QCompere @ REPERE 2013

BY quaero

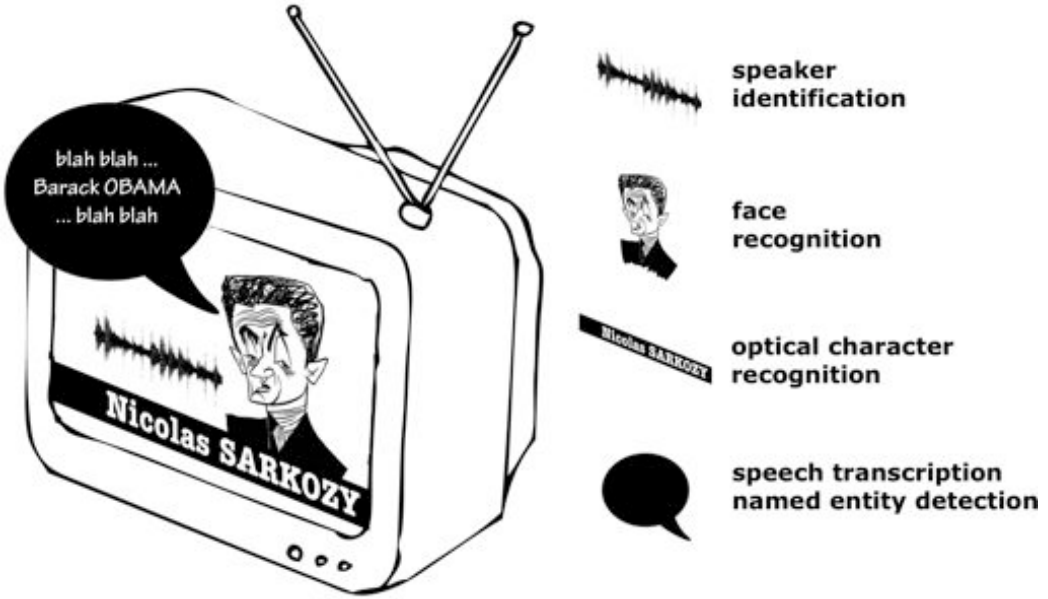
Hervé BREDIN *et al.*

bredin@limsi.fr
herve.niderb.fr

REPERE challenge

"Who speaks when?"

"Who appears when?"



QCompere consortium

BY quaero

CNRS LIMSI

L I G

KIT Karlsruhe Institute of Technology

VOCAPIA research

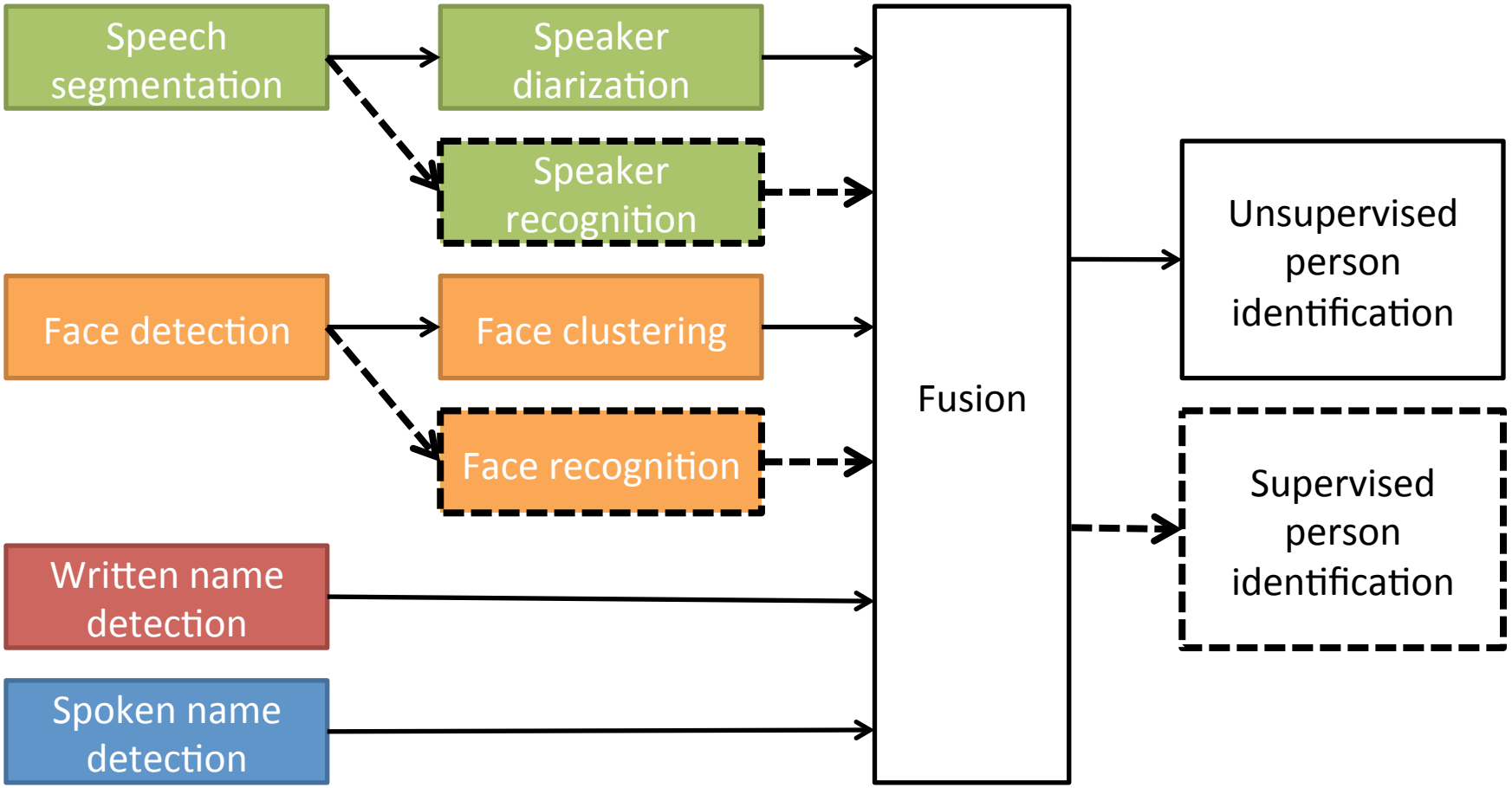
informatiques mathématiques Inria

UNICAEN université de Caen Basse-Normandie

yacast

Multimodal fusion

Supervised *vs.* unsupervised identification



Outline

- Monomodal approaches
  - Speaker recognition
  - Face recognition
- Person name detection
  - Written name detection
  - Spoken name detection
- Multimodal fusion
  - Propagation-based fusion
  - Graph-based fusion
  - Classifier-based fusion
  - Hybrid fusion

Mono-modal person recognition

Speaker diarization

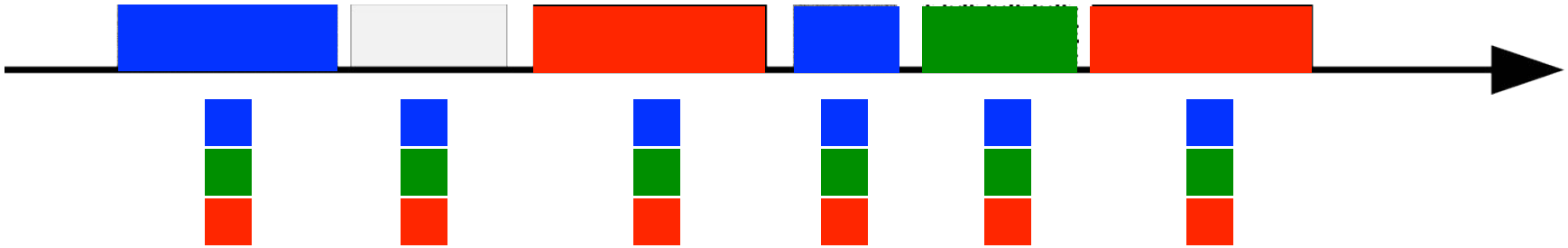
- Speech activity detection
- Divergence-based segmentation

- Two-steps speech turns clustering
  - BIC agglomerative clustering
  - CLR agglomerative clustering

- Cross-show speaker diarization

Mono-modal person recognition

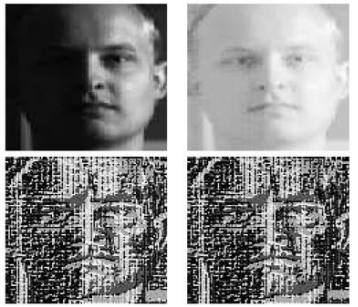
Open-set speaker identification



- GMM-UBM approach
  - Gaussian Mixture Models with 128 diagonal Gaussians
  - by MAP adaptation of gender-matching UBM
  - Log-likelihood ratio
- GSV-SVM approach
  - Mean supervectors of the adapted GMM with the UBM
  - Support Vector Machine with linear kernel
  - Negative samples come from other target speakers

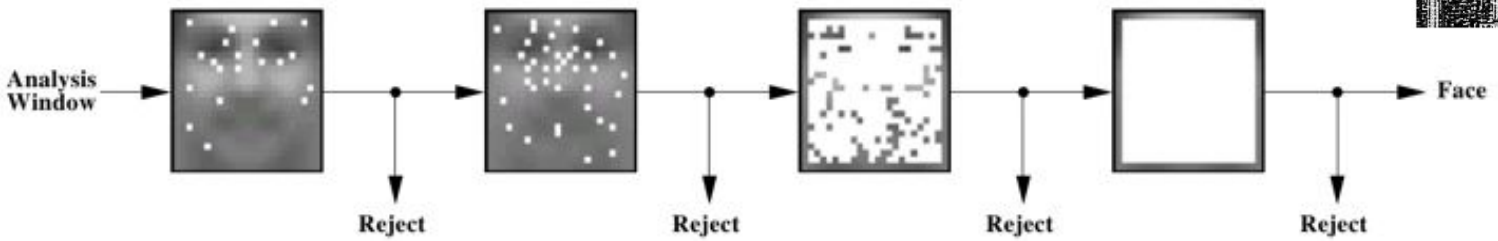
Mono-modal person recognition

Face detection & tracking



Feature: Modified Census Transform (Lighting conditions)

Classification: Boosting Cascade



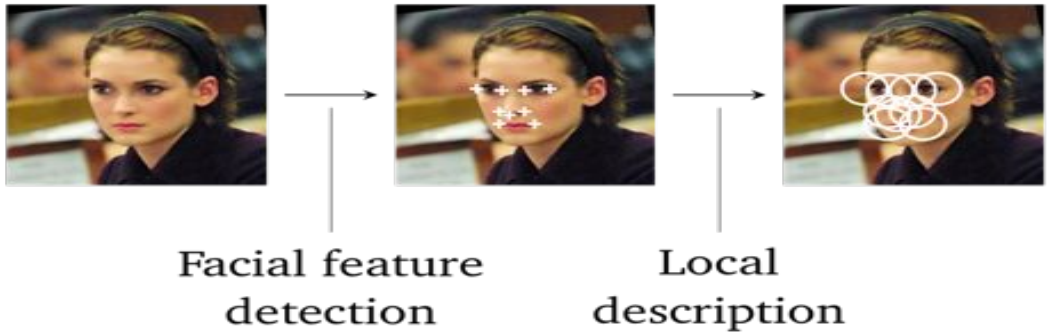
Tracking by Detection, using Particle Filter

State = {x, y, s, α}; Location (x, y); Face size (s) and Angle (α)

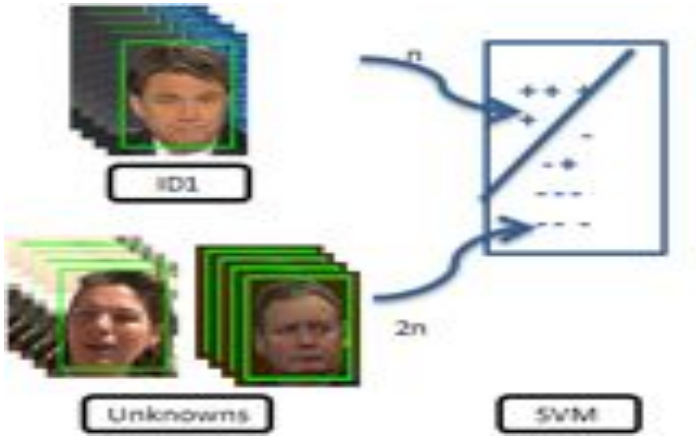
Mono-modal person recognition

Face recognition

- Frontal face descriptor using Histogram of Gradient



- Support Vector Machine in One-vs-All for classification



TRAINING phase

TESTING phase

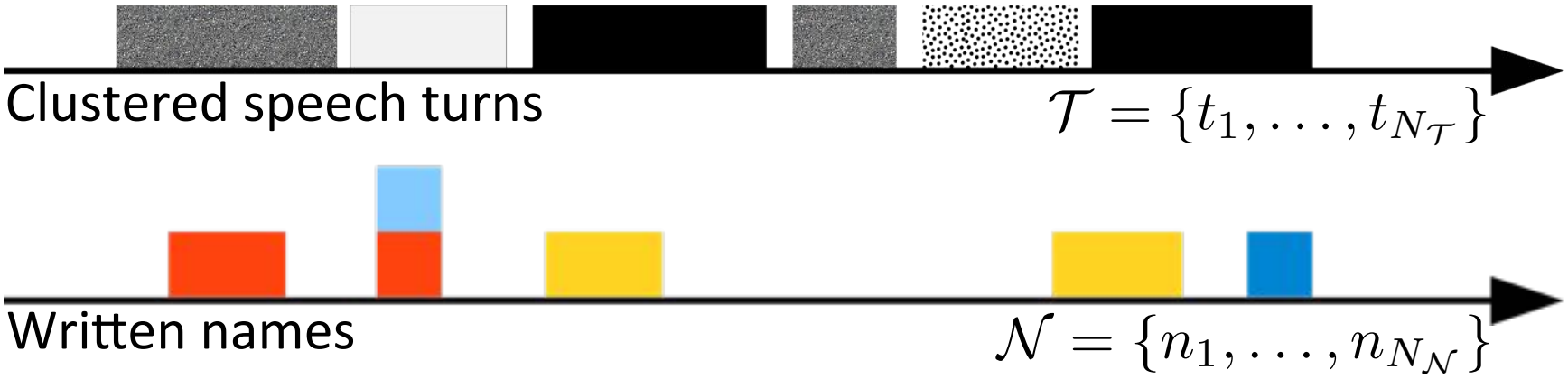
Cross-modal person recognition
Written name detection

- Video Optical Character Recognition
  - LOOV: LIG Overlaid OCR in Video
- Detection of title blocks
  Layout depends on the show
  - Use Wikipedia list of 175k names to learn spatial position of title blocks
- Detection of person names
  Blocks may contain other text
  - Filter false positives based on linguistic rules

Cross-modal person recognition
Spoken name detection

- Automatic speech recognition
  - state-of-the-art off-the-shelf STT system
  - 17% Word Error Rate
- Named entity recognition
  - applied on ASR output
  - based on Conditional Random Field
  - 60% Slot Error Rate (SER)
- Post-processing
  - Look for first or last name to get full names: 53% SER
  - Look for similar names in training set: 52% SER

Propagation-based (late) fusion
for unsupervised speaker identification



Find the best mapping function

$$m : \mathcal{T} \to \mathcal{N} \cup \varnothing$$

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

Propagation-based (late) fusion
for unsupervised speaker identification

$$m^*(s) = \underset{n \in \mathcal{N}}{\operatorname{argmax}} \, \mathrm{TF}(s,n) \cdot \mathrm{IDF}(n)$$

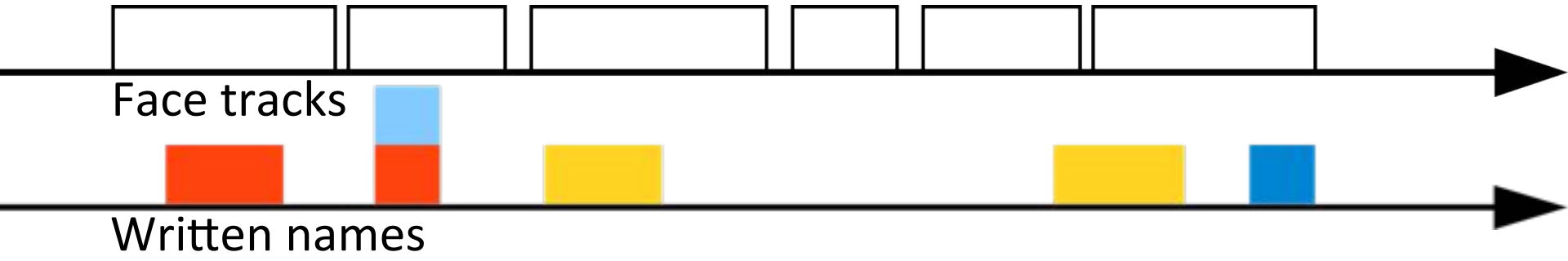
$$\mathrm{TF}(s,n) = \frac{\text{duration of name } n \text{ in cluster } s}{\text{total duration of all names in cluster } s}$$

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

J. Poignant, H. Bredin, V.B. Le, L. Besacier, C. Barras, G. Quénot
Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast
Interspeech 2012

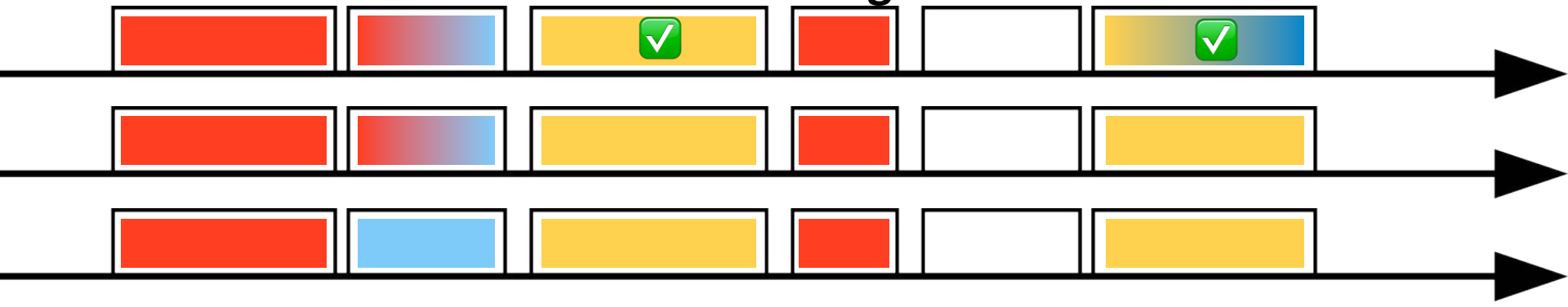
Propagation-based (early) fusion

for unsupervised face identification



1. Direct name propagation

2. Name-constrained face clustering



Graph-based fusion

Person Instance Graph

- Two types of vertices
  - Person instances
    e.g. speech turns, face tracks or written names
  - Person identities
- Three types of edges
  weighted by the probability that vertices are the same person
  - Intra-modal edges
    e.g. speech turn / speech turn
  - Cross-modal edges
    e.g. speech turn / written name
  - Instance/identity edges
    e.g. speech turn / identity

Graph-based fusion

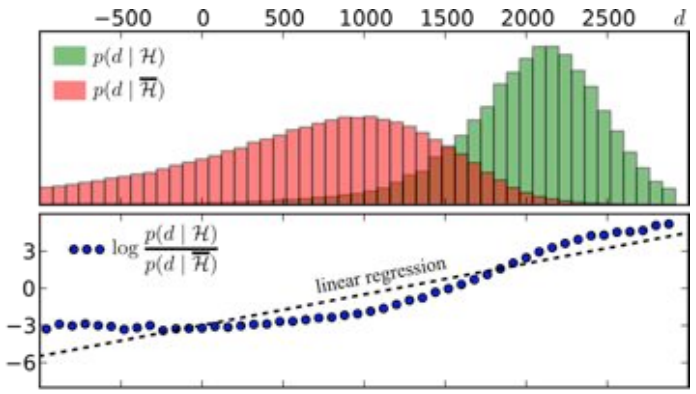
Edges

Intra-modal edges

e.g. between speech turns

- Bayesian Information Criterion as distance between speech turns
- Distance to probability with Bayes' theorem

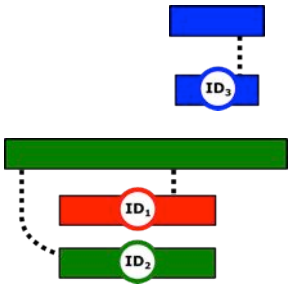
$$p\left(\mathcal{H} \mid d\right) = \frac{1}{1 + \frac{p(d \mid \overline{\mathcal{H}})}{p(d \mid \mathcal{H})} \frac{p(\overline{\mathcal{H}})}{p(\mathcal{H})}}$$



Cross-modal edges

e.g. between speech turn and written name

- Only between cooccurring person instances
- Probability is learned automatically from the training set
- Probability depends on the number of simultaneous "tracks" in both modalities



1 speech turn
1 written name

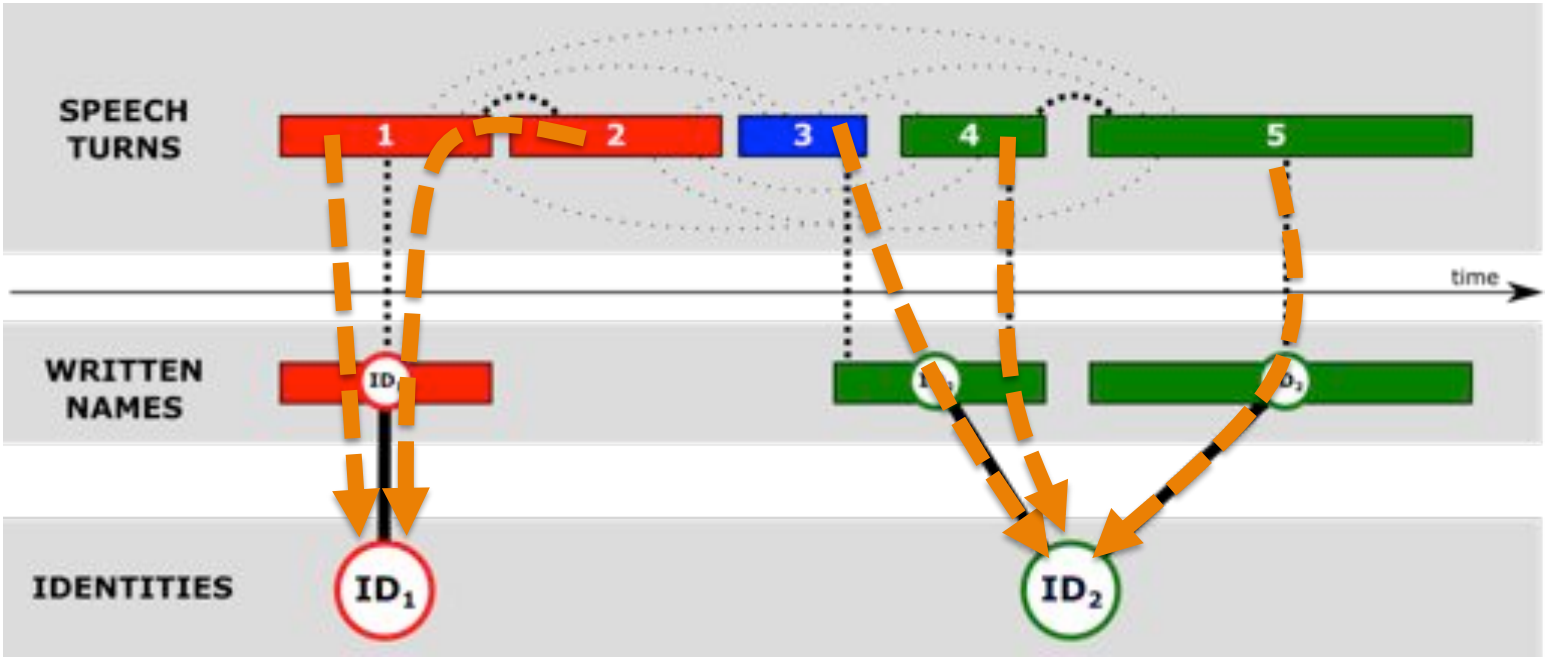
$P_{11} = 0.956$

1 speech turn
2 written names

$P_{12} = 0.996$

Graph-based fusion

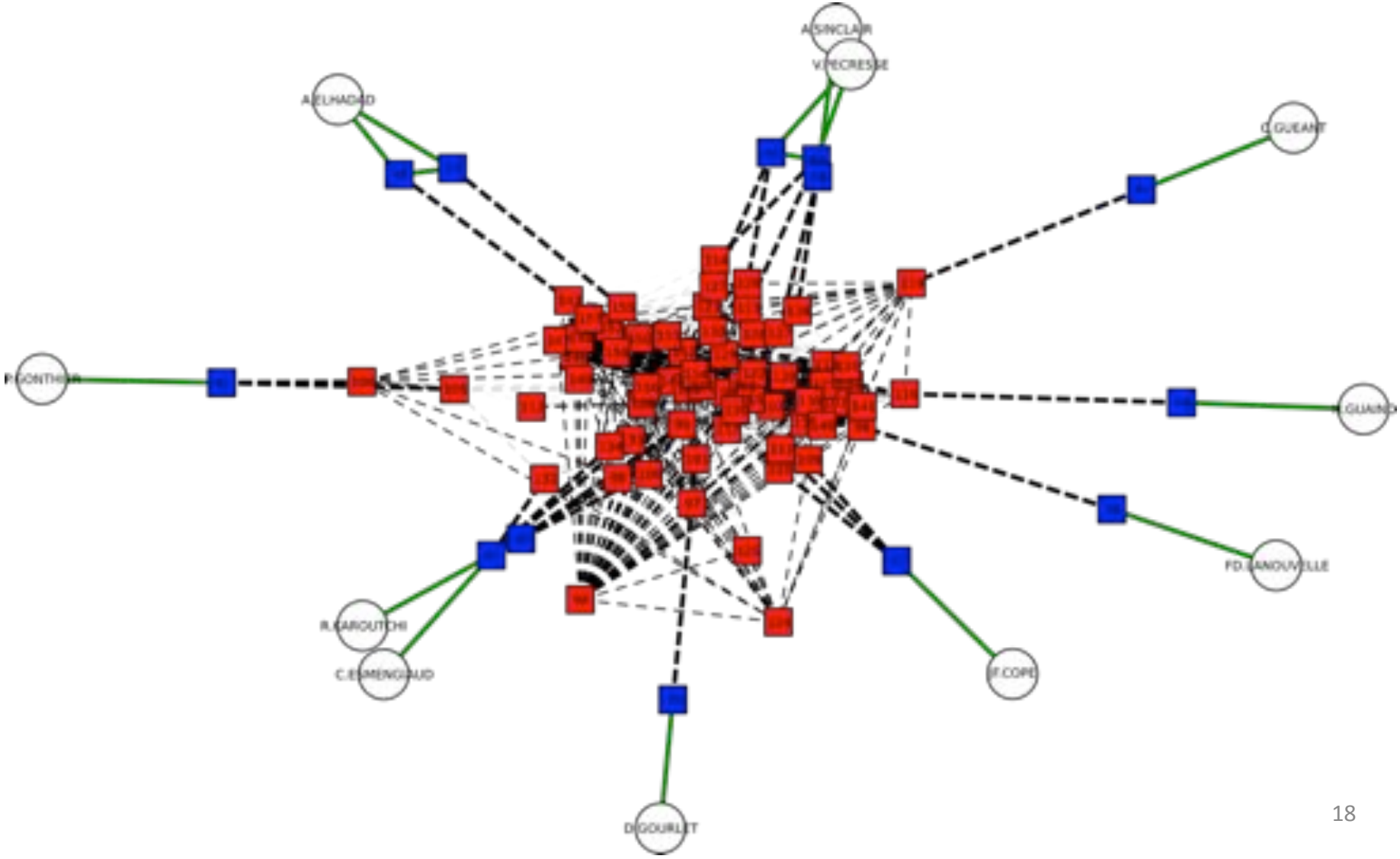
Maximum Probability Path



- For each (speech turn, identity) pair,
  - find the maximum probability path
- For each speech turn,
  - find the identity with highest maximum probability path

Graph-based fusion

"Real-life" Person Instance Graph



Results

Unsupervised person recognition

TASK	APPROACH	EGER
Who speaks when?	Propagation-based (late) fusion	26.2 %
	Graph-based fusion	38.1 %
Who appears when?	Propagation-based (early) fusion	46.2 %
	Graph-based fusion	50.3 %

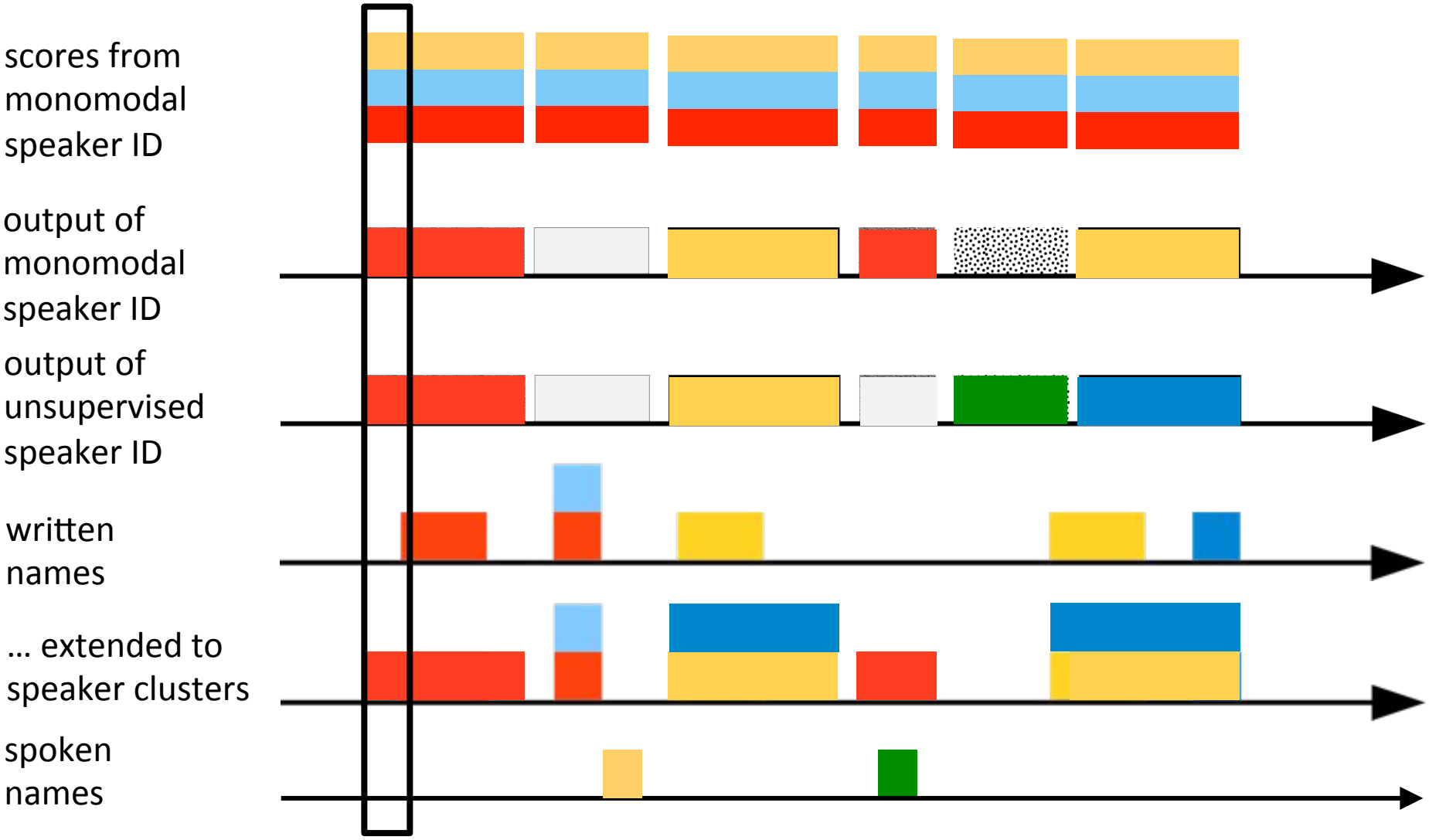
Classifier-based fusion
Multilayer perceptron

- Binary classification problem
  - Train a multilayer perceptron to answer the following: « *is speaker S speaking at time T?* »
  - Select the identity S with the highest score

- Short list of potential speaker S
  - from targets of monomodal speaker identification system
  - from detected written names at time T
  - from detected spoken names in previous/next speech turns

Classifier-based fusion

Feature vector



Hybrid fusion
for supervised face identification

- Heuristic cascading approach
  - Supervised monomodal face recognition for anchors only
  - Unsupervised propagation-based face recognition to remaining face tracks
  - Supervised monomodal face recognition for remaining face tracks
  - Propagation of multimodal speaker recognition to remaining face tracks

Results

Supervised person recognition

TASK	APPROACH	EGER
Who speaks when?	Best mono-modal approach	44.2 %
	Classifier-based fusion	17.8 %
	Graph-based fusion	35.3 %
Who appears when?	Best mono-modal approach	61.1 %
	Hybrid fusion	37.3 %
	Graph-based fusion	48.1 %

Conclusion

- OCR- and ASR-based name detection
  significantly improves person identification in TV broadcast

- Unsupervised multimodal *beats* supervised monomodal

- Face recognition is not as good as speaker recognition
  - Focus on face recognition for REPERE 2014
  - Use audio to achieve face recognition

Hervé BREDIN, Anindya ROY, Claude BARRAS, Sophie ROSSET, Achintya SARKAR

Johann POIGNANT, Laurent BESACIER, Georges QUENOT

Guillaume FORTIER, Jakob VERBEEK

Makarand TAPASWI, Qian YANG, Hua GAO, Hazim Kemal EKENEL, Rainer STIEFELHAGEN

Viet Bac LE

Alexis MIGNON