Evaluation is not just a score

Olivier Galibert, Juliette Kahn, Ilya Oparin

Firstname.lastname@lne.fr

MESURES & RÉFÉRENCES
Clés de la COMPÉTITIVITÉ et d'un MONDE PLUS SÛR

Laboratoire national de métrologie et d'essais

Evaluations we are used to

Modular

- Modules are evaluated independently
- Traditional and “simple” metrics
  - WER, TER, BLEU…

Evaluation is about computing an overall score

- Little to no performance and error analysis
- Test data selected to get a global score
  - Application is the primary guide

Evaluations we would like to have

Evaluations that help improving systems

- Provide clues to system performance
  - Weak and strong points
  - Factors of variation
- Comprehensive evaluation report

Evaluations that are not fully blackbox

- Inter-related modules
- Enhanced metrics and evaluation protocols
- Coherent data selection and annotation

In-depth evaluations : explain and help

In-Depth evaluation: an evaluator's view

How to build the perfect evaluation (ha!)

- Task definition and decomposition
- Data definition
- Metrics definition
- Results analysis

Défi REPERE – Building a complex task

Financé par

- Bring together several consortia to develop systems for person recognition in audiovisual data

Head detection

Head Identification

Text detection

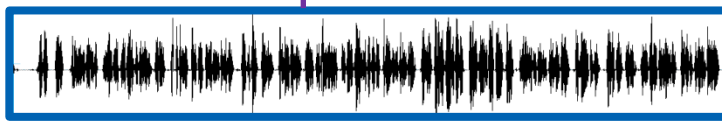
Optical character recognition

Written name recognition

FUSION : Who is in the video?

Speaker recognition

Speaker segmentation



Speech recognition

Recognition of spoken names

Défi REPERE – Building a corpus

Corpus representative for the task

- Classify the difficulties of the corpus
- Composed from documents corresponding to applicative tasks

Annotations

- Identify phenomena to be found
- Annotate potential explicative factors of system errors

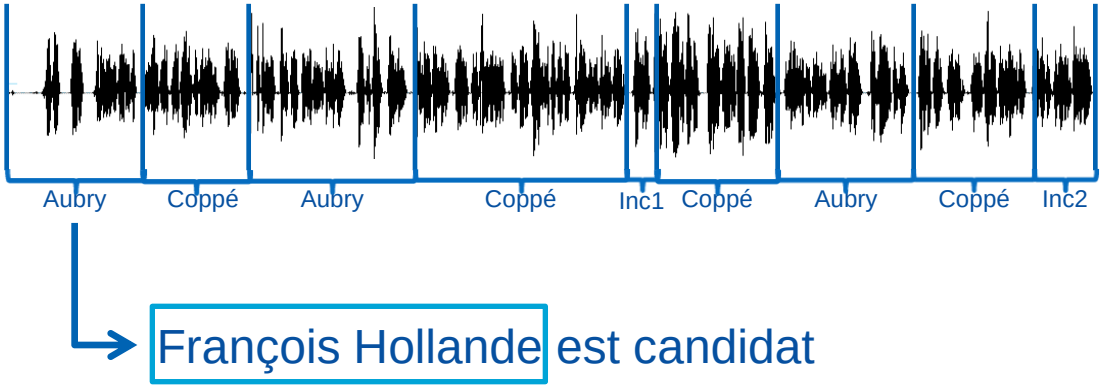
Examples of annotations

Annotations for the tasks

- Speech transcription
- Annotation of persons names in the speech transcription, text
- Identification of the speakers and the heads

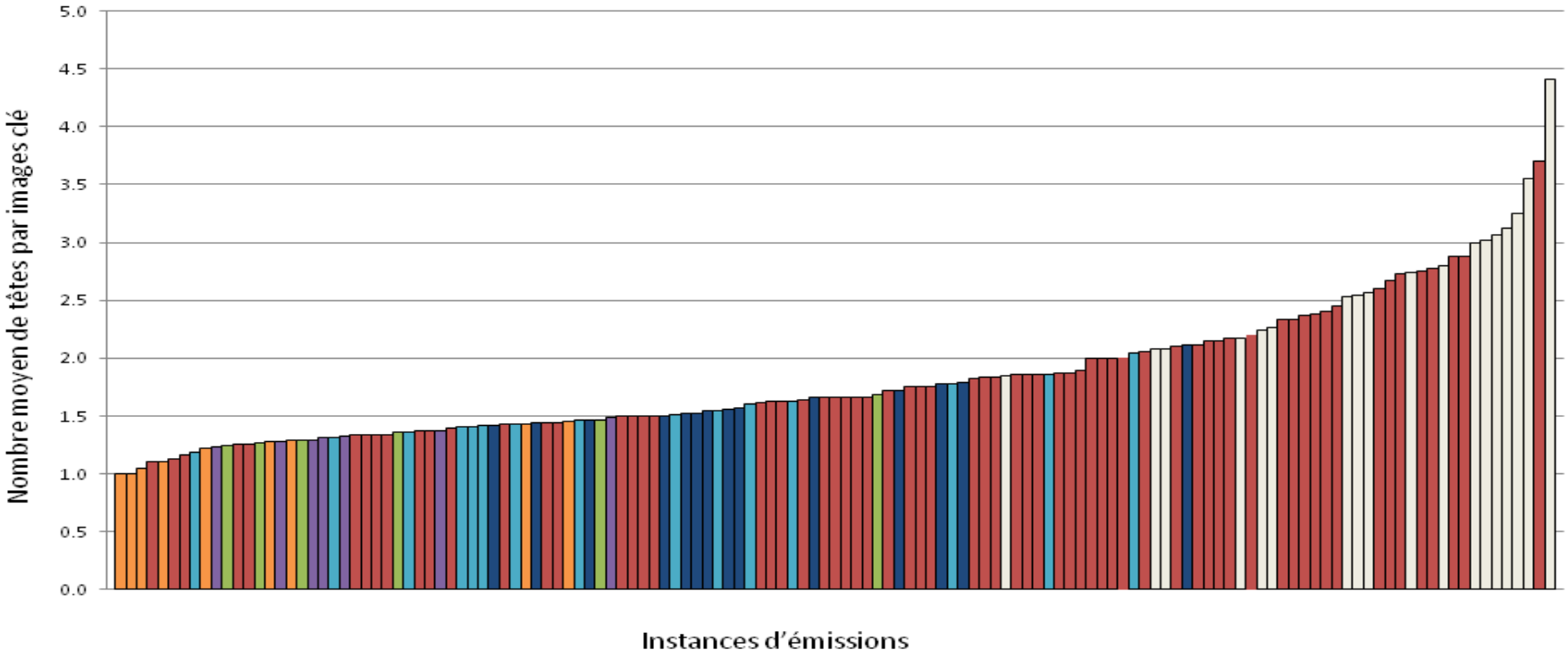
Annotations for the analysis

- Gender
- Persons' roles
- Head descriptions



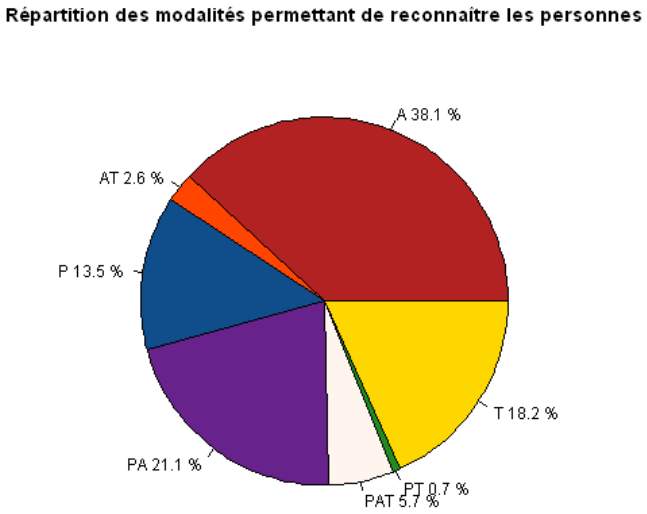
Classifying the difficulties

Number of heads per annotated image



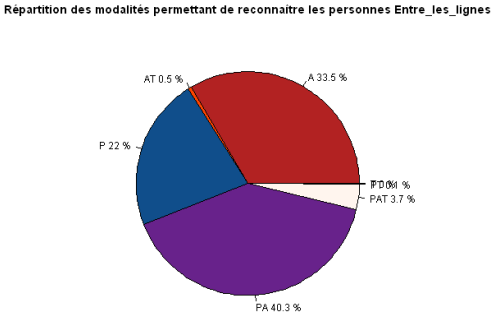
[Kahn J. et al, 2012]

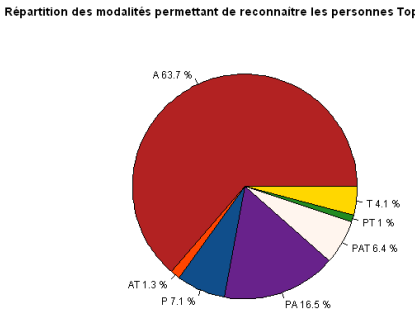
Classifying the difficulties

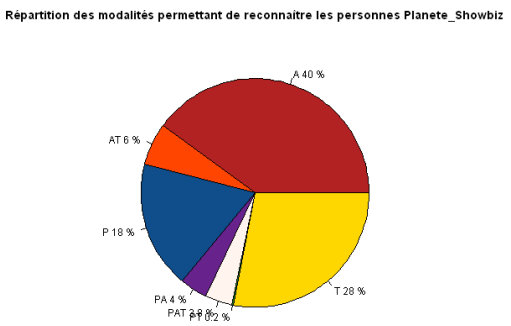


Modalities that can be used

A : Seen
P : Spoken
T : His/her name shown on screen

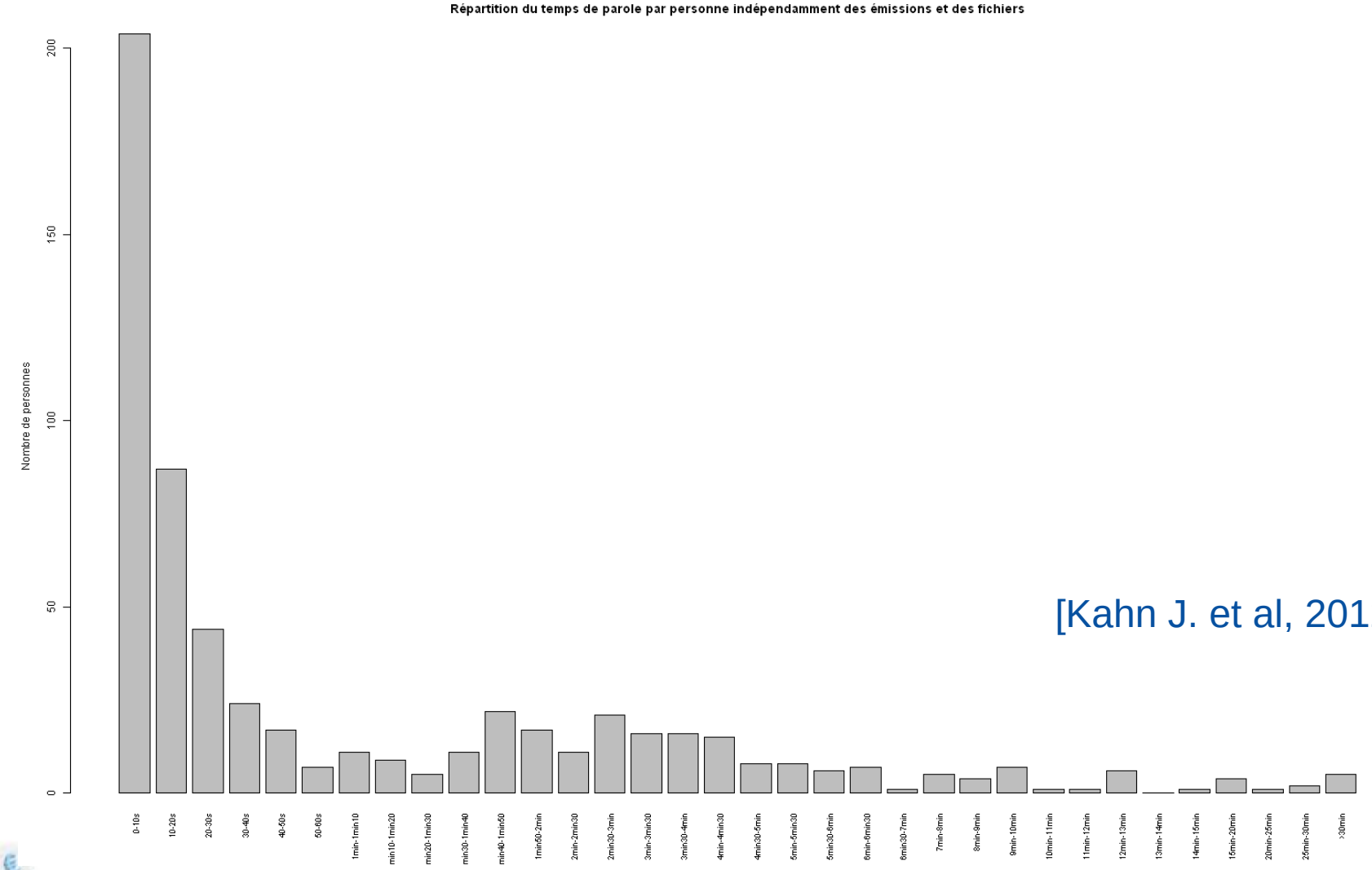






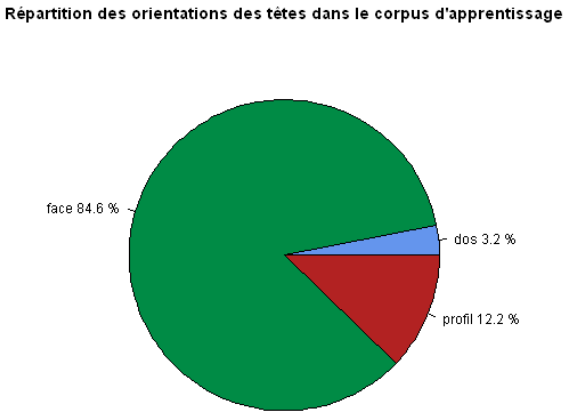
[Kahn J. et al, 2012]

New data, known task or not ?

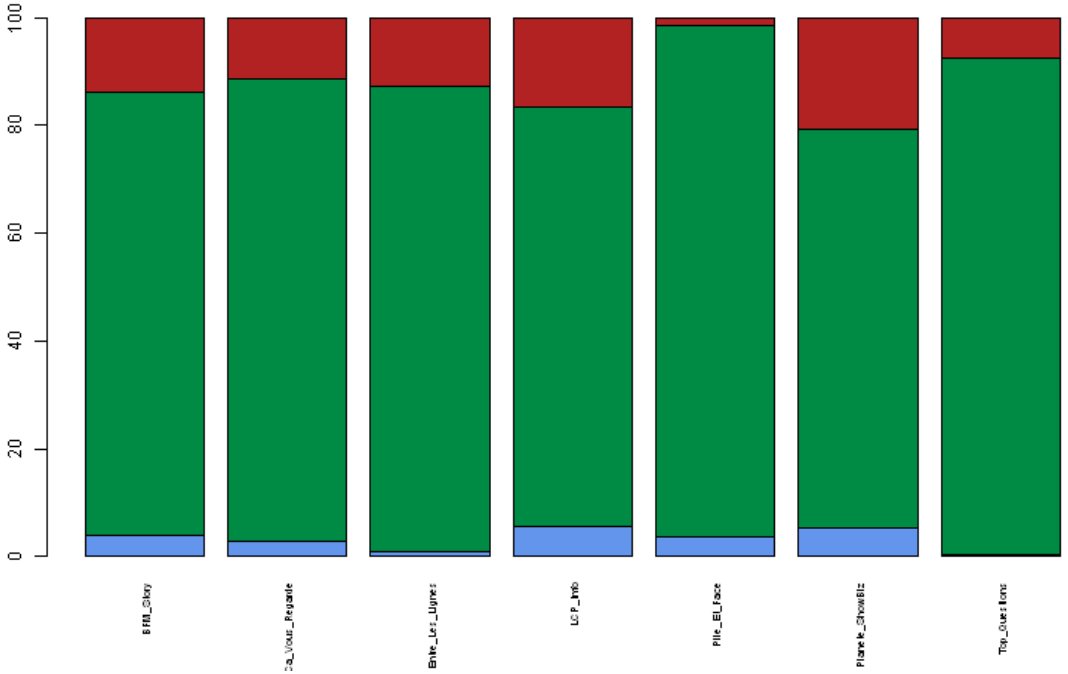


[Kahn J. et al, 2012]

Annotation for factor analysis

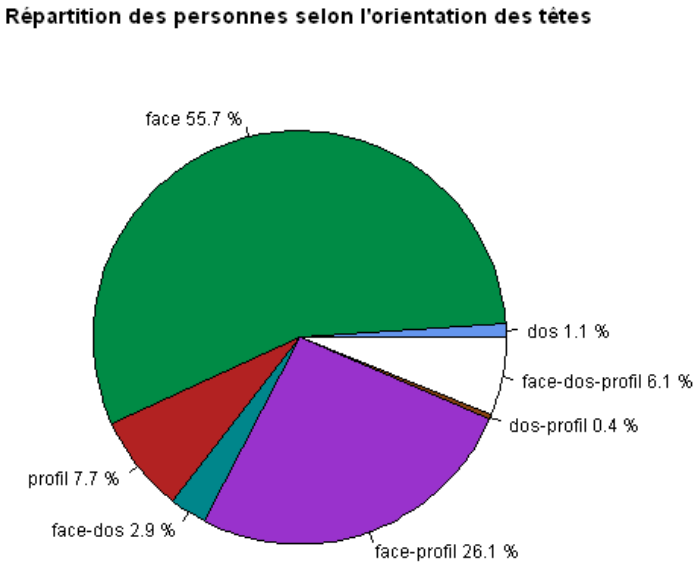


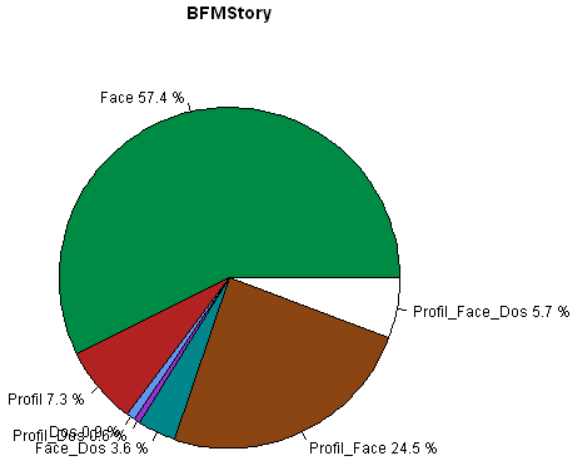
Head orientation : Face, profile, back

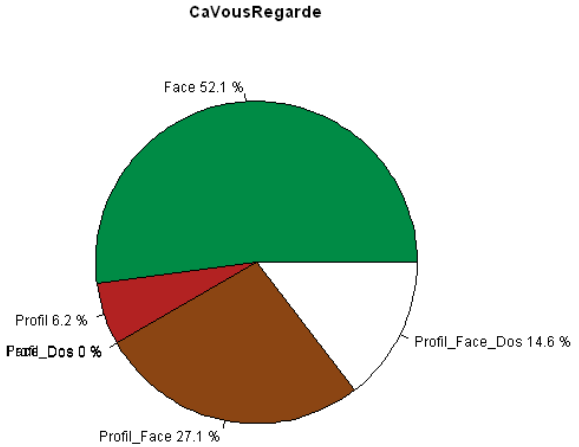


[Kahn J. et al, 2012]

Annotation for factor analysis







[Kahn J. et al, 2012]

Annotation for factor analysis

Marginal phenomenon except for the factor of glasses

Caractéristique binaire	NON	OUI	Pourcentage de OUI
Synthétique	13190	4	0.03%
Lunettes	9 674	3520	26.68%
Coiffes	12987	207	1.57%
Moustache	12981	213	1.61%
Barbe	12940	254	1.93%
Piercing	12434	760	5.76%
Autre	13165	29	0.22%

[Kahn J. et al, 2012]

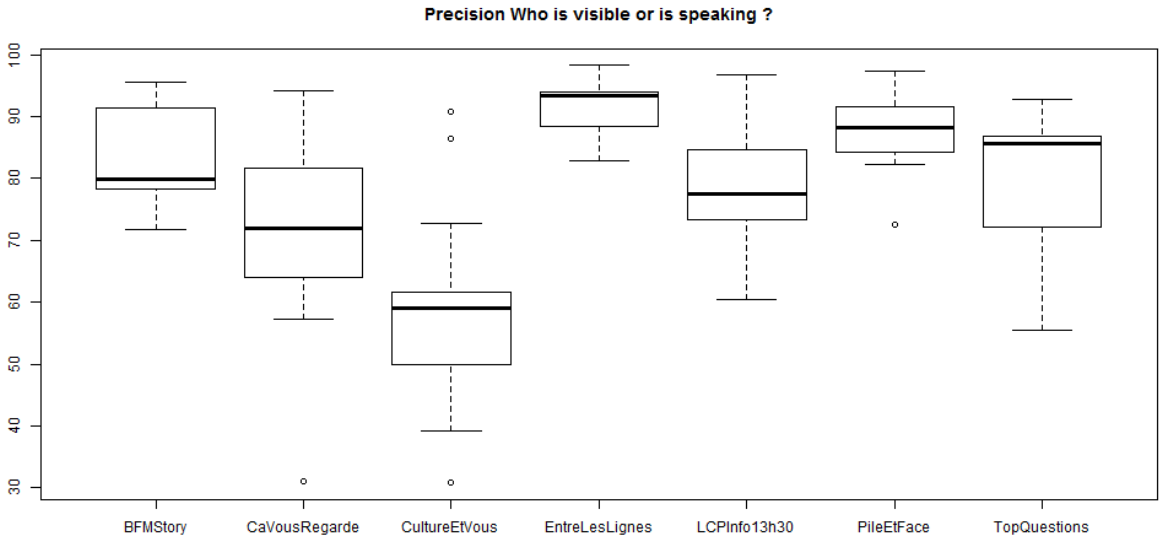
Metrics design

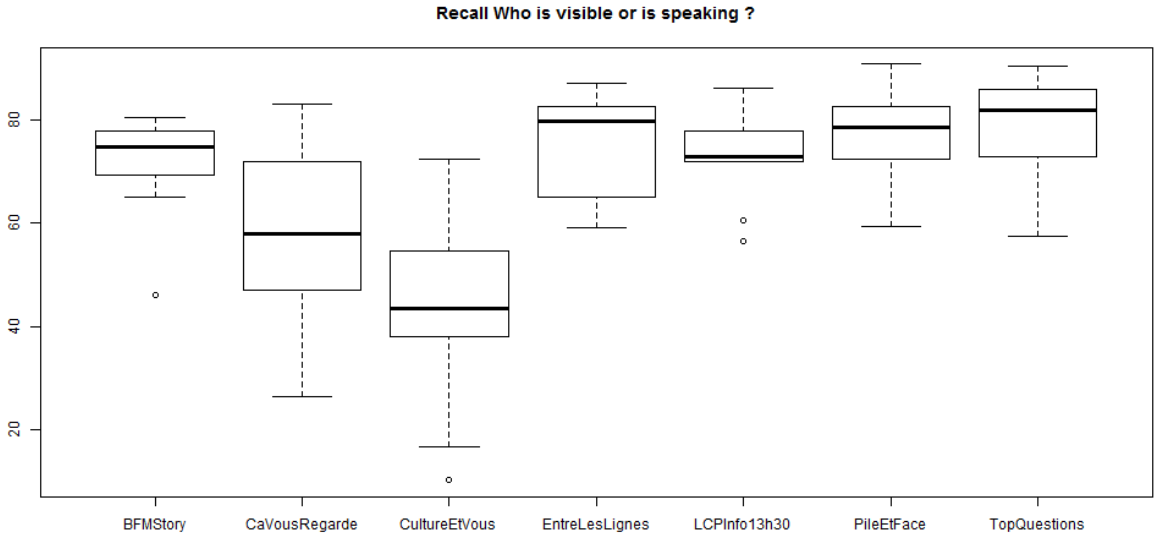
- Some metrics compute a global opaque score, some try to enumerate errors
  - BLEU vs. TER
  - F-measure vs. SER
  - Jaccard vs. DetEval/ZoneMap
- Global scores are often correctness scores, which people often feel are more understandable
- Error-enumeration based metrics usually give error rates, which enable classification and importance tweaking
  - Classes are often a little large though

Error analysis

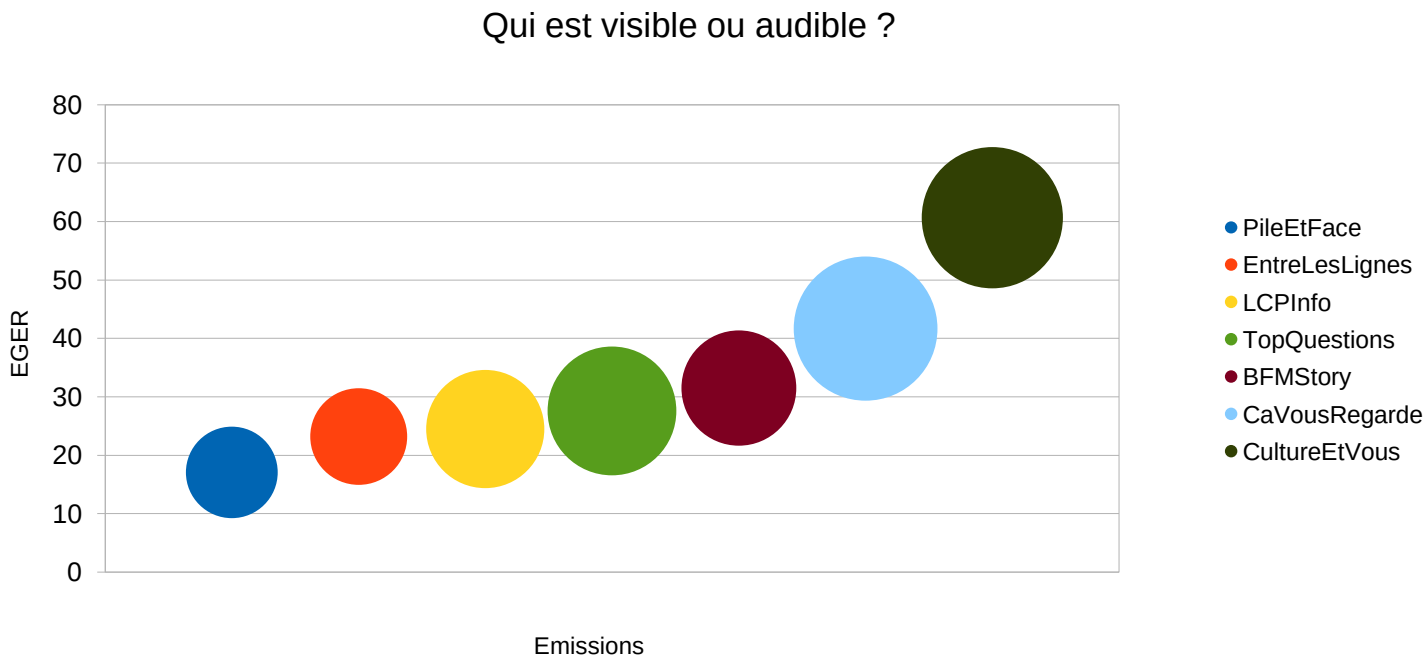
- Try to classify errors
- Try to extract explanation factors, discriminative factors between system outputs
- Representation is key… but also very hard
  - Information overload vs. oversimplification
- And data sparseness is quickly a killer
  - You don’t do statistics on a couple of examples…
  - …but building confidence intervals is an open problem

Box plot vs…


Precision Who is visible or is speaking ?
100
90
80
70
60
50
40
30
BFMStory
CaVousRegarde
CultureEtVous
EntreLesLignes
LCPInfo13h30
PileEtFace
TopQuestions

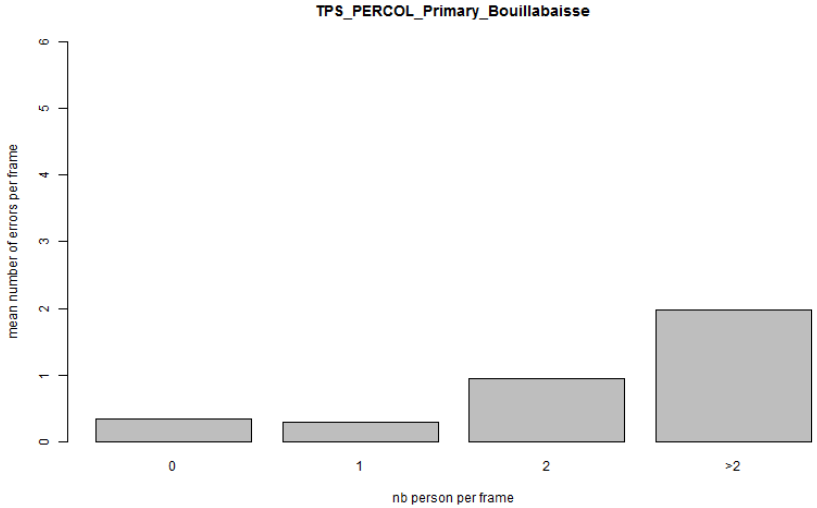

Recall Who is visible or is speaking ?
80
60
40
20
BFMStory
CaVousRegarde
CultureEtVous
EntreLesLignes
LCPInfo13h30
PileEtFace
TopQuestions

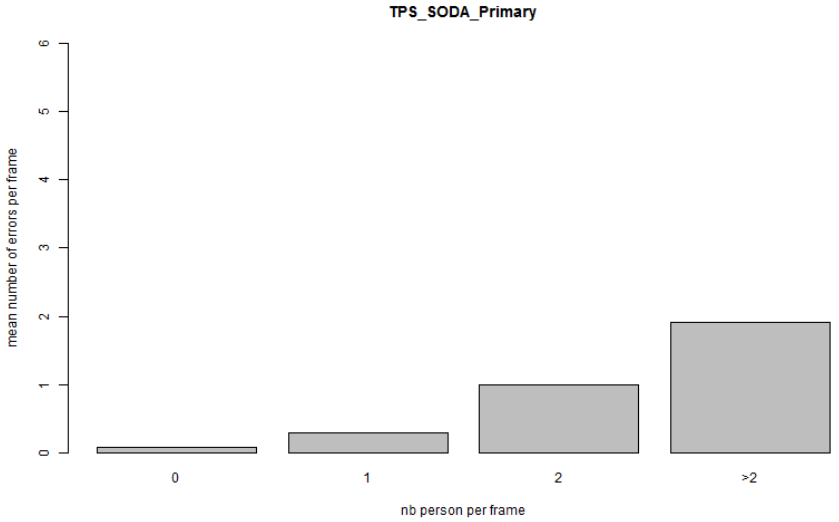
…Mean and standard deviation

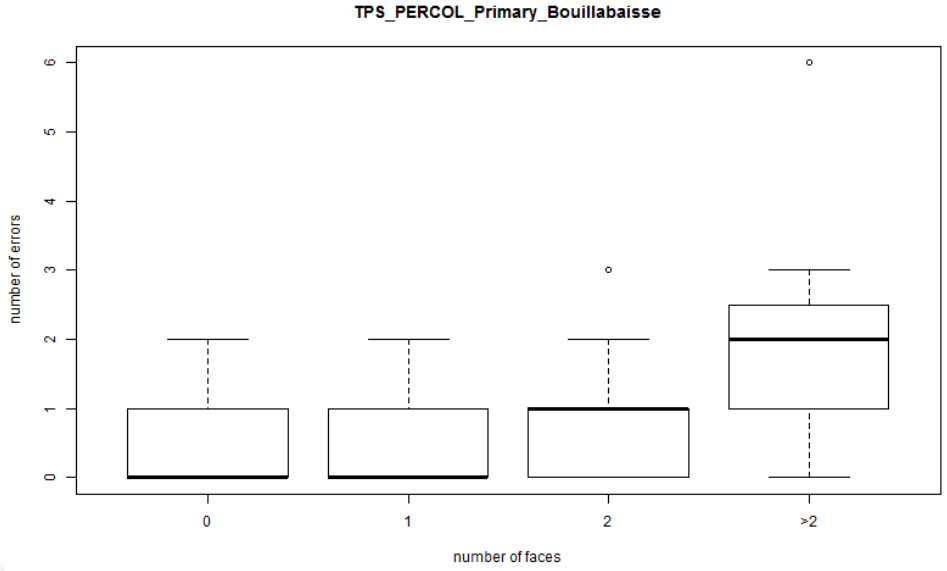


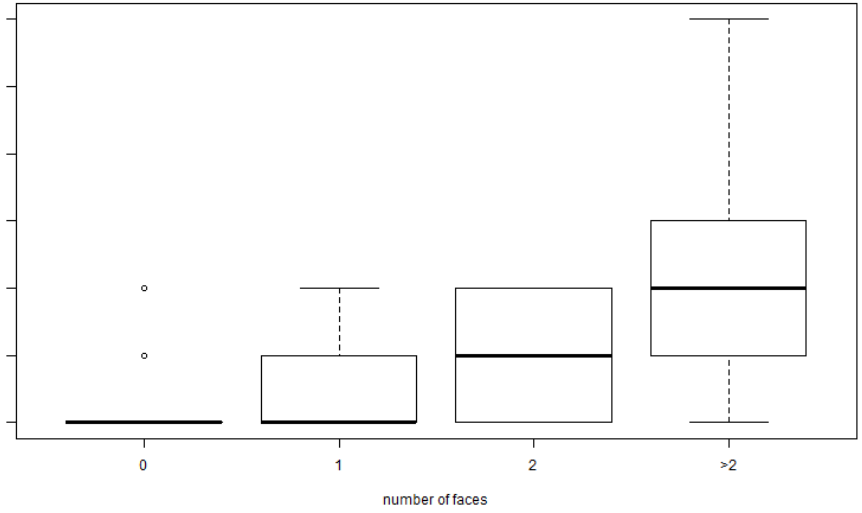
Variation representation

Influence of the number of heads by frame









Conclusions

- Corpus definition
  - Factor-based structure vs. usefulness ratio
  - Annotation for analysis… if there is enough data
- Metrics definition
  - Metrics that enumerate errors
  - Tweakability
- Evaluations
  - Factor analysis and sparsity
  - Pertinence of representations

Thanks for your attention

Firstname.lastname@lne.fr

MESURES & RÉFÉRENCES
Clés de la COMPÉTITIVITÉ et d'un MONDE PLUS SÛR