Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast

Johann Poignant, Hervé Bredin, Viet-Bac Le, Laurent Besacier, Claude Barras, Georges Quénot

Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast

Johann Poignant[1], *Hervé Bredin*[2], *Viet Bac Le*[3],
Laurent Besacier[1], *Claude Barras*[2], *Georges Quénot*[1]

[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
[3]Vocapia Research, 3 rue Jean Rostand, Parc Orsay Université, 91400 Orsay, France
[1]first.lastname@imag.fr, [2]first.lastname@limsi.fr, [3]levb@vocapia.com

Abstract

We propose an approach for unsupervised speaker identification in TV broadcast videos, by combining acoustic speaker diarization with person names obtained via video OCR from overlaid texts. Three methods for the propagation of the overlaid names to the speech turns are compared, taking into account the co-occurence duration between the speaker clusters and the names provided by the video OCR and using a task-adapted variant of the TF-IDF information retrieval coefficient. These methods were tested on the REPERE dry-run evaluation corpus, containing 3 hours of annotated videos. Our best unsupervised system reaches a F-measure of 70.2% when considering all the speakers, and 81.7% if anchor speakers are left out. 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 57.5% F-measure when considering all the speakers and 45.7% without anchor.

Index Terms: unsupervised speaker identification, multimodal fusion, speaker diarization, optical character recognition, reproducible results

1. Introduction

With the growing amount of audio-visual content available nowadays, automatic person identification becomes a very valuable tool for searching and browsing data, relying on face detection or speaker identification for instance. However, speaker identification requires expensive manual annotations for training the models, and these models need to be adapted with data matching the actual acoustic condition for better efficiency. Since one can not consider the manual annotation of each new video source as a viable option, an interesting alternative is to use unsupervised approaches for building speaker models. To this end, one can combine an automatic clustering of the audio track into anonymous speakers (i.e. speaker diarization) and a source of information providing the true speaker name for at least some of the clusters. A first system using automatic speech transcription (ASR) and manual rules for extracting speaker names from the transcription was proposed in [1]. An evolution of this approach using semantic classification trees (SCT) was presented in [2]. Both methods detect named entities in the ASR and use linguistic information to find the true name of a speaker. However, the frequent errors in the automatic transcription of proper names limit this approach. To prevent this issue, [3] combined subtitles and manual audio transcripts for face recognition in TV series. These sources of information can be found in TV series or movies, but generally not in news or talk shows. Automatic transcription of overlaid texts can provide names information with a high reliability (see figure 1). Indeed, TV broadcast news, reports or talk shows often use an overlaid text to introduce a person. In this paper, we address an unsupervised speaker identification method in videos with the help of overlaid texts obtained via video OCR. The core aspect of this approach is that no a priori speaker models are needed. Only speaker diarization and video OCR technologies are used.

Figure 1: Two sample screen shots

The first section presents the systems used for speaker diarization and video OCR, the second section describes three methods to name the speech turns (without speaker models trained in a supervised fashion) and finally the last section compares supervised and unsupervised speaker identification results on a three-hours data set.

2. Monomodal Components

Our proposed approach for unsupervised speaker identification relies on two components: acoustic-based speaker diarization and video-based overlaid name detection.

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).

2.1. 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 conditon, only anonymous speaker labels can be provided at this stage. We use in our experiments the LIMSI multi-stage speaker diarization system for broadcast news data [4]. After splitting the signal into acoustically homogeneous segments, the clustering into speaker classes is performed in two steps: a first agglomerative clustering stage uses the BIC criterion with single full-covariance Gaussians and is optimized for providing pure clusters; then, a second clustering stage takes advantage of an increased amount of data per cluster and uses more complex models and a cross-likelihood ratio (CLR) as cluster distance. This system provides 14.1% Diarization Error Rate (DER) on our test data (see section 4.1) and 9.9% if overlapping speech is discarded.

2.2. Overlaid Name Detection

We present in [5] a video OCR system for overlaid texts in videos. Text detection is performed in three steps, first a coarse detection finds text box candidates, then text box coordinates are refined and finally a temporal tracking is performed. After adapting images (resolution, binarisation) to a standard OCR system, a post-processing combines multiple transcriptions of the same text box.

In addition to these methods of text detection and text recognition, we use a simple technique to find spatial positions of the text boxes used by the show to introduce a person. This technique is based on the recurrent spatial positions of famous people names.A preliminary analysis of our data has shown that when a speaker speaks and one or more text box is found by this technique, the speaker has his name written in 95% of the cases in our evaluation corpus.

3. Name Propagation

Let us denote $\mathcal{T} = \{t_1, \ldots, t_K\}$ the set of speech turns and $\mathcal{S} = \{s_1, \ldots, s_L\}$ the set of L speaker clusters found by our speaker diarization system. $\mathcal{N} = \{n_1, \ldots, n_M\}$ is the short list of M names detected by our video OCR approach. Figure 2 illustrates an example that will be referred to throughout the rest of the paper.

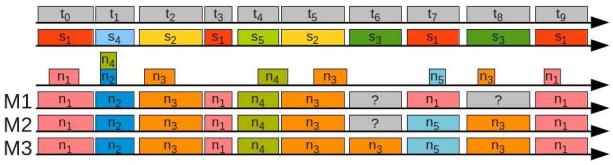

Figure 2: Speech turns, speaker clusters, overlaid names and results of the proposed name propagations.

We propose three different approaches to name propagation whose differences are illustrated in Figure 3. Their shared objective is to find the optimal mapping function m defined as:

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

3.1. One-to-One Speaker Tagging

This first method (denoted `M1` thereafter) relies on the strong assumption that speaker diarization provides perfect speaker clusters. Therefore, it consists in finding the one-to-one mapping $f\colon \mathcal{S} \to \mathcal{N} \cup \varnothing$ that maximizes the co-occurrence duration between speaker 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. $f(s) = \varnothing$ means the name of speaker s remains unknown and $\mathbb{K}(s, \varnothing) = 0$. The so-called Hungarian algorithm (also known as Munkres assignment algorithm) is used to solve this problem in polynomial time [6].

Figure 3 shows the output of this approach `M1` when applied on our running example: $s_1 \mapsto n_1$, $s_4 \mapsto n_2$, $s_5 \mapsto n_4$ and $s_2 \mapsto n_3$. Speaker s_3 remains unknown and name n_5 is not associated with any speaker.

3.2. Direct Speech Turn Tagging

The second approach (denoted `M2`) is based on the observation that, when one name n written alone on screen is detected, any co-occurring speech turn is very likely (91.5% precision on the test set) to be uttered by this person n. Therefore, our second approach is performed in two steps. First, speech turns with exactly one co-occurring name n are tagged with the latter. Then, the previous method `M1` is applied on the remaining unnamed speech turns. As a result, Figure 3 shows that speech turn t_7 is correctly renamed from n_1 (with method `M1`) to n_5 (with direct speech turns tagging).

3.3. One-to-Many Speaker Tagging

Our third (and last) proposed approach (denoted `M3`) 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 speaker clusters s_2 and s_3 in Figure 3. Therefore, this approach allows the propagation of an overlaid name to two or more speaker clusters.

First, direct speech turn tagging is applied similarly to method `M2`. Then, each remaining unnamed speech turns are tagged cluster-wise using the following criteria:

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

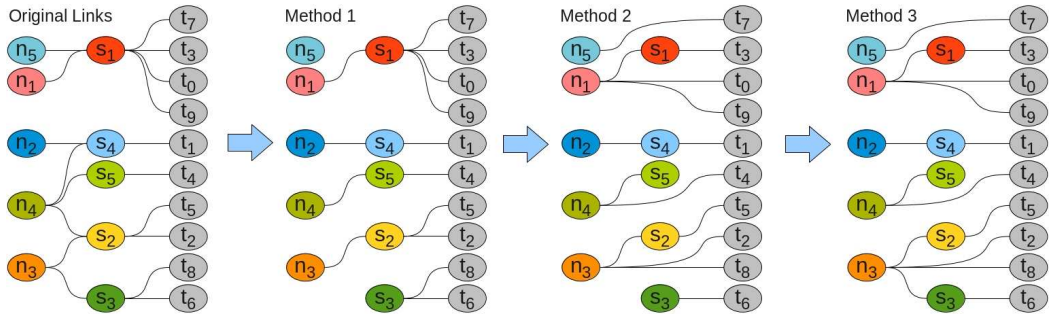

Figure 3: Name propagation methods `M1`, `M2` and `M3`

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:

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

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

where speaker clusters are analogous to textual documents, whose words are detected overlaid names. Figure 3 shows how speaker clusters s_2 and s_3 can be correctly merged using this approach.

4. Evaluation protocol

The REPERE[1] evaluation campaign dry-run took place in January 2012. The main objective of this challenge is to answer to 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 in an unsupervised way.

4.1. REPERE Corpus

The data used for our experiments are extracted from a corpus created for the REPERE challenge [7], which addresses multi-modal person identification in videos. The videos are recorded from seven different shows (including news and talk shows) broadcast on two French TV channels. An overview of the data is presented in Table 1.

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. Therefore, two processing conditions can be opposed. In the **full** condition, systems are allowed to use the whole videos to perform speaker identification. In the **standard** condition, only the annotated sections of the videos are available.

The development set is used to build speaker models for the contrastive supervised experiments, and the evaluation is performed on test set. Though the whole test set is processed, the performance is only measured on the annotated frames.

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

	Development	Test
Full condition	14h16	13h14
Standard condition	3h00	3h00
Number of annotated frames	1088	1107

Table 1: Development and test sets statistics

4.2. Evaluation Metrics

Alongside the usual precision P, recall R and corresponding F_1-measure F, the official REPERE metric is also used for evaluation, called the Estimated Global Error Rate (EGER). This metric is defined by:

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

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.

4.3. Speaker statistics

The distribution of speech duration in the test set is detailed in Table 2 for the standard condition. Due to the speech turns and duration imbalance between anchors and other speakers, results are systematically reported for all speakers vs. all but anchor speakers.

Type of speakers	#	Number of speech turns		Speech duration (minutes)	
Anchors	9	404	$\approx$ 45 ps.	45	$\approx$ 5 ps.
All others	113	1067	$\approx$ 10 ps.	133	$\approx$ 1 ps.

Table 2: Speech turns and duration (and average per speaker – ps.) in the test set

5. Reproducible[2] results & discussions

Table 3 summarizes the performance of the proposed methods in the full condition. Though `M2` has the best precision, `M3` is the best performing approach – thanks

[2] All the necessary material (source code and data) to reproduce the results reported in Tables 3 to 6 is freely available online at http://code.niderb.fr/

mostly to its higher recall. Furthermore, the overall performance is much better when anchors are not considered. The main reason is that the name of the anchor is seldom written on screen (as opposed to guests names) and therefore difficult to find.

Speakers	Propagation	%EGER	$\%P$	$\%R$	$\%F$
All	M1	44.4	80.5	58.2	67.5
	M2	41.9	82.1	60.7	69.8
	M3	38.7	77.7	63.9	70.2
No anchor	M3	28.4	89.2	75.3	81.7

Table 3: Name propagation performance, full cond.

In order to highlight the efficiency of our proposed unsupervised algorithm, a supervised mono-modal speaker identification baseline SID was also evaluated. It is based on the GSV/SVM modeling [8] with a total of 535 target speaker models trained using several TV and radio sources (including the REPERE development set).

Table 4 compares the performance of our best unsupervised approach with this supervised baseline. Because only 50% of all test set speakers actually had a corresponding SID model[3], the unsupervised approach greatly outperforms the supervised one. A simple combination of these two approaches (unsupervised identification followed by supervised identification on still-unnamed speech turns) allows to get even better results, especially in terms of recall.

Speakers	Approach	%EGER	$\%P$	$\%R$	$\%F$
All	SID	48.8	60.1	55.1	57.5
	M3	38.7	77.7	63.9	70.2
	M3 + SID	27.2	77.9	77.0	77.5
No anchor	SID	61.2	47.0	44.4	45.7
	M3	28.4	89.2	75.3	81.7
	M3 + SID	22.7	80.7	83.4	82.0

Table 4: Supervised (SID) vs. unsupervised (M3) speaker identification and their combination (M3+SID), full cond.

Table 5 shows that the knowledge of the full video (rather than just the annotated/evaluated part) allows to obtain names that might have been missed otherwise. It is especially true for anchor speakers whose name are usually written only once at the beginning of the video. The use of the OCR names extracted from the full video is worthwhile anyway.

Speakers	Condition	%EGER	$\%P$	$\%R$	$\%F$
All	Standard	46.7	82.0	55.6	66.3
	Full video	38.7	77.7	63.9	70.2
No anchor	Standard	30.9	88.5	72.4	79.7
	Full video	28.4	89.2	75.3	81.7

Table 5: Effect of condition on M3 performance.

[3]For those speakers, SID commits half as many errors as M3.

Table 6 allows to apprehend the impact of mistakes made by both speaker diarization and name propagation modules. The first two lines show that propagation errors have little impact when speaker diarization is perfect[4]. However, speaker diarization errors yield a significant performance decrease (-7% F_1-measure). Yet, as expected, M2 and M3 are less sensitive to diarization errors than M1.

SD	Propagation	EGER	$\%P$	$\%R$	$\%F$
Perfect	Perfect	23.5	100.0	76.5	86.7
	M1	23.6	98.0	76.4	85.8
Auto	M1	33.0	89.1	70.3	78.6
	M2	30.3	91.0	73.1	81.0
	M3	30.9	88.5	72.4	79.7

Table 6: Effect of speaker diarization (SD) and name propagation errors (standard condition, without anchors).

6. Conclusion & future work

We present in this article a method for unsupervised speaker identification in TV broadcasts, using person names obtained via video OCR on overlaid texts. Three methods for OCR names propagation on the speaker diarization clusters are proposed and evaluated on a 3 hours corpus of video including news and talk shows. Our best unsupervised system reaches a F-measure of 70.2% when considering all the speakers, and 81.7% if anchor speakers are left out, showing the relevance of our approach. The results also show that our unsupervised approach (intrinsically multi-modal) clearly overpasses a (mono-modal) SID baseline; their combination leads to the best results. Future works will focus on early fusion methods and on the use of cross shows information. Unsupervised training of collection of speaker models and application of these methods to face recognition are other promising extensions.

7. References

[1] L. Canseco, L. Lamel, and J.-L. Gauvain, "A Comparative Study Using Manual and Automatic Transcriptions for Diarization," in *IEEE Workshop on Automatic Speech Recognition*, 2005.

[2] V. Jousse, S. Petit-Renaud, S. Meignier, Y. Esteve, and C. Jacquin, "Automatic Named Identification of Speakers Using Diarization and ASR Systems," in *IEEE ICASSP*, 2009, pp. 4557–4560.

[3] M. Everingham, J. Sivic, and A. Zisserman, "Taking the Bite out of Automatic Naming of Characters in TV Video," *Image and Vision Computing*, vol. 27, no. 5, 2009.

[4] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multi-Stage Speaker Diarization of Broadcast News," *IEEE TASLP*, vol. 14, no. 5, pp. 1505–1512, September 2006.

[5] J. Poignant, L. Besacier, G. Quénot, and F. Thollard, "From Text Detection in Videos to Person Identification," in *IEEE ICME*, 2012.

[6] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[7] A. Giraudel, M. Carré, Valérie Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The REPERE Corpus : a Multimodal Corpus for Person Recognition," in *LREC*, 2012.

[8] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support Vector Machines Asing GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308–311, May 2006.

[4]Recall is not perfect since some target speakers names never appear.