Conclusions

This deliverable has summarized the research carried out in workpackage 4 of the ALERT project. The main focus of this work has been to investigate techniques to automatically detect topics in audiovisual data from the audio signal. As such, the developed algorithms have had to deal with specificities of the audio data, such as locating the speech portions, segmenting the continuous stream into stories and identifying relevant topics. Different approaches have been explored, from using user-specified keywords to implicit topic specification using samples of on-topic stories. Despite word error rates over 20%, topics can be detected in the recognizer outputs with performances close to those that can be obtained using manually produced transcripts. This is in part due to redundancy in the story (topic related words are often repeated or may have synonyms) and the use of information retrieval and normalization techniques such as stopping, stemming and query expansion which reduce the effects of recognition errors.

7 References

[1] D. Abberley, S. Renals, D. Ellis, T. Robinson, "The THISL SDR System at TREC-8," *Proc. of the 8th Text Retrieval Conference TREC-8*, pp. 699-706, Gaithersburg, November 1999.

[2] A. Allauzen, J.L. Gauvain "Mise à jour automatique du modèle de langage d'un système de transcription," *JEP'02*, Nancy, June 2002.

[3] A. Allauzen, J.L. Gauvain "Adaptation automatique du vocabulaire et du modèle de langage d'un système de transcription," submitted to *TAL*.

[4] P. Clarkson and R. Rosenfeld, "Statistical Language Modeling using the CMU-Cambridge Toolkit", in *Proc. EUROSPEECH'97*, Rhodes, Greece, 1997.

[5] J. Fiscus, G. Doddington, J. Garofolo and A. Martin, "NIST'S 1998 Topic Detection and Tracking Evaluation (TDT2)", in *Proc. DARPA Broadcast News Workshop*, Feb. 1999.

[6] M. Franz, J.S. McCarley, T. Ward, W.J. Zhu, "Segmentation and detection at IBM: Hybrid statistical models and two-tiered clustering," *TDT 1999 workshop notebook*, 1999.

[7] J.S. Garofolo, C.G.P. Auzanne, E.M. Voorhees, "The TREC spoken document retrieval track: A success story," *Proc. of the 6th RIAO Conference*, Paris, April 2000. Also J.S. Garofolo et al., "1999 Trec-8 spoken document retrieval track overview and results," *Proc. 8th Text Retrieval Conference TREC-8*, pp. 107-130, Gaithersburg, November 1999. (http://trec.nist.gov).

[8] Demonstration of the *LIMSI-Vecsys AudioSurf system*, presented by J.L. Gauvain and L. Lamel, at the *Workshop on Automatic Speech Recognition and Understanding - ASRU 2001*, Madonna di Campiglio, Italy, December 2001.

[9] J.L. Gauvain, L. Lamel, C. Barras, G. Adda, and Y. Kercadio, "The LIMSI SDR system for TREC-9," *Proc. of the Text Retrieval Conference, TREC-9*, pages 335-341, Gaithersburg, November 2000.

[10] A. Gelbukh, G. Sidorov and A. Guzmán-Arenas, "Document Indexing With a Concept Hierarchy," In *New Developments in Digital Libraries. Proceedings of the 1st International Workshop on New Developments in Digital Libraries (NDDL - 2001)*. ICEIS PRESS, Setubal, 2001.

[11] C. Hagège, "SMORPH: um analisador/gerador morfológico para o portugue°s, Lisboa, Portugal, 1997.

[12] G. Hernndez-brego, X. Menéndez-Pidal and L. Olorenshaw, "Robust And Efficient Confidence Measure For Isolated Command Recognition," *IEEE Automatic Speech Recognition and Understanding Workshop*, 2001, Madonna di Campiglio, Italy.

[13] D. Hiemstra, W. Kraaij, "Twenty-One at TREC-7: Ad-hoc and cross-language track," *Proc. of the 7th Text Retrieval Conference TREC-7*, pp. 227-238, Gaithersburg, November 1998.

[14] U. Iurgel, S. Werner and A. Kosmala, "Automatische Auswertung von Radio- und Fernsehnachrichten: Fortschritte in der Spracherkennung und Themenidentifikation," *ESSV 2002*, Germany.

[15] U. Iurgel, S. Werner, A. Kosmala, F. Wallhoff, and G. Rigoll, "Audio-Visual Analysis of Multimedia Documents for Automatic Topic Identification," *SPPRA 2002*, Greece.

[16] S.E. Johnson, P. Jourlin, K. Spärck Jones, P.C. Woodland, "Spoken document retrieval for TREC-8 at Cambridge University," *Proc. of the 8th Text Retrieval Conference TREC-8*, pp. 197-206, Gaithersburg, November 1999.

[17] Y.Y. Lo and J.L. Gauvain, "The LIMSI Topic Tracking System for TDT2001," *Proc. DARPA Topic Detection and Tracking Workshop*, Gaithersburg, November 2001.

[18] Y.Y. Lo and J.L. Gauvain, "The LIMSI Topic Tracking System for TDT2002," *Proc. DARPA Topic Detection and Tracking Workshop*, Gaithersburg, November 2002.

[19] Y.Y. Lo and J.L. Gauvain, "Tracking Topics in Broadcast News Data," submitted to the *ISCA ITRW on Multilingual Spoken Document Retrieval*.

[20] H. Meinedo, N. Souto and J. Neto, "Speech Recognition of Broadcast News for the European Portuguese language," *Proceedings ASRU'2001 - IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Italy, December 2001.

[21] D. Miller, T. Leek, R. Schwartz, "BBN at TREC-7: Using hidden Markov models for information retrieval," *Proc. of the 8th Text Retrieval Conference TREC-7*, pp. 133-142, Gaithersburg, November 1998.

[22] NIST Speech Group, "The 2001 Topic Detection and Tracking (TDT2001) Task Definition and Evaluation Plan," `ftp://jaguar.ncsl.nist.gov//tdt/tdt2001/evalplans/TDT01.Eval.Plan.v1.2.ps`, 15 November 2002.

[23] K. Ng, "A maximum likelihood ratio information retrieval model," *Proc. of the 8th Text Retrieval Conference TREC-8*, pp. 413-435, Gaithersburg, November 1999.

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

[25] B. Prouts, M. Garnier-Rizet, C. Barras, P. Paroubek, J.J. Gangolf, G. Adda, M. Adda-Decker and Y. de Kercadio, Demonstration of "An Audio Transcriber for Broadcast Document Indexation" - accompanying paper authored by J.L. Gauvain and B. Prouts. In RIAO'2000 Content-based Multimedia Information Access, Paris, France, April 12-14, 2000.

[26] G. Rigoll, "ALERT System for Selective Dissemination of Multimedia Information," *ISCA Tutorial and Research Workshop on Automatic Speech Recognition (ASR2000)*, Paris, France, September 18-20, 2000.

[27] G. Rigoll, "The ALERT System: Advanced Broadcast Speech Recognition Technology for Selective Dissemination of Multimedia Information" Invited talk at the *Workshop on Automatic Speech Recognition and Understanding - ASRU 2001*, Madonna di Campiglio, Italy, December 2001.

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

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

[30] S. Werner, U. Iurgel, A. Kosmala, and G. Rigoll, "Automatic Topic Identification in Multimedia Broadcast Data," *ICME 2002*, Switzerland.

[31] J.P. Yamron, I. Carp, L. Gillick and S. Lowe, "A Hidden Markov Model Approach to Text Segmentation and Event Tracking", in *Proceedings of ICASSP-98*, Seattle, May 1998.