**Noname manuscript No.**
(will be inserted by the editor)

# Naming multi-modal clusters to identify persons in TV Broadcast

**Johann Poignant**[1,2] · **Guillaume Fortier**[1,2] · **Laurent Besacier**[1,2] · **Georges Quénot**[2,1]

Received: date / Accepted: date

**Abstract** Persons' identification in TV broadcast is one of the main tools to index this type of videos. The classical way is to use biometric face and speaker models, but, to cover a decent number of persons, costly annotations are needed. Over the recent years, several works have proposed to use other sources of names for identifying people, such as pronounced names and written names. The main idea is to form face/speaker clusters based on their similarities and to propagate these names onto clusters.

In this paper, we propose a method to take advantage of written names during the diarization process, in order to both name clusters and prevent the fusion of two clusters named differently. First, we extract written names with the LOOV tool [27]; these names are associated to their co-occurring speaker turns / face tracks. Simultaneously, we build a multi-modal matrix of distances between speaker turns and face tracks. Then agglomerative clustering is performed on this matrix with the constraint to avoid merging clusters associated to different names. We also integrate the prediction of few biometric models (anchors, some journalists) to directly identify speaker turns / face tracks before the clustering process.

Our approach was evaluated on the REPERE corpus and reached an F-measure of 68.2% for speaker identification and 60.2% for face identification. Adding few biometric models improves results and leads to 82.4% and 65.6% for speaker and face identity respectively. By comparison, a mono-modal, supervised person identification system with 706 speaker models trained on matching development data and additional TV and radio data provides 67.8% F-measure, while 908 face models provide only 30.5% F-measure.

**Keywords** Multimodal fusion · videoOCR · face and speaker identification · TV Broadcast

[1] Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
[2] CNRS, LIG, F-38000 Grenoble, France
E-mail: firstname.lastname@imag.fr

## 1 Introduction

Knowing "who is seen" and "who is speaking" in TV Broadcast programs is very useful to provide efficient information access to large video collections. Therefore, person identification is important for search and browsing in this type of data. Conventional approaches are supervised with the use of face and voice biometric models built on manually annotated data. However, these approaches face two main problems: 1) manual annotations are very costly because of the great number of recognizable persons in video collections; 2) due to that lack of prior knowledge on persons appearing in videos (except for journalists and anchors), a very large amount of a priori trained models (several hundred or more) is needed for covering only a decent percentage of persons in a show.

A solution to these problems is to use other information sources for naming persons in a video. Two modalities, intrinsic to the video (see figure 1), can provide the name of person present in TV Broadcast: pronounced names and names written on the screen to introduce the corresponding person.


Fig. 1: Pronounced names and written names in a TV Broadcast video

Most approaches that use these sources can be decomposed into three following steps:

1. Segmentation of speech/face tracks into clusters;
2. Extraction of hypothesis names from the video;
3. Association between hypothesis names and clusters.

The first step is the process of partitioning the audio stream into speaker turns / extracting face tracks from the image; and then grouping them into homogeneous clusters without prior knowledge on the speakers' voice or face appearance. Each cluster must correspond to only one person and *vice versa*.

For the second step, the quality of the names extraction depends on the source. Most state-of-the-art approaches rely on pronounced names due to the poor quality of written names transcription observed in the past.

The third step (mapping) also depends on the name source used. Pronounced names can refer to a person visible (or speaking) when the name is pronounced in the current, next or previous shot (or speaker turn); whereas a written name in a title block introduces the corresponding person in the

screen (or in the current audio stream).

Improving the quality of video over the recent years has changed the usability of written names extracted automatically. We describe in [27] a method to extract these names with very few transcription errors. Then, we have proposed, in [30], [29] and [32] a fair comparison between written names and pronounced names on their ability to provide names of persons present in TV Broadcast. We have shown that for videos where written names are available, like for news TV or debat show, it is more interesting to try to name persons with written names. Pronounced names show a potential with manual annotation but speech transcription and named-entity detection errors reduce this potential for naming persons with good precision.

This paradigm shift enables us to develop new methods to identify people visible/speaking using the written names as a primary source of identity

The main goal of this paper is to describe a process that uses only written names on screen as unique source of names, that can jointly and simultaneously name speakers and faces without a priori knowledge of the target TV show (person present, ad-hoc rules, ...). Secondly, it is shown that this process can easily integrate, as complementary source, some supervised biometric models (persons of the classical casting of a show: anchors, some journalists).

After this introduction the outline is as follows: after presentation of the related works in section 2, section 3 is dedicated to the description of our multi-modal person identification system. Section 4 shows how we integrated biometric models inside this system. In section 5 we describe the experimental setup and in section 6 we analyze results obtained on the REPERE corpus [20]. Finally, we conclude this paper and give some perspectives.

## 2 State-of-the-art

Until very recently, research works dealt only with one of the two tasks (we could not find early attempts were both face and speaker identification tasks were treated simultaneously).

For speaker identification, the closest modality for extracting the names of speakers was used: the pronounced names from speech transcription. We can mention the works of *Canseco et al.* [9] and [10] as the first paper that did not use biometric models to identify speakers. They set manually linguistic patterns to determine a link between a pronounced name and a speech segment. Following works improved this idea: [37] replaced manual rules by learning sequence of n-grams with associated probabilities, [23], [13] and [19] used semantic classification trees to calculate the probabilities, [24] replaced the decision by belief functions. However, in all the above studies, the identification error rate is very high when automatically recognized (and noisy) pronounced names are used as source of naming information.

Written names were first used for a face identification task in broadcast news ([36], [18]), but due to a high word error rate (respectively 52 % and 65 %), these names were detected and corrected with the help of a dictionary (limiting identification to a closed list of persons). Despite these corrections, the error rate was still very high (45 % after correction in [18]) and consequently greatly limited the use of this source of information. Later, *Yang et al.* [39], [40] also tried to use written names, but again, the video OCR system [35] used to process the overlaid text produced highly erroneous results (e.g. "Newt Gingrich" was recognized as "nev j ginuhicij"). Their studies were limited to a single show (ABC World News Tonight) and they only tries to label faces in monologue-style speech

We can also cite *Pham et al.* [25], [26], which first extracted a list of names from speech transcription. Then, they manually associated a set of face images to these names. These identities are then propagated to the entire video from the similarity between faces. They have therefore concentrated their works on the face identification and thus did not taken advantage of the video multimodality with the use of speakers as indice for the propagation. Indeed, in [21] *Khoury et al* show that a multimodal audio-visual diarization obtains better results than monomodal diarization. In our work, we want to add an extra dimension to the multimodal diarization by integrating the knowledge of written names to both identify audio-visual clusters and also help the diarization process.

In 2011 began the REPERE challenge [20], it aimed at supporting research on person recognition in multi-modal conditions. To assess the technology progress, annual evaluation campaigns were organized from 2012 to 2014. In this context, the REPERE corpus [16], a French video corpus with multimodal annotation, was developed. 3 consortiums composed of multiple teams (including ourselves) participated to this challenge. Thanks to this corpus, new research contributions appeared using written names on screen wisely.

We proposed an unsupervised speaker/face identification system ([28] and [6]) based only on written names as source of names (extracted using the tool LOOV [27]) in TV broadcast. The main idea was to build mono-modal clusters (faces or speakers) and to associate written names to these clusters based on their co-occurrences (un-supervised approach). In this former work, faces and speakers were treated separately. Supervised biometric models were also integrated in the person identification system to boost the performance: a classifier was trained to take the decision to name speaker/face. The unsupervised system was extended in [7], [31] with the modification of the agglomerative clustering process. This process integrated directly the knowledge of written names to both identify clusters and also to prevent the merging of clusters named differently. This actual journal paper is somehow an extension of [31] but clustering is no longer mono-modal (faces and speakers are treated jointly).

Another work that is worth being mentioned is using Integer Linear Programming (ILP) for speech clustering [5], [8]. The main idea is to replace the classical agglomerative BIC clustering by an ILP clustering and at the same

time integrating written names to identify speech clusters. First, multi-modal similarity graph is built, were intra-modal links correspond to the similarity of mono-modal elements (speech turns: BIC distance, written names: identical or not) and cross-modal links correspond to the temporal co-occurrence between written names and speech turns. As a written name has a high probability to match the current speaker, identification of speech turns via the ILP solver is equivalent to find the less expensive way to connect names and speech turns. The main limitation of this method is the large computation time for solving ILP (as well as the basic cross-modal link used). For this reason, ILP clustering is not used in this proposed contribution.

In [14], speakers are named in the first place and then identities are propagated to visible persons. Speakers are named by the propagation of written names and pronounced names and also with biometrics speaker models. After a face diarization step, written names and speaker identities are propagated to faces based on their co-occurrence. Authors also integrate a set of manual rules for a specific show to post-process their output (e.g. *if one of the anchors is speaking and two faces are detected, then the second face is given the identity of the second anchor*). They extended these works in [3], [34] with the use of automatic scene analysis (camera identification and scene classification as studio or report). This system needs additional annotations (scene type, camera position) for a specific show. Once a camera has been identified, they can deduct those who are visible on the screen (e.g., if the camera filming the anchor has been identified, they infer that the anchor is visible in screen). Finally, [4] proposed to integrate a lip activity detector to propagate speakers identities to face. Again, rules are used to propagate a name to a speaker/face. In our work, we try to keep a generic method (no ad-hoc camera identification for specific shows) but the idea to use lip activity detector is kept as a way to obtain reliable cross-modal links.

Last but not least, *Gay et al.* [15] proposed to propagate written names onto multi-modal speaker/face clusters. First, speakers and face diarization are performed in parallel, then speaker and face clusters are grouped based on their co-occurrence. They are associated to written names with two methods. The first one relies on co-occurrence information between written names and speaker/face clusters, and rule-based decisions which assign a name to each mono-modal cluster. The second method uses a Conditional Random Field (CRF) which combines different types of co-occurrence statistics and pairwised constraints to jointly identify speakers and faces.

Our proposed approach is presented in the next section. It differs from the previous works by the fact that we try to propose a generic (no ad-hoc approaches dedicated to specific shows), fast (no ILP, simultaneous face and speaker id), scalable (evaluated on a large database, with different type of shows and a large and open set of person to recognize), using advanced cross-modal links between faces and speakers (lip-activity detection) and with a multimodal diarization process that integrated the knowledge of written names (extracted automatically).

## 3 Multi-modal person identification

### 3.1 System overview

Figure 2 shows a global overview of our proposition. From a video, we extract names written in a title box. In parallel, we split the audio signal into speaker turns and compute a distance matrix $D_{tt'}$ based on the BIC criterion [12].

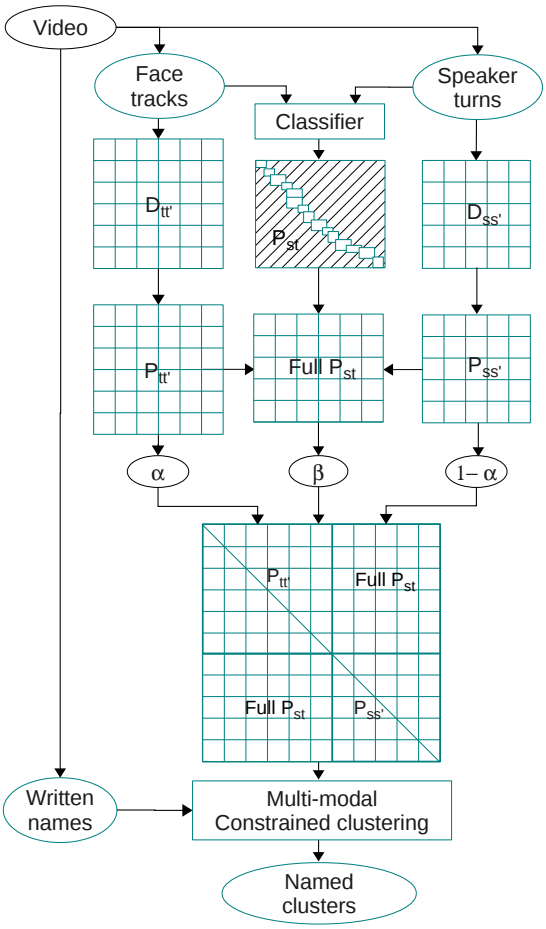


Fig. 2: System overview

For the image stream, after a face detection and tracking step, we find facial landmarks [38] and then compute an HoG descriptor on these points. These descriptors are projected using the LDML approach [17]. The matrix $D_{ss'}$ correspond to the l2 distance between descriptors projected. From these distances we calculate the probability that a face track (respectively a speaker

turn) correspond to another ($P_{tt'}$, respectively $P_{ss'}$). This probability is based on the confidence score of two logistic regressions trained with positive and negative pairs of speakers and faces.

In parallel we calculate the probability that a speaker correspond to a co-occurring face track ($P_{st}$). This last matrix is enriched with distances between speaker turns and face tracks that do not co-occur but could be linked by transitivity through a third face track / speaker turn ($FullP_{st}$). More details on this aspect will be given later in the paper.

These three probability matrices are combined into a big multi-modal matrix using weights $\alpha$ and $\beta$ to give more or less importance to a modality. Finally an agglomerative clustering is performed on this multi-modal matrix. The goal of this clustering is to merge all face tracks / speaker turns of same person into a unique cluster. This clustering early integrates the knowledge of written names to identify clusters and also to prevent wrong merges (avoiding the fusion of clusters named differently).

For clarity, notations used in the rest of the paper are introduced in table 1.

| ID | link each element to its identity |
|---|---|
| $\mathcal{N}$ | set of written names |
| $\mathcal{O}$ | set of written name occurrences |
| $\mathcal{S}$ | set of speaker turns |
| $\mathcal{T}$ | set of face tracks |
| $\mathcal{G}$ | set of clusters |
| $\mathcal{K}$ | set of named clusters |
| $\mathcal{U}$ | set of unknown clusters |
| h | function from $\mathcal{O}$ to $\mathcal{N}$ |
| q | function from $\mathcal{S}$ to $\mathcal{T}$ |
| f | function from $\mathcal{G}$ to $\mathcal{O}$ |

Table 1: Notations used in the rest of the paper

### 3.2 Written names (WN) extraction

To detect the names written on the screen used for introducing a person, a detection and transcription system is needed. For this task, we used LOOV [27] (LIG Overlaid OCR in Video). This system has been previously evaluated on another broadcast news corpus with low-resolution videos (352×288). We obtained a character error rate (CER) of 4.6% for any type of text and of 2.6% for names written on the screen to introduce a person.

From the transcriptions, we use a simple technique in order to detect the spatial positions of title blocks. This technique compares each transcript with a list of famous names. This list (175k names) has been extracted from Wikipedia pages using associated tags corresponding to person names. It has been completed with the names of person present in the train part of the REPERE

corpus (see section 5.1)). Whenever a transcription corresponds to a famous name, we add its spatial position to a list. With the repeating positions in this list we find the spatial positions of title blocks used for introducing a person. However, these text boxes detected do not always contain a name. A simple filtering based on some text questions (does the first word correspond to a first name? is the text sequence longer than a five words? is the name is near from a famous name in terms of edit distance ? ...) allows us to filter false positives. Transcription errors are corrected using the Wikipedia list when the edit distance is small (based on a ratio of the Levenshtein distance).

In table 2 we evaluate how a written name extracted automatically corresponds to the current speaker and to the visible face, i.e. when a name is written, we hypothesize that the corresponding person speaks and is visible.

| Modality | %P | %R | %F |
|---|---|---|---|
| Face | 93.3 | 7.4 | 13.7 |
| Speaker | 92.7 | 9.1 | 16.5 |

Table 2: Speaker and face identification based on written names only on REPERE train corpus (described in the section 5.1)

We can see that of course the recall is very low since names are not always written, but the precision is very high which means that we can assume the strong hypothesis: when a name is written in a title box, the speaker and one of the faces appearing on screen has a great probability to correspond to this name.

To use these names, we define the set of names $\mathcal{N}$ and the set of name occurrences $\mathcal{O}$:

$$\begin{aligned} \mathcal{N} &= \{a, b, \ldots, n\} \\ \mathcal{O} &= \{o_i\} \end{aligned} \tag{1}$$

These two sets are linked using the application $h : \mathcal{O} \to \mathcal{N}$, defined by:

$$h(o_i) \in \mathcal{N} \tag{2}$$

### 3.3 Building the multi-modal probability matrix

To propagate the name of an element $e$ (speaker turn or face track) to another element $e'$, we need to know their similarity in terms of biometric measures. The goal is to have a high probability $p_{ee'}$ when $ID(e) = ID(e')$ where $ID(x)$ correspond to the real identity of $x$, and a low probability in other cases.

*3.3.1 Score between speaker turns ($D_{ss'}$ and $P_{ss'}$ matrices)*

We first split the audio signal into acoustically homogeneous segments $\mathcal{S}$:

$$\mathcal{S} = \{s_1, s_2, \ldots, s_M\} \tag{3}$$

Then, we calculate a matrix of similarity scores between each pair of segments using the BIC criterion [12]. Segments are modeled with one Gaussian with full covariance matrix $\sum$ trained on the D = 12-dimensional Mel Frequency Cepstral Coefficients (MFCC) and energy. $\Delta\text{BIC}_{s,s'}$ defines the similarity $d_{ss'}$ between two segments $s$ and $s'$:

$$\begin{aligned} d_{ss'} = (n_s + n_{s'}) \log |\Sigma| - n_s \log |\Sigma_s| - n_{s'} \log |\Sigma_{s'}| \\ - \frac{1}{2} \cdot \lambda \cdot \left( D + \frac{1}{2} D \left( D + 1 \right) \right) \log \left( n_s + n_{s'} \right) \end{aligned}$$

where $n_k$ is the number of samples in segment $k$ and $\lambda$ the penalty weighting coefficient. We combine all distances into the similarity matrix $D_{ss'}$. It is worth mentioning that the matrix is not updated after each merging of clusters, as this is usually the case for regular BIC clustering.

We are aware that agglomerative clustering based on BIC distance is less efficient than clustering with CLR distance [1] or *I-vector*+ILP [33] but our goal, here, is to integrate this distance into a multi-modal matrix between faces and speakers. We therefore do not use directly this distance but transform it into a probability using logistic regression trained on a subset of the REPERE corpus (which will be detailed later on).

$$p_{ss'} = p(ID(s) = ID(s') \mid d_{ss'}) \tag{4}$$

*3.3.2 Scores between face tracks ($D_{tt'}$ and $P_{tt'}$ matrices)*

A face detection and tracking step build the face tracks $\mathcal{T}$:

$$\mathcal{T} = \{t_1, t_2, \ldots, T_M\} \tag{5}$$

A track corresponds to a sequence of face images of a person that appear on consecutive frames in a shot. The tracking is performed using particle filtering approach [2], initialized from face detections. The first frame of each shot, and every subsequent fifth frame is scanned and face tracks are initialized from frontal, half-profile and full-profile face detections. Tracking is performed in an online fashion, i.e., using the state of the previous frame to infer the location and head pose of the faces in the current frame. The face tracking was obtained by one of our partner in REPERE. So we did not control the detailed settings used. Nevertheless, they are given here, just for information.

On every face image of a track, our face detector identifies seven landmarks on the face [38], around the eyes, the nose and the mouth. Once the landmarks are detected, faces are aligned using an affine transformation, and an HoG descriptor is computed around each of the seven facial landmarks. The descriptor

quantizes local image gradients into 10 orientation bins, and computes a gradient orientation histogram for each cell in a $7 \times 7$ spatial grid over image region around the landmark. The global descriptor of an image concatenates the local gradient orientation histograms to form a $9 \times 10 \times 7 \times 7 = 4410$ dimensional feature vector per face (9 landmarks $\times$ 10 orientation bins $\times$ a grid of $7 \times 7$ spatial bins).

These descriptors are projected onto a 100 dimensional descriptor using Logistic Discriminant Metric Learning approach (LDML) [17]. The final descriptor of a track corresponds to the median elements of each part of all descriptors of this track. $l2$ distance between two projected descriptors provides us the distance $d_{tt'}$ between two tracks $t$ and $t'$.

$$d_{tt'} = l2(HoG_t, HoG_{t'}) \tag{6}$$

We also use logistic regression to compute the probability that two face tracks correspond to the same person.

$$p_{tt'} = p(ID(t) = ID(t') \mid d_{tt'}) \tag{7}$$

*3.3.3 Multi-modal Score ($P_{st}$ and $FullP_{st}$)*

To propagate a name from a speaker to a face or vice versa, we link $\mathcal{T}$ and $\mathcal{S}$ with the function $q\colon \mathcal{S} \to \mathcal{T}$ and $q'\colon \mathcal{T} \to \mathcal{S}$:

$$\begin{aligned} q(s) &= \{t \in \mathcal{T} \mid t \text{ co-occur with } s\} \\ q'(t) &= \{s \in \mathcal{S} \mid s \text{ co-occur with } t\} \end{aligned} \tag{8}$$

We first calculate the probability $p_{st}$ that a speaker turn corresponds to one of its co-occurring face tracks. This probability is based on a descriptor $desc_{st}$ built with a basic lip activity detector (difference of color histogram between two consecutive frames around the lip area of a face). However, on several faces, it is difficult to correctly positioning the lip on the face. Therefore, we also use other spatio-temporal features such as:

- size of the face
- centrality of the face on the screen
- co-occurring duration between the face track and the speaker turn
- feature indicating if the face moves (or not) in the image
- feature indicating if the face is on the front
- number of face tracks that co-occur the speaker turn
- number of shots that appear during the speaker turn

We train logistic regression with all these features to obtain the probability that a face corresponds to the co-occurring speaker.

$$p_{st} = p(ID(s) = ID(t) \mid desc_{st}) \tag{9}$$

With the above equation, we only link speaker turns and face tracks that co-occur. Additional distances might be obtained by transitivity through a

third element. To obtain these distance between a speaker turn $s$ and a face track $t \notin q(s)$ (that do not co-occur) we enrich the matrix using transitive elements ($FullP_{spk/face}$ matrix) by computing two types of score:

- $pt'_{st}$ via a third face track $t' \in q(s)$
- $ps'_{st}$ via a third speaker turn $s' \in q'(t)$

As several face tracks $\{t'_1, ..., t'_n\}$ (respectively $\{s'_1, ..., s'_m\}$) can co-occur $s$ (respectively $t$) we find the best transitive face track and speaker turn to compute these scores:

$$\begin{aligned} pt'_{st} &= \max_{t' \in q(s) \text{ and } p_{st'} > 0.5} \left(\frac{p_{st'} + p_{t't}}{2}\right) \\ ps'_{st} &= \max_{s' \in q'(t) \text{ and } p_{s't} > 0.5} \left(\frac{p_{ss'} + p_{s't}}{2}\right) \end{aligned} \tag{10}$$

If neither $t'$ (respectively $s'$) can be used then $pt'_{st}$ (respectively $ps'_{st}$) is not defined.

To better understand, let's take the toy example in the figure 3. Two timelines are drawn in this figure, one for speaker turns and one for face tracks. To compute the score between $s_5$ and $t_2$ we should find the best path via a transitive face track, here $t_5$ or $t_6$, and the best path via a transitive speaker turn, here $s_2$ or $s_3$.

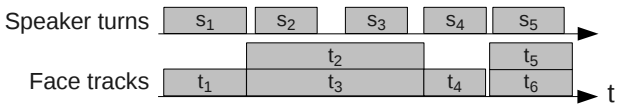


Fig. 3: Toy example for transistive score

Then the two parts of the transitive score are defined by:

$$\begin{aligned} pt'_{s5t2} &= \max\left(\frac{p_{s5t5} + p_{t5t2}}{2}, \frac{p_{s5t6} + p_{t6t2}}{2}\right) \\ &\text{with} \quad p_{s5t5} > 0.5 \text{ and } p_{s5t6} > 0.5 \\ ps'_{s5t2} &= \max\left(\frac{p_{s5s2} + p_{s2t2}}{2}, \frac{p_{s5s3} + p_{s3t2}}{2}\right) \\ &\text{with} \quad p_{s2t2} > 0.5 \text{ and } p_{s3t2} > 0.5 \end{aligned} \tag{11}$$

Finally the score is:

$$\begin{aligned} p_{st} &= \frac{pt'_{st} + ps'_{st}}{2} && \text{if } t' \text{ and } s' \text{ are defined} \\ p_{st} &= pt'_{st} && \text{if } t' \text{ only is defined} \\ p_{st} &= ps'_{st} && \text{if } s' \text{ only is defined} \\ p_{st} &= \text{is not defined} && \text{if there is no } t' \text{ nor } s' \text{ defined} \end{aligned} \tag{12}$$

*3.3.4 Combine all matrices into a multi-modal matrix*

These three probability matrices are combined into a big multi-modal matrix using the weights $\alpha$ and $\beta$. As in an agglomerative clustering process, we find iteratively the two nearest clusters (i.e. with the highest probability to be the same person in the matrix), these weights can delay or advance the fusion of some elements of a sub-matrix relative to the two other sub-matrices.

## 3.4 Multimodal constrained clustering

As already stated, when a name is written on the screen, there is a very high probability that the current speaker and one of the co-occurring face tracks correspond to this name. Therefore, we use the information provided by written names during the clustering process to name clusters and also to constrain the clustering process (prevent the fusion of two clusters with different associated names).

The main idea is that before clustering, we associate each written name to co-occurring speaker turns and face tracks. Then, a regular agglomerative clustering (based on similarity) is performed with the constraint that merging two (already named) clusters without at least one name in common is forbidden.

We divide this process into three steps:

1. **Initialization of the clustering**: prior to the clustering, we create links between speaker turns/face tracks and written names.
2. **Constraints on the clustering**: during the agglomerative clustering based on the multi-modal matrix, we prevent some merging to avoid the propagation of wrong names on already named clusters.
3. **Update after each merge**: merging two clusters can change the link between written names and clusters; so these links need to be updated. We also recalculate the scores between the new cluster (created by the merging) and all other clusters.

1) **Initialization of the clustering**

As clustering will merge speaker turns / face tracks into clusters, we define the set $\mathcal{G}$ of clusters. A cluster corresponding to a subset of $\mathcal{S} \cup \mathcal{T}$. Before the clustering, there is only one speaker turn/face track per cluster. Therefore, initially $\mathcal{G}$ is the set of singletons of $\mathcal{S} \cup \mathcal{T}$:

$$\mathcal{G} = \{\{e\}, e \in (\mathcal{S} \cup \mathcal{T})\} \tag{13}$$

Then, we create links between the two modalities with the function $f : \mathcal{G} \to P(\mathcal{O})$ with $P(\mathcal{O})$ corresponding to all partitions of $\mathcal{O}$, defined by

$$f(g) = \{o \in \mathcal{O} \mid o \text{ co-occur with } g\} \tag{14}$$

After that, $\mathcal{G}$ is divided into two subsets:

$$\begin{array}{l} \mathcal{K} = \{g \in \mathcal{G} \mid f(g_i) \neq \{\varnothing\}\} \\ \mathcal{U} = \mathcal{G} \backslash \mathcal{K} \end{array} \tag{15}$$

$\mathcal{K}$ corresponds to the set of clusters associated to at least one written name and $\mathcal{U}$ corresponds to the set of unknown clusters.

We consider that a speaker turn co-occurring with a written name must be identified by this name. As a speaker turn can co-occur with several names (this happens sometimes in our video collection when two or more names are written on screen at the same time), we associate all these names to the speaker turn and we assume that the name of the speaker is one of them. For faces, since several heads can appear at the same time, we do not use the same assumption. For each face co-occurring with a written name we create an association without being sure that the association is good. Therefore we also divide $\mathcal{K}$ into 2 subsets ($\mathcal{K} = \mathcal{K}_1 \cup \mathcal{K}_0$):

- $\mathcal{K}_1$: set of clusters for which we are sure to have the corresponding name. At the beginning, this set only contains speaker turns.

$$\mathcal{K}_1 = \{g \in \mathcal{K} \cap \mathcal{S}\} \tag{16}$$

- $\mathcal{K}_0$: set of clusters for which we are not sure to have the corresponding name. At the beginning, this set contains only face tracks.

$$\mathcal{K}_0 = \{g \in \mathcal{K} \cap \mathcal{T}\} \tag{17}$$

With links established between face tracks/speaker turns and written names, we perform an agglomerative clustering of elements of the set $\mathcal{G}$ with the multi-modal distance matrix.

The aim of this clustering is to find the equivalence classes that minimize the identification error rate, but also to associate to each cluster a unique name; this goal is formalized as:

$$card(\{h(o) \mid o \in f(g)\}) = 1 \tag{18}$$

2) **Constraints on the clustering**

We use the relationship between clusters and written names to constrain this agglomeration. Thus, two clusters $g$ and $g'$ of $\mathcal{K}_1$ (i.e. already named clusters) cannot be merged if:

$$\nexists (o \in f(g), o' \in f(g')) \mid h(o) = h(o') \tag{19}$$

which means they cannot be merged if they do not have a name in common among all of their associated names.

3) **Update after each merge**

After each agglomeration step, the merging of two clusters $g$ and $g'$ in a cluster $g''$ changes the function $f$. Four cases can be listed for this function:

– The two clusters are in $\mathcal{K}_1$, then:

$$f(g'') = \{o \in f(g), o' \in f(g') \mid h(o) = h(o')\} \tag{20}$$

– The two clusters are in $\mathcal{K}_0$, then:

$$f(g'') = f(g) \cup f(g') \tag{21}$$

– Only $g \in \mathcal{K}_1$ or $g \in \mathcal{K}_0$ and $g' \in \mathcal{U}$ (or vice versa) then:

$$f(g'') = f(g) \text{ (respectively } f(g") = f(g')) \tag{22}$$

– None is in $\mathcal{K}$, then the function $f$ is unchanged.

After each merge, we recalculate the score between the new cluster $g''$ and all other clusters $g$ of $\mathcal{G}$. This new score is the average score between elements of each cluster:

$$score(g, g'') = \frac{\sum\limits_{e \in g, e' \in g"} p_{ee'}}{card(g'') * card(g)} \tag{23}$$

We stop this process according to a threshold learned on the training set. At the end, clusters in $\mathcal{K}$ can be named, others are still unknown.

## 4 Adding information from supervised biometrical models

Some person names rarely appear in the extracted written names; this is mostly the case for anchors and some journalists. Therefore, we used a subset of the training data to build biometric speaker and face models. These models are used to directly name some speaker turns and face tracks (if the confidence of the identification is big enough) before the clustering step (see figure 4).

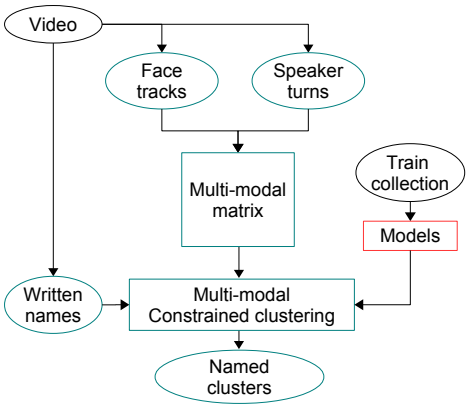


Fig. 4: Adding supervised biometric models into the system

### 4.1 Supervised biometric speaker models (Models_spk_706)

The supervised speaker identification system relies on two steps:

1. Diarization step: an agglomerative clustering based on the BIC criterion provide pure clusters, then a second clustering stage is proceed using cross-likelihood ratio (CLR) as distance between the clusters [1].
2. Identification step: we used the GSV-SVM system described in [22]. This system uses the super-vector made of the concatenation of the UBM-adapted GMM means to train one Support Vector Machine classifier per speaker. Each cluster provide by the diarization step is scored against all gender-matching speaker models, and the best scoring model is chosen if its score is higher than the decision threshold.

Three data sources were used for training models for 706 speakers in our experiments: a part of the REPERE training set, the ETAPE training and development data[1] and additional French politicians data extracted from French radios.

### 4.2 Supervised biometric face models (Models_face_908)

We also built 908 face models in the same way on a part of the REPERE training set. Due to face annotation cost, only 1 image every 10 seconds are annotated in this corpora. After the face tracking step we propagate these annotations along the face tracks. We therefore obtain several annotation face images for a person.

To extract a unique descriptor per person, we use the same way as for face tracks (find landmark, compute HoG descriptors, project these descriptor with the LDML matrix and use the median descriptor). Then, all face tracks of the test set is score against all face models with the $l2$ distance between the 2 median HoG descriptors projected and the best scoring model is chosen if its score is higher than the decision threshold..

### 4.3 (Optionally) Reducing the number of models

Having too many models can generate missidentification on the test set. Therefore we also experiment an experimental setting where we reduce these initial models sets using the information related to the classical casting of each show (anchor, some reporters). For speaker identity we select 84 models and for face identity we select 52 models. In the rest of the paper we call these reduced systems Models_spk_84 and Models_face_52.

[1] http://www.afcp-parole.org/etape.html

## 5 Experimental setup

The REPERE challenge [20] is an evaluation campaign on multimodal person recognition (phase 1 took place in January 2013 and phase 2 in January 2014). The main objective of this challenge is to answer the two following questions at any instant of the video: *"who is speaking?" "who is seen?"*. All modalities available (audio, image, external data ...) can be used for answering these questions. In this paper, we try to answer both questions simultaneously.

### 5.1 *REPERE* corpus [16]

The dataset used in our experiments is composed of videos recorded from eight different shows (including news and talk shows) broadcasted from two French TV channels in 720×576, MPEG-2. Thanks to the quality of the videos, written names are extracted automatically with relatively few errors. This corpus was made up in three phases, at each phase new data was added. In table 3 we detail the size of the different sets. We grouped the official `train` and `dev` sets of the second phase to cut them into two parts: a `TRN.m` used to build biometric models, a `TRN.f` used to tune $\alpha$, $\beta$ and the stopping criterion of the clustering process.

| Part | `train` ∪ `dev` | | `test1` | `test2` |
|---|---|---|---|---|
| | `TRN.m` (75%) | `TRN.f` (25%) | | |
| Raw | 58 h. | 28 h. | 14 h. | 37 h. |
| Annotated duration | 28 h. | 9 h. | 3 h. | 10 h. |
| size of $F$ | 23442 | 7800 | 2458 | 8272 |
| # faces | 13085 (887) | 4374 (379) | 1277 (145) | 4690 (433) |
| # speakers | 10631 (639) | 3517 (272) | 1101 (120) | 3844 (309) |

Table 3: Corpus size for the annotated part, number of faces and speakers annotated, in parenthesis is the number of different person appearing and speaking

Though raw videos were provided to the participants (including the whole show, adverts and part of surrounding shows), only excerpts of the target shows were manually annotated. On these segments, the speaker identities were completely annotated for the audio signal. For the image modality, only the faces that have a higher surface than 2000 pixels$^2$ are considered. Due to the cost of image annotation, only one image per shot (and every ten seconds maximum was manually annotated). Therefore to compare the results obtained for the two tasks we used the REPERE protocol where the presence of speakers and faces are evaluated on a set of annotated frames $F$. In the second part of the table 3 we detail the utterance number of faces and speakers identified on annotated frame (unknowns are not evaluated). In parenthesis we detail the number of different persons appearing and speaking for each data set.

In table 4 we detail the average duration and the number of named speakers/faces in the annotated segments. There are important variations in this corpus (e.g. 13.3 faces/video annotated in ⓗ for 592 seconds of annotation whereas there are only 3 faces annotated for ⓖ for a longer duration). These differences in person density will probably lead to a variable quality of identification. A surprise show (ⓒ) which is not present in the train set was included in `text 2`. Depending on the show, there are several faces that do not play a role in the videos (e.g. persons appearing in the background). Therefore, these faces are be very hard to identify.

| Show | Type | dur in s. | # video | dur s. average per video | # spk average per video | # face average per video |
|---|---|---|---|---|---|---|
| ⓐ BFMStory | news | 14395 (40%) | 8 | 1799 | 16.5 | 21 |
| ⓑ CultureEtVous | people news | 2898 (8%) | 25 | 115 | 4.7 | 7.6 |
| ⓒ RuthElkrief | news | 1267 (3.5%) | 4 | 317 | 5.3 | 7.5 |
| ⓓ CaVousRegarde | debate | 911 (2.5%) | 3 | 304 | 5 | 7.7 |
| ⓔ EntreLesLignes | debate | 3558 (10%) | 5 | 712 | 4.8 | 7.6 |
| ⓕ LCPInfo | news | 6337 (18%) | 8 | 792 | 10.9 | 14 |
| ⓖ PileEtFace | debate | 4626 (13%) | 6 | 771 | 3 | 3 |
| ⓗ TopQuestions | Q to Assembly | 1777 (5%) | 3 | 592 | 6.3 | 13.3 |
| all | | 35769 | 62 | 577 | 7 | 10 |

Table 4: Details per shows for set `test 2`

The response time of the clustering process with a single core at 2.00 Ghz is 128 minutes for 37 hours of the `test2`. The efficiency however depends upon the number of speaker turns / face tracks to proceed for a video (every videos are treated separately).

## 5.2 Evaluation metrics

### *5.2.1 Evaluate person identification*

As already stated, although the whole test set is processed, the person identification performance is measured only on the annotated frames $F$. We use the official REPERE metric defined as the proportion (in number of person utterances to be detected in annotated frames) of the reference $r$ incorrectly identified by the hypothesis $h$:

$$EGER(r,h) = \frac{\sum_{f \in F} \#\text{fa}(f) + \#\text{miss}(f) + \#\text{conf}(f)}{\sum_{f \in F} r(f)}$$

Where we count the number of errors: confusions (#conf), miss (#miss) and false alarm (#fa) in the hypothesis $h$ for a given annotated frame $f \in F$.

It is worth mentioning that, in the REPERE protocol, only named persons are taken into account for evaluation, which leads to systems that are trying

to name all the face tracks and speaker turns (because confusion and miss have the same weight in the evaluation metric). That's why we also report the precision $\%P$, recall $\%R$ and F-measures $\%F$.

$$\%P(r,h) = \frac{\sum_{f \in F} C_r^h(f)}{\sum_{f \in F} h(f)} \quad \%R(r,h) = \frac{\sum_{f \in F} C_r^h(f)}{\sum_{f \in F} r(f)} \quad \%F(r,h) = \frac{2 * P * R}{P + R}$$

where $C_r^h(f)$ is the number of utterances correctly identified, $h(f)$ the number of utterance in the hypothesis and $r(f)$ the number of utterances in the reference for a given annotated frame $f \in F$.

#### *5.2.2 Evaluate clustering*

To evaluate the speaker clustering quality, we use the classical diarization error rate (DER) defined by:

$$\text{DER} = \frac{d_{fa} + d_{miss} + d_{conf}}{d_{total}}$$

where $d_{total}$ is the total speech time and $d_{fa}$, $d_{miss}$, $d_{conf}$ are the duration errors of false alarm, miss and confusion. As identities of speakers are not considered for diarization, hypothesis and reference are aligned 1-to-1 to minimize $d_{conf}$.

As the annotation for head is not complete (1 annotation every 10 seconds) we cannot directly use the DER that take into account the duration. We, therefore, propose the EDER for Estimated Diarization Error Rate as a particular case of the EGER that not consider the name of clusters in the hypothesis:

$$\text{EDER}(r,h) = \text{EGER}(r, m(h)))$$

where $m(h)$ is the optimal mapping function 1-to-1 between clusters from the hypothesis and persons from the reference.

### 5.3 Upper bounds of performance

In table 5 we summarize some information about speakers. The second part of the table shows (between parentheses) the maximum number of speaker turns identifiable, i.e. those that correspond to either an existing biometric model or to an extracted written name. Our multi-modal person identification system should try to be the closest to the recall shown on the two last lines, while keeping the best precision.

Reducing the list of speaker models (Models_spk_84) leads to the reduction of speaker turns identifiable (test1:-33.9% ; test2:-28.6%) but when adding the written names (WN) information, this reduction is very low (test1:-3.4% ;

| | test1 | test2 |
|---|---|---|
| # speaker / corpus | 120 | 309 |
| total duration speaker turns | 9742 s. | 33172 s. |
| # speaker turns / corpus | 1224 (8 s.) | 2949 (11.2 s.) |
| Way of naming | max %recall for speaker turns | |
| Models_spk_706 | 73.4 (898) | 73.2 (2158) |
| Models_spk_84 | 39.5 (484) | 44.6 (1315) |
| WN of the video | 62.7 (768) | 69.2 (2042) |
| Models_spk_706 ∪ WN | 93.8 (1148) | 90.8 (2677) |
| Models_spk_84 ∪ WN | 90.4 (1107) | 88.4 (2606) |

Table 5: Statistics for speaker identification (without unknown) ; maximum recall expected

test2:-2.4%) which shows the complementary of these two sources of information (in the REPERE corpus, anchors and journalists are rarely presented by their name written on the screen).

| | test1 | test2 |
|---|---|---|
| # person appearing / corpus | 145 | 433 |
| # annotated frames | 2458 | 8272 |
| # face / corpus | 1277 | 4690 |
| Way of naming | max %recall for faces | |
| Models_face_908 | 74.2 (947) | 69.7 (3267) |
| Models_face_52 | 31.9(408) | 35.1 (1646) |
| WN of the video | 68 (868) | 79.5 (3727) |
| Models_face_908 ∪ WN | 91.9 (1174) | 91.4 (4286) |
| Models_face_52 ∪ WN | 84.5 (1079) | 85.8 (4025) |

Table 6: Statistics for face identification (without unknown) ; maximum recall expected

The table 6 shows the same information for faces. It may be noted that we have to identify more faces on `test2` (4690) that speaker turns (2949), while this number is close on `test1` (1277 *vs* 1224). This difference may come from the different distribution of shows between the two test sets. The maximum recall that we can expect behaves the same way than for speakers because we observe a good complementarity between written names and biometric models.

## 6 Experimental Results

### 6.1 Clustering quality

Before assessing the identification, we evaluate the clustering quality (see table 7). Overall, identifying clusters using written names information increases the quality of clustering. It may be noted that the error rate for faces between `test1` and `test2` increases about 12%, while for the speakers there are no significant differences.

It is important to observe that adding written names knowledge during the diarization process reduces systematically the diarization error rates.

| Set | WN | Speaker %DER | Head %EDER |
|---|---|---|---|
| test1 | | 19.72 | 23.7 |
| | X | 18.89 | 19.7 |
| test2 | | 21.96 | 35.2 |
| | X | 18.51 | 33.4 |

Table 7: Quality of the clustering with and without the identification by written names (WN)

### 6.2 Supervised biometric models

In table 8 we evaluate the speaker and face identification based only on biometric models. In the first part of the table, the speaker identification with the 706 models (Models_spk_706) obtains about 38% EGER. If we reduce the number of models to the classical casting of the shows (Models_face_52) the EGER is higher but the precision is improved about 10%. This better precision will allow us to better identify the speaker turns of anchors/journalists.

| system | Set | %P | %R | %F | %EGER |
|---|---|---|---|---|---|
| Models_spk_706 | test1 | 83.4 | 60.3 | 70.0 | 36.9 |
| | test2 | 81.7 | 58.0 | 67.8 | 38.9 |
| Models_spk_84 | test1 | 94.2 | 33.0 | 48.8 | 59.9 |
| | test2 | 90.4 | 40.2 | 55.6 | 54.8 |
| Models_face_908 | test1 | 56.8 | 33.1 | 41.8 | 65.5 |
| | test2 | 52.6 | 21.5 | 30.5 | 71.6 |
| Models_face_52 | test1 | 94.5 | 19.0 | 31.7 | 68.9 |
| | test2 | 91.9 | 16.9 | 28.6 | 70.8 |

Table 8: Speaker and face identification based only on biometric models

The second part of the table shows the results of face identification based only on the biometric models. The EGER is more important than for speakers due to the difficulty to have a good descriptor for a face. Indeed, actually, only front faces with good lighting can be recognized correctly. This is also true for the models construction, only those faces that appeared several times front face allow us to filter out bad images. We note also that if we use only 52 models of anchors/journalists, the EGER does not increase. Recall decreases while precision greatly increases. Again this good precision will allow us to better identify the faces of anchors/journalists.

### 6.3 Multi-modal person identification

Identification results of our multi-modal system are presented in table 9. The first information that we can see is that the propagation of written names (first and fourth line of each subtable) performs better than a system based only on biometric models (table 8).

| Speaker identification | | | | | | | |
|---|---|---|---|---|---|---|---|
| Set | Models spk_84 | Models face_52 | WN | %P | %R | %F | %EGER |
| test1 | | | X | 80.5 | 69.7 | 74.7 | 29.0 |
| | | X | X | 85.8 | 76.2 | 80.7 | 22.7 |
| | X | | X | 89.4 | 83.5 | 86.3 | 16.2 |
| | X | X | X | 89.2 | 83.7 | 86.4 | 16.0 |
| test2 | | | X | 76.2 | 61.7 | 68.2 | 35.9 |
| | | X | X | 79.5 | 66.9 | 72.6 | 31.4 |
| | X | | X | 86.0 | 79.1 | 82.4 | 20.1 |
| | X | X | X | 85.7 | 78.9 | 82.1 | 20.5 |
| Face identification | | | | | | | |
| Set | Models spk_84 | Models face_52 | WN | %P | %R | %F | %EGER |
| test1 | | | X | 86.1 | 61.5 | 71.7 | 32.9 |
| | | X | X | 88.6 | 65.8 | 75.5 | 28.9 |
| | X | | X | 86.0 | 65.4 | 74.3 | 30.3 |
| | X | X | X | 86.5 | 65.2 | 74.4 | 29.9 |
| test2 | | | X | 77.4 | 49.3 | 60.2 | 44.3 |
| | | X | X | 82.4 | 54.5 | 65.6 | 39.3 |
| | X | | X | 79.7 | 53.7 | 64.2 | 40.8 |
| | X | X | X | 80.2 | 53.6 | 64.2 | 40.5 |

Table 9: Multi-modal speaker and face identification

When we add the direct identification of some speaker turns / face tracks by biometric models, the EGER decreases. This diminution is more important for speakers than for faces (around -15% for speaker, -5% for face). What is interesting is that this reduction occurs even if the biometric models used come from another modality (use face models for speaker identification and vice versa, e.g. on `test2` for speaker identification WN: 35.9% of EGER, WN+Models_face_52: 31.4% of EGER). This reduction comes from the augmentation of the recall (e.g. on `test2` for speaker identification WN: 61.7% of recall, WN+Models_face_52: 66.9% of recall). Finally, the joint use of the two types of biometric models (face+speaker) does not improve results relative to the use of only one (best) biometric model.

### 6.4 Deeper analysis of the results and effect of crossmodality

In this section we show how the links between faces and speakers improve results. The data from table 10 must be compared to those from table 9. Usually, $\beta \simeq 0.4$ whatever the target task, while, $\alpha \simeq 0.4$ for speaker identification and $\alpha \simeq 0.5$ for face identification (see figure 2). In the table below, we set $\beta = 0.0$ (i.e. the $FullP_{st}$ matrix is set to zero) to cancel the cross-modal effect (no links between faces and speakers).

The links between faces and speakers not much change the results for speaker identification, with only a slight increase of the recall (+1.2% in average). For faces, the cross-modal effect is bigger on `test1` (+8.3% of recall in average) than on `test2` (+4.7% of recall in average). This increase of the

number of faces correctly named comes from the propagation of the identity of speaker names to faces that have a bad face descriptor (e.g. profile faces).

| Speaker identification | | | | | | |
|---|---|---|---|---|---|---|
| Set | Models_spk_84 | WN | %P | %R | %F | %EGER |
| test1 | | X | 81.6 | 68.3 | 74.3 | 29.9 |
| | X | X | 89.7 | 82.8 | 86.1 | 16.4 |
| test2 | | X | 74.8 | 60.1 | 66.7 | 37.3 |
| | X | X | 87.1 | 78.1 | 82.4 | 20.9 |
| Face identification | | | | | | |
| Set | Models_face_52 | WN | %P | %R | %F | %EGER |
| test1 | | X | 78.9 | 52.2 | 62.8 | 41.5 |
| | X | X | 83.9 | 58.6 | 69.0 | 35.1 |
| test2 | | X | 80.0 | 45.0 | 57.6 | 46.9 |
| | X | X | 81.6 | 49.5 | 61.6 | 42.8 |

Table 10: Speaker and face identification with $\beta = 0.0$ (no crossmodal links used))

The cross-modality can also be useful if decisions taken by biometric models are replaced by human annotations (e.g. in a active learning process). In this context we can ask a human to annotate faces (which is an easy and quick task than annotate audio) to identify speakers.

### 6.5 Per show analysis

In the table 11, we detail results per show, the second part takes into account only biometric models corresponding to target task (matching modality). There are important variations of performance between shows.

In debate shows ⓓ, ⓔ, ⓖ), guests are those most present in the video. Their names are written several times during the shows, this makes easier the identification than for news shows (ⓐ, ⓒ, ⓕ) where anchors speak/appear in a large part of the show. Biometric models correct this difference with an important reduction of EGER for news shows.

People news show (ⓑ) has specific characteristics (many people appearing, dubbing of foreign voices, shouts and laughs) which makes difficult the person identification. Biometric models add important information that leads to a better identification and more particularly for speakers.

For the show ⓗ (questions to the french Assembly), we note that it is easier to recognize speakers than faces. In this show, the questioner and the person who answers are introduced by their written names. This explains why speakers are easy to identify. But, around them, other persons appear with big enough face size to be annotated in the reference; therefore these faces have to be named too while it is difficult to do. For this show, adding biometric models does not improve performance since there is no anchors nor journalists which correspond to our pre-trained biometric models.

| Speaker identification | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Show | test 1 | | | | test 2 | | | |
| | %P | %R | %F | %EGER | %P | %R | %F | %EGER |
| | Without biometric models | | | | | | | |
| ⓐ BFMStory | 77.8 | 72.9 | 75.3 | 26.9 | 71.2 | 63.7 | 67.2 | 34.6 |
| ⓑ CultureEtVous | 56.4 | 26.8 | 36.4 | 70.7 | 23.3 | 10.4 | 14.4 | 85.2 |
| ⓒ RuthElkrief | - | - | - | - | 64.2 | 52.7 | 57.9 | 45.3 |
| ⓓ CaVousRegarde | 55.6 | 44.6 | 49.5 | 49.5 | 74.1 | 59.4 | 65.9 | 39.6 |
| ⓔ EntreLesLignes | 85.4 | 79.6 | 82.4 | 18.4 | 90.9 | 80.0 | 85.1 | 17.3 |
| ⓕ LCPInfo | 75.8 | 54.6 | 63.5 | 43.2 | 78.4 | 59.9 | 67.9 | 37.5 |
| ⓖ PileEtFace | 82.9 | 76.7 | 79.7 | 23.3 | 91.7 | 82.0 | 86.6 | 15.2 |
| ⓗ TopQuestions | 96.5 | 92.8 | 94.6 | 7.7 | 93.3 | 85.2 | 89.1 | 16.5 |
| all | 80.5 | 69.7 | 74.7 | 29.0 | 76.2 | 61.7 | 68.2 | 35.9 |
| | With biometric models (matching modality) | | | | | | | |
| ⓐ BFMStory | 88.9 | 86.7 | 87.8 | 13.6 | 86.4 | 82.1 | 84.2 | 16.8 |
| ⓑ CultureEtVous | 71.4 | 67.1 | 69.2 | 36.6 | 74.5 | 65.0 | 69.4 | 40.6 |
| ⓒ RuthElkrief | - | - | - | - | 64.2 | 52.7 | 57.9 | 45.3 |
| ⓓ CaVousRegarde | 74.1 | 59.4 | 65.9 | 34.7 | 83.7 | 76.2 | 79.8 | 23.8 |
| ⓔ EntreLesLignes | 93.5 | 87.8 | 90.5 | 10.9 | 89.0 | 80.9 | 84.8 | 17.0 |
| ⓕ LCPInfo | 94.5 | 84.2 | 89.0 | 15.3 | 88.5 | 81.2 | 84.7 | 16.9 |
| ⓖ PileEtFace | 87.4 | 80.8 | 84.0 | 19.2 | 91.4 | 85.5 | 88.3 | 12.1 |
| ⓗ TopQuestions | 96.5 | 92.8 | 94.6 | 7.7 | 93.3 | 85.2 | 89.1 | 16.5 |
| all | 89.4 | 83.5 | 86.3 | 16.2 | 86.0 | 79.1 | 82.4 | 20.1 |
| Face identification | | | | | | | | |
| Show | test 1 | | | | test 2 | | | |
| | %P | %R | %F | %EGER | %P | %R | %F | %EGER |
| | Without biometric models | | | | | | | |
| ⓐ BFMStory | 81.5 | 70.1 | 75.4 | 30.3 | 63.6 | 54.1 | 58.4 | 50.1 |
| ⓑ CultureEtVous | 65.0 | 12.3 | 20.6 | 86.8 | 47.7 | 13.0 | 20.4 | 81.7 |
| ⓒ RuthElkrief | - | - | - | - | 67.2 | 40.6 | 50.6 | 50.7 |
| ⓓ CaVousRegarde | 88.4 | 67.3 | 76.4 | 29.2 | 82.4 | 54.0 | 65.2 | 41.6 |
| ⓔ EntreLesLignes | 99.2 | 55.8 | 71.4 | 25.3 | 98.7 | 48.6 | 65.1 | 33.3 |
| ⓕ LCPInfo | 87.1 | 57.7 | 69.4 | 37.1 | 84.5 | 52.3 | 64.6 | 42.2 |
| ⓖ PileEtFace | 93.6 | 65.6 | 77.2 | 15.3 | 95.3 | 59.2 | 73.0 | 25.0 |
| ⓗ TopQuestions | 82.3 | 70.8 | 76.1 | 31.3 | 72.5 | 42.3 | 53.4 | 60.9 |
| all | 86.1 | 61.5 | 71.7 | 32.9 | 77.4 | 49.3 | 60.2 | 44.3 |
| | With biometric models (matching modality) | | | | | | | |
| ⓐ BFMStory | 85.2 | 73.9 | 79.2 | 27.0 | 74.0 | 64.6 | 69.0 | 40.1 |
| ⓑ CultureEtVous | 74.5 | 35.8 | 48.4 | 66.0 | 59.3 | 21.0 | 31.0 | 74.3 |
| ⓒ RuthElkrief | - | - | - | - | 66.7 | 39.6 | 49.7 | 50.7 |
| ⓓ CaVousRegarde | 88.2 | 66.4 | 75.8 | 30.1 | 84.9 | 64.6 | 73.4 | 31.0 |
| ⓔ EntreLesLignes | 99.2 | 55.4 | 71.1 | 25.3 | 98.5 | 49.0 | 65.5 | 33.1 |
| ⓕ LCPInfo | 96.7 | 66.3 | 78.6 | 28.6 | 85.1 | 54.5 | 66.4 | 39.9 |
| ⓖ PileEtFace | 99.1 | 71.3 | 83.0 | 9.6 | 98.6 | 63.7 | 77.4 | 19.8 |
| ⓗ TopQuestions | 81.0 | 70.4 | 75.3 | 31.7 | 71.6 | 40.4 | 51.6 | 62.8 |
| all | 88.6 | 65.8 | 75.5 | 28.9 | 82.4 | 54.5 | 65.6 | 39.3 |

Table 11: Speaker and face identification, score per show

## 7 Conclusion

In this paper, we presented strategies for unsupervised person identification in TV broadcast. This approach uses written names on screen and some biometric models as source of names into an agglomerative clustering process. Our process tries to merge speaker turns and face tracks together to form multi-modal clusters. The use of written names and speaker/face models allows to name these clusters but also to constrain the clustering process with rules that prevent merging two clusters with different associated names. This additional knowledge leads to an improvement of the diarization.

On a blind test set (`test2`), our system obtained an error rate (EGER) of 20.1% for speaker identification and 39.3% for face identification, which is much better than supervised mono-modal approaches, showing the relevance of our approach. Adding biometric models is useful to identify some particular persons (anchors, journalists).

Future works will focus on other types of clustering and on the integration of a more state-of-the-art diarization module (*I-vector*+ILP). Adding other types of visual information (clothes and backgrounds for person without a good face descriptor, e.g. profile face) is also a promising perspective. In the context of the camomile project and as the organizer of the new task "Multimodal Person Discovery in Broadcast TV" at Mediaeval 2015, we will try this method on an other language (Catalan) and provide the source code as a baseline for participants. Finally, we are currently trying to integrate our method into a semi-supervised scenario where manual annotations are available from a collaborative annotation platform.

**Acknowledgements** This work was partly realized as part of the Quaero Program and the QCompere project, respectively funded by OSEO (French State agency for innovation) and ANR (French national research agency

## References

1. C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multi-Stage Speaker Diarization of Broadcast News," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, p. 1505-1512, 2006.
2. M. Buml, K. Bernardin, M. Fischer, H. K. Ekenel, and R. Stiefelhagen, "Multi-pose face recognition for person retrieval in camera networks," in *7th International Conference on Advanced Video and Signal-Based Surveillance, AVSS*, p. 441-447, 2010.
3. F. Béchet, M. Bendris, D. Charlet, G. Damnati, B. Favre, M. Rouvier, R. Auguste, B. Bigot, R. Dufour, C. Fredouille, G. Linares, J. Martinet, G. Senay, and P. Tirilly, "Multimodal understanding for person recognition in video broadcasts," *15th Annual Conference of the Internationnal Speech Communication Association (INTERSPEECH)*, 2014.
4. M. Bendris, B. Favre, D. Charlet, G. Damnati, R. Auguste, J. Martinet, and G. Senay, "Unsupervised Face Identification in TV Content using Audio-Visual Sources," in *Proceedings of the 11th International Workshop on Content-Based Multimedia Indexing (CBMI)*, p. 243-249, 2013.
5. H. Bredin and J. Poignant, "Integer Linear Programming for Speaker Diarization and Cross-Modal Identification in TV Broadcast,", in *the 14th Annual Conference of the International Speech Communication Association, (INTERSPEECH)*, 2013.
6. H. Bredin, J. Poignant, M. Tapaswi, G. Fortier, V. B. Le, T. Napoleon, H. Gao, C. Barras, S. Rosset, L. Besacier, J. Verbeek, G. Quénot, F. Jurie, and H. Kemal Ekenel, "Fusion of speech, faces and text for person identification in TV broadcast," in *Workshop on Information Fusion in Computer Vision for Concept Recognition, ECCV-IFCVCR*, p. 385-394, 2012.
7. H. Bredin, J. Poignant, G. Fortier, M. Tapaswi, V. B. Le, A. Sarkar, C. Barras, S. Rosset, A. Roy, Q. Yang, H. Gao, A. Mignon, J. Verbeek, L. Besacier, G. Quénot, H. Kemal Ekenel, and R. Stiefelhagen, "QCompere at REPERE 2013," in *First Workshop on Speech, Language and Audio in Multimedia - the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH-SLAM*, 2013.
8. H. Bredin, A. Roy, V.B. Le, and C.Barras, "Person instance graphs for mono-, cross- and multi-modal person recognition in multimedia data: application to speaker identication in TV broadcast," in *International Journal of Multimedia Information Retrieval*, 2014

9. L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "Speaker diarization from speech transcripts," in *the 5th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2004.
10. L. Canseco, L. Lamel, and J.-L. Gauvain, "A Comparative Study Using Manual and Automatic Transcriptions for Diarization," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, p. 415-419, 2005.
11. D. Charlet, C. Fredouille, G. Damnati, and G. Senay, "Improving Speaker Identification in TV-shows using person name detection in overlaid text and speech,", in *14th Annual Conference of the Internationnal Speech Communication Association (INTERSPEECH)*, 2013.
12. S. S. Chen and P. S. Gopalakrishnan, "Speaker, Environment And Channel Change Detection And Clustering Via The Bayesian Information Criterion," in *DARPA Broadcast News Transcription and Understanding Workshop*, p. 127-132, 1998.
13. Y. Estève, S. Meignier, P. Deléglise, and J. Mauclair, "Extracting true speaker identities from transcriptions," in *the 8th Annual Conference of the International Speech Communication Association, INTERSPEECH*, p. 2601-2604, 2007.
14. B. Favre, G. Damnati, and F. Béchet, and M. Bendris, and D. Charlet, R. Auguste, and S. Ayache, B. Bigot, A. Delteil, R. Dufour, C. Fredouille, G. Linares, J. Martinet, G. Senay, and P. Tirilly, "PERCOLI: a person identification system for the 2013 REPERE challenge," in *First Workshop on Speech, Language and Audio in Multimedia - the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013.
15. P. Gay, G. Dupuy, C. Lailler, J.-M. Odobez, S. Meignier, and P. Deléglise, "Comparison of Two Methods for Unsupervised Person Identification in TV Shows," in *12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2014.
16. A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert and L. Quintard, "The REPERE corpus: a multimodal corpus for person recognition," in *the 8th International Conference on Language Resources and Evaluation, LREC*, 2012.
17. M. Guillaumin, J. Verbeek and C. Schmid, "Is that you? Metric learning approaches for face identification," in *the IEEE 12th International Conference on Computer Vision*, p. 498-505, 2009.
18. R. Houghton, "Named faces: putting names to faces," *IEEE Intelligent Systems*, vol. 14, p. 45-50, 1999.
19. V. Jousse, S. Petit-Renaud, S. Meignier, Y. Estève, and C. Jacquin, "Automatic named identification of speakers using diarization and ASR systems," in *the 34th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 4557-4560, 2009.
20. J. Kahn, O. Galibert, L. Quintard, M. Carré, A. Giraudel, and P. Joly, "A presentation of the REPERE challenge," in *the 10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, p. 1-6, 2012.
21. E. Khoury, C. Snac, and P. Joly. "Audiovisual Diarization Of People In Video Content," in *Multimedia Tools and Applications*, 2012.
22. V. B. Le, C. Barras, and M. Ferràs, "On the use of GSV-SVM for Speaker Diarization and Tracking," in *Odyssey - The Speaker and Language Recognition Workshop*, p. 146-150, 2010.
23. J. Mauclair, S. Meignier, and Y. Estève, "Speaker diarization: about whom the speaker is talking?" in *IEEE Odyssey 2006 - The Speaker and Language Recognition Workshop*, 2006.
24. S. Petit-Renaud, V. Jousse, S. Meignier, and Y. Estève, "Identification of speakers by name using belief functions," in *the 13th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods, IPMU*, p. 179-188, 2010.
25. P.T. Pham, M.-F. Moens, and T. Tuytelaars, "Naming persons in news video with label propagation," in *IEEE international conference on Multimedia and Expo, ICME*, p. 15281533, 2010.
26. P.T. Pham, T. Tuytelaars, and M.-F. Moens, "Naming people in news videos with label propagation," in *IEEE MultiMedia*, 18(3) p. 4455, 2011.

27. J. Poignant, L. Besacier, G. Quénot, and F. Thollard, "From text detection in videos to person identification," in *IEEE International Conference on Multimedia and Expo, ICME*, p. 854-859, 2012.
28. J. Poignant, H. Bredin, V. B. Le, L. Besacier, C. Barras, and G. Quénot, "Unsupervised speaker identification using overlaid texts in TV broadcast," in *the 13rd Annual Conference of the International Speech Communication Association, INTERSPEECH*, p. 2650-2653, 2012.
29. J. Poignant, L. Besacier, and G. Quénot, "Nommage non-supervisé des personnes dans les émissions de télévision: une revue du potentiel de chaque modalité," in *la 10ème COnférence en Recherche d'Information et Applications, CORIA*, 2013.
30. J. Poignant, L. Besacier, V. B. Le, S. Rosset, and G. Quénot, "Unsupervised naming of speakers in broadcast TV: using written names, pronounced names or both ?" in *the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013.
31. J. Poignant, H. Bredin, L. Besacier, G. Quénot, and C. Barras, "Towards a better integration of written names for unsupervised speakers identification in videos," in *First Workshop on Speech, Language and Audio in Multimedia - the 14th Annual Conference of the International Speech Communication Association, INTERSPEECH-SLAM*, 2013.
32. J. Poignant, L. Besacier, and G. Quénot, "Nommage non-supervisé des personnes dans les émissions de télévision: utilisation des noms écrits, des noms prononcés ou des deux?," in *Documents numriques*, p. 37-60, 2014.
33. M. Rouvier and S. Meignier, "A Global Optimization Framework For Speaker Diarization," in *Odyssey - The Speaker and Language Recognition Workshop*, 2012.
34. M. Rouvier, B. Favre, M. Bendris, D. Charlet, and G. Damnati, "Scene understanding for identifying persons in TV shows: beyond face authentication," in *12th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2014.
35. T. Sato, T. Kanade, T.K. Hughes, M.A. Smith, S. Satoh, "Video OCR: Indexing digital news libraries by recognition of superimposed caption," in *ACM Multimedia Systems*, 1999.
36. S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: naming and detecting faces in news videos," *IEEE Multimedia*, vol. 6, p. 22-35, 1999.
37. S. E. Tranter, "Who really spoke when? finding speaker turns and identities in broadcast news audio," in *the 31st IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, p. 1013-1016, 2006.
38. M. Uřičář, V. Franc, and V. Hlaváč, "Detector of Facial Landmarks Learned by the Structured Output SVM," in *the 7th International Conference on Computer Vision Theory and Applications*, p. 547-556, 2012.
39. J. Yang and A. G. Hauptmann, "Naming every individual in news video monologues," in *the 12nd ACM International Conference on Multimedia, ACMMM*, p. 10-16, 2004.
40. J. Yang, R. Yan, and A. G. Hauptmann, "Multiple instance learning for labeling faces in broadcasting news video," in *the 13th ACM international conference on Multimedia, ACMMM*, p. 31-40, 2005.