Comparison of two methods for unsupervised person identification in TV shows

Paul Gay, Grégor Dupuy, Carole Lailler, Jean-Marc Odobez, Sylvain Meignier, Paul Deléglise

Comparison of Two Methods for Unsupervised Person Identification in TV Shows

Paul Gay*†, Grégor Dupuy*, Carole Lailler*, Jean-Marc Odobez†, Sylvain Meignier*, and Paul Deléglise* *LUNAM
Université, LIUM, Le Mans, France
†IDIAP Research Institute, Martigny, Switzerland

***Abstract*—We address the task of identifying people appearing in TV shows. The target persons are all people whose identity is said or written, like the journalists and the well known people, as politicians, athletes, celebrities, etc. In our approach, overlaid names displayed on the images are used to identify the persons without any use of biometric models for the speakers and the faces. Two identification methods are evaluated as part of the REPERE French evaluation campaign. The first one relies on co-occurrence times between overlay person names and speaker/face clusters, and rule-based decisions which assign a name to each monomodal cluster. The second method uses a Conditionnal Random Field (CRF) which combine different types of co-occurrence statistics and pair-wised constraints to jointly identify speakers and faces.**

I. Introduction

The main purpose of this paper is to answer the questions "who is speaking?" and "who is seen?" at any time of the videos. The target persons are both journalists and guests, which can refer either to experts in a specific field, or to politicians, or celebrities. In other words, we try to identify by their names the people of a video using only the information of the video, without any biometric model related to the target persons.

Without any a priori information, person identities can be retrieved either from the speech transcripts or from the overlaid person names (OPN) commonly used to introduce the current speaker (e.g. see Fig 1). Several speaker identification approaches using speech transcripts have already been presented in [1][2][3]. Person identification from transcripts generally gives good results when transcripts are well segmented and close to what was said. However, results deteriorates as soon as automatic transcripts, produced by an automatic speech recognition (ASR) system, are used: the error rate in speaker identification increases from 16.66% to 75.15% [3]. On the other hand, OPNs can be reliably extracted using Optical Character Recognition (OCR) techniques, and their association with people in the videos is easier than analysing whether or not pronounced names in ASR transcripts refer to people appearing in the video.

Fig. 1. Example of annotation with an OPN introducing the speaker.

Head: Claude Géuant, unknown
OPN: Claude Géant
Speaker: Claude Géant

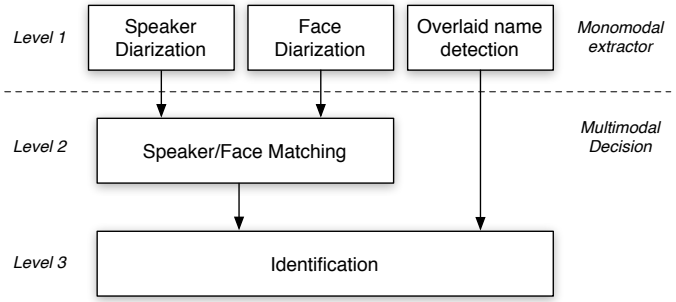


Fig. 2. Overview of the proposed system

The identification using OPN, which has started to be investigated 15 years ago [4], has since then raised a large amount of work, especially in face clustering tasks [5], face naming in captioned images [6], and recently, automatic naming within broadcast videos [7], [8], [9], [10], [11]. Regarding broadcast news videos, the use of OPNs raises two main challenges i) the sparsity of OPN information which requires to rely on diarization methods to propagate the identification to all person apparitions ii) the ambiguities that arise when several names and clusters co-occur together.

The work we present there has been carried out during the REPERE French evaluation campaign [12], which focused on multimedia people recognition in television documents. The proposed system, illustrated in Figure 2, consists in a 3-tier architecture.

The first level detects speakers, faces and overlaid person names. The second level associates speakers to faces using co-occurrence statistics between the speaker and the face clusters. In the last level, the overlaid person names are propagated to the speakers, or faces, in order to give, at any time, the identities of the persons in the show. Two identification methods, developped in parallel in different labs, are compared in this paper. The first one relies directly on the co-occurence times between names, faces and speakers and uses a set of rules to assign a name to each monomodal cluster. The second one uses a CRF which jointly performs the naming of

TABLE I. NUMBER OF KEY FRAMES, SPEAKERS, FACES AND OVERLAID NAMES IN CORPORA

Corpus	Key frames	Speakers	Heads	Overlaid names
dev	1229	1229	1449	197
test	1040	1039	1201	146
both	2269	2268	2650	343

all person clusters, thus allowing to account for pair-wised constraints and heterogeneous co-occurrence statistics. Two other consortiums participated to this campaign [9], [11]. Our approach differs from those previous works in, first, the choice of using face diarization directly as a pre-process to work both with speaker and face clusters, and second, the CRF-based identification method.

II. CORPUS

This study was conducted on two corpora from the REPERE French evaluation campaign [12]. The development corpus *(dev)* is composed of 28 TV shows. This corpus corresponds to the test corpus of the first evaluation (January 2013). The test corpus *(test)* is composed of 27 TV shows. It corresponds to the development set of the second evaluation (January 2014). Shows are recorded from the two digital French terrestrial television stations BFM and LCP (respectively MPG2 720×576 and MPEG2 544×576).

Thoses corpora are balanced between prepared speech, with 15 broadcast news, and more spontaneous speech, from 40 political discussions or street interviews. Only a part of the recordings are annotated, giving respectively a total duration of 3 hours for both corpus.

A keyframe of the corpora is annotated every 10 seconds, giving both identity of speakers and faces. The overlaid text corresponding to a target person is also annotated. The Table I reports some statistics about the corpora, and the Fig 1 shows an example of annotation.

III. LEVEL 1: SPEAKER, FACE AND OPN

The audio processing, known as the speaker diarization task (*cf.* sub-section III-A), is performed without any prior information regarding speakers: neither the number of speakers, nor their identities, nor voice samples are available. However, speaker diarization only tags audio segments with anonymous automatically-generated labels. The audio stream is split into small segments, which are clustered speakers. The video processing, called face diarization task (*cf.* sub-section III-B), consists in the same kind of process as the speaker diarization, except it is performed on face tracks in the video stream instead of speech segments in the audio stream. The OPN set (*cf.* sub-section III-C) is extracted using, first, an OCR on the video stream, then, a named entity detector which uses external ressources (predefined lists, freebase database, google hits,...).

A. Speaker Diarization

The speaker diarization system ("who speak when?") is based on the LIUM Speaker Diarization system [13], freely distributed[1]. This system has achieved the best or second best results in the speaker diarization task on French broadcast news evaluation campaigns, such as ESTER2, ETAPE (2011) [14] and REPERE (january 2012 and january 2013) [15].

The diarization system is composed of an acoustic Bayesian Information Criterion (BIC)-based segmentation followed by a BIC-based hierarchical clustering. Each cluster represents a speaker and is modeled with a full covariance Gaussian. A Viterbi decoding re-segments the signal using GMMs with 8 diagonal components learned by EM-ML, for each cluster. Segmentation, clustering and decoding are performed with 12 MFCC+E, computed with a 10ms frame rate. Music and jingle regions are removed using a Viterbi decoding with 8 GMMs (trained on french broadcast news data) for music, jingle, silence, and speech (with wide/narrow band variants for the last two, and clean/noised/musical background variants for wideband speech).

Gender and bandwidth are detected before transcribing the signal using the LIUM ASR [16]. A new non-speech segmentation is built from the filler words of the transcript. Non-speech segments longer than 0.5 second are removed in the speaker diarization.

In the previous steps, features were used unnormalized in order to preserve information on the background environment, which may help differentiating between speakers. At this point however, each cluster contains the voice of only one speaker, but several clusters can be related to a same speaker. The background environment contribution must be removed from each GMM cluster, through feature gaussianization, in order to obtain a one-to-one relationship between clusters and speakers.

Finally, the system is completed with clustering method based on the i-vectors paradigm and Integer Linear Programming (ILP). This new clustering method is fully described in [17]. The ILP clustering along with i-vectors speaker models gives better results than the usual hierarchical agglomerative clustering based on GMMs and cross-likelihood distances [18].

B. Face Diarization

The face diarization ("who appears when?") process consists of four main steps. First, shot boundary detection is performed to split the video stream into homogenous video clips. Second, frontal faces are detected within each shot using Viola And Jones algorithm [19]. False alarms are filtered out, based on temporal continuity and skin color detection. Then, face tracking is used to temporally extends the face detections within each shot into face tracks and increase the recall detection rate. Eventually, a face clustering step groups all face tracks that belong to the same person together. This last step is carried out using the method from [20]. It is based on a bottom-up algorithm, which computes the face similarities by combining *Speeded Up Robust Features* based distances, and statistical models built on block-based DCT features. This method reached state of the art results on the publicly available BUFFY dataset [21].

C. Overlaid name detection

The process is based on two steps: the first extracts hypotheses sentences from the video overlaid, the second searches for speaker names in the extracted sentences. The system used [22] to detect texts in the image sequences and

[1] http://www-lium.univ-lemans.fr/en/content/liumspkdiarization

transcribe it. The error rates of that system on the January 2012 dry-run corpus of the REPERE campaign is about 12% in terms of character, and 31% in terms of words.

The name research in overlaids is a task of named entity detection reduced to a single entity type: the “persons”. Compared to the same task in journalistic texts or audio transcripts, the announcement of the “persons” in the overlaids is well formatted, making the detection easier. Persons are mostly announced over two written lines, where the first refers to the person identity, and the second, the person function. The identity of a person is generally constituted of a firstname followed by a lastname. In some rare cases, the identity of a person only consists in a firstname, as for people interviewed in the street, or a stage name, as for artists.

Usually, named entity detectors use the linguistic context bordering the named entity to help its detection and categorization. The only context in the video overlaids could be the second line which presents the function. We chose to develop a rule-based identities detection system along with identity dictionaries. In that way, transcripts errors are taken into account much more easily than using a CRF-based method, as it was done in name entity detection. Three identity dictionaries were built:

1) *target*: the first dictionary includes 7345 identities. Most entries refer to journalists of the processed television stations, political personnalities, athletes and artists recurrently appearing in French media and REPERE training corpus.
2) *freebase*: The second dictionary is composed of more than 1.7 million of identities extracted from freebase [23]. We only extracted identities of people born after 1900, as well as those for which the birthdate is unknown.
3) *firstname*: The third and last dictionary consists in more than 17000 firstnames extracted from a specialized web site.

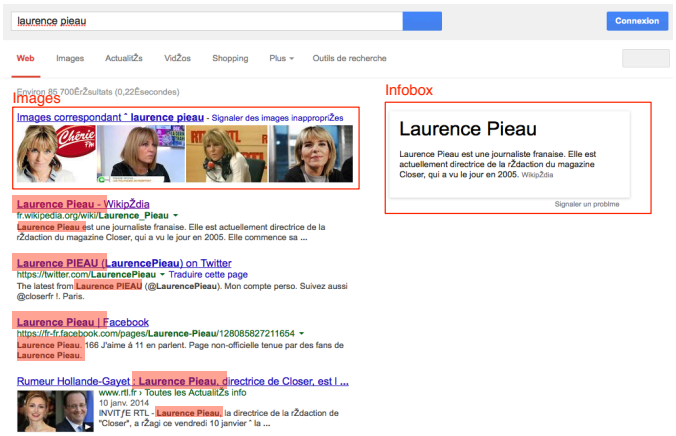
The detection process is only concerned by the first line of the video overlaids (c in the following):

1) Rejection step: c is rejected if the number of characters is less than 3 or more than 30, if c is composed of more than 10 words, or if c includes a keyword belonging to a pre-definite list. That list mostly includes conjugated action verbs. This step aims to reject lines c matching everything else than only an identity, like information messages which start by a person name.
2) c is accepted if it matches an entry from the *target* dictionary. The search into dictionary is performed using an error tolerance: the identity n of the dictionary is substituted to c if $s(c) > 0.8$ (eq. 1) such as:

$$s(c) = \min_{n \in N}(1 - \frac{d(c,n)}{max(l(c), l(n))}) \qquad (1)$$

where N is the *target* dictionary, $d(c,n)$ the Levenshtein distance between c and n, and $l(x)$ the string length of x.
3) c is accepted if it matches an entry of the *freebase* dictionary.

Fig. 3. Google roquets for ”Laurence Piau”



4) Finally, c is accepted if it starts with a firstname available in the *firstname* dictionary, and if the Google search engine gives clues to confirm the hypothetical identity. The analysis of the HTML page returned by the Google search engine, which answers to a request on c, is used to validate, or invalidate the hypothetical identity (Fig. 3). Analysis criteria on the Google HTML page relies on the frequency of c in the page, the presence of pictures associated to c, the presence of an informative box describing an individual, and the alternative proposed spelling.

IV. Level 2: Matching

The next step, using the speaker and face diarization, consists in matching the speaker $i \in \{1, \ldots, N\}$ with it's face $j \in \{1, \ldots, M\}$. In the REPERE training corpus, 78.4% of the speakers have their faces appearing on the video. We made the assumption that the more a speaker voice is heard in the same time a face is seen, the higher the chance that speaker and that face correspond to the same person. Therefore, co-occurrence time between speaker i and face j is computed for each pair (i,j). In the following, we refer to that co-occurrence value as $\delta t(i,j)$. It should be noted that this could be completed with lip activity detection as in [11]. The matching problem can be considered as the search of the largest set of pairs (i,j), for which the sum of the co-occurrences $\delta t(i,j)$ is maximal, and where each speaker i can only be associated to a single face j, et vice versa. Formally, this problem can be formulated as an Integer Linear Programming problem, given in 2, where $x_{i,j}$ is a binary variable equals to 1 when the speaker i is associated with the face j (*i.e.*, (i,j) has been selected).

$$\begin{aligned} \text{Maximize:} \quad & \textstyle\sum_{i=1}^{n}\sum_{j=1}^{m} \delta t(i,j) x_{i,j} \\ \text{Subject to:} \quad & \textstyle\sum_{i=1}^{n} x_{i,j} = 1, \quad (i = 1, \ldots, n) \\ & \textstyle\sum_{i=1}^{n} x_{i,j} = 1, \quad (j = 1, \ldots, m) \\ & x_{i,j} \in 0, 1 \end{aligned} \qquad (2)$$

The number of solution of this problem is $p!$, with p the higher number between speaker and face cluster numbers. However, an optimal solution to that matching problem can be found in a polynomial time using the Kuhn-Munkres algorithm [24]. In order to use the Kuhn-Munkres algorithm, we transform the maximisation problem into a minimisation problem, by changing the co-occurrence value in:

$$\delta t'(i,j) = \max_{i,j} \delta t(i,j) - \delta t(i,j) \tag{3}$$

After applying the Kuhn-Munkres algorithm, the speaker clusters which have been matched to face clusters are renamed with the same label, in their respective diarization segmentations. To account for segment boundary errors, speakers and faces are only matched if they overlap more than 0.1 second.

V. Level 3: Identification

At this level, audio and video diarization have been already performed. Several speaker clusters have been associated to faces clusters. In addition, the OPNs have been extracted. OPNs are characterized by their appearance and disappearance times, which make them constitute segments. We propose two identification methods. The first one, described in sub-section V-A, uses co-occurrence statistics between clusters and OPNs to choose the best identity for each clusters, without any consideration about the identities attributed to the other clusters. The second method (*cf.* sub-section V-B), inspired from [25], relies on Conditional Random Fields (CRFs) and performs a joint identification of all person clusters (speakers and faces), allowing to account for uniqueness constraints and co-occurrence statistics between clusters and OPNs (see [26] for more details).

A. Direct identification

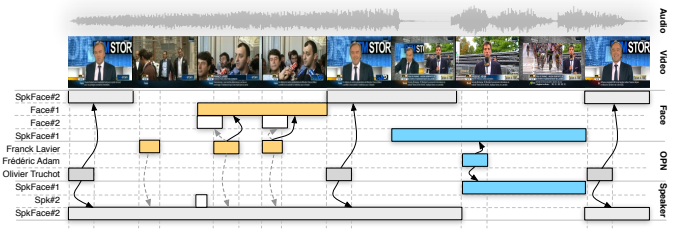
This identification method is first applied to speaker clusters. For speaker clusters that have been associated to a face cluster, the identity is propagated to both speaker and face clusters. Then, this identification method is applied to the face clusters for which no speaker cluster have been matched.

As for the speaker/face clusters matching step (*cf.* section IV), we used the co-occurrence times $\delta t(C_i, P_j^k)$ between the occurrences P_j^k with $k = 1 \dots K_j$ of a OPN P_j from the set of OPN $P = \{P_j, j = 1 \dots N^P\}$, and the segments of a speaker or face cluster C_i. Each $\delta t(C_i, P_j^k)$ is normalized by appearance duration d_k^j of the occurrence P_j^k. Each occurrence score of the name P_j are then summed. The cluster C_i is therefore identified by the name P_j giving the highest score among the OPN set P, which refers to all the OPN in co-occurrence with the cluster C_i (equation 4).

$$C_i = \underset{P_j \in P}{\operatorname{argmax}} \sum_{k=1}^{K_j} \frac{\delta t(C_i, P_j^k)}{d_k^j} \tag{4}$$

d_k^j, which is the appearance duration normalization of the occurrence P_j^k, favours a name P_j which recurrently appear during the various person C_i interventions, instead of favouring the total appearance durations of OPNs related to the person C_i.

Fig. 4. Speaker diarization, Face diarization et OPN



B. CRF identification

CRF models enable to efficiently combines heterogeneous statistics by and have been used in a speaker identification task using transcripts [27]. Another advantage is the possibility to introduce pairwise relationships between cluster pairs. Let us denote by $C^{av} = \{C_i^{av}, i = 1 \dots N^{C^{av}}\}$ the set of audiovisual clusters obtained after the speaker/face matching step described in Section IV. Here, each cluster C_i^{av} consists either in a speaker/face pair, or a face alone, or a speaker alone. Conversely to the previous method which takes a decision about each mono-modal cluster separately, that method considers audiovisual cluster pairs, thus allowing for more meaningful co-occurrence statistics.

The identification can be seen as an estimation of the label field $E = \{e_i, i = 1 \dots N^{C^{av}}\}$, such as the label e_i corresponds to the name of the cluster C_i^{av}. e_i takes value from the set of OPNs P, augmented by an *anonymous* label which should be assigned to anonymous persons. Let G be an undirected graph over the set of random variables C^{av}, P, and E. We then seek to maximize the CRF posterior probability $P(E|C^{av}, P)$ defined as:

$$P(E|C^{av}, P) = \frac{1}{Z(P, C^{av})} \times exp\left\{ \sum_{i=1}^{6} \sum_{Clq \in G_i} \lambda_i f_i(Clq) \right\} \tag{5}$$

where $Z(P, C^{av})$ is a normalization constant, and each triplet (f_i, G_i, λ_i) is composed of a feature function f_i, the set G_i of cliques where this function is defined, and its CRF weight λ_i learned at training time.

This naming model exploits four different co-occurrence statistics between clusters and OPNs. First, for each triplet (e_i, C_i^{av}, P_j), we define:

$$f_{audio}(e_i, C_i^{av}, P_j) = \sum_{k=1}^{K_j} \frac{\delta t(C_i^a, P_j^k)}{d_k^j} \tag{6}$$

Where C_i^a is the audio part of C_i^{av} (which can be null). We define similarly two other feature functions for the visual part of C_i^{av}: f_{v_alone} which counts the co-occurrences when the face is alone in the image, and f_{v_multi} which accounts for multi-face images. Thus, the CRF will learn a different λ weight for each case and will be able to weight each type of information. Eventually, we followed the assumption that a person does not appear or speak before the first apparition of his name in an OPN, and defined a function $f_{before}(e_i, C_i^{av}, P)$

which returns the number of audio segments from cluster C_i^{av} which occurs before the first apparition of the name e_i. We also introduce prior knowledge over the $anonymous$ label by defining a fifth feature function $f_{ano}(e_i, C_i^{av})$ which returns 1 if e_i is the $anonymous$ label. Lastly, we define a uniqueness function $f_{uniq}(e_i, C_i^{av}, e_j, C_j^{av})$ over visually overlapping clusters. For such pair C_i^{av}, C_j^{av}:

$$f_{uniq}(C_i^{av}, C_j^{av}, e_i, e_j) = \begin{cases} -Inf & \text{if } e_i = e_j \\ 0 & \text{otherwise} \end{cases}$$

VI. Experiments

A. Evaluation metrics

The evaluation metric chosen to measure the identification performance is the official REPERE Estimated Global Error Rate (EGER). This metric is defined as follow:

$$EGER = \frac{\#fa + \#miss + \#conf}{\#total} \quad (7)$$

where $\#total$ is the number of persons to be detected, $\#conf$ the number of persons wrongly identified, $\#miss$ the number of missed persons and $\#fa$ the number of false alarms. It should be noted that this metric does not take into account the spatial position of the faces but rather focus on evaluating the ability of the system to answer the question "Who is seen?" and "Who is speaking".

To evaluate the impact of the identification, we report the Diarization Error Rate (DER). The metric was introduced by the NIST as the fraction of speaking time which is not attributed to the correct speaker using the best match between references and hypothesis speaker labels. Both DER and EGER are computed using the scoring tool developed by the LNE[2] as part of the REPERE campaigns.

B. Overlaid person name detection results

The Table II shows the evaluation results of the overlaid person names, for the both corpora. The EGER is 15.5%. Most of the errors are due to missed overlaid names, which were too close to the overlaid segment boundaries.

The 8 confusion errors come from erroneous transcripts, thus generating a substitution of the transcript name with the most similar target name.

TABLE II. Evaluation of overlaid name detection

Corpus	EGER	#conf	#miss	#fa
DEV	14.2%	0	22	6
TEST	17.1%	8	16	1
both	15.5%	8	38	7

C. Association and Identification results

The speaker DER is 13.12% on the *dev* set and 15.00% on the *test* set. A third of DER errors come from missed detection. The reason is that BFM is a television station which continuously broadcast, frequently using jingles to capture the public attention. Those are detected as non-speech areas, even if the host is announcing headlines. Regarding the Face diarization, the DER is 77% person on the *dev* set and 77% on the *test* set. Overall 75% of the errors come from miss detections due to non profile faces and person seen from the back. Qualitatively, we found that the face diarization method provides pure face clusters, but might split one person into different clusters if the head pose variations are important.

TABLE III. Precision, Recall and F-mesure on dev and test corpora. Results are reported while evaluating the speaker only, the head only and both

Corpus	EGER	#conf	#miss	#fa
(1) DEV: Oracle				
Speaker	20.2%	25 (2.0%)	206 (16.8%)	17 (1.4%)
Head	31.5%	8 (0.5%)	446 (30.8%)	3 (0.2%)
Spk + Head	26.3%	33 (1.2%)	652 (24.3%)	20 (0.7%)
(2) TEST: Oracle				
Speaker	29.9%	23 (2.2%)	284 (27.3%)	4 (0.4%)
Head	38.6%	2 (0.2%)	461(38.4%)	0 (0%)
Spk + Head	34.6%	25 (1.1%)	745 (33.3%)	4 (0.2%)
(3) DEV: System "Direct" with OPN groundtruth				
Speaker	25.0%	114 (11.7%)	181(18.4%)	12 (0.8%)
Head	37.5%	49 (3.4%)	465 (32.1%)	29 (2%)
Spk + Head	31.7%	163 (6.1%)	646 (24.1%)	41 (1.5%)
(4) TEST System "Direct" with OPN groundtruth				
Speaker	37.8%	152 (14.6%)	236 (22.7%)	5 (0.5%)
Head	45.0%	51 (4.3%)	482 (40.1%)	7 (0.6%)
Spk + Head	41.7%	203 (9.1%)	718 (32.1%)	12 (0.5%)
(5) DEV: System "Direct"				
Speaker	25.0%	114 (11.7%)	181 (14.7%)	12 (1.0%)
Head	37.4%	50 (3.5%)	465 (32.1%)	27 (1.8%)
Spk + Head	31.7%	164 (6.1%)	646 (24.1%)	39 (1.5%)
(6) TEST: System "Direct"				
Speaker	39.2%	166 (16%)	236 (22.7%)	5 (0.5%)
Head	46.4%	68 (5.7%)	482 (40.1%)	7 (0.6%)
Spk + Head	43.0%	234 (10.5%)	718 (32.0%)	12 (0.5%)
(7) DEV: System "CRF"				
Speaker	24.9 %	102(8.3%)	192(15.6%)	12(1.0%)
Head	38.4%	55(3.4%)	480(33.1%)	21(1.4%)
Spk + Head	32.2%	157(5.9%)	672(25.1%)	33(1.2%)
(8) TEST: System "CRF"				
Speaker	37.2%	150(14.4%)	233(22.4%)	4(0.4 %)
Head	46.0%	54(4.5%)	492(41.0%)	7(0.6%)
Spk + Head	42.0%	204(9.1%)	725(32.4%)	11(0.5%)

The identification results for the different systems are presented in Table III. Table (1-2) is an oracle simulating the best possible results we would reach with perfect diarization and identification systems, while keeping automatic speech segmentation, face and OPN detections. For the speaker part, errors are mostly due to journalist off-voices, which are not announced by OPNs. Considering the faces, miss errors are mainly divided into non-frontal faces and figurative persons not announced by OPNs. The few confusions and false alarms are due to inaccuracies of segment boundaries. In addition to provide an upper bound for the identification results, those tables highlight the differences among the *dev* and *test* sets. The identification on the *test* set is considerably more difficult with an increase of the EGER from 26.3% to 34.6%.

The tables (3-4) and (5-6) present the identification results of the direct method. The system in tables (3-4) uses OPN ground truth while the other system uses the OPNs automatically extracted. While performances are stable for the *dev* test, the use of automatic OPN detection increase the EGER from 41.7% to 43%. This change could probably be explained by the confusions errors in the OPN detections only observed in the *test* set, in table II.

The increase in the miss rate, with respect to the oracle, is mainly due to diarization errors where a person is split into several clusters. Indeed, those smaller clusters are likely to not

[2]The French National Laboratory of Metrology and Testing

co-occur with any OPN, and thus, it is not possible to identify them. Confusion errors arise when a cluster co-occurs with wrong names.

Regarding comparisons with the Direct method and the CRF method (tables 5 to 8), two differences can be highlighted. First, the CRF tends to give a better performance on the speaker identification, with an absolute reduction of the EGER from 39.2% to 37.2% on the *test* set. Improvements mainly come from a reduction of the confusion errors, meaning that the considered co-occurrence statistics, and the joint audio-visual decision process used by the CRF, can help to solve ambiguities. Secondly, considering the head identification, the Direct method gives more confusion and less miss errors. Actually, this method try to identify all possible faces, and may label two co-occurring clusters with the same name. We found this strategy optimal for the EGER metric which does not take into account the spatial positions of the faces as explained previously. The CRF, by using the uniqueness constraint, provides a more accurate naming at the cost of more miss errors. Although performances are similar regarding EGER metric, the CRF results are more suitable for other tasks such as collecting faces of a given person for unsupervised model training.

VII. Conclusions

In this work, a method for identifying people using overlaid person names in TV shows has been described. Results shows that such OPNs can be reliably extracted and used to identify face and speaker clusters. Two approaches are explored to associate names and clusters. The first identification method relies on co-occurrence times between overlay person names and speaker/face clusters; the second method uses a CRF to jointly identify the speaker and the face clusters. By jointly exploiting the audiovisual information and additional context clues, the CRF provides the best results, mainly by solving ambiguous speaker cases where a same person co-occurs with different OPNs.

The difference of performances between the oracle and the automatic systems show that there is room for improvements in the diarization and identification steps. First, the speaker and face diarizations could be optimised for the identification tasks, for instance by taking into account the OPNs information [28]. Second, the CRF method could easily be extented to include biometrics models learned in an supervised or unsupervised way [29], [30]. Last but not least, saliency and namedness features could be used in multi-face images to improve name/face association.

Acknowledgment

This research was supported by ANR (French National Research Agency) under contract number ANR-2010-CORD-101-01 (SODA project), as well as by the European Union (project FP7-611057 EUMSSI).

References

[1] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "A comparative study using manual and automatic transcriptions for diarization," in *Proc. of ASRU*, 2005.

[2] S. Tranter, "Who really spoke when? Finding speaker turns and identities in broadcast news audio," in *Proc of ICASSP*, 2006.

[3] V. Jousse, S. Petit-Renaud, S. Meignier, Y. Estève, and C. Jacquin, "Automatic named identification of speakers using diarization and ASR systems," in *Proc. of ICASSP*, 2009.

[4] S. Satoh, Y. Nakamura, and T. Kanade, "Name-it: Naming and detecting faces in news videos," *Proc. of IEEE Multimedia*, 1999.

[5] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is... buffy automatic naming of characters in tv video," in *Proc. of BMVC*, 2006.

[6] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller, and . Forsyth, "Names and faces in the news," in *Proc. of CVPR*, 2004.

[7] M. Pham, P. and T. Tuytelaars, "Naming persons in news video with label propagation," in *Proc. of ICME*, 2010.

[8] B. Jou, H. Li, J. G. Ellis, D. Morozoff-Abegauz, and S.-F. Chang, "Structured exploration of who, what, when, and where in heterogeneous multimedia news sources," in *Proc.of ACM Multimedia*, 2013.

[9] 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?" 2013.

[10] J. Poignant, H. Bredin, V. Le, L. Besacier, C. Barras, G. Quénot *et al.*, "Unsupervised speaker identification using overlaid texts in tv broadcast," in *Proc. of Interspeech*, 2012.

[11] M. Bendris, B. Favre, D. Charlet, G. Damnati, G. Senay, R. Auguste, and J. Martinet, "Unsupervised face identification in tv content using audio-visual sources," in *Proc. of CBMI*, 2013.

[12] J. Kahn, O. Galibert, L. Quintard, M. Carré, A. Giraudel, and P. Joly, "A presentation of the repere challenge," in *Proc. of CBMI*, 2012.

[13] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization," in *Proc. of Interspeech*, 2013.

[14] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, O. Galibert *et al.*, "The etape corpus for the evaluation of speech-based tv content processing in the french language," in *Proc. of ICLREC*, 2012.

[15] O. Galibert and J. Kahn, "The first official REPERE evaluation," in *Proc. of SLAM*, Marseille, France, 2013.

[16] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "Improvements to the lium french asr system based on cmu sphinx: what helps to significantly reduce the word error rate?" in *Proc. of Interspeech*, 2009.

[17] M. Rouvier and S. Meignier, "A global optimization framework for speaker diarization," in *Proc. of Odyssey*, 2012.

[18] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Improving speaker diarization," in *Proc. of DARPA RT04*, 2004.

[19] P. Viola and M. Jones, "Robust real-time face detection," *Proc. of IJCV*, 2004.

[20] E. Khoury, P. Gay, and J. Odobez, "Fusing matching and biometric similarity measures for face diarization in video," in *Proc. of ICMR*, 2013.

[21] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *Proc of ICCV*, 2009.

[22] D. Chen and J. Odobez, "Video text recognition using sequential monte carlo and error voting methods," *Pattern Recogn. Lett.*, vol. 26, no. 9, pp. 1386–1403, 2005.

[23] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. of ACM SIGMOD*, 2008.

[24] J. Munkres, "Algorithms for the assignment and transportation problems," *Proc. of JSIAM*, 1957.

[25] P. Gay, E. Khoury, S. Meignier, J.-M. Odobez, and P. Deleglise, "A conditional random field approach for audio-visual people diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, 2014.

[26] ——, "Face identification from overlaid texts using local face recurrent patterns and crf models," in *IEEE International Conference on Image Processing (ICIP)*, 2014.

[27] C. Ma, P. Nguyen, and M. Mahajan, "Finding speaker identities with a conditional maximum entropy model," in *Proc. of ICASSP*, 2007.

[28] H. Bredin and J. Poignant, "Integer linear programming for speaker diarization and cross-modal identification in tv broadcast," 2013.

[29] C. Lailler, G. Dupuy, M. Rouvier, and S. Meignier, "Semi-supervised and unsupervised data extraction targeting speakers: From speaker roles to fame?" in *In Proc. of SLAM*, 2013.

[30] C. Liu, S. Jiang, and Q. Huang, "Naming faces in broadcast news video by image google," in *Proc. of ACM Multimedia*.