Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

The REPERE Challenge: finding people in a multimodal context

Juliette Kahn[1], *Olivier Galibert*[1], *Matthieu Carré*[2], *Aude Giraudel*[3], *Philippe Joly*[4], *Ludovic Quintard*[1]

[1]Laboratoire National de métrologie et d'Essais, Trappes, France
[2]ELDA, Paris, France
[3]Direction Générale de l'Armement, Bagneux, France
[4]IRIT, Toulouse, France

firstname.lastname@lne.fr, carre@elda.fr, aude.giraudel@dga.gouv.defense.fr, joly@irit.fr

Abstract

The REPERE Challenge aims to support research on people recognition in multimodal conditions. To assess the technology progress, annual evaluation campaigns will be organized from 2012 to 2014. In this context, the REPERE corpus, a French video corpus with multimodal annotation, has been developed. The systems which participated in the dry run had to answer the following questions : Who is speaking ? Who is present in the video ? What names are cited ? What names are displayed ? The first results obtained during a dry run show that significant progress are quite possible. The challenge is to combine the various information coming from the speech and the images.

1. Introduction

Finding people on video is a major issue at a time when various information come from television and from the Internet. The challenge is to understand how to use the information about people that comes from the speech and the image and combine them so as to determine who is speaking and who is present in the video.

Some evaluation campaigns [1] or [2] worked on people multimodal recognition on English databases.

Started in 2011, the REPERE Challenge (REconnaissance de PERsonnes dans des Emissions audiovisuelles) aims at supporting the development of automatic systems for people recognition in a multimodal context. Funded by the French research agency (ANR) and the French defense procurement agency (DGA), this project has started in March 2011 and ends in March 2014.

To assess the systems' progress, two international campaigns will be organized at the beginning of each year by the Evaluation and Language resources Distribution Agency (ELDA) and the Laboratoire national de métrologie et d'essais (LNE). The first evaluation, which is a dry run, has occured at the beginning of 2012. The two main 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.

This paper presents the protocol that estimates the systems progress and the first results of the dry run. Section 2 describes the different tasks of the REPERE Challenge. Section 3 presents the data used to assess the systems. Section 4 is dedicated to the metrics description. Section 5 shows the first results. We conclude in Section 6, proposing some perspectives.

2. Questions and tasks

The goal of the REPERE Challenge is to support the development of automatic systems for people recognition in video. Video frames and speech signal are extracted from each video. The challenge is to extract from these two information sources the relevant features to know who is speaking and who appears in the video.

2.1. Main tasks

The first tasks in the REPERE Challenge are to determine every person who is visible and/or is speaking in the video. The goal is to combine the idiosyncratic information that comes from the speech and the video frames to answer those questions. These tasks are conducted in supervised and unsupervised modes. In supervised mode, participants can use other videos than those from the testing data. They are allowed to build voice models or head models for famous people who may be in the data. In unsupervised mode, the participants cannot use other data. They have to say who is on the video only with the clues included in the testing video

The secondary tasks are to determine the people who are cited in the video. The people can be cited in speech. For example, a speaker can mention another person or he can name his interlocutor. In addition, the names of the people may be displayed on the video frames. Those two tasks are conducted in unsupervised mode.

To sum up, in the REPERE Challenge, the systems try to answer to the four following questions using information coming from the speech and from the video frames :

1. Who is speaking ?
2. Who is present in the video ?
3. What names are cited ?
4. What names are displayed ?

To answer those questions, the sources may be only the speech, only the video frames or a combination of both, as summarized in Table 1.

2.2. Sub-tasks

Answering the four previous questions requires to combine multiple technologies. Some of them, as presented in table 2, are assessed in the REPERE Challenge as sub-tasks.

For example, to determine who is speaking, a speaker diarization system and a speech transcription system may be used. To determine who is present in the video frames a head detection and segmentation system may be useful. To know what

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

TABLE 1 – Tasks and sources

names are cited in speech, a speech transcription system is needed. To determine what names are displayed, an Overlaid Words Text Detection system and an optical character recognition system may be used.

	Who is speaking ?	Who is present in the video frames ?	What names are cited ?	What names are displayed ?
Head detection and segmentation		•		
Speech transcription	•		•	
Speaker diarization	•			
Overlaid words text detection and segmenting		•		•
Optical Character Recognition				•

TABLE 2 – What task for what question

The following sub-tasks which are useful to answer to the four main questions are assessed in the REPERE Challenge :

- Speaker diarization
- Speech transcription
- Head detection and segmenting
- Overlaid words text detection and segmentation
- Optical Character Recognition (OCR)

People who are interested in the REPERE Challenge and decide to participate to the official campaign will have access to the REPERE Corpus which is described in the next section.

3. The REPERE Corpus

3.1. Sources

The January 2012 dry-run corpus represented 3 hours of development data and 3 hours of evaluation data and is described in Table 3.

More training data and new evaluation data will be provided for the 2013 and 2014 evaluations for a total volume of 60 hours at the end of the project. At this point, the video are selected from two french TV channels, BFM TV and LCP, for which ELDA has obtained distribution agreements. The shows are varied, as shown in Table 3.

Top Questions is extracts from parliamentary "Questions to the government" sessions, featuring essentially prepared speech.

Ca vous regarde, *Pile et Face* and *Entre les lignes* are variants of the debate setup with a mix of prepared and spontaneous but relatively policed speech.

LCP Info and *BFM Story* are modern format information shows, with a small number of studio presenters, lots of on-scene presenters, interviews with complex and dynamic picture composition.

Planete Showbiz is a celebrity news show with a voice over, lots of unnamed known people shown and essentially spontaneous speech. The database consists on different utterances of

Show	Channel	total duration (mn)
BFM Story	BFM	60
Plante 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 3 – TV shows currently present in the corpus

the same show so as to measure the intra-show and the inter-show variability.

These video were selected to showcase a variety of situation in both the audio and video domains. A first criteria has been to reach a fair share between prepared and spontaneous speech. A second one was to ensure a variety of filming conditions (luminosity, head size, camera angles...). For instance, the sizes of the heads the annotators would spontaneously segment varied from 936 pixels^2 to 192,072 pixels^2. Some example frames are given Figure 1.

FIGURE 1 – Some example frames from the video corpus

3.2. Annotations

Two kinds of annotations are produced in the REPERE corpus : audio annotation with rich speech transcription and visual annotation with head and embedded text annotation.

3.2.1. *Speech annotations*

Speech annotations are produced in *trs* format using the Transcriber software [3]. The annotation guidelines are the ones created in the ESTER2 [4] 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.

3.2.2. *Visual annotations*

In complement to the audio annotation, the visual annotation has necessitated the creation of specific annotation guidelines [1]. The VIPER-GT video annotation tool has been selected for its ability to segment objects with complex shapes and to enable specific annotation schemes. The visual annotations consist in the six following tasks :

- Head segmentation : all the heads that have an area larger than 2500 pixels[2] are isolated. Heads are delimited by polygons that best fit the outlines. Figure 2 is an example of head segmentation. It is worth noting that it is head segmentation and not face segmentation. Sideways poses are annotated too.
- 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.
- People identification : The name of the people is indicated. Only well-known people and the people named in the video are annotated. Unknown people are identified with a unique numerical ID.
- Embedded text segmentation and transcription : the transcription of the segmented text is a direct transcript of what appears in the video. All characters are reproduced with preservation of capital letters, word wrap, line break, etc. Targeted texts are segmented with rectangles that fit best the outlines (see Figure 3)
- Named-entities (type "person") annotation in transcripts of embedded texts
- The annotation of appearance and disappearance timestamps : the aim is to identify the segments where the annotated object (head or text) is present.

The visual annotation is conduted on 1,074 key-frames choosen every 10 seconds in average.

3.2.3. *Harmonization of the 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 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.

1. Guidelines are available for participants on the REPERE website. They will be distributed with the REPERE corpus at the end of the project.

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 strategy is to keep the 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.

FIGURE 2 – Polygonal head segmentation

FIGURE 3 – Segmentation example

4. Metrics

4.1. EGER

The main evaluation metric is the *Estimated Global Error Rate* (EGER). This metric is based on a comparison between the person names in the references and in the system outputs. EGER is a solution to take in count the fact that the systems have found the correct number of people. For each annotated frame, i, the list of the names of speaking and/or visible persons is built for the reference on one side and for the hypothesis on the other side. Both lists are compared by associating the names one-on-one, each name being associated at most once.

An association between two identical names, or between two anonymous persons is considered correct.

An association between persons with two different names or between a named person and an anonymous one is a confusion noted C_i. Each person with no association in the hypothesis is a false alarm FA_i, and in the reference a miss, M_i.

A cost is associated to each error type, in our case 0.5 for confusion and 1 for miss/false alarm. Among all possible association sets the one with the lowest cost is chosen. Adding up all these costs gives us the total error count, which is divided by the number of expected names (i.e. sum of the size of the reference lists) to get the error rate.

For N annotated frames, EGER is defined as :

$$EGER = \frac{\sum_{i=0}^{i=N} 0.5 * C_i + FA_i + M_i}{\sum_{i=0}^{i=N} P_i} \quad (1)$$

where P_i is the number of people in the i frame.

This metric, with adapted list building methodologies, is used for four tasks :

- Who is speaking or is present in the video frame ?
- Who is speaking ?
- Who is present in the video frame ?
- What names are displayed ?

4.2. SER : What names are cited ?

The expected answer to the *what names are cited ?* question takes the form of a list of temporal segments to which an identity is associated. Obviously, anonymous identities do not exist in that task. We decided to use the *Slot Error Rate* as a metric. The list reference temporal segments to find is built from the audio and the annotated transcriptions through a forced alignment procedure. The hypothesis and reference intervals lists are then compared, and an error enumeration is built :

- I : For every interval of the hypothesis without an intersection with the reference we count an *Insertion* error, with a cost of 1
- D : For every interval of the reference without an intersection with the hypothesis we count an *Deletion* error, with a cost of 1
- T : For an (hypothesis, reference) interval pair in intersection where the identity is different we count a *Type* error, with a cost of 0.5
- F : For an (hypothesis, reference) interval pair in intersection where the frontiers are different by more than 250ms, we count a *Frontier* error, with a cost of 0.5

Note that a pair can end up counting as both a type and a frontier error. The SER is them computed by cumulating the error costs and dividing by the number of intervals in the reference. In other words, noting R the number of intervals in the reference :

$$SER = \frac{I + D + 0.5 \times (T + F)}{R}$$

4.3. ELDM : Detection and segmentation

The detection and segmentation tasks are regrouped under the *Erreur de Localisation/Detection Moyenne* (ELDM, mean detection/segmentation error) metric, which is parametrized by a local error function M. Noting N the number of annotated frames, $R(k)$ the set of reference zones for frame k and $H(k)$ the set of hypothesis zones, the mean error is defined as the sum of the per-frame errors divided by the total number of zones :

$$ELDM_{head} = \frac{\sum_{k=1}^{N} M(H(k), R(k))}{\sum_{k=1}^{N} |R(k)|}$$

The error function M depends on the task. For the head detection task the error metric turns around whether zones of the hypothesis and the reference have more than 50% overlap. Unmatcheable zones cost one point. More precisely, with the comparators yielding one when verified and 0 otherwise :

$$M_{dh}(R, H) = \sum_{r \in R} (\max_{h \in H} \frac{|r \cap h|}{|r \cup h|}) < \frac{1}{2} + \sum_{h \in H} (\max_{r \in R} \frac{|r \cap h|}{|r \cup h|}) < \frac{1}{2}$$

For the head segmentation metric, the precise amount of pixels in error is taken into account :

$$M_{sh}(R, H) = \sum_{r \in R} \min_{h \in H} \frac{|r \cup h - r \cap h|}{|r \cup h|} + \sum_{h \in H} \min_{r \in R} \frac{|r \cup h - r \cap h|}{|r \cup h|}$$

Finally, for text, keeping the zones separate did not seem to make any sense, so the segmentation metric only is used with the reference set to the union of all regions, and identically for the hypothesis. The total is eventually divided by the number of frames.

4.4. WER and CER : OCR

For the overlaid text transcription task , images, with timecodes, and bounding rectangles are provided and the system is expected to write down into text form the words present in each zone. A zone may span multiple lines as long as they're thematically and graphically homogeneous. The natural metrics for the task are the *Character Error Rate* (CER) and *Word Error Rate* (WER). These are computed by taking the character or word levenshtein distance between hypothesis and reference and dividing by the number of reference elements. Spacing is normalized before evaluation :

- Spaces at beginning and end of line are removed.
- Multiple consecutive spaces are reduced into one.
- Typographic norms are applied to the reference (spaces vs. commas, periods, parenthesis, etc).

At caracters level, end-of-line is considered a character by itself. At words level, words are characters between spaces and line extremities.

4.5. DER

The speaker segmentation task requires to extract the speech from the recordings and split it into speaker-attributed segments. Some segments have overlapping speech and must be associated to all pertinent speakers. The naming of the speakers does not need to be related to their real name, abstract labels are plenty. Two conditions are evaluated : one where each show is considered independant, and one called *cross show* where speakers coming back from one show to another should be labelled identically.

The standard metric for the task is the *Diarization Error Rate* (DER). The metric counts the time in error and divides it by the total reference speech time. The time in error is divided in three categories :

- False alarm, where the hypothesis puts a speaker but nobody actually talks
- Miss, where the reference indicates the presence of a speaker but not the hypothesis
- Confusion, where reference and hypothesis disagree on who the speaker is

The speaker labels being abstract, establishing the confusion time requires some effort. It is done through a *mapping*, where speakers in the reference are associated 1 :1 with the hypothesis speakers. Some may remain unassociated. Among all possible mappings the one that gives the best (smallest) DER is the one chosen for the evaluation. A 250ms tolerance on the reference speaker segment boundaries is taken into account to reduce the impact of the intrinsic ambiguousness of their setup.

4.6. WER : Speech transcription

For the speech transcription task, the systems have to transcribe every word spoken in a show. Segments where speech from multiple people overlap are ignored in the evaluation. The usual ASR metric, the *Word Error Rate*, is similar to the OCR one : a levenshtein distance between the words of the reference and the hypothesis. A normalisation process is used :

- Punctuation removal and downcasing.
- Substitution of dashes by spaces.
- Separation of the words at the apostrophe (l'autre becomes l' autre) except on occasions (aujourd'hui).

Homophones are handled on a case-by-case basis.

5. Dry run first results

A dry run was organized during January 2012. Three consortia participated in the dry run. In the REPERE challenge's participation rules, "participants are free to publish results for their own system but participants will not be allowed to name other participants or cite another sites results without permission from the other site". That is why we do not indicate the consortia's names.

5.1. Corpus description

Table 4 summaries the annotations done on the first six hours of corpus created for that run, 3 hours dev and 3 hours test, and the number of persons that can be found through audio or visual clues.

		Dev	Test
Visual clues	Heads on screen	1,421	1,534
	Words in texts	13,240	14,764
Audio clues	Speech segments	1,571	1,602
	Transcribed words	33,205	33,247
Persons	Head appears on screen	216	145
	Name appears on screen	200	141
	Unnamed seen on screen	177	138
	Speaking	141	122
	Named cited in speech	242	191
	Unnamed speaking	45	33
	Total count of persons	237	171

TABLE 4 – Some number about the REPERE dry-run corpus

In the development set (3h), 45% of the persons to be found have their name appearing on the screen, and 55% have their name cited at some point in the speech. In total 33% of the persons to find are no cited either way, meaning that in the unsupervised condition only 67% of the identities are findable. 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.

The distribution on the test set is 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.

Figure 4 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.

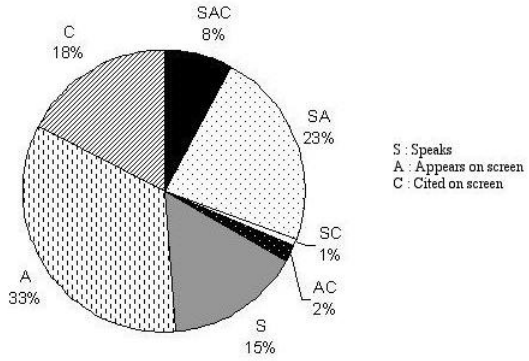

FIGURE 4 – Clues distribution for people recognition

The 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 5.

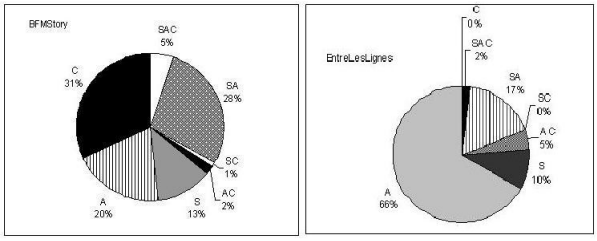

FIGURE 5 – Clues distribution in 2 TV shows

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

In total 351 persons are present in the dry-run corpus, with only 57 common to development and test sets. The per-person speech durations are very uneven, as shows Figure 6. Speech segmentation span from almost 10 minutes down to less than 20 seconds. The situation requires systems robust solutions for when a low amount of data is available.

Regarding the number of head by person in the dry-run corpus, 26% of the people appears only on one video frame when 4% of the people appears on more than 30 video frames. Figure 7 is the people count according to the number of video frames where they appear.

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 8. 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

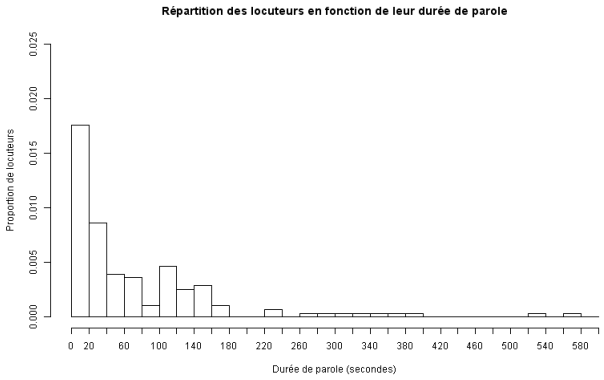

FIGURE 6 – Speakers counts depending on speech duration

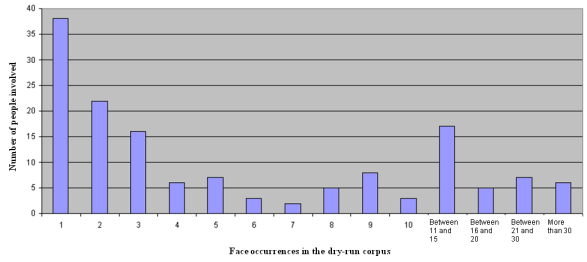

FIGURE 7 – Speakers counts depending on the number of video frames where they appear

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 9.

Another important element included in heads attributes is the presence of objects that can hide a part of the segmented head. Figure 10 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 10).

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.

5.2. Dry run results

Participants usually submitted multiple runs for each task. In this part, we present the first results obtained so as to show the difficulties of the tasks.

We present the results obtained to answer the five main questions.

– Who is speaking or is present in the video frame ?

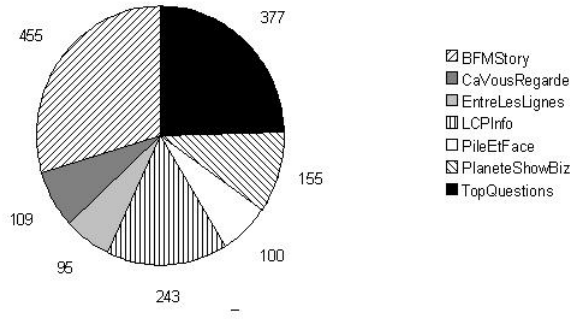

FIGURE 8 – Head distribution

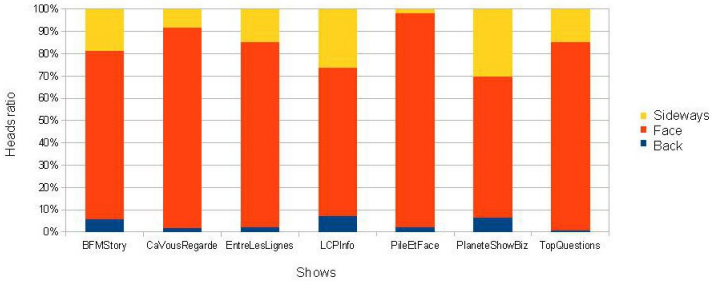

FIGURE 9 – Head orientation

– Who is speaking ?
– Who is present in the video frame ?
– Who is cited in speech ?
– What names are displayed ?

We present the results obtained for the sub-tasks so as to underline the difficulty of the systems combination.

5.2.1. *Who is speaking and who is visible in the video frame*

Global results

In the supervised condition the global EGER ranges between 43.0% and 63.9%. These results show that the task is not too hard, while leaving room for progress. It is interesting to note that the best EGER for speaking person detection (19.1%) is clearly better than the EGER for shown person detection (51.8%), as show in Figures 12 and 13. The head detection and identification is clearly the hardest part at this point. That effect is compounded by the relatively large ratio of persons seen on screen but not talking.

Speaker Diarization

The first step to know who is speaking is to conduct a speaker diarization task. The systems are assessed in two conditions : one where each show is considered independant, and one where speakers coming back from one show to another should be labelled identically.

With independant shows (standard measure), the DER varies from 14.12 to 17.14% according to the system assessed. These results are comparable to ESTER results [4]. These results are in line with the EGER scores on the audio side.

In cross-shows, the DER varies from 36.63% to 64.79% according to the system assessed. The task in the cross-show condition is more difficult than the task with independant shows. These results confirm that it is important to have a mea-

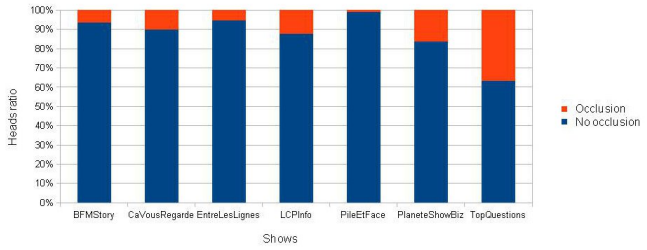

FIGURE 10 – Distribution of hidden head in TV shows

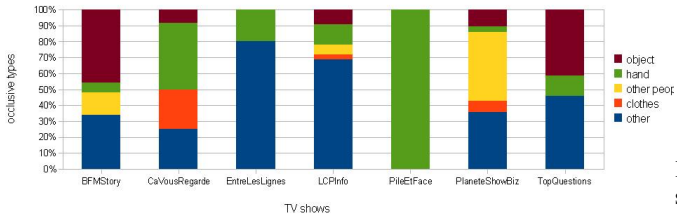

FIGURE 11 – Distribution of hiding objects in TV shows

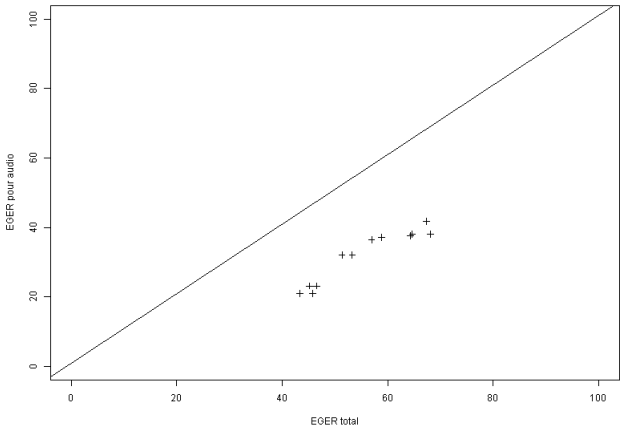

FIGURE 12 – Global EGER vs speaking-person EGER for each system

sure that take into account the intra-speaker variability. The difficulty of the task increases using different shows are recorded at different times.

The results obtained in speaker diarization show that the systems do not have too many difficulties to determine where a speaker speaks in a given file but the systems answers are not consistent in the different files. Regarding the DER and the EGER results for the *who is speaking* question, the difficulties in this context is to diarize the speaker across the files and, for EGER, to find the correct names for each speaker.

Head detection and segmentation

$ELDM_{Head}$ for head detection fluctuates from 64.7% and 109%. $ELDM_{Head}$ for head segmentation fluctuates from 46.8% et 63.9%. Important variation of performance is observed according to the show. The performance observed for *Ca vous regarde* is very different according to the utterance of the show in segmentation ($\mu = 103\%$ and $\sigma = 51$) and in detection

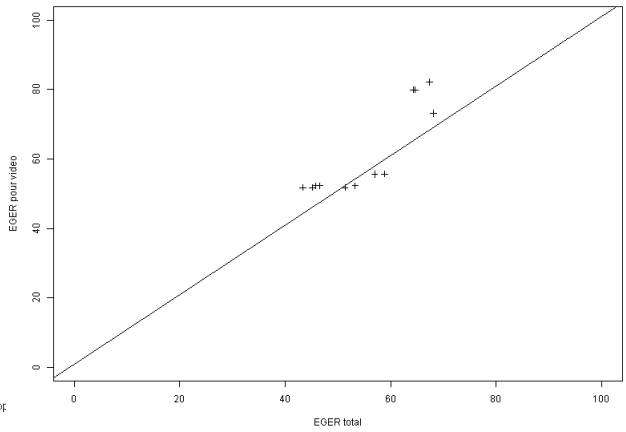

FIGURE 13 – Global EGER vs shown-on-screen EGERfor each system

($\mu = 131\%$ and $\sigma = 54$). It may be explained by the fact that in this show, there are a lot of little head and a lot of people. Head detection may be clearly improved and may explain the EGER results for the question Who is present in the video ?

5.2.2. What names are cited in speech ?

SER results

To assess the ability of the systems to determine what names are cited in speech, the first metric used is the SER. The results are summarized in Table 5.

		System1	System2	System3
SER		83.6%	62.1%	55.3%
Errors distribution	Miss	42%	83%	66%
	Insert	22%	3%	12%
	Bad frontier	15%	2%	5%
	Bad name	10%	7%	11%
	Both bad	11%	5%	6%
Correct answers		37.5%	37.2%	47.2%

TABLE 5 – Errors distribution : The major error is the Miss

For the dry run, the SER ranges between 55.3% and 83.6% depending on the system. The main error made by the systems is miss errors : that kind of error (when the systems miss the person) represent between 83% and 42% of the errors made by the systems. The correct answers ratio fluctuate between 47.2% and 37.2%. The systems favour the accuracy to the recall.

The systems clearly have a room for progress for this task too. Part of the problem may be the well-known propensity of automatic speech recognition systems at incorrectly transcribing proper names. Such analysis have not been done at this point.

Speech transcription

The speech transcription task asks a system to transcribe every word spoken in a show. Segments where speech from multiple people overlap are ignored in the evaluation. The usual ASR metric is the *Word Error Rate* (WER).

The WER varies from 15.62% to 28.96%. These results are in line with to the results obtained in the ESTER Campaings [4] on transcription of radio speech.

It is worth noting that, as expected, the WER fluctuates according to the shows. Figure 14 presents the WER for two systems according to the shows. It can be seen that, as expected from the level of spontaneousness, the WER for the PlaneteShowBiz show is higher than those obtained for the TopQuestion show ($WER_{Best} = 30.7\%$ and $WER_{Worst} = 56.9\%$ for the PlaneteShowBiz show vs $WER_{Best} = 7.19\%$ and $WER_{Worst} = 17.5\%$ for TopQuestion show).

The fact that the quality of the transcription fluctuate according to the show underlines that the explanation of the SER results may depend of the kind of show. The speech transcription is not the only explanation of the errors done though. It will be important to analyze the error of names detection too.

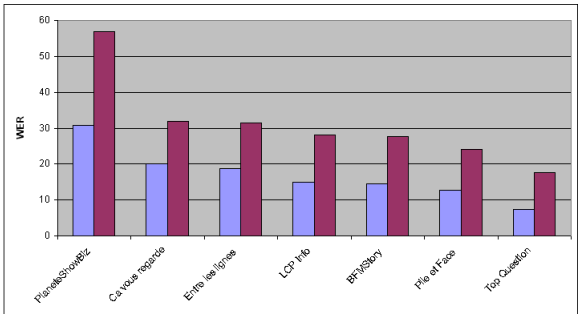

FIGURE 14 – WER for 2 participants according to the show

5.2.3. What names are displayed ?

EGER Results

To assess the ability of the systems to determine what names are displayed, we used the EGER measure.

The EGER varies from 55.6% to 83.4% according to the system. The task is very difficult for the systems assessed. The consortiums have to work on this question so as to obtain better result during the first official campaign.

Text Segmentation

The systems have been relatively successful for the Text segmentation task. The best system obtained 19.7% of $ELDM_{Text}$. The text is correctly detected but the question is to determine if the text has been correctly recognized.

Optical Character Recognition

Taking in count the case and the diacritics, the CER fluctuates from 9.4% to 12.62% when the WER fluctuates from 29.52% and 31.54%. A sligth decrease is observed if the metric is non-sensitive to the diacritics and the case.The performances of the OCR systems may be improved but the most important question is the influence of the error on the name detection system.

6. Conclusions and perspectives

The REPERE Challenge aims to support research on people recognition in multimodal conditions. It focuses on four main questions :

1. Who is speaking ?
2. Who is present in the video ?
3. What names are cited ?
4. What names are displayed ?

Some more usual technologies which are useful to answer these questions are also assessed during the evaluation campaign.

At the end of the project, the REPERE corpus will consist on 60 hours of French video annotated with visual indications (heads and embedded texts) and audio information (transcription, speaker). Seven different kinds of shows are recorded. The database consists on different utterances of the same show so as to measure the intra-show and the inter-show variability. The creation of this corpus is an important step to understand where are the difficulties for the systems.

Metrics have been developped so as to measure the systems progress. The majority of these metrics are the usual metrics except EGER which has been choosen because of its very easy implementation. In future works, the correlation with FA and FR measure may be developped. The goal of this evaluation, is to lead end-to-end evaluation (Tasks) and unitary evaluations (sub-tasks).

The first results show significant progress is quite possible. Part of the task difficulty is that people are sometimes present only a few seconds on screen or in the speech signal. The solutions developed must be robust when presented with a limited amount of data. Moreover, this challenge opens the question of the combination of several technologies. This challenge is an opportunity to develop multimodal solutions to find people.

How to improve the detection of names ? How to merge the relevant information in speech and video frames ? These are some questions that the future campaigns of 2013 and 2014 will attempt to answer.

7. References

[1] A.F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM, 2006, pp. 321–330.

[2] J. Ortega-Garcia, J. Fierrez, F. Alonso-Fernandez, J. Galbally, M.R. Freire, J. Gonzalez-Rodriguez, C. Garcia-Mateo, J.L. Alba-Castro, E. Gonzalez-Agulla, E. Otero-Muras, et al., "The multiscenario multienvironment biosecure multimodal database (bmdb)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1097–1111, 2010.

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

[4] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J-F. Bonastre, and G. Gravier, "The ester phase ii evaluation campaign for the rich transcription of french broadcast news," in *European Conference on Speech Communication and Technology*, 2005, pp. 1149–1152.