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

Multimodal Person Discovery in Broadcast TV: lessons learned from MediaEval 2015

Johann Poignant, Hervé Bredin and Claude Barras

Received: date / Accepted: date

Abstract We describe the "Multimodal Person Discovery in Broadcast TV" task of MediaEval 2015 benchmarking initiative. Participants were asked to return the names of people who can be both seen as well as heard in every shot of a collection of videos. The list of people was not known *a priori* and their names had to be discovered in an unsupervised way from media content using text overlay or speech transcripts. The task was evaluated using information retrieval metrics, based on *a posteriori* collaborative annotation of the test corpus. The first edition of the task gathered 9 teams which submitted 34 runs. This paper provides quantitative and qualitative comparisons of participants submissions. We also investigate why all systems failed for particular shots, paving the way for future promising research directions.

Keywords benchmark, information retrieval, unsupervised person recognition, multimodal fusion, error analysis

1 Introduction

TV archives maintained by national institutions such as the French INA, the Dutch Institute for Sound & Vision, or the British Broadcasting Corporation are rapidly growing in size. The need for applications that make these archives searchable has led researchers to devote concerted effort to developing technologies that create indexes. Human nature leads people to be very interested in other people. Therefore, knowing "who is seen" and "who is speaking" in TV Broadcast programs is very useful to provide efficient information access to large video collections. In this regard, benchmarking initiatives provide a framework to evaluate and compare person identification algorithms applied to these large archives.

LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay, France.
E-mail: firstname.lastname@limsi.fr

In the past, several corpora and evaluation campaigns were proposed for person identification. Based on a very large collection of short videos from different types, *TRECVID* Semantic Indexing task [35,54] aims at searching video segments for a pre-determined list of concepts (including a limited set of persons). The *REPERE* challenge aimed at supporting research on multimodal person recognition [5,28,22]. Its main goal was to answer the two questions *"who speaks when?"* and *"who appears when?"* using any available source of information (including pre-existing biometric models and person names extracted from text overlay and speech transcripts). To assess the technology progress, annual evaluations were organized in 2012, 2013 and 2014.

Conventional approaches for person recognition rely on the supervison of prior (face and/or voice) biometric models. Because the actual list of people appearing in a TV show is usually missing when it is first broadcasted, a very large amount of trained models (hundreds or more) is needed to cover only a decent percentage of all the people in the show. Regular anchors and interviewers may be known, but one cannot assume that biometric models will be available at indexing time for all participants. In addition, it is not always possible to predict which people will be the most important to find in the future. A model may not be available in advance, simply because the person is not (yet) famous. It is even possible that archivists annotating content by hand do not even know the name of the person and may not have the time to perform additional investigations.

A more challenging task is thus indexing people in the archive, under real-world conditions –*i.e.* when there is no pre-set list of people to index. A solution to these problems is to use other sources of information (see figure 1) for discovering people names: speech transcripts (*i.e.* pronounced names) and text overlay. (*i.e.* written names).

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

Such approaches can usually be decomposed into the three following steps.

1.1 Speech turns and/or face tracks clustering

The first step aims at grouping speech turns (resp. face tracks) into homogeneous clusters without prior knowledge on the voice (resp. the face) of the person. Each cluster must correspond to exactly one person and *vice versa*.

Though speaker diarization and face clustering have been addressed by both the speech and computer vision communities, there are still a lot of issues that need to be addressed: noisy audio or low video resolution, overlapping speech, sensitivity to changes in face pose, lighting and occultation, *etc.*. In addition, joint speech turns and face tracks multi-modal clustering is still an on-going and difficult research problem due to the difficult cross-modal speaker/face comparison problem [57].

1.2 Extraction of names candidates

In the past, most state-of-the-art approaches relied on pronounced names because they were dealing with audio recordings [14], [33] or because of optical character recognition errors when applied to poor quality videos. Recent video quality improvement made automatically extracted written names more reliable [41]. [39, 42] and [26] provide a fair comparison of written and pronounced names on their ability to provide names of people present in TV broadcast. When written names are available (*e.g.* for TV news or debates), it is more suitable to use them rather than pronounced names. Pronounced names, on the other hand, offer great theoretical naming potential (when using manual speech transcription). Though accuracy of automatic speech transcription has improved a lot recently [**?**], the out-of-vocabulary problem remains a significant issue, leading to named entities (and person names, in particular) not being transcribed correctly, nor recognized as such.

1.3 Assocation between clusters and name candidates

While a name written in a title block usually introduces the person currently on screen [41], pronounced names can refer to both the current speaker, their addressee or even someone else [14]. Therefore, propagating pronounced names to speech turns or face tracks clusters is a much more difficult and error-prone task than propagating written names [39, 42]. Efficiently combining these two approaches is also a research area that has yet to be explored [10].

In a nutshell, a lot remains to be done towards improving unsupervised person recognition in TV broadcast, in order to support more efficiently archives indexing and search. This was our motivation for proposing a new benchmarking task, as a follow-up of the now completed REPERE evaluation campaigns, in the context of the MediaEval. Section 2 defines the task, the development and test datasets, and the evaluation metric. Section 3 provides the state-of-the-art of existing methods relevant to the task. Section 4 describes the architecture of the baseline system that was distributed to participants and Section 5 reports the experimental results of the participants and discusses their performances. In Section 6, detailed error analysis aims at uncovering the main limitations of current approaches. Feedback on the organizational aspects of the first edition of *"Multimodal Person Discovery"* at MediaEval 2015 are given in Section 7. Finally, Section 8 concludes the paper.

2 Definition of the task

Participants were provided with a collection of TV broadcast recordings pre-segmented into shots. Each shot $s \in \mathbb{S}$ had to be automatically tagged with the names of people both speaking and appearing at the same time during the shot: this tagging algorithm is denoted by $\mathcal{L} : \mathbb{S} \mapsto \mathcal{P}(\mathcal{N})$ in the rest of the paper. We choose to only take into account persons present both in audio and image stream as they can be considered as persons of interest.

The main novelty of the task is that the list of persons was not provided *a priori*, and person biometric models (neither voice nor face) could not be trained on external data. The only way to identify a person was by finding their name $n \in \mathcal{N}$ in the audio (*e.g.* using speech transcription – ASR) or visual (*e.g.* using optical character recognition – OCR) streams and associating them to the correct person. This made the task completely unsupervised (*i.e.* using algorithms not relying on pre-existing labels or biometric models).

Because person names were detected and transcribed automatically, they could contain transcription errors to a certain extent (more on that later in Section 2.2). In the following, we denote by $\mathbb{N}$ the set of all possible person names in the universe, correctly formatted as `firstname_lastname` – while $\mathcal{N}$ is the set of hypothesized names.

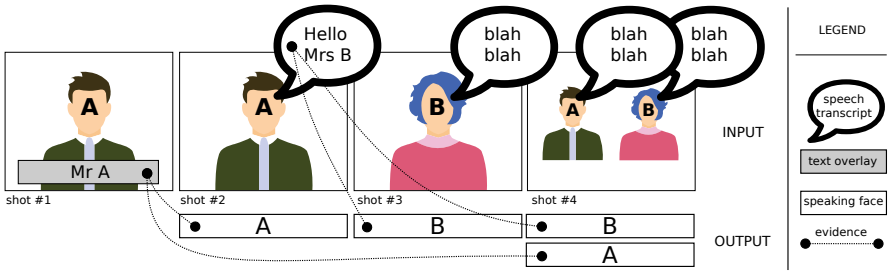

Fig. 2: For each shot, participants had to return the names of every speaking face. Each name had to be backed up by an evidence.

To ensure that participants followed this strict *"no biometric supervision"* constraint, each hypothesized name $n \in \mathcal{N}$ had to be backed up by a carefully selected and unique shot prooving that the person actually holds this name n: we call this an evidence and denote it by $\mathcal{E} : \mathcal{N} \mapsto \mathbb{S}$. In real-world conditions, this evidence would help a human annotator double-check the automatically-generated index, even for people they did not know beforehand.

Two types of evidence were allowed: an *image* evidence is a shot during which a person is visible, and their name is written on screen; an *audio* evidence is a shot during which a person is visible, and their name is pronounced at least once during a [shot start time $- 5s$, shot end time $+ 5s$] neighborhood. For instance, in Figure 2, shot #1 is an *image* evidence for Mr A (because his

name and his face are visible simultaneously on screen) while shot #3 is an *audio* evidence for Mrs B (because her name is pronounced less than 5 seconds before or after her face is visible on screen).

2.1 Datasets

The REPERE corpus – distributed by ELDA[1] – served as development set. It is composed of various TV shows (around news, politics and people) from two French TV channels, for a total of 137 hours. A subset of 50 hours is manually annotated. Audio annotations are dense and provide speech transcripts and identity-labeled speech turns. Video annotations are sparse (one image every 10 seconds on average) and provide overlaid text transcripts and identity-labeled face segmentation. Both speech and overlaid text transcripts are tagged with named entities.

The test set – distributed by INA[2] – contains 106 hours of video, corresponding to 172 editions of evening broadcast news *"Le 20 heures"* of French public channel *"France 2"*, from January 1st 2007 to June 30st 2007. As the test set came completely free of any annotation, it was annotated *a posteriori* based on participants' submissions. Hence, among the 64k shots it is made of, 27k have been annotated collaboratively and used for evaluation.

In the following, task groundtruths are denoted by function $\mathbb{L} : \mathbb{S} \mapsto \mathcal{P}(\mathbb{N})$ that maps each shot s to the set of names of every speaking face it contains, and function $\mathbb{E} : \mathbb{S} \mapsto \mathcal{P}(\mathbb{N})$ that maps each shot s to the set of person names for which it actually is an evidence.

2.2 Evaluation metric

Precision and recall are usual evaluation metrics for information retrieval tasks; but as recall would require an exhaustive annotation of the corpus not available in our case, the task was evaluated using a variant of Mean Average Precision (MAP), that took the quality of evidences into account. For each query $q \in \mathbb{Q} \subset \mathbb{N}$ (`firstname_lastname`), the hypothesized person name n_q with the highest Levenshtein ratio ρ to the query q is selected ($\rho : \mathbb{N} \times \mathcal{N} \mapsto [0, 1]$) We chose not to provide a predefined thesaurus in order to evaluate the capabilities of systems to actually detect and transcribe person names. Combined with the evidence-weighting introduced in the following paragraph, this also had the side effect of enforcing the strict "no-supervision" rule to participants:

$$n_q = \arg\max_{n \in \mathcal{N}} \rho\,(q, n) \text{ and } \rho_q = \rho\,(q, n_q)$$

[1] ISLRN: 360-758-359-485-0

[2] `http://dataset.ina.fr`

Average precision $\mathrm{AP}(q)$ is then computed classically based on relevant and returned shots:

$$\begin{aligned}\mathrm{relevant}(q) &= \{s \in \mathbb{S} \mid q \in \mathbb{L}(s)\} \\ \mathrm{returned}(q) &= \{s \in \mathbb{S} \mid n_q \in \mathcal{L}(s)\}_{\substack{\text{sorted by} \\ \text{confidence}}}\end{aligned}$$

Proposed evidence is Correct if name n_q is close enough to the query q and if shot $\mathcal{E}(n_q)$ actually is an evidence for q:

$$C(q) = \begin{cases} 1 & \text{if } \rho_q > 0.95 \text{ and } q \in \mathbb{E}(\mathcal{E}(n_q)) \\ 0 & \text{otherwise} \end{cases}$$

To ensure participants do provide correct evidences for every hypothesized name $n \in \mathcal{N}$, standard MAP is altered into EwMAP (Evidence-weighted Mean Average Precision), the official metric for the task:

$$\mathrm{EwMAP} = \frac{1}{|\mathbb{Q}|} \sum_{q \in \mathbb{Q}} C(q) \cdot \mathrm{AP}(q)$$

In practice, we observed that MAP and EwMAP values were all very close to each other (see Table 1 in Section 5.2), highlighting the fact that participants did follow the "no-supervision" rule. A discrepancy between those two values could have been partially explained by the use of external biometric models: how else could a participant be able to correctly recognize a person without providing a correct evidence?

Based on all submissions, 1642 person names having a corresponding speaking face in the test set were selected as queries $\mathbb{Q}$. Figure 3 shows that most queried people appear only once in the corpus.

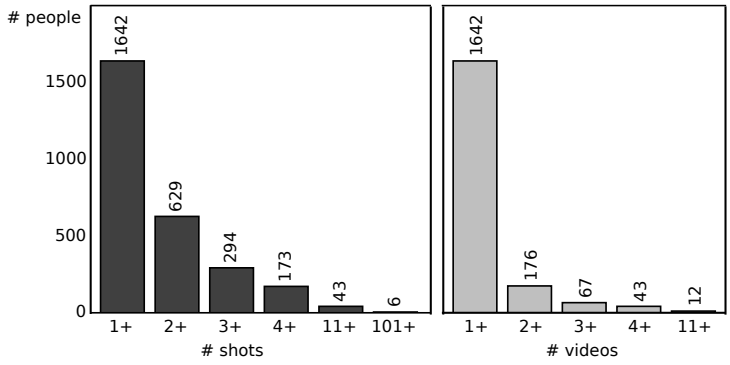

Fig. 3: Number of queried people, as a function of the minimum number of shots (or videos) they appear in. For instance, 1642 people appear in one or more shots.

3 State-of-the-art of existing methods

Until very recently, research work relevant for the described task was dealing mainly with speaker or face recognition, with few early attempts were both face and speaker identification tasks were treated simultaneously; the following review reflect this situation.

For speaker identification, the closest modality for extracting the names of speakers was used: the pronounced names from speech transcription. We can mention the works of *Canseco et al.* pioneered approaches relying on pronounced names instead of biometric models for speaker identification [14] and [13]. They set manually linguistic patterns to determine a link between a pronounced name and a speech segment. Following works improved this idea: [55] replaced manual rules by learning sequence of n-grams with associated probabilities, [33,19] and [27] used semantic classification trees to calculate the probabilities, [36] replaced the decision by belief functions. However, due to relatively high speech transcription and named entity detection errors, all these audio-only approaches did not achieve good enough identification performance.

Written names were first used for a face identification task in broadcast news ([53,24]), but due to a high word error rate (respectively 52 % and 65 %), these names were detected and corrected with the help of a dictionary (limiting identification to a closed list of persons). Despite these corrections, the error rate was still very high (45 % after correction in [24]) and consequently greatly limited the use of this source of information. Later, *Yang et al.* [58,59] also tried to use written names, but again, the video OCR system [52] used to process the overlaid text produced highly erroneous results (e.g. "Newt Gingrich" was recognized as "nev j ginuhicij"). Their studies were limited to a single show (ABC World News Tonight) and they only tries to label faces in monologue-style speech

We can also cite *Pham et al.* [37,38], which first extracted a list of names from speech transcription. Then, they manually associated a set of face images to these names. These identities are then propagated to the entire video from the similarity between faces. They have therefore concentrated their works on the face identification and thus did not taken advantage of the video multimodality with the use of speakers as indice for the propagation. Indeed, in [29] *Khoury et al* show that a multimodal audio-visual diarization obtains better results than monomodal diarization.

Thanks to the REPERE challenge, significant progress was achieved in either supervised or unsupervised mulitmodal person recognition. We proposed an unsupervised speaker/face identification system ([46,9]) based only on written names as source of names (extracted using the tool LOOV [41]) in TV broadcast. The main idea was to build mono-modal clusters (faces or speakers) and to associate written names to these clusters based on their co-occurrences (un-supervised approach). In this former work, faces and speakers were treated separately. This system was extended in [8,45,40,48] with the modification of the agglomerative clustering process. This process integrated

directly the knowledge of written names to both identify clusters and also to prevent the merging of clusters named differently. We used this method during the presented campaign as the LIMSI system (*Poignant et al.* [43] in Section 5).

Another work that is worth being mentioned is using Integer Linear Programming (ILP) for speech clustering [7,6,10]. The main idea is to replace the classical agglomerative BIC clustering by an ILP clustering and at the same time integrating written names to identify speech clusters. First, multi-modal similarity graph is built, were intra-modal links correspond to the similarity of mono-modal elements (speech turns: BIC distance, written names: identical or not) and cross-modal links correspond to the temporal co-occurrence between written names and speech turns. As a written name has a high probability to match the current speaker, identification of speech turns via the ILP solver is equivalent to find the less expensive way to connect names and speech turns. The main limitation of this method is the large computation time for solving ILP (as well as the basic cross-modal link used).

In [20], speakers are named in the first place and then identities are propagated to visible persons. Speakers are named by the propagation of written names and pronounced names and also with biometrics speaker models. After a face diarization step, written names and speaker identities are propagated to faces based on their co-occurrence. Authors also integrate a set of manual rules for a specific show to post-process their output (e.g. *if one of the anchors is speaking and two faces are detected, then the second face is given the identity of the second anchor*). They extended these works in [1,50] with the use of automatic scene analysis (camera identification and scene classification as studio or report). This system needs additional annotations (scene type, camera position) for a specific show. Once a camera has been identified, they can deduct those who are visible on the screen (e.g., if the camera filming the anchor has been identified, they infer that the anchor is visible in screen). Finally, [4] proposed to integrate a lip activity detector to propagate speakers identities to face. Again, rules are used to propagate a name to a speaker/face.

Last but not least, *Gay et al.* [21] proposed to propagate written names onto multi-modal speaker/face clusters. First, speakers and face diarization are performed in parallel, then speaker and face clusters are grouped based on their co-occurrence. They are associated to written names with two methods. The first one relies on co-occurrence information between written names and speaker/face clusters, and rule-based decisions which assign a name to each mono-modal cluster. The second method uses a Conditional Random Field (CRF) which combines different types of co-occurrence statistics and pairwised constraints to jointly identify speakers and faces.

4 Baseline and metadata

This task targeted researchers from several communities including multimedia, computer vision, speech and natural language processing. Though the task was multimodal by design and necessitated expertise in various domains, the tech-

nological barriers to entry was lowered by the provision of a baseline system described in Figure 4 and available as open-source software[3]. For instance, a researcher from the speech processing community could focus its research efforts on improving speaker diarization and automatic speech transcription, while still being able to rely on provided face detection and tracking results to participate to the task.

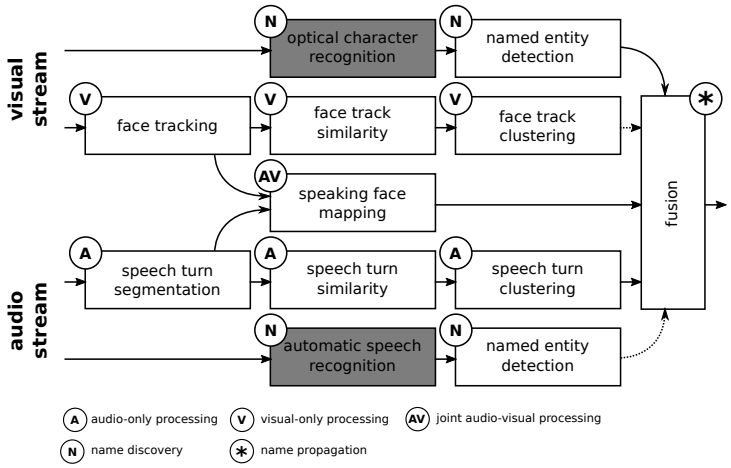

Fig. 4: Multimodal baseline pipeline. Output of all but greyed out modules is provided to the participants.

The audio stream was segmented into speech turns, while faces were detected and tracked in the visual stream. Speech turns (resp. face tracks) were then compared and clustered based on MFCC and the Bayesian Information Criterion [15] (resp. HOG [16] and Logistic Discriminant Metric Learning [23] on facial landmarks [56]). The approach proposed in [48] was also used to compute a probabilistic mapping between co-occuring faces and speech turns. Written (resp. pronounced) person names were automatically extracted from the visual stream (resp. the audio stream) using open source LOOV Optical Character Recognition [41] (resp. Automatic Speech Recognition [30,18]) followed by Named Entity detection (NE).

The fusion module (Figure 5) was a two-steps algorithm: From the written names and the speaker diarization, we used the "Direct Speech Turn Tagging" method described in [46] to identify speaker: we first tagged speech turns with co-occurring written name. Then, on the remaining unnamed speech turns, we find the one-to-one mapping that maximizes the co-occurrence duration between speaker clusters and written names (see [46] for more details). Finally, we propagate the speaker identities on the co-occurring face tracks based on the speech turns/face tracks mapping. Note that this fusion scheme did not

[3] `http://github.com/MediaEvalPersonDiscoveryTask`

use the pronounced names as input while this information was provided to participants

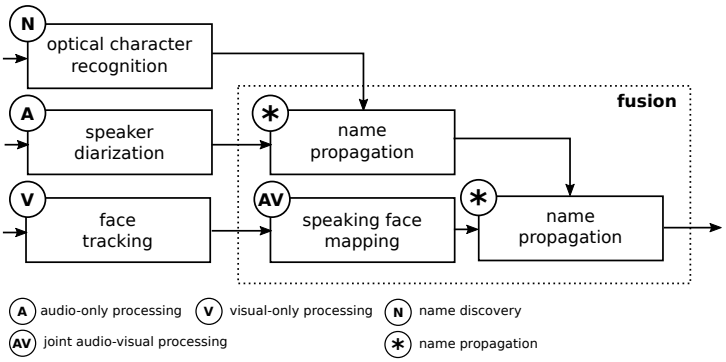

Fig. 5: Baseline fusion system overview

For each person who speaks and appears in a shot (following the shot segmentation provided to all participants), we compute a confidence score. This score is based on the temporal distance between the speaking face and its closest written name. This confidence equals to:

$$\text{confidence} = \begin{cases} 1 + d & \text{if the speaking face co-occurs} \\ & \text{with the written name} \\ 1/\delta & \text{otherwise} \end{cases}$$

where d is the co-occurrence duration and δ is the duration of the gap between the face track (or speech turn) and the written name.

5 System overview

Nine different teams [31,51,3,12,43,32,34,25] submitted a total of 34 runs. They were allowed to submit one primary run and up to four contrastive runs. Unless otherwise stated, this section focuses on primary runs.

5.1 Tested approaches

Figure 6 summarizes the three main families of approaches that were tested by participants.

The first family of approaches rely on the strong assumption that "voice-over" is not very common and that speakers' faces are usually visible. Therefore, they do not rely on any face processing module. Though they did experiment with other approaches in their contrastive runs, the primary runs of

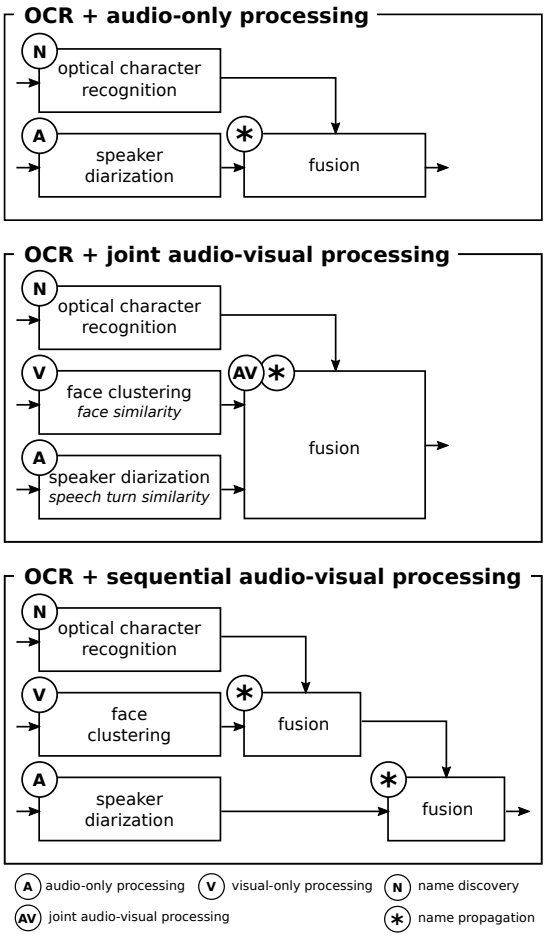

Fig. 6: Three types of systems.

the winning and both first and second runner-ups teams use this kind of approach. *Le et al.* combined their own state-of-the-art speaker diarization [49] with the simple written name propagation technique introduced in [46] (found to perform very well when applied to the output of a good speaker diarization module). *Dos Santos et al.* [51] investigated the use of late fusion of face clustering and speaker diarization but ended up choosing an approach very close to the one from *Le et al.* – leading to a difference in terms of EwMAP which is not statistically significant. *Bendris et al.* [3] tried to improve the propagation of written names onto speaker diarization clusters using the quite rigid structure of TV broadcast news. Hence, chapter segmentation [2] was used to prevent names from being propagated to speech turns from other chapters, and they relied on speaker role classification [17] to deal with anchors separately.

Primary runs by *Budnik et al.*, *Poignant et al.*, *Lopez-Otero et al.* and *Nishi et al.* and the baseline are members of the second family of approaches: joint audio-visual processing. *Budnik et al.* started from the baseline system and focused on replacing the speaking face mapping module and improving the fusion module [12]. For each pair of co-occurring speech turn and face track (*i.e.* speaking face candidates), the speech turn (resp. face track) is compared to speech turns (resp. face tracks) of all other pairs. The speaking face mapping score is then computed as the correlation between the resulting similarity vectors. Finally, contrary to what the baseline fusion does, they chose to propagate names to face first and then to mapped speech turns. Instead of performing speaker diarization and face clustering separately, *Poignant et al.* investigated joint clustering of speech turns and face tracks. Name-constrained agglomerative clustering was applied to an audio-visual similarity matrix built from the provided speech turn similarity matrix, face track similarity matrix and speaking face mapping scores [48]. *Lopez-Otero et al.* kept the baseline general architecture unchanged. Instead, they focused on the improvement of speaker diarization, face clustering and speaking face mapping modules [32]. *Nishi et al.* tried improving the speech turn similarity matrix of the baseline system with similarity scores between speaking faces. Unfortunately, it did not bring any significant improvement over the baseline [34].

Finally, the third family of approaches (including *India et al.* primary run [25]) start by performing speaker diarization and face clustering separately. Then, names obtained through optical character recognition propagated to face tracks clusters and, then only, to cooccurring speech turns (or *vice-versa*). Most participants tried this kind of approach in their contrastive runs.

Note that all participants (but one) have only used the written names as source of identity, although it was previously shown [39,26] that the pronounced names can bring additional information about persons present in the show.

5.2 Results

Table 1 reports the performance obtained by the best primary run campared to the baseline[4]. Most teams managed to outperform the baseline, thanks to the provision of the output of its constituent modules. The winning and runner-up teams (EwMAP $\approx 83\%$) even bridged half of the gap between the baseline (EwMAP $\approx 67\%$) and a perfect system (EwMAP $= 100\%$).

Table 4 (which is discussed in details in Section 7) reveals that some teams chose to focus their research on the multimodal fusion side [12,43], while others chose to improve monomodal components [31,32,34]. Furthermore, a quick

[4] These figures depart slightly from the ones reported during the MediaEval campaign: negative queries (i.e., person names without a corresponding speaking face in the test set) were initially included in the MediaEval queries but have been discarded here; also, complementary annotations were performed since the campaign.

Participants	EwMAP (%)	MAP (%)	C. (%)
Le et. al [31]	82.6	82.8	95.7
Dos Santos et al. [51]	82.6	82.8	96.1
Bendris et al. [3]	80.6	80.6	96.9
Budnik et al. [12]	77.2	77.3	94.0
Poignant et al. [43]	75.2	75.3	91.5
Lopez-Otero et al. [32]	70.0	70.1	88.8
Baseline (Section 4)	66.8	66.9	89.7

Table 1: Primary runs sorted by EwMAP (C. = correctness). Participants appear in this table on a voluntary basis.

poll among participants revealed that video quality (including encoding and resolution) was not good enough to really benefit from face recognition modules. Therefore, the best performance were obtained by the simplest *OCR + audio-only processing* approaches, ignoring face processing and speech transcripts entirely.

Except for one team, there are no significant difference between the MAP and EwMAP values: EwMAP is always only slightly smaller than MAP. This means that, when a person is correctly identified, the associated evidence is most often correct. Unfortunately, none of the participating teams provided a description of how evidences were chosen. Yet, looking at their submissions reveals that all of them (but the one team who did not provide the description of their system) relied solely on named discovered thanks to optical character recognition. Only one team (*Bendris et al.* [3]) actually used spoken names extracted from automatic speech transcripts, though very marginally to improve written name propagation scores.

5.3 Oracle experiments

Table 2 provides insight into whether systems could have been improved using spoken names or cross-show propagation (*i.e.* where names are propagated from one video to another). Based on the list of names discovered by the OCR and ASR baseline modules, it reports the best results that could have been obtained in case they had always been perfectly propagated.

Names	Mono-show propagation	Cross-show propagation
Spoken names	17.1	22.0
Written names	93.2	94.6
Both	94.8	95.7

Table 2: Optimal EwMAP with oracle propagation

The improvement brought by spoken names propagation is marginal. This is mostly due to a specificity of the provided test corpus, where people are almost systematically introduced by overlaid title blocks. Combined with the fact that propagating spoken names is much more difficult than propagating written names, this probably led participants to not even bother focusing on this research aspect.

The improvement brought by cross-show propagation is also very small. This can be explained by Figure 3 showing that most people (approximately 90%) appear in one video only. This theoretical observation might explain why no participant tried cross-show approaches.

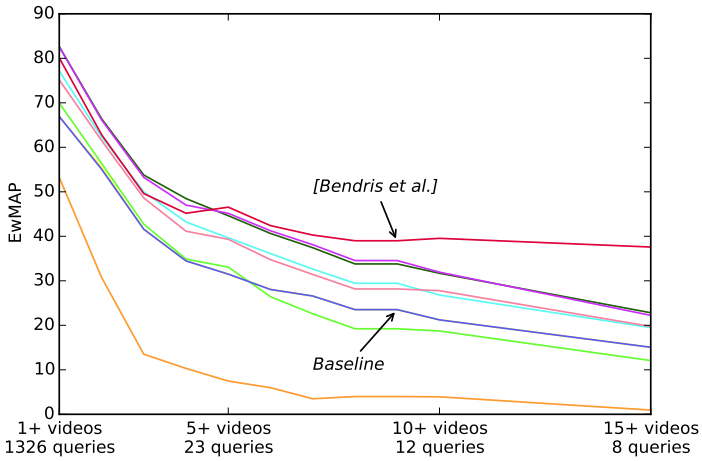

Fig. 7: Evolution of the EwMAP as the number of queries is decreased progressively from 1642 people appearing in at least 1 video, to the only 8 people appearing in more than 15 videos.

However, in practice, Figure 7 shows that all approaches tend to perform worse with queries for which relevant shots are spread over more videos (*i.e.* the very queries for which cross-show propagation could be very beneficial). It also means that the identification of recurring people is harder than people appearing only locally. *Bendris et al.*'s approach stands out from the other approaches because it includes a module dedicated to detecting anchor persons, whose relative importance increases as the number of queries decreases.

6 Error analysis

Leaving the four main anchors aside (they would probably not be the primary targets for unsupervised person identification), a total of 710 shots were missed by all participants, out of the 27k manually annotated shots.

This section aims at uncovering why person identification was more difficult in these cases. To this end, we manually annotated those 710 missed shots with attributes potentially useful for this analysis (e.g. duration of speech in the shot, presence of noise or music). Two contrastive sets of shots (where at least one participant did return the correct identity) were annotated with the same attributes:

- 569 shots carefully selected so that the same person was missed in another shot of the same video
- 630 more shots where the person appears in at least two other shots

In a nutshell, 1909 shots were manually annotated: 710 shots missed by all participants, 1199 found by at least one participant. Then, we trained decision trees to predict whether a shot would be correctly identified or not, hoping that the list of attributes selected automatically during training would help us determine the reason why some shots were incorrectly identified.

6.1 Optical character recognition

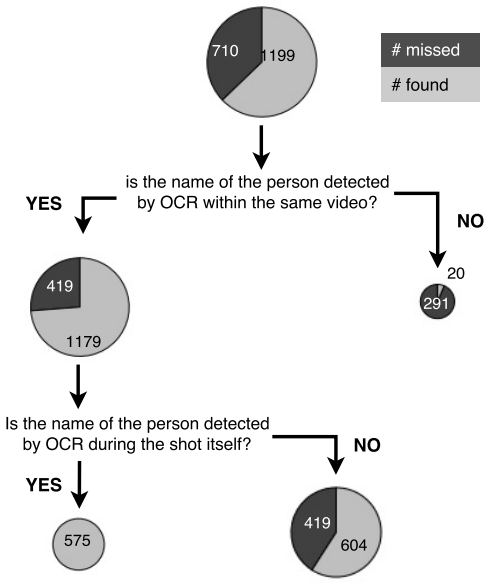

Fig. 8: Impact of OCR-based person name detection

Among all attributes that we investigated, the most significant ones are related to whether the name of the person is detected by optical character

recognition. Figure 8 summarizes our findings. In case the person name is not detected by OCR neither in the shot itself, nor in any other shot of the same video, person identification fails almost systematically (291 missed out 311 shots). When the person name is detected by OCR during the shot itself, person identification is perfect (575 shots).

In the next section, we focus on the remaining shots (419 missed and 604 correct) where the person name is not detected in the shot itself but is detected in at least one other shot of the same video. It means that propagating the person name is possible and we would like to understand why it failed for nearly half of those 1023 shots.

6.2 Variability and noise

Attributes (**boolean** or numerical)	Importance	Average on missed shots	Average on found shots
Duration of front-facing speech	28.6%	2.7s	4.7s
Is there background noise?	16.7%	32%	15%
Is the person singing?	16.0%	12%	3%
Number of detected faces	13.8%	3.1	2.9
Number of other speaking faces	8.7%	0.1	0.03
Duration of profile-facing face	7.6%	1.01s	0.58
Is the face moving?	6.3%	16%	6%
Duration of overlapped speech	2.3%	0.96	1.24

Table 3: Average value of attributes on missed and found shots, sorted by their importance in a cross-validated classification tree.

As stated at the beginning of this section, shots were manually annotated with a quite large set of attributes, some of which might be strongly correlated. Therefore, attributes were first grouped into families using DBSCAN clustering based on their correlation. The most central attribute of each family was kept, others were removed. For instance, attributes such as *duration of the shot*, *duration of speech* and *duration of front-facing person* were all grouped into the *duration of front-facing speech* family. A classification tree was then trained for discriminating missed from found shots, using those attributes as input data. Cross-validation experiments led to a classification accuracy of 67%.

Table 3 summarizes the final list of attributes ranked by their Gini importance [11] and also provides their average values in both categories (missed vs. found shots). The average value of boolean attributes (**bold type**) is simply the ratio of positive answers. For instance, the target person is actually singing in 12% of the the missed shots, and only 3% of the ones that were found.

As anticipated, the most important attributes are either related to the variability of the input signal, or the noise it contains. Hence, visual attributes such as the duration of front- or profile-facing person and the fact that the

face is moving are all related to face pose variation (which is known to be a difficult problem for face recognition). Similarly, the large importance given to the *singing person* attribute can be explained by the difficulty of overcoming acoustic variability to match a singing voice with a speaking voice during the diarization process. As far as noise is concerned, appart from the obvious acoustic background noise, spurious detected (frontal or speaking) faces can also be considered as visual noise that may lure the propagation algorithms into choosing the wrong person.

6.3 On the importance of context

z Another aspect to take into account in this study is the actual temporal context of each shot. Among shots where the same person is detected in at least one of the two previous or following shots, 78% are correctly identified. The fact that speech turns can overlap several shots might be the main explanation. Reciprocally, it appears that 79% of isolated shots are missed. It means that the tested approaches are good at propagating identities locally in neighboring shots, but are left in the dark when the temporal context does not provide meaningful information.

7 Organizational aspects

In this section, we discuss the usefulness and cost of several organizational aspects of the campaign, including the provision of a complete baseline, the availability of a live leaderboard and the amount of a posteriori collaborative annotation needed for this test dataset to compare systems.

7.1 Baseline

Module / Team	[31]	[32]	[34]	[25]	[51]	[3]	[12]	[43]
Face tracking	●	●						
Face clustering	●	●	●	●	●			
Speaker diarization	●	●	●	●	●	●		
Written names	●			●		●		
Spoken names								
Speaking face detection	●	●					●	
Fusion				●	●	●	●	●

Table 4: Was the baseline useful? Teams that relied on the provided (resp. designed their own) modules are shown with ■ (resp. ●).

As already mentioned in Section 4, the main motivation behind providing a baseline system was to lower the technological barriers to entry into a task

requiring expertise in a large range of domains (including computer vision, speech processing and natural language processing). Table 4 provides insight into whether the baseline was actually useful for participants. The answer is *"yes"* – every single team relied on at least two modules constituting the baseline, though with differing strategies:

- teams with past expertise in computer vision or speech processing chose to focus on the improvement of the monomodal components and used the baseline fusion module (mostly) as is;
- other teams kept most monomodal components unchanged and focus their research on the multimodal fusion problem.

Had the baseline not be provided, the second category of teams would (most likely) not have been able to participate to the task. The main conclusion that we can draw from this analysis is that the provided baseline really lowered the barrier to entry for the task and attracted participants who would not have participated otherwise. We should therefore keep providing a baseline system for future editions of the task.

Furthermore, notice how the written name extraction module has been used by all participants, whereas only *Bendris et al.* [3] played around with the spoken name extraction module. Two reasons explain this lack of interest for the spoken name module. The first one relates to the nature of the test dataset itself, in which most people are actually introduced by their written name – reducing the need for other sources of person names. The second reason is that only partial audio transcription was provided to the participants (five words before and after the detected names), preventing the development of advanced natural language processing techniques, and making the propagation of spoken names unreliable. In future editions, this part of the baseline should be improved (by providing the complete audio transcription, for instance) and the test corpus should be carefully selected so that the predominance of written names is reduced in favour of spoken names.

7.2 Leaderboard

Two submission deadlines were proposed to the participants, at one week interval. For the first one, participants could only tune their algorithms using the REPERE corpus as development set. This development set did not match the test set perfectly: different video quality, different channels, different type of shows. This miss-match allows to show which methods generalize well. Right after the first deadline, a leaderboard was opened, providing live feedback to the participants, until the second deadline one week later. During a full week, participants were allowed to submit up to five runs every six hours. In return, the leaderboard would provide them with scores computed on a secret subset of the test set with a subset of queries. Participants would have only access to the scores of their own runs, along with their current rank among all participants.

Figure 9 shown that four participants took advantage of the additional time and of the leaderboard in order to either detect and correct a bug in their system, or to improve it.

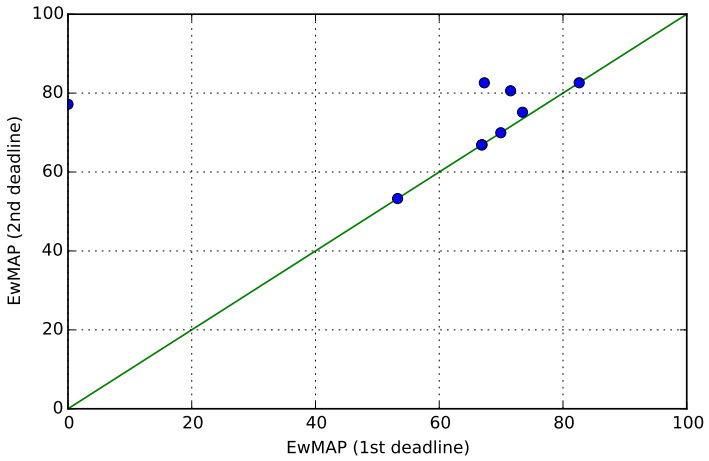

Fig. 9: What the leaderboard useful? Evolution of the systems performance between the 1st and the 2nd deadline, allowed by the leaderboard feedback.

Note that the systems performance presented in section 5.2 and elsewhere in this paper exclude the shots which were involved in this leaderboard.

7.3 Collaborative annotation

For this first campaign we chose to perform a dense annotation on the first half of the corpus, through the use of a collaborative annotation interface [47, 44]. Among the 240 hours spent on the actual annotation of the corpus, only 50 hours were spent by participants themselves (around 6 hours per team, on average). The other 190 hours were spent by people specially hired for this task (and thus paid to do so). For this task to become sustainable in the long term, we should make sure that the (free, but limited) annotations produced by participants are enough to obtain statistically significant performance comparison.

Figure 10 plots the mean of EwMAP and its standard deviation for 400 different random subsets of the complete query list, as a function of the number of selected queries. Less than 700 queries seem enough for discriminating between the performance of the submitted primary runs. Combined with lessons learned from previous research about Inferred Mean Average Precision [60], it should therefore be possible to annotate only a subset of the shots returned by

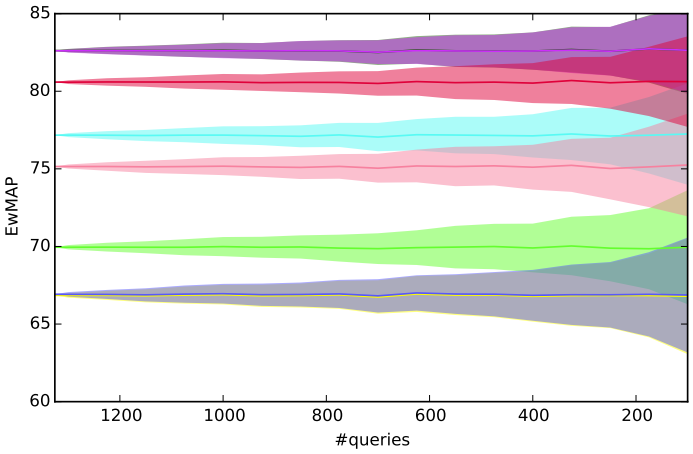

Fig. 10: EwMAP as a function of the number of queries (mean and standard deviation over 400 random query subsets).

participant for a limited set of carefully-chosen queries. This would drastically reduce the cost of annotation and make the task sustainable.

8 Conclusion and future works

The organization of a new task on multimodal person discovery at MediaEval 2015 benchmarking initiative was well received by the community, with 9 participating teams interested in the idea of unsupervised person identification. The provision of a modular baseline system lowered the entry gap for the competing teams with varying scientific backgrounds. The live leaderboard allowed to overcome the condition mismatch between the development and the test set. The evaluation workflow relied on the CAMOMILE framework with a posteriori annotation of the corpus by the organizers and the participants. Very good results could thus be achieved with 82.6% MAP for the best system. However, the analysis of the results shows that a simple strategy of propagating the written names to the speaker diarization output was sufficient, mainly due to the nature of the test corpus. This leads us to several directions for future editions of the task. First, we should improve the video quality and increase the content variety of the corpus; the respective weight of face tracking and speaker diarization would be better balanced, and the detection of pronounced names would become more useful. Also, the queries may be restricted to persons occuring in at least two shows, making the task definitively harder but also reducing the annotation need.

Acknowledgements This work was supported by the French National Agency for Research under grant ANR-12-CHRI-0006-01. The open source CAMOMILE collaborative

annotation platform was used extensively throughout the progress of the task: from the run submission script to the automated leaderboard, including a posteriori collaborative annotation of the test corpus. We thank ELDA and INA for supporting the task by distributing development and test datasets.

References

1. Bechet, F., Bendris, M., Charlet, D., Damnati, G., Favre, B., Rouvier, M., Auguste, R., Bigot, B., Dufour, R., Fredouille, C., Linarès, G., Martinet, J., Senay, G., Tirilly, P.: Multimodal Understanding for Person Recognition in Video Broadcasts. In: INTERSPEECH (2014)
2. Bechet, F., Bendris, M., Charlet, D., Damnati, G., Favre, B., Rouvier, M., Auguste, R., Bigot, B., Dufour, R., Fredouille, C., et al.: Multimodal understanding for person recognition in video broadcasts. In: INTERSPEECH, pp. 607–611 (2014)
3. Bendris, M., Charlet, D., Senay, G., Kim, M., Favre, B., Rouvier, M., Bechet, F., Damnati, G.: Percolatte : A multimodal person discovery system in tv broadcast for the medieval 2015 evaluation campaign. In: MediaEval (2015)
4. Bendris, M., Favre, B., Charlet, D., Damnati, G., Auguste, R., Martinet, J., Senay, G.: Unsupervised Face Identification in TV Content using Audio-Visual Sources. In: CBMI (2013)
5. Bernard, G., Galibert, O., Kahn., J.: The First Official REPERE Evaluation. In: SLAM-INTERSPEECH (2013)
6. Bredin, H., Laurent, A., Sarkar, A., Le, V.B., Rosset, S., Barras, C.: Person Instance Graphs for Named Speaker Identification in TV Broadcast. In: Odyssey (2014)
7. Bredin, H., Poignant, J.: Integer Linear Programming for Speaker Diarization and Cross-Modal Identification in TV Broadcast. In: INTERSPEECH (2013)
8. Bredin, H., Poignant, J., Fortier, G., Tapaswi, M., Le, V.B., Sarkar, A., Barras, C., Rosset, S., Roy, A., Yang, Q., Gao, H., Mignon, A., Verbeek, J., Besacier, L., Quénot, G., Ekenel, H.K., Stiefelhagen, R.: QCompere at REPERE 2013. In: SLAM-INTERSPEECH (2013)
9. Bredin, H., Poignant, J., Tapaswi, M., Fortier, G., Le, V.B., Napoleon, T., Gao, H., Barras, C., Rosset, S., Besacier, L., Verbeek, J., Quénot, G., Jurie, F., Ekenel, H.K.: Fusion of speech, faces and text for person identification in TV broadcast. In: ECCV-IFCVCR (2012)
10. Bredin, H., Roy, A., Le, V.B., Barras, C.: Person instance graphs for mono-, cross- and multi-modal person recognition in multimedia data: application to speaker identification in TV broadcast. In: IJMIR (2014)
11. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
12. Budnik, M., Safadi, B., Besacier, L., Quénot, G., Khodabakhsh, A., Demiroglu, C.: Lig at mediaeval 2015 multimodal person discovery in broadcast tv task. In: MediaEval (2015)
13. Canseco, L., Lamel, L., Gauvain, J.L.: A Comparative Study Using Manual and Automatic Transcriptions for Diarization. In: ASRU (2005)
14. Canseco-Rodriguez, L., Lamel, L., Gauvain, J.L.: Speaker diarization from speech transcripts. In: the 5th Annual Conference of the International Speech Communication Association, INTERSPEECH, p. (2004)
15. Chen, S., Gopalakrishnan, P.: Speaker, Environment And Channel Change Detection And Clustering Via The Bayesian Information Criterion. In: DARPA Broadcast News Trans. and Under. Workshop (1998)
16. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
17. Damnati, G., Charlet, D.: Robust speaker turn role labeling of tv broadcast news shows. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pp. 5684–5687 (2011)
18. Dinarelli, M., Rosset, S.: Models Cascade for Tree-Structured Named Entity Detection. In: IJCNLP (2011)

19. Estève, Y., Meignier, S., Deléglise, P., Mauclair, J.: Extracting true speaker identities from transcriptions. In: INTERSPEECH (2007)
20. Favre, B., Damnati, G., Béchet, F., Bendris, M., Charlet, D., Auguste, R., Ayache, S., Bigot, B., Delteil, A., Dufour, R., Fredouille, C., Linares, G., Martinet, J., Senay, G., Tirilly, P.: PERCOLI: a person identification system for the 2013 REPERE challenge. In: SLAM-INTERSPEECH (2013)
21. Gay, P., Dupuy, G., Lailler, C., Odobez, J.M., Meignier, S., Deléglise, P.: Comparison of Two Methods for Unsupervised Person Identification in TV Shows. In: CBMI (2014)
22. Giraudel, A., Carré, M., Mapelli, V., Kahn, J., Galibert, O., Quintard, L.: The REPERE Corpus : a Multimodal Corpus for Person Recognition. In: LREC (2012)
23. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Face recognition from caption-based supervision. IJCV (2012)
24. Houghton, R.: Named Faces: Putting Names to Faces. IEEE Intelligent Systems **14** (1999)
25. India, M., Varas, D., Vilaplana, V., Morros, J., Hernando, J.: Upc system for the 2015 mediaeval multimodal person discovery in broadcast tv task. In: MediaEval (2015)
26. J. Poignant and L. Besacier and G. Quénot: Nommage non-supervisé des personnes dans les émissions de télévision: utilisation des noms écrits, des noms prononcés ou des deux? In: Documents numériques (2014)
27. Jousse, V., Petit-Renaud, S., Meignier, S., Estève, Y., Jacquin, C.: Automatic named identification of speakers using diarization and ASR systems. In: the 34th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 4557–4560 (2009)
28. Kahn, J., Galibert, O., Quintard, L., Carré, M., Giraudel, A., P.Joly: A presentation of the REPERE challenge. In: CBMI (2012)
29. Khoury, E., Sénac, C., Joly, P.: Audiovisual Diarization Of People In Video Content. MTAP (2012)
30. Lamel, L., Courcinous, S., Despres, J., Gauvain, J., Josse, Y., Kilgour, K., Kraft, F., Le, V.B., Ney, H., Nussbaum-Thom, M., Oparin, I., Schlippe, T., Schlëter, R., Schultz, T., da Silva, T.F., Stüker, S., Sundermeyer, M., Vieru, B., Vu, N., Waibel, A., Woehrling, C.: Speech Recognition for Machine Translation in Quaero. In: IWSLT (2011)
31. Le, N., Wu, D., Meignier, S., Odobez, J.M.: Eumssi team at the mediaeval person discovery challenge. In: MediaEval (2015)
32. Lopez-Otero, P., Barros, R., Docio-Fernandez, L., González-Agulla, E., Alba-Castro, J., Garcia-Mateo, C.: Gtm-uvigo systems for person discovery task at mediaeval 2015. In: MediaEval (2015)
33. Mauclair, J., Meignier, S., Estève, Y.: Speaker diarization: about whom the speaker is talking? In: IEEE Odyssey 2006 - The Speaker and Language Recognition Workshop, p. (2006)
34. Nishi, F., Inoue, N., Shinoda, K.: Combining audio features and visual i-vector at mediaeval 2015 multimodal person discovery in broadcast tv. In: MediaEval (2015)
35. Over, P., Awad, G., Fiscus, J., Antonishek, B., Michel, M., Smeaton, A., Kraaij, W., Quénot, G.: Trecvid 2010–an overview of the goals, tasks, data, evaluation mechanisms, and metrics (2011)
36. Petit-Renaud, S., Jousse, V., Meignier, S., Estève, Y.: Identification of speakers by name using belief functions. In: IPMU (2010)
37. Pham, P., Moens, M.F., Tuytelaars, T.: Naming persons in news video with label propagation. In: ICME (2010)
38. Pham, P., Tuytelaars, T., Moens, M.F.: Naming people in news videos with label propagation. IEEE MultiMedia (2011)
39. Poignant, J., Besacier, L., Le, V., Rosset, S., Quénot, G.: Unsupervised naming of speakers in broadcast TV: using written names, pronounced names or both ? In: INTERSPEECH (2013)
40. Poignant, J., Besacier, L., Quénot, G.: Unsupervised Speaker Identification in TV Broadcast Based on Written Names. ASLP (2015)
41. Poignant, J., Besacier, L., Quénot, G., Thollard, F.: From text detection in videos to person identification. In: ICME (2012)
42. Poignant, J., Besacier, L., Quénot, G.: Nommage non-supervisé des personnes dans les émissions de télévision: une revue du potentiel de chaque modalité. In: CORIA (2013)

43. Poignant, J., Bredin, H., Barras, C.: Limsi at mediaeval 2015: Person discovery in broadcast tv task. In: MediaEval (2015)
44. Poignant, J., Bredin, H., Barras, C., Stefas, M., Bruneau, P., Tamisier, T.: Benchmarking multimedia technologies with the CAMOMILE platform: the case of Multimodal Person Discovery at MediaEval 2015. In: LREC (2016)
45. Poignant, J., Bredin, H., Besacier, L., Quénot, G., Barras, C.: Towards a better integration of written names for unsupervised speakers identification in videos. In: SLAM-INTERSPEECH (2013)
46. Poignant, J., Bredin, H., Le, V., Besacier, L., Barras, C., Quénot, G.: Unsupervised speaker identification using overlaid texts in TV broadcast. In: INTERSPEECH (2012)
47. Poignant, J., Budnik, M., Bredin, H., Barras, C., Stefas, M., Bruneau, P., Adda, G., Besacier, L., Ekenel, H., Francopoulo, G., Hernando, J., Mariani, J., Morros, R., Quénot, G., Rosset, S., Tamisier, T.: The CAMOMILE Collaborative Annotation Platform for Multi-modal, Multi-lingual and Multi-media Documents. In: LREC (2016)
48. Poignant, J., Fortier, G., Besacier, L., Quénot, G.: Naming multi-modal clusters to identify persons in TV broadcast. MTAP (2015)
49. Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., Meignier, S.: An open-source state-of-the-art toolbox for broadcast news diarization. In: INTERSPEECH (2013)
50. Rouvier, M., Favre, B., Bendris, M., Charlet, D., Damnati, G.: Scene understanding for identifying persons in TV shows: beyond face authentication. In: CBMI (2014)
51. dos Santos Jr, C.E., Gravier, G., Schwartz, W.: Ssig and irisa at multimodal person discovery. In: MediaEval (2015)
52. Sato, T., Kanade, T., Hughes, T., Smith, M., S.Satoh: Video ocr: Indexing digital news libraries by recognition of superimposed caption. In: ACM Multimedia Systems (1999)
53. Satoh, S., Nakamura, Y., Kanade, T.: Name-It: Naming and Detecting Faces in News Videos. IEEE Multimedia **6** (1999)
54. Smeaton, A., Over, P., Kraaij, W.: High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In: Multimedia Content Analysis, Theory and Applications, pp. 151–174 (2009)
55. Tranter, S.E.: Who really spoke when? Finding speaker turns and identities in broadcast news audio. In: ICASSP (2006)
56. Uřičář, M., Franc, V., Hlaváč, V.: Detector of facial landmarks learned by the structured output SVM. In: VISAPP (2012)
57. Vallet, F., Essid, S., Carrive, J.: A multimodal approach to speaker diarization on tv talk-shows. IEEE Transactions on Multimedia **15**(3), 509–520 (2013)
58. Yang, J., Hauptmann, A.: Naming every individual in news video monologues. In: ACM Multimedia (2004)
59. Yang, J., Yan, R., Hauptmann, A.: Multiple instance learning for labeling faces in broadcasting news video. In: ACM Multimedia (2005)
60. Yilmaz, E., Aslam, J.: Estimating Average Precision with Incomplete and Imperfect Judgments. In: 15th ACM International Conference on Information and Knowledge Management (2006)