The first official REPERE evaluation

Olivier Galibert, Juliette Kahn

Laboratoire national de métrologie et d'essais, Trappes, France

firstname.name@lne.fr

Abstract

The REPERE Challenge aims to support research on people recognition in multimodal conditions. Following a 2012 dry-run [1], the first official evaluation of systems has been conducted at the beginning of 2013. To both help system development and assess the technology progress a specific corpus is developed. It current totals at 30 hours of video with multimodal annotations. The systems have to answer the following questions: Who is speaking? Who is present in the video? What names are cited? What names are displayed? The challenge is to combine the various informations coming from the speech and the images.

Index Terms: REPERE, multimodality, evaluation, fusion, person recognition

1. Introduction

Finding people on video is a major issue when various informations 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 [2] or [3] worked on people multimodal recognition on English databases.

Started in 2011, the REPERE Challenge aims to support 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, the first of two international campaigns has been organized at the beginning of 2013 by the Evaluation and Language resources Distribution Agency (ELDA) and the Laboratoire national de métrologie et d'essais (LNE). The second official campaign is open to external consortia who want to participate in this challenge and will take place at the beginning of 2014.

People who are interested in the REPERE Challenge and decide to participate to the second official campaign will have access to the REPERE Corpus and to the metrics tools.

This paper presents the protocol used to estimate the systems progress and the results of the evaluation. Section 2 describes the different tasks that form the REPERE Challenge. Section 4 presents the data used to assess the systems. Section 3 is dedicated to the metrics description. Section 5 presents an overview of the evaluation results. Section 6, concludes this paper.

2. Questions and tasks

2.1. Main tasks

The first tasks in the REPERE Challenge is the identify 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 (a-priori models of voice and face allowed) and unsupervised modes (a-priori models of voice and face not allowed).

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 as show in Figure 2. Those two tasks are conducted in unsupervised mode.

2.2. Sub-tasks

Answering the four previous questions requires to combine multiple technologies. The following sub-tasks which may be useful are assessed in the REPERE Challenge:

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

During the 2013 REPERE Evaluation campaign, only the Speaker diarization and Speech transcription tasks had system outputs submitted.

3. Metrics

3.1. 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 account 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 is considered correct. An association between persons with two different names 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.

An uniform cost of 1 is associated to every error type. Among all possible association sets the one with the lowest cost is selected. 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.

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

For N annotated frames, EGER is defined as :

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

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

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

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

We also created two variants of the metric. One variant takes the persons the annotators (and systems) were not capable of naming into account. The other builds the lists per-show instead of per keyframe, measuring the capability of the systems as input to a full-show search task.

3.2. 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 500ms, 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:

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

3.3. 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.

3.4. 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 for a small number of exceptions (aujourd'hui).

Homophones are handled on a case-by-case basism through normalization tables and by putting alternatives directly in the reference in some cases.

4. The REPERE Corpus

4.1. Sources

The January 2013 corpus represented 24 hours of training data, 3 hours of development data and 3 hours of evaluation data and is described in Table 1.

The videos are selected from two French TV channels, BFM TV and LCP, for which ELDA has obtained distribution agreements. The shows are varied:

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.

Culture et vous, previously named *Planète Showbiz*, is a celebrity news show with a voice over, lots of unnamed known people shown and essentially spontaneous speech.

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 146 pixels^2 to 96,720 pixels^2 for an image resolution of 720x576. Some example frames are given Figure 1.

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

Show	Train	Dev	Test
BFM Story	7:57:49	1:00:50	0:59:48
Culture et Vous	2:09:28	0:15:00	0:15:03
Ça vous regarde	2:00:05	0:15:39	0:15:01
Entre les lignes	1:59:43	0:15:00	0:15:02
Pile et Face	2:01:26	0:15:04	0:15:01
LCP Info	4:07:09	0:30:08	0:29:56
Top Questions	3:57:41	0:30:02	0:27:01
Total	24:13:23	3:01:46	2:56:55

Table 1: TV shows currently present in the corpus

Figure 1: Some example frames from the video corpus

4.2. Annotations

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

4.2.1. Speech annotations

Speech annotation are produced in *trs* format using the Transcriber software [4]. The annotation guidelines are the ones created in the ESTER2 [5] 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 with a normalized label for each identity.

4.2.2. Visual annotations

In complement to the audio annotation, the video 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 video annotations consist in the six following tasks:

- Head segmentation: all the heads that have an area larger than 1000 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 have 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 2). Also whether a text is part of an identification cartouche is also annotated.
- 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.

Figure 2: Segmentation example

4.2.3. Global annotations

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.

In addition two per-person annotation are provided for both video and audio: the gender of the person, and its role in the show under a 5-class taxonomy.

4.3. First evaluation corpus

Table 2 summaries the annotations done on the 30 hours of corpus created for that run, and the number of persons that can be found through audio or visual clues.

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

		Train	Dev	Test
Visual	Heads seen	13188	1534	2081
	Words seen	120384	14811	15844
Speech	Segments	12833	1602	1514
	Words	275276	34662	36489
Persons	Seen known	725	146	141
	Speaking known	556	122	126
	To find	811	172	162
	Seen unknown	1907	238	160
	Speaking unknown	1108	163	179
	Names on screen	729	138	160
	Names cited	870	190	161
Clues modalities	Name appears	504	83	83
	Name cited	544	116	101
	Never named	178	39	36
	Not speaking	255	50	36
	Not seen	86	26	21
	Speaking and seen	470	96	105

Table 2: Some number about the REPERE first evaluation corpus

We can see that in the test corpus 51% of the people to find have their name appearing on screen and 62% are introduced in the speech. In practice the OCR is much more reliable than the speech recognition for proper names, making these 51% is primary information source for the global system. Interestingly, 22% of the persons are never named, limiting the reachable level for unsupervised systems.

A number of persons appear only in one modality. In the test 22% are only visible, which is a little lower than in the rest of the corpus, and 13% are only heard.

5. Evaluation results

5.1. Participants

Three consortium participated to the evaluations. SODA is a combination of the LIUM (Computer technology lab of the Université du Maine, France) and the Idiap Research Institute. QCOMPERE is made of the LIMSI (Computer technology lab for mechanics and engineering sciences), the INRIA research centre Grenoble (Rhône-Alpes, France), the LIG (Computer technology lab of Grenoble, France), YACAST, Vocapia Research, the GREYC (Research group for computer science, image, automatic and instrumentation of Caen, France) and the Karlsruhe Institute of Technology. Finally the PERCOL consortium is composed of the Laboratoire d'Informatique Fondamentale de Marseille (LIF), the Université d'Avignon et des Pays de Vaucluse (UAPV), the Laboratoire d'Informatique Fondamentale de Lille (LIFL) and France Télécom.

5.2. Main supervised task results

The main supervised task is to find who is present and who talks in the videos by any (automatic) means necessary. The anonymized primary results for each consortium are presented in table 3 using the three EGER variants of section 3.1.

We can see that the results are quite close, with around a third of the identities incorrect. Evaluating the task as finding who is present in a given show degrades the results a little but not by much, with interestingly a different loss for different systems.

Declining per media the results for the speaker identification task are presented in table 4.

Partner	Main EGER	With unnamed	Full-show
A	32.9	43.0	34.7
B	27.9	38.0	32.8
C	29.6	37.5	35.0

Table 3: Main supervised task results

Partner	Main EGER	With unnamed	Full-show
A	22.8	23.1	25.5
B	17.6	18.0	21.7
C	17.7	18.5	21.1

Table 4: Speaker identification task results

Unsuprisingly, the results are much better for the speech side of the multimedia problem. Not only speech technologies are more mature but the task is much simpler, speech overlap being rare compared to the presence of multiple persons in the same image. That particularly shows in the results taking into account the unnamed people: it's much easier to detect whether someone is present in the speech and cluster his interventions than detecting persons in the image and clustering their apparitions.

This is confirmed by the person presence in the picture results presented in table 5.

Partner	Main EGER	With unnamed	Full-show
A	41.5	54.2	42.0
B	36.7	50.0	41.5
C	39.8	48.2	45.9

Table 5: Visible person identification task results

The results are as expected much worse than on the audio side, with the unnamed persons being particularly problematic. Image processing is the achille's heel of these integrated systems.

5.3. Main unsupervised task results

The unsupervised variant of the main task still requires the system to identify the persons speaking and present on the screen, but precludes the use of a-priori trained biometric models. The names are to be found in the signal, either pronounced or written on the screen. The results are presented in table 6.

Partner	Main EGER	With unnamed	Full-show
A	39.5	48.2	36.1
B	37.2	45.2	43.2
C	44.2	49.9	50.8

Table 6: Main unsupervised task results

The loss due to the lack of pre-trained biometric models is around 10% absolute, which isn't bad. Especially since 22% of the persons are never named, putting a hard limit to the minimum possible error rate.

We decline the results per media in tables 7 for the speakers and 8 for the persons present on screen.

The system behaviour is similar than for the supervised task, with a higher loss in the speech case showing that acoustic

Partner	Main EGER	With unnamed	Full-show
A	31.8	32.0	25.5
B	26.3	26.9	36.6
C	40.1	42.8	44.1

Table 7: Unsupervised speaker identification task results

Partner	Main EGER	With unnamed	Full-show
A	46.1	57.3	44.4
B	46.4	55.5	48.3
C	47.8	53.9	56.1

Table 8: Unsupervised visible person identification task results

biometric models are currently more efficient than visual biometric models.

5.4. Monomodal task results

The two monomodal tasks aim at measuring the quality of biometric models by asking of the participant to only use them for the identification and avoiding any fusion process. Hence the name *monomodal*, since only the speech signal modality (without ASR) is used for speaker identification, and only the images (without OCR information) is used for visible person recognition. The results are given in tables 9 for speaker identification and 10 for visible person identification.

Partner	Main EGER	With unnamed	Full-show
A	48.3	48.3	54.0
B	44.2	45.2	43.5
C	37.3	37.2	41.0

Table 9: Monomodal speaker identification task results

Partner	Main EGER	With unnamed	Full-show
B	62.2	62.6	65.9

Table 10: Monomodal visible person identification task results

The speaker identification results go from a 36% to a 49% error rate, which shows a good use of what models were pretrained. The visible person identification is worse as expected.

5.5. Speaker diarization

The speaker diarization task consists in detecting the speech segments in the audio and associating them abstract speaker labels, where the same label is used for multiple interventions of the same speaker. Two conditions were evaluated, one where labels are local to an individual show, and the cross-show one where the same label must be used for a speaker recurring in multiple shows. The results are given table 11.

Partner	DER-ind	DER-cross
A	13.70	33.09
B	13.35	16.05
C	11.10	14.20

Table 11: Speaker diarization task results

We can see that the individual show results are quite good, and at the state of the art for this kind of data. Interestingly, with one exception the cross-show results are very close to the individual-show ones. Since not taking the cross-show condition into account would have given error rates in the 60+% the problem really had to be tackled, and it has been done rather succesfully. These good results have made the cross-show diarization in combination with the OCR of names (not evaluated this year) the backbone of the information fusion efforts of the participants.

5.6. Speech transcription

The speech transcription performance is roughly state-of-the-art, as shown in table 12.

Partner	WER
A	28.03
B	16.43
C	15.18

Table 12: Speech transcription task results

The participants did not consider the speech transcription as a reliable primary information source, given how easy it is for an ASR system to make errors on proper nouns. They seem to plan to work on it more for the next evaluation.

The per-show results, table 13, confirm our expectations on the relative shows difficulties.

	A	B	C
Culture et Vous	54.53	34.56	37.87
Ça vous regarde	36.10	21.75	21.14
Entre les lignes	27.83	17.77	14.92
LCP Info	20.76	11.26	10.10
BFM Story	26.69	15.11	13.03
Pile et Face	27.81	16.27	14.34
Top Question	18.33	10.22	9.26

Table 13: Per-show speech transcription task results

6. Conclusions and Perspectives

The REPERE project focuses on identifying speakers and visible persons in multimodal conditions.

Specific metrics has been implemented. Evaluation tools are made available to interested persons to participate in the next evaluation.

30 hours of data have been created for that evaluation. The annotations are rich and useful for both training systems and evaluating their results. The corpus will double in size for the second evaluation, with the amount put aside for the test still to be decided.

The first evaluation has shown that reasonably good results are possible but a large margin of progress exists, especially on the image side. The influence of the types of programs will be discussed soon.

The sub-tasks will be redefined for the next campaign to better meet the developers needs of modular analysis (specially for video treatment)

7. Acknowledgments

This work was funded by the the ANR/DGA Repere project.

8. References

[1] J. Kahn, O. Galibert, L. Quintard, M. Carré, A. Giraudel, and P. Joly, "A presentation of the repere challenge," in *CBMI*, P. Lambert, Ed. IEEE, 2012, pp. 1–6.

[2] A. 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.

[3] J. Ortega-Garcia, J. Fierrez, F. Alonso-Fernandez, J. Galbally, M. Freire, J. Gonzalez-Rodriguez, C. Garcia-Mateo, J. 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.

[4] 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*, vol. 33, January 2000.

[5] 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.