PERCOLI: a person identification system for the 2013 REPERE challenge

Benoit Favre[1], *Geraldine Damnati*[4], *Frederic Bechet*[1], *Meriem Bendris*[1],
Delphine Charlet[4], *Remi Auguste*[2], *Stephane Ayache*[1], *Benjamin Bigot*[3],
Alexandre Delteil[4], *Richard Dufour*[3], *Corinne Fredouille*[3], *Georges Linares*[3],
Jean Martinet[2], *Gregory Senay*[3], *Pierre Tirilly*[2]

[1]Aix Marseille Université, LIF-CNRS ; [2]Université de Lille, LIFL
[3]Université d'Avignon, LIA; [4]Orange Labs, France

Abstract

The goal of the PERCOL project is to participate to the REPERE multimodal challenge by building a consortium combining different scientific fields (audio, text and video) in order to perform person recognition in video documents. The two main scientific issues addressed by the challenge are firstly multimodal fusion algorithms for automatic person recognition in video broadcast ; and secondly the improvement of information extraction from speech and images thanks to a combine decoding using both modalities to reduce decoding ambiguities. This paper describes the system PERCOLI that participated to the REPERE 2013 challenge and presents the results obtained on the main person recognition tasks.

Index Terms : multimodal fusion, person identification, video processing.

1. Introduction

The *Repere* challenge consists in identifying persons in video shows using cues from spoken content (speaker identity and words), and video content (faces and overlaid text) [1]. Systems participating in the challenge must generate a list of segments with person names according to the presence of said persons in the visual and audio modalities, using both biometric models and context analysis. The challenge provides a set of videos manually annotated with speaker segmentation, speech transcription, overlaid text transcription and face outline. All image-related annotations are sampled every 10 seconds on so-called key-frames.

Most visual indexing methods are based on face detection and recognition. Those methods require large databases of facial models trained to recognize each person who could appear in a video. However, the variability of face appearance in TV content (pose, facial expressions, lighting, occlusions) makes identification using facial models very unreliable. In addition, maintaining up-to-date large dictionaries of face models is prohibitively expensive. In this paper, we are interested in methods for naming faces in TV content with no face models.

Such person identification methods are often performed in two steps : (1) names are extracted from a range of sources and (2) an association strategy assigns each detected name to a person. In the first step, the identities can be extracted from speech (using Automatic Speech Recognition [2, 3]), image (with Optical Character Recognition [4] on overlaid text) and text content (such as scripts and subtitles [5]). In the second step, the extracted identities are propagated via clustering methods [4, 6]. This step is the focus of our paper. Figure 1 illustrates that process on a video from the *REPERE* [1] corpus [7].

We propose to directly associate OCR and speech detected names with current faces and speakers, and then propagate that information within and cross modalities with face and speaker similarities and talking face detection. This paper is organized as follows : Section 2 describes related work ; Section 5 describes person name acquisition from *OCR* and *ASR* output ; Section 6 similarity measures for speaker and face clustering ; Section 7 presents our identity propagation method based on direct and indirect association. Finally, Section 8 presents the *REPERE* corpus, results of experiments and a discussion.

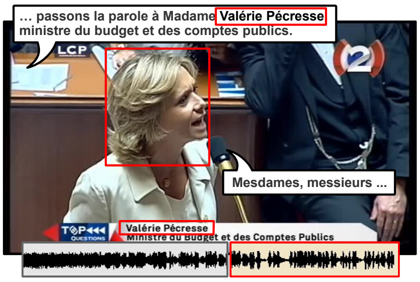
... passons la parole à Madame Valérie Pécresse
ministre du budget et des comptes publics.
LCP
Mesdames, messieurs ...
Valérie Pécresse
Ministre du Budget et des Comptes Publics

FIGURE 1 – The *REPERE* corpus. The identity appears in multiple sources.

2. Related work

Several studies have addressed the problem of association-propagation strategies for face identification. Name-it [8] proposed to find face-name associations by maximizing the co-occurrence between similar faces and names extracted from OCR output. [9] proposed to name faces in images using a graphical model for face clustering. Nodes represent detected faces and edges are weighted by SIFT-based similarity. Then, for each name detected in OCR, greedy search is applied to find the sub-graph that maximizes face similarities within the set of faces associated to the name. However, this approach cannot identify faces if no name is detected in the image. In [10], authors proposed to identify faces in *TRECVID* news videos using training data obtained automatically from Google image search. Names were extracted from both OCR and ASR output. In [5], authors proposed to align detected faces with names from the script and used rules based on lip activity and gender detection to resolve ambiguities. In [6], names are extracted from movie scripts and subtitles and associated to faces

[1] Reconnaissance de PERsonnes dans des Emissions Audiovisuelles : www.defi-repere.fr

according to lip activity. Identities are then propagated using face-level and clothes-level similarities. Although preliminary results are promising, face and clothes variability (pose, expression, color. . .) hamper the robustness of the similarity measure. In this case, audio information can be used in addition to visual cues to associate names to faces through speaker identity. In fact, in TV content speaker diarization appears to be more robust than face clustering [11]. [2, 3] proposed to extract names using *ASR* output and associated them to speakers using lexical rules on speaker clusters. In [4], names are extracted from OCR output and propagated to speaker clusters in order to maximize co-occurrence.

3. System architecture

The general architecture of the PERCOLI system is displayed in figure 2. The goal of this system is to predict, for each video frame, who is talking (*SPEAKER* hypotheses) and who appears on the screen (*FACE* hypotheses).

There are three main steps in the system :

1. **Person name hypotheses generation** : this step is in charge of producing all the person name identities that can be associated with a voice or a face in a given time window in a video .
2. **Multimodal speaker identification** : this process gives an identity, when possible, to each speaker segment produced during the speaker diarization process thanks to the person name identities given by the previous step.
3. **Multimodal face identification** : this second multimodal fusion process gives an identity, when possible, to each face segment produced by the face tracking process thanks to person name identities, face similarity, and speaker hypothesis provided by the previous step.

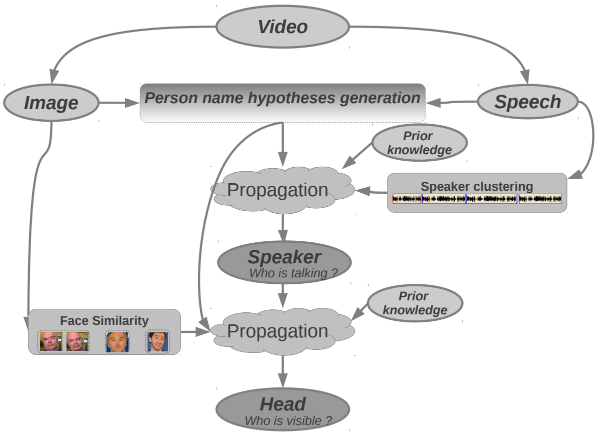

FIGURE 2 – General architecture of the Percoli system

The **Person name hypotheses generation** process is displayed in figure 3. There are three sources of person name identities : person names written in a text box, called *Overlay Person Name* hypotheses, and obtained through an Optical Character Recognition (OCR) process ; person names occurring in the speech channel, called *Uttered Person Name* hypotheses, and extracted from the Automatic Speech Recognition (ASR) of speech segments ; speaker recognition hypotheses obtained thanks to *a priori* speaker models corresponding to the main presenters, journalists and politicians likely to occur in the news. For the first two sources of identities, an *entity linking* process is needed in order to obtain a full identity from the occurrences of names detected in either text or speech.

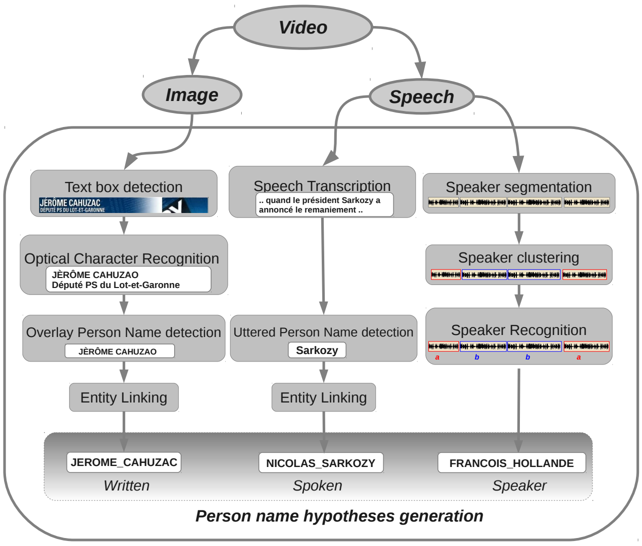

FIGURE 3 – Person name hypotheses generation according to three modalities : text in overlay text box ; speech ; speaker

4. Prior knowledge

Our system relies on three knowledge sources for performing identification : person name databases, speaker models and show-specific knowledge. From these sources, speaker models are the only biometric source of identity which is not allowed in unsupervised REPERE tasks.

4.1. Person name linking

Person name linking consists in linking name hypotheses manipulated in the system with a large database, in order to (1) discard unlikely names, and (2) account for meta-data such as the gender or spoken language of a person. In particular, we are interested at determining names which are likely to appear in the show being processed. This likelihood is estimated from entity linking and the aired date of the show. Entity linking is performed in two steps : clustering of person name variants and matching with database entries. The first step is achieve by collecting newswire articles around the dates of the broadcast shows, detecting names with a Named Entity Tagger [12] and clustering mentions. We applied this process to a large corpus of newswire spreading from 2004 to 2012. About 9.2M NEs were automatically detected. After the clustering and filtering process we obtained 117K clusters containing 162K mentions of person names. The second step consists in linking each cluster with a unique identifier. We used for this purpose the *ALEDA* entity database [13], which is a structured version of Wikipedia (about 225K person entities). From this database, heuristics are used to extract biographical elements such as gender, spoken language, topic, and whether the person is alive. When processing name hypotheses, they are linked to the mention which corresponds to the most frequent person cluster.

4.2. Speaker models

The speaker identification system is a standard GMM/UBM system (512 gaussians). We have collected audio for 345 speakers, mainly on journalists and politicians, from REPERE training data and various BN sources. Speakers with less than 30 seconds of speech are discarded. The generated models cover 30% of the training data speakers, 50% of the development data speakers and 54% of the test data speakers. Post-campaign evaluation has shown that the system has robustness issues because speakers with moderate quantity of training data are three times more likely to be incorrect on the test set than they are on the development set.

4.3. Show-specific constraints

The idea is to build models of who is likely to appear in recurring TV shows, and take advantage of show structure in order to capture names that would otherwise be difficult to detect. In particular, our system relies on two sources of information : lists of per-show presenters, journalists, columnists, commentators, and the setting of a show as its type (talk-show on a fixed stage, news with field reports), the number of invited speakers. In addition, for specific stage shows, an online component uses the order in which guests are presented to determine their location on stage and deduce probable coocurrence on a give camera angle.

5. Name detection

In addition to prior knowledge sources, person names are extracted from overlaid texts by using optical character recognition and from spoken content thanks to automatic speech recognition.

5.1. Optical Character Recognition

The Overlay Person Name (OPN) recognition process is made of 3 steps in our approach : text box detection ; Optical Character Recognition producing a confusion network of characters ; person name recognition in the character hypotheses.

Text box detection is achieved with a convolutional neural net approach described in [14], then OCR is performed with Tesseract [2], a standard open-source OCR system. Frame-to-frame tracked text boxes lead to different OCR hypotheses because of background changes and animations. The consecutive transcripts are merged in a Confusion Network (CN) in order to compute the most maximum posterior probability character sequence on the whole track. A few heuristics are used to locate actual person names in text boxes that contain other information (occupation, etc) and hypothesized names are filtered according to their Levenstein distance, computed efficiently with finite state transducers, to the large list of person names described in Section 4. If linking the name to the database fails, we back off to a web search and filter names returning less than 400 hits.

5.2. Automatic Speech Recognition

Automatic transcription of all speech content is not adequate for finding person names because word error rate can be relatively high (near 30% in our system) and names tend to be out-of-vocabulary and therefore never hypothesised by the system. Our name spotting component searches for names in phoneme confusion networks generated by a first pass of ASR. Given a list of potential person names likely to appear in the processed show, the system ranks them according to the average phonetic posterior after alignment of the phonetic transcription of the name (with a cutoff on the Levenstein distance). Name spotting and ASR 1-best are hybridized in order to retrieve first names which are easier for ASR.

6. Speaker and Face Diarization

The task of diarization aims at determining for each pair of (visual or acoustic) frames whether it contains the same person. This task is often referred to as clustering.

6.1. Speaker diarization

The diarization system used in this work is the one presented in [15]. It is a sequential processing using firstly Bayesian Information Criterion and then Cross-likelihood Criterion, with special attention paid for overlapped speech. Overlapped speech segments are first detected and discarded from the clustering process, and then reassigned to the 2 nearest speakers, in terms of temporal distance between speech segments. Processing overlapped speech is particularly interesting for shows including debates.

6.2. Face diarization

Faces are detected using OpenCV's cascade classifier [16] for frontal and profile faces. The resulting detections are tracked until shot boundaries using bounding box overlap. Then, the upper body is detected using a background subtraction algorithm based on Grabcut [17], initialized with detected face. The background subtraction algorithm yields a very accurate silhouette of the person, even in presence of a dynamic background. Each extracted person is then modelled using a space-time color histogram [18]. This model stores color along with geometric and time information. It allows to retain the aspect of the person as it moves throughout the shot. A similarity matrix is built between person tracks using a combination of Bhattacharyya coefficient and Mahalanobis distance [18]. In the PERCOLI system, the similarity matrix is directly used in the face identification process as described in section 7.2 without requiring a specific clustering process.

7. Multimodal Fusion

As mentioned in section 3, our system identifies speaker identities in a first step and identifies face identities in a second step. Both identification steps are achieved thanks to a multimodal fusion system composed of two modules :

- Local identities propagation
- Show-specific post-processing

The following subsections describe the nature of local identities that are propagated for both steps, along with the generic propagation approach and the specific post-processing stage. Note that the different strategies have been designed in order to minimize the EGER metric (defined in section 8) for which all types of errors are equally weighted. In particular, substitutions and omissions having the same cost, we have chosen to try to give an identity to every detected speaker or face.

7.1. Multimodal speaker identification

Local identities for speaker identification are OPN hypothesis and scored speaker recognition (SR) hypothesis. The score

[2] https ://code.google.com/p/tesseract-ocr/

of an SR hypothesis is obtained thanks to a re-ranking process applied on the speaker recognition n-best list provided by speaker models. Re-ranking is based on the acoustical score and the presence of the speaker name in overlaid text as described in details in [19]. UPNs extracted from the spoken content are used in the post-processing and validation steps, along with show-specific *a priori* knowledge.

The core **identity propagation** method consists in attributing local identities to speakers in the following way :

1. direct identification : SR hypothesis are attributed to their corresponding speaker turns if their score is above a given threshold ; then for each local OPN hypothesis, the speaker turn which has the maximum duration overlap with the OPN span is given the identity carried by the OPN hypothesis ;
2. indirect identification : each unidentified speaker turn is given the identity of its speaker cluster, i.e. the OPN hypothesis which has the maximum duration overlap with the whole cluster or the SR which has the maximal score over the whole cluster if there is no such OPN hypothesis along the cluster.

The first **post-processing** step applies on speaker turns that have not been identified in the previous propagation step. It consists in using specific knowledge about the shows to identify speakers that cannot be identified by the core propagation step. It is applied for the unsupervised system where some speaker turns can remain unnamed after the propagation step. It is particularly designed for the identification of voice-overs for shows that contain reports commented by journalists (LCP_CaVousRegarde, LCP_LCPInfos, BFMTV_BFMStory, BFMTV_PlaneteShowbiz). The identity of such journalists is usually not displayed in the overlaid text and can only be retrieved from the spoken content. To this purpose, we use a predefined list of potential journalists for each type of show, and perform a specific name spotting in the audio content leading to a set of specific UPN hypotheses. A show can potentially contain several voice-over reports and we make the assumption the voice-over journalists are introduced before their report. Hence, after each specific UPN hypothesis, we attribute the corresponding identity to every unidentified speaker turn until the next specific UPN hypothesis.

Finaly a *global validation* step is performed for two particular shows for which the number of speakers is known in advance. It is the case of debate shows that only contain on-stage debates without any additional reports (LCP_PileEtFace with only three speaker and LCP_EntreLesLignes with five speakers). If the overall number of speaker identity hypotheses is above the a priori number of speakers N, the N most frequent hypotheses are kepts and the others are simply replaced by the most frequent hypothesis.

7.2. Multimodal face identification

Local identities for face identification are OPN hypothesis and speaker identities output by the previous multimodal speaker identification stage. UPNs extracted from the spoken content are used in the post-processing step, along with show-specific *a priori* knowledge.

The core **identity propagation** method is also the succession of a direct and an indirect identification steps. The *direct face identification* step follows the assumption that most OPNs occur while the corresponding face appears on the screen. Statistics on the *REPERE* corpus corroborate this idea, showing that 98.5% of the annotated frames containing an OPN also contain the corresponding face. Consequently in unambiguous shots where only one face is detected, we locally propagate the OPN to the face track. Then, for ambiguous shots where multiple faces could be identified by an OPN, we make a global decision using bipartite matching. For a given OPN, potential face sets are formed by gathering all face tracks that do not co-occur in the same shot. Then, that name is associated to the purest cluster containing all shots it occurs in.

The *indirect face identification* approach implemented in the PERCOLI system includes $Face \rightarrow Face$ propagation and cross-modal $Speaker \rightarrow Face$ propagation. First, the $Face \rightarrow Face$ propagation corresponds to the "*similarity based*" approach described in [20]. It is based on the principle that directly-named faces are very reliable, and can be considered as *models* in an open-set face identification paradigm. Let $\hat{g}_n$ be the set of faces directly associated to name n, for each face f with no direct naming, a distance D is defined between this face and $\hat{g}_n$. The name $\hat{n}(f)$ given to face f is the name for which the distance is minimal, if the distance $< \theta_1$.

$$\hat{n}(f) = \left\{ \begin{array}{l} \mathrm{argmin}_{n \in N} D(\hat{g}_n, f) \,\mathrm{if}\, D(\hat{g}_n, f) < \theta_1 \\ \emptyset \,\mathrm{otherwise} \end{array} \right.$$

with $D(g, f) = \frac{1}{|g|} \sum_{f_i \in g} d(f, f_i)$ being computed on the basis of the similarity matrix described in section 6.2. Note that for the moment, it has been applied only for shows from LCP channel (except LCP_TopQuestions) because similarity matrix were not reliable enough for the other shows that have a much more complex video editing policy.

The second aspect of our *indirect face identification* approach consists in attributing an identity to the remaining face tracks from the speaker modality. The identity of the speaker which has the maximal temporal overlap with the face track is given to the track. Note that if the system had to be designed in order to optimize identity precision, this *Speaker* → *Face* propagation should be refined and conditioned for instance to the response of a talking face detector.

The **post-processing** module relies on *a priori* knowledge of the shows, regarding their structure and their staging. If the structure of reports is difficult to model, studio set parts of a show usually follow a regular staging process. In order to exploit this specific knowledge, a set of rules has been manually designed for 4 different shows.

BFMTV_PLANETESHOWBIZ is dedicated to show business news : the show starts with an introduction by two anchor journalists in studio followed by several voice-over reports

→if one of the anchors is speaking and two faces are detected, then the second face is given the identity of the second anchor.

LCP_LCPINFO is a classical Brodcast News show with an alternation of reports introduced by an anchor speaker (identified as the first OPN hypothesis), and studio interviews between the anchor and the principal guest (identified as the most frequent OPN hypothesis). The staging of this show implies that during the interviews, the anchor and the guest can appear simultaneously on screen with smaller face sizes.

→if the principal guest is identified by the identity propagation step and has a small size, a second face track corresponding to the anchor is systematically added.

LCP_PILEETFACE is a debate between two politicians, their names are detected as being the two most frequent OPN hypotheses along the show. They appear on screen whether alone or together with a smaller head size.

→*if one politician face track is identified by the identity propagation step and has a small size, a second face track corresponding to the second guest is systematically added.*

LCP_EntreLesLignes is a debate between four journalists which are sitting two by two on both sides of a square table. Their names are detected as being the four most frequent OPN hypotheses along the show. A specific spotting of these four names in the audio content of the very beginning of the show (when they are presented by the debate animator), allows to infer their position around the table. Actually they are always presented in the same order, and it is possible to infer who is sitting next to who and who is facing who.

→*if one guest face track is identified by the identity propagation step and has a small size, a second face track corresponding to his neighbour is systematically added.*

These rules are very specific but can cover a large proportion of shots in a studio show which follows a regular and structured staging.

8. Evaluation

Metric		**EGER**		**Pre.**	**Rec.**	**F-m**
Modality	**speak.**	**head**	**s+h**		**speaker+head**	
Sup. local	24.4	53.5	40.2	75.4	62.7	68.5
+ post-proc	24.5	50.4	**38.6**	75.7	64.6	69.7
Unsup. local	36.3	58.8	48.5	81.6	51.9	63.4
+ post-proc	34.1	55.4	**45.7**	67.5	57.6	62.2

FIGURE 4 – Global results in terms of EGER and F-measure for the two fusion strategies (local fusion and local fusion + post-processing) for the supervised and unsupervised tasks. The post-processed systems correspond to REPERE submissions.

EGER	**local**	**+ post-proc**
BFMTV_BFMStory	46.0	43.9
BFMTV_CultureEtVous	93.5	81.5
LCP_CaVousRegarde	65.8	67.5
LCP_EntreLesLignes	61.6	57.1
LCP_LCPInfo	51.5	48.0
LCP_PileEtFace	51.5	38.0
LCP_TopQuestions	64.8	69.3
All	58.8	55.4

FIGURE 5 – Impact of the fusion strategy on each show, for the head modality (unsupervised mode).

The results presented in this section have been obtained on the 2013 REPERE test corpus during the challenge. The output of the automatic systems participating to the challenge is a list, for each video file, of temporal segments representing the identities of detected persons in the video with the corresponding modality, such as :

```
s1 227.6  240.1  speaker Valerie_PECRESSE
s1 237.9  256    head    Nicolas_SARKOZY
s1 249.2  252.7  speaker Nicolas_SARKOZY
s1 282.2  284.1  head    Valerie_PECRESSE
```

The first field is the show id, then the time window, the modality (*head, speaker*) and the name in a normalized form (first name/last name). The 2013 REPERE test corpus contains 2 hours of video from 2 TV channels and 7 different shows. The evaluation was performed on 1187 manually annotated key-frames containing 1165 speaker identities and 1386 face identities. The official scoring metric of the REPERE challenge is an error metric called *Estimated Global Error Rate* (EGER). This metric compares the list of person names produced by the automatic systems on the key-frames with the reference list. One or several modalities can be considered in the scoring. EGER computes the error rate by adding three kinds of errors : *confusion*, *false alarm* and *missed detection*. The cost of each error is set to 1 and the following score is computed :

PRIMARY	**Supervised**		**Unsupervised**	
Origin	**%Test**	**%Corr.**	**%Test**	**%Corr.**
Direct OCR	12.7	98.5	12.7	98.5
Face similarity	20.8	84.3	20.8	84.3
Speaker →Face	49.7	67.8	49.7	59.1
Post-processing	16.6	86.0	16.6	84.0
Total	100.0	77.3	100.0	72.7

FIGURE 6 – Origin of face identities in the primary system output for LCP (except LCP_TopQuestions).

$$EGER(m) = \frac{Conf(m) + FA(m) + Miss(m)}{\#\ of\ person\ name\ in\ modality\ m}$$

In the official results, the main results are given for the *head+speaker* modality. The performance of the PERCOLI system presented in this paper are displayed in table 4 in term of EGER and F-score. We present two variants : one only with only local propagation, and the submitted output with the post-processing presented in section 7. Compared to other participants, a stratified shuffling test [21] shows that our supervised system is significantly worse than the best participant ($\Delta = 5.1, p = 0.026$), and that our unsupervised submission is not significantly different from the best submission ($\Delta = 3.4$, $p = 0.433$). In addition, Table 5 shows the impact of post-processing on each show. Clearly, the strategy only pays off for a few shows which have a stable structure. Figure 6 shows the origin of face naming decisions in the system for the subset of shows we were able to process with face similarity matrix. OCR direct naming and Face similarity-based naming are very accurate but only cover a third of the faces that we were able to identify. Naming from the co-occurent speaker is not very precise, but enables to name a large set of faces, for which no other information is available. Finally, the show specific post-processing is fairly accurate for these shows.

9. Conclusions

This paper presents the PERCOL system for the first phase of the REPERE challenge. The system is focused on the unsupervised task which precludes the use of prior biometric models. Person identification is achieved by (1) detecting names in overlaid text and speech, (2) linking those name hypotheses to large databases of known people, (3) propagating them to detected speakers and faces through clustering and show-specific heuristics. In addition the variant of the system for supervised identification uses speaker models as another source of identity. In particular, propagation is achieved first on speakers, and then on faces, because of the confidence we have in those two modalities. In the official evaluation, this approach performed on par with the best system on the main unsupervised task and not significantly worse than the second best system on the main supervised task. Future work includes taking advantage of training data to learn how to merge identity hypotheses from the various components of the system, as well as inserting face identification in the pipeline.

10. Acknowledgements

This work is supported by the ANR project PERCOL, under ANR contract number 2010-CORD-102-01.

11. References

[1] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The repere corpus : a multimodal corpus for person recognition," in *LREC*, 2012.

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

[3] F. Liu and Y. Liu, "Identification of soundbite and its speaker name using transcripts of broadcast news speech," *ACM*, 2010.

[4] J. Poignant, H. Bredin, V.-B. Le, L. Besacier, C. Barras, and G. Quénot, "Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast," *Interspeech*, 2012.

[5] T. Cour, C. Jordan, and B. Taskar, "Learning from ambiguously labeled images," *CVPR*, 2009.

[6] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in tv video," *Image Vision Comput.*, 2009.

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

[8] S. Satoh and T. Kanade, "Name-it : Association of face and name in video," *CVPR*, 1997.

[9] D. Ozkan and P. Duygulu, "A graph based approach for naming faces in news photos," *CVPR*, 2006.

[10] C. Liu, S. Jiang, and Q. Huang, "Naming faces in broadcast news video by image google," *ACM*, 2008.

[11] H. Bredin, J. Poignant, M. Tapaswi, G. Fortier, V. B. Le, T. Napoléon, G. Hua, C. Barras, S. Rosset, L. Besacier *et al.*, "Fusion of speech, faces and text for person identification in tv broadcast," in *ECCV Workshop on Information fusion in Computer Vision for Concept Recognition*, 2012.

[12] F. Bechet and E. Charton, "Unsupervised knowledge acquisition for extracting named entities from speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, Dallas, USA, 2010.

[13] B. Sagot and R. Stern, "Aleda, a free large-scale entity database for French," in *Proceedings of LREC 2012*, Istanbul, Turquie, 2012, p. 4 pages. [Online]. Available : http ://hal.archives-ouvertes.fr/hal-00699300

[14] M. Delakis and C. Garcia, "Text detection with convolutional neural networks," *VISAPP*, 2008.

[15] D. Charlet, C. Barras, and J. Lienard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates," *ICASSP*, 2013.

[16] P. Viola and M. Jones, "Robust real-time object detection," *IJCV*, 2002.

[17] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut : Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, 2004.

[18] R. Auguste, A. Aissaoui, J. Martinet, and C. Djeraba, "Les histogrammes spatio-temporels pour la ré-identification de personnes dans les journaux télévisés," *CORESA*, 2012.

[19] D. Charlet, C. Fredouille, G. Damnati, and G. Senay, "Improving speaker identification in tv-shows using person name detection in overlaid text and speech," *submitted to Interspeech*, 2013.

[20] 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," *CBMI*, 2013.

[21] E. Yücesan, "Evaluating alternative system configurations using simulation : A nonparametric approach," *Annals of Operations Research*, vol. 53, no. 1, pp. 471–484, 1994.