# QCompere @ REPERE 2013

*Hervé Bredin*[1], *Johann Poignant*[2], *Guillaume Fortier*[3], *Makarand Tapaswi*[4], *Viet Bac Le*[5], *Anindya Roy*[1], *Claude Barras*[1], *Sophie Rosset*[1], *Achintya Sarkar*[1], *Qian Yang*[4], *Hua Gao*[4], *Alexis Mignon*[6], *Jakob Verbeek*[3], *Laurent Besacier*[2], *Georges Quénot*[2], *Hazim Kemal Ekenel*[4], *Rainer Stiefelhagen*[4]

[1]LIMSI-CNRS, Université Paris-Sud, BP 133, 91403 Orsay, France
[2]UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS-LIG UMR 5217, F-38041 Grenoble, France
[3]INRIA Rhone-Alpes, 655 Avenue de l'Europe, F-38330 Montbonnot, France
[4]KIT, Karlsruhe Institute of Technology, Karlsruhe, Germany
[5]Vocapia Research, 28 rue Jean Rostand, Parc Orsay Université, 91400 Orsay, France
[6] Université de Caen / GREYC UMR 6072, F-14050 Caen Cedex, France

## Abstract

We describe QCompere consortium submissions to the REPERE 2013 evaluation campaign. The REPERE challenge aims at gathering four communities (face recognition, speaker identification, optical character recognition and named entity detection) towards the same goal: multimodal person recognition in TV broadcast. First, four mono-modal components are introduced (one for each foregoing community) constituting the elementary building blocks of our various submissions. Then, depending on the target modality (speaker or face recognition) and on the task (supervised or unsupervised recognition), four different fusion techniques are introduced: they can be summarized as propagation-, classifier-, rule- or graph-based approaches. Finally, their performance is evaluated on REPERE 2013 test set and their advantages and limitations are discussed.

**Index Terms**: speaker identification, face recognition, named entity detection, video optical character recognition, multimodal fusion

## 1. Introduction

The REPERE challenge[1] aims at gathering four communities (face recognition, speaker identification, optical character recognition and named entity detection) towards the same goal: multimodal person recognition in TV broadcast. It takes the form of an annual evaluation campaign and debriefing workshop. In this paper we describe the submissions of the QCompere consortium to the 2013 REPERE evaluation campaign [1]

Given TV broadcast videos such as news or talk-shows, the main objective of the REPERE challenge is to answer two questions: *who speaks when?* and *who appears when?*. We distinguish two subtasks: either supervised (when prior identity models are allowed) or unsupervised recognition (when prior models are forbidden and person names must be automatically extracted from the test videos themselves). Speaker and face recognition both rely on a priori models of each person to be recognized: they fall in the supervised recognition category. Our mono-modal (audio or visual) person recognition modules are introduced in Section 2. However, other sources of information are available in TV broadcast and can be used to achieve unsupervised person recognition; such as named entities detected in automatic speech transcription, and block titles usually written on screen to introduce reporters and interviewees. Our efforts in this direction are described in Section 3.

The main contributions of the QCompere consortium lie in the way these modules are combined into a multimodal person identification framework. In Section 4, we propose a classifier-based late fusion approach and another one modeling person recognition as a shortest path problem in a multimodal probability graph. QCompere runs submitted to the 2013 campaign are evaluated and compared in Section 5. Finally, Section 6 concludes the paper.

## 2. Supervised Person Recognition

In this section, we only describe the mono-modal supervised person recognition approaches (speaker and face).

### 2.1. Speaker Recognition

Speaker diarization (SD) is the process of partitioning the audio stream into homogeneous clusters without prior knowledge of the speaker voices and serves as a pre-processing step for the speaker identification module. Two SD systems were developed, respectively by LIMSI and KIT.

LIMSI's SD system relies on two steps: agglomerative clustering based on the BIC criterion to provide pure clusters followed by a second clustering stage using cross-likelihood ratio (CLR) as distance between the clusters [2]. Additionally, since the corpus contains several shows for each recorded program, the same identifier has to be associated to a given speaker across all the shows. Following previous experiments on cross-show speaker diarization, a first, local clustering stage is followed by a CLR clustering across all the shows; this hybrid approach was found to provide a good performance while being computationally acceptable for a corpus lasting a few hours [3].

KIT's SD system contains the following components. Audio segmentation first discriminates speech from non-speech segments. It is implemented using a HMM segmenter with 4 GMMs for speech, silence, noise and music. Speaker turn detection [4] is then applied on the segments longer than $5s$. A first-pass BIC clustering groups the segments from the same speaker together. Viterbi re-segmentation refines the segment

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). Corresponding author: bredin@limsi.fr

[1] http://www.defi-repere.fr

boundaries. The speaker models are trained on the clustering results. The features are 20-dimensional MFCC plus their first derivatives. Feature warping is applied to compensate channel effects. GMMs with 64 Gaussians are used to model the speakers. A second-pass BIC clustering and Viterbi re-segmentation further refines the segment boundaries and clustering results. Finally, the post processing merges the adjacent segments from the same speakers which are separated by the silence shorter than $0.5s$.

Unsupervised speaker diarization is followed by a cluster-wise speaker identification. We implemented two systems [5]. Our baseline system follows the standard Gaussian Mixture Model-Universal Background Model (GMM-UBM) paradigm, and the GSV-SVM system uses the super-vector made of the concatenation of the UBM-adapted GMM means to train one Support Vector Machine classifier per speaker. For both systems, each cluster 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 648 speakers in our experiments: the REPERE training set, the ETAPE training and development data[2] and additional French politicians data extracted from French radios.

### 2.2. Face Recognition

The supervised face recognition process is divided into face detection and tracking stage, and the recognition stage.

Face tracking is performed using particle filtering approach [6], 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 recognition uses a frontal face descriptor. First a detector locates nine landmarks on the face, around the eyes, the nose and the mouth. We use a tree-structured constellation model [7] that computes Histogram of Gradient (HoG) features [8] for detection of these facial landmarks. Once the landmarks are detected, faces are aligned using an affine transformation, and a second HoG descriptor is computed around each of the nine 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 final descriptor concatenates the local gradient orientation histograms to form a $9 \times 10 \times 7 \times 7 = 4410$ dimensional feature vector per face (9 landmarks × 10 orientation bins × a grid of $7 \times 7$ spatial bins). For each track, we compute a mean HoG descriptor from all the frontal face detections found along the track. A database is automatically generated using a training set of annotated faces for learning the face recognition models. A Support Vector Machine (SVM) classifier is trained for each person, using one-versus-rest approach.

For the test set, we score the mean descriptor of each track using the learned models. The best scoring model is chosen and the face is tagged with the corresponding name, provided its score is higher than the decision threshold. The initial face recognition stage is followed by an unsupervised face clustering stage that is used to extend the labels to faces not named in the previous step. For each track, the mean HoG descriptor is projected on to a 200 dimensional descriptor using Logistic Discriminant Metric Learning approach (LDML) [9]. The learned face metric is used in a nearest neighbor classifier to assign names to tracks that were unlabeled so far; but only if the ratio of distances to the first and the second neighbor is sufficiently small.

## 3. Person Name Detection

Speaker and face recognition both rely on a priori models of each person to be recognized: they fall in the "supervised recognition" category. However, other sources of information are available in TV broadcast and can be used to achieve unsupervised person recognition.

### 3.1. Written Name Detection

In order to detect the names written on the screen used to introduce a person, a detection and transcription system is needed. For this task we used LOOV [10] (LIG Overlaid OCR in Video). This system has been previously evaluated on another broadcast news corpus with low-resolution videos. 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 for detecting the spatial positions of title blocks. This technique compares each transcript with a list of famous names (list extracted from Wikipedia, 175k names). Whenever a transcription corresponds to a famous name, its spatial position is added to a list. The repeating positions in this list provide the spatial positions of title blocks used to introduce a person. However, the detected text boxes do not always contain a name. A simple filtering based on some linguistic rules allows to filter false positives. Transcription errors are corrected using our Wikipedia list when the edit distance is small. The use of LOOV pipelined with our written names detection technique provides an F1-measure of 97.5% (see Table 1). The few remaining errors are due to transcription or filtering errors.

### 3.2. Spoken Name Detection

The aim of this task is to detect all person names spoken during a TV program and link each instance of a spoken name to the identity of a real person in terms of a normalized identifier (in the form `Firstname_LASTNAME`). In the first step, the acoustic data is processed by a Speech-To-Text (STT) module. Second, the transcripts produced by the STT module are processed by a Named Entity Recognizer (NER) to detect person names. Note that the name in its spoken form may include only *part* of the name (first, middle or last name, eg. "Hollande" instead of "François Hollande"), an acronym or even a nickname. From these incomplete forms, the correct full name has to be guessed. This necessitates a post-processing step applied to the output of the ASR-NER modules. 6427, 1555 and 1947 spoken names were present in the training set, the development set and the test set, respectively.

A state-of-the-art off-the-shelf STT system for French [11] was used to transcribe the audio data. No task-specific adapta-

| Modalities | Precision | Recall | F1-measure |
|---|---|---|---|
| written names | 99.4% | 95.7% | 97.5% |

Table 1: Quality of written names extraction for names written in title blocks

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

| Post-processing | dev | test |
|---|---|---|
| None | 61.1 | 60.0 |
| Approach A | 51.9 | 53.4 |
| Approach B | **49.3** | **52.2** |

Table 2: Spoken name detection performance in terms of SER (%).

tion was made for the REPERE evaluation (i.e., the REPERE training dataset was not used to adapt the acoustic models or the language models). The system obtained a word error rate of 16.43% (on around 36k words) during the first evaluation campaign of the REPERE challenge. In the NER module, two independent CRF models were trained on data annotated for the Quaero project: (1) a model to detect the mention of person with at least a first or a last name, and (2) a model to detect the different part of a person mention (e.g. first name or last name). These models use the same features as in [12]. In the final post-processing module to complete or correct the output of NER, two distinct approaches were studied.

**Approach A** used information from the *NER output itself*. Each name `N` in the output corresponding to one audio document (or TV show) was first checked if it is *full* (i.e. in the form `Firstname LASTNAME`). If not, `N` was searched inside the output. If `N` was found as *part* of another name `M` which itself is *full*, as its first, middle or last name, then each instance of `N` was replaced by `M`. For example, if the NER output contained both `MONTEBOURG` and `Arnaud_MONTEBOURG`, then each instance of the former was replaced by the latter. After this step, all remaining names which were still not full were searched in the Wikipedia. If a corresponding full name is found, it was used to replace the original name. All names remaining which were not full were discarded.

**Approach B** used information from the *groundtruth training data*. A Lookup Table (LUT) is created where each row contained (1) a name as it appears in the groundtruth training data, and (2) the corresponding name as it appears in the output of the ASR-NER system. When evaluating on dev or test, the LUT was used to translate each NER output to its corresponding correct form. Note that this method works only if the name occurred in the training data.

The task was evaluated by using the Slot Error Rate (SER) defined as: $\mathrm{SER} = [\mathrm{I} + \mathrm{D} + 0.5 \times (\mathrm{T} + \mathrm{F})]/\mathrm{R}$ where I is the Insertion error, D the Deletion error, T the Type error (i.e. a name was detected at the correct position but not the right name), F the Frontier error (i.e. the correct name was detected but not at the right time point) and R is the number of reference intervals. Table 2 shows the results obtained by the two approaches in terms of the SER. In the table, “None” refers to the case where the output of ASR-NER was directly used for evaluation. Note that the post-processing steps reduced the SER by about 10% absolute and 16.7% relative. Also, Approach B performed about 1% absolute better than A. This shows the role of training data in the performance of the system. It was also found that (1) combining A and B did not improve the scores more than B alone, and (2) about 70% of the deletion errors were a result of the ASR module.

# 4. Multimodal Fusion

In this section, we describe the runs submitted to the main multimodal tasks (supervised and unsupervised)

## 4.1. Propagation-based fusion

**Unsupervised speakers recognition**: This method is based on our previous work [13] (method M3). Speaker diarization and overlaid names recognition are run independently from each other. Speaker diarization is tuned to achieve the best diarization performance (i.e. minimize the diarization error rate, DER). The mapping between written names and speaker clusters is based on the following observations:

- when only one name is written on screen, any co-occurring speech turn is very likely (95% precision according to the train set) to be uttered by this person;
- the speaker diarization system can produce over-segmented speaker clusters, i.e. split speech turns from one speaker into two or more clusters.

Therefore, this method proceeds in two steps. First, speech turns with exactly one co-occurring name are tagged. Then, each remaining unnamed speech turn is tagged cluster-wise using an approach similar to the classical *Term-Frequency Inverse Document Frequency* (TF-IDF). We made two slight updates to this method: we reduce the temporal scope of each written names to the more co-occurring speech turn, this can correct the time offset between audio and written names segmentation. We also add the information of pronounced names: we name each remaining unnamed speech turn with closest pronounced names; this increases the number of speech turns named by our method.

**Unsupervised faces recognition**: As already stated, when one or more names are written on the screen, there is a very high probability that the name of one of the appearing face corresponds to the name written on screen. Therefore we use the information provided by written names during the face clustering process.

Before clustering, we associate each written name $n$ to the co-occurring face. At this stage, a face can have several names if several names are written on the screen at the same time. Then, regular agglomerative clustering (based on face similarity) is performed with the constraint that merging two clusters $s$ without at least one name $n$ in common is forbidden.
For example, two clusters $s_1$ and $s_2$ **can** be merged into a new one $s_{new}$ in the following case (the list of associated names is shown between brackets):

- $s_1(\emptyset) \cup s_2(\emptyset) \Rightarrow s_{new}(\emptyset)$
- $s_1(n_1) \cup s_2(\emptyset) \Rightarrow s_{new}(n_1)$
- $s_1(n_1, n_2) \cup s_2(\emptyset) \Rightarrow s_{new}(n_1, n_2)$
- $s_1(n_1, n_2) \cup s_2(n_1) \Rightarrow s_{new}(n_1)$

Below are examples where the two clusters **cannot** be merged:

- $s_1(n_1) \cup s_2(n_2) \Rightarrow$ Forbidden
- $s_1(n_1, n_3) \cup s_2(n_2) \Rightarrow$ Forbidden

The clustering is stopped according to the optimal threshold on the training set (minimizing the EGER, see Section 5.1).

## 4.2. Classifier-based Fusion

**Speaker identification**: Once all monomodal components have been run on a video, their outputs can be combined to improve the overall person recognition performance. Figure 1 draws up their list, along with two slightly modified versions of OCR: extended to the whole speech turns (`OCR`$^+$) or speaker diarization clusters (`OCR`$^*$).

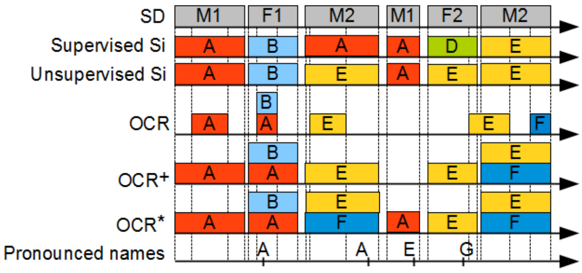

Figure 1: Several annotation timelines

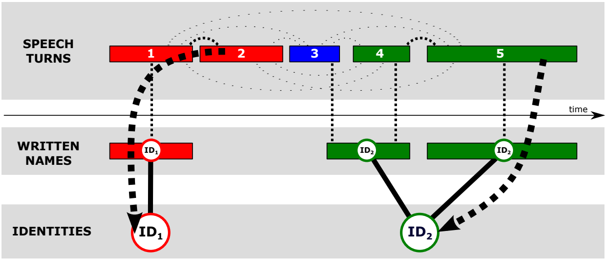

Figure 2: Multimodal probability graph for unsupervised speaker recognition, and two maximum probability paths.

Since each modality relies on its own temporal segmentation, the first step consists in aligning the various timelines onto the finest common segmentation. The final decision is taken at this segmentation granularity. For each resulting segment $\mathcal{S}$, a list of possible identities is built based on the output of all modalities. For each hypothesis identity $\mathcal{P}$, a set of features is extracted:

- Does the name of $\mathcal{P}$ appear in OCR? in OCR$^+$? in OCR$^*$?
- Duration of appearance of names in OCR$^+$, in OCR$^*$ and their ratio.
- Speaker recognition scores for identity $\mathcal{P}$ provided by GSV-SVM SID and their difference to best competing scores.
- Is $\mathcal{P}$ the name proposed by the unsupervised speaker recognition system?
- Is $\mathcal{P}$ the most likely identity according to GSV-SVM SID?
- Has $\mathcal{P}$'s name been pronounced by the previous or the next speaker.

Based on these features, we trained a Multilayer Perceptron classifiers using Weka[3] to answer the following question: *"is $\mathcal{P}$ speaking for the duration of $\mathcal{S}$?"* Since these features can be either boolean or (unbounded) float, several classifiers insensitive to numerical types were used. The identity with the highest score is selected for the speaker identification task.

### 4.3. Rules-based Fusion

**Supervised face recognition**:

Several sources of information are exploited for multimodal and supervised face identification. They are combined using a set of simple rules, ordered by priority:

1. mono-modal face recognition for anchor persons;
2. names written on the screen;
3. unsupervised face recognition;
4. mono-modal face recognition for non-anchor persons;
5. multi-modal speaker recognition.

### 4.4. Graph-based Fusion

Alongside classifier-based approaches, the QCompere consortium also submitted a few contrastive runs based on a graphical representation of the person identification problem. For each video, a multimodal probability graph is built as illustrated in Figure 2. Each person utterance (e.g. a speech turn, a face track or a written name) is added as a vertex to this graph. For each target of supervised recognition systems (speaker identification or face recognition) and for each name found by name detection systems (written or spoken name detection), an identity vertex is added containing the normalized identifier of the person (e.g. `Nicolas_SARKOZY` or `Francois_HOLLANDE`).

Two vertices $i$ and $j$ are connected by an edge weighted by the probability $p_{ij}$ that they correspond to the same person. This probability is obtained differently depending on the vertices it connects:

**Intra-modal edges** connect vertices of the same modality (speech turns-to-speech turns, or face tracks-to-face tracks). Probabilities are derived from the similarity scores $d$ (BIC criterion for speech turns, learned metric for face tracks) using Bayes' theorem: $p\left(\mathcal{H} \mid d\right) = 1/(1+r)$ with $r = \frac{p(d|\overline{\mathcal{H}})}{p(d|\mathcal{H})}\frac{p(\overline{\mathcal{H}})}{p(\mathcal{H})}$ where $\mathcal{H}$ is the hypothesis that connected vertices are from the same person, and $\frac{p(d|\overline{\mathcal{H}})}{p(d|\mathcal{H})}$ and $\frac{p(\overline{\mathcal{H}})}{p(\mathcal{H})}$ are estimated using the annotated training set.

**Cross-modal edges** connect co-occurring vertices with two different modalities (e.g. a speech turn and a co-occurring written name) with a fixed probability estimated using the training set. For instance, two co-occurring speech turn and written name have more than 97% chance to correspond to the same person.

**Identity edges** connect detected names (written or spoken) to the corresponding identity with probability $p = 1$. They also connect speech turns and face tracks to target models (from supervised recognition system) with a probability derived from the identification scores.

Finally, person identification is achieved by looking for the maximum probability path between every speech turn (or face track) and all available identities. The probability of the path is simply defined as the product of the probability of its edges. It is straightforward to show that this maximum probability path problem can be modeled as a shortest path problem in the dual graph where edges are weighted by $-\log p_{ij}$ instead of $p_{ij}$. In Figure 2, speech turn #2 (resp. # 5) is given the identity `ID1` (resp. `ID2`).

The same framework can be used for speaker or face recognition; and for both supervised and unsupervised recognition. However, for the latter, one must remove *identity edges* coming from mono-modal speaker identification and face recognition system introduced in Section 2. Furthermore, one does not have to use all available edges to achieve the best performance. We only report on the best combination in Section 5.

The supervised run contains speech turn-to-identity (s-to-i), speech turn-to-written name (s-to-w) and w-to-i edges for

[3] `http://www.cs.waikato.ac.nz/ml/weka`

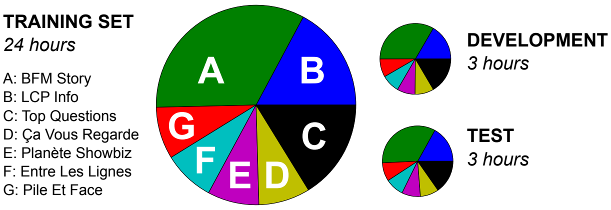

Figure 3: Training, development and test sets each contain 7 different types of shows (A to G).

speaker recognition, augmented with speech turn-to-face tracks (s-to-f) and f-to-w edges for face recognition. The unsupervised run contains s-to-w, h-to-w, w-to-i, s-to-h and h-to-h edges for both speaker and face recognition.

## 5. REPERE Evaluation Campaign 2013

### 5.1. Corpora & Metrics

Figure 3 provides a graphical overview of the REPERE video corpus 2013 [14] (training, development and test sets). Overall, it contains 188 videos (30 hours) recorded from 7 different shows broadcast by the French TV channels *BFM TV* and *LCP*.

While the audio annotation is dense (who speaks when?), the visual annotation (whose head appears when?) is only provided from one video frame every 10 seconds on average. [14] provides a more detailed description of the corpus and the associated annotation process.

Though the whole test set is processed, evaluation is only performed on the annotated frames $\mathcal{F}$. For each frame $f$, let us denote #total($f$) the number of persons in the reference. The hypothesis proposed by an automatic system can make three types of errors: false alarms (#fa) when it contains more persons than there actually are in the reference; missed detections (#miss) when it contains less persons than there actually are in the reference; confusions (#conf) when the detected identity is wrong. For evaluation purposes, and because unknown people cannot – by definition – be recognized in any way, they are excluded from the scoring. The Estimated Global Error Rate (EGER) is defined by:

$$\text{EGER} = \frac{\sum_{f \in \mathcal{F}} \#\text{conf}(f) + \#\text{fa}(f) + \#\text{miss}(f)}{\sum_{f \in \mathcal{F}} \#\text{total}(f)}$$

### 5.2. Experimental protocol

For our experiments, the training set was split in two balanced subsets *train A* and *train B*. Target models (for speaker identification and face recognition of Section 2) are obtained using *train A*. *train B* is used to train classifiers and graph probabilities introduced in Section 4. The *development* set allows to tune various fusion parameters, and the final evaluation is done on the *test* set.

### 5.3. Supervised Recognition

Table 3 summarizes the performance achieved by our submissions to the supervised recognition task. Looking at the monomodal tasks, the speaker recognition system performs significantly better than the face recognition system (44.2% vs. 61.1% in EGER), probably due to more important variability factors in the image: face size, orientation, exposition, etc. The classifier-based fusion is very effective and reduces the speaker EGER to 17.8%, a 60% relative reduction compared to the mono-modal performance. The improvement brought by the rule-based fusion for faces is also important with a 39% relative reduction of errors, from 61.1% to 37.3%. The graph-based fusion is less effective but still reduces the EGER by about 20% relative compared to the mono-modal systems.

| Approach | | EGER (%) |
|---|---|---|
| speaker | mono-modal speaker recognition | 44.2 |
| | classifier-based fusion | 17.8 |
| | graph-based fusion | 35.3 |
| head | mono-modal face recognition | 61.1 |
| | rule-based fusion | 37.3 |
| | graph-based fusion | 48.1 |

Table 3: Performance of the QCompere submissions to the supervised person recognition tasks.

### 5.4. Unsupervised Recognition

The performance achieved by our submissions to the unsupervised recognition tasks are presented in Table 4; they are of course worse than the performance of a supervised multi-modal fusion, roughly 8.8% absolute above them. But they are also significantly better than the mono-modal identification scores, with 26.2% EGER for speakers and 46.2% for heads; this had already been shown for speakers after the REPERE dry-run evaluation [13]. Interestingly, the performance of unsupervised graph-based fusion for speakers and faces is almost similar to the performance observed for the supervised case (38.1% vs. 35.3% for speakers and 50.3% vs. 48.1% for faces), showing that there is room for improvement for this approach with a better integration of the person identification scores.

| Approach | | EGER (%) |
|---|---|---|
| spk. | propagation-based fusion | 26.2 |
| | graph-based fusion | 38.1 |
| head | propagation-based fusion | 46.2 |
| | graph-based fusion | 50.3 |

Table 4: Performance of the QCompere submissions to the unsupervised person recognition tasks.

## 6. Conclusion and Future Work

In this paper, we described, evaluated and discussed QCompere consortium submissions to the REPERE 2013 evaluation campaign. As expected, we showed that speaker identification and face recognition can be greatly improved when combined with name detection through video optical character recognition and automatic speech transcription available in TV broadcast. Moreover, it should be highlighted that the unsupervised person recognition approaches that we proposed perform much better than state-of-the-art supervised mono-modal ones (for both speaker and face identification). However, results show a strong performance discrepancy in favor of speaker recognition for all three participating consortia [1] as well as for QCompere various approaches. Therefore, for next year evaluation (scheduled in January 2014), a strong effort should be focused on face recognition.

# 7. References

[1] O. Galibert and J. Kahn, “The First Official REPERE Evaluation,” in *First Workshop on Speech, Language and Audio for Multimedia (SLAM 2013)*, August 2013.

[2] 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, pp. 1505–1512, 2006.

[3] V.-A. Tran, V. B. Le, C. Barras, and L. Lamel, “Comparing Multi-Stage Approaches for Cross-Show Speaker Diarization,” in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011).*, Florence, Italy, August 2011, pp. 1053–1056.

[4] Q. Jin, K. Laskowski, T. Schultz, and A. Waibel, “Speaker segmentation and clustering in meetings,” 2004.

[5] V.-B. Le, C. Barras, and M. Ferràs, “On the use of GSV-SVM for Speaker Diarization and Tracking,” in *Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010, pp. 146–150.

[6] M. Bäuml, K. Bernardin, M. Fischer, H. K. Ekenel, and R. Stiefelhagen, “Multi-Pose Face Recognition for Person Retrieval in Camera Networks,” in *International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2010.

[7] M. Everingham, J. Sivic, and A. Zisserman, ““Hello! My name is... Buffy” Automatic Naming of Characters in TV Video,” in *British Machine Vision Conference*, 2006.

[8] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *CVPR*, 2005.

[9] M. Guillaumin, J. Verbeek, and C. Schmid, “Is that you? Metric learning approaches for face identification,” in *ICCV*, 2009.

[10] J. Poignant, L. Besacier, G. Quénot, and F. Thollard, “From Text Detection in Videos to Person Identification,” in *International Conference on Multimedia & Expo (ICME)*, 2012.

[11] L. Lamel, S. Courcinous, J. Despres, J.-L. Gauvain, Y. Josse, K. Kilgour, F. Kraft, V. B. Le, H. Ney, M. Nußbaum-Thom, I. Oparin, T. Schlippe, R. Schlüter, T. Schultz, T. F. da Silva, S. Stüker, M. Sundermeyer, B. Vieru, N. T. Vu, A. Waibel, and C. Woehrling, “Speech Recognition for Machine Translation in Quaero,” in *IWSLT*, San Francisco, CA, USA, 2011.

[12] M. Dinarelli and S. Rosset, “Models Cascade for Tree-Structured Named Entity Detection,” in *IJCNLP’11, 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, 2011.

[13] J. Poignant, H. Bredin, V. Le, L.Besacier, C.Barras, and G.Qunot, “Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast,” in *Interspeech 2012*, 2012.

[14] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, “The REPERE Corpus: a Multimodal Corpus for Person Recognition,” in *International Conference on Language Resources and Evaluation (LREC)*, 2012.