# TALKING FACES INDEXING IN TV-CONTENT

*Meriem Bendris*[1,2]*, Delphine Charlet*[1]*, Gérard Chollet*[2]

[1] France Télécom *R&D* - Orange Labs, France
[2] CNRS LTCI, TELECOM-ParisTech, France
{meriem.bendris,delphine.charlet}@orange-ftgroup.com
gerard.chollet@telecom-paristech.fr

## ABSTRACT

*Our objective is to index talking faces in a TV-Context: build a description of TV-content, in terms of talking people, without any pre-defined dictionary of identities. In TV-content, because of multi-face shots and non-speaking face shots, it is difficult to determine which face is speaking. In this work, a method is proposed which clusters people independently by the audio and by the visual information and combines these clusterings of people (audio and visual) in order to detect sequences of talking faces. The audio indexing system is based on agglomerative clustering with the Bayesian Information Criterion. The visual indexing system is based on costume detection and clustering of color histograms. The combination of both indexes is based on searching for the best match between both clusterings, to obtain a correspondence between the automatic audio labels and the automatic video labels. The talking faces are then determined by the intersection of the segments of the associated audio and video labels. Results of experiments on a TV-Show database show that a high correct detection rate can be achieved by the proposed method.*

***Index Terms***— Talking faces indexing, speaker clustering, video clustering, audio-visual indexing.

## 1. INTRODUCTION

With the increase of internet use, a proliferation of multimedia content (video on Demand, TV websites interfaces) is observed. It is necessary to develop technologies to facilitate access to this multimedia data. One way is to use audio-visual indexing of people, allowing a user to locate sequences of a certain person. In particular, we are interested in locating sequences in popular TV-programs in which a certain person is speaking and visible. In the literature, detecting people in video content is done using two different kinds of information (Audio and Visual) [1]. In the audio modality, most approaches are based on speech detection and speaker recognition [2]. For the visual case, most approaches separate the problem in 2 steps [3]: face detection [4] and face recognition [5]. There are also methods based on the exploitation of both modalities to correctly identify people in video [6]. In our case, the objective is describe TV-content in terms of talking faces, without any pre-defined dictionary of people. Structuring content in terms of people in a TV-context is a difficult problem due to many ambiguities in audio, in video and in their association. First, in the audio modality, speech is spontaneous, speaker turns can be very short, and often people are speaking at the same time, making speaker analysis very difficult. Secondly, concerning the visual modality, faces appear with many variations in lighting conditions, position and facial expressions, also making accurate face analysis difficult. Finally, associating audio and visual information in TV-context introduces many ambiguities in the case of multi-face shots, or shots where the speaker face is not detected because it is partially visible or not at all filmed.

In this paper, a method of indexing talking faces in a TV-Context based on the fusion of the results of audio and visual clustering is proposed. First, audio index of people is build through speaker clustering and visual index of people is build through visual information clustering. Then, a correspondence is found between the set of audio labels and the set of visual labels. Finally, the talking faces structure is given by the intersection of the segments of the associated audio and video labels.

This paper is organized as follows: in section 2, methods of clustering people using the audio and visual information, and their combination are presented. Section 3 presents the TV-Show database with the annotation methods employed. Finally, experiments are reported in section 4.

## 2. TV-CONTENT CLUSTERING METHODS

### 2.1. People clustering using audio information

In this section, the objective is to automatically build an index of the TV-content based on speaker turns, without any pre-defined set of audio labels (identities of speaker). First, the audio signal is segmented in speech/non speech parts, and each speech segment is segmented in speaker turns, so as to get segments which are supposed to contain only the speech of one speaker. The segmentations are performed according to the algorithms described in [7] and were kindly provided

by *C.Barras* from LIMSI-CNRS.

Then, speaker clustering is performed: it consists in merging together speech segments that are supposed to come from the same speaker. The clustering is performed according to an agglomerative clustering based on Bayesian Information Criterion (BIC), which is one of the most robust speaker clustering method [7]. The BIC is a criterion that measures how well a model fits to the data, by combining the likelihood of the data on the model and a penalty factor about the complexity of the model. Here, the model is assumed to be a monogaussian on the acoustical coefficients (cepstral coefficients). When comparing two segments $X_i$ and $X_j$, the variation of the BIC ($\Delta_{BIC}$) is measured when the 2 models (one for each segment) are replaced with one model (obtained with the fusion of the segments $X_{i\cup j}$):

$$\begin{aligned}\Delta BIC(X_i, X_j) = \frac{1}{2}((n_i + n_j)\log(|\Sigma_{i\cup j}|) \\ -n_i \log(|\Sigma_i|) - n_j \log(|\Sigma_j|)) \\ -\frac{\lambda}{2}(d + \frac{d(d+1)}{2})\log(n_i + n_j)\end{aligned}$$

where $n_i$, $n_j$ are the numbers of frames of segments $X_i$ and $X_j$, $\Sigma_i$ (resp. $j$, $i \cup j$) is the full covariance matrix computed for segment $X_i$ (resp. $X_j$, $X_{i\cup j}$), $d$ is the number of acoustical parameters per frame, and $\lambda$ is a tuning factor (equal here to 1). The segments that give minimal loss according to the $\Delta_{BIC}$ are then clustered together in the agglomerative process. This criterion also provides a stopping criterion in the clustering: when the loss is above a certain threshold (here 0) the clustering is stopped.

### 2.2. People clustering using the visual information

In this section, the objective is to automatically build an index of the TV-content based on people appearance, without any pre-defined set of visual labels (identities of people). In the literature, there are several methods of structuring audiovisual document by person using the visual information. Most of these methods are based on face detection and recognition [3, 1]. In TV-context, because of the low quality of face appearance, it is very difficult to detect and cluster faces with a great reliability. We chose to build the index by grouping all shots in which the same person appears using a signature of costume's colors. A shot with several people can belong to several sets. This technique assumes that within a single document, there is a bijection between people and their costumes.

#### *2.2.1. Costume detection*

One of the most popular methods used to detect costumes from an image is to look for the person who wears it [8]. First, faces are detected using the OpenCV implementation of the *Viola&Jones* face detection algorithm [4]. Next, a rectangle under the detected face determine the costume (see example in the figure 1). This rectangle is proportional to face size( $\times 3.6$ width face and $\times 1.5$ height of face). This size is chosen to take the best costume area without background. The features vector is represented by the concatenation of the color histogram of the costumes.


**Fig. 1**. Example of costume detection using the face

#### *2.2.2. Costume clustering*

Most approaches for audio and video clustering are based on agglomerative hierarchical algorithms in which clusters are constructed by combining iteratively the closest elements. In this study, the Ward's method is used. Initially, each element is associated to a cluster. At each step, all combinations of clusters are studied, the two elements which present the minimum information loss are grouped. Information loss between two clusters $A$ and $B$ is calculated as follows:

$$\begin{aligned}\Delta(A, B) = \sum_{i\in A\cup B} ||X_i - \bar{X}_{A\cup B}||^2 - \sum_{i\in A} ||X_i - \bar{X}_A||^2 \\ - \sum_{i\in B} ||X_i - \bar{X}_B||^2\end{aligned}$$

where $\bar{X}_{A\cup B}$ is the centroid of the cluster $A \cup B$. In clustering methods, the elements (costumes of people detected in shots) are grouped without taking into account temporal information. However, costumes detected in the same shot should not be associated to the same person. Thus, the clustering algorithm is modified so as to make impossible the merging of costumes coming from the same shot.

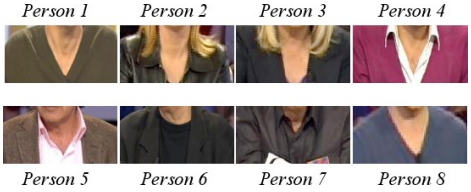


**Fig. 2**. Examples of costumes in the *Show1*

Note that this method may introduce some errors in costume clustering in case of similar costumes in terms of color histograms. Figure 2 shows an example of typical costume for each individual in the same show.

### 2.3. Audio-visual structuring

Our goal is to index talking faces in TV-content. To accomplish that, first, two indexes of person are constructed independently using audio and visual information. After that, each visual cluster is associated to an audio cluster as follows: explore all possible combinations and select one that maximizes the total duration of the intersection of associated pairs. This method assumes that the most frequently visible person when someone is speaking is the speaker himself. After associating each audio cluster to a visual one, sequences of talking faces are obtained by the intersection of associated clusters segments. This method does not require same numbers of clusters for each modality. In our case, some clusters in one modality may not be associated to any cluster in the other modality.

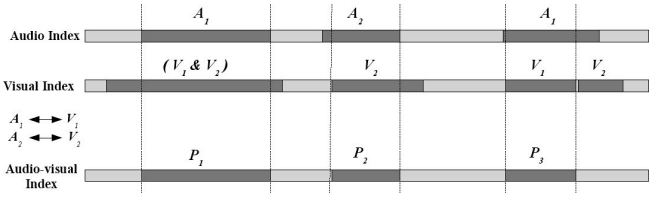


**Fig. 3**. Example of indexes fusion

Figure 3 shows an example of the method of obtaining the talking faces sequences from the audio index of person and the visual one after association of the audio-person $A_1$ to the visual person $V_1$ and the audio-person $A_2$ to the visual person $V_2$. The segment $(V_1 \& V_2)$ means that two faces ($V_1$ and $V_2$) are detected simultaneously in the same shots.

## 3. TV-SHOW DATABASE

Few works have been focused on real TV-Context database. To our best knowledge, there are no public data annotated with both voice and facial appearance. It was necessary to collect and annotate a real TV-Context database.

### 3.1. Presentation

In order to have a large number of examples, a TV-Show in which people appear often is selected. The experiments are done on the live TV-Show *"on n'a pas tout dit"*, a French TV-program presented by *Laurent Ruquier* on the public channel "France 2" broadcast between September 2007 and July 2008 from Monday to Friday at 7PM. Many commentators discussed, with some celebrities, the news of the moment in a good mood. Four shows are annotated manually as follows:

- *Audio annotation: transcriber* format files *XML* containing information from beginning and end of each speech sequence, the identity of the speaker and the spoken text. The applause and overlapping of two or more speakers are also annotated with text if it is understandable.
- *Video annotation:* each participant (anchor, commentators and guests) was annotated from the time he appears to the shot end. The information manually annotated is the identity of the person, coordinates of the face region in the shot, and the position of the face relative to the camera (Right, left, front, quarter right, quarter left,top, bottom and the face occultation). The tool used for the audio annotation is the *Elan*[1] software.

**Fig. 4**. TV-Show Database - examples of shots collected

Figure 4 shows typical examples of shots in the TV-database. There are several types of shots: focus on one face, multi-faces and public shots, general around table and edited shots between people.

### 3.2. Corpus analysis

The duration of each show is about 50 minutes where each person intervenes at different times. During the show, the anchor define the topics of conversation separated by jingles. There are also sequences of reports, clips and generic. One show contains typically ten personalities. The figure 5 summarizes general statistics of the TV-database by show. In TV-context, shots are very short, dialogues are interactive and several people appear in the same shot. The total duration of people annotated by the visual information is larger than the duration of the show. This is due to the fact that in one shot, several people can appear, which involves counting the segment as many times as the number of people who appear in it.

| *Shows* | *S1* | *S2* | *S3* | *S4* |
|---|---|---|---|---|
| *Number of person* | 8 | 9 | 7 | 7 |
| *Total duration of speaking(sc)* | 2347 | 2568 | 2014 | 2288 |
| *Total duration of faces(sc)* | 3548 | 3720 | 3668 | 3049 |
| *Total duration of talking faces(sc)* | 1409 | 1505 | 981 | 1456 |

**Fig. 5**. Corpus analysis - general statistics

In shows, people speak and appear in a structured way. The figure 6 summarizes the distribution of speaking duration and appearance for each personality in the *show1*. In audio reference of each show, except the anchor which occurs very often (175 turns in the *show1*), each person speaks approximatively 35 times, each with an average duration of $6sc$. In

[1] http://www.lat-mpi.eu/tools/elan/download

| | *SpeakingDuration (#speaker turns)* | *Appearance Duration (#visual shots)* | *Talking Faces Duration (#talking face segments)* | *Non-visible Speaker Duration ( % on time of speaking)* | *Silent Face Duration (% on time of appearance)* |
|---|---|---|---|---|---|
| *Person 1 (anchor)* | 918 (175 turns) | 812 (297 shots) | 471 (188 Seg) | 447 ( 48.7%) | 341 ( 42.0%) |
| *Person 2* | 193 (41 turns) | 383 (126 shots) | 133 (45 Seg) | 59 (30.8%) | 249 ( 65.1%) |
| *Person 3* | 240 (56 turns) | 545 (174 shots) | 162 (64 Seg) | 78 (32.4%) | 382 (70.2%) |
| *Person 4* | 304 (38 turns) | 529 (160 shots) | 190 (52 Seg) | 114 (37.5%) | 339 (64.1%) |
| *Person 5* | 252 (38 turns) | 240 (69 shots) | 156 (62 Seg) | 96 (38.1%) | 84 (34.9%) |
| *Person 6* | 119 (25 turns) | 395 (152 shots) | 80 (26 Seg) | 39 ( 32.6%) | 315 (79.6%) |
| *Person 7* | 239 (28 turns) | 483 (150 shots) | 163 (49 Seg) | 76 (31.7%) | 320 (66.2%) |
| *Person 8* | 81 (17 turns) | 163 (53 shots) | 52 (22 Seg) | 29 (35.4%) | 110 (67.6%) |

**Fig. 6**. Corpus analysis - speaking and appearance duration (sc) for each person in the *show1*

visual reference, each person (except the anchor) appears approximately in $140$ shots, each with an average duration of $3sc$. The anchor appears more than other people ($297$ times in *show1*). In speaking face reference of each show, except the anchor which occurs very often ($447$ times in the *show1*), each person speaks and appears simultaneously approximatively $70$ times. The average duration of a speaking face shots is $2.7sc$. For each person, the face associated to a voice is visible more than $60\%$ of his/her speaking time, whereas the speaking time of a visible face is about $35\%$ of the total duration of the appearance of this face. Thus, for these TV-shows, the probability that a speaker is visible is much higher (almost twice as much) than the probability that a face is speaking.

## 4. RESULTS AND DISCUSSION

### 4.1. Evaluation

The evaluation is done using the available tools proposed by NIST for speaker clustering(*SpkrSegEval-v23.pl*), and also used in the ESTER evaluation campaign (Rich Transcription of French Radio Broadcast) [9]. This tool enables to find the best match between the set of reference labels and the automatic labels obtained by an automatic segmentation and clustering. From this match, the following metrics are computed:

- $\mathrm{FalseAlarmDuration}$ : total duration of people automatically detected but not referenced. In audio clustering, it corresponds to the duration of speech segments detected automatically and annotated as non speaker in the reference. These segments are in fact non-speech or multi-speakers turns.
- $\mathrm{MissedDuration}$ : total duration of people referenced but not detected automatically.
- $\mathrm{ClusterDuration}$ : total duration of people detected automatically.
- $\mathrm{ReferenceDuration}$ : total duration of people annotated manually.
- $\mathrm{ErrorDuration}$ : total duration of segments associated to the wrong identity.
- $\mathrm{CorrectDuration}$ : total duration of segments associated to the correct identity.

As we attach more importance to the reliability of the talking faces indexing systems than to their ability to detect all talking faces in a show, we propose to focus on the cluster composition IC to specifically evaluate our indexing system; thus following rates are computed:

$$\mathrm{FalseAlarmRate(FAR)} = \frac{FalseAlarmDuration}{ClusterDuration}$$
$$\mathrm{ErrorDurationRate(EDR)} = \frac{ErrorDuration}{ClusterDuration}$$
$$\mathrm{CorrectDurationRate(CDR)} = \frac{CorrectDuration}{ClusterDuration}$$
$$\mathrm{ClusterComposition(IC)} = CDR + EDR + FAR$$

The undetected segments are calculated in the metric MDR (Missed Duration Rate) as follow:

$$\mathrm{MissedDurationRate(MDR)} = \frac{MissedDuration}{ReferenceDuration}$$

Two types of evaluation are done: the first one is the *Full Evaluation* where all the duration of the show is taken into account to evaluate our system. The second evaluation is the *Restricted* one, done on a selected part of the show: first, a window of $0.25$ seconds is removed in the segment borders (in the reference and test) in order not to count as error a small lags ($< 0.25s$) between reference borders and automatic borders. Second, segments not referenced as speaker segments are removed from the automatic outputs, and thus are not taken into account in evaluation. These segments are annotated as non-speech, overlapping speech, reports... This restriction is usually performed in speaker clustering in order to focus on one specific problem and not be masked by another problem: here, to focus on speaker clustering and not on errors in speaker clustering due to erroneous speech detection for instance. Thus, per definition, there cannot be any false alarm in restricted evaluation (except in the fusion of audio and video, see section 4.4). In video reference, only the *Restricted Evaluation* is done (shots where no personality is annotated are removed) because, in TV-Context, many

shots are general or public views making face detection more complex in terms of time computing. Thus, automatic face detection was restricted to the shots annotated by a presence of a personality. In the *Restricted Evaluation*, the decrease of the total duration of reference time due to the removal of windows is $2 \times 0.25sc$ for each segment. Thus, the loss in total duration of reference time is $0.5\times$ of number of the segments, which makes an amount of about $200 - 300sc$ per show for audio clustering.

### 4.2. Audio Clustering Results

Figures 7 and 8 summarize the results of the audio-clustering system by show in the TV-Database using the two evaluation methods. In the first evaluation (*Full Evaluation*), the duration correctly associated to the right audio-person varies from $43.1$ to $68.7\%$ according to the shows. The missed duration rate varies also widely (from $5.4$ to $12.4\%$) and the audio clustering system outputs between 16 and $21\%$ of false alarms which is relatively high. These variations are due to the interactivity of dialogues in this context which multiplies the ambiguous speech segments. In particular, the *CDR* in the show 3 is $40.2\%$ and the *MDR* is $16.7\%$, because this show has more short interactions between people and overlapping speech than the other shows.

| | *IC = CDR + EDR + FAR* | *MDR* | *Correct/TotalCluster/Reference(sc)* |
|---|---|---|---|
| *S1* | $59.8\% + 18.8\% + 21.4\%$ | $5.4\%$ | $1687/2822/2347$ |
| *S2* | $54.0\% + 28.8\% + 17.2\%$ | $6.3\%$ | $1570/2907/2568$ |
| *S3* | $43.1\% + 40.2\% + 16.7\%$ | $12.4\%$ | $913/2199/2014$ |
| *S4* | $68.7\% + 19.6\% + 16.7\%$ | $9.8\%$ | $1579/2480/2288$ |

**Fig. 7**. Audio-Clustering - *Full Evaluation*

In the second evaluation (see figure 8), the Correct Duration Rate (*CDR*) is improved compared to the *full Evaluation*, as the non-speech or ambiguous parts that might be in the clusters are discarded from evaluation. As conclusion, the audio clustering system is not robust to speech ambiguities in this context but when the audio is clear, the accuracy of the output of the proposed method is acceptable.

| | *IC = CDR + EDR + FAR* | *MDR* | *Correct/TotalCluster/Reference(sc)* |
|---|---|---|---|
| *S1* | $78.7\% + 21.3\% + 0.0\%$ | $4.9\%$ | $1602/2036/2140$ |
| *S2* | $67.9\% + 32.1\% + 0.0\%$ | $5.8\%$ | $1502/2214/2349$ |
| *S3* | $55.9\% + 44.1\% + 0.0\%$ | $11.6\%$ | $835/1494/1690$ |
| *S4* | $79.7\% + 20.3\% + 0.0\%$ | $9.6\%$ | $1512/1897/2098$ |

**Fig. 8**. Audio-Clustering - *Restricted Evaluation*

### 4.3. Visual Clustering Results

#### *4.3.1. Protocol*

In a given shot, face detector is applied in all the frames, then a costume is obtained for each detected face. Each person detected is associated to a collection of costumes detected in all the frames of the shot. The centroid costume of the collection (in term of color histograms correlation) is selected as the representative costume of the shot, to initialize the clustering process. No stopping criterion was developed. In our clustering method, the number of people annotated in the show is considered as known a priori. So the clustering is stopped when the number of clusters reaches the number of people.

#### *4.3.2. Results*

Figure 9 summarizes the results of the visual-clustering system by show in the TV-Database. The performance obtained are quite good, with an error duration rate between 9 and $22\%$. These errors can be explained by the fact that in these TV-shows, sometimes two people are dressed with similar costumes (see the figure 2). The missed duration rate varies from $8$ to $15\%$ explained by shots where people are not detected automatically.

| | *IC = CDR + EDR + FAR* | *MDR* | *Correct/TotalCluster/Reference(sc)* |
|---|---|---|---|
| *S1* | $86.15\% + 13.85\% + 0.0\%$ | $15.07\%$ | $2192 / 2544 / 2958$ |
| *S2* | $81.68\% + 18.32\% + 0.0\%$ | $8.00\%$ | $2285 / 2798 / 3022$ |
| *S3* | $90.11\% + 9.89\% + 0.0\%$ | $12.19\%$ | $2426 / 2692 /3030$ |
| *S4* | $77.86\% + 22.14\% + 0.0\%$ | $15.57\%$ | $1692 / 2173 / 2556$ |

**Fig. 9**. Visual-Clustering - *Restricted Evaluation*

### 4.4. Audio-Visual Clustering Results

For each show, a reference index of talking faces is obtained by the intersection of the audio and visual annotation as described in the figure 3. A speaking face segment $S_i$ is composed on the cluster $A_i$ from the audio and cluster $V_i$ from the visual information.

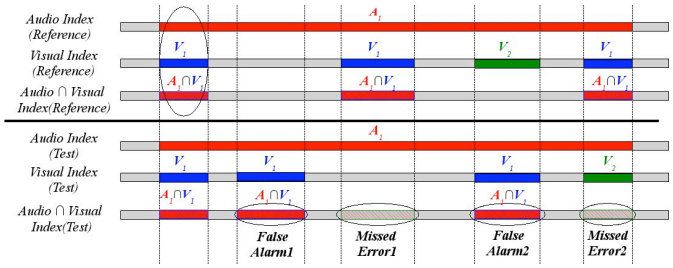


**Fig. 10**. Example of errors in audio-visual clustering

In a speaking face evaluations, a segment is identified as a false alarm when audio or visual clustering makes a false alarm or both (example *False Alarm1*). An output segment is also a false alarm in case of a labeling error that favors the association (*False Alarm2)*. A missed speaking face error occurs when the person is not detected by the audio or by the visual information or by both (*Missed Error1*). An output segment is also missed in case of a labeling error without

favoring an association (*Missed Error2)*. The figure 11 summarizes the results of the audio-visual clustering system by show in the *Full Evaluation*. In the *Full Evaluation*, FAR is relatively high (between 8 and 33%). This rate reflects the impact of false alarms from the two Clustering systems in the intersection step. A consistent part of talking faces segments is lost as a result of non-detection or labeling errors made by audio and visual systems.

| | *IC = CDR + EDR + FAR* | *MDR* | *Correct/TotalCluster/Reference(sc)* |
|---|---|---|---|
| *S1* | $78.6\% + 3.5\% + 17.9\%$ | $43.8\%$ | 758 / 965 / 1409 |
| *S2* | $71.6\% + 2.9\% + 25.5\%$ | $45.6\%$ | 786 / 1097 / 1505 |
| *S3* | $53.9\% + 12.6\% + 33.5\%$ | $48.9\%$ | 406 / 754 / 981 |
| *S4* | $87.8\% + 4.1\% + 8.1\%$ | $53.4\%$ | 649 / 739 / 1456 |

**Fig. 11**. Audio-Visual Clustering - *Full Evaluation*

Figure 12 summarizes the results of the audio-visual clustering system by show using in the *Restricted Evaluation*. By removing audio ambiguous segments, audio clustering system outputs less false alarms (see section 4.2.). This decrease is reflected in the FAR of the Audio-visual system which decreased significantly for each show compared to the first evaluation. The origin of false alarm comes from labeling error in favor of an association (case of *False Alarm2* in figure 10). Unfortunately, the missed duration rate does not change because it comes from the labeling errors committed by both audio and visual clustering systems. In the evaluation *Restricted Evaluation*, FAR reduces, the correct duration increases but the duration of talking faces returned decreases.

| | *IC = CDR + EDR + FAR* | *MDR* | *Correct/TotalCluster/Reference(sc)* |
|---|---|---|---|
| *S1* | $90.3\% + 3.6\% + 6.1\%$ | $41.6\%$ | 659 / 730 / 1174 |
| *S2* | $87.6\% + 3.2\% + 9.2\%$ | $44.0\%$ | 668 / 762 / 1235 |
| *S3* | $66.3\% + 13.4\% + 20.3\%$ | $45.8\%$ | 334 / 503 / 739 |
| *S4* | $91.8\% + 3.7\% + 4.5\%$ | $51.3\%$ | 569 / 620 / 1216 |

**Fig. 12**. Audio-Visual Clustering - *Restricted Evaluation*

## 5. CONCLUSION AND PERSPECTIVES

Our objective is to automatically annotate sequences of talking faces in TV-Content. In this context, determining which face is speaking is very difficult due to many ambiguities in the audio, image and their association. In this paper, we propose to cluster people using the audio and the visual information independently and combine these clusterings to obtain sequences of talking faces. The system of indexing people by the visual information is based on costume detection and clustering. Indexing people by the audio is based on a Bayesian Information Criterion (BIC). Talking faces sequences are determined by searching for the best association between each detected visual-cluster and audio-cluster, and taking the intersection of these associated segments. To perform our experiments, it was necessary to collect and annotate a corpus of TV-shows. Results on this database show that only 55% of total duration time of talking faces referenced is detected automatically by our method, but with a good reliability (about 90% of the indexed time is correct for 3 shows out of 4). Hence, when both audio and visual clusters agree, it is very likely to be correct. A way to improve the results is to improve our core audio and video clustering systems. For video clustering, the costume clustering can be improved introducing information about the shape of the costumes. Face clustering can also be used. For speaker clustering, there has been until now little work on such interactive TV-shows, and this type of contents deserve efforts in this area. Finally, for the fusion of audio and video clustering, all cases where the audio and video disagree must be addressed. As these disagreement can be due either to actual differences between audio and video reference (the face which is visible is not speaking) or to errors in audio and video clustering, they deserve very careful treatment. In future work, as the reliability of our outputs is high, we want to use the detected speaking faces clusters to train model of speaking faces in order to recover missed segments.

## 6. REFERENCES


[1] Cees G. M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 2005.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 2000.

[3] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy – automatic naming of characters in tv video. *Proc. British Machine Vision Conference*, 2006.

[4] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition*, 2001.

[5] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 1991.

[6] M. Bendris, D. Charlet, and G. Chollet. Introduction of quality measures in audio-visual identity verification. *Proc. ICASSP'09*, 2009.

[7] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain. Multistage speaker diarization of broadcast news. *IEEE Trans. on Audio, Speech and Language Processing*, 2006.

[8] G. Jaffre and P. Joly. Costume: A new feature for automatic video content indexing. *Proc. RIAO*, 2004.

[9] S.Galliano, G. Gravier, and L. Chaubard. The ester2 evaluation campaign for the rich transcription of french radio broadcast. *Proc. Interspeech'09*, 2009.