The REPERE Corpus : a multimodal corpus for person recognition

Aude Giraudel[1], Matthieu Carré[2], Valérie Mapelli[2]
Juliette Kahn[3], Olivier Galibert[3], Ludovic Quintard[3]

[1] Direction générale de l'armement, France
[2] ELDA, France
[3] Laboratoire national de métrologie et d'essais, France
aude.giraudel@dga.defense.gouv.fr, {lastname}@elda.org, {firstname.lastname}@lne.fr

Abstract

The REPERE Challenge aims to support research on people recognition in multimodal conditions. To assess the technology progression, annual evaluation campaigns will be organized from 2012 to 2014. In this context, the REPERE corpus, a French videos corpus with multimodal annotation, has been developed. This paper presents datasets collected for the dry run test that took place at the beginning of 2012. Specific annotation tools and guidelines are mainly described. At the time being, 6 hours of data have been collected and annotated. Last section presents analyses of annotation distribution and interaction between modalities in the corpus.

Keywords: People recognition, French, Multimodal corpora

1. Introduction

The REPERE Challenge[1] aims to support the development of automatic systems for people multimodal recognition in videos. Funded by the French research agency (ANR) and the French defence procurement agency (DGA), this project has started on March 2011 and ends on March 2014.

An evaluation is organized at the beginning of each year by ELDA and LNE. The first evaluation is a dry run and will occur at the beginning of 2012. The other two campaigns will be organized respectively at the beginning of 2013 and 2014. These official campaigns are open to external consortia who want to participate in this challenge.

During this three-year project, a major effort will be dedicated to the production of the REPERE corpus. The goal is to build a corpus of 60 hours of videos with multimodal annotations, i.e rich speech transcription and rich video annotation. At the time being, 6 hours have been produced, corresponding to the development and test sets for the dry run. This paper describes the corpus building. The second section presents the data and the third section describes in detail the needful annotations.

2. Data

2.1. Data identification and IPR issues

When collecting data, one has to face questions related to the expected use of such data on the one side and their allowed use on the other side (these aspects will be developed in the final paper). For the REPERE corpus, ELDA, who has a long-lasting experience in that activity, could obtain agreements with two French TV channels (BFM TV and LCP), so as to collect their self-produced TV shows concerning news and debates.

[1]Additional information is available on dedicated website: www.defi-repere.fr

2.2. The REPERE dry run corpus

Development and test sets are respectively made of 3 hours data sets with a similar repartition of TV shows presented in table 1.

TV Show	Channel	Dev/test set (minutes)
BFM Story	BFM	60
Planete Showbiz	BFM	15
Ca vous regarde	LCP	15
Entre les lignes	LCP	15
Pile et Face	LCP	15
LCP Info	LCP	30
Top Questions	LCP	30

Table 1: TV shows for the REPERE corpus (dry run)

A variety of TV shows with gradual difficulties in audio and video contents has been selected (see figure 1). The selection criteria is the situation diversity so as to have the larger panel of examples. The focus is put on prepared vs. spontaneous speech, head size and orientation, camera motion and angle, lighting, etc. This first corpus should be considered as a baseline for future parts of the REPERE corpus both in terms of video and audio content and annotation guidelines.

Figure 1: Visual examples of TV shows

3. Annotation of people in videos

In the REPERE Challenge, competitive systems try to answer the four following questions using information that come from audio and video frames :

1. who is speaking?
2. who is present in the video?
3. which name is cited?
4. which name is displayed?

To answer those questions, the sources may be only the audio frame, only the video frame or a combination of both, as summarized in table 2.

	Audio frame	Video frame	Both
Who is speaking ?	•		•
Who is present in the video ?		•	•
What names are cited?	•		•
What names are displayed?		•	•

Table 2: Tasks and sources

Two kinds of annotations are thus produced in the REPERE corpus : audio annotation with rich speech transcription and video annotation with head and embedded text annotation. Table 3 summarizes annotations of the dry run corpus.

Annotations		Dev set	Test set
Visual	Segmented head	1421	1534
	Words in text transcription	13240	14764
	Named entities in text transcription	200	141
Audio	Speech segments	1571	1602
	Words in speech transcription	33205	33247
	Named entities in speech transcription	242	191

Table 3: Annotations in the REPERE dry run corpus

3.1. Audio annotation

Rich speech transcriptions is a well-known task for which reference guidelines and annotation tools exist. For the REPERE corpus, audio annotation is produced with Transcriber[2] annotation tool (Barras et al., 2000) in *trs* format. We chose to use a version of Transcriber that includes the visualization of the video during the audio transcription process. It is of great help for annotators that can take advantage of visual clues for the audio annotation.
The annotation guidelines are the ones created in the ESTER2[3] (Galliano et al., 2005) project for rich speech transcription.
The following elements are annotated :

- Speaker turn segmentation
- Speaker naming
- Rich speech transcription tasks gather segmentation, transcription and discourse annotation (hesitations, disfluences...)
- The annotation of named-entities of type ”person” in the speech transcription.

The annotation of named-entities is naturally focused on entities of type ”person”. To facilitate the gathering of person names, we created a shared database of people names. A web application allows annotator to query the database to get normalized form of a given name present in the database (see figure 2 for illustration).The normalized form is pasted directly into the transcription. If not present in the database, the person name is added in a form where annotators should give firstname, lastname and optional variants of namings. The normalized form is automatically generated and then available for all annotators.

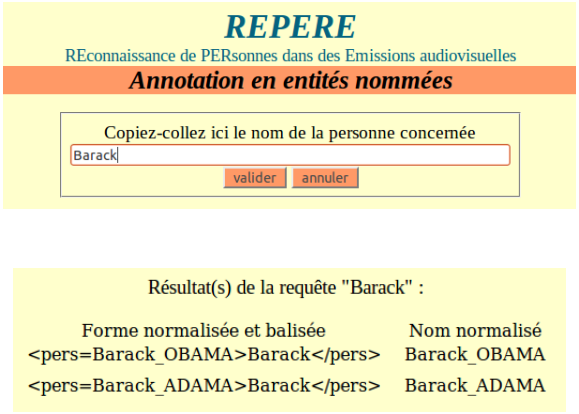

Figure 2: Person database query and normalized results

3.2. Video annotation

In complement to audio annotation, video annotation has necessitated the creation of specific annotation guidelines[4]. The video annotation task is composed of:

- Head segmentation
- Head description
- People identification
- Embedded text segmentation and transcription
- Named-entities (type ”person”) annotation in transcripts of embedded texts

VIPER-GT[5] video annotation tool has been selected for its ability to segment objects with complex shapes and to enable specific annotation schemes. The video annotation is

[2]http://trans.sourceforge.net
[3]http://www.afcp-parole.org/camp_eval_systemes_transcription
[4]Guidelines are available for participants on the REPERE website. They will be distributed with the REPERE corpus at the end of the project.
[5]http://viper-toolkit.sourceforge.net

a two phase process. First, head and embedded texts are segmented in selected key frames. The selection of key frames is achieved by annotators. The objective is to keep one key frame by scene. We observe that this process leads to the selection of one key frame every ten seconds on average. Then, for each segmented object, appearance and disappearance timestamps are annotated thanks to a tool specifically developed for the project (see section 3.2.3.). Annotations are produced in *xgtf* format.

3.2.1. Head annotation

By considering people's head, and not only face, the annotation also concerns people from behind or in profile (in the selected key frames). Note that, out of the key frame, those people can turn and face the camera during the rest of their apparition.

Head segmentation.

The first phase in head annotation is the segmentation step. Heads are segmented when their area in pixels is greater than a threshold that determines the ability to recognize people in videos. In practice, heads that have an area larger than 2500 pixels^2 are isolated. Heads are delimited by polygons that best fit the outlines. For those heads that are too small to be segmented, the key frame is annotated with the presence of small heads. Figure 3 illustrates the head segmentation and the presence of small heads in the frame.

Figure 3: Polygonal head segmentation

Head description.

Each segmented head may have physical attributes (glasses, headdress, moustache, beard, piercing or other). The head orientation is also indicated: face, sideways, back. The orientation choice is based on the visible eyes count. Finally, the fact that some objects hide a part of the segmented head is indicated specifying the object's type.

Identification of segmented head.

To identify segmented head, annotators have access to several types of information. First, the video annotation guideline contains a description of major participants in selected TV shows. For occasional speakers, annotators have to search in the entire video when the speaker is introduced either in speech or in the video. Furthermore, the database of person names presented in 3.1. is available to help annotators in the identification task and to ensure the normalization of names in the corpus.

People who are not named in the video are identified with a unique ID of type *unknown_XXX* (*XXX* incremental counter).

3.2.2. Embedded text annotation

Like for head annotation, embedded text annotation is to segment, describe and transcript foreground texts. For this annotation task, permanent texts, such as logos, are left out.

Embedded text segmentation and description.

Targeted texts are segmented with rectangles that fit best the outlines (see figure 4).

Figure 4: Foreground text segmentation (green rectangles)

Texts are segmented as coherent blocks and not only in words. Each block of text is associated with an attribute which describes how readable the text is. Values of this attribute are: complete, incomplete, unreadable, distorted.

Embedded text transcription.

The transcription of segmented text is a fair view of what appears in the video. All characters are reproduced with preservation of capital letters, word wrap, line break, etc. Annotation of named-entities of type "person" is directly added into transcriptions of text via VIPER-GT. The normalization process is still active with the access to the centralized database of person names.

3.2.3. Appearance and disappearance timestamps

The annotation of appearance and disappearance timestamps is the next step. The aim is to identify the segments where the annotated object is present. This annotation step concerns both head and embedded texts, and extends keyframe discrete annotation. Some criteria have been defined to select the best timestamps for each annotated object (head or text) as follows:

1. The object is present along the whole segment despite some optional camera movements (same object with different angles).
2. The object is always big enough to be annotated
3. Appearance timestamp is the first frame where the object appears in the video
4. Disappearance timestamp is the last frame where the object is present in the video

This work has required a specific tool to help annotators quickly find beginning and ending frames. The timestamps annotator is a semi-automatic tool based on dichotomous search (see figure 5).

Figure 5: appearance timestamp detection

The use of this tool will be detailed in the final version of the article.

3.3. Harmonization of people names

Beyond the parallel annotation of audio and visual content, the corpus creation pays special attention to the multimodal annotation consistency. A people names database presented in section 3.1. ensures the coherence of given names in audio and visual annotations. Moreover, unknown people IDs are harmonized when the same person appears both in audio and video annotations.

The annotation of people whose name is not obviously present in the video is also managed. Those people named as unknown are given separate IDs in audio and video annotations. The harmonization process enables the matching between the two lists of people. The stategy is to keep video ID when available.

The separation between audio and video annotation may lead to incoherence issues in the naming of annotated people. To avoid such problems, two verification procedures have been put in place. The first one enables annotators to share normalized naming of annotated people and the second one give access to a harmonisation process in the identification of unnamed people.

3.3.1. Named people database

For audio and video annotation, annotators can have access to a database of named entities of type ”person”. This database is accessible via a web application and has been created for this project. It is first filled with normalized names of TV shows speakers. At any time, annotators can get back existing normalized names or add new names when the person is not already in the database.

3.3.2. Unknown people

The annotation of people whose name is not present, the audio and video process may lead to separate list of unknown people. The harmonisation process is as follows. General process leads to the preservation of the video ID. In case of absence of video annotation, the audio ID is kept.

3.4. Clues for people recognition

In this section, we go further into the analysis of annotations distribution to understand the relative importance of different audio and visual clues. The objective is to focus on elements that could optimize and guide systems development.

3.4.1. Audio vs. video clues for people identification

To find people names present in each evaluated frames, systems can take advantage of audio or visual direct citations, that is people names cited in speech signal or appearing on the screen. Table 4 presents this distribution in dev and test sets.

		Dev	Test
People to find	Head appears on screen	216	145
	Name appears on screen	200	141
	Unnamed seen on screen	177	138
	Speaking	141	122
	Name cited in speech	242	191
	Unnamed speaking	45	33
	Total count of persons	237	171

Table 4: Audio/visual clues in the REPERE dry-run corpus

In the development set of the dry run corpus, the distribution of audio and visual citations is as follows: 45% of the persons to be found have their neme appearing on the screen, and 55% have their name cited at some point in the speech. Furthermore, 33% of people to recognize are not cited either way, meaning that in unsupervised conditions, only 67% of identities are findable. Figure 6 presents an illutration of this distribution. In addition 51% of the person both appear on screen and speak, 40% only appear on screen and 9% only speak. As a consequence, a system looking for who is present needs a good head detection capability.

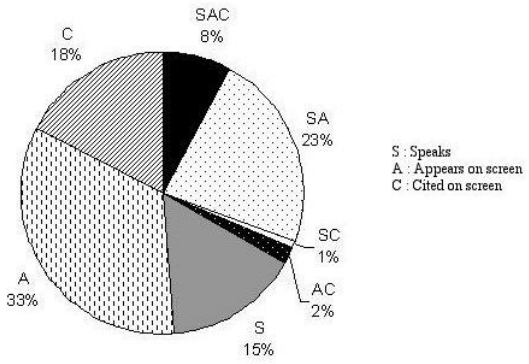

Figure 6: Clues distribution for people recognition

The numbers on the evaluation corpus are similar. 49% of the persons have their name showing up on the screen, 69% are cited. 22% are not cited at all, giving a 78% upper-bound for unsupervised approaches. 56% of the persons both appear in the image and talk, 29% only appear and 15% only talk.

First conclusion is that audio and visual clues are not equally distributed in the corpus. Moreover, distinct analysis on different TV shows, leads to the conclusion that this distribution is also very uneven between them as shown in figure 7.

We may conclude that different recognition strategies could be relevant to deal with different shows.

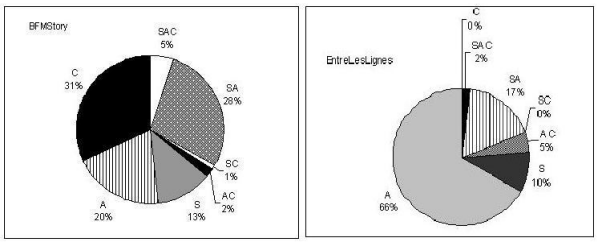

Figure 7: Clues distribution in 2 TV shows

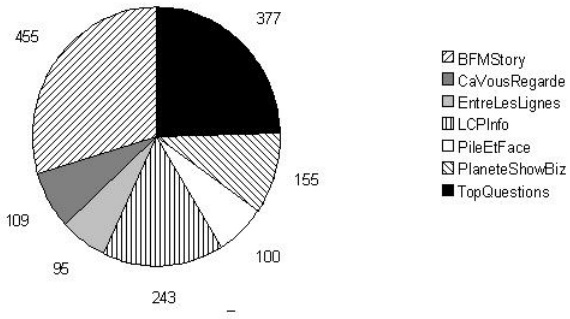

Figure 9: Head distribution

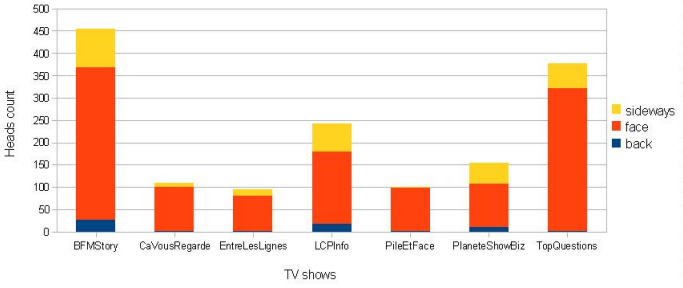

Figure 10: Head orientation

3.4.2. Speech duration

In terms of speakers distribution, the per-person speech durations are very uneven, as shown in figure 8. Speech times span from almost 10 minutes down to less than 20 seconds. The situation requires systems systems robust solutions for when a low amount of data is available.

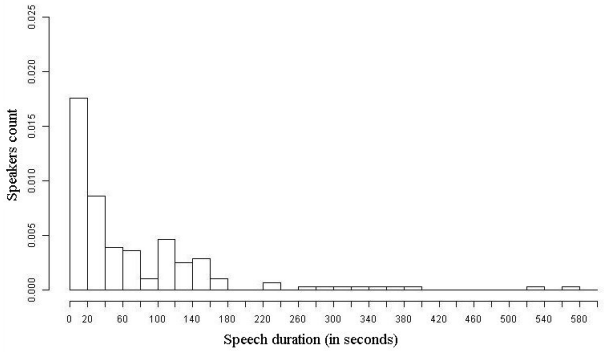

Figure 8: Speakers counts depending on speech duration

3.4.3. Head attributes

A study on heads attributes has also been conducted. 1534 heads have been annotated in the test set. The distribution of heads count through TV shows is represented in figure 9. We notice that the amount of people to recognize is largely uneven between all TV shows. It is quite logical if we consider that TV shows that have been annotated differ in duration and style. To be more precise, *BFM Story* contains almost 30% of people to be found while only 6% of them appear in *Entre les lignes*.

Concerning heads orientation, full-face heads are the most numerous in all TV shows. The amount of heads in profile is quite important in half of the shows while there is very few people from the back. Details of the distribution are shown in figure 10.

Another important element included in heads attributes is the presence of objects that can hide a part of the segmented head. Figure 11 shows that a great majority of heads are not hidden at all. For those that are partly hidden, the distribution of hiding objects varies between different TV shows (see figure 11).

The presence of a majority of full-face heads and not hidden ensures that in most cases it is possible to take advantage of complete heads characteristics to recognize people.

4. Conclusion and perspectives

The constitution of the REPERE corpus is the first step for the REPERE challenge success. The REPERE challenge aims to support research on people recognition in videos. Within the scope of the project, the creation of multimodal corpus is essential. The REPERE corpus consists on 60 hours of French videos annotated with visual (heads and embedded texts) and audio information. At the time being, the dry run corpus (6h) has been created and the first evaluation took place at the beginning of 2012.

There is a strong willingness to make all the data produced during this three-year project available for the research community. Consequently, we can already announce that the REPERE corpus will be distributed at reasonable cost at the end of the project.

Next step is the organization of the official upcoming evaluation campaigns in 2013 and 2014. We strongly invite external consortia willing to participate in this challenging competition to contact us.

5. References

C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. 2000. Transcriber: development and use of a tool for assisting speech corpora production. In *Speech Communication special issue on Speech Annotation and Corpus Tools*, volume 33, January.

S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J-F. Bonastre, and G. Gravier. 2005. The ester phase ii evaluation campaign for the rich transcription of french broad-

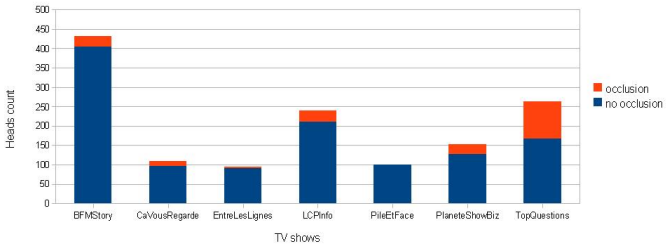

Figure 11: Distribution of hidden head in TV shows

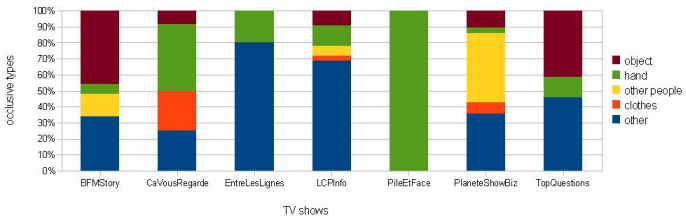

Figure 12: Distribution of hiding objects in TV shows

cast news. In *European Conference on Speech Communication and Technology*, pages 1149–1152.