Unsupervised Speaker Identification in TV Broadcast Based on Written Names

Johann Poignant, Laurent Besacier, Georges Quénot

Unsupervised Speaker Identification in TV Broadcast Based on Written Names

Johann Poignant[1,2], Laurent Besacier[1,2] and Georges Quénot[2,1]

[1] Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
[2] CNRS, LIG, F-38000 Grenoble, France

***Abstract*—Identifying speakers in TV broadcast in an unsupervised way (i.e. without biometric models) is a solution for avoiding costly annotations. Existing methods usually use pronounced names, as a source of names, for identifying speech clusters provided by a diarization step but this source is too imprecise for having sufficient confidence. To overcome this issue, another source of names can be used: the names written in a title block in the image track.**

We first compared these two sources of names on their abilities to provide the name of the speakers in TV broadcast. This study shows that it is more interesting to use written names for their high precision for identifying the current speaker.

We also propose two approaches for finding speaker identity based only on names written in the image track. With the "late naming" approach, we propose different propagations of written names onto clusters. Our second proposition, "Early naming", modifies the speaker diarization module (agglomerative clustering) by adding constraints preventing two clusters with different associated written names to be merged together.

These methods were tested on the REPERE corpus phase 1, containing 3 hours of annotated videos. Our best "late naming" system reaches an F-measure of 73.1%. "early naming" improves over this result both in terms of identification error rate and of stability of the clustering stopping criterion. By comparison, a mono-modal, supervised speaker identification system with 535 speaker models trained on matching development data and additional TV and radio data only provided a 57.2% F-measure.

***Index Terms*—Speaker identification, speaker diarization, written names, multimodal fusion, TV broadcast.**

I. Introduction

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

A solution to these problems is to use other information sources for naming speakers in a video. This is called unsupervised naming of speakers and most approaches for that can be decomposed into the three following steps:

1) Segmentation of speech into clusters, a cluster must correspond to one person and vice-versa (diarization);
2) Extraction of hypothesis names from the video (or from the collection of videos);
3) Mapping (or association) between hypothesis names and clusters.

Speaker diarization [1] is the process of partitioning the audio stream into homogeneous clusters without prior knowledge on the speakers' voice. Each cluster must correspond to only one speaker and *vice versa*. Most systems use a bottom-up agglomerative approach which tries to merge speech turns into clusters that are as pure as possible using a distance metric (with a distance-based criterion for stopping the clustering).

Two modalities, intrinsic to the video, (see figure 1) can provide the name of speakers in broadcast TV: pronounced names and names written on the screen to introduce the corresponding person (i.e. written names).

The third step depends on the name source used. We will see in the next section that most state-of-the-art approaches rely on pronounced names because of the poor quality of written names transcription observed in the past.

Objectives of this paper are to show that written names over the image in TV broadcast can provide the name of speakers for a cross-modal identification. The main idea is to directly name some speech turns and then propagate these identities through the diarization process.

The outline is as follows: section II presents the related works and shows how previous studies have integrated pronounced names and/or written names to identify speakers in radio and TV broadcast. Then, in section III, we describe the experimental setup: REPERE corpus, audio and video processing modules, and a comparison between written names and pronounced names to provide speaker identities. Section IV is dedicated to our different unsupervised speaker naming methods based on written names only. Section VI compares our propositions with a state-of-the-art method and with two methods based on biometric models in the framework of the REPERE challenge. Finally, we conclude this paper and give some perspectives.

II. State of the art

A. Naming Speakers with Pronounced Names

Until very recently, the state-of-the-art works on unsupervised speaker identification used the closest modality for extracting the names of speakers: the pronounced names from speech transcription. The first works were proposed by *Canseco et al.* [2], [3]. They used linguistic patterns set manually to determine which referred to a pronounced name: the current, next or previous speaker (e.g. "thank you Candy Crowley" inferred that the name corresponds to the previous speaker). *Charhad et al.* [4] used the same method but with automatic diarization and automatic speech recognition [5].

In 2006, *Tranter* [6] replaced manual rules by learning sequence of n-grams with associated probabilities. She showed

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

that less speakers are nameable with automatic systems (47.3 %) than with manual annotations (76.8 %) on the Hub-4 corpus. In [7], *Ma et al.* extended [6] by using a maximum entropy model. The naming strategy was enriched with features such as the name position in the sentence and genre correspondence between names and speakers.

Mauclair et al. in [8] used semantic classification trees (SCT) to calculate the probabilities (according to the terms around a pronounced name) that a name correspond to the previous, current, next speaker or another speaker. This method has been tested on the ESTER corpus with manual transcripts and manual diarization. 70% (approximately) of the total broadcasting time was correctly identified (18% error and 12% unidentified). *Estève et al.* [9] show that SCT are less sensitive to the use of automatic speech recognition than rules based on n-grams. *Jousse et al.* [10] improved the use of SCT with a local decision (affiliation of pronounced names to adjacent speech turns) and a global decision (propagation of pronounced names to speech clusters). They also observed an increase of the identification error rate duration from 16.66% (manual annotations) to 75.15% (full automatic systems) on the ESTER corpus phase 2, (213 speakers).

In 2010, *Petitrenaud et al.* in [11], used the same context as *Jousse et al.*. However, they replaced the decision by belief functions. These functions have the characteristic to take into account the consistency of the information between adjacent speech turns. The authors observed that the identification error rate was reduced from 16.6% to 13.7% with manually annotated data. *El-Khoury et al.* [12] applied these belief functions with the use of automatic systems. The error rate increased from 10%, based on manual annotation, to 41.1% with the use of automatic system. They also integrated scores transformed into belief function from a speaker recognition system based on biometric GMM-UBM models. This reduces the error to respectively 4.6% and 32.7%. Given that the error rate for biometric models alone was 63%. This work shows that biometric models and unsupervised naming systems tend to identify different speakers.

A recent study on the capability of pronounced names to identify persons present in a video was proposed by *Béchet et al.* [13]. They asked two human judges to choose for all pronounced names in the corpus *REPERE*, phase 0, whether they were present, absent or whether it was not possible to determine their presence using only the spoken transcription of the turns (containing people name). Only 43.4% of name occurrences corresponded to a speaker name. For 51.2% of them the judges couldn't determine their presence (24.1% were absent, 7.3% spoke, 19.8% were only visible). This means that only 37.1% of name occurrences provided a relevant information. This work shows that the names pronounced, in addition to being hard to extract, provide information for which we cannot have a strong confidence.

B. Naming persons with written names

Written names were first used for a face identification task in broadcast news ([14], [15]), 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 [15]) and consequently greatly limited the use of this source of information. Later, *Yang et al.* [16], [17] also tried to use this source of names, but again, the video OCR system [18] used to process the overlaid text produced highly erroneous results (e.g. "Newt Gingrich" was recognized as "nev j ginuhicij").

Improved video quality in recent years allows us to extract written names on screen (used to introduce the corresponding person) with very few errors. We were the first to propose a speaker identification system based only on this source of names (extracted using the tool LOOV [19]) in TV broadcast [20], [21], [22], [23]. Our methods are described and extended in this paper.

We also collaborated with the LIMSI laboratory for integrating identities provided by written names in an Integer Linear Programming (ILP) speech clustering [24]. The main idea was 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 was build, were intra-modal links correspond to the similarity of mono-modal elements (speech turns: BIC distance, written names: identical or not) and multi-modal links 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 corresponded 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 the ILP clustering. A comparison of identification results obtained with this method and with the methods that we propose in this paper will be given later in section VI.

In [26], [27], we also proposed a fair comparison between written names and pronounced names to identify speakers. We summarize the main results in the section III-C.

C. Lessons learned

Two types of names that can be automatically extracted were used in the literature: pronounced names and written names. The automatic extraction of names from these two sources generates several types of error and, in this context, the quality of speech transcription and speaker names detection is an important issue. For overlaid texts on the image, improvement of the video quality these recent years reduces errors as we will show in section III-B2. In addition, it should be noted that the character models (which can be multilingual) for the transcription of overlaid texts are more generic than language models used for speech transcription.

There are also errors coming from the detection of the names in these transcriptions. The detection of named entities in speech is not an easy task when no a priori knowledge of names to detect is available whereas the detection of written names corresponds to the simple detection of the template used in the show for writing text on the screen, this task is much simpler.

Both sources provide temporal information of the possible intervention of a speaker. But it remains unclear for pronounced names (a name can correspond to the previous, current, next speaker or to another) while a written name correspond to the current speaker in 95% of the cases in the corpus used for our experiments (see next section).

Association methods between pronounced names and clusters take into account the uncertainty of this source of names. So far for written names, there are few association methods proposed due to the difficulty of extracting them. But as the quality of the videos increased significantly in recent years, we can now extract them with very few errors. We therefore propose, in section IV, different association methods using written names to identify speakers in TV broadcast.

III. Experimental setup and preliminary study

The REPERE challenge [28] is an evaluation campaign on multimodal person recognition (phase 1 took place in January 2013 and phase 2 in January 2014). The main objective is to answer the two following questions at any instant of the video: *"who is speaking?" "who is seen?"*. All modalities available (audio, image, external data ...) can be used for answering these questions. In this paper, we try to answer the first question without using speaker biometric models.

A. REPERE corpus [29]

The dataset used in our experiments is composed of videos recorded from seven different types of shows (including news and talk shows) broadcasted from two French TV channels. Data is split between 3 sets (training, development and test). An overview of the data is presented in Table I.

Though raw videos were provided to the participants (including the whole show, adverts and part of surrounding shows), only excerpts of the target shows were manually annotated for the evaluation. Our evaluations are performed on the test set. It is important to note that, although the whole test set is processed, the performance is measured only on the annotated frames on it.

	Train	Dev.	Test
Raw video	58h	13h	15h
Annotated part	24h	3h	3h
# speech turns in the annotated part	19208	2010	2220
# named spk in the annotated part	555	122	126
Number of annotated frames	8766	1108	1229

TABLE I: Train and test sets repartition of the *REPERE* corpus phase 1

B. Audio and video processing modules

1) Speaker diarization:

Speaker diarization consists in segmenting the audio stream into speech turns and tagging each turn with a label specific of the speaker. Given that no a priori knowledge of the speaker's voice is available in the unsupervised condition, only anonymous speaker labels can be provided at this stage.

After splitting the signal into acoustically homogeneous segments, we calculate a similarity score matrix between each pair of segments using the BIC criterion [30]. Segments are modeled with one Gaussian with full covariance matrix $\sum$ trained on the D = 12-dimensional Mel Frequency Cepstral Coefficients (MFCC) and energy. ΔBIC_{ij} defines the similarity between two segments i and j:

$$\Delta\text{BIC}_{ij} = (n_i + n_j)\log|\Sigma| - n_i \log|\Sigma_i| - n_j \log|\Sigma_j| - \frac{1}{2}\cdot\lambda\cdot\left(D + \frac{1}{2}D\,(D+1)\right)\log(n_i + n_j)$$

where n_k is the number of samples in segment k and λ the penalty weighting coefficient. A similarity matrix between all segments is then given as input of a complete-link agglomerative clustering. Depending on the similarity threshold used as stopping criterion, several clustering results can be obtained.

It is worth mentioning that the matrix is not updated after each merging of clusters, as this is usually the case for regular BIC clustering. We are aware that hierarchical clustering based on BIC distance is less efficient than hierarchical clustering with CLR distance [31] or *I-vector*+ILP [32] but our goal, here, is to do a fair comparison of several speaker naming methods, independently of the similarity measure used for clustering.

2) Written names (WN):

In order to detect written names used for introducing a person, a detection and transcription system is needed. For this task we used LOOV [19] (LIG Overlaid OCR in Video). This system has been previously evaluated on another broadcast news corpus with low-resolution videos (352×288). We obtained a character error rate (CER) of 4.6% for any type of text and of 2.6% for names written in a title block.

From the transcriptions, we use a simple technique in order to detect the title blocks spatial positions. This technique compares each transcript with a list of famous names (175k names extracted from Wikipedia). Whenever a transcription corresponds to a famous name, we add its spatial position to a list. With the repeating positions in this list we find the spatial positions of title blocks used for introducing a person. However, these text boxes detected do not always contain a name. A simple filtering based on some text questions (does

the first word correspond to a first name? is the sentence longer than a threshold? ...) allows us to filter false positives (4779 text boxes candidates, 1315 after filtering, 11 should not have been filtered, 13 should have). Transcription errors are corrected using the Wikipedia list when the edit distance is small (207 corrections with 4 errors only).

3) Pronounced names (PN):

A state-of-the-art off-the-shelf Speech-To-Text system for French [33] was used for transcribing the audio data without specific model adaptation to our corpus. The recognizer uses the same basic statistical modeling techniques and decoding strategy as in the LIMSI English BN system [5]. Prior to transcription, segmentation and clustering [34] are performed based on acoustic features. Word decoding is carried out in a 1×RT single decoding pass. Each decoding pass produces a word lattice with cross-word, word-position dependent acoustic models, followed by consensus decoding with a 4-gram language model and pronunciation probabilities (35-phone set, 65k word vocabulary). This system obtained a word error rate of 16.87% (on around 36k words) during the first evaluation campaign of the REPERE challenge.

For named-entity detection, specific independent Conditional Random Field (CRF) models were trained on other data from Quaero project. These models used the same features as those presented in [35]: (1) Standard features like word prefixes and suffixes. (2) Morpho-syntactic features extracted as in [36]. (3) Features extracted from a multilevel analyzer used in the LIMSI question-answering systems [37].

C. Proportion of nameable speakers by written names (WN) and pronounced names (PN)

To analyze the capability of our two sources of names to provide the real speaker names, we first compare their intrinsic qualities, and then confront their abilities for providing the name of persons that speak in the *REPERE* corpus, phase 1. For the experiments, we use the whole training part to extract more significant statistics.

With LOOV piped with our written names detection technique, we obtain 97.7% of names (see Table II), with a precision of 95.7%. The few remaining errors are due to transcription or filtering errors. Extracting pronounced names generates more errors. The main difficulty lies in the transcription and detection of unknown names (we do not have any a priori knowledge of names that could be pronounced).

Modalities	Precision	Recall	F1-measure
WN	95,7%	97,7%	96,7%
PN	73,5%	50%	59,5%

TABLE II: Quality of written (WN) and pronounced names (PN) extraction

Despite the lower precision and recall of the PN relative to WN, the automatic system extracts more hypothesis names from speech (see Table III). We can observe that there are about twice more pronounced names compared to written names, whether we analyze raw videos or annotated part only. This is observed for the number of names occurrences and the number of different unique names.

Modalities	Segment	#Occurrences of names	#Persons w/o duplicates
WN	Ann. part (24h)	1407	458
	Raw (58h)	2090	629
PN	Ann. part (24h)	2905	736
	Raw (58h)	4922	1156

TABLE III: Number of written (WN) and pronounced names (PN)

To compare these two modalities, we also count the number of nameable speakers for each video. We first count the proportion of videos where a person p is nameable:

$$\%Nameable_p = \frac{\text{\#videos where } p \in Phr}{\text{\#videos where } p \in Pr}$$

With:
p : a person
Pr : a set of persons p speaking
Phr : Pr with their names written/pronounced

The ($\%Nameable_p$) of a person p is the ratio of the number of videos where the name of p is written/pronounced and where p speaks by the number of videos where p speaks. Overall, for all persons we calculate:

$$\%Nameable = \frac{\sum_{p \in Pr} \%Nameable_p}{\#p \in Pr}$$

The third column of the table IV shows the percentage of nameable speakers depending on the modalities used as source of names.

PN	WN	$\%Nameable$
$Manu$	-	62.2
-	$Manu$	60.5
$Manu$	$Manu$	80.4
$Auto$	-	26.7
-	$Auto$	73.5
$Auto$	$Auto$	75.8

TABLE IV: Percentage of nameable speakers with PN (pronounced names) and/or WN (written names) modalities. $Manu$: manual annotations, $Auto$: automatic annotations,

We observe that the written names extracted automatically can name 73.5% of the 555 speakers. The manual annotation of WN is not complete (1 image / 10 sec only), which explains the higher score of the automatic system (73.5%) compared to manual annotations (60.5%). The combined use of the two modalities (WN+PN) enhances the score (+19.9 % in the case of manual annotations - $Manu$ but lower improvement when automated systems are used (+2.3 % for $Auto$)).

This study can be retrieved integrally in [27] (in english) or [26] (in french). We can conclude that when the written names are available, it is more interesting to try to name speakers with the help of written names. Pronounced names show a potential with manual annotation but speech transcription and named-entities errors reduce this potential for naming speakers.

In the next section, we propose two naming strategies for speakers identification in TV broadcast where the knowledge of written names is integrated at different steps of the naming process.

IV. Unsupervised naming of speakers

Our different approaches are based on the strong assumption that when a speech turn/cluster and one (or more) written name(s) occur simultaneously, the probability that the latter corresponds to the former is very high (>95%). The main idea of our propositions is to:

1) Identify speech turns/clusters co-occurring with overlaid names.
2) Propagate these identities to the remaining speech turns/clusters.

We propose two different approaches for unsupervised speakers naming with written names. The first, "late naming", tries to identify clusters provided by a diarization process. This approach proposes different "questioning" of the choice made during the diarization. The second one, "early naming", integrates the knowledge provided by written names during the diarization process itself.

A. Late naming

In the late naming approach, speaker diarization and written names extraction are run independently from each other and association is performed later as shown in Figure 2.

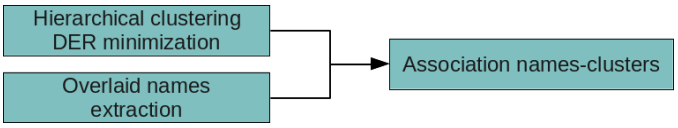

Fig. 2: Late naming approach

Speaker diarization is tuned for achieving the best diarization performance (i.e. minimize the diarization error rate, DER). Written names are extracted as described in the section III-B2. The objective of the "association names-clusters" step is to find the optimal mapping function m defined as:

$$m\colon \mathcal{T} \to \mathcal{N}$$
$$t \mapsto \begin{cases} n & \text{if name of speech turn } t \text{ is } n \in \mathcal{N} \\ \perp & \text{if it is unknown or not in } \mathcal{N} \end{cases}$$

Where $\mathcal{T} = \{t_1, \ldots, t_M\}$ the set of speech turns. $\mathcal{N} = \{n_1, \ldots, n_I\}$ is the list of I names detected by the video OCR.

Figure 3 illustrates an example that will be referred in the remainder of the section. $\mathcal{S} = \{s_1, \ldots, s_L\}$ corresponds to the set of L clusters found by the speaker diarization system.

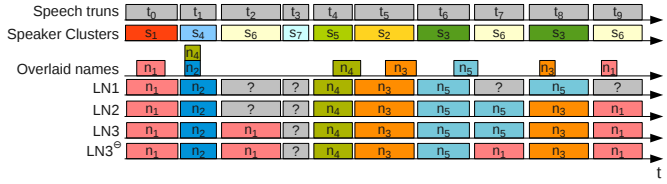

Fig. 3: Example of a timeline and the resulting name propagation for our different late naming methods

Figure 4 may be another representation of the co-occurrence links between written names, cluster and speech turns:

This two figures also illustrates the resulting name propagation of our different late naming methods.

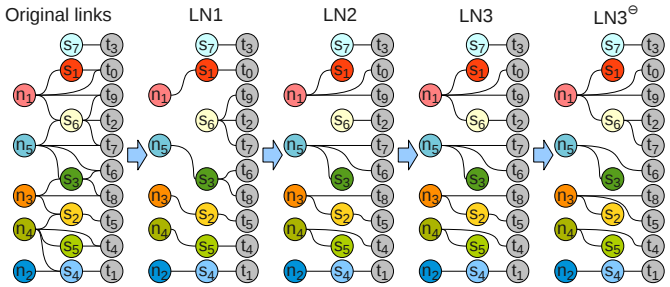

Fig. 4: Graph with original links and the resulting name propagation for our different late naming methods

1) One-to-one speaker tagging (`LN1`):

This first method (denoted `LN1` thereafter) relies on the strong assumption that speaker diarization provides perfect clusters. Therefore, `LN1` consists in finding the one-to-one mapping $f\colon \mathcal{S} \to \mathcal{N} \cup \perp$ that maximizes the co-occurrence duration between clusters and the names provided by the video OCR component:

$$f = \underset{f}{\operatorname{argmax}} \sum_{s \in \mathcal{S}} \mathbb{K}\left(s, f(s)\right)$$

where $\mathbb{K}(s, n)$ is the total duration of segments where speaker s talks and name n appears simultaneously. $f(s) = \perp$ means the name of speaker s remains unknown and $\mathbb{K}(s, \perp) = 0$. The so-called Hungarian algorithm (also known as Munkres assignment algorithm) is used for solving this problem in polynomial time [38].

In our running example (figure 3 and 4) $n_1 \mapsto s_1$, $n_2 \mapsto s_4$, $n_4 \mapsto s_5$, $n_3 \mapsto s_2$, $n_5 \mapsto s_3$. Clusters s_6 and s_7 remain unknown.

2) Direct speech turn tagging (`LN2`):

The second approach (denoted `LN2`) is based on the observation that, when one name n written alone on screen is detected, any co-occurring speech turn is very likely (95% precision on the train set) to be uttered by this person n. Therefore, our second approach is performed in two steps:

- Speech turns with exactly one co-occurring name n are tagged with the latter.
- The previous method `LN1` is applied on the remaining unnamed speech turns.

The first step of this method can question the choice made during the diarization process: it can name speech turns from different clusters with the same name or name a speech turn with a name different than the name of its cluster. As a result, in our example, speech turn t_8 is renamed from n_5 (with method `LN1`) to n_3. We also named speech turn t_7 and t_9.

3) One-to-many speaker tagging (`LN3`):

Our third proposed approach (denoted `LN3`) no longer blindly trusts the speaker diarization system. In particular, it assumes that it may produce over-segmented speaker clusters, i.e. split speech turns from one speaker into two or more clusters. This is likely to be the case for clusters s_1 and s_6 in our example. Therefore, this approach allows the propagation of a written name to two or more clusters.

First, direct speech turn tagging is applied similarly to method LN2. Then, each remaining unnamed speech turn is tagged cluster-wise using the following criterion:

$$f(s) = \underset{n \in \mathcal{N}}{\operatorname{argmax}} \, \mathrm{TF}\,(s,n) \cdot \mathrm{IDF}\,(n)$$

where the *Term-Frequency Inverse Document Frequency* (TF-IDF) coefficient – made popular by the information retrieval research community – is adapted to our problem as follows:

$$\mathrm{TF}\,(s,n) = \frac{\text{duration of name } n \text{ in cluster } s}{\text{total duration of all names in cluster } s}$$

$$\mathrm{IDF}\,(n) = \frac{\text{\# clusters}}{\text{\# clusters co-occurring with } n}$$

where clusters are analogous to textual documents, whose words are detected written names. The IDF part has a very little influence. It plays a role only if two names are associated to the same cluster with the same TF score and if one of them is associated to another cluster. The IDF part will influence the global score to name the clusters with the least used name.

Figure 4 shows how clusters s_1 and s_6 can be correctly merged using this approach.

4) Temporal re-alignment between modalities (LN3$^{\ominus}$):

In figure 5 we find the timeline of our running example. We can see that there is a bad alignment between information from the audio and the image, there is actually a good chance that the name n_5 refers to t_6 only and not to t_7.

Two reasons can explain this bad alignment, firstly, the use of video decoders using different decoding strategies can produce a time lag between these two sources of information. Secondly, segmentation of written names does not always match speech segmentation. For example, when a person interrupts another, the name of the first person may not have disappeared.

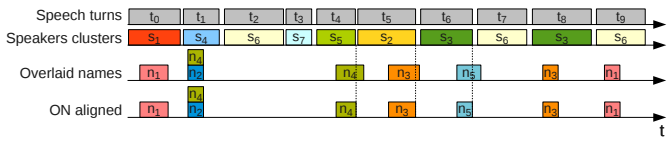

Fig. 5: Temporal re-alignment of each written names to the more co-occurring speech turn

To avoid name propagation on a bad cluster, we reduced the temporal scope of written names to the more co-occurring speech turn and apply the method LN3 with this new segmentation, we denoted this method LN3$^{\ominus}$.

In our example, the segmentation of the three names n_3, n_4 and n_5 co-occurring speech turns t_4, t_5, t_6 have been reduced. Now, n_4 et n_5 do not co-occur respectively with the speech turn t_5 and t_7. We can see, in figure 3 that the speech turn t_7 is not named directly by n_5.

5) Integrate naming (IN):

One limitation of the late naming method is that the threshold used to stop agglomerative clustering is optimized in terms of diarization error rate (DER), while the ultimate objective is speaker identification, not diarization. Obviously, optimizing DER does not necessarily lead to the lower identification error rate (EGER). Therefore, "integrated naming" (denoted IN) is a simple extension of late naming where the stopping criterion threshold is tuned in order to minimize the EGER. The main idea is to take advantage of the multiple writing of a name: we can use each occurrence of a same name to name different clusters. We choose LN3$^{\ominus}$ since this method can name different clusters with different occurrences of the same name, which allows us to stop the agglomerative clustering earlier (clusters are smaller but purer).

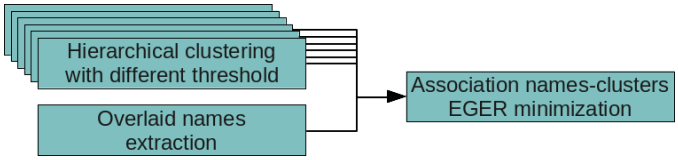

Fig. 6: Integrated naming

In practice, as shown in Figure 6, we keep multiple clustering outputs depending on the stopping criterion threshold, on which we apply the LN3$^{\ominus}$ methods, since this is the method that best suits an imperfect clustering. The threshold optimizing the EGER on the training set is chosen to stop the process on the test set.

B. Early naming

As already stated, when at least one name is written on the screen, there is a very high probability that the name of the current speaker corresponds to one of the name written on screen. Therefore, in "early naming", we use the information provided by written names during the clustering process to name clusters and also to constrain the clustering process (by forbidding the fusion of two clusters with different associated names).

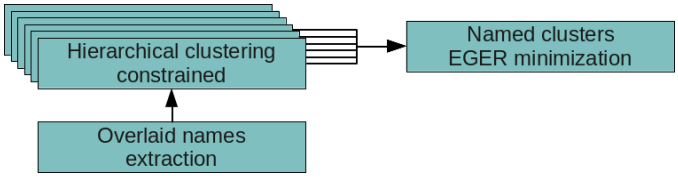

Fig. 7: Early naming

The main idea is that before clustering, we associate each written name n to the more co-occurring speech turns. At this stage, a speech turn can have several names if several names are written on the screen at the same time. Then, regular agglomerative clustering (based on speech turn similarity) is performed with the constraint that merging two clusters s without at least one name n in common is forbidden.

We can divide this process into four steps:

- **Initialization of the clustering**: prior to the clustering into clusters, we create links between the two modalities.
- **Constraints on the clustering**: during the hierarchical clustering based on a similarity matrix of speech turns, we prevent some cluster merging to avoid the propagation of names on clusters already named.
- **Update after each merge**: merging two clusters can change the link between written names and clusters. We

should recalculate the similarity scores between the new cluster (created by the merging) and all other clusters.
- **Final association between written names and clusters**: the final step chooses the best names-clusters association.

Initialization of the clustering

We first define the set of names $\mathcal{N}$ and the set of name occurrences $\mathcal{O}$:

$$\begin{array}{rcl} \mathcal{N} & = & \{a, b, \ldots, n\} \\ \mathcal{O} & = & \{o_i\} \end{array} \quad (1)$$

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

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

We also define the set of speech turns $\mathcal{T}$:

$$\mathcal{T} = \{t_1, t_2, \ldots, t_M\} \quad (3)$$

Clustering will merge speech turns into clusters, we define the set $\mathcal{G}$ of clusters. A cluster corresponding to a subset of $\mathcal{T}$. Before the clustering, there is only one speech turn per cluster. Therefore, initially $\mathcal{G}$ is the set of singletons of $\mathcal{T}$:

$$\mathcal{G} = \{\{t\}, t \in \mathcal{T}\} \quad (4)$$

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

$$f(g) = \{o \in \mathcal{O} \mid o \text{ co-occur with } g\} \quad (5)$$

With g a cluster of $\mathcal{G}$. This function allows us to divide the set $\mathcal{G}$ into two subsets:

$$\begin{array}{rcl} \mathcal{K} & = & \{g \in \mathcal{G} \mid f(g_i) \neq \{\varnothing\}\} \\ \mathcal{U} & = & \mathcal{G} \backslash \mathcal{K} \end{array} \quad (6)$$

$\mathcal{K}$ correspond to the set of clusters associated at least to a written name and $\mathcal{U}$ correspond to the set of unnamed clusters. It is important to note that, for each element of $\mathcal{O}$, a written name is only related to the more co-occurring speech turn. Therefore, every element of $\mathcal{O}$ corresponds to only one cluster when a cluster may correspond to several elements of $\mathcal{O}$.

From now, with links established between these two modalities, we can perform hierarchical clustering of elements of the set $\mathcal{G}$ with a similarity matrix between each speech turn. The aim of this clustering is to find the equivalence classes that minimize the identification error rate, but also have associated to each cluster a unique name; this goal is formalised as:

$$card(h(f(g))) = 1 \quad (7)$$

The cardinality ($card$) gives the number of different names associated to a cluster.

Constraints on the clustering

We use the relationship between clusters and written names to constrain this agglomeration. Thus, two clusters g_1 and g_2 of $\mathcal{K}$ (i.e. already named clusters) can be merged if and only if:

$$h(f(g1)) \cap h(f(g2)) \neq \varnothing \quad (8)$$

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

Update after each merge

After each agglomeration step, the merging of two clusters g_1 and g_2 in a cluster g_{12} changes the function f. Three scenario can occur for this function:

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

$$f(g_{12}) = \{o_1 \in f(g_1), o_2 \in f(g_2) \mid h(o_1) = h(o_2)\} \quad (9)$$

- Only g_1 (respectively g_2) is in $\mathcal{K}$, then:

$$f(g_{12}) = f(g_1) \text{ (respectively } f(g_{12}) = f(g_2)) \quad (10)$$

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

After each merging, we must recalculate the similarity score between the new cluster g_{12} and all other clusters g of $\mathcal{G}$. This new score is the average similarity score (ΔBIC) between elements of each cluster:

$$score(g_{12}, g) = \frac{\sum_{t_1 \in g_{12}, t_2 \in g} score(t_1, t_2)}{card(g_{12}) * card(g)} \quad (11)$$

Illustration with a toy example

Consider an example (see figure 8) with $\mathcal{K} = \{g_1, g_2, g_3, g_4\}$ and $\mathcal{U} = \{g_5, g_6\}$. 3 names are displayed $\mathcal{N} = \{a, b, c\}$ with $h(a_1) = h(a_2) = h(a_3) = a$, $h(b_1) = h(b_2) = b$ et $h(c_1) = c$.

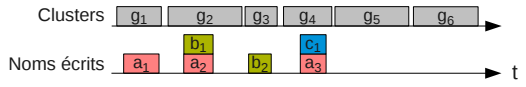

Fig. 8: Example of a timeline

Another representation is given in figure 9 with the two functions f and h:

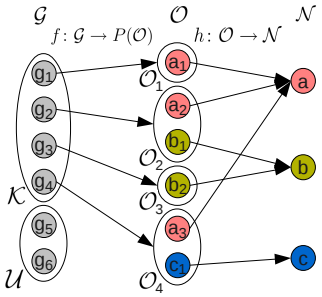

Fig. 9: Another representation of the example

The co-occurrences allow us to define that:

$$\begin{array}{ll} f(g_1) = \{a_1\} & f(g_2) = \{a_2, b_1\} \\ f(g_3) = \{b_2\} & f(g_4) = \{a_3, c_1\} \end{array}$$

Examples of fusion of the following classes give as result:

$\cup$ classes	$f : \mathcal{G} \rightarrow P(\mathcal{O})$	Sets $\mathcal{K}$ and $\mathcal{U}$
$g_5 \cup g_6 \rightarrow g_{56}$		$\mathcal{K} = \{g_1, g_2, g_3, g_4\}$ $\mathcal{U} = \{g_{56}\}$
$g_1 \cup g_6 \rightarrow g_{16}$	$f(g_{16}) = \{a_1\}$	$\mathcal{K} = \{g_{16}, g_2, g_3, g_4\}$ $\mathcal{U} = \{g_5\}$
$g_2 \cup g_6 \rightarrow g_{26}$	$f(g_{26}) = \{a_2, b_1\}$	$\mathcal{K} = \{g_1, g_{26}, g_3, g_4\}$ $\mathcal{U} = \{g_5\}$
$g_1 \cup g_2 \rightarrow g_{12}$	$f(g_{12}) = \{a_1, a_2\}$	$\mathcal{K} = \{g_{12}, g_3, g_4\}$ $\mathcal{U} = \{g_5, g_6\}$
$g_1 \cup g_3$ $g_3 \cup g_4$	Forbidden	

TABLE V: Example of fusion and the corresponding results

Final association between written names and clusters

When the stopping criterion is reached, for each g of $\mathcal{K}$ which have only one name associated ($card(\{h(o) \mid o \in f(g)\}) = 1$), we identify g directly by the name. For other clusters belonging to $\mathcal{K}$, we select the name that has the best TF.IDF score for the cluster (see section IV-A3).

In fact, in the *REPERE* corpus, only the show "Pile et face" regularly displayed two names simultaneously. But these names are also written alone at different moments during the show. So in most cases, the clustering will produce clusters associated with a single name. The clustering is stopped according to the optimal threshold (minimizing EGER) tuned on the training set.

V. Experimental Results

In this section we show the difference of behavior between late naming (LN1, LN2, LN3, LN3$^{\ominus}$), integrate naming IN and early naming EN, their ability to correctly identify speakers in TV broadcast and their sensitivity to the diarization stopping criterion threshold.

A. Evaluation metrics

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

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

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

To evaluate the identification quality we used the REPERE protocol where one sample every ten seconds is evaluated. The official REPERE metric is defined as:

$$\text{EGER} = \frac{\text{\#fa} + \text{\#miss} + \text{\#conf}}{\text{\#total}}$$

where we count the number of errors: confusions ($\#conf$), miss ($\#miss$) and false alarm ($\#fa$) divided by the number of person utterances to be detected ($\#total$).

We also used the precision, recall and F-measure with the same protocol (note as $\%P$, $\%R$, $\%F$ in the following tables).

B. Tuning the stopping criterion threshold

Two clusters are merging if the BIC score between them is higher than a threshold (note as Thr. in the following tables). A lower threshold means that the agglomeration stop later and therefore more clusters will merge together. We used our training set to tune the stopping criterion threshold. However, in order to be less dependent on manual annotations, we did not use the whole 24 hours training set and selected 100 subsets of 3 hours randomly from it. These subsets were chosen to match the test set characteristics (duration, balance between shows, and number of videos for each show).

As expected, table VI shows that the optimal threshold for IN is higher than those for LN. It means that IN stops earlier the agglomerative clustering, so it can split some clusters but name them with different occurrences of the same name.

Naming strategy	Thr.			
	median	min	max	stand. dev.
LN: lower DER	1540	1440	1680	54
IN: lower EGER	1620	1520	1740	44
EN: lower EGER	1260	300	1640	277

TABLE VI: Threshold tuned on 100 subsets of the train set, to minimize the DER or the EGER, LN: Late naming, IN: Integrate naming, EN: Early naming

The constrained clustering of EN stops at a lower threshold. The standard deviation for EN threshold is very high compared to the two others methods. We conclude that EN is less sensitive to the threshold value. For the rest of the paper, we chose to use the median of the table VI as global threshold.

For all the following experiences, it is important to note that stopping criterion thresholds are tuned on the training set while the results are displayed for the test set.

C. Comparison between different naming methods

Table VII summarizes the performance of the late naming methods with a diarization based on BIC.

Methods	Thr.	$\%P$	$\%R$	$\%F$	%EGER
LN1	1540	82.3	60.5	69.8	35.9
LN2	1540	82.7	62.7	71.3	34.1
LN3	1540	80.9	66.3	72.9	31.5
LN3$^{\ominus}$	1540	81.9	66.1	73.1	31.7
IN	1620	81.6	65.3	72.6	32.1
EN	1260	80.4	68.3	73.9	29.9

TABLE VII: Name propagation performance for the different late naming methods.

The different proposals that we do in the late naming method improves the results differently. First, we can see that LN2 (adding direct speech turns tagging step) names correctly more speech turns than LN1 (recall increased from 60.5% to 62.7%). This recall is improved by the one-to-many speaker tagging of LN3 (from 62.7% to 66.3%). Temporal re-alignment between audio and image (LN3$^{\ominus}$) improves the precision of LN3 (+1% in absolute).

The maximization of the final metric (the EGER) on all the corpus in IN with a high threshold do not improve the results. However, we will see at the end of the section that this method overpass LN3$^{\ominus}$ for particular shows.

The clustering constraint (EN) helps keeping the same precision (80.4%) though the threshold is lower. It allows to correctly merge some additional clusters and therefore increases the recall to 68.3%.

D. Contribution of written names

1) Speaker diarization:

To better understand the effect of the naming, we show the evolution of DER as a function of the threshold (see figure 10). This figure should be read from right to left as a smaller threshold value means that the agglomerative clustering stops later. The baseline "before naming" corresponds to an audio-only diarization. As explained in the previous section the diarization is different before and after the late naming.

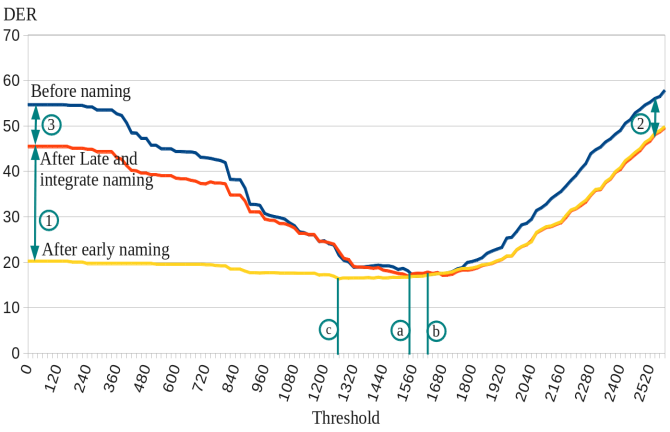

Fig. 10: Influence of the stopping criterion threshold (tuned on training set) on diarization error rate on test set, before and after naming.

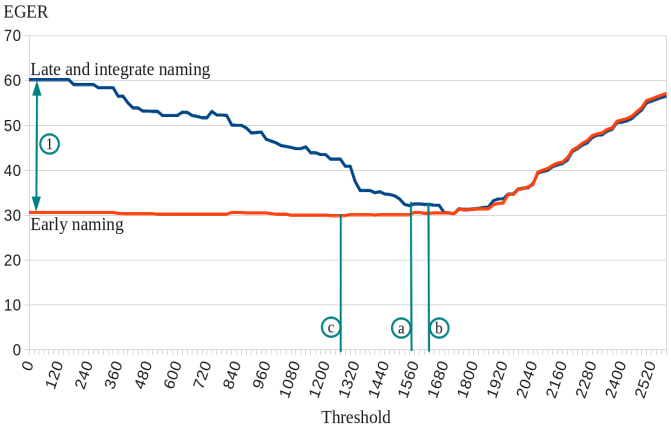

Fig. 11: Influence of the stopping criterion threshold (ⓐ, ⓑ, ⓒ tuned on train set) on identification error rate on test set, for the three naming strategies.

② and ③ in the figure 10 show the influence of the direct speech turn tagging step. At the start of the clustering ②, this step merges speech turns with the same name. At the end of the clustering ③, this step names directly some speech turns with a name different than the cluster name. ① shows the effect of the constraints preventing clusters with different names from being merged.

As shown in figure 10 and table VIII, ⓐ corresponds to the threshold tuned to minimize the DER. We obtain an 18.11% DER on the test set without written names (see Table VIII). "Late naming" merged some clusters with identical associated names, leading to a lower DER of 16.95%. The constrained clustering shows only a small variation of DER (from 18.7% to 20.2%, with a minimum of 16.37%) over the [0-1800] threshold range: it appears to be much less sensitive to the threshold choice (see figure 10).

	Thr.	DER
Before naming	ⓐ 1540	18.11
After late naming	ⓑ 1540	16.95
After early naming	ⓒ 1260	16.37

TABLE VIII: DER depending on the threshold

2) Speaker identification according to the threshold:

Figure 11 shows the evolution of EGER with respect to the selected threshold. `LN` and `IN` curves overlap but differ in the optimal stopping criterion threshold: threshold ⓐ aims at minimizing the DER (late naming) while ⓑ focuses on minimizing EGER (integrated naming). `EN` behaves very differently. ① shows the impact of the written name constraints and ⓒ the threshold tuned to minimize the EGER for the early naming method.

E. Sensitivity to the threshold tuning set

Threshold tuning is achieved by randomly selecting 100 subsets from the training set and choosing the best threshold value for each of them.

The x-axis of Figure 12 summarizes the range of variation of this optimal threshold over the 100 training subsets (e.g. 1440 to 1680 for late naming strategy), as already introduced in Table VI. The y-axis reports the corresponding average identification error rate (EGER) and its standard deviation on the test set.

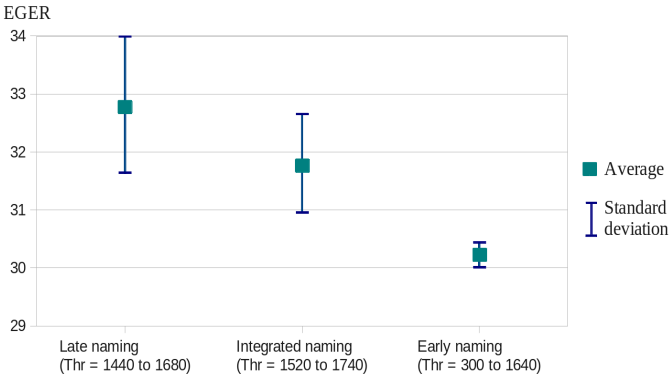

Fig. 12: Average and standard deviation of the EGER on test set depending on the subsets used to tuned the threshold

This figure points out that late and integrated naming strategies are more dependent on the training set and may therefore suffer from over fitting. Their respective identification error rates (EGER) have a standard deviation of 1.2% and 0.8%, while standard deviation of early naming EGER is only 0.2% (though the range of optimal thresholds over the 100 training subsets is much bigger).

F. Portability: tuning the threshold from a different show

The REPERE corpus is composed of seven different types of shows (`A` to `G`). Some shows in the corpus correspond to the same type. `A` and `E` are news programs, `C`, `D` and `F` are debate programs, `B` is a short show (2 minutes) of people news and `G` corresponds to debates at to French Assembly.

While a global threshold can be tuned on the training set, we also investigate the use of a show-dependent threshold (only a part of the training set is used). The robustness of a particular naming strategy can be inferred by the difference between the optimal threshold *Thr oracle* (corresponds to the

best possible performance in case an oracle is able to predict the best threshold, with the DER or the EGER, for a particular show) and the thresholds tuned on:

- The whole tuning set: *Thr. Global*.
- Same shows of the tuning set: *Thr. tuned on* A, test on A.
- Same type of shows of the tuning set: *Thr. tuned on* E and test on A, for news programs for example.
- Different types of shows of the tuning set: *Thr. tuned on* B and test on A.

In the following tables, we report the EGER scores of each show (column, computed on the test set) as a function of threshold selected (in parenthesis, tuned on the whole or a part of the training set). The last line of each table corresponds to the standard deviation of the EGER score for each show.

Show		Thr. apply on A	B	C	D	E	F	G
Oracle DER:								
Threshold		1700	1600	1560	1800	1300	1540	1000
Corres. EGER		31.6	70.7	25.7	20.4	54.6	22.5	11.5
Thr. Global (1540)		31.6	73.2	24.8	22.4	54.6	22.5	13.6
Thr. tuned on	A (1800)	**32.7**	69.5	28.7	20.4	43.2	31.7	14.5
	B (1580)	31.9	**73.2**	26.7	25.2	54.6	21.7	13.6
	C (1660)	31.9	70.7	**27.7**	21.8	54.6	23.3	13.6
	D (1620)	31.9	70.7	26.7	**25.2**	54.6	22.5	13.6
	E (1640)	31.9	70.7	27.7	21.8	**54.6**	23.3	13.6
	F (1300)	35.2	75.6	51.5	49.7	54.6	**40.0**	13.2
	G (1500)	35.2	75.6	24.8	21.8	54.6	25.8	**13.6**
Stand. dev.		1.6	2.5	9.3	10.4	4.3	6.7	0.4

TABLE IX: EGER for late naming LN3$^{\ominus}$ depending on the set used to tune the threshold

The table IX shows results for the "late naming" strategy (LN3$^{\ominus}$). The sensitivity of this method is once again demonstrated: tuned threshold on E (news program) leads to 54.6% of EGER for this same show. Whereas tune threshold on a same type of show (A), we obtained a lower EGER (43.2%). This can be also observed for the debates program F (Thr. tuned on F: EGER = 40%, Thr. tuned on C: EGER = 23.3%, Thr. tuned on D: EGER = 22.5%). In F, only one anchor and two guests are talking during the video, which allows to have a low threshold (1300) for the minimization of the DER. Contrary for the show C and D, one anchor, four or five guests and some reportage with voice-over are found, which lead to a higher threshold on the training set (1660, 1620). These higher thresholds lead to a lower EGER if we used them for the show F. These examples demonstrate the difficulty of using a threshold initially tuned to reduce the DER.

The oracle which selects the threshold that minimizes EGER (table X, first line) leads to better results than the oracle which minimizes DER (table IX, first line). This can be explained by the difference behavior of these two metrics:

- On one hand, DER minimization aims at associating one specific cluster to each speaker, whether they can be named or not.
- On the other hand, EGER minimization tries to associate a name to every speaker. Anonymous speakers can remain in the same cluster or split into several clusters as it has no influence on the final value of the identification error rate (EGER).

Show		Thr. apply on A	B	C	D	E	F	G
Oracle EGER:								
Threshold		1720	2100	1540	1780	1720	1600	1000
Corres. EGER		31.6	68.3	24.8	19.7	42.6	21.7	11.5
Thr. Global (1620)		31.9	70.7	26.7	25.2	54.6	22.5	13.6
Thr. tuned on	A (2000)	**32.4**	68.3	46.5	29.3	48.1	38.3	17.0
	B (2580)	47.4	**70.7**	76.2	61.2	60.1	60.0	46.4
	C (1620)	31.9	70.7	**26.7**	25.2	54.6	22.5	13.6
	D (1640)	31.9	70.7	27.7	**21.8**	54.6	23.3	13.6
	E (1640)	31.9	70.7	27.7	21.8	**54.6**	23.3	13.6
	F (1560)	31.9	73.2	25.7	25.2	54.6	**22.5**	13.6
	G (1520)	31.6	74.4	24.8	21.8	54.6	25.8	**13.6**
Stand. dev.		5.9	2	19.1	14.2	3.5	14.1	12.3

TABLE X: EGER for Integrate naming IN depending on the set used to tune the threshold

As a matter of fact, since speaker names are written multiple times, it is not worth trying to get exactly one cluster per speaker. A cluster can be split into multiple smaller purest clusters as long as those clusters are named correctly. For example, during the show F, guest names are displayed 24 times on average over the duration of each show. For this particular show, the optimal DER threshold is 1300 (leads to 40% of EGER, see table IX) while the EGER threshold tuned is 1560 (leads to 22.5% of EGER) very close to the oracle one (1600).

Show		Thr. apply on A	B	C	D	E	F	G
Oracle EGER:								
Threshold		920	2020	1100	1380	1780	860	580
Corres. EGER		32.1	69.5	24.8	15.0	42.6	22.5	11.1
Thr. Global (1260)		32.1	75.6	25.7	15.0	44.8	23.3	13.2
Thr. tuned on	A (940)	**32.4**	79.3	26.7	15.6	44.8	23.3	13.2
	B (1500)	32.1	**76.8**	25.7	15.0	45.4	22.5	13.6
	C (1500)	32.1	76.8	**25.7**	15.0	45.4	22.5	13.6
	D (1160)	32.1	76.8	24.8	**15.6**	44.8	23.3	13.2
	E (1220)	32.1	76.8	24.8	15.0	**44.8**	23.3	13.2
	F (1420)	32.1	76.8	25.7	15.0	45.4	**22.5**	13.6
	G (660)	32.4	80.5	26.7	15.6	45.4	23.3	**11.1**
Stand. dev.		0.2	1.6	0.8	0.3	0.3	0.4	0.9

TABLE XI: EGER for early naming EN depending on the show set to tune the threshold

Globally, results are better for IN when threshold is tuned and tested on the same show (diagonal of the matrix). However, the standard deviation of each column is more important. The use of the IN seems to be limited for tuning threshold on the same data type (same show, same type of shows).

Standard deviation values for the "early naming" are very low (see table XI). This demonstrates that, whatever the show used to tune thresholds, results will be very close, and furthermore, very close to the oracle one.

Finally, we highlight that oracle results show almost identical performance for the three strategies. However, since EN is less sensitive to the chosen threshold, it leads to significantly better identification performance.

G. Computation time

The response time of the constraint clustering of the early naming process with a single core at 2.00 Ghz is 81 seconds for 15 hours of the test set (MPEG2, 720x576, 25 frames/sec). This duration is 40 to 90 seconds depending on the criterion threshold.

The efficiency however depends upon the number of speaker turns to proceed for a video (every videos are treated separately).

This calculation time is negligible compared to the computation time of the pre-process (speech segmentation, distance matrix calculation between speech turns, overlaid names extraction).

VI. Comparison with state-of-the-art methods

A. Supervised mono-modal system

In order to highlight the efficiency of our proposed unsupervised algorithms, two supervised mono-modal speaker identification baseline (`GMM-UBM` and `GSV-SVM` [40], [39]) were also evaluated. The first one follows the standard Gaussian Mixture Model-Universal Background Model (GMM-UBM) paradigm, and the second one, 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[1] and additional French politicians data extracted from French radios.

System	$\%P$	$\%R$	$\%F$	%EGER
Oracle model	100.0	70.0	82.3	26.9
GMM-UBM	52.9	49.8	51.3	49.6
GSV-SVM	60.5	54.2	57.2	44.2
Our best system: `EN`	80.4	68.3	73.9	29.9

TABLE XII: Comparison of out best system with two supervised mono-modal systems

On the test set, 111 different persons spoke on the annotated images, but only 65 of them have a biometric model. We also add a line "Oracle model" in the table XII. This oracle allows us to know what is the result obtained by a perfect system but limited to models trained from the three sources mentioned above. The GSV-SVM system obtains better results than the GMM-UBM system, but with only a recall of 49.8% and 54.2% respectively, these methods obtained worse results than our best unsupervised system using early naming (`EN`).

B. Comparison with ILP clustering

In a previous work ([24]) we have tried to name speaker clusters with an ILP clustering. The main idea is to replace the classical agglomerative BIC clustering by an ILP clustering. This method finds the optimal clustering solution by maximizing the intra-class similarity and minimizing inter-class similarity. Integration of written names in this process constrains the clustering and simultaneously identifies clusters.

Results in table XIII are computed using only the annotated part of videos, due to computation issue of the ILP clustering. Therefore, some written names displayed outside the annotated part cannot be used. This explains the different scores of our method compared to those presented before (EGER increase from 29.9% to 34.7%). The first line corresponds to the maximum score than we can obtain with an oracle. This oracle identifies speakers when their name has been extracted from the written texts at least once in the video.

Approach	$\%P$	$\%R$	$\%F$	%EGER
Oracle: perfect cluster and perfect propagation	100.0%	62.1%	76.6%	32.8%
ILP [24]	90.6%	58.2%	70.9%	-
Our best system: `EN`	85.3%	61.5%	71.5%	34.7%

TABLE XIII: Identification of our methods compare to the ILP method [24]

These methods obtain essentially identical results in terms of F-measure. ILP method obtains a better accuracy while "early naming" gets very close recall to the oracle one. However, the main issue with ILP method is the computation time that limits the video duration that can be processed, while algorithms presented in this paper are is under one minute for one hour of video.

VII. Conclusion and Future Works

In this paper, we introduced and analyzed strategies for unsupervised speaker identification in TV broadcast. These approaches use exclusively written names on screen as source of names. Compared to pronounced names, usually used in state-of-the-art approaches, written names identify co-occurring speech turns with a very high precision.

In our propositions we integrate information provided by written names at different levels within the clustering process. With the "late naming" approach we propose different propagation of written names onto clusters provided by a diarization module. In our second proposition, "Early naming", we modified the speaker diarization module by adding constraints preventing two clusters with different written names from being merged together.

Our best unsupervised system reaches a F-measure of 73.9%, showing the pertinence of our approach. "Early naming" compared to "late naming" is less sensitive to the choice of the stopping criterion threshold of the clustering process. These methods intrinsically multi-modal clearly overpasses a (mono-modal) supervised system baseline.

Future works will focus on the integration of a diarization module at the-state-of-the-art (*I-vector*+ILP) and how a semi-supervised scenario, where manual annotations are added, can take advantage of our naming strategies.

Acknowledgment

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

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

REFERENCES

[1] Barras, C., Zhu, X., Meignier, S. and Gauvain, J.-L. Multi-stage speaker diarization of broadcast news. ASLP, 2006

[2] Canseco-Rodriguez, L., Lamel, L. and Gauvain, J.-L. Speaker diarization from speech transcripts. INTERSPEECH, 2004

[3] Canseco, L., Lamel, L. and Gauvain, J.-L. A comparative study using manual and automatic transcriptions for diarization. ASRU, 2005

[4] Charhad, M., Moraru, D., Ayache, S. and Quénot, G. Speaker identity indexing in audio-visual documents. CBMI, 2005

[5] Gauvain, J.-L., Lamel, L. and Adda, G. The limsi broadcast news transcription system. SC, 2002

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

[7] Ma, C., Nguyen, P. and Milind, M. Finding speaker identities with a conditional maximum entropy model. ICASSP, 2007

[8] Mauclair, J., Meignier, S. and Estève, Y. Speaker diarization: about whom the speaker is talking ? Odyssey, 2006

[9] Estève, Y., Meignier, S., Deléglise, P. and Mauclair, J. Extracting true speaker identities from transcriptions. INTERSPEECH, 2007

[10] Jousse, V., Petitrenaud, S., Meignier, S., Estève and Y., Jacquin, C. Automatic named identification of speakers using diarization and asr systems. ICASSP, 2009

[11] Petitrenaud, S., Jousse, V., Meignier, S. and Estève, Y. Identification of speakers by name using belief functions. IPMU, 2010

[12] El-Khoury, E., Laurent, A., Meignier, S. and Petitrenaud, S. Combining transcription-based and acoustic-based speaker identifications for broadcast news. ICASSP, 2012.

[13] Bechet, F., Favre, B. and Damnati, G. Detecting person presence in tv shows with linguistic and structural features. ICASSP, 2012

[14] Satoh, S., Nakamura, Y. and Kanade, T. Name-it: naming and detecting faces in news videos. Multimedia, 1999

[15] Houghton. R. . Named faces: putting names to faces. IS, 1999

[16] Yang, J. and Hauptmann, A.G. Naming every individual in news video monologues. ACMMM, 2004

[17] Yang, J., Yan, R. and Hauptmann, A.G. Multiple instance learning for labeling faces in broadcasting news video. ACMMM, 2005

[18] Sato, T., Kanade, T., Hughes, T.K., Smith, M.A. and Satoh, S. Video OCR: Indexing digital news libraries by recognition of superimposed caption. ACM Multimedia Systems, 1999.

[19] Poignant, J., Besacier, L., Quénot, G. and Thollard, F. From text detection in videos to person identification. ICME, 2012

[20] 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. and Ekenel, H.K. Fusion of speech, faces and text for person identification in tv broadcast. IFCVCR, 2012

[21] Poignant, J., Bredin, H., Le, V.B., Besacier, L., Barras, C. and Quénot, G. Unsupervised speaker identification using overlaid texts in tv broadcast. INTERSPEECH, 2012

[22] 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. and Stiefelhagen, R. Qcompere at repere 2013. SLAM, 2013

[23] Poignant, J., Bredin, H., Besacier, L., Quénot, G. and Barras, C. Towards a better integration of written names for unsupervised speakers identification in videos. SLAM, 2013

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

[25] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. and Ouellet, P. Front-End Factor Analysis for Speaker Verification. ASLP, 2011

[26] Poignant, J., Besacier, L. and Quénot, G. Nommage non-supervisé des personnes dans les émissions de télévision : une revue du potentiel de chaque modalité. CORIA, 2013

[27] Poignant, J., Besacier, L., Le, V.B., Rosset, S. and Quénot, G. Unsupervised naming of speakers in broadcast TV: using written names, pronounced names or both ? INTERSPEECH, 2013

[28] Kahn, J., Galibert, O., Quintard, L., Carré, M., Giraudel, A. and Joly, P. A presentation of the REPERE challenge. CBMI, 2012

[29] Giraudel, A., Carré, M., Mapelli, V., Kahn, J., Galibert, O. and Quintard, L. The REPERE corpus: a multi-modal corpus for person recognition, LREC, 2012.

[30] Chen, S.S. and Gopalakrishnan, P.S. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In DARPA Broadcast News Transcription and Understanding Workshop, 1998

[31] Barras, C., Zhu, X., Meignier, S. and Gauvain, J.-L. Multi-stage speaker diarization of broadcast news. ASLP, 2006

[32] Rouvier, M. and Meignier, S. A Global Optimization Framework For Speaker Diarization. Odyssey, 2012

[33] Lamel, L., Courcinous, S., Despres, J., Gauvain, J.-L., Josse, Y., Kilgour, K., Kraft, F., Le, V.B., Ney, H., Nubaum-Thom, M., Oparin, A., Schlippe, T., Schluter, R., Schultz, T., Fraga da Silva, T., Stuker, S., Sundermeyer, M., Vieru, B., Vu, N.T., Waibel, A. and Woehrling, C. Speech Recognition for Machine Translation in Quaero. IWSLT, 2011

[34] Gauvain, J.-L., Lamel, L. and Adda, G. Partitioning and transcription of broadcast news data. ICSLP, 1998

[35] Dinarelli, M. and Rosset, S. Models Cascade for Tree-Structured Named Entity Detection. IJCNLP, 2011

[36] Allauzen, A. and Bonneau-Maynard, H. Training and evaluation of pos taggers on the French multitag corpus. LREC, 2008

[37] Bernard, G., Rosset, S., Galibert, O., Bilinski, E. and Adda, G. Limsi participation in the QAst 2009 track: Experimenting on answer scoring. CLEF, 2009

[38] H. W. Kuhn. The hungarian method for the assignment problem. Naval Research Logistics Quarterly, 1955

[39] Campbell, W.M., Sturim, D.E. and Reynolds, D.A. Support Vector Machines Asing GMM Supervectors for Speaker Verification. Signal Processing Letters, 2006

[40] Le, V.B., Barras, C. and Ferràs, M. On the use of gsv-svm for speaker diarization and tracking. Odyssey, 2010

Johann Poignant is actually assistant lecturer (ATER) at the university Pierre-Mendès France (UPMF). He has a first Master diploma in 2004 in computer science at the University Claude Bernard Lyon 1 (France). Then, he taught for four years in a higher education institution. After a second master's degree, he defended his PhD thesis in Computer Science in October 2013 on "Unsupervised identification of persons in TV broadcasts" at the University of Grenoble (France). His research, conducted in the Laboratoire d'informatique de Grenoble (LIG), focuses on the use of multi-modality present in multimedia documents to identify people and particularly in TV broadcasts.

Laurent Besacier is Professor at UJF since September 2009. He defended his PhD thesis in Computer Science in April 1998 on "A parallel model for automatic speaker recognition" at the University of Avignon (France). Then he spent one and a half year at IMT (Switzerland) as an associate researcher working on M2VTS European project (Multimodal Person Authentication). Since September 1999 he is associate professor (professor since 2009) in Computer Science at the University Joseph Fourier (Grenoble, France). From September 2005 to October 2006, he was invited scientist at IBM Watson Research Center working on Speech to Speech Translation. His research interests can be divided in two main parts: 1. speech and audio processing in a multimodal framework, 2. multilingual speech recognition and (recently) translation. Laurent Besacier has supervised or co-supervised 12 PhD students and 15 Master students. Finally, he has been involved in several national and international projects: among others, NESPOLE European project on speech-to-speech translation, M2VTS European project on multimodal biometrics, as well as evaluation campaigns organized by NIST, DARPA or other organizations: RT, TRECvid, TRANSTAC, WMT, IWSLT.

Georges Quénot is Senior Researcher at CNRS (French National Centre for Scientific Research). He has an engineer diploma of the French Polytechnic School (1983) and a PhD in computer science (1988) from the University of Orsay. He is currently leading the Multimedia Information Indexing and Retrieval group (MRIM) of the Laboratoire d'informatique de Grenoble (LIG) where he is also responsible for their activities on video indexing and retrieval. His current research activity is about semantic indexing of image and video documents using supervised learning, networks of classifiers and multimodal fusion.