

Video Story Segmentation with Multi-Modal Features: Experiments on TRECvid 2003

Laurent Besacier
CLIPS / IMAG

UJF BP 53
F-38041 Grenoble cedex 9

(+33)476635695

Laurent.Besacier@imag.fr

Georges Quénot
CLIPS / IMAG

CNRS BP 53
F-38041 Grenoble cedex 9

(+33)476635855

Georges.Quenot@imag.fr

Stéphane Ayache
CLIPS / IMAG

INPG BP 53
F-38041 Grenoble cedex 9

(+33)476514629

Stephane.Ayache@imag.fr

Daniel Moraru
CLIPS / IMAG

UJF BP 53
F-38041 Grenoble cedex 9

(+33)476514879

Daniel.Moraru@imag.fr

ABSTRACT

This paper describes the first steps of CLIPS/IMAG on the TREC video story segmentation task. We mostly describe the multi-modal features used and their respective performance for the story segmentation task. These features are based on the audio, video and text modalities. The preliminary system, which has the advantage to be relatively free with respect to the use of training data, is also presented in this paper. First experiments on the TRECVID 2003 evaluation set lead to a recall rate of 0.613 and a precision rate of 0.467. We plan to participate to the official TRECVID 2004 story segmentation task with this system.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *Retrieval models, Selection process.*

General Terms

Algorithms, Experimentation.

Keywords

Video story segmentation, TREC video, multi-modal features, macro-segmentation

1. INTRODUCTION

The main goal of the TREC Video Retrieval Evaluation (TRECVID¹) is to promote progress in content-based retrieval from digital videos via open metrics-based evaluation [11]. Among the different TRECVID tasks, the story segmentation task is defined as: given a test collection, identify the story boundaries

¹ <http://www-nlpir.nist.gov/projects/tv2003/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MIR'04, October 15–16, 2004, New York, New York, USA.
Copyright 2004 ACM 1-58113-940-3/04/0010...\$5.00.

with their location (time) and optionally their type (miscellaneous or news) in the given video clip(s). This was a new task for 2003.

A news story is defined as a segment of broadcast news with a coherent news focus which contains at least two independent, declarative clauses. Other coherent segments are labeled as miscellaneous. These non-news stories cover a mixture of footage: commercials, lead-ins and reporter chit-chat.

A story can be composed of multiple video shots, e.g. an anchorperson introduces a reporter and the story is finished back in the studio-setting. On the other hand, a single video shot can contain story boundaries, e.g. an anchorperson switching to the next news topic.

In TRECVID 2003, a participant had to detect all the transition boundaries from news to following news, news to non-news and non-news to news segments. CLIPS/IMAG did not participate to the official story segmentation task in 2003 (we however participated to other TRECVID 2003 tasks). However, we report here the first experiments we made on the TRECVID 2003 story segmentation evaluation data (on 105 video files of 30mn each), in order to prepare the TRECVID 2004 evaluation.

The different multi-modal features used by CLIPS (extracted from audio, video and text obtained with automatic speech recognition) are presented in section 2. The overall CLIPS system is explained in section 3. The experiments, evaluation metrics and discussions are listed in section 4. Finally, section 5 concludes this work and gives some perspectives.

2. MULTIMODAL FEATURES

Our approach to story bound segmentation is to use a range of different feature detectors and in this section we describe each of them in turn (the different performance measures given in this section in order to illustrate each feature are gathered in Table 1, section 4 of this paper; the evaluation metrics used are also detailed in section 4).

2.1 Long pauses detection

A silence detection is applied on the audio channel. It is only based on an energy bi-Gaussian distribution and on a detection threshold between the two Gaussians. The silence segment minimal length is set to 1 second in order to only catch relatively

“long” silence segments. It is interesting to note that this basic feature alone is already interesting for story segmentation. We have tested it on the reference boundaries and found its $F1^2$ measure to be 0.44, when all the long pauses were assigned to a boundary in the story segmentation system output.

2.2 Shot boundary detection

The CLIPS-IMAG shot segmentation system detects two types of transition effects: “cuts” and “dissolves” (including fades in and out) [9]. It detects cut transitions by direct image comparison after motion compensation and dissolve transitions by comparing the norms of the first and second temporal derivatives of the images. It also has a special module for detecting photographic flashes and filtering them as erroneous cuts and a special module for detecting additional cuts via a motion peak detector. The precision versus recall or noise versus silence compromise is controlled by a global parameter that coherently modifies the system internal thresholds. The system is globally organized according to a (software) dataflow approach and Figure 1 shows its architecture.

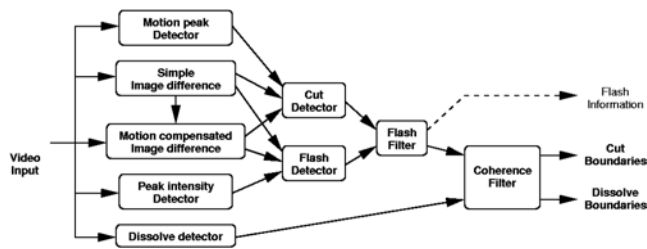


Figure 1 Shot segmentation system architecture

Using this feature alone lead to a F1 measure of 0.25 with a recall of 0.934. This recall result, different from 1, confirms that a single video shot can contain multiple story boundaries. Thus, selecting all the shot boundaries as candidate points for story boundaries is not sufficient. So, we take the union of shot boundaries and long pauses as candidate points for the story segmentation task, but we remove duplications within a 5s fuzzy window. A similar proposal was made in [5] and yielded 100% recall rate. Our union however leads only to 0.963 recall rate.

2.3 Audio change detection

Audio change detection may be a useful feature since many story boundaries correspond to an audio change on the audio channel. Examples of “audio change” are : speaker changes, speech to music transitions, speech to speech-over-music transitions, etc... These audio changes can be automatically obtained by detecting abrupt changes on the audio channel.

At the moment, the CLIPS audio change system is based on a BIC [3] (Bayesian Information Criterion) detector. It is important to note that the BIC criterion has been often used for speaker change detection whereas it should be able to detect any other abrupt change on the audio signal. Thus, we called our feature “audio change detection” instead of “speaker change detection”, even if a large part of the changes found with the BIC criterion may be actually speaker changes.

The signal is characterized by 16 mel cepstral features (MFCC) computed every 10ms on 20ms windows using 56 filter banks. Then the cepstral features are augmented by energy. No frame removal or any coefficient normalization is applied. The idea of the audio change detection is to find audio signal discontinuities that will help us to distinguish between two consecutive audio sources (speech followed by music ; speaker X followed by speaker Y ; ...). We can use two adjacent windows and a similarity measure between them. For the similarity measure we use the Bayesian Information Criterion (BIC) [3].

In order to apply the BIC we consider that the sound signal is a Gaussian process in the space of acoustic parameters. This kind of approach is based on the decision theory.

Let us consider two consecutive segments of speech, each of them being characterised by a sequence of spectral acoustic parameters (ex: coefficients MFCC, LPCC, etc) denoted by x_n ($n=1..N1$) and respectively by y_n ($n=1..N2$). We suppose that every sequence could be modeled by a multidimensional Gaussian distribution and that the vectors are statistically independent.

The question that we are asking regarding the two consecutive sequences is: do they belong or not to the same fundamental model or do both sequences correspond to the same acoustic source or not.

We must test the next two hypotheses:

H_0 : the two sequences correspond to the same acoustic source;

H_1 : the two sequences correspond to two different acoustic sources.

We can evaluate these two hypotheses using the generalised likelihood ratio. We will compute the ratio using maximum likelihood estimated models for the two sequences.

Lets say that $L(x; \mu_1, \Sigma_1)$ is the probability that the sequence x was generated by the Gaussian model characterized by the mean vector μ_1 and covariance matrix Σ_1 , and $L(y; \mu_2, \Sigma_2)$ is the same probability for the sequence y ; then the probability L_1 of the two sequences being generated by two different models is:

$$L_1 = L(x; \mu_1, \Sigma_1) L(y; \mu_2, \Sigma_2) \quad (1)$$

The probability of the two sequences being generated by the same model is:

$$L_0 = L(z; \mu, \Sigma) \quad (2)$$

where z is the joint sequence of x and y , and μ and Σ are the parameters of the model estimated from sequence z .

² $F1 = (2 \cdot P \cdot R) / (P + R)$, where P and R are precision and recall rates

If we consider that λ is the generalised likelihood ratio, then

$\lambda = \frac{L_0}{L_1}$ and we have:

$$\lambda = \frac{L(z; \mu, \Sigma)}{L(x; \mu_1, \Sigma_1) L(y; \mu_2, \Sigma_2)} \quad (3)$$

If we use log-likelihood then we have:

$$R = -\log \lambda = -\log(z; \mu, \Sigma) + \log(x; \mu_1, \Sigma_1) + \log(y; \mu_2, \Sigma_2) \quad (4)$$

It was proven that for mono-Gaussian distributions we have:

$$R = -\frac{N_1 + N_2}{2} \log |\Sigma| - \frac{N_1}{2} \log |\Sigma_1| - \frac{N_2}{2} \log |\Sigma_2| \quad (5)$$

We will compute the ratio for all available data and we will obtain a sequence of values $R(t)$. The estimate of the audio change point is the value that maximizes $R(t)$.

$$\hat{i} = \underset{i}{\operatorname{argmax}} R(t) \quad (6)$$

Looking for audio change points means looking for maximum points of the curve $R(t)$, called the BIC curve.

Figure 2 illustrates the whole audio change detection process. To select the maximum points of the BIC curve we use a sliding window that goes along the curve. The window is centred on the potential speaker change point. The point is selected if it has the highest BIC value in the window and if its BIC value is superior to

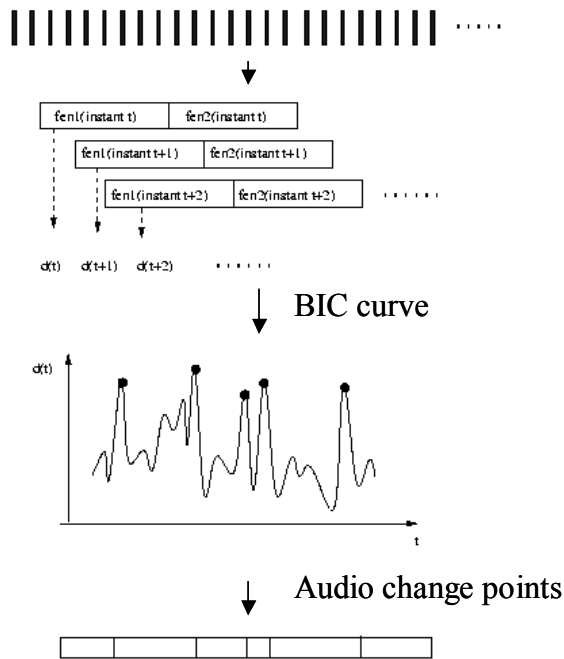


Figure 2 The audio change detection process

1.3 x average BIC curve value. The size of the window is 0.3 seconds. The use of the average BIC curve value gives us a data independent threshold. The use of the sliding window selects only the highest maximum among multiple close maximum points giving us a better precision.

For the story segmentation task, this feature alone gives a F1 score of 0.29 with 0.78 recall rate. That confirms our hypothesis that many story boundaries correspond to audio changes, but of course there are much more audio changes than story boundaries which explains the relatively low precision rate obtained with this feature alone (0.18).

2.4 Speaker segmentation

From the list of informative features that are provided in the ASR transcript, speaker information is available : for each speaker turn, a speaker label is assigned. This kind of output is generally called speaker segmentation [6] [7].

This speaker segmentation output may be useful : for instance, since we do not have yet a visual anchor face detector at CLIPS, a complete speaker segmentation output could be interesting to retrieve anchor person shots which are known to be very useful for story segmentation [5].

To illustrate the interest of speaker segmentation, Figure 3 shows as example the speaker segmentation of a complete 30mn video file, that can be obtained from the LIMSI ASR XML files. Each line corresponds to a speaker occurring on the audio channel.

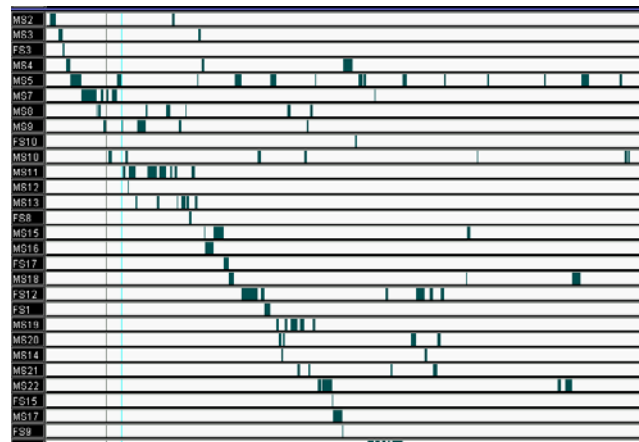


Figure 3. Speaker segmentation of a complete 30mn video

Each speaker intervention is given by the black segments on each line. We clearly see here that most of the speakers occur occasionally and on a limited period, except for the news presenter (on the 5th line) whose segments are spread over the whole video file. Thus, such an output segmentation is interesting for finding the anchorperson shots without using the image channel. Of course, an automatic segmentation makes some errors on the speaker interventions. For instance, best speaker

segmentation systems obtain around 15% of speaker segmentation error on broadcast news data, as shown in [7] and [10].

From the speaker segmentation output obtained with LIMSI ASR files, we extract automatically the news presenter line. There is no speaker training data available for any of the news presenters. The extraction is entirely based on the following empirical rules:

- The news presenter always speaks in the first five minutes of a broadcast news document.
- The news presenter interventions are spread over the entire news document. This is different from a usual reporter who speaks only during his news story.
- Finally, the news presenter is generally the main speaker (the speaker who has the highest total speech duration).

The first rule is applied as is, meaning that we first select all the speakers that speak during the first five minutes as “presenter candidates”. For the second rule we split the audio document in fixed length intervals (3 minutes each). For each speaker we count the number of intervals where he actually does not speak. We select the speaker(s) with the minimum number of “no speech” intervals.. Finally if at this point we still have more than one speaker as a potential news presenter, the final criterion is the quantity of uttered speech (third rule).

When we have extracted the news presenter line, we know the start and the end of each of his interventions. The story boundaries then correspond to each start point of the news presenter interventions. On the contrary, the end points of the news presenter interventions correspond generally to a reporter start inside the same story ; thus, these end points are used as “anti-story-boundaries” to remove story detection points that could be found by other multi-modal features at the same time.

For the story segmentation task, this feature alone gives a F1 score of 0.32 with 0.205 recall rate and 0.702 precision rate. It is not as efficient as a complete anchor face detector (the one from [5] leads to 0.51 F1 score on CNN data for instance), but this “presenter information”, extracted only from the audio channel helps to extract the general structure of a video document. It could also be very useful for story segmentation on radio broadcast, for instance, where no video channel is available.

2.5 Jingle detection

Detecting some key sounds (so called jingles) on the audio channel can reveal the beginning or the end of a particular sequence or announce it. This may be useful for the story segmentation part. Though most jingles include music, our jingle detector is not a music detector.

To detect and locate such jingles we used only one example of each jingle, taken on a separate video set. Each reference jingle was described by low level descriptors based on a spectral analysis while dissimilarity was measured between the target jingles and the whole video test set with an Euclidian distance, as done in [8] for instance.

More precisely, our low level descriptors were 8 coefficients corresponding to the spectral flatness feature computed on 8 frequency bands. This spectral flatness feature is part of the

MPEG-7 low level description and has shown to give interesting results for audio fingerprinting [1].

We have selected 10 jingles from ABC and CNN (CNN headline news jingles, CNN top stories, CNN sport, ABC short jingles, etc...) which have a length between 2s and 10s.

Of course, for the story segmentation task, this feature gives very few boundaries which results in a very low recall rate when used alone (0.028). However, the boundaries obtained generally correspond to effective story changes, since the precision obtained with this feature alone is 0.735.

2.6 ASR Text output

The LIMSI laboratory provided to all participants of TRECVID the output of their automatic speech recognition (ASR) system [4] for the whole 2003 video database. Our ASR-based feature was based on the selection of a set of lexical sequences likely to correspond to story transitions. To extract our list of lexical sequences, we calculated on the development data the most frequent N-grams (N=1 to 5) computed from ASR outputs located around reference story boundaries. From this, we manually made a list of 27 transition word sequences. Examples of word sequences extracted are : “A. B. C. News”, “C. N. N.”, “JUST BEFORE WE LEAVE”, “BACK WITH MORE NEWS”, “COMING UP IN TWO MINUTES”, ... This feature is rather similar to the “cue phrases” proposed and used in [2].

To find the story boundaries using the ASR output, we have selected all the speaker turns containing at least one of our “transition word sequences”. Then, the story boundaries were obtained by selecting the beginning or the end of each selected speaker turn according to the transition sequence concerned.

The use of this feature alone gives a F1 score of 0.41 with a relatively good precision (0.73).

3. SYSTEM OVERVIEW

3.1 Candidate Points

A good candidate set should have a very high recall rate on the reference boundaries. As seen in section 2.2, we decided not to use only shot boundaries, but the union of shot boundaries and long pauses which lead to 0.963 recall rate.

3.2 Overall strategy

At the moment, our strategy is very basic, but it has the advantage to be free of any development set which is not the case when some SVM-based combination schemes are used, for instance. It could however benefit from training when there is time and opportunity for it. This will be the case for additional features currently considered.

The general idea is to evaluate, for each candidate point, the output of each separate detector (described in section 2) which indicates the presence or not of a story boundary. For audio change (AC) and Pauses (P) features, a boundary is considered to

be detected if it is found inside a 2s fuzzy window around a candidate point. For ASR, Speaker Segmentation (SS) and Jingles (J), a boundary is considered to be detected if it is found inside a 4s fuzzy window around a candidate point. Then, the combination of features is based on logical operations between each separate detectors. For instance, in Table 1, *(AC AND P) OR J* means that a candidate point is considered to be a story boundary if one of the following cases is encountered :

- the audio change and pause detectors both found a boundary around it,
- the jingle detector found a boundary around it.

4. EXPERIMENTS ON TRECVID 2003

In this experiment we have used the 105 half-hour video programs of TRECVID 2003 evaluation. Our approach did not need training data since Pauses, Shots, Speaker Segments and Audio changes are obtained on-line from the test signals without any tuning necessary. Our ASR-based boundary detection has been developed by choosing “transition word sequences” . This was done by using other ABC/CNN videos present in the development set, as explained in section 2.6. The only detector that needed some examples was the jingle detector. Here again, we have extracted jingle samples from the development data.

We have made our experiments using the data and methodology proposed by TRECVID but we did so after the official evaluation period and the results presented here were not submitted to NIST. Therefore our results should not be directly compared to the official TRECVID 2003 official results because a) we would compare our system to systems that are older and could have evolved in the interval and b) though we have followed the methodology and we have made an appropriate use of development and test data respectively, we had knowledge of the results of other systems (which feature worked and didn’t work for instance). We have ranked our system within TRECVID 2003 evaluated systems however but this is just indicative.

4.1 Story boundary detection metric

The segmentation measure metrics are precision P and recall R and are defined as follows. According to the TRECVID 2003 metrics, each reference boundary is expanded with a fuzzy window of 5 seconds in each direction, resulting in an evaluation interval of 10 seconds. A reference boundary is *detected* when one or more computed story boundaries lie within its evaluation period. If a computed boundary does not fall in the evaluation interval of a reference boundary, it is considered as a *false alarm*. The precision P and recall R are defined in equations 1 and 2 were |.| means the number of boundaries.

$$P = \frac{|computed\ boundaries| - |false\ alarms|}{|computed\ boundaries|} \quad (7)$$

$$R = \frac{|detected\ reference\ boundaries|}{|reference\ boundaries|} \quad (8)$$

P and R are the official TRECVID measures for the story segmentation task. Since there is no ranking considered in this task, it is not possible to compute the classical Mean Average Precision (MAP) for system ranking. Since there are two values with very variable P versus R compromises between system, it is not easy to compare systems. In order to obtain a single measure to permit such comparison and ranking, we chose the classical F-measure (harmonic mean between P and R) and, more precisely, the F1 measure (giving equal weight to P and R in the mean).

4.2 Experimental results

The story boundary detection performance of the different detectors alone, as well as their combination with logical operators are given in Table 1.

Table 1. Story boundary detection performance on TRECVID 2003 evaluation data (105*30mn files)

	Recall	Precision	F1
Pauses (P)	0.613	0.344	0.44
Shots (S)	0.934	0.142	0.25
<i>P OR S</i> (candidate points)	<i>0.963</i>	<i>0.146</i>	<i>0.25</i>
Audio Change (AC)	0.782	0.176	0.29
Speaker Segmentation (SS)	0.205	0.702	0.32
Jingles (J)	0.028	0.735	0.05
ASR	0.280	0.734	0.41
AC AND P	0.495	0.382	0.43
(AC AND P) OR J	0.516	0.394	0.45
(AC AND P) OR SS OR J	0.567	0.405	0.47
(AC AND P) OR SS OR J OR ASR	0.616	0.450	0.52
(AC AND P) OR SS OR J OR ASR + commercial detection	0.613	0.467	0.53

The comments concerning the detectors alone are to be found in section 2. The association of audio change and pause feature (AC AND P) is slightly disappointing since it leads only to a F1 score of 0.43 which is approximately the same performance as the

pauses used alone. However, we kept the boundary points found because the precision is improved in that case. Adding the jingle detector (AC AND P OR J) improves the overall performance which shows the interest of this detector. It seems to be able to find boundaries that are not redundant with the boundaries found by other detectors. Adding now the news presenter information obtained from speaker segmentation (AC AND P OR SS OR J) improves again the overall performance since we reach 0.567 recall and 0.405 precision rate. It is also interesting to note that the association of these 4 features (AC, P, SS and J) leads to a system with acceptable performances without using the ASR text output (this corresponds to condition 1 of TRECVID 2003 evaluation plan).

If we add the ASR-based boundary detector, we reach a F1 score of 0.52. At this point, an analysis of the errors shows some false alarms occurring during commercial sequences. We have done a final experiment to detect and remove candidate points which are inside a sequence of commercials. The sequences of commercials were detected by applying a black frames detector on the video channel, since we noticed that commercials are generally separated by a variable number of consecutive black frames. This final process allowed to slightly increase the precision rate, leading to a F1 score of 0.53. If we look at the official results of TRECVID 2003 story segmentation evaluation, we would be ranked as the 4th system among 9 with such a performance (see note at the beginning of section 4).

5. CONCLUSION AND FUTURE WORK

We have presented our first steps on the TRECVID story segmentation task. We have mostly described the multi-modal features used and their respective performance for the story segmentation task. These features are based on the audio, video and text modalities. The preliminary system, which has the advantage to be relatively free with respect to the use of training data, has been tested on TRECVID 2003 evaluation set and leads to a recall rate of 0.613 and a precision rate of 0.467.

We plan to participate to the official TRECVID 2004 story segmentation task with this system. In the near future, we notably plan to use our own speaker segmentation system [6] [7] instead of the LIMSI one and to improve our commercial detection system in order to reduce false alarms. We also plan to include more features from the image track and from ASR analysis. We are currently developing a multi-modal story classification tool (politics, sports, weather, commercials, ...) and to integrate story segmentation and story classification together with a feedback to each other. We finally consider the integration of external feature detectors (developed elsewhere than at CLIPS) and the use of a more flexible, more analog (non-Boolean) and less ad hoc fusion procedure.

6. REFERENCES

- [1] E. Allamanche, J. Herre, O. Hellmuth, B. Fröba, T. Kastner, M. Cremer, "Content-based identification of Audio Material Using MPEG-7 Low Level DeACription", *ISMIR 2001*.
- [2] L. Chaisorn, C. Koh, Y. Zhao, H. Xu, T.-S. Chua, T. Qi "Two-level multimodal framework for news story segmentation of large video corpus", *12th Text Retrieval Conference*, Gaithersburg, MD, USA, 2003.
- [3] P. Delacourt and C. Wellekens, "DISTBIC: a speaker-based segmentation for audio data indexing," *Speech Communication*, Vol. 32, No. 1-2, September 2000.
- [4] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37(1-2) 89-108, 2002.
- [5] W. Hsu, L. Kennedy, C.-W. Huang, S.-F. Chang, C.-Y. Lin and G. Iyengar, "News video story segmentation using fusion of multilevel multimodal features in TRECVID 2003", *ICASSP'2004*, Montréal, Canada, Mai 2004.
- [6] D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, and I. Magrin-Chagnolleau, "The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation". *ICASSP'03*, Hong Kong.
- [7] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.-F. Bonastre, "The ELISA consortium approaches in Broadcast News speaker segmentation during the NIST 2003 Rich Transcription evaluation". *ICASSP'04*, Montreal, Canada, May 2004.
- [8] J. Pinquier and R. André-Obrecht, "Jingle detection and identification in audio documents" , *ICASSP'2004*, Montréal, Canada, Mai 2004.
- [9] G.-M. Quénot, D. Moraru, L. Besacier, "CLIPS at TRECvid: Shot Boundary Detection and Feature Detection", *12th Text Retrieval Conference*, Gaithersburg, MD, USA, 2003.
- [10] <http://nist.gov/speech/tests/rt/rt2004/spring/>
- [11] A. Smeaton, W. Kraaij and P. Over, "TRECVID 2003 – An introduction", *12th Text Retrieval Conference*, Gaithersburg, MD, USA, 2003.