Towards a better integration of written names for unsupervised speakers identification in videos

Johann Poignant[1], *Hervé Bredin*[2]
Laurent Besacier[1], *Georges Quénot*[1], *Claude Barras*[2]

[1]UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France
[2]Univ Paris-Sud, LIMSI-CNRS, Spoken Language Processing Group, BP 133, 91403, Orsay, France
[1]first.lastname@imag.fr, [2]first.lastname@limsi.fr

Abstract

Existing methods for unsupervised identification of speakers in TV broadcast usually rely on the output of a speaker diarization module and try to name each cluster using names provided by another source of information: we call it "late naming". Hence, written names extracted from title blocks tend to lead to high precision identification, although they cannot correct errors made during the clustering step.

In this paper, we extend our previous "late naming" approach in two ways: "integrated naming" and "early naming". While "late naming" relies on a speaker diarization module optimized for speaker diarization, "integrated naming" jointly optimize speaker diarization and name propagation in terms of identification errors. "Early naming" modifies the speaker diarization module by adding constraints preventing two clusters with different written names to be merged together.

While "integrated naming" yields similar identification performance as "late naming" (with better precision), "early naming" improves over this baseline both in terms of identification error rate and stability of the clustering stopping criterion.

Index Terms: speaker identification, speaker diarization, written names, multimodal fusion, TV broadcast.

1. 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. However, the use of biometric models faces two main problems: 1) manual annotations: generating biometric models 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 hundreds 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 steps:

1. Speaker clustering (or diarization),
2. Extraction of hypothesis names from the video (or from the collection of videos),
3. Mapping (or association) between hypothesis names and speaker clusters.

Speaker diarization 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 approach which tries to merge speech turns into clusters that are the purest as possible using a distance metric (with a distance-based criterion to stop the clustering).

Two modalities, intrinsic to the video, can provide the name of speakers in broadcast TV: pronounced names and names written on the screen (see figure 1). Most state-of-the-art approaches rely on pronounced names due to the poor quality of written names transcription observed in the past. Naming speakers with pronounced names has been proposed by *Canseco et al.* [1, 2] and *Charhad et al.* [3]. Manually-designed linguistic patterns indicate whether a name refers to the speaker of the current speech turn, the following or the previous one. *Tranter et al.* [4] learn these patterns as sequences of *n-grams*. *Mauclair et al.* [5] use semantic classification trees (SCT) to match names and speaker turns. *Estève et al.* [6] compare these two techniques. They conclude that SCTs are less sensitive to automatic speech transcriptions errors than sequences of *n-grams*. *Jousse et al.* [7] improved over the SCT baseline: first, each name is attached locally to a nearby speech turn; names are then propagated globally to speaker clusters. They also show a performance drop from 19.5% to 85% in speaker identification error rate when using automatic speech transcription instead of (perfect) manual transcriptions and named entities detection. More recently, we proposed three propagation methods to propagate written names to speaker clusters [8]. These unsupervised multi-modal methods yield much better performance than mono-modal ones. We also show that these methods lead to 98.9% accuracy with perfect speaker diarization.

(a) (b)

Figure 1: Example of written names on the screen

The use of automatically extracted pronounced names faces several challenges: (i) transcription errors; (ii) named entity detection errors (missing first/last name, false alarms, etc.); (iii)

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).

mapping errors (current, previous or next speech turn).

The use of automatically extracted written names faces similar difficulties: (a) transcription errors – though better video quality reduces these errors; (b) detection errors – fewer because each TV show uses specific spatial position for title blocks[1]; (c) mapping errors – though a name is usually written on the screen while the person is talking, yielding easier affiliation.

This paper addresses other errors that can impact results: the errors made during the clustering process (during speaker diarization). For instance, the incorrect merging of two clusters containing different speakers can severely impact the speaker naming performance. Tuning the stopping criterion for hierarchical clustering is important to avoid such a problem. In this paper, we rely on the hypothesis that the high precision of written names to identify the current speaker can help us improve the diarization process in order to avoid the problems mentioned earlier. We limit our study to the use of written names for unsupervised speaker identification in videos and propose an extension of [8]. In this previous work, we proposed three methods for "late naming" of speakers which are highly dependent on the quality of speaker diarization. In this article, we present two novel approaches to overcome this issue: "integrated naming" aims at better choosing the value of the stopping criterion in order to minimize the speaker identification error while "early naming" adds written names-driven constraints to speaker diarization.

The outline of the paper is as follows. Section 2 presents the experimental setup as well as the speaker diarization module and the written names extraction module used in our experiments. Then, we describe our speaker naming methods in Section 3. Section 4 presents our experiments. Finally, we conclude this work and give some perspectives.

2. Experimental setup

The REPERE [9] evaluation campaign phase 1 took place in January 2013. The main objective of this challenge is to answer the two following questions at any instant of the video: *"who is speaking?" "who is seen?"*. In this paper, we try to answer the first question in an unsupervised way.

2.1. REPERE Corpus

The dataset used in our experiments is extracted from a corpus created for the REPERE challenge [10], which addresses multimodal person identification in videos. Videos are recorded from seven different shows (including news and talk shows) broadcasted on two French TV channels. An overview of the data is presented in Table 1.

	Train	Test
Raw video	58h	15h
Annotated part	24h	3h
Number of annotated frames	8766	1229

Table 1: Train and test sets statistics

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.

[1]Title block: spatial position used in the TV show to write a name and introduce the corresponding person.

Our evaluation is performed on test set. It is important to note that, although the whole test set is processed, the performance is measured only on the annotated frames. Figure 2 shows some statistics of the test set (duration and number of videos) for each TV show available in the REPERE corpus.

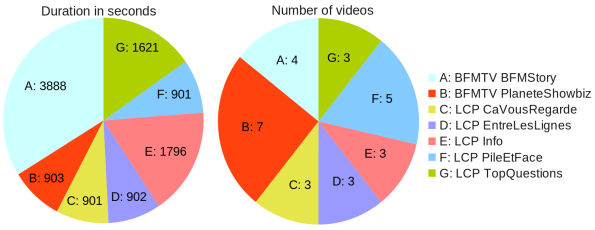

Figure 2: Duration and number of videos for the various TV shows available in the REPERE collection

2.2. Evaluation Metrics

Alongside the usual precision P and recall R, the official REPERE metric is also used for evaluation. It is called the Estimated Global Error Rate (EGER). This metric is defined as:

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

where #total is the number of person utterances to be detected, #conf the number of utterances wrongly identified, #miss the number of missed utterances and #fa the number of false alarms.

To evaluate speaker diarization performance, we also 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, d_{fa} the duration of false alarm, d_{miss} the duration of missed speech and d_{conf} the duration of the speech time where hypothesis and reference disagree. As identities of speakers are not considered, hypothesis and reference are aligned 1-to-1 to minimized d_{conf}.

2.3. Audio and Video Processing Modules

2.3.1. Speaker Diarization

Speaker diarization consists in segmenting the audio stream into speaker 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 [11] with single full-covariance Gaussians. This similarity matrix 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 [12] but our goal, here, is to do a fair comparison of several speaker naming methods, independently of the similarity measure (BIC or CLR).

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

2.3.2. Written names extraction

To detect the names written on the screen used to introduce a person, a detection and transcription system is needed. For this task we used LOOV [13] (LIG Overlaid OCR in Video). This system has been previously evaluated on another broadcast news corpus with low-resolution videos. We obtained a character error rate (CER) of 4.6% for any type of text and of 2.6% for names written on the screen to introduce a person.
From the transcriptions, we use a simple technique in order to detect the spatial positions of title blocks. This technique compares each transcript with a list of famous names (list extracted from Wikipedia, 175k names). Whenever a transcription corresponds to a famous name, we add its spatial position to a list. With the repeating positions in this list we find the spatial positions of title blocks used to introduce a person. However, these text boxes detected do not always contain a name. A simple filtering based on some linguistic rules allows us to filter false positives.

3. Unsupervised Naming of Speakers

We propose three methods for unsupervised (i.e. with no prior biometric models) naming of speakers with written names.

3.1. Late naming (`LN`)

Late naming is based on our previous work [8] (method M3). Speaker diarization and overlaid names recognition are run independently from each other. Speaker diarization is tuned to achieve the best diarization performance (i.e. minimize the diarization error rate, DER) as shown in Figure 3.

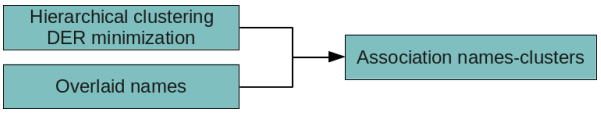

Figure 3: Late naming

The mapping between written names and speaker clusters is based on the following observations:

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

Therefore, this method proceeds in two steps. First, speech turns with exactly one co-occurring name are tagged. Then, each remaining unnamed speech turn is tagged cluster-wise using the following criteria:

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

where the *Term-Frequency Inverse Document Frequency* (TF-IDF)[14, 15] coefficient – made popular by the information retrieval research community – is adapted to our problem as follows:

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

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

In other words, speaker clusters are analogous to textual documents, whose words are detected written names.

Late naming is based on this method but there is a slight update that needs to be mentioned: we reduce the temporal scope of each written name to the more co-occurring speech turn, this can correct the time offset between audio and written names segmentation. It is important to note that the diarization can be different before and after the name-clusters association: some clusters may be merged (same name) or split (speech turn with a different name). Therefore, the scoring of the diarization can marginally change.

3.2. Integrated naming (`IN`)

One limitation of the late naming method is that the threshold used to stop hierarchical 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" is a simple extension of "late naming" where the stopping criterion threshold is tuned in order to minimize the EGER. We will show later in the experiments that the resulting threshold is generally higher than the one selected to minimize DER (i.e. agglomerative clustering is stopped earlier)

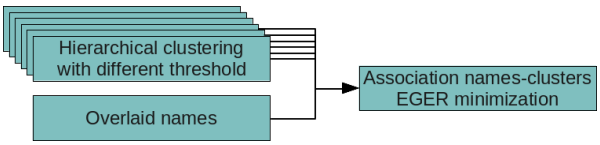

Figure 4: Integrated naming

In practice, as shown in Figure 4, we keep multiple clustering outputs, on which we apply the same method as in the "late naming" strategy described before. The threshold optimizing EGER on the training set is chosen.

3.3. Early naming (`EN`)

As already stated, when one or more names are written on the screen, there is a very high probability that the name of the current speaker corresponds to the written name on screen. Therefore, in "early naming", we use the information provided by written names during the clustering process.

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. For example, two clusters s_1 and s_2 **can** be merged into a new one s_{new} in the following case (the list of associated names is shown between brackets):

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

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

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

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

The clustering is stopped according to the optimal (minimizing EGER) threshold learned on the training set.

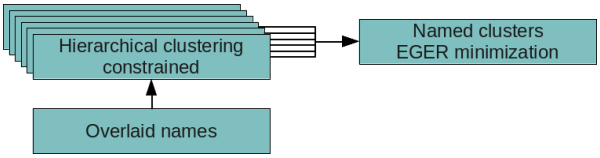

Figure 5: Early naming

4. Results

In this section we compare the ability of our naming methods to correctly identify speakers in TV broadcast and more particularly their sensitivity to the value of the stopping criterion threshold.

4.1. Learning the threshold as stop criterion

We used the training set to learn 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 randomly from it. These subsets were chosen to match the test set (duration, balance between shows, and number of videos for each show).

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

Table 2: Threshold learned on 100 subsets of the train set, to minimize the DER or the EGER, LN: Late naming, IN: Integrate naming, EN: Early naming

As expected, Table 2 shows that the optimal threshold for IN is higher than those for LN. It means that IN stops earlier in the agglomerative clustering though split clusters may end up with the same name.

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

4.2. Speaker Identification

For all the following experiences, it is important to note that the stopping criterion thresholds are learned on the training set while the results are displayed for the test set. Figure 6 shows the evolution of EGER with respect to the selected threshold and should be read from right to left as a smaller threshold value means that the agglomerative clustering stops later. 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 learned to minimize the EGER.

Table 3 summarizes the performance of the three methods. The integrated naming has a lower EGER but the difference is very small, yet this method has better precision due to its higher threshold. As far as EN is concerned, the clustering constraint 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%. For IN and EN, minimizing the EGER still allows to maximize other metrics like precision, providing at least enough speech duration to build speakers models.

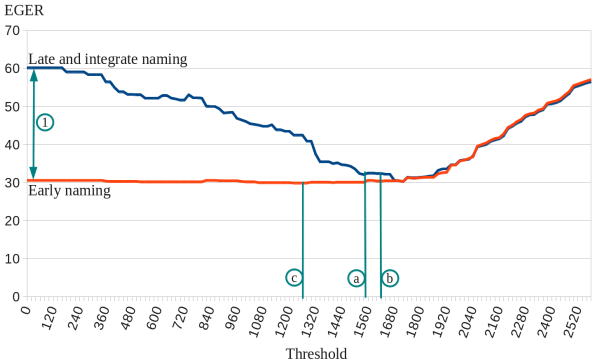

Figure 6: Influence of the stopping criterion threshold (ⓐ, ⓑ, ⓒ learned on train set) on identification error rate on test set, for the three naming strategies.

Naming strategy	Thr.	EGER (%)	P (%)	R (%)
Late (LN)	ⓐ 1540	32.1	80.4	66.0
Integrated (IN)	ⓑ 1620	32.4	81.5	65.3
Early (EN)	ⓒ 1260	29.9	80.4	68.3

Table 3: Trained stopping criterion threshold learned on the train set and the corresponding identification error (EGER), precision (P) and recall (R) obtained on test set.

4.3. Speaker Diarization

Figure 7 shows the evolution of DER as a function of the threshold. The baseline "before naming" corresponds to an audio-only diarization. As explained in section 3.2 the diarization is different before and after the late naming.

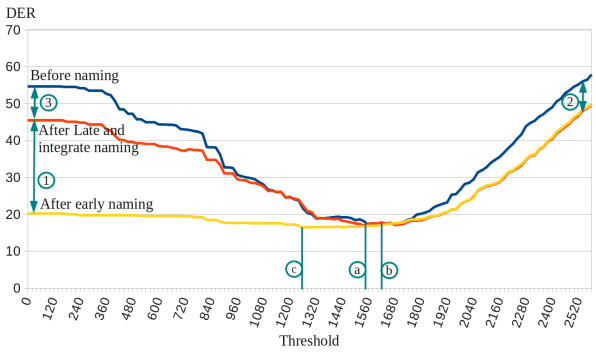

Figure 7: Influence of the stopping criterion threshold on diarization error rate on test set, before and after naming.

② and ③ show the influence of the direct speech turn tagging step. At the start of the clustering ②, this step merges

Proceedings of the First Workshop on Speech, Language and Audio in Multimedia (SLAM), Marseille, France, August 22-23, 2013.

speech turns with the same name. At the end of the clustering ③, this step removes from clusters some speech turns with a different name. ① shows the effect of the constraints preventing clusters with different names from being merged.

ⓐ corresponds to the threshold tuned to minimize the DER. We obtain an 18.11% DER on the test set without written names (see Table 4). "Integrated naming" has a higher threshold but some clusters end up merged (thanks to their identical associated names), leading to a lower DER of 17.5%. 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 7).

	Thr	DER
Before naming	ⓐ 1540	18.11
After late and integrated naming	ⓑ 1620	17.51
After early naming	ⓒ 1260	16.37

Table 4: DER depending on the threshold

4.4. Sensitivity to the training 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 8 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 2. The y-axis reports the corresponding average identification error rate (EGER) and its standard deviation on the test set.

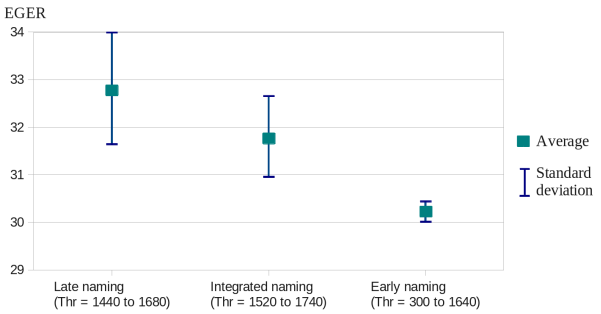

Figure 8: Average and standard deviation of the EGER on test set depending on the subsets used to learn the threshold

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) has 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).

4.5. Show-dependent threshold

The test corpus is composed of seven different types of shows (as illustrated in Figure 2). While a global show-independent threshold (*Thr. corpus*) can be trained, we also investigate the use of a show-dependent threshold (*Thr. per show*) and report the outcome of this experiment in Figure 9. *Thr. oracle* corresponds to the best possible performance in case an oracle is able to predict the best threshold. The robustness of a particular naming strategy can be inferred by the difference between the thresholds tuned on the whole training set (*Thr. corpus* and *Thr. per show*) and the optimal threshold (*Thr oracle*).

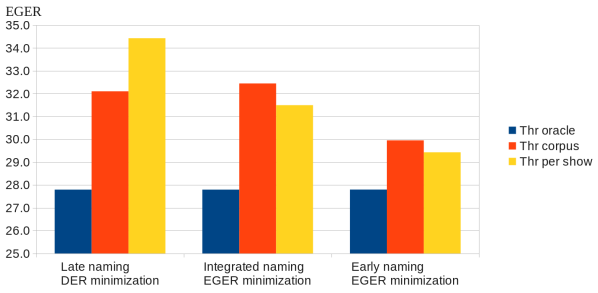

Figure 9: Identification error rate (EGER) for a show-dependent or show-independent stopping criterion.

Figure 9 shows that there is a difference of behavior between DER minimization (*late naming*) or EGER minimization (*integrated* or *early naming*). 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 its 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).

The REPERE corpus is composed of various types of shows. Some contains numerous speakers (up to 18 for news show *BFM Story*) whose names are usually displayed only once. Others, like the debate *Pile Et Face*, only have three speakers (two guests and the anchor) whose names are displayed 24 times on average over the duration of each show. For this particular type of show, the optimal DER threshold is 1300 while the EGER one is 1560. 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 speaker cluster can be split into multiple smaller clusters as long as those clusters are named correctly.

Finally, we highlight that oracle results show almost identical performance for the three strategies. However, since early naming is less sensitive to the chosen threshold, it leads to much better identification performance (very close to the oracle one).

5. Conclusions

In this paper, we introduced and analyzed two naming strategies for unsupervised speaker identification in TV broadcast. *Integrated naming* is a simple extension of our previous work [8] that improves precision (+1.1%) while keeping the same identification error rate (32.4%). *Early naming* relies on the knowledge of overlaid names during the clustering process. This information is used to constrain clustering by preventing two clusters named by different written names from being merged. This method leads to better identification error rate (29.9%) and is less sensitive to the choice of the stopping criterion threshold. These two methods allow maximizing the metric associated to the target task. Future works will focus on the integration of additional sources of information like pronounced names or face clustering.

6. References

[1] Canseco-Rodriguez L., Lamel L., Gauvain J.-L., Speaker diarization from speech transcripts, *the 5th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2004, p. 1272-1275, Jeju Island, Korea.

[2] Canseco L., Lamel L., Gauvain J.-L., A Comparative Study Using Manual and Automatic Transcriptions for Diarization, *Workshop on Automatic Speech Recognition and Understanding (ASRU*, 2005, p. 415-419, Cancun, Mexico.

[3] Charhad M., Moraru D., Ayache S., Quénot G., Speaker Identity Indexing In Audio-Visual Documents, *Content-Based Multimedia Indexing (CBMI)*, 2005, Riga, Latvia, 2005

[4] Tranter S. E., Who Really Spoke When? Finding Speaker Turns and Identities in Broadcast News Audio, *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006, p. 1013-1016, Toulouse, France.

[5] Mauclair J., Meignier S., Estève Y., Speaker diarization : about whom the speaker is talking?, *IEEE Odyssey 2006 - The Speaker and Language Recognition Workshop*, 2006, p. 1-6, San Juan, Porto Rico.

[6] Estève Y., Meignier S., Deléglise P., Mauclair J., Extracting true speaker identities from transcriptions, *the 8th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2007, p. 2601-2604, Antwerp, Belgium.

[7] Jousse V., Petit-Renaud S., Meignier S., Estève Y., Jacquin C., Automatic named identification of speakers using diarization and ASR systems, *the 34th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2009, p. 4557-4560, Taipei, Taiwan.

[8] Poignant J., Bredin H., Le V.B., Besacier L., Barras C., Quénot G., Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast, *the 13rd Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2012, Portland, USA.

[9] Kahn J., Galibert O., Quintard L., Carr M., Giraudel A., Joly P., A presentation of the REPERE challenge, *Workshop on Content-Based Multimedia Indexing (CBMI)*, 2012, p 1-6, Annecy, France.

[10] Giraudel A., Carré M., Mapelli V., Kahn J., Galibert O., Quintard L., The REPERE Corpus : a Multimodal Corpus for Person Recognition, *the 8th International Conference on Language Resources and Evaluation (LREC)*, 2012, p. 1102-1107, Istanbul, Turkey.

[11] Chen S. S. and Gopalakrishnan P., Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion, in *DARPA Broadcast News Transcription and Understanding Workshop*, 1998, Virginia, USA.

[12] Barras C., Zhu X., Meignier S., Gauvain J.-L., Multi-Stage Speaker Diarization of Broadcast News, *IEEE Transactions on Audio, Speech and Language Processing*, 2006, vol. 14, no. 5, pp. 1505–1512.

[13] Poignant J., Besacier L., Quénot G., Thollard F., From Text Detection in Videos to Person Identification, IEEE *International Conference on Multimedia & Expo (ICME)*, 2012, p. 854-859, Melbourne, Australia.

[14] Robertson S. E., Jones K.S., Relevance weighting of search terms, Journal of the American Society for Information Science, 1976, p. 129-146.

[15] Fang H., Tao T., Zhai C., A formal study of information retrieval heuristics, the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004, p. 49-56, Sheffield, UK