Multimedia scenario extraction and content indexing for e-learning

Thomas Martin[1,2], Alain Boucher[3], Jean-Marc Ogier[1], Mathias Rossignol[2], Eric Castelli[2]

[1]L3i - Univ. of La Rochelle
17042 La Rochelle cedex 1
La Rochelle, France

[2]MICA Center
C10, Truong Dai Hoc Bach Khoa
1 Dai Co Viet
Hanoi, Vietnam

[3]IFI-MSI
ngo 42 Ta Quan Buu
Hanoi, Vietnam

thomas.martin@mica.edu.vn, alain.boucher@auf.org, jean-marc.ogier@univ-lr.fr
mathias.rossignol@mica.edu.vn, eric.castelli@mica.edu.vn

Abstract

In this paper, we present the use of multimodal content analysis in the MARVEL (Multimodal Analysis of Recorded Video for E-Learning) project. In this project, we record teachers giving their lectures in class and semi-automatically analyze the video-audio content in order to help transfer this lecture into a multimedia course for e-learning. We distinguish two primary goals in this application: scenario extraction (mostly from video) and content indexing (mostly from text and speech). Three objects take place in these goals: the teacher, the screen (for slide projection) and the whiteboard (for handwriting). These goals and the roles of all objects are explained in details, as well as our preliminary results. Through this application, we are giving some ideas about multimodality analysis and its formalization.

1 Introduction

Nowadays, as the available multimedia content grows every day, the need for automatic content analysis is becoming increasingly important. For example, information retrieval in broadcast news archives requires indexing the different medias available. Many projects currently focus on these research topics (content analysis, media enrichment...) but most of these works are focused on one single media, and are unaware of other medias. Because information is not concentrated in one media but distributed among all the medias, such approaches are losing important parts of that information, and ignore media interactions. Recently, many research works[14] have focused on the use of multiple modalities to increase the potentiality of analysis. However, to our knowledge, there is no existing framework for multimodal analysis, and there is only little serious study of the possibilities of interaction between modalities. In this paper, we present our ideas and framework on multimodal analysis, followed by our application in e-learning with the MARVEL project, which is divided into two goals: scenario extraction (mostly from the video) and content indexing (mostly from text and speech). This is still an on-going project, with some parts more developped than others. For each section, we will present our main ideas or detailed results, depending on work achievement.

2 multimodality

There is often confusion in the literature between the concept of media and the concept of modality. In many papers, the authors use both words to refer to the same concept. This does not seem to be exact, as we can see the two different concepts in the context of content analysis. We propose to define a modality as a refinement of the media concept. A media is characterized mostly by its nature (for example audio, video, text), while a modality is characterized by both its nature and the physical structure of the provided information (for example video text *vs* motion). One media can then be divided into multiple modalities, following two criteria: the semantic structure of the information and the algorithms involved in the analysis process. While the concept of media is independent from the application, the concept of modality is application dependant.

As proposed in [7], we will use generic modalities listed in three main families. First, the audio family includes different modalities in terms of structure such as speech, music or sound. Second, we distinguish between still image and motion (video) in the visual family. While both are acquired from a camera, motion contains time structure and is richer in term of content than still images. Third, the text family

includes printed text and handwritten text.

This split of media into modalities can surely be discussed and different organizations can be proposed. We will use this scheme through this paper using several examples taken from some applications to illustrate our choice. We insist on the fact that the information contained in each modality has a different structure, regarding the algorithms that can be used, the difficulty for content extraction and for the semantic that can be given to it.

Once modality is defined, the next step is to define multimodality. In video indexing context, Snoek and Worring [14] have proposed to define multimodality from the author's point of view: it is "the capacity of an author of the video document to express a semantic idea, by combining a layout with a specific content, using at least two information channels". The inter-modal relation is then located at a high level using semantics. On the contrary, in the context of speech recognition, Zhi *et al.* [20] have implemented the multimodal integration just after the feature extraction phase and an alignment step. In this case, multimodal integration takes place at a low level. Both these definitions are incomplete. Furthermore, several multimodal applications found in the literature use two modalities, audio and video, and the multimodal part of these application is often limited to a fusion step. Examples of such works include applications for video indexing such as [17] where a high level fusion step is processed after speaker segmentation in audio and shot detection in video. Shao *et al.*[13] have performed multimodal summary of musical video using both audio and video contents. In the same domain, Zhu *et al.*[21] perform video text extraction and lyrics structure analysis in karaoke contents using multimodal approaches. Song *et al.*[15] recognize emotions using a fusion step just after feature extraction in audio and video. Zhu and Zhou [22] combine audio and video analysis for scene change detection. They have classified audio shots into semantic types and process shot detection in video They integrate then these results to have robust detection. Zhi *et al.*[20] and Murai *et al.*[10] use facial analysis (video) to improve speech recognition (audio). [10] detects shots in video containing speech whereas [20] combines lip movements and audio features to process speech recognition. Zotkin *et al.*[23] propose a tracking method based on multiple cameras and a microphone array. Bigün *et al.*[4] proposed a scheme for multimodal biometric authentication using three modalities: fingerprint, face and speech. Fusion is processed after individual modality recognition.

We propose a more general definition of multimodality as an interaction process between two or more modalities. This process is based on an inter-modal relation. We have identified three different types of inter-modal relations [8]: trigger, integration and collaboration. The trigger relation is the simplest relation: an event detected in one modality activates an analysis process to start in another modality. The integration relation is already widely used and is mainly characterized by its interaction level. The analysis processes are done separately for each modality, but followed by a process of integration (fusion or others) of their results. Snoek and Worring [14] present a more complete review of existing works widely using the integration relation for the application of multimodal video indexing. The third relation is collaboration, and it is the strongest multimodal relation, consisting in a close interaction of two modalities during the analysis process itself. The results of the analysis of one modality are used for analyzing a second one.

3 Video analysis for e-learning

Our main application for multimodality is e-learning through the MARVEL project. The goal of MARVEL (Multimodal Analysis of Recorded Video for E-Learning) is the production of tools and techniques for the creation of multimedia documents for e-learning.

The complete course of a professor is recorded live. Furthermore, textual sources such as course slides may be available. The recorded material from live courses is analyzed and used to produce interactive e-courses. This can be seen as an application of video analysis to produce rich media content. The slides used by the professor in the class can be automatically replaced by an appropriate file in the e-course, being synchronized with the professor's explanations. The course given by the professor is indexed using various markers from speech, text or image analysis. The main aim of this project consists in providing semi-automatic tools to produce e-learning courses from recorded live normal courses.

In this project, three different medias are available: audio, video and lecture material (essentially the slides). Following the model proposed in section 2, we have identified five different modalities: *i) printed text* which contains text from the slides and, if available, from other external textual sources. This modality is present in both video and lecture material media; *ii) handwritten text* which consists in the text written on the whiteboard; *iii) graphics* which include all the graphics and images present in the slides. *iv) motion* which contains the motion content of the video media; *v) speech* which gathers the teacher's explanations.

To simplify the explanations in this paper, we will not take into account the *graphic* modality and we consider only the textual parts of the slides. A difference must be made between *handwritten text* and *printed text* for two reasons. First, as presented in section 2, the nature of both modalities is different (handwritten *vs* printed text). The second reason is specific to this application: the two modalities do not contain the same data. Even if the contents of both modalities

are related to the course, one (*printed text*) is more structured than the other.

The *printed text* modality is available in two different medias: video and text. It is a good example to illustrate our distinction between media and modality (section 2). Even if it is available in two different medias, the *printed text* still contains the same information, with the same structure. Once detected and extracted from the video media, the analysis processes involved are similar whatever the media.

The application is divided into two distinct parts, which represents two different, but complementary, goals to achieve: *i scenario extraction* (section 4): The scenario is given mainly by the video. The teacher's behavior (see *fig.* 1) is analyzed to extract the course scenario (explaining the current slide, writing on whiteboard, talking to the class,...). This will be used later as a layout during the e-course production. Other regions of interest such as the screen or the whiteboard are detected. Detections of slide changes or new writing on the whiteboard are events that will be used; *ii content indexing* (section 5): The content indexing of available media has to be done using the speech given by the teacher, the printed text on the slides and the handwritten text on the whiteboard. These three sources are complementary to show all the content of the course. Different inter-modal interactions are identified here.

During the first part of the application (scenario extraction), 3 trigger relations are involved. These relations are directly related to the actors who interact in a course: teacher, whiteboard and screen. The trigger source is the *motion* modality. First, the "slide transition" event triggers the *printed text* detection and recognition. Second, the "teacher points at screen" event triggers the point of interest search. Third, similar to the first, the "teacher writes on whiteboard" event triggers the *handwritten text* recognition process.

The second part of the application (content indexing) contains most of the inter-modal relations. First, the *speech-printed text* interaction is a bimodal and bidirectional collaboration interaction, with its main direction from *printed text* to *speech*. As used in [20, 10], *motion-speech* interaction can be also useful . Recognition of *handwritten text* is a difficult task, especially in video. We propose to help recognition of *handwritten text* using both *speech* and *printed text* modalities. Both relations, *speech-handwritten text* and *speech-printed text*, are bimodal and unidirectional.

4 Scenario extraction

Scenario extraction aims at retrieving the structure of the lecture. We have identified three elements in the MARVEL application (see *fig.* 1: the screen, the whiteboard and the teacher. Both the screen and the whiteboard are passive elements, whereas the teacher interacts with the others. The

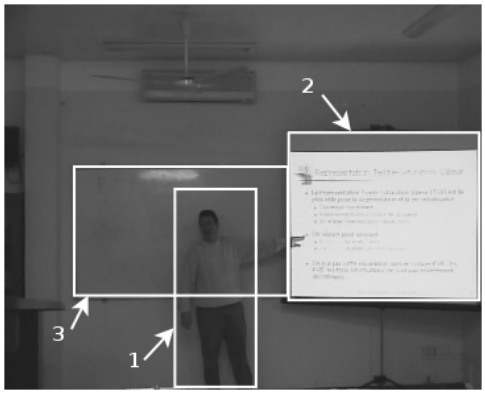

Figure 1. Frame extracted from a recorded course. White shapes highlight identified actors of the application: the teacher (1), the screen (2) and the whiteboard (3).

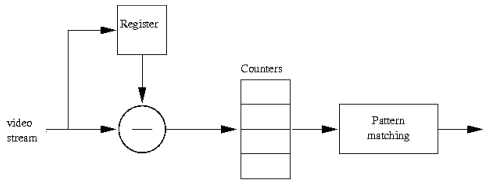

Figure 2. Slide change detection schema

screen is the support to display slides. In the rest of this paper both words, screen and slides, will be used referring to the same object. The information displayed on the screen is structured and contains text and graphics.

4.1 Slide change detection

Similarly to shot detection in more general videos, slide change detection aims at segmenting the video into logical units, each containing a different slide. A slide change is defined as the replacement of a slide by another. Such change is characterized by an image modification in the screen zone. This modification is more or less important according to the considered slide change, and can be considered as the result of an edit operation on the current slide. Slide changes have a global effect to the screen, whereas slide modifications are more located.

During the course the teacher's interactions with the screen can temporally occlude the slide. Another source of motion is inherent to the compressed video: as video compression algorithms often suppress high frequency information, small patterns such as letters are affected by temporal noise. Such patterns are obviously frequent in slides.

Our slide change detection algorithm is based on image differences. However, we introduce a priori knowledge

sequence	occured	detected	false detections
seq1	24	23 (98.8%)	0
seq2	10	10 (100%)	1
seq3	13	13 (100%)	0

Figure 3. Some results for our slide transition detection algorithm on three video sequences taken from three different class courses. For each sequence, the actual number of transitions occurring in the sequence is shown, followed by the numbers of detected transitions and false detections.

by restricting this to the slide zone in the stream. As the camera position and view are fixed during the recording, we manually fix this zone. That screen zone is extracted from each 25th frame (number decided upon experimentally), and an image difference is computed. The resulting image is thresholded to eliminate the compression noise. To avoid problems due to the teacher's interaction with the slide, we divide the image into quarters and count the modified pixels in each quarter. Indeed, modifications will be detected when the teacher interacts with the screen. However, if the screen is divided in 4 parts, he will not interact with all quarters at the same time. The 4 resulting values are normalized. Thus, we obtain 4 temporal curves describing motion in the slide zone.

The slide transition detection is performed through simple pattern matching on these curves. If modifications are simultaneously detected in the 4 quarters within less than 2 seconds, we consider that a slide transition has occurred. Simultaneous modifications in one, two or three quarters are pieces of evidence but are not sufficient to detect slide modification.

Tests have been performed on three sequences (see table 3). These results are quite satisfactory. However, in the case of slide changes occurring on three or less quarters of the slide, the algorithm will not detect them. We will see in the next section that the teacher detection algorithm permits to solve this problem.

4.2 Teacher detection

The teacher detection aims at getting the position of the teacher in the classroom and to determine with which zone he is interacting: the screen or the whiteboard. For this task, we use an algorithm based on image difference (see *fig.* 4). To improve its results, we extract an image of the background, which is substracted from the current frame. However, the screen zone in the image difference is very noisy and causes many false detections. To avoid this, we

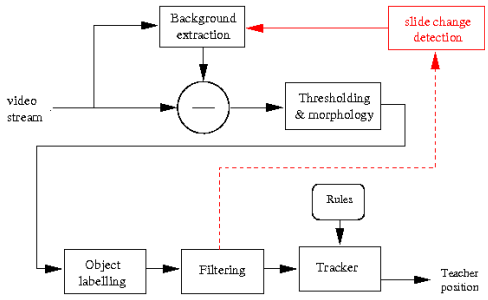

Figure 4. Teacher detection schema

use the slide transition detection results (see section 4.1) to correctly update the background. More precisely, the screen zone of the background is not updated, except if a slide change is detected. After a morphological step, bounding-box candidates are extracted and filtered. Collaborations between these two algorithms lead to more accurate results for the course scenario extraction.

4.3 Teacher gesture detection

Depending on the people, gesture may take an important part in the communication process. In our specific context, two main groups of gestures can be identified: *i)* free gestures; as an example when the teacher is interacting with the classroom, *ii)* constrained gestures; we put in this group the interactions with the whiteboard or the screen.

Even if gestures can provide useful information, due to high variability of gesture types, the first category does not seem to be usable. On the contrary, the second one does not have this problem, and specific gestures, such as pointing something, can be identified. Moreover, the succession of gestures can provide relevant indices for whiteboard or slide content ordering (scenario extraction giving the order used by the teacher to present the content).

5 Content indexing

5.1 Text detection and recognition

As presented, our strategy relies on the inter-modality cooperation for the course content indexing process. Here, the purpose is to use text recognition from the slides in the video for guiding the indexing process, especially through providing information from this text recognition module to the speech recognition one.

The problem of text recognition is a very well known problem, for which many contributions can be found in the litterature [6], and industrial software quite reliable. Most of these recognition tools have been developed in the context of "high resolution" images, and have to be re-visited

and adapted in the context of our problem, because of the quality of the images.

In the context of the MARVEL project, two categories of text information have to be considered: the text which is handwritten by the teacher on the blackboard and the text which is presented on slides prepared on ICT tools such as PowerPoint. The indexing process can also rely on graphic parts, drawn by the teacher on the blackboard or presented on the slides. So far, we have not considered the question of the recognition and indexing of all the information drawn or written by the teacher on the blackboard.

Our first developments deal with the slides-based indexing process, through a recognition process of the information which is presented on these supports. In this kind of context, the usual document processing chain proposes a first stage whose aim is to separate (segmentation stage) the different layers of information of the document. Generally this "physical" segmentation process depends on the a priori knowledge concerning the information of the document: text, size... In the context of our project, we have decided to apply a blind segmentation process, based on very relevant tools developped in the context of ancient document processing [5]. The segmentation process relies on the computation of the auto-correlation function, allowing to detect regular orientations, some of them being highly representative of the presence of text.

Using these tools in the context of slide segmentation is found to be a very relevant approach, since it is very difficult to have reliable a priori knowledge concerning text features. Concerning the text recognition engine, we decided to develop our own recognition tools. This decision was motivated by a strong competence in this domain in our lab, and also because we wanted to take the benefits of all the intermediate information concerning the recognition process, which is rarely available in industrial tools. As a consequence, we developed a "classic text recognitionengine", based on relevant features [11]. These features are introduced as input of a KNN classifier [1], allowing to provide a confidence associated with each decision, information that can be re-used in a feedback process, in the context of inter-modality cooperation. The exploitation of this text recognition tool in the context of slide recognition is very encouraging.

A syntactic analysis tool allows to increase this recognition rate, in relation with a dictionary which is available in our system. This text recognition tool provides some information that can be considered as indexes for the indexing process, and that can be transmitted to the speech recognition module to increase its performances. This intermodality indexing process allows to increase the quality of the index in a very significant manner.

5.2 Speech recognition

In the MARVEL project, we aim at indexing available data streams for further use such as audio-video and slides synchronization. The most direct way to obtain semantic indexing is through linguistic data, which can in particular be obtained using speech recognition techniques. However, in such a project, full continuous speech recognition is not useful, since we do not intend to perform a complete automatic transcription, but only audio content indexing. Thus, our aim is to detect keywords in the speech recording. In a first research step, we perform tests in order to evaluate what we can recognize using an existing automatic speech recognition (ASR) tool. The ASR software used is Raphael [2], which is *a priori* not well adapted at all to the kind of speech we are dealing with, but is quite representative of the state of the art in voice analysis.

ASR tools typically use a three step process. First, potential phonemes are extracted from the signal using an acoustic model, then a lexicon of phonetic transcriptions is consulted to find which words may correspond to those phonemes, and finally the lattice of word hypotheses is processed through a language model (LM) to find which sequence of words is the most linguistically plausible one. We cannot affect directly the first of those steps, since developing acoustic models is a huge, very technical task, but the two subsequent ones exhibit weaknesses which we can amend. A first problem is the incompleteness of the phonetic lexicon, from which words used during the course may be absent. Since the absent words are typically the most specialized, technical ones, which are also the most likely to be interesting keywords, this is a very critical problem. A second difficulty appears with the language model: in the tool we are using, as in most existing ASR software, the LM consists in a database of three-word sequences (3-grams) probabilities. Such probabilities are difficult to compute reliably for general spoken language—if such a thing even exists—and in the ideal case, a specialized language model adapted to the considered speaker and topic must be used. We must find an inexpensive way to develop such a specialized LM without the data usually exploited or that task (a consequent transcript of spoken language dealing with the considered topic).

The chosen approach is to mix an existing generic spoken language model with a small specialized language model extracted from textual data related to the course: the text of the slides used by the teacher. This text is very simple and features prominently the keywords of the lessons, which is precisely what we are interested in. From the same text, we shall also extract all words that are absent from the phonetic lexicon and add them to it thanks to an automatic phonetization tool.

A similar idea is followed in [18], but the authors of that

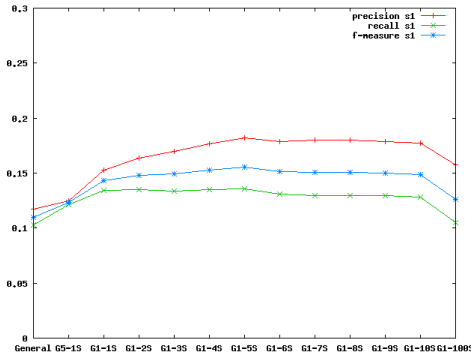

Figure 5. Speech recognition results for all words with different models mixing weights being used. In that case, recognition rates vary between 10% and 20%.

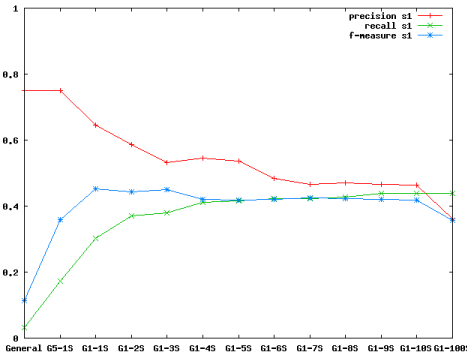

Figure 6. Speech recognition results restricted to keywords extracted from slides with different model mixing weights being used. On these domain keywords, results are significantly higher and vary around 50%.

work evaluate the interest of the additional textual information in terms of theoretic modeling power (perplexity) of the hybrid LM. We shall adopt a more practical approach, and directly evaluate the impact of that transformation of the LM on speech recognition rates. In order to perform our tests, we have used two sequences named *speaker1* and *speaker2*, each with a distinct speaker and topic. *Speaker1* is a 20 minutes presentation of approx. 3,500 words, whose attached slides gather 748 words, while *speaker2* lasts 35 minutes, counts 6,000 words, and has slides gathering 751 words.

The generic LM is a French spoken language model with a vocabulary containing about 16,500 words. The specialized LM is automatically built from the text slides using the SRILM language modeling toolkit [16], and the phonetization of new words is performed by the LIA-PHON automatic phonetizer for French [3]. Tests have been performed with different language models: *i* the original, generic model of Raphael only; *ii* various mixed models obtained by performing a weighted average of transition probabilities between the generic and specialized model with the following weights: 5-1, 1-1, 1-2... 1-10, and 1-100 (first number is the generic model weight and second is the specialized model weight).

5.2.1 Results

The two sequences have been manually transcribed. For each sequence the recognition result has been aligned to this transcription. Recognition rates have been computed on: *i)* all words, *ii)* keywords manually selected by the teacher, representative of the course content.

Figures 5 and 6 present the results obtained for the sequence *speaker1* depending on the model mixing weights used. The results obtained for *speaker2* are nearly identical. We can observe that the use of a mixed model significantly improves the results of speech recognition for all words relatively to the original performance, but since that one was very low, the results remain not reliable enough to be exploitable. However, recognition rate on keywords is greatly improved. The best obtained f-measure using a mixed model is about 50% (with 1-4 and 1-5 weights) while it was only about 10% using the general model. In both cases (and on *speaker2*), we can see a peak in performance when the weight for the specialized LM is about 5 times higher as that of the general LM. That seems to correspond to an optimal level of specialization, above which the LM loses too much generality to be able to model "ordinary" speech.

5.3 A video-text driven speech keyword detection

As the results presented above show, the result of continuous speech recognition is not usable as is. However, the recognition results can be improved by introducing a specialized knowledge. In the context of the MARVEL project, such knowledge can be provided as automatically as possible. We propose to automatically select keywords in the slide text and to use them to improve speech recognition. Provided that the slides are not available as input of our process, their text has to be extracted from the video stream.

Instead of building a LM with this text, we propose to stop the continuous speech recognition process after phoneme extraction. At this step, the output is a lattice of phonemes hypothesis. Selected keywords will be searched in this lattice.

Keywords are selected in the text of the slides. Depending on the teacher, this text is more or less concise. The au-

tomatic keyword selection can be performed using a *stoplist* or more complex methods using morpho-syntactic analysis with tools such as *TreeTagger*[12]. After this selection step, these keywords will be phonetized using the LIA-PHON phonetizer, then searched in the phoneme lattice.

5.4 Text and speech attributes

Slide text information is not only borne by words. Text can have many attributes that participate in characterizing the content, such as size, position of text, style (title, subtitle or item), color, font weight, slanting, underlining, etc. These attributes can be used to order the different ideas presented in the slide and to stress on some important ideas. A complete slide text representation model must include these attributes. They will be used later for multimedia representation of the course content, but also for content retrieval (section 5.5).

Similarly to text, speech can also have many attributes to characterize its content. These attributes can be relative to the prosody of the speech or to emotions expressed when speaking. In the first case, changes in prosody can be used to determine between interrogative and affirmative sentences for example [9]. In the second case, emotions in the speech can be used to emphasize on a word or to discriminate between two ideas [19].

Text and speech attributes do not contribute to the content recognition process. They are mostly recognized independently and associated to the (spoken or written) words that they characterize, but they will be of importance as driving factors for the indexing and retrieval of the course contents.

5.5 Content indexing and retrieval

So far we have worked on speech and text detection and recognition, with some experiments on attribute recognition. The final objective of the content recognition for this application is to be able to index and retrieve the course content. A user (student for example) should be able to query a course database and retrieve links to audio-video records fulfilling his needs. In this section, we present our preliminary ideas on course content indexing and retrieval.

An accurate model for content indexing and retrieval must include four aspects: text, speech and their respective attributes. The speech recognition model based on text slides (section 5.2) is limited on domain key words. Following this model, content indexing is limited to key words, both for text and for speech, and do not include the whole word content. This limitation restricts the retrieval scheme to key words or key concepts in the course domain. It is acceptable for the application, where content indexing and retrieval should help the user to browse into the course content.

The currently developed content retrieval model is based on the combination of text and speech, plus attributes when available. The time unit used to index the text and speech content is based on slide change detection (section 4.1), which defines a time interval [t1,t2] for each slide. Text is naturally associated with slide display. But speech can also be indexed using the same scheme, given the hypothesis that the teacher's speech is always related to the displayed content. This hypothesis is not always true, but sufficiently to allow indexing of all speech content following that scheme. To be more specific, speech associated to a slide lasts from the last audio silence before the slide change (marked as the beginning of a sentence) to the first silence following the next slide change (marked as the end of a sentence).

Undergoing work bears on the weights in the retrieval model to be associated with text and speech attributes. It sounds natural that the co-presence of a word in both the speech and the text content indicates a high relevance of this part of the video regarding to the query. But the influence of attributes on the relevance of a word is less obvious, and depends on each attribute.

Regarding the text attributes, the position and the size of a word gives a good idea of its relevance. But other attributes such as color, bold or italic need to be tested, as there are no given rules on how to use these attributes. They also depend on each person who can mean different things using the same attributes. While some people use many attributes on their slides, others may never use them. Moreover, the attributes may not be on the indexed keywords, but on neighbor words. A possible example of this is the emphasis that can be made on some words like *do*, *must* or *never*, which are not domain keywords but used to characterize the text preceding of following them.

Regarding the speech attributes, emphasis on words seems to be the most important attribute to take into account. Such emphasized words should be more relevant. The role of other attributes in the retrieval model is not clear and not defined yet. As for the text attributes, the speech retrieval model should take into account emphasis made on generic words not part of the content domain but used to emphasis preceding or following domain content.

6 Conclusion and future works

In this article, we have presented our work on multimodal analysis through the two main goals of the MARVEL project: scenario extraction and content indexing. The interactions between the different modalities for the indexing process rely on a device based on different triggers allowing starting the cooperation between the different recognition modules. Of course, our future works will deal with the

improvement of each recognition module, but the theoretical works will also consider a formal description of these interactions through adapted mathematical tools. For these points, some current studies deals with Petri nets combined with Bayesian networks.

So far, we have skipped the whiteboard text analysis and indexing. As it is handwritten text and low resolution images, this work is more difficult than for printed text.

As presented, we are currently developing a keyword detection tool. Our aim is to automatically select keywords from the slides. If an electronic version of the slides is available, direct access to the text is available. If not, we have to perform video text recognition to access this text. One possibility to select the keywords is to analyze de teacher's gestures. For example, when the teacher points at a zone in the current slide, that information can be used to characterize and to stress on a given content. Consequently, specific words can be highlighted and selected as relevant keywords.

References

[1] K. Aas and L. Eikvil. Text categorisation: A survey. Technical Report 941, Norwegian Computing Center, 1999.

[2] M. Akbar and J. Caelen. Parole et traduction automatique: le module de reconnaissance RAPHAEL. In *17th International Conference on Computational Linguistics (COLING 98)*, pages 36–40, Montreal, Québec, Canada, 1998.

[3] F. Béchet. LIA-PHON : Un système complet de phonétisation de textes. *TAL (Traitement Automatique de Langues)*, 42(1):47–67, 2001.

[4] J. Bigün, J. Fiérrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Multimodal biometric authentication using quality signals in mobile communications. In *Proc. of ICIAP*, pages 2–11. IEEE Computer Society, 2003.

[5] N. Journet, R. Mullot, V. Eglin, and J. Ramel. Dedicated texture based tools for characterisation of old books. *dial*, 0:60–69, 2006.

[6] L. Li, G. Nagy, A. Samal, S. C. Seth, and Y. Xu. Integrated text and line-art extraction from a topographic map. *IJDAR*, 2(4):177–185, 2000.

[7] T. Martin, A. Boucher, and J.-M. Ogier. Multimodal analysis of recorded video for e-learning. In *Proc. of the 13th ACM Multimedia Conference*, pages 1043–1044. ACM Press, 2005.

[8] T. Martin, A. Boucher, and J.-M. Ogier. Multimodal interactions for multimedia content analysis. In *Proc. of ICTACS 2006*, pages 68–73. World Scientific, 2006.

[9] V. Minh-Quang, T. Do-Dat, and E. Castelli. Prosody of interrogative and affirmative sentences in vietnamese language: Analysis and perceptive results. In *Interspeech 2006*, Pitsburg, Pennsylvania, US, 2006.

[10] K. Murai, K. Kumatani, and S. Nakamura. Speech detection by facial image for multimodal speech recognition. In *Proc. of ICME*, page 149. IEEE Computer Society, 2001.

[11] S. Nicolas, T. Paquet, and L. Heutte. Markov random field models to extract the layout of complex handwritten documents. In *IWFHR-10*, pages 563–568, La Baule, France, 2006.

[12] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

[13] X. Shao, C. Xu, and M. S. Kankanhalli. Automatically generating summaries for musical video. In *Proc. of ICIP*, volume 2, pages 547–550, 2003.

[14] C. G. M. Snoek and M. Worring. Multimodal video indexing: a review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[15] M. Song, J. Bu, C. Chen, and N. Li. Audio-visual based emotion recognition - a new approach. In *Proc. of CVPR*, volume 2, pages 1020–1025. IEEE Computer Society, 2004.

[16] A. Stolcke. SRILM – an Extensible Modeling Toolkit. In *ICSLP 02*, pages 901–904, Denver, CO, USA, 2002.

[17] S. Tsekeridou and I. Pitas. Audio-visual content analysis for content-based video indexing. In *Proc. of ICMCS*, volume 1, pages 667–672. IEEE Computer Society, 1999.

[18] L. Villaseor-Pineda, M. M. y Gmez, M. Prez-Coutio, and D. Vaufreydaz. A Corpus Balancing Method for Language Model Construction. In *4th International Conference of Computational Linguistics and Intelligent Text Processing*, pages 393–401, Mexico City, Mexico, 2003.

[19] L. Xuan-Hung, G. Quenot, and E. Castelli. Speaker-dependent emotion recognition for audio document indexing. In *The 2004 International Conference on Electronics, Informations and Communications (ICEIC 2004)*, Hanoi, Vietnam, 2004.

[20] Q. Zhi, M. Kaynak, K. Sengupta, A. D. Cheok, and C. C. Ko. HMM modeling for audio-visual speech recognition. In *Proc. of ICME*, pages 201–204. IEEE Computer Society, 2001.

[21] Y. Zhu, K. Chen, and Q. Sun. Multimodal content-based structure analysis of karaoke music. In *Proc. of the 13th ACM Multimedia Conference*, pages 638–647. ACM Press, 2005.

[22] Y. Zhu and D. Zhou. Scene change detection based on audio and video content analysis. In *Proc. of ICCIMA*, page 229, 2003.

[23] D. Zotkin, R. Duraiswami, and L. S. Davis. Multimodal 3-d tracking and event detection via the particle filter. In *Proc. of Event*, 2001.