Audiovisual Diarization Of People In Video Content

Elie El Khoury[1,2], Christine Sénac[3], Philippe Joly[3]

[1]Idiap Research Institute, Martigny, Switzerland

[2]Laboratoire d'Informatique de l'Université du Maine, Le Mans, France

[3]Institut de Recherche en Informatique de Toulouse, Toulouse, France

Abstract

Audio-Visual People Diarization (AVPD) is an original framework that simultaneously improves audio, video, and audiovisual diarization results. Following a literature review of people diarization for both audio and video content and their limitations, which includes our own contributions, we describe a proposed method for associating both audio and video information by using co-occurrence matrices and present experiments which were conducted on a corpus containing TV news, TV debates, and movies. Results show the effectiveness of the overall diarization system and confirm the gains audio information can bring to video indexing and *vice versa*.

1 Introduction

Audio-Visual People Diarization (AVPD) aims to identify the people that talk and/or appear in a video document and to quantify their talk/appearance time. Much existing work has addressed this problem by using only one modality. In the audio domain, AVPD is often known as speaker diarization: it aims to segment the audio stream into turns by speakers, then cluster all turns that belong to the same speaker. Its goal is to answer the questions "who spoke?" and "when?". In the video domain, AVPD typically refers to visual people detection, tracking, and clustering. In other words, it aims to answer the questions "who appeared?" and "when?". Some research activities have addressed the problem of AVPD from a multimodal point of view but their applications have often been limited.

AVPD can be used in many different kinds of applications by both professionals and the general public. One of the most interesting applications of people diarization to video documents is the detection of major casts and their roles, for example, the anchor persons in TV news or principal characters in movies [6, 10, 11, 14, 18, 19]. Their occurrences provide good indices for organizing and presenting video content. This enables many applications of such "intelligent fast-forwards" where users easily digest the main scheme of visual media by skimming through clips associated with major casts.

The task of people diarization encounters many difficulties:
- the number of people in the document is unknown;
- there is no *a priori* knowledge about the identity of the people in the document;
- there may be different lighting conditions;
- many people may appear at the same time;
- the size of face may vary from the very small to the very large;
- there may be different audio recording conditions;
- many speakers may speak at the same time;
- the audio channel may contain not only speech, but also music and other non-speech sources (applause, laugher, etc.).

This paper is organized as follows: recent literature on people diarization done in both audio and video domains including our own contributions is reviewed in sections 2 and 3. In section 4, we briefly describe existing work on audiovisual fusion, and then we outline the framework for associating audio and video information using co-occurrence matrices. Experiments done on news, debates and movies are discussed in section 5.

2 People diarization in audio domain

In the audio domain, AVPD is known as speaker diarization. It consists of segmenting and clustering an audio recording into its different speakers without *a priori* knowledge of their numbers or identities. Speaker diarization is a necessary step in a majority of applications such as speech recognition, speaker recognition, and document content structuring. All these applications are part of the Rich Transcription (RT) domain and are regularly evaluated by the National Institute of Standards and Technology (NIST)[1].

Domains that initially received special research attention were telephone speech and broadcast news (radio, TV) while today, meetings (debates, lectures, etc.) are predominantly studied because they bring a number of new challenges to speaker diarization. While broadcast news is mainly recorded in a studio with a lapel microphone, the recording conditions for meetings can vary considerably due to many factors: different far-field microphones (single or multiple), variable distance between speakers and/or microphones that leads to different speech volume levels, possible reverberations, background noise, etc. In addition, meetings sometimes contain spontaneous or overlapping speech while broadcast news speech is often read and speech turns may be of very short duration in meetings.

Fig. 1 shows the general modules that make up most speaker diarization systems. The preprocessing step is the traditional parameterization of speech data into acoustic features; in our work, we use the Mel Frequency Cepstrum Coefficients (MFCCs) and 4 Hz modulation energy. Next, there is a module of speech activity detection (see section 2.1) which can be preceded, in the case of "difficult data", by noise reduction and multichannel acoustic beamforming.

Speaker segmentation (see section 2.2) aims at splitting the audio stream into homogeneous segments by speaker. This module is generally applied before the clustering one, but new speaker diarization systems for meetings try to employ them simultaneously, and in-

[1] http://www.itl.nist.gov/iad/mig/tests/rt/

deed, our own approach tends to combine these two modules in an iterative way. Cluster initialization depends on the clustering approach, i.e. the choice of an initial set of clusters in bottom-up clustering [1] or a single segment in top-down clustering [7] (see section 2.3). Finally, the distance between clusters and a split/merging mechanism is used to iteratively merge clusters [43] or to introduce new ones [21] until the optimum number of clusters has been reached using stopping criteria. Optionally, data purification algorithms can be used to make clusters more discriminant [7, 43]. In the following sections, we review each of these steps and describe the method adopted in our work.

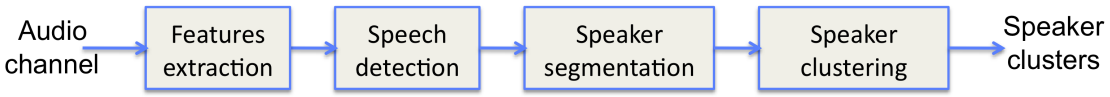

Figure 1: General architecture for speaker diarization.

2.1 Speech activity detection

Speech Activity Detection (SAD) is a fundamental task that involves the separation of speech and non-speech segments. SAD can have a significant impact on speaker diarization performance because the speaker acoustic models involved in the process can be distorted by the presence of non-speech segments. Many different approaches have been reported in the literature [49]. They are mainly based on models (such as Gaussian Mixture Models) and rely on a two-class detector. The models are pre-trained with external speech and non-speech data [36, 68].

The drawback of this model-based approach is the need for new training for every new data especially in the case of changes in acoustic conditions. It is for this reason that in our work we combine the model-based approach with an unsupervised speech detector based on 4 Hz modulation energy [51, 47]. This fusion technique produced positive results during the French competition ESTER-1 [24].

However, we have found that in segments where two people talk simultaneously or where speech overlaps with music, the value of 4 Hz modulation energy is not always relevant. Due to a threshold decision, this method may introduce additional missed detections and imprecise boundary locations of speech regions. To avoid these errors, we propose to apply our GLR/BIC segmentation (cf. section 2.2) before using the speech detection module. This improvement was validated during the French competition ESTER-2 [25].

2.2 Speaker segmentation

Speaker segmentation consists in splitting the audio recording into homogeneous segments. Each segment must be as long as possible and must only contain the speech of one speaker. This segmentation is closely related to acoustic change detection. Classic methods perform hypothesis testing by using the acoustic segments in two sliding and possibly overlapping, consecutive windows. They generally use metric approaches (such as symmetric Kullback-

Leibler [53] or Hottelings T2-Statistic [67]), or approaches based on model selection such as GLR [26] or BIC [12] which lead to the best systems [8, 54].

However, we have found that the usual GLR and BIC methods present some disadvantages: too many parameters are required to tune the algorithm, and detecting the boundaries of small segments is often imprecise. In a previous paper [32], we presented a different method of segmentation that provides more accurate segments: a GLR algorithm is applied several times until it converges to the best repartition of Gaussian distributions. Then a BIC algorithm chooses the points that correspond to speaker changes. Due to the shifted variable size window introduced in this GLR/BIC method [33], processing from "left to right" may detect different points of change than processing from "right to left", and therefore, there is a chance that a missed boundary in the first direction will be detected in the other direction and *vice versa*. Thus, the output is the union of both segmentations.

2.3 Speaker clustering

Clustering consists of collecting all segments corresponding to the same speaker. Ideally, there will be one cluster for each segment. Most existing clustering methods for speaker diarization have either bottom-up or top-down architectures as illustrated in Fig. 2. Top-down architecture is initialized with few clusters (usually one) whereas the bottom-up approach (the most common in the literature because of its results) is initialized with many clusters that are usually the segments provided by speaker segmentation. In the hierarchical bottom-up manner, the closest clusters - in the sense of a matching and/or similarity measure - are merged iteratively. Three scenarios are possible depending on the threshold used to stop the clustering: over-clustering, under-clustering, or optimal clustering (see Fig. 2). Many matching measurements such as BIC [12] or EVSM [61] (Eigen Vector Space Model) are proposed in the literature.

In our work, we use the bottom-up BIC clustering to which we have applied some improvements in order to fit recordings in which there is high interaction between speakers: this corresponds to scenarios where many people speak simultaneously and the average segment duration is relatively short. These scenarios decrease segment purity, and thus, introduce a risk of cumulative errors in the clustering process. To deal with this problem, we previously [33] applied local clustering that helps construct a first set of "good clusters" of balanced size, before applying global clustering to the whole document.

At the end of the clustering process, each segment is theoretically assigned to the cluster providing the highest BIC similarity. However, due to the hierarchical bottom-up manner, there are still some segments that do not follow this hypothesis. To correct these errors and therefore enhance cluster purity, we compute the similarity matrix between segments $\{S_i\}$ $(1 \leq j \leq N_S)$ and clusters $\{C_j\}$ $(1 \leq j \leq N_C)$ and then reclassify segments according to this matrix. The clusters are updated by assigning each segment S_i to $arg\max_{C_j}(-\Delta BIC(S_i, C_j))$ $(1 \leq j \leq N_C)$.
"Unstable segments" are split using bidirectional GLR/BIC segmentation: we consider "unstable segments" as those segments for which $-\Delta BIC(S_i, C_j)) < 0$, (i.e. the similarity between segment S_i and its corresponding cluster is low). If at least one segment is split, a new step of speech detection is processed and another loop of similarity matrix computation, cluster updating and unstable segment splitting is performed. Otherwise, a final clustering is processed in order to group clusters corresponding to the same speaker but under different

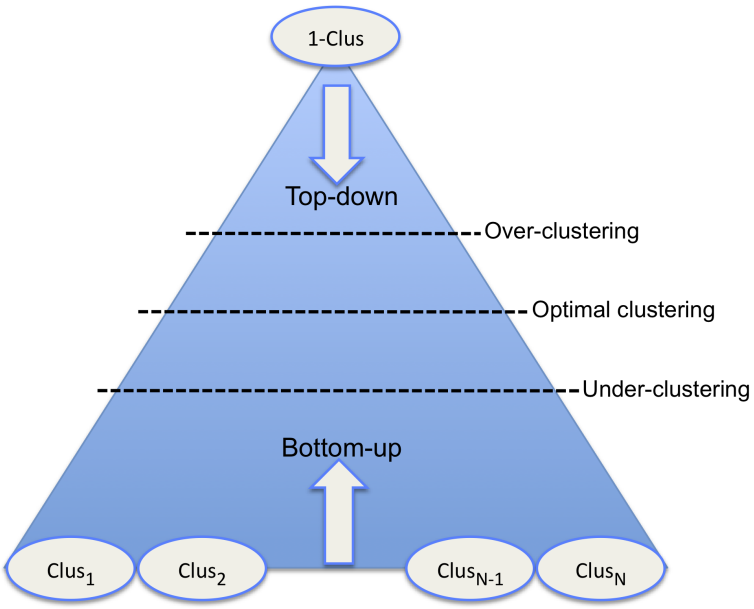

Figure 2: Top-down and bottom-up hierarchical clusterings.

backgrounds. This method has shown excellent results on the ESTER-2 corpus [25].

3 People diarization in video domain

AVPD in the video domain aims at annotating video documents according to the people appearing in those documents using only visual information.

In [17], the authors present an overview of the current approaches that provide an appearance-based person "re-identification" using camera networks. These methods are based on the use of the overall appearance of an individual as opposed to passive biometrics such as face and gait. In such applications of video surveillance, only people detection and people tracking are required. However, we are interested in the case of edited documents such as TV content and movies, in which visual people diarization requires many steps as illustrated in Fig. 3: shot segmentation, people detection, people tracking, and people clustering. In the following sections, we will present a state-of-the-artstate-of-the-art system and the method adopted for each of the processing steps.

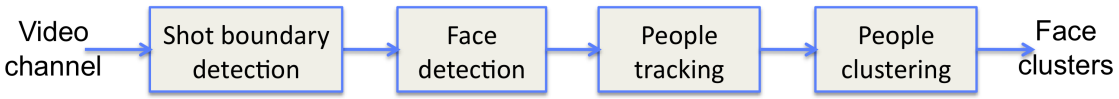

Figure 3: General architecture for visual people diarization.

3.1 Shot boundaries detection

Shot boundary detection is a well-known segmentation process. It aims to break down the massive volume of video into smaller chunks. Quite a lot of approaches have been proposed in the literature [38, 60]. Readers can see the TRECVid report [55] for a detailed review and a comparison of state-of-the-art systems.

In our work, we applied our generic segmentation method that combines the generalized likelihood ratio (GLR) and the Bayesian information criterion (BIC) [34]. The main idea behind this method is to chunk any audio or video stream into homogeneous segments. For shot boundary detection, this method gives results comparable to state-of-the-art systems.

3.2 Visual people detection

Once the video shots are extracted, the next goal is to detect people in each of those shots. People detection consists in identifying and locating humans in an image regardless of their position, scale and illumination. Many methods that aim to detect people have been proposed in the literature which are often based on full-body detection, partial-body detection (upper and lower body) or face detection [2, 30, 64]. As we are interested in methods applied to TV, faces and upper-body are the most relevant for this kind of data.

3.2.1 Face detection

Given an arbitrary image, the goal of face detection is to determine whether or not there are any faces in the image, and if so, to provide the location and the size of each face. Many existing approaches aim to detect faces in images and/or sequences of images [66] by carrying out the task through extracting some properties (e.g. local features) of a set of training images acquired in a fixed pose (e.g. upright frontal pose). Based on the extracted properties, the face detection system scans through the entire tested image at every possible location and scale in order to locate faces.

In our work, we use the AdaBoost method [65] to detect frontal faces thanks to the OpenCV toolbox[2]. This method contains three major phases: a rectangular feature extraction, a training data classifier using boosting techniques and a multi-scale detection algorithm. To cope with sequences of frames, we bring a trivial improvement by considering that a face must be present in at least n consecutive frames (e.g. $n = 5$, corresponding to 200 milliseconds if the frequency is 25 frames/second) in order to be visible.

3.3 Visual people tracking

Once a person is detected, the next goal is to follow this person in scenarios where the face detector fails. Numerous approaches for non-rigid object (such as human) tracking have

[2] http://opencvlibrary.sourceforge.net/

been proposed in the literature. They generally differ in the way an object is represented and image features are selected, and/or on the algorithm used for tracking.

We focus on tracking faces and clothes because unlike video surveillance, movies, TV talk-shows, TV game shows and TV news frequently feature scenes containing people in which their upper-bodies are the most visible part.

3.3.1 Face-based people tracking

Tracking is essentially motion estimation. However, general motion estimation has fundamental limitations such as the aperture problem. In face recognition systems, each face must be tracked over the video sequence in order to extract appropriate information. Existing approaches can be divided into three categories: (1) head tracking, where the entire face is tracked as a single rigid entity (such as in [4]); (2) facial feature tracking (such as in [59],) where features like eyes, ears, nostrils, eyebrows, lips, mouth and nose are limited by the anatomy of the head that is considered here as a non-rigid object influenced by motion due to speech or facial expressions; (3) complete tracking, which involves tracking both the head and facial features (such as in [57]). In addition, many of those methods are able to handle challenging situations such as facial deformations, lighting changes , partial occlusions, pose variation and facial resolution.

In order to deal with the large variation in face sizes, we consider the face as a single non-rigid entity with no need to track face features (eyes, lips, etc.). Based on facial skin color, two tracking processes are done: backward tracking and forward tracking.

3.3.1.1. Skin color extraction. The most difficult issue for skin color extraction is to separate chrominance from lighting effect. As reported in [63], the most interesting descriptors are the chrominance components (C_r and C_b) of the YC_rC_b color space, and the hue (H) component of the HSV color space. In our work, we apply a thresholding method on C_r and C_b that are coded on 1 byte, and H that is normalized between 0 and 1, using the following expressions:

$$\begin{cases} 135 \leq C_r \leq 170 \\ 130 \leq C_b \leq 200 \\ 0.01 \leq H \leq 0.1 \end{cases} \quad (1)$$

Those thresholds are parametered on a training set of faces of various skin colors ranging from very light to very dark.

3.3.1.2. Skin modeling. Once the skin color is extracted, the corresponding normalized r and b are computed and, are used to set up a 2D Gaussian model. It has been shown [64] that the rgb normalized space is better than RGB, YC_rC_b and HSV spaces because it handles lighting variation.

3.3.1.3. Backward-forward tracking. For each detected face, the bounding box is defined by two points: the top-left corner (Pt_1) and the bottom-right corner (Pt_2). Supposing that a shot contains n frames and that the face is only detected in the sequence of

frames $\{I_s, \ldots, I_e\}$, the goal is to verify if that face is also present throughout the shot in $\{I_1, \ldots, I_{s-1}\}$ on the left side, and in $\{I_{e+1}, \ldots, I_n\}$ on the right side as seen in Fig. 4.

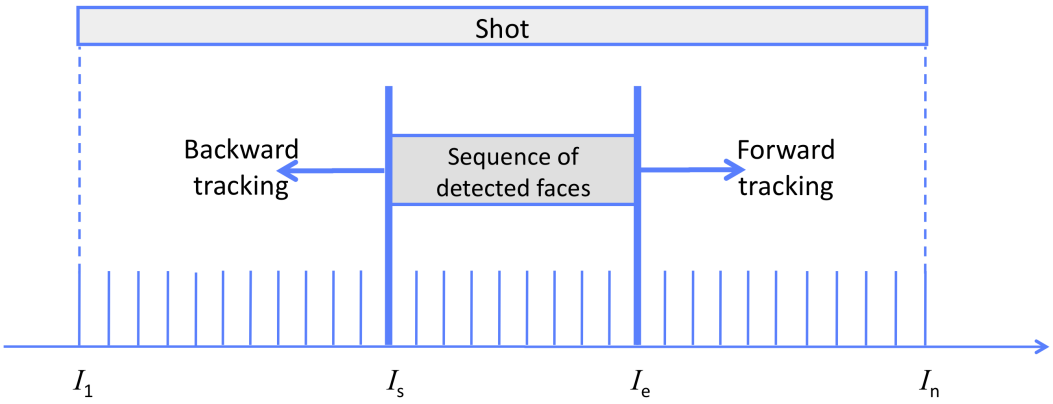

Figure 4: The backward-forward tracking scheme.

The proposed algorithm is an iterative process and can be divided into 4 steps:

1. For the backward (*respectively* forward) tracking, two points are estimated in the frame I_{s-1} (*respectively* I_{e+1}) as follows:

$$\begin{aligned} Pt'_1 &= Pt_1 - \alpha(Pt_2 - Pt_1) \\ Pt'_2 &= Pt_2 + \alpha(Pt_2 - Pt_1) \end{aligned} \tag{2}$$

   where Pt_1 and Pt_2 are the corners of the face box obtained in the starting frame I_s (*respectively* I_e) and α a fixed coefficient (e.g. $\alpha = 0.1$).
   Pt'_1 and Pt'_2 delimit the estimated box in which the candidate face is present.

2. Each pixel $x = (x_i, x_j)$ within the box is classified (skin/non-skin) using the probability function:

$$p(x) = \frac{1}{2\pi\,|\Sigma|^{\frac{1}{2}}} \exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)] \tag{3}$$

   where the mean μ and covariance Σ are adapted to the skin color of the frame I_s (*respectively* I_e).

3. Since the face is considered as a single entity, pixels are processed using dilation and erosion morphological filters.

4. If the Ratio of the Skin Part (RSP) within the box is higher than a threshold (Thr_{RSP}), the face is considered as visible and the points Pt_1 and Pt_2 are updated according to the proper box in the candidate image; the 2D Gaussian model is also updated with the new data and the process is repeated for frame I_{s-2} (*respectively* I_{e+2}) starting from step 1. If the RSP is lower than Thr_{RSP} or if the boundaries of the shot are reached, the process is stopped.(See Table 2 for values of Thr_{RSP}).

3.3.2 Clothing-based people tracking

Sometimes, the face tracker fails because the face may be occluded or the skin color model is not accurate enough. One way to overcome these problems is to track clothing instead of the face.

Even though researchers do not give clothing special attention in many publications, it remains one of the important cues for people tracking because it has an amount of color information that is trackable in difficult situations like occlusion [37]. In [28], the authors used clothes tracking in order to re-texture it for real-time virtual clothing applications. More sophisticated research ontracking clothed people can be found in [50] where the authors used it for motion capture.

In our work, since no precise characterization of clothing is needed, we propose a simpler clothing tracker. First, for a given detected face, we estimate the clothing box by the area under the face. The size of this clothing box is proportional to the size of the face: as in [31], we suppose that the width of the clothing box is equal to 2.3 times the width of the face box, and its height is equal to 2.6 times the height of the face box. This is illustrated in Fig. 5. Next, we use a tracking technique similar to the one used for the face. However, instead of using a pre-selected set of pixels as we did for skin color, we use the entire set of pixels within the clothing box.

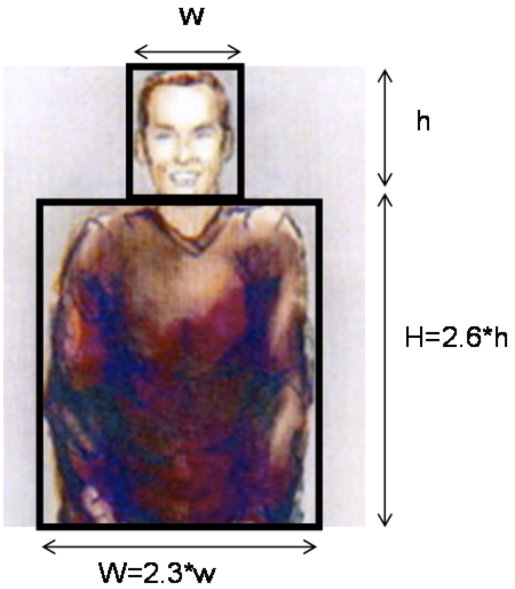

Figure 5: Clothing localization based on the face box.

This multi-people visual tracker we have been discussing is suitable for offline processing of TV data especially for debates and news. However, it may fail in some cases where 1) faces are subject to fast motion, and 2) the background is of similar color to the skin.

3.4 Visual people clustering

At the end of the tracking step, a list of all "face-tracks" is available. Every track T_i corresponds to a sequence of frames where the face F_i is visible. The next goal is to cluster the tracks that contain appearances of the same person. The task of visual people clustering is relatively new. It can be applied to both still images (e.g. organizing consumer photos) as in [13] or moving sequences of images like in[15] since the basic technique is often the same.

3.4.1 Review of existing methods

Researchers often view visual people clustering as a recognition problem [3, 45], an identification [5, 14] or also re-identification problem [17]. In [3], the authors develop a recognition method based on a cascade of processing steps that normalize the effects of the changing environment: they first suppress the background surrounding the face, enabling the maximum facial area to be retained. Then, they add a pose refinement step to optimize registration (using facial features like eyes and mouth detected using SVM) between the test image and a sample face. They use a distance inherent to a subspace to allow for partial occlusion and expression change.

In [15], after detecting faces using an iterative algorithm that gives a confidence measure for the presence or absence of faces within video shots, the authors process the clustering of those faces using a PCA-based dissimilarity measure in conjunction with spatio-temporal correlation. In [20], a distance which is invariant to affine transformations, is introduced for clustering and classification. This is applied to face clustering in order to produce an automatic cast listing in movies. In [13] Chu *et al.* present a clustering method for consumer photos by matching images using local features. It represents matching situations using visual sentences. Then, visual language models are constructed to describe the dependency of image patches on faces.

In [14], the authors propose an unsupervised metric learning method for face identification, recognition and clustering. Their method learns a Mahalanobis distance without manually labeled examples. They use pairs of faces within tracks as positive examples, while negative examples are generated from frames where different people appear together. However, this unsupervised learning may lead to over-fitting because there is not a lot of variability in the positive examples.

3.4.2 Proposed method for people clustering

The face is the most reliable entity that is used to visually cluster people. However, other high-level visual concepts like clothing can be helpful. In the following paragraphs, we will present our face-based and clothing-based matching methods and then our hierarchical bottom-up clustering algorithm.

3.4.2.1. Face-based matching. The face contains many discriminative features: skin color, hair, ears, eyes, mouth, nose and even shape. All these features can be used to recognize people. However, the variations in illumination, face scales, head pose, partial

occlusions, etc., are constraints that make the task of face-based matching difficult. In our study, face-based matching relies on two features: SIFT matching and skin color matching both of which are outlined below. Moreover, instead of processing the whole sequence of faces which is time consuming, we decided to work only on *keyfaces*: for every sequence of frames, we chose one frame in which the facial image is the most representative and contains the maximum amount of useful information [35].

- *SIFT matching.* SIFT features introduced by Lowe in 2004 [41] are known to be robust to variations in scale, rotation, and illumination. Nowadays, they are used as baseline features in most successful object recognition systems such as [9]. These systems need huge quantities of both positive and negative training data. However, in our case, the clustering must work in an unsupervised manner. The challenge here is not to match unlabeled to labeled images in order to detect a face in the unlabeled image (as the supervised systems do), but to verify if the two faces are assigned to the same person or not.

  In our previous work [35], we defined a new distance named "Average N-Minimal Pair Distance" (ANMPD). If we consider two tracks T_1 and T_2 to which the *keyfaces* F_1 and F_2 are associated, their corresponding sets of SIFT keypoints K_1 and K_2 are:

$$\begin{cases} K_1 = k_1^1, k_2^1, \ldots, k_L^1 \\ K_2 = k_1^2, k_2^2, \ldots, k_M^2 \end{cases}$$

  We define d_p as the Euclidean distance between each pair of keypoints $P = (k_i^1, k_j^2)$. After sorting the distances of all possible pairs, the first N minimum ones expressed by $\{d_p\}$ $(1 \leq p \leq N)$ are selected.

  Then, the ANMPD is computed by:

$$D_1(T_1, T_2) = ANMPD(K_1, K_2) = \frac{1}{N} \sum_{p=1}^{N} d_p \tag{4}$$

- *Skin color matching.* Under the same illumination conditions (especially for debates), skin color can be used as an additional cue to help merging or separating between people. Inside the face box, we select the pixels that correspond to the skin. To do this, we use the thresholding method applied to the C_r and C_b components (from YC_rC_b color space), and the hue H as described in section 3.3. Then matching between the skin colors of two keyfaces F_1 and F_2 is done by computing the variation between their corresponding histograms h^{F_1} and h^{F_2}. Here, we use the Bhattacharyya distance to obtain slightly better results than Euclidean and Manhattan distances.

$$D_2(T_1, T_2) = d_{Bhat}(h_1^{F_1}, h_2^{F_2}) = -\log \left[\sum_i \sum_j \sum_k h^{F_1}(i,j,k)(h^{F_2}(i,j,k) \right] \tag{5}$$

3.4.2.2. Clothing-based matching. Since within video documents such as debates, TV contests, movies and series a character often wears the same clothing throughout the whole document or at least for a considerable duration (e.g. a scene), clustering that uses clothing information is a appropriate/useful solution. The extraction of the clothing part is done as explained in [31]. After this extraction, we investigate two clothing descriptors: 3D histograms and texture.

- *Comparing Histograms.* After computing the 3D histograms h^{C_1} and h^{C_2} that correspond to clothes C_1 and C_2, their comparison is made using Bhattacharyya distance:

$$D_3(T_1, T_2) = d_{Bhat}(h^{C_1}, h^{C_2}) \tag{6}$$

- *Texture.* We use the Gabor texture features that are introduced in [42]. In order to compute the distance $dTexture$ between the textures of two different clothes C_1 and C_2, we compute the normalized distance in the feature space between the corresponding feature vectors X_1 and X_2.

$$\begin{cases} X_1 = [x_1^1, x_2^1, ..., x_Q^1] \\ X_2 = [x_1^2, x_2^2, ..., x_Q^2] \end{cases} \tag{7}$$

The distance is defined by:

$$D_4(T_1, T_2) = d_{Texture}(C_1, C_2) = \sum_{q=1}^{Q} \left| \frac{x_q^1 - x_q^2}{\alpha(v_q)} \right| \tag{8}$$

where $\alpha(x_q)$ is the standard deviation of the q^{th} coefficient of the feature vector over all the database.

3.4.2.3. Hierarchical bottom-up clustering. After listing the different kinds of face and clothing matching that can be used to help group(cluster) tracks if they correspond to the same person, the issue is to find an appropriate way to combine all this information. It is obvious that pairs of tracks that verify all the merging criteria listed above are preferable at the beginning of the clustering process. However, in some cases where illumination, background clutter and clothing are subject to change, some of the above matchings may not be verified at all. In this case, the next clustering steps should be done by using fewer merging criteria. Thus, we propose to adopt a 3-level hierarchical bottom-up clustering.

- *First-level hierarchical clustering.* From the four distances obtained from the $Sift$, $Skin$, $Clothing - histogram$ and $Texture - clothing$ descriptors, we can compute the similarity between two tracks T_i and T_j as:

$$S(T_i, T_j) = \prod_{a=1}^{4} \max(Thr_a - D_a(T_i, T_j), 0) \tag{9}$$

Where $D_a(T_i, T_j)$ is the distance between T_i and T_j in terms of the a^{th} descriptor. Thr_a is the threshold that corresponds to the a^{th} descriptor (see Table 2 for the value of these thresholds). $S(T_i, T_j)$ may even be positive if there is good matching, or be equal to 0 if at least one of the descriptors does not confirm the matching. Then, the clustering is done in a hierarchical bottom-up manner, (i.e. starting from the most similar tracks/clusters), using the complete linkage property. Each time two tracks T_i and T_j are merged, the matrix is updated as explained in [35].

- *Second-level hierarchical clustering.* After a first clustering for which merging confidence is very high, a second clustering is done with more tolerance. In this case, two conditions are sufficient:

1. at least one of the two clothing descriptors works: the second descriptor may fail if there are partial occlusions in which case the texture descriptor will fail) or lighting variations (in which case the color histogram comparison will fail);
2. at least one of the two face descriptors works: this condition is taken into account to prevent merging in case two people are wearing the same clothing.

- *Third-level hierarchical clustering.* When the illumination varies or the clothing of the person changes, color-based descriptors and texture descriptors are subject to change. In this case, the only reliable descriptors that remain useful are the SIFT descriptors of faces. For this reason a final clustering step is done based only on SIFT descriptors.

4 People diarization in audiovisual domain

In previous sections, we reviewed techniques that handle each of the audio and video media separately. In this section, the challenge we are facing is the fusion of different modalities. By its nature, an audiovisual document contains a set of information generally synchronized like frames, sound and sometimes textual information. More particularly, we give special care to the problem of associating voices from the audio channel to characters from the video channel. We will then use this association to improve the results of "video-only" AVPD, and the results of "audio-only" AVPD.

The next sub-sections, use the following **notations**:
- n_a is the number of audio clusters;
- n_v is the number of video clusters;
- $\{A_i\}_{i=1...n_a}$ is the set of audio clusters;
- $\{V_j\}_{j=1...n_v}$ is the set of video clusters;
- Q_i is the number of utterances of the audio cluster A_i;
- R_j is the number of tracks of the video cluster V_j;
- $\{U_q^i\}_{q=1...Q_i}$ is the set of utterances that correspond to the audio cluster A_i;
- $\{T_r^j\}_{r=1...R_j}$ is the set of tracks that correspond to the video cluster V_j.

4.1 Related work

Using the different modalities to create cross-modal correspondences in an unsupervised manner is an advantage of multimodal systems that has not been adequately explored in the existing literature.

One domain where audiovisual diarization has been studied is meeting scenarios. The challenge here is to use far-field cameras and microphones to analyze human activity in a meeting scene which typically has multiple subjects. The CLEAR 2006-2007 evaluations [56] focused on this domain.

In [29] the authors propose an audiovisual online diarization of participants in group meetings. They develop an unsupervised approach based on the analysis of pairwise correlations between speaker clusters and visual activity features extracted from multiple video channels. An iterative association is made between pairwise the audio and video streams with the highest correlation, until all audiovisual streams are associated. This system tries to solve the

task incrementally and on-the-fly. This work is extended in [22], where a multimodal speaker diarization of real-world meetings is proposed. This system, not on-the-fly, makes use of a single far-field microphone and any collection of available uncalibrated cameras and is tested on 4-person meetings where participants behave naturally. Instead of using a lip activity detector, the authors prefer a motion vector magnitude to construct an estimate of personal activity levels. This estimate has been shown to correlate well with speaking activity patterns. In [23], the same authors present an audiovisual approach for unsupervised speaker localization in both time and space, called "dialocalization". Using recordings from a single, low-resolution room overview camera and a single far-field microphone, a state-of-the-art audio-only speaker diarization system is extended so that both acoustic and visual models are estimated as part of a joint unsupervised optimization problem. After the speaker diarization step, the visual models are used to infer the location of the speakers in the video. The multimodal integration is made so that, during every agglomerative clustering iteration, each speaker cluster is modeled by two GMMs, one for the audio features and one for the video activity features. In the segmentation step and in the merging step, the weighted sum of the log-likelihood scores of the two models is used.

In [52], the authors present an online diarization of streaming audiovisual data for smart environments. That system, which requires a training step, integrates components for speaker change detection, speaker identification, speaker localization and face identification. It is divided into a video sub-system that performs face detection and identification, and an audio sub-system that localizes and identifies the speakers. The video system incorporates a single camera, while the audio system contains multiple microphone arrays.

Contrary to our work, these previously reviewed studies focus on analyzing multi-channel recordings rather than edited content. Existing work with the same purpose as ours are [16, 39, 40]. In [16], an unsupervised detection of multimodal clusters in edited recordings (such as talk-shows and sitcoms) is presented. This detection avoids making assumptions about the recording content, such as the presence of specific participant voices or faces. In this approach, the video stream is segmented into shot clusters and the audio stream is segmented into audio clusters using a diarization framework. Then AV-clusters are built based on the co-occurrences between shot and audio clusters: a selection criterion based on χ^2 (chi-squared distribution) test [48] is used to this end.

In all this research, we can see the difficulty of associating audio and visual features due to two main factors. First, the data to model are often heterogeneous (color histograms, SIFT features, presence of the face, size of the face, etc.) and correspond to different levels of granularity. Second, there is the problem of stream synchronization due to the fact that the extractions of low-level features are not done on the same timestamps for audio and video. These factors make early fusion of audio and video not particularly appropriate in this case. In addition, as shown in [62] in the case of a multimodal speaker diarization system, early fusion did not improve the diarization performance compared to using audio or video alone. With late fusion, the authors showed that by modeling audio and video features separately, they improved upon audio-only speaker diarization when video features were also used.

The work the most similar to our topic is that done by Liu and Wang [39, 40] to detect the major casts in video content. In their work, they assume that the majority of speech that accompanies the appearances of each character is from the same person. Thus, the correlation between the audio cluster A_i and the video cluster V_j can be expressed by the overlapping time between all the utterances of A_i and all the tracks of V_j.

$$m_{ij} = \sum_{q=1}^{Q_i} \sum_{r=1}^{R_j} OL(U_q^i, T_r^j) \tag{10}$$

where $OL(U_q^i, T_r^j)$ is the overlapping of audio utterance U_q^i and face track T_r^j.

They improve this association by assuming that large faces sizes are most likely to be talking. Thus the correlation between A_i and V_j becomes:

$$c_{ij} = \sum_{q=1}^{Q_i} \sum_{r=1}^{R_j} OL(U_q^i, T_r^j) \times FS(T_r^j) \tag{11}$$

where $FS(T_r^j)$ is the face size of a track T_r^j corresponding to the video cluster V_j. The use of face size is helpful when more than one face appears during a speech segment, where the larger face is more likely to be the real speaker.

One limitation of this method is that it cannot handle the case where the video image of one person is accompanied by the speech of another person (voice over). Our first contribution will focus on solving this problem.

4.2 Proposed audiovisual association

4.2.1 Baseline system

As seen previously, audiovisual people association methods such as [40] consider both visual and speech features to be simultaneously relevant in video subsequences and assume that the current voice corresponds to a face present in the frame. In real sequences, this hypothesis is often violated. It is very common to find sequences where the people appearing do not talk for many frames or many shots. Furthermore, it is also possible that the current voice belongs to a person whose face is not in the current frame.
In this work, we propose to compute co-occurrences between audio and video indexes, i.e. we match up the voices with the faces. This approach is suitable to handle cases where the usual assumptions are not verified.

Before describing our method, let us illustrate how a person A can occur in a document. As seen in Fig. 6, there are 7 scenarios for person A. These scenarios depend on the way person A is visible on the screen or talking.

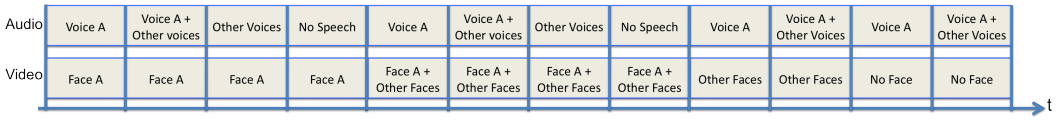

Figure 6: 7 different scenarios where a person A may occur in a document.

First, we compute a matrix which represents the intersection between the audio and video indexes. We consider the two indexes, frame by frame. For every frame, if the voice of A_i

is heard and the visual person V_j is present, then the number of occurrences m_{ij} of the pair (A_i, V_j) is incremented. Thus, we obtain the following matrix:

$$M = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_{n_a} \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & \dots & V_{n_v} \end{array} \\ \begin{pmatrix} m_{11} & m_{12} & \dots & m_{1n_v} \\ m_{21} & m_{22} & \dots & m_{2n_v} \\ \vdots & \vdots & \vdots & \vdots \\ m_{n_a1} & m_{n_a2} & \dots & m_{n_an_v} \end{pmatrix} \end{array} \tag{12}$$

where the value m_{ij} means that in all the frames where the voice A_i is heard, the visual person V_j appears m_{ij} times. Conversely, in all the frames where the person V_j is present, the voice A_i is heard m_{ij} times.

The idea in [40] would be to sort the resulting matrix M by rows (or by columns).

However, this solution makes the assumption that: when a voice is heard, the corresponding face is the one most present in the overlapping time (sorting by rows). Conversely, sorting by columns means that for each face its corresponding voice is the one mostly heard when the features appear. For example, in some TV talk-shows and debates, this assumption is not valid: the person who speaks the most is usually the host. In this case,the host's voice is often heard the most even when the guest(s) appears on screen. Thus, even if the matrix M is a good starting point to associate audio and video indexes, it cannot be directly used if there is no a priori information about the people. A post-processing procedure is required.

One way to bypass the problem is to read M both by rows and columns, and to retain the most significant information. This fusion is carried out by computing two new matrices, M_a and M_v where the overlapping time is replaced by one of the frequencies:

$$f_{ij}^a = \frac{m_{ij}}{\sum_{k=1}^{n_v} m_{ik}}, \qquad f_{ij}^v = \frac{m_{ij}}{\sum_{k=1}^{n_a} m_{kj}} \tag{13}$$

in M_a, the sum of all frequencies of a row is equal to 1.

$$M_a = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_{n_a} \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & \dots & V_{n_v} \end{array} \\ \begin{pmatrix} f_{11}^a & f_{12}^a & \dots & f_{1n_v}^a \\ \boxed{f_{21}^a \quad f_{22}^a \quad \dots \quad f_{2n_v}^a} \\ \dots & \dots & \dots & \dots \\ f_{n_a1}^a & f_{n_a2}^a & \dots & f_{n_an_v}^a \end{pmatrix} \end{array} 100\,\% \tag{14}$$

Similarly, the sum of all frequencies of a column in M_v is equal to 1. The matrix M_a (respectively M_v) gives the probability density of each audio cluster A_i (respectively each video cluster V_j).

Therefore, a new matrix M_{av} that combines these two matrices is defined. To compute the coefficients of M_{av}, we can choose a conjunction operator ("AND") as the minimum operator or the probabilistic operator (product). The latter is used in this work:

$$f(A_i, V_j) = f_{ij}^a \times f_{ij}^v \tag{15}$$

From this matrix, an association between pair wise audio and video clusters is performed as follows:

1. Search- Delete step: Select the pair (A_i, V_j) with the highest co-occurrence and eliminate the two corresponding clusters from the matrix (i.e. eliminate row I and column J);
2. Repeat the search-delete step until all clusters are associated (i.e. until an empty matrix remains).

At the end of this process, we obtain a list of all the clusters which can be classified into three categories: talking-faces, face-only and voice-only.

Other algorithms could have been used to associate the audio and visual clusters. Here, we assume that associating the best co-occurring pairs of clusters first can help to accurately associate the remaining co-occurring ones by process of elimination. However, there are two limitations to the above proposal:

- If a non-talking person appears while a voice is heard, we should not allow the association between face and voice. To solve this problem, we use a lip activity detector.
- If two or more persons appear at the same time, the decision of who is talking is difficult. In the matrix, this corresponds to the following scenario: in the same row, there are two or more similar frequencies. To cope with this problem, we use information on face size.

4.2.2 The use of lip activity

As previously noted, an additional feature must be added to deal with the case where one person appears when another is talking. In this case, to eliminate any confusion, it is better to detect lip activity. Even though the literature reveals much work done to detect the lip activity, the majority deal specifically with large faces and their goal is to deal with the problem of audiovisual speech recognition [44, 58].

In this work, and given the range of face sizes in our data, we propose an easier way to estimate lip activity from the automatically detected face.
Assuming that the face is frontal and that the mouth is located in the middle-bottom of the face box [46], the bounding box of the mouth is selected as illustrated in Fig. 7. In order to quantify lip activity, we proceed by pairs of frames as follows: considering two consecutive face boxes F_1 and F_2 of the same person that are detected within two consecutive frames, and after localizing the region of the mouth m_1 inside F_1, we build a search zone around m_1 inside F_2. Then, we move a window m_2 of the same size of m_1 into this zone. Therefore, the best matching and the lip activity rate are both obtained by computing the Minimal Mean Square Error (MMSE) of the Hue values between m_1 and m_2 pixels.

Since head motions are generally related to speaking expressions, we assume that: if a person is not moving his/her lips or his/her head, we can be certain that this person is not talking. This corresponds to the case where LA is lower than a fixed threshold Thr_{la} (See

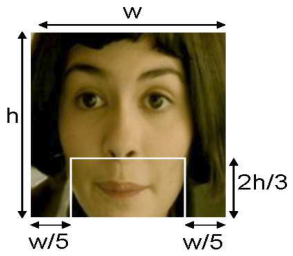

Figure 7: Mouth localization (thumbnail taken from the movie "*Amélie*").

Table 2 for the value of Thr_{la}). Lip activity can be represented by a coefficient δ_{V_j}:

$$\delta_{V_j} = \begin{cases} 0 & if \quad LA < Thr_{la} \\ 1 & if \quad LA \geq Thr_{la} \end{cases} \tag{16}$$

4.2.3 The use of face size

In a track (or shot) where many faces appear, often the person with the relatively larger face size is more likely to be the real speaker. However, this assumption is not true in the case where there are many faces with different sizes and each one appears alone in its track. In the clustering process, there should be no difference between those faces. Therefore, we define the normalized weight of a face size that is computed compared to the other faces in the image.

$$W^k = \frac{size(F^k)}{\sum_{l=1}^{L} size(F^l)} \tag{17}$$

where L is the total number of faces within the image. This formula can be extended to the track level by assuming that the size of a face in a track is almost always the same. The overall normalized weight of the face that corresponds to the visual cluster V_j is:

$$\omega_{V_j} = \frac{\sum_{r=1}^{R_j} Dur(T_r^j) \times W_r^j}{\sum_{r=1}^{R_j} Dur(T_r^j)} \tag{18}$$

Then, these two coefficients are introduced into the co-occurrence matrix M'_{av}:

$$M'_{av} = M_{av} \bullet [\delta_1 \times \omega_1, \delta_2 \times \omega_2, \ldots, \delta_{n_v} \times \omega_{n_v}]^T \tag{19}$$

4.3 Audiovisual system for people indexing

At the end of the audio (respectively video) processing, a list of audio (respectively video) clusters as well as similarity measures for each pair of clusters are provided. Above we

studied the association between these audio and video clusters by computing a co-occurrence matrix. Since the confidence level of the bottom-up clustering process decreases gradually as it approaches the top of the clustering hierarchy, the use of mutual information in the later stages such as the co-occurrence matrix will help to improve clustering performance. A good way to implement our proposal is to apply the following algorithm that is illustrated in Fig. 8:

1. The first step in confident audio clustering and video clustering is applied using restrictive decision that ensures high cluster purity but potentially more clusters than in reality. The n_a audio clusters, the n_v video clusters, as well as the similarity matrices S_a and S_v computed for each pair of clusters, are retained.
2. Using these clusters, calculate the co-occurrence matrix M of $n_a \times n_v$ dimension. Then, deduce the matrices M_a and M_v as previously explained.
3. Using M_a, compute $\alpha(A_i, A_j)$ for each pair (A_i, A_j):

$$\alpha(A_i, A_j) = \sum_{v=1}^{n_v} m_a(A_i, V_v).m_a(A_j, V_v) \tag{20}$$

and the new similarity measure:

$$S_a^{'}(A_i, A_j) = \tau_1.S_a(A_i, A_j) + \tau_2.\alpha(A_i, A_j) \tag{21}$$

Then, find the pair (A_I, A_J) that corresponds to the maximum similarity:

$$(A_I, A_J) = \arg\max_{(A_i, A_j)}(S_a^{'}(A_i, A_j)) \tag{22}$$

If $\max(S_a^{'}(A_I, A_J))$ is higher than a fixed threshold Thr_a, then merge the two clusters (see Table 2 for Thr_a). In this case, the matrices S_a, S_v, M, M_a and M_v are updated. Similarly, using matrix M_v, compute $\beta(V_k, V_l)$ and $S_v^{'}(V_k, V_l)$ for each pair (V_k,V_l):

$$\beta(V_k, V_l) = \sum_{a=1}^{n_a} m_v(A_a, V_k).m_v(A_a, V_l) \tag{23}$$

$$S_v^{'}(V_k, V_l) = \rho_1.S_v(V_k, V_l) + \rho_2.\beta(V_k, V_l) \tag{24}$$

Then, find the pair (V_K, V_L) that corresponds to the maximum similarity:

$$(V_K, V_L) = \arg\max_{(V_k, V_l)}(S_v^{'}(V_k, V_l)) \tag{25}$$

If $\max(S_v^{'}(V_K, V_L))$ is higher than a threshold Thr_v, then merge the two clusters (see Table 2 for Thr_v). In this case, the matrices S_a, S_v, M, M_a and M_v are updated. During our experiments, τ_1, τ_2, ρ_1 and ρ_2 were respectively valued at $\frac{1}{2}$, 2, $\frac{1}{2}$, 2.

4. Next, return to second step of this algorithm. The three steps are repeated until the stopping criteria for both audio and video clustering have been reached. In this case, we compute the weighted co-occurrence matrix M'_{av} in terms of face size and lip activity detection using Eq. 19. Using M'_{av}, we can deduce the voice and/or the face of each person. Consequently, three types of clusters emerge: talking faces, non-talking faces and the voice-only clusters.

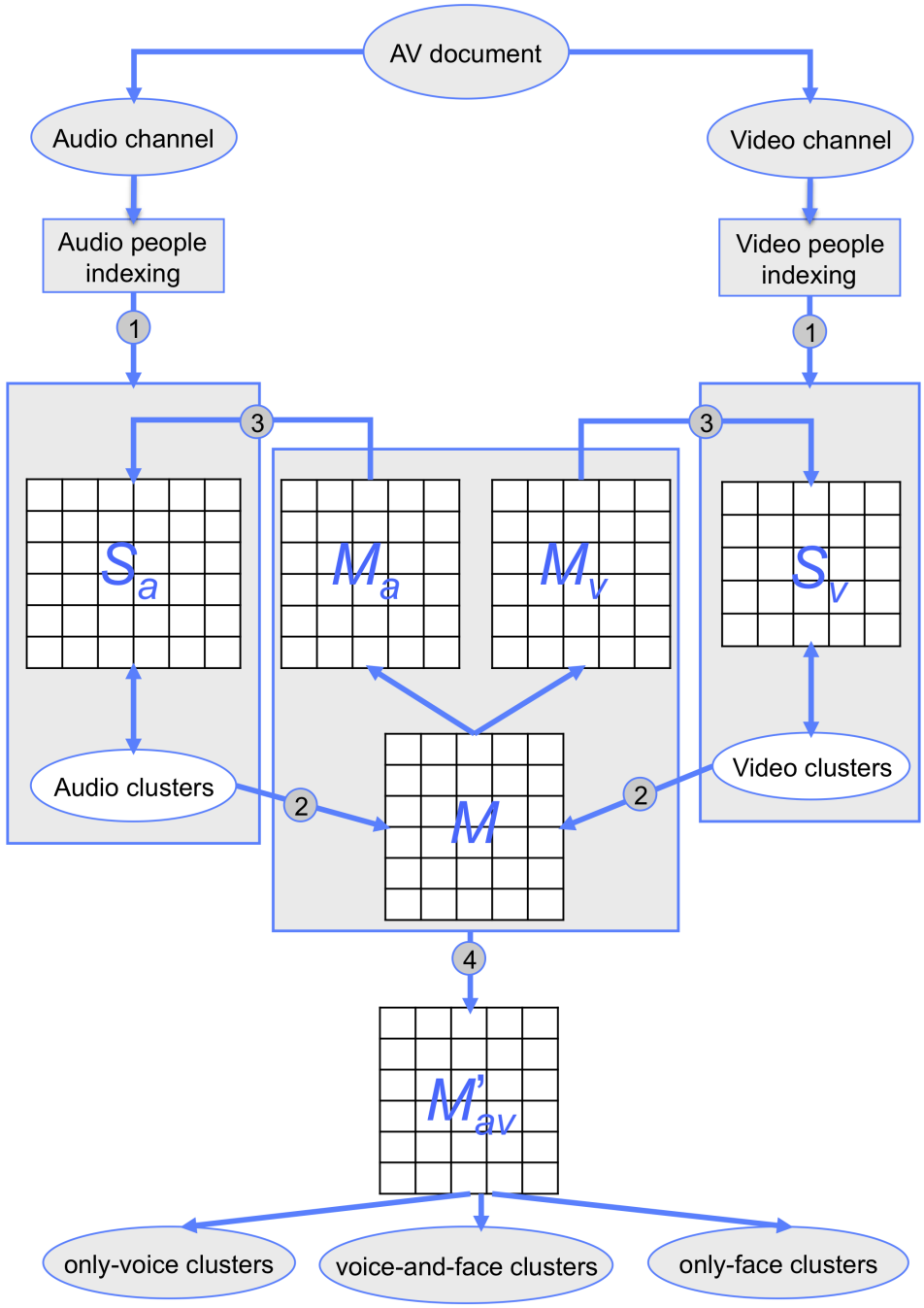

Figure 8: Architecture of the audiovisual people diarization system. S_a is the similarity matrix for audio clusters, S_v is the similarity matrix for video clusters, M is the co-occurrence matrix, M_a and M_v are the normalized co-occurrence matrices, and M_{av} is the final association matrix.

5 Experiments and results

Table 1 describes our audiovisual corpus of overall duration of 10.6 hours. This corpus is divided into three subsets: news, debates and movies. For each subset, the total duration,

Table 1: Details of the corpus.

	Dur.	Speech dur.	Ref. spkrs	Faces dur.	Ref. faces
News	4h05'	3h04'	311	1h48'	626
Debates	3h30'	2h41'	129	2h25'	311
Movies	3h05'	1h13'	128	1h03'	378
Total	**10h40'**	**6h58'**	**568**	**5h16'**	**1315**

Table 2: The set of thresholds used in all the experiments

Threshold	Value	Description
Thr_{RSP}	0.35	corresponds to the optimal stopping criterion of the face-based tracker.
Thr_1	0.41	corresponds to the stopping criterion that provides the optimal face-based clustering using SIFT matching.
Thr_2	3.20	corresponds to the stopping criterion that provides the optimal face-based clustering using Skin matching.
Thr_3	3.30	corresponds to the stopping criterion that provides the optimal clothing-based clustering using 3D-Histogram matching.
Thr_4	0.13	corresponds to the stopping criterion that provides the optimal clothing-based clustering using Texture matching.
λ_{BIC}	0.80	corresponds to the penalty coefficient that provides the optimal audio-people clustering using BIC matching.
Thr_{la}	6.80	is used for lip activity to decide whether a person is speaking or not.
Thr_a	0.50	corresponds to the stopping criterion that provides the optimal audio-people clustering using all audiovisual cues.
Thr_v	0.50	corresponds to the stopping criterion that provides the optimal video-people clustering using all audiovisual cues.

the speech duration, the number of speakers in the reference, the total duration of appearing faces, and the number of appearing faces in the reference are reported.

Table 2 sums up the thresholds used in these experiments as well as their corresponding value. They are trained on a development set of video of about 40 minutes.

5.1 Results of the audio people diarization

In this section, we measure the performance of the diarization system with and without video information. To do this, we use the diarization error rate[3] (DER). The output of a speaker diarization system consists of a list of speech segments described with starting time, ending time and speaker cluster name (this list is called the hypothesis). It is evaluated against a manually annotated ground truth (called reference). The evaluation performs an optimum one-to-one mapping between the hypothesis segments and the reference segments so that the total overlap time between the reference speaker and the corresponding mapped speaker cluster returned by the hypothesis is maximized. The DER is the sum of three errors: speech/non speech errors, where speech is present in the hypothesis but not in the reference ($False$ detection), non speech/speech errors, the contrary ($Missed$ detection),

[3]http://www.itl.nist.gov/iad/mig/tests/rt/

Table 3: DER of "Audio-only" and "Audiovisual" processing.

	Audio-only processing	**Audiovisual processing**
News	18.68%	15.85%
Debates	25.96%	14.89%
Movies	40.81%	39.70%
Overall DER	**25.35%**	**19.64%**

and speaker errors ($SpkrErr$), where the mapped reference is not the same as the speaker found by the system.

$$DER = SpkrErr + Miss + False \tag{26}$$

Table 3 shows that the overall weighted DER decreases from 25.35% to 19.64% when our audiovisual association is applied. For TV news, the gain is about 2.83%. For debates, the decrease of the error rate is very significant (from 25.96% to 14.89%). This can be explained by the fact that clustering based on audio information is more difficult for debates than for news, however, the use of video information resolves this problem. For movies, there is a poor gain of only 1.11% (from 40.81% to 39.70%). This can be explained by the fact that both audio and video error rates are high.

5.2 Results of the video people diarization

In this section, we evaluate our system in terms of face clustering. Although, [27] have defined a cost metric that computes the number of clicks a user would need to correctly label all images and this metric is fine for still images (e.g. personal photo albums), it is not suitable for face tracks in video because it fails to take into account the duration of face tracks. Therefore, we have defined a new metric that we call the "clustering error rate" (CER). This metric, inspired from the DER for audio, determines the optimal mapping between the hypothesis face clusters and the reference face clusters in terms of time:

$$CER = \frac{\sum\limits_{Allseqs} (dur(s) \times (min(N_R(s), N_S(s)) - N_C(s)))}{\sum\limits_{Allseqs} (dur(s) \times N_R(s))} \tag{27}$$

where for each sequence s: dur(s) is the duration of s, $N_R(s)$ is the number of people appearing in s according to the **reference**, $N_S(s)$ is the number of people appearing in s according to the **system**, and $N_C(s)$ is the number of **correct** matches, i.e. the number of correct corresponding matches between the two.

Table 4 shows the CER values before and after using the audio information with the overall CER decreasing from 19.75% to 17.22%. For TV news, the gain is 1.46% (from 9.10% to 7.64%). For debates, the gain is 3.32% (from 15.73% to 12.41%). For movies, the gain is 3.23% but the CER is still high (40.49%).

Table 4: CER of "video-only" and "Audiovisual" processing.

	video-only processing	**Audiovisual processing**
News	9.10%	7.64%
Debates	15.73%	12.41%
Movies	43.72%	40.49%
Overall CER	**19.75%**	**17.22%**

5.3 Results of the audiovisual association

In this section, we test the robustness of our proposed audiovisual association. To do this, we compute the precision and recall measures of "talking faces", "non-talking faces", and "off" voices. For each measure, the number of false positives, false negatives, true positives, and true negatives are computed with respect to positive and negative people annotated in the ground truth.

First, we evaluate our baseline system where only the co-occurrence matrix M_{av} is used. Then, we evaluate the benefits of using "lip activity" and "face size". Finally, we take into account the overall measures of our proposed systems and compare them to the system proposed in [40].

5.3.1 Results of our baseline system

Table 5 shows the detailed results of our baseline system obtained for different subsets (news, debates and movies) as well as the weighted overall scores. Talking faces are detected with a precision of 80% despite the low recall score (32%). On the other hand, non-talking faces are detected with a precision of 65% and a recall of 92%. Furthermore, "off" voices are detected with a precision of 43% and a recall of 55%. The results obtained for TV news are generally better than for debates and movies (except for "off" voices detection). This is mainly due to the fact that the purity of audio and video clusters is higher in the case of TV news data (as seen in previous sections).

Table 5: Results of our baseline system for audiovisual association: detection of talking faces, non-talking faces and "off" voices.

	Talking faces			**Non-talking faces**			**Voices-only**		
	Num.	Prec.	Rec.	Num.	Prec.	Rec.	Num.	Prec.	Rec.
News	132	87%	58%	565	86%	96%	82	44%	45%
Debates	78	65%	34%	387	75%	91%	52	57%	58%
Movies	52	90%	15%	354	23%	77%	60	30%	72%
Overall	**262**	**80%**	**32%**	**1306**	**65%**	**92%**	**194**	**43%**	**55%**

5.3.2 Results of the improved system

Here,we evaluate the benefits of adding either **lip activity** detection (S2), **face size** (S3), or both (S4). Table 6 shows that all results are higher than those of the baseline system.

Table 6: Results of the improved system.

	Talking faces			**Non-talking faces**			**Voices-only**		
	Num.	Prec.	Rec.	Num.	Prec.	Rec.	Num.	Prec.	Rec.
S2	**280**	**83%** (+3%)	**35%** (+3%)	**1292**	**68%** (+3%)	**95%** (+3%)	**177**	**45%** (+2%)	**62%** (+7%)
S3	**278**	**84%** (+4%)	**36%** (+4%)	**1294**	**67%** (+2%)	**93%** (+1%)	**182**	**46%** (+3%)	**62%** (+7%)
S4	**331**	**90%** (+10%)	**46%** (+14%)	**1232**	**72%** (+7%)	**96%** (+4%)	**120**	**78%** (+35%)	**70%** (+15%)

As the most interesting people in the document are generally those who appear and talk within that document, we detail in Fig. 9 the results of our proposed system on the task of "talking faces" detection. In TV news (Fig. 9.a), the precision increases from 87% (S1) to 92% (S4), and the recall increases from 58% (S1) to 80% (S4). In debates (Fig. 9.b), the overall gain is 15% for both precision (from 65% to 80%) and recall (from 34% to 49%). In movies (Fig. 9.c), S4 outperforms S1 by 7% for precision (from 90% to 97%) and by 8% for recall (from 15% to 23%).

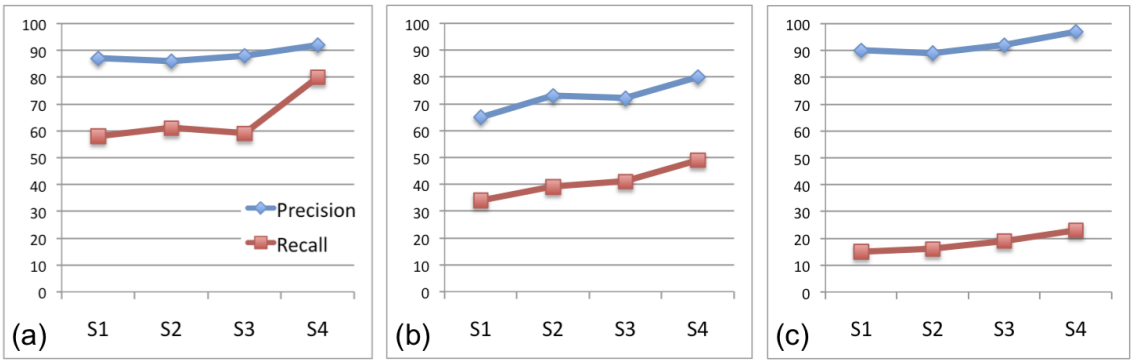

Figure 9: Results of our proposed system for the task of "Talking faces" detection in (a) news, (b) debates and (c) movies.

5.3.3 Comparison with the state-of-the-art system

In the final experiment, we compared our proposed system to the system proposed by Liu and Wang [40] (S0). To do this, for each system we computed the total precision and recall: the total precision (respectively recall) of a system is the sum of the precisions (respectively recalls) of talking faces, non-talking faces and "off" voices of that system.

Table 7 shows that our baseline system **(S1)** results in both higher precision and recall than the method proposed in [40] **(S0)**. This is due to the normalization of matrices (M_a and M_v). This table also shows that our system **(S4)** which uses both "lip activity" and "face size" outperforms **(S0)** by 14% for precision and 12% for recall.

Table 7: Comparison between the system proposed in [40] (S0) and our proposed system (S1, S2, S3, S4).

	S0	**S1**	**S2: S1 + Lip activity**	**S3: S1 + Face size**	**S4: S2 + Face Size**
Prec.	62%	65%	71%	67%	**76**%
Rec.	63%	67%	68%	69%	**75**%

5.4 Analysis of errors

After combining all the audio and video components, several different types of errors still remain. From the audio point of view, we have found that:

- In TV news, errors are often due to confusion between different people who can be heard with the same background noise. And sometimes, they are due to the dissimilarity between the different speech turns of the reporter who is either talking in the studio or in a noisy environment.
- In debates, errors are especially due to the high interaction rate between people.
- In movies, errors are due to the high variations in the background noise (music, indoors, outdoors, etc.), the short duration of speech turns, and the high interaction rate between actors.

From the video point of view, we have found that:

- In TV news, errors are especially due to confusion between small faces of similar size, similar lighting or clothes.
- In debates, errors are often due to TV reports that are shown during the program.
- In movies, they are due to variations in the lighting, poses and face sizes.

6 Conclusions

In this paper, we address the problem of audiovisual people diarization using both audio and video cues. After describing our contributions to cope with this problem by studying each medium separately, we present our proposed method for audiovisual association using a co-occurrence matrix as well as enhancements through additional modules such as face size and lip activity rate. In addition, we describe a framework that simultaneously improves

audio, video and audiovisual diarization output. The results obtained on a corpus of TV news, debates and movies show the robustness of this association method, and confirm the gains that one modality can bring to the other.

One drawback of our audiovisual diarization system is that there are many parametered thresholds that may not be optimal for all types of documents. Future work will focus on finding solutions to automatically compute the optimal thresholds for each type of document. In addition, we will focus on extending the work presented in this paper to "inter-documents" audiovisual people diarization, and on designing dynamic audiovisual models of people in a collection of documents.

7 Acknowledgments

This work was supported by a 3-year individual fellowship from the French Ministry of High Education and Research, and by the SODA project funded by the National French Research Agency (ANR).

References

[1] X. Anguera, C. Wooters and J. Hernando. Robust speaker diarization for meetings: ICSI RT06 evaluation system. *International Conference on Spoken Language Processing*, 2006.

[2] M. Andriluka, S. Roth and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[3] O. Arandjelovic and A. Zisserman. Automatic face recognition for film character retrieval in feature-length films. *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[4] A. Azarbayejani, T. Starner, B. Horowitz and A. Pentland. Visually controlled graphics. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1993.

[5] M. Bicego, A. Lagorio, E. Grosso and M. Tistarelli. On the use of sift features for face authentication. *Computer Vision and Pattern Recognition Workshop*, 2006.

[6] B. Bigot, I. Ferrané and J. Pinquier. Exploiting speaker segmentations for automatic role detection. An application to broadcast news documents. *International Workshop on Content-Based Multimedia Indexing*, 2010.

[7] S. Bozonnet, N. Evans and C. Fredouille. The LIA-EURECOM RT09 Speaker diarization system: anhancements in speaker modelling and cluster purification *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010.

[8] M. Cettolo and M. Vescovi. Efficient audio segmentation algorithms based on the bic. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[9] S.F. Chang, J. He, Y.G. Jiang, E. El Khoury, C. W. Ngo, A. Yanagawa and E. Zavesky. Columbia University/VIREO-CityU/IRIT TRECVID2008 High-Level Feature Extraction and Interactive Video Search. *TREC Video Retrieval Workshop,* NIST, 2008.

[10] U.V. Chaudhari, G.N. Ramaswamy, G. Potamianos and C. Neti. Audio-visual speaker recognition using time-varying stream. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003.

[11] U.V. Chaudhari, G.N. Ramaswamy, G. Potamianos and C. Neti. Information fusion and decision cascading for audio-visual speaker recognition based on time-varying stream reliability prediction. *IEEE International Conference on Multimedia and Expo*, 2003.

[12] S.S. Chen and P.S. Gopalakrishnan. Clustering via the bayesian information criterion with applications in speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998.
[13] W.T. Chu, Y.L. Lee and J.Y. Yu. Visual language model for face clustering in consumer photos. *ACM international conference on Multimedia*, 2009.
[14] G. Cinbis, J. Verbeek and C. Schmid. Unsupervised metric learning for face identification in TV video. *IEEE International Conference on Computer Vision*, 2011.
[15] C. Czirjek, S. Marlow and N. Murphy. Face detection and clustering for video indexing applications. *Advanced Concepts for Intelligent Vision Systems*, 2003.
[16] A. Dielmann. Unsupervised detection of multimodal clusters in edited recordings. *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2010.
[17] G. Doretto, T. Sebastian, P. Tu and J. Rittscher. Appearance-based person reidentification in camera networks: Problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing*, 2011.
[18] M. Everingham, J. Sivic and A. Zisserman. Hello! my name is... buffy – automatic naming of characters in TV video. *British Machine Vision Conference, BMVC06*, 2006.
[19] M. Everingham, J. Sivic and A. Zisserman. Taking the bite out of automated naming of characters in TV video. *Journal of Image and Vision Computing,* 2009.
[20] A.W. Fitzgibbon and A. Zisserman. On affine invariant clustering and automatic cast listing in movies. *ECCV '02: European Conference on Computer Vision*, 2002.
[21] C. Fredouille, S. Bozonnet and N. Evans. The LIA-EURECOM RT09 Speaker diarization system. *NIST Rich Transcription Workshop*, 2009.
[22] G. Friedland, H. Hung and Chuohao Yeo. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.
[23] G. Friedland, C. Yeo and H. Hung. Dialocalisation: Acoustic Speaker Diarization and Visual Localization as Joint Optimization Problem. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2010.
[24] S. Galliano, E. Geofrois, D. Mosterfa, J.F. Bonastre and G. Gravier. The ESTER phase II evaluation campaign for the rich transcription of the French broadcast news. *European Conference on Speech Communication and Technology*, 2005.
[25] S. Galliano, G. Gravier and L. Chaubard. The ester 2 evaluation campaign for the rich transcription of French radio broadcasts. *INTERSPEECH,* 2009.
[26] H. Gish, M.H. Siu and R. Rohlicek. Segregation of speakers for speech recognition and speaker identification. *International Conference on Acoustics, Speech, and Signal Processing*, 1991.
[27] M. Guillaumin, J. Verbeek and C. Schmid. Is that you? Metric learning approaches for face identification. *ICCV*, 2009.
[28] A. Hilsmann and P. Eisert. Tracking and retexturing cloth for real-time virtual clothing applications. *International Conference on Computer Vision/Computer Graphics CollaborationTechniques*, 2009.
[29] H. Hung and G. Friedland. Towards audio-visual on-line diarization of participants In group meetings. *Workshop on Multi-camera and Multi-modal Sensor Fusion,* 2008.
[30] S. Ioffe and D.A. Forsyth. Human tracking with mixtures of trees. *ICCV01*, 2001.
[31] G. Jaffré and P. Joly. Costume: A new feature for automatic video content indexing, *RIAO*, 2004.
[32] E. El Khoury, C. Senac and R. André-Obrecht. Speaker Diarization: Towards a more robust and portable system. *IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2007.
[33] E. El-Khoury, C. Senac and J. Pinquier. Improved speaker diarization system for meetings. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2009.
[34] E. El Khoury, C. Senac and P. Joly. Unsupervised segmentation methods of TV contents.

International Journal of Digital Multimedia Broadcasting, 2010.
[35] E. El Khoury, C. Senac and P. Joly. Face-and-clothing based people clustering in video content. *ACM International Conference on Multimedia Information Retrieval*, 2010.
[36] D. A. V. Leeuwen and M. Konecný. Progress in the AMIDA speaker diarization system for meeting data. *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*, 2008.
[37] C. Lerdsudwichai, M. Abdel-Mottaleb and A.N. Ansari. Tracking multiple people with recovery from partial and total occlusion. *Pattern Recognition*, 2005.
[38] Z. Liu, D. Gibbon, E. Zavesky, B. Shahraray and P. Haffner. A fast, comprehensive shot boundary determination system. *IEEE International Conference on Multimedia and Expo*, 2007.
[39] Z. Liu and Y. Wang. Major cast detection in video using both audio and visual information. *IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2001.
[40] Z. Liu and Y. Wang. Major cast detection in video using both speaker and face information. *IEEE Transactions on Multimedia*, 2007.
[41] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.
[42] B.S. Manjunath and W.Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996.
[43] T. H. Nguyen, H. Sun, S. Zhao, S. Z. Khine, H. D. Tran, T. L. Ma, B. Ma and E. S. Chng, H. Li. The IIR-NTU speaker diarization systems for RT 2009. *NIST Rich Transcription Workshop*, 2009.
[44] H.J. Nock, G. Iyengar and C. Neti. Speaker localisation using audio-visual synchrony: an ampirical study. In *CIVR: ACM International Conference on Image and Video Retrieval*, 2003.
[45] J. Peng and Q.X. Lin. Automatic classification video for person indexing. *Congress on Image and Signal Processing*, 2008.
[46] J. Philippeau, J. Pinquier and P. Joly. Intervenant classification in an audiovisual document. *International Conference on Signal Processing and Multimedia Applications*, 2006.
[47] J. Pinquier, J.L. Rouas and R. André-Obrecht. A fusion study in speech/music classification. *IEEE International conference on Acoustics, Speech and Signal Processing*, 2003.
[48] R.L. Plackett. Karl Pearson and the chi-squared test. *International Statistical Review*, 1983.
[49] J. Ramirez, J.M. Girriz and J. C. Segura. Voice activity detection. Fundamentals and speech recognition system robustness. *Robust Speech Recognition and Understanding*, 2007.
[50] B. Rosenhahn, U. Kersting, K. Powell, T. Brox and H.P. Seidel. Tracking clothed people. *Human Motion - Understanding, Modeling, Capture, and Animation*. Springer, 2007.
[51] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.
[52] J. Schmalenstroeer and R. Haeb-Umbach. Online Diarization of Streaming Audio-Visual Data for Smart Environments. *IEEE Journal of Selected Topics in Signal Processing*, 2010.
[53] M. A. Siegler, U. Jain, B. Raj and R.M. Stern. Automatic segmentation, classification and clustering of broadcast news audio. *DARPA Speech Recognition Workshop*, 1997.
[54] P. Sivakumaran, J. Fortuna and A.M. Ariyaeeinia. On the use of the bayesian information criterion in multiple speaker detection. *The 7th European Conference on Speech*

Communication and Technology (Eurospeech'01), 2001.
[55] A. F. Smeaton, P. Over and A. R. Doherty. Video shot boundary detection: Seven years of trecvid activity. *Computer Vision and Image Understanding*, 2009.
[56] R. Stiefelhagen, R. Bowers and J. Fiscus. Multimodal technologies for perception of humans: International Evaluation Workshops CLEAR 2007 and RT 2007. *Springer-Verlag, ser. Lecture Notes in Computer Science*, 2008.
[57] J.W. Sung, T. Kanade and D.J. Kim. Pose robust face tracking by combining active appearance models and cylinder head models. *International Journal of Computer Vision*, 2008.
[58] S. Tamura, K. Iwano and S. Furui. Multimodal speech recognition using optical-flow analysis for lip images. *Real World Speech Processing*, 2004.
[59] D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1993.
[60] B.T. Truong, C. Dorai and S. Venkatesh. New enhancements to cut, fade, and dissolve detection processes in video segmentation. *ACM international conference on Multimedia*, 2000.
[61] W. H. Tsai, S. S. Cheng, Y. H. Chao and H. M. Wang. Clustering speech utterances by speaker using eigenvoice-motivated vector space model. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.
[62] H.Vajaria, T.Islam, S.Sarkar, R.Sankar, and R.Kasturi. Audio segmentation and speaker localization in meeting videos. *ICPR'06: International Conference on Pattern Recognition*, 2006.
[63] V. Vezhnevets, V. Sazonov and A. Andreeva. A survey on pixel-based skin color detection techniques. *in Proc. Graphicon*, 2003.
[64] P. Viola, M.J. Jones and D. Snow. Detecting pedestrians using patterns of motion and appearance. *ICCV '03: IEEE International Conference on Computer Vision*, 2003.
[65] P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 2004.
[66] M.H. Yang. Face Detection. *Encyclopedia of Biometrics*, Springer, 2009.
[67] B. Zhou and J.H.L. Hansen. Efficient audio stream segmentation via the combined $T2$ statistic and the bayesian information criterion. *IEEE Trans. Speech Audio Processing*, 2005.
[68] X. Zhu, C. Barras, L. Lamel and J.L. Gauvain. Multi-stage speaker diarization for conference and lecture meetings. *Multimodal Technologies for Perception of Humans*, Springer, 2008.