Unsupervised Speaker Identification in TV Broadcast

Hervé BREDIN
Claude BARRAS

Johann POIGNANT
Laurent BESACIER
Georges QUENOT

Viet-Bac LE

AGENCE NATIONALE DE LA RECHERCHE

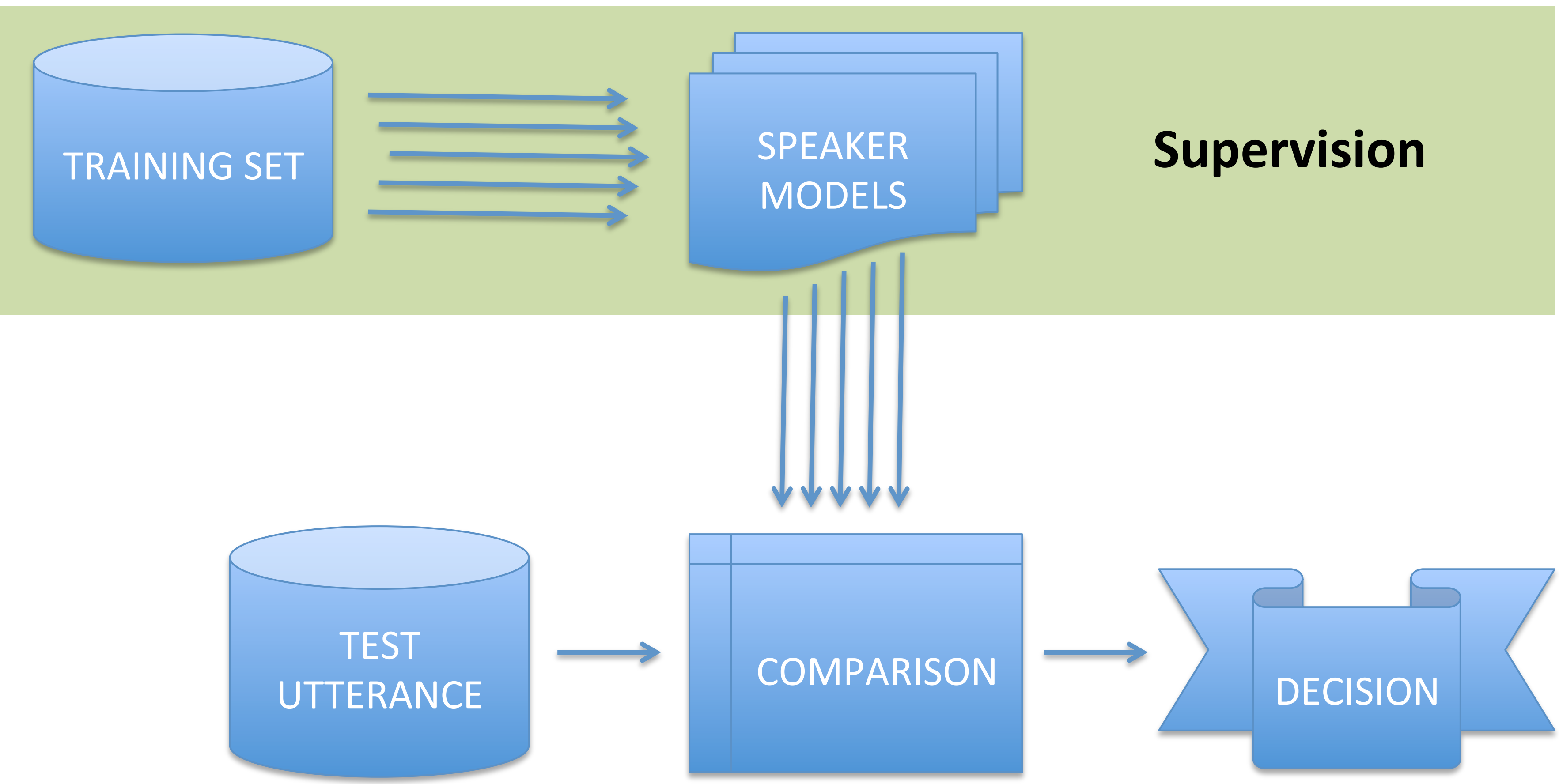
Speaker verification vs. recognition

- Let **S** be a set of **N** speakers $\mathbf{\{s_1, \ldots, s_N\}}$

- Speaker verification (= authentication)
    - « is person $\mathbf{s_i}$ speaking? »

- Speaker recognition (= identification)
  – *Closed-set* conditions
    - « which speaker $\mathbf{s_i}$ is speaking? »
  – *Open-set* conditions
    - « is any speaker $\mathbf{s_i}$ speaking? »
    - « if so, which one? »

``*Supervised*'' paradigm



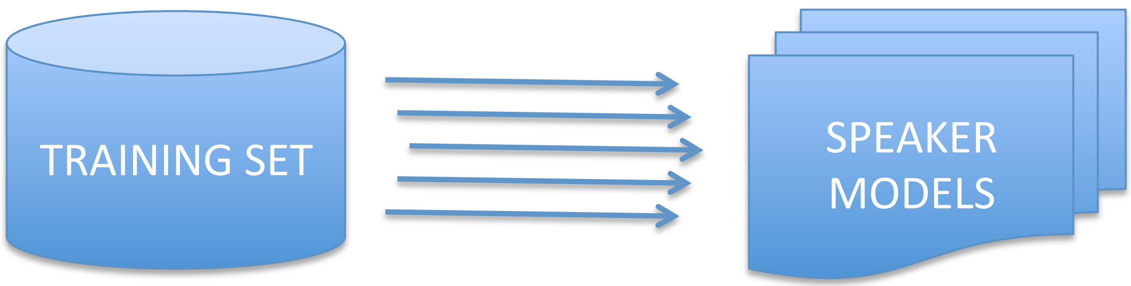
Outline

- Context
- Multimodal & unsupervised identification
  – Overlaid name detection
  – Speaker diarization
  – Name propagation
- Results & discussion
- Conclusion

REPERE challenge

- Speaker identification in TV broadcast
  - ~~*Closed-set* conditions~~
    ~~« which speaker $\mathbf{s_i}$ is speaking? »~~
  - ~~*Open-set* conditions~~
    ~~« is any speaker $\mathbf{s_i}$ speaking? »~~
    ~~« if so, which one? »~~
  - ***REPERE challenge* conditions**
    « who is speaking and when? »

Reaching the limits



600+ hours
manually annotated

535 different speakers

The REPERE test set contains 116 different speakers…

… of which only 57 have a corresponding model

Even a ``supervised oracle'' could not get 100 % accuracy.

Think multimodal

Any source of information can be used.

The REPERE test set contains 116 different speakers…

- … 49 % have a corresponding speaker model
- … 34 % have a corresponding face model
- … **64 % have their name written on screen at least once**

Overlaid name detection

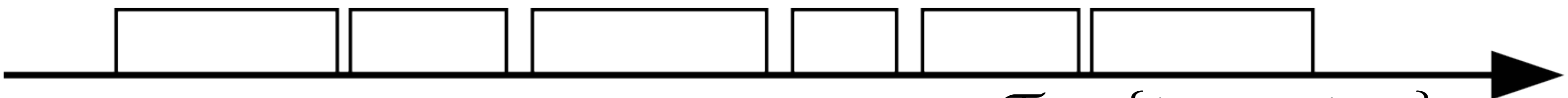
- Text detection
- Optical Character Recognition (OCR)
  - *Tesseract*
    `http://code.google.com/p/tesseract-ocr/`
- Temporal filtering & smoothing
  - `MA1T DAMUN | MATT DAMUN | MA1T DAMON | MATT DAMON`
  - $\Longrightarrow$ `MATT DAMON`

Johann Poignant, Laurent Besacier, Georges Quénot, and Franck Thollard
From Text Detection in Videos to Person Identification
IEEE International Conference on Multimedia and Expo, 2012.

Multimodal fusion

- Speech activity detection & temporal segmentation

  into homogeneous segments (a.k.a. speech turns t)



$$\mathcal{T} = \{t_1, \ldots, t_{N_\mathcal{T}}\}$$ time

- Overlaid name detection

$$\mathcal{N} = \{n_1, \ldots, n_{N_\mathcal{N}}\}$$ time

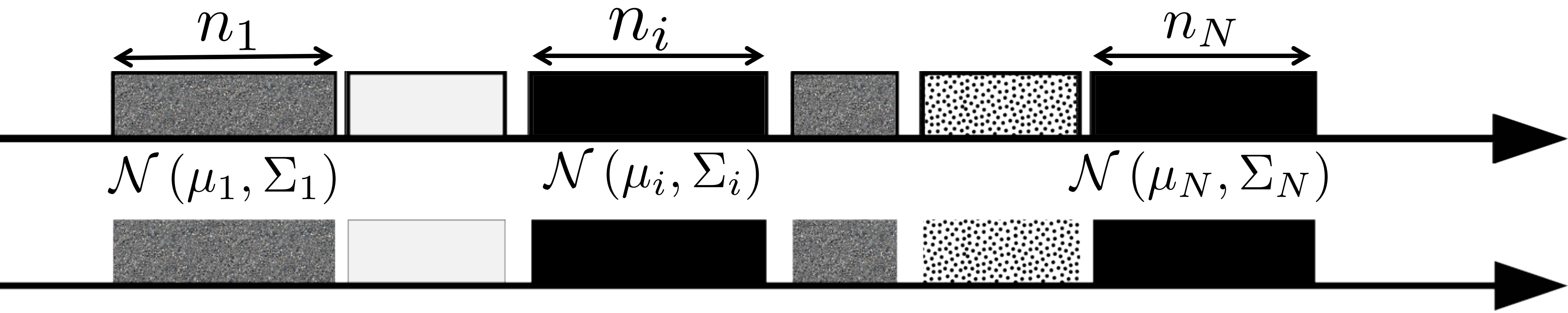
- Name propagation

  find the best mapping function

$$m \colon \mathcal{T} \rightarrow \mathcal{N} \cup \varnothing$$

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

Speaker diarization



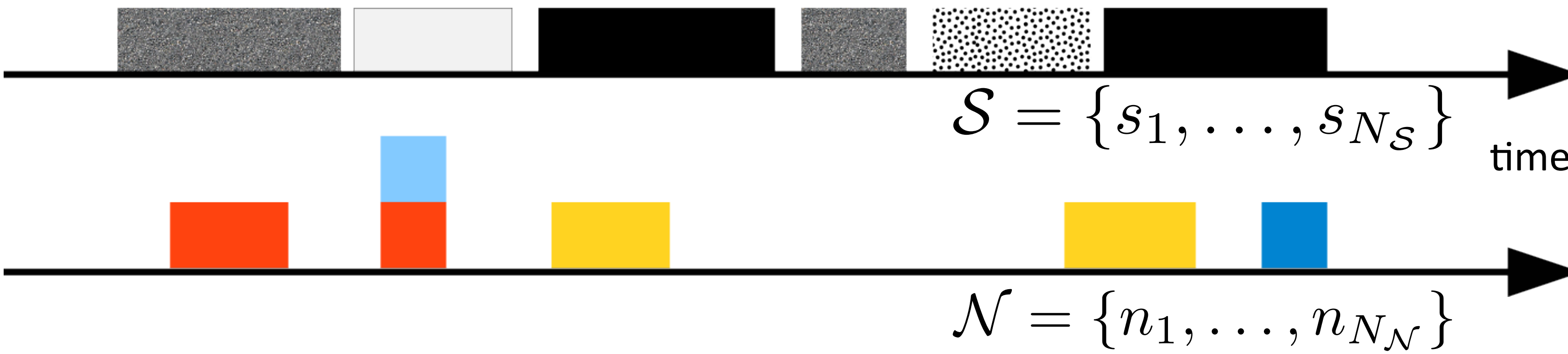
- Agglomerative clustering
  - MFCC coefficients extracted every 10 ms (d=13)
  - Bayesian information criterion (BIC)

$$\Delta\mathtt{BIC} = (n_i + n_j)\log|\Sigma| - n_i \log|\Sigma_i| - n_j \log|\Sigma_j| - \lambda P$$

$$P = \frac{1}{2}\left(d + \frac{1}{2}d\,(d+1)\right)\log(n_i + n_j)$$

  - Stop merging when $\Delta\mathtt{BIC} > 0$

Multimodal fusion



- Smaller (easier?) problem

  find the best mapping function

$$m \colon \mathcal{S} \to \mathcal{N} \cup \varnothing$$

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

1-to-1 mapping / $\texttt{M1}$

- Assumption
  - speaker diarization is perfect
  - one name *n* ⟺ one speaker *s*
- Assignment problem

$$\texttt{M1} = \underset{m:\, \mathcal{S} \to \mathcal{N} \cup \varnothing}{\operatorname{argmax}} \sum_{s \in \mathcal{S}} \mathbb{K}\left(s, m(s)\right)$$

where $\mathbb{K}\left(s, n\right)$ is the cooccurrence duration of *s* & *n*

- Solved using the Hungarian algorithm

Speech turn tagging / `M2`

- Observation
  - when a single name cooccurs with a speech turn, p = 95 % that this is the correct name.

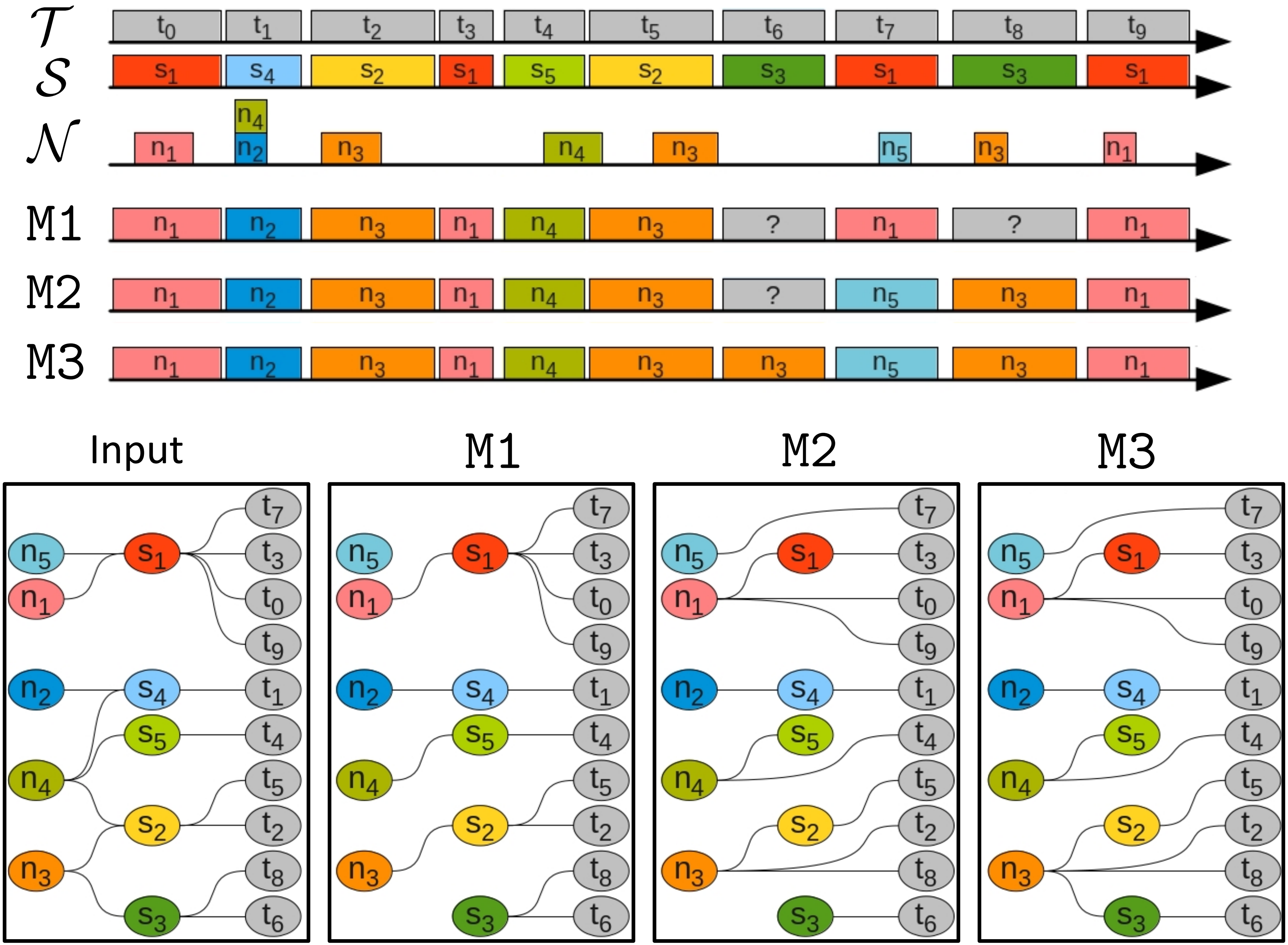
- Principle
  - first, tag unambiguously named speech turns
  - then, apply previous approach on remaining speech turns

1-to-N mapping / M3

- Assumption
  - over-segmented speaker diarization
  - one name $\Leftrightarrow$ **multiple** speaker clusters
- Apply speech turn tagging first
- $f(s) = \underset{n \in \mathcal{N}}{\operatorname{argmax}} \, \mathrm{TF}(s,n) \cdot \mathrm{IDF}(n)$

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

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



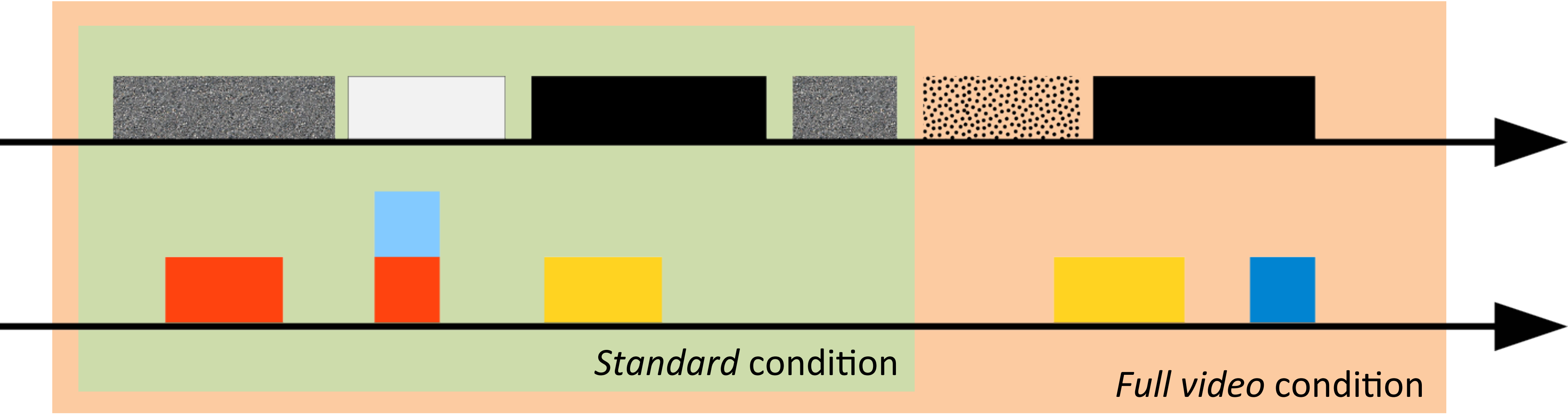
$\mathcal{T}$
$\mathcal{S}$
$\mathcal{N}$
M1
M2
M3
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9
s_1 s_4 s_2 s_1 s_5 s_2 s_3 s_1 s_3 s_1
n_1 n_4 n_2 n_3 n_4 n_3 n_5 n_3 n_1
n_1 n_2 n_3 n_1 n_4 n_3 ? n_1 ? n_1
n_1 n_2 n_3 n_1 n_4 n_3 ? n_5 n_3 n_1
n_1 n_2 n_3 n_1 n_4 n_3 n_3 n_5 n_3 n_1
Input
M1
M2
M3

Experiments on REPERE test set

- 7 TV shows
  - from BFM TV and LCP
  - talk shows and news
- 27 hours of videos
  - 6 hours annotated
- 9 anchors
  - 45 speech turns / anchor
  - 5 minutes / anchor
- 113 other speakers
  - 10 speech turns / speaker
  - 1 minute / speaker

Speakers	Propagation	Precision (%)	Recall (%)	F_1-measure
All	M1	80.5	58.2	67.5
	M2	82.1	60.7	69.8
	M3	77.7	63.9	70.2
No anchor	M3	89.2	75.3	81.7

The longer, the better



Speakers	Condition	Precision (%)	Recall (%)	F_1-measure
All	Standard	82.0	55.6	66.3
	Full video	77.7	63.9	70.2
No anchor	Standard	88.5	72.4	79.7
	Full video	89.2	75.3	81.7

Error analysis

- Speaker diarization is not perfect

  Diarization Error Rate ≃ 10 %

- How does it impact the overall system?

Speaker Diarization	Propagation	Precision (%)	Recall (%)	F_1-measure
Perfect	Perfect	100.0	76.5	86.7
	M1	98.0	76.4	85.8
Automatic	M1	89.1	70.3	78.6
	M2	91.0	73.1	81.0
	M3	88.5	72.4	79.7

No anchor

The best of both worlds

- Gathering training data for anchors is easy.
- Why not combine both supervised (SID) and unsupervised (M3) approaches?

Speakers	Approach	Precision (%)	Recall (%)	F_1-measure
All	SID	60.1	55.1	57.5
	M3	77.7	63.9	70.2
	M3 + SID	77.9	77.0	77.5
No anchor	SID	47.0	44.4	45.7
	M3	89.2	75.3	81.7
	M3 + SID	80.7	83.4	82.0

Conclusion

- Unsupervised multimodal identification better than its supervised monomodal counterpart.
- … and they are complementary

- Reproducible research
  http://herve.niderb.fr/reproducible_research

J. Poignant, H. Bredin, V.B. Le, L. Besacier, C. Barras, G. Quénot
Unsupervised Speaker Identification using Overlaid Texts in TV Broadcast
Interspeech 2012, 13th Annual Conference of the International Speech Communication Association

Hervé BREDIN – bredin@limsi.fr

http://herve.niderb.fr/