Audio/Video Fusion: a Preprocessing Step for Multimodal Person Identification

Gaël JAFFRÉ Julien PINQUIER

Université Paul Sabatier - 118 route de Narbonne
SAMOVA - IRIT
31062 Toulouse Cedex 9 - France
{jaffre, pinquier}@irit.fr

Abstract

In the audiovisual indexing context, we propose a system for automatic association of voices and images. This association can be used as a preprocessing step for existing applications like person identification systems. We use a fusion of audio and video indexes (without any prior knowledge) in order to make the information brought by each of them more robust. If both audio and video indexes are correctly segmented, this automatic association yields excellent results. In order to deal with oversegmentation, we propose an approach which uses one index to improve the segmentation of the other. We show that the use of the audio index improves an oversegmented video index on a corpus composed of French TV broadcasts.

1 Introduction

By analogy with textual documents which can be easier to handle (storage, data mining, accessible by anybody, etc), multi-media document treatment is only at its beginning. For example, to find a video which contains the first steps of Armstrong on the Moon (without prior information) is rather critical. It would require to find semantics from the video and/or the audio.

Many works were carried out on the audiovisual content characterization, and particularly on person detection. The majority of these studies are monomedia and allow the detection of a person either by his visual appearance in a frame (like a face) or by his voice.

On one hand, the image study is based on visual features, like face detection (many applications are indexed in [1]), or on costume detection [2]. On the other hand, the audio analysis is based on homogeneous segments, which follow a speaker segmentation via the Bayesian Information Criterion (BIC) for example [3].

Sometimes, the objective is to improve exclusively audio-based systems with video (like Automatic Speech Recognition (ASR) [4]) or sometimes, it is the opposite [5, 6].

More recent works start video content analysis by integrating both audio and visual features, which are the two inseparable parts of a video bitstream. Thus, an adaptive speaker identification system that employs audiovisual cues which is based on a probabilistic framework is proposed in [7]. In [8], a rather similar approach based on confidence values is presented. The goal of these applications is to find, starting from voice and face models, in which sequences a given person appears: each face is associated with a voice.

However, in some applications, this audio/video association is not available. For example, in our case, we do not have any prior model: they are computed on the fly, when the persons appear. That is why, in this paper, we present a framework for audio/video fusion in order to compute automatically these association models. Our goal is not to improve descriptor or segmentation method qualities: we use the fusion of the audio and video indexes in order to make the information brought by each of them more robust. This work can be used before the preceding approaches (like a pretreatment, for example): our goal is to merge and associate effectively results (or indexes) of audio and video parameters of each person without any prior knowledge. The interest of this work is multiple. On the one hand, the idea is to limit analysis tool oversegmentation (audio and/or video) and on the other hand to associate a voice to each visual character.

First, we propose in section 2 a common index to compare audio and video. Then, in section 3 we make a description of the automatic algorithm of voice/image association. Finally, section 4 show experiments in order to validate our audiovisual index merging. The corpus is composed of French TV broadcasts (like TV games) and could be generalizable with other broadcasts.

2 Common index for audio and video

In this section, we propose a common scheme for audio and video indexes so that we can compare them. Moreover, we present a general framework for audio and video fusion that will be more detailed in section 3.

2.1 Drawback of existing applications

Some recent methods use both audio and video cues for improving person identification [9, 10]. Their goal is to identify persons using both visual features (the face is often used) and speech recognition. However, this processing is carried out in video subsequences where both visual features and speech are present, and the assumption that the current voice corresponds to a visual feature in the frame is made.

In real sequences, this hypothesis is often violated. It is very common to find sequences where the appearing persons do not speak during many frames (or many shots). Moreover, it is also usual that the current voice belongs to a person whose visual feature is not in the current frame. For example, figure 1 presents the number of appearances of the ten main characters in a TV talk show, for both audio and video channels. We can see that these probability distributions are quite different for audio and video indexes, which is involved by the violation of the usual assumptions. If these assumptions were verified, the number of occurrences of each character would be similar in the two indexes.

In our application, we propose to compute co-occurrences between audio and video indexes, i.e. to compute the intersection between these indexes. Then, from this fusion we determine the matching between voices and images. This approach is well suited to take into account the cases where the usual assumptions are not verified.

2.2 Comparison of audio and video indexes

When audio and video indexes are generated from different applications, their structure can be quite different. For example, video index can refer to frames or to shots, whereas audio index can refer to speech segments. An example is given in figure 2. So, a direct comparison of the two indexes is not possible.

That is why we propose to use a common index, in order to be able to directly compare audio and video. To make easier statistic comparison between these indexes, we propose to use discrete indexes using frame by frame decomposition. Both audio and video indexes are written frame by frame, as shown in figure 3. To obtain this audio index from a traditional audio index (as the one presented in figure 2), we convert each speech segment by computing the frame which corresponds to the beginning of the segment, as well as the one which corresponds to the end. To compute them, we use the video frequency (25 Hz in our examples). Then, the voice which is heard in this segment is associated to every frame of this subsequence. As an illustration, the transformation of the indexes of figure 2 is given in figure 3. Using this normalization, a direct comparison is now conceivable. It will be presented in next section.

3 Automatic matching of audio and video

In some applications, as in [9], each voice must be associated with a visual feature, like a face. Actually, it is often a hand-made association, computed with learning data. In this section, we propose a framework to automatically realize this association, using a statistical analysis of audio and video indexes.

For the purpose of the presentation, we make the assumption that indexes are perfectly segmented, i.e. there is not oversegmentation. So, each character has only one voice and exactly one visual feature. Moreover, each voice is associated to exactly one face, and conversely. However, in section 4, we will show a framework to deal with oversegmentation.

3.1 Index intersection

First, we compute a matrix which represents the intersection between audio and video indexes. We use the following notations:

- n_a is the number of different voices in the audio index,
- n_v is the number of different visual characters in the video index,
- $\{A_i\}_{i=1\ldots n_a}$ is the set of voices of all the characters,
- $\{V_j\}_{j=1\ldots n_v}$ is the set of visual features of each character.

To compute this intersection matrix, we go through the two indexes, frame by frame. For each frame, if the voice A_i is heard and the visual character V_j is present, the number of occurrences m_{ij} of the pair (A_i, V_j) is incremented. Then, we obtain the following matrix:

$$m = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_{n_a} \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & \ldots & V_{n_v} \end{array} \\ \left(\begin{array}{cccc} m_{11} & m_{12} & \ldots & m_{1n_v} \\ m_{21} & m_{22} & \ldots & m_{2n_v} \\ \vdots & \vdots & \vdots & \vdots \\ m_{n_a1} & m_{n_a2} & \ldots & m_{n_an_v} \end{array} \right) \end{array} \quad (1)$$

In this matrix, the value m_{ij} means that in all the frames where the voice A_i is heard, the visual character V_j appears

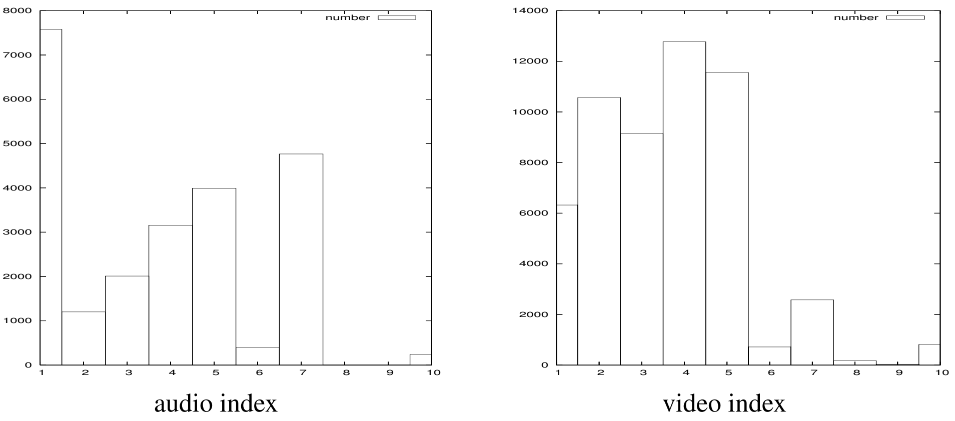

Figure 1: Number of frames for each character appearance, on a TV talk show.

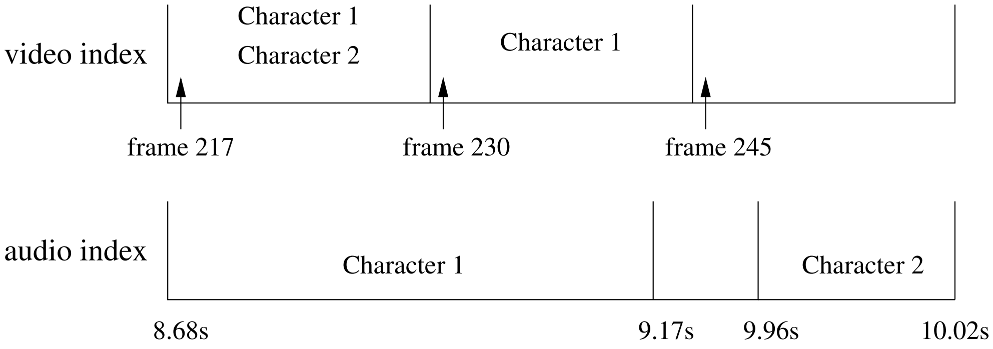

Figure 2: Direct comparison between audio and video indexes.

m_{ij} times. Conversely, in all the frames where the character V_j is present, the voice A_i is heard m_{ij} times.

An intuitive idea would be to sort this matrix by rows (or by columns). However, this solution is often wrong, because it makes the assumption that while a voice is heard, its corresponding visual feature is the most present in the frames (sorting by rows). Sorting by columns would mean that for each visual feature its corresponding voice is the most heard while the feature appears.

In real TV shows, this assumption is often wrong. For instance, in TV games or TV talk shows, the character who speaks the most is usually the presenter. In this case, his voice can be the most heard even when the players appear on screen. Thus, even if this intersection matrix is interesting for the fusion of audio and video, it cannot be directly used. A postprocessing is required: it will be presented in the next section.

3.2 Index fusion

With some special contents, like TV talk shows and TV games, the matrix m can be directly read if we have some prior information about the characters. For example, in a TV talk show, if a character is assumed to be the presenter, his voice is the most heard when he appears on screen (which is wrong for a guest). Conversely, if the character is assumed to be a guest, his visual feature is the most seen while he is speaking (which is wrong for a presenter).

With real data, when there is no learning stage we cannot have prior information about this “class” of characters, which makes direct reading of the matrix m impossible because we cannot determine for each character if the matrix must be sorted by rows or by columns. So, we propose in this section to read m both by rows and by columns, and to keep the most significant information.

To carry out this fusion, we compute two new matrices, m_a and m_v, where the frame numbers are replaced with

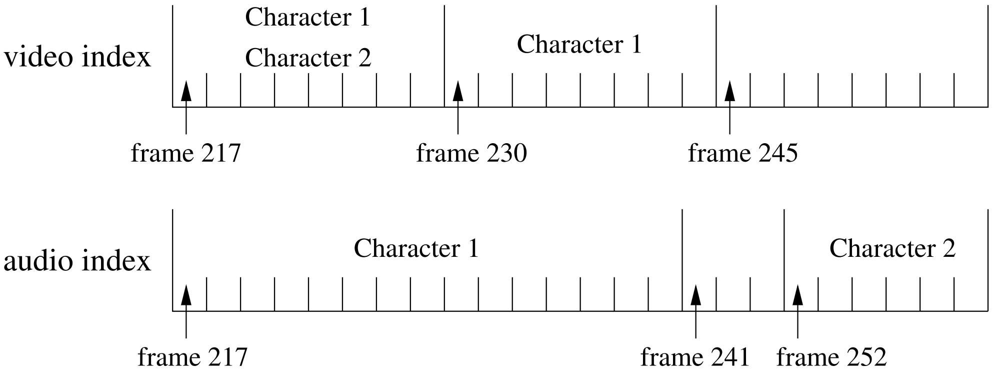

Figure 3: Normalization of the indexes of figure 2.

percentage by rows and by columns:

$$m_a = \begin{array}{cc} & \begin{array}{cccc} V_1 & V_2 & \cdots & V_{n_v} \end{array} \\ \begin{array}{c} A_1 \\ A_2 \\ \vdots \\ A_{n_a} \end{array} & \left(\begin{array}{cccc} f^a_{11} & f^a_{12} & \cdots & f^a_{1n_v} \\ \hline f^a_{21} & f^a_{22} & \cdots & f^a_{2n_v} \\ \hline \cdots & \cdots & \cdots & \cdots \\ f^a_{n_a1} & f^a_{n_a2} & \cdots & f^a_{n_an_v} \end{array}\right) \end{array} 100\,\% \tag{2}$$

$$m_v = \begin{array}{cc} & \begin{array}{cccc} V_1 & V_2 & \cdots & V_{n_v} \end{array} \\ \begin{array}{c} A_1 \\ A_2 \\ \vdots \\ A_{n_a} \end{array} & \left(\begin{array}{c|c|cc} f^v_{11} & f^v_{12} & \cdots & f^v_{1n_v} \\ f^v_{21} & f^v_{22} & \cdots & f^v_{2n_v} \\ \cdots & \cdots & \cdots & \cdots \\ f^v_{n_a1} & f^v_{n_a2} & \cdots & f^v_{n_an_v} \end{array}\right) \\ & 100\,\% \end{array} \tag{3}$$

Matrix m_a gives the probability density of each voice A_i, whereas m_v gives the one of each visual feature V_j. From these matrices, we define the fusion matrix F, by computing, for each pair (i, j), a fusion between f^a_{ij} and f^v_{ij} with a fusion operator like maximum, mean or product. If we note $C(A_i, V_j)$ the fusion coefficient between A_i and V_j, expression of matrix F is given by:

$$F = \begin{pmatrix} C(A_1, V_1) & \cdots & C(A_1, V_{n_v}) \\ C(A_2, V_1) & \cdots & C(A_2, V_{n_v}) \\ \cdots & \cdots & \cdots \\ C(A_{n_a}, V_1) & \cdots & C(A_{n_a}, V_{n_v}) \end{pmatrix} \tag{4}$$

This matrix F can be directly used to realize the association. When the number of voices and visual features is the same ($n_a = n_v$), it is read equally by rows or by columns. For instance, for each row i, we search the column j which provides the maximum value: then the voice A_i is automatically associated to the visual feature V_j.

In the next section, we present some experiments to illustrate this fusion algorithm. Moreover, we show how to deal with videos where the numbers n_a and n_v are different.

4 Experiments

To experiment this algorithm, we manually indexed audio and video channels of a broadcast, and ran several experimentations. In section 4.1 we present the corpus that we used. Then, the experiments are divided in two parts. First, in section 4.2 we show the results of audio/video associations with manual indexes. Then, section 4.3 deals with the problem of oversegmentation.

4.1 Corpus

To estimate the accuracy of the fusion algorithm, we made several tests on a french TV game. This game lasts thirty one minutes, which provides 46 464 frames. The format of this video is MPEG-1, with a frame size of 352 × 288. In order to obtain the ground truth we manually indexed this game by describing, in each frame, all the characters who are present. This annotation was made in both audio and video channels. For both channels, we used the same labels to identify the characters. Finally, we obtained ten characters who visually appear in the video, and eight voices. This difference is due to two characters who never speak, thus they cannot be associated to any voice.

Then, we use this annotated corpus for the evaluation of our algorithm. First, we process these manual indexes to check the results of the audio/video association. If the association is perfect, each voice should be associated with the visual feature having the same label.

Secondly, we process the manual audio index with an automatically generated video index, and try to solve the oversegmentation problem. Regarding the automatic video index, we can find in literature many methods for video segmentation [1]. In our application, we implement a costume segmentation, which produces the video index using the clothes of the characters (for more details, the reader can refer to [2]). The various characters of the manual index are presented in figure 4.

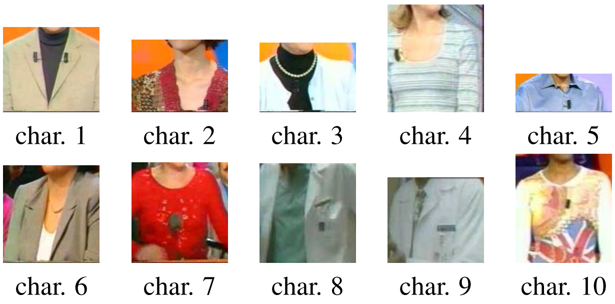

Figure 4: Costumes of the main characters in the video sequence for the manual index. For each character who speaks at least one time during the game, his corresponding voice is associated.

costume	1	2	3	4	5	6	7	8	9	10
voice	1	2	3	4	5	6	7	-	-	10

Table 1: Automatic voice/image association with the "product" operator, for the two manual indexes.

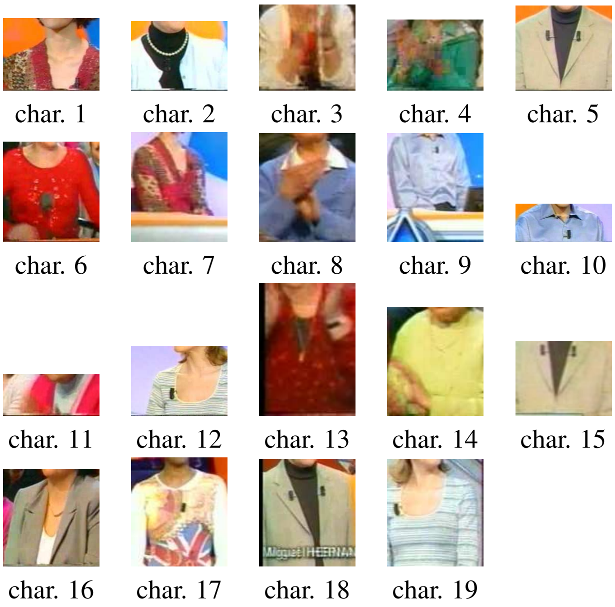

Figure 5: Costumes of the characters in the automatic video index.

4.2 Audio/video association

First, we only process the manual indexes, to check the automatic association. As each character has the same label for audio and video indexes, each costume i should be assigned to the voice i. This first step is required, because it would not be conceivable to use an algorithm on noisy indexes if it does not work with perfect indexes.

We computed the matrix F, using the "product" as a fusion operator. Figure 6 shows the value of F on this example. As F is a square matrix (because the set of labels is the same for both indexes), it could be read equally by rows or by columns. In this case, we can use another approach: we look for the maximum value in this matrix, which is 0.63 (line 6, column 6). Thus, the voice 6 is associated with the costume 6. Then, we delete this row and this column, and we repeat this search, to obtain another association, until having an empty matrix. Results are given in table 1. We can notice that each voice is correctly assigned to its corresponding costume, and costumes 8 and 9 are not associated to any voice (this was foreseeable because they corresponds to characters who never speak).

4.3 Resolution of the oversegmentation problem

We presented in the previous paragraph two correctly segmented indexes. We now replace one of them by an automatic index. If we take for example the costume index: each character presents several models, as seen in figure 5. In [2], this problem was solved manually by affecting his real name to each detected person. Thus, all models of the same person were merged under the same label. With our audio/video fusion algorithm, each costume model is associated to the character voice which corresponds to him. The idea is to use this information to propose an automatic fusion of the oversegmented index. After computing matrix F, given in formula (4), we obtain for each costume model the name of the person who speaks during his appearance.

To test our fusion system on real data, we compute and compare accuracy results on our corpus when the oversegmentation problem is solved manually and automatically. We have two oversegmentation examples (noted *index1* and *index2*) resulting from the costume detection system [2], using different thresholds. In the *index1*, 19 different costumes are detected (they are shown in figure 5) and 32 in the *index2*. For each of them, the corresponding fusion matrix F is read by columns, in order to associate a voice to each costume. The automatic associations (costume/voice) are given respectively in tables 2 and 3.

In table 2, we can notice that costumes 5, 15 and 18 are corresponding to the same voice: thus they will be grouped (see figure 7). In some cases, no voice (noted "-" in previous tables) is associated with some costumes: the costume is regarded as "not assigned" (noted "NA"). In table 2, it is the case for the costumes 11 and 16. However, the number of NA is very weak, as seen in table 4. By using the same evaluation method as [2], we obtain the results summed up in table 4.

$$F_{\text{product}} = \begin{pmatrix} 0.55 & 0.25 & 0.19 & 0.20 & 0.16 & 0.12 & 0.07 & 0.00 & 0.00 & 0.16 \\ 0.01 & 0.28 & 0.10 & 0.15 & 0.12 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.01 & 0.31 & 0.50 & 0.02 & 0.02 & 0.00 & 0.02 & 0.00 & 0.00 & 0.00 \\ 0.03 & 0.20 & 0.17 & 0.34 & 0.22 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.04 & 0.01 & 0.00 & 0.34 & 0.51 & 0.04 & 0.09 & 0.00 & 0.00 & 0.00 \\ 0.07 & 0.00 & 0.00 & 0.05 & 0.05 & \mathbf{0.63} & 0.00 & 0.00 & 0.00 & 0.12 \\ 0.11 & 0.08 & 0.03 & 0.12 & 0.13 & 0.00 & 0.54 & 0.00 & 0.00 & 0.13 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.27 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.42 \end{pmatrix}$$

Figure 6: Fusion matrix computed with the product operator, for the two manual indexes.

voice	1	2	3	4	5	6	7	10	-
costume	5,15,18	1,7	12,19	2	9,10	3,4,8,13,14	6	17	11,16

Table 2: Automatic voice/image association with the “product” operator, for the *index1*.

voice	1	2	3	4	5	6	7	10	-
costume	6,16,19,22,29,30	1,9	15	2,28	12,13,21,23,27	11	7,8,10,17,18,24	5,25	3,4,14,20,26,31,32

Table 3: Automatic voice/image association with the “product” operator, for the *index2*.

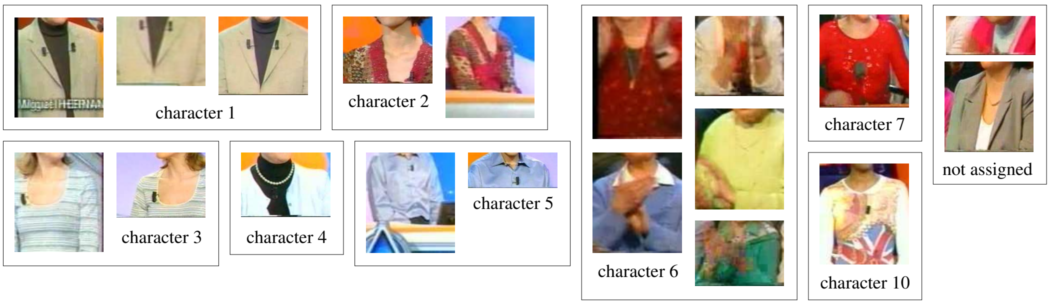

Figure 7: Automatic fusion of the costumes which corresponds to the same character. The label “character 6” is associated with the audience noise. That is why the audience members have the same label.

index	fusion	class 1	class 2	class 3	recognition errors	false alarm detection	NA
index1	manual	94.87 %	5.86 %	3.87 %	2.42 %	1.12 %	-
	auto	95.56 %	6.46 %	3.61 %	1.12 %	1.14 %	0.41 %
index2	manual	94.52 %	7.08 %	3.99 %	0.21 %	1.29 %	-
	auto	94.52 %	6.85 %	3.78 %	0.54 %	1.19 %	0.84 %

Table 4: Recognition result comparison, according to the oversegmentation resolution problem (manual or automatic).

The various classes correspond to the size of the character in the frame, as explained in [2]. The first one is a centered character who has sufficient size to be the most important visual interest in the frame whereas the third one corresponds to background characters. The automatic fusion results do not take into account "NA". We notice that the results obtained with an automatic fusion are as good as the manuals (even better in some cases).

5 Conclusions and directions of further research

We proposed in this paper a framework for automatic fusion of audio and video indexes. We showed that speech and visual features can be automatically associated using a common index scheme. If both audio and video indexes are correctly segmented, i.e. without oversegmentation, this automatic association yields good results as shown in the experiment section. This association can be used as a preprocessing step for existing applications, as [7, 8]. In order to deal with oversegmentation, we proposed an approach which uses one index to improve the segmentation of the other. For example, we showed that using a correctly segmented audio index, it is possible to improve an oversegmented video index.

For further research, we first plan to analyze the real case where both audio and video indexes are oversegmented. At the moment, we did not have an automatic application of speaker segmentation, that is why we could not try to merge two oversegmented indexes. Moreover, we also plan to work on different kinds of audiovisual contents, for example some TV broadcasts that contain a voice over. In this case, it would be very interesting to study how to take a character who never appears in any frame into consideration. Finally, we think that audio/video fusion can be a step to automatically determine the function of the characters in a TV broadcast, for instance to automatically determine in a TV game or in a TV talk show who is the presenter or who are the guests.

References

[1] C. Snoek and M. Worring, "Multimodal Video Indexing: A Review of the State-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, Jan. 2005.

[2] G. Jaffré and P. Joly, "Costume: A New Feature for Automatic Video Content Indexing," in *Coupling approaches, coupling media and coupling languages for information retrieval (RIAO)*, Avignon, France, Apr. 2004, pp. 314–325.

[3] B. Zhou and J. H. L. Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian Information Criterion," in *Proceedings of the International Conference on Spoken Language Processing*, 2000.

[4] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-Visual Automatic Speech Recognition: An Overview," in *Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. E. Perrier, Eds. MIT Press, 2004.

[5] E. Kijak, G. Gravier, L. Oisel, and P. Gros, "Audio-visual Integration for Tennis Broadcast Structuring," in *International Workshop on Content-Based Multimedia Indexing*. Rennes, France: GDR-PRC ISIS, Sept. 2003, pp. 421–428.

[6] A. Albiol, L. Torres, and E. J. Delp, "Combining Audio and Video for Video Sequence Indexing Applications," in *IEEE International Conference on Multimedia and Expo*, vol. 2, Naples, Italy, July 2002, pp. 353–356.

[7] Y. Li, S. Narayanan, and C. Jay Kuo, "Adaptive speaker identification with audiovisual cues for movie content analysis," *Pattern Recognition Letters*, vol. 25, no. 7, pp. 777–791, May 2004.

[8] C. Taskiran, A. Albiol, L. Torres, and E. J. Delp, "Detection of unique people in news programs using multimodal shot clustering," in *Proceedings of the International Conference on Image Processing*, Singapore, Oct. 2004.

[9] A. Albiol, L. Torres, and E. Delp, "Two are better than one: when audio comes to the rescue of video," in *Proceedings of the 5^{th} European Workshop on Image Analysis for Multimedia Interactive Services*, Lisbon, Portugal, Apr. 2004.

[10] S. Tsekeridou and I. Pitas, "Content-Based Video Parsing and Indexing Based on Audio-Visual Interaction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 4, pp. 522–535, Apr. 2001.