

# person instance graphs for person **recognition** in **multimedia** data

application to **speaker** identification in TV broadcast

**Hervé Bredin**

Anindya Roy - Viet-Bac Le - Claude Barras

the **QCOMPERE** project



# the partners



# the REPERE challenge

- multimodal person recognition in TV shows

- three  projects  
co-funded by 

QCOMPERE

SODA

PERCOL

**LIMSI** et al.

**LIUM** et al.

**LIF** et al.

- two evaluation campaigns in 2013 and 2014

- data collection & annotation by 
- evaluation by 

# the REPERE challenge



*who speaks when?*

*who appears when?*

# the REPERE challenge



*who speaks when?*

*who appears when?*

# the holy grail

## *free annotated data*

- **dense audio annotation**
  - speaker identification
  - speech transcription
- **sparse video annotation**
  - head recognition
  - optical character recognition
- **text annotation** named entity detection (person names only)

# the holy grail

## *free annotated data*

- **dense audio annotation**

speaker identification  
speech transcription

- **sparse video annotation**

head recognition  
optical character recognition

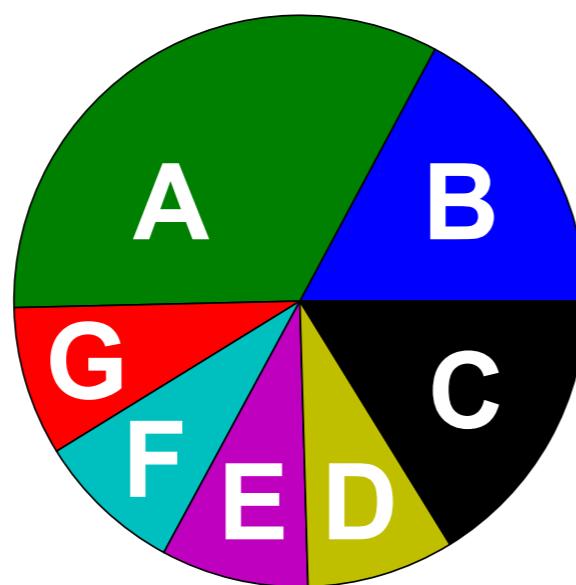
- **text annotation** named entity detection (person names only)

2013

### TRAINING SET

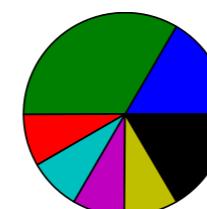
*24 hours*

- A: BFM Story
- B: LCP Info
- C: Top Questions
- D: Ça Vous Regarde
- E: Planète Showbiz
- F: Entre Les Lignes
- G: Pile Et Face



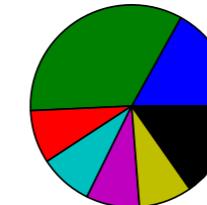
### DEVELOPMENT

*3 hours*



### TEST

*3 hours*



# the holy grail

## *free annotated data*

- **dense audio annotation**

speaker identification  
speech transcription

- **sparse video annotation**

head recognition  
optical character recognition

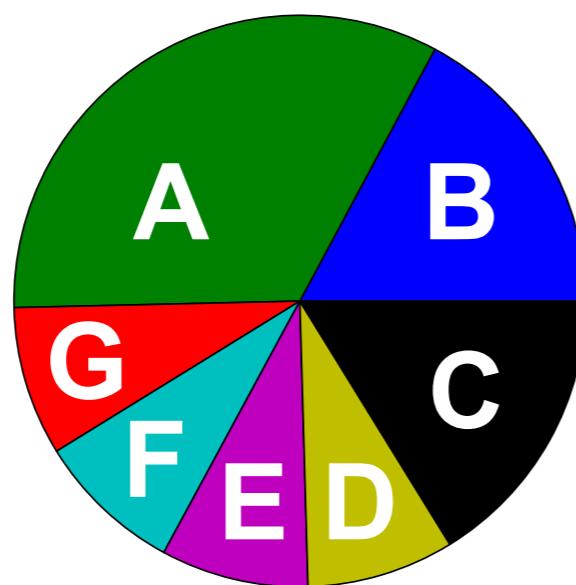
- **text annotation** named entity detection (person names only)

2013

### TRAINING SET

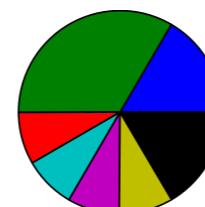
*24 hours*

A: BFM Story  
B: LCP Info  
C: Top Questions  
D: Ça Vous Regarde  
E: Planète Showbiz  
F: Entre Les Lignes  
G: Pile Et Face



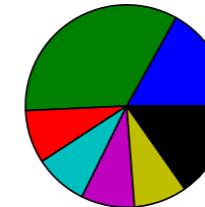
### DEVELOPMENT

*3 hours*



### TEST

*3 hours*



2014 X 2

# multimodal person recognition

- Multiple sources of information
  - **Audio stream**  
Speaker diarization & identification  
Speech transcription (ASR)
  - **Visual stream**  
Face clustering & recognition  
Optical character recognition (OCR)
  - **Text stream (from ASR and OCR)**  
Named entity detection  
Name normalization



# **TLP — who does what?**

- Multiple sources of information
  - **Audio stream**  
Speaker diarization & identification  
Speech transcription (ASR)
  - **Visual stream**  
Face clustering & recognition  
Optical character recognition (OCR)
  - **Text stream** (from ASR and OCR)  
Named entity detection  
Name normalization

# TLP — who **does** what?

- Multiple sources of information
  - **Audio stream**  
Speaker diarization & identification  
Speech transcription (ASR)
  - **Visual stream**  
Face clustering & recognition  
Optical character recognition (OCR)
  - **Text stream** (from ASR and OCR)  
Named entity detection  
Name normalization

**Claude Barras**  
**Achintya Sarkar**

# TLP — who **does** what?

- Multiple sources of information
  - **Audio stream**  
Speaker diarization & identification  
Speech transcription (ASR)
  - **Visual stream**  
Face clustering & recognition  
Optical character recognition (OCR)
  - **Text stream** (from ASR and OCR)  
Named entity detection  
Name normalization

**Claude Barras**  
**Achintya Sarkar**

**Sophie Rosset**  
**Anindya Roy**  
**Camille Guinaudeau**

# TLP — who does what?

- Multiple sources of information

**Hervé Bredin**

- **Audio stream**

Speaker diarization & identification

Speech transcription (ASR)

**Claude Barras**  
**Achintya Sarkar**

- **Visual stream**

Face clustering & recognition

Optical character recognition (OCR)

- **Text stream (from ASR and OCR)**

Named entity detection

Name normalization

**Sophie Rosset**  
**Anindya Roy**  
**Camille Guinaudeau**

# person instance graphs for person **recognition** in **multimedia** data

application to **speaker** identification in TV broadcast

**Hervé Bredin**

Anindya Roy - Viet-Bac Le - Claude Barras

# outline

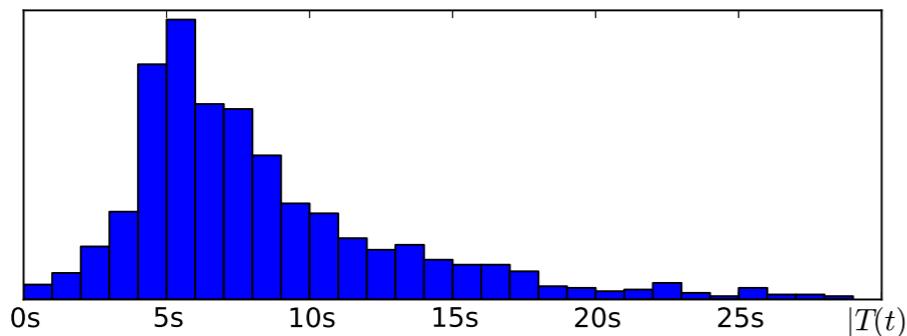
- integer linear programming  
for speaker diarization
- building person instance graphs
- mining person instance graphs
- experiments
- conclusion

# **integer linear programming**

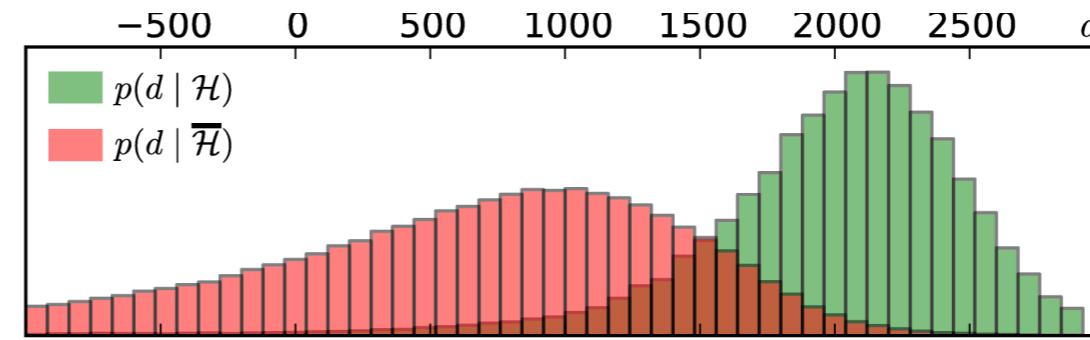
## for speaker diarization

# speaker diarization

- the task of **partitioning** and **labeling** an audio stream into homogeneous **speech segments** according to the **identity of the speaker**
- usually addressed as a **segmentation** problem followed by a **clustering** problem



distribution of **speech turns duration**



distribution of **intra-cluster** and **inter-cluster** speech turn similarity

# three limitations

- k-means (et al.) rely on the assumption that the **number of clusters** is known *a priori*.
- hierarchical agglomerative clustering does not guarantee **global optimality** does not provide clear **stopping criterion**
- cannot deal with **incomplete similarity matrices**
- *Finkel & Manning* formulate clustering as an **integer linear programming** problem (for co-reference resolution)

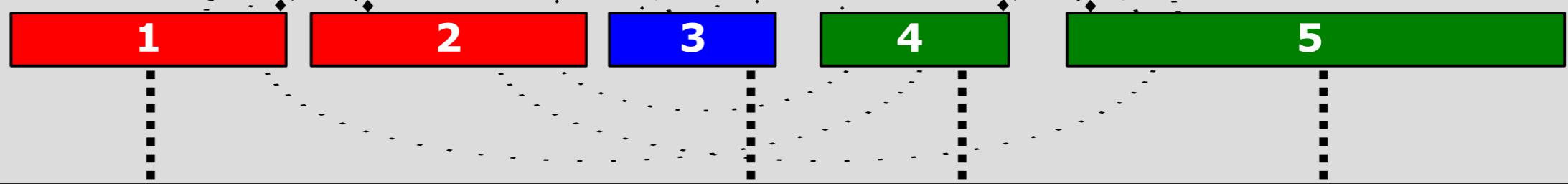
# notations

- $\mathcal{G} = (\mathcal{V}, \mathcal{E}, p)$  is an **undirected weighted** graph
- $\mathcal{V}$  is the set of **vertices (one per item** to cluster)
- $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$  is the set of **edges**  
two vertices **may or may not be connected**
- edges are weighted by the **probability** that  
connected vertices belong to the **same cluster**

$$p: \mathcal{E} \rightarrow [0, 1]$$

$$(v, v') \mapsto p_{vv'} = p(\text{ID}(v) = \text{ID}(v') \mid v, v')$$

SPEECH  
TURNS



# clustering function

- Any output of a valid clustering algorithm can be described by a clustering function

$$\delta: \mathcal{V} \times \mathcal{V} \rightarrow \{0, 1\}$$

$$(v, v') \mapsto \begin{cases} 1 & \text{if } v \text{ and } v' \text{ are in the same cluster,} \\ 0 & \text{otherwise.} \end{cases}$$

# clustering function

- Any output of a valid clustering algorithm can be described by a clustering function

$$\delta: \mathcal{V} \times \mathcal{V} \rightarrow \{0, 1\}$$

$$(v, v') \mapsto \begin{cases} 1 & \text{if } v \text{ and } v' \text{ are in the same cluster,} \\ 0 & \text{otherwise.} \end{cases}$$

- Reciprocally, additional constraints are needed in order to guarantee a valid clustering

$$\Delta_{\mathcal{V}} = \left\{ \begin{array}{ll} \delta \in \{0, 1\}^{\mathcal{V} \times \mathcal{V}} & \text{s.t. } \forall (v, v', v'') \in \mathcal{V}^3, \\ \text{(a) } \delta_{vv} = 1 & \text{reflexivity} \\ \text{(b) } \delta_{vv'} = \delta_{v'v} & \text{symmetry} \\ \text{(c) } \delta_{vv'} = 1 \wedge \delta_{v'v''} = 1 \implies \delta_{vv''} = 1 & \text{transitivity} \end{array} \right.$$

# clustering function

- Any output of a valid clustering algorithm can be described by a clustering function

$$\delta: \mathcal{V} \times \mathcal{V} \rightarrow \{0, 1\}$$

$$(v, v') \mapsto \begin{cases} 1 & \text{if } v \text{ and } v' \text{ are in the same cluster,} \\ 0 & \text{otherwise.} \end{cases}$$

- Reciprocally, additional constraints are needed in order to guarantee a valid clustering

$$\Delta_{\mathcal{V}} = \left\{ \begin{array}{ll} \delta \in \{0, 1\}^{\mathcal{V} \times \mathcal{V}} & \text{s.t. } \forall (v, v', v'') \in \mathcal{V}^3, \\ \text{(a) } \delta_{vv} = 1 & \text{reflexivity} \\ \text{(b) } \delta_{vv'} = \delta_{v'v} & \text{symmetry} \\ \text{(c) } \delta_{vv'} = 1 \wedge \delta_{v'v''} = 1 \implies \delta_{vv''} = 1 & \text{transitivity} \end{array} \right.$$

# clustering function

- Any output of a valid clustering algorithm can be described by a clustering function

$$\delta: \mathcal{V} \times \mathcal{V} \rightarrow \{0, 1\}$$

$$(v, v') \mapsto \begin{cases} 1 & \text{if } v \text{ and } v' \text{ are in the same cluster,} \\ 0 & \text{otherwise.} \end{cases}$$

- Reciprocally, additional constraints are needed in order to guarantee a valid clustering

$$\Delta_{\mathcal{V}} = \left\{ \begin{array}{ll} \delta \in \{0, 1\}^{\mathcal{V} \times \mathcal{V}} & \text{s.t. } \forall (v, v', v'') \in \mathcal{V}^3, \\ \text{(a) } \delta_{vv} = 1 & \text{reflexivity} \\ \text{(b) } \delta_{vv'} = \delta_{v'v} & \text{symmetry} \\ \text{(c) } \delta_{vv'} = 1 \wedge \delta_{v'v''} = 1 \implies \delta_{vv''} = 1 & \text{transitivity} \end{array} \right.$$

# clustering function

- Any output of a valid clustering algorithm can be described by a clustering function

$$\delta: \mathcal{V} \times \mathcal{V} \rightarrow \{0, 1\}$$

$$(v, v') \mapsto \begin{cases} 1 & \text{if } v \text{ and } v' \text{ are in the same cluster,} \\ 0 & \text{otherwise.} \end{cases}$$

- Reciprocally, additional constraints are needed in order to guarantee a valid clustering

$$\Delta_{\mathcal{V}} = \left\{ \begin{array}{ll} \delta \in \{0, 1\}^{\mathcal{V} \times \mathcal{V}} & \text{s.t. } \forall (v, v', v'') \in \mathcal{V}^3, \\ \text{(a) } \delta_{vv} = 1 & \text{reflexivity} \\ \text{(b) } \delta_{vv'} = \delta_{v'v} & \text{symmetry} \\ \text{(c) } \delta_{vv'} = 1 \wedge \delta_{v'v''} = 1 \implies \delta_{vv''} = 1 & \text{transitivity} \end{array} \right.$$

# clustering function

- Any output of a valid clustering algorithm can be described by a clustering function

$$\delta: \mathcal{V} \times \mathcal{V} \rightarrow \{0, 1\}$$

$$(v, v') \mapsto \begin{cases} 1 & \text{if } v \text{ and } v' \text{ are in the same cluster,} \\ 0 & \text{otherwise.} \end{cases}$$

- Reciprocally, additional constraints are needed in order to guarantee a valid clustering

$$\Delta_{\mathcal{V}} = \left\{ \begin{array}{ll} \delta \in \{0, 1\}^{\mathcal{V} \times \mathcal{V}} & \text{s.t. } \forall (v, v', v'') \in \mathcal{V}^3, \\ \text{(a) } \delta_{vv} = 1 & \text{reflexivity} \\ \text{(b) } \delta_{vv'} = \delta_{v'v} & \text{symmetry} \\ \text{(c) } \delta_{vv'} = 1 \wedge \delta_{v'v''} = 1 \implies \delta_{vv''} = 1 & \text{transitivity} \end{array} \right.$$

# clustering function

- Any output of a valid clustering algorithm can be described by a clustering function

$$\delta: \mathcal{V} \times \mathcal{V} \rightarrow \{0, 1\}$$

$$(v, v') \mapsto \begin{cases} 1 & \text{if } v \text{ and } v' \text{ are in the same cluster,} \\ 0 & \text{otherwise.} \end{cases}$$

- Reciprocally, additional constraints are needed in order to guarantee a valid clustering

$$\Delta_{\mathcal{V}} = \left\{ \begin{array}{ll} \delta \in \{0, 1\}^{\mathcal{V} \times \mathcal{V}} & \text{s.t. } \forall (v, v', v'') \in \mathcal{V}^3, \\ \text{(a) } \delta_{vv} = 1 & \text{reflexivity} \\ \text{(b) } \delta_{vv'} = \delta_{v'v} & \text{symmetry} \\ \text{(c) } \delta_{vv'} = 1 \wedge \delta_{v'v''} = 1 \implies \delta_{vv''} = 1 & \text{transitivity} \end{array} \right.$$

# objective function

- Find the **clustering function** that maximizes the objective function:  $\delta^* = \underset{\delta \in \Delta_V}{\operatorname{argmax}} \mathcal{L}^\alpha(\delta, \mathcal{E}, p)$

# objective function

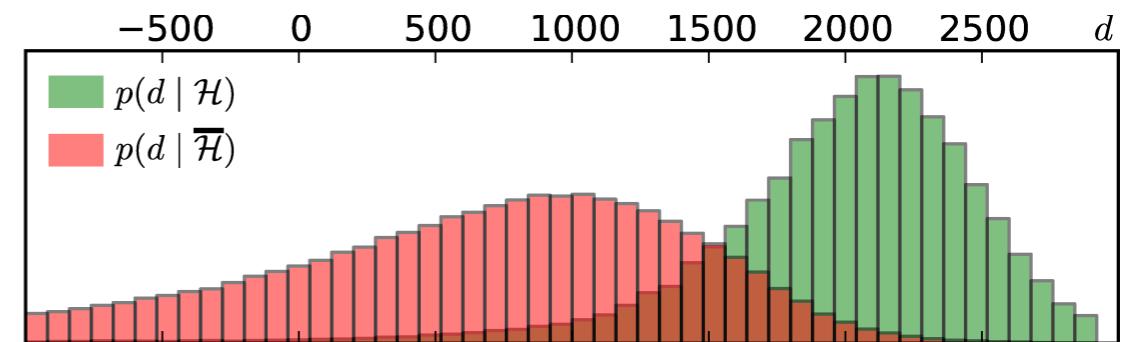
- Find the **clustering function** that maximizes the objective function:  $\delta^* = \underset{\delta \in \Delta_{\mathcal{V}}}{\operatorname{argmax}} \mathcal{L}^\alpha(\delta, \mathcal{E}, p)$

$$\mathcal{L}^\alpha(\delta, \mathcal{E}, p) = |\mathcal{E}|^{-1} \left[ \alpha \cdot \overbrace{\sum_{(v,v') \in \mathcal{E}} \delta_{vv'} \cdot p_{vv'}}^{\text{intra-cluster similarity}} + (1 - \alpha) \cdot \overbrace{\sum_{(v,v') \in \mathcal{E}} (1 - \delta_{vv'}) \cdot (1 - p_{vv'})}^{\text{inter-cluster dissimilarity}} \right]$$

# speech turns similarity

- Standard MFCC features
- Bayesian information criterion (BIC)

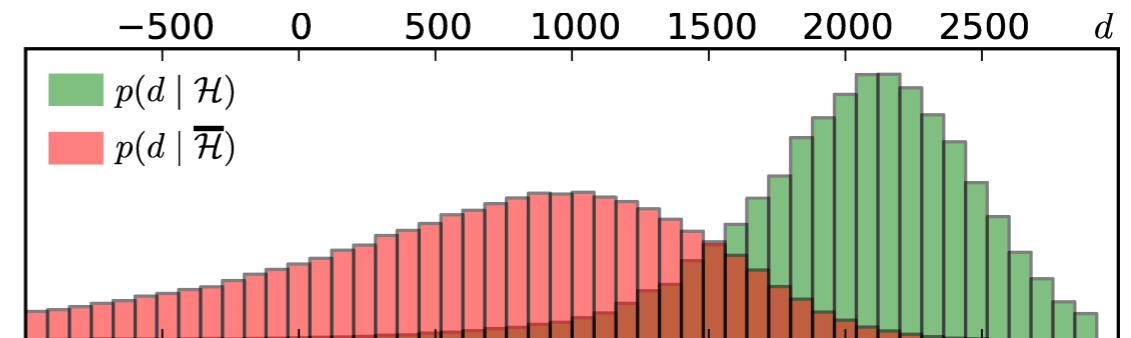
$$\begin{aligned} d_{tt'} = & (n_t + n_{t'}) \log |\Sigma_{t+t'}| \\ & - n_t \log |\Sigma_t| - n_{t'} \log |\Sigma_{t'}| \\ & - \frac{1}{2} \cdot \lambda \cdot \left( D + \frac{1}{2} D(D+1) \right) \log (n_t + n_{t'}) \end{aligned}$$



# speech turns similarity

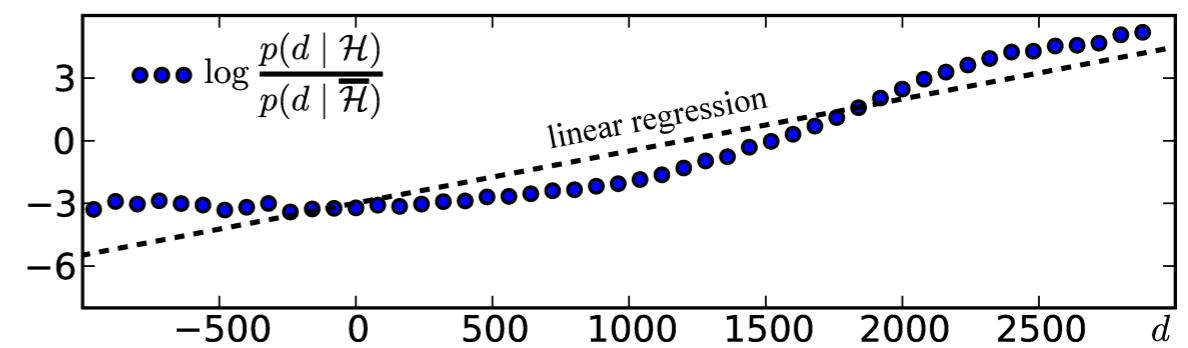
- Standard MFCC features
- Bayesian information criterion (BIC)

$$\begin{aligned}
 d_{tt'} &= (n_t + n_{t'}) \log |\Sigma_{t+t'}| \\
 &\quad - n_t \log |\Sigma_t| - n_{t'} \log |\Sigma_{t'}| \\
 &\quad - \frac{1}{2} \cdot \lambda \cdot \left( D + \frac{1}{2} D(D+1) \right) \log (n_t + n_{t'})
 \end{aligned}$$



- Bayes' theorem
- Linear regression for log-likelihood ratio estimation

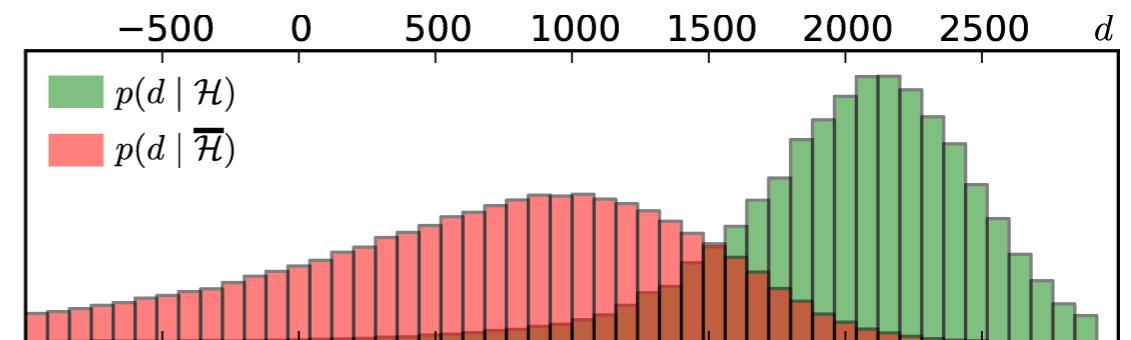
$$\begin{aligned}
 p_{tt'} &= p(\text{ID}(t) = \text{ID}(t') \mid d_{tt'}) \\
 &= \frac{1}{1 + \frac{\pi_{\neq}}{\pi_{=}} \cdot \frac{p(d_{tt'} \mid \text{ID}(t) \neq \text{ID}(t'))}{p(d_{tt'} \mid \text{ID}(t) = \text{ID}(t'))}}
 \end{aligned}$$



# speech turns similarity

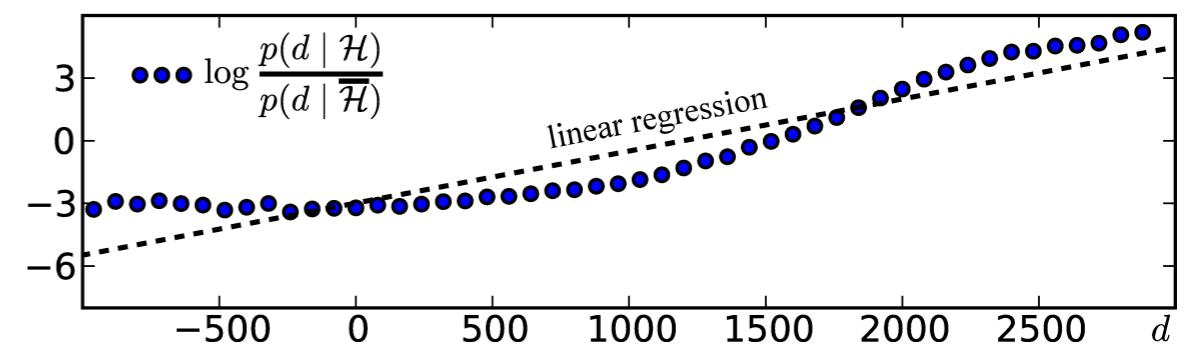
- Standard MFCC features
- Bayesian information criterion (BIC)

$$\begin{aligned}
 d_{tt'} &= (n_t + n_{t'}) \log |\Sigma_{t+t'}| \\
 &\quad - n_t \log |\Sigma_t| - n_{t'} \log |\Sigma_{t'}| \\
 &\quad - \frac{1}{2} \cdot \lambda \cdot \left( D + \frac{1}{2} D(D+1) \right) \log (n_t + n_{t'})
 \end{aligned}$$



- Bayes' theorem
- Linear regression for log-likelihood ratio estimation

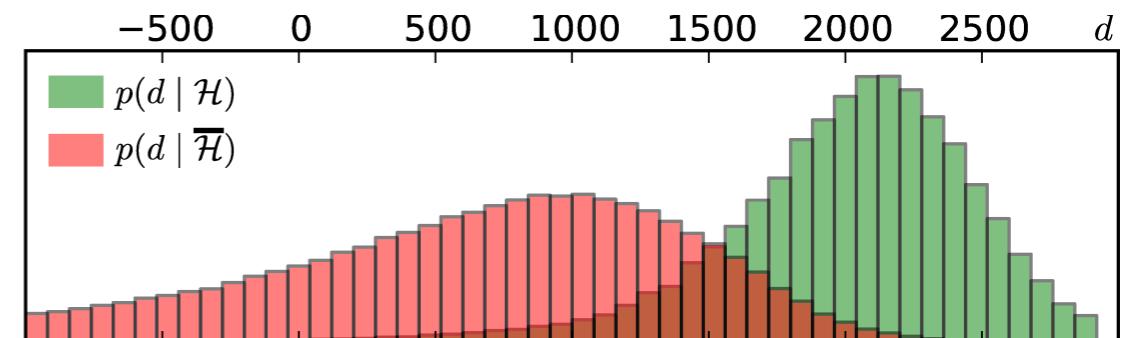
$$\begin{aligned}
 p_{tt'} &= p(\text{ID}(t) = \text{ID}(t') \mid d_{tt'}) \\
 &= \frac{1}{1 + \frac{\pi_{\neq}}{\pi_{=}} \cdot \frac{p(d_{tt'} \mid \text{ID}(t) \neq \text{ID}(t'))}{p(d_{tt'} \mid \text{ID}(t) = \text{ID}(t'))}}
 \end{aligned}$$



# speech turns similarity

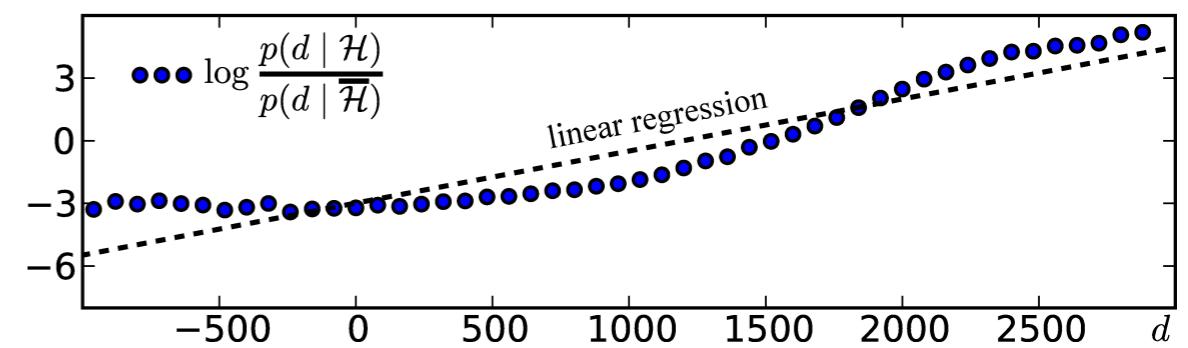
- Standard MFCC features
- Bayesian information criterion (BIC)

$$\begin{aligned}
 d_{tt'} &= (n_t + n_{t'}) \log |\Sigma_{t+t'}| \\
 &\quad - n_t \log |\Sigma_t| - n_{t'} \log |\Sigma_{t'}| \\
 &\quad - \frac{1}{2} \cdot \lambda \cdot \left( D + \frac{1}{2} D(D+1) \right) \log (n_t + n_{t'})
 \end{aligned}$$



- Bayes' theorem
- Linear regression for log-likelihood ratio estimation

$$\begin{aligned}
 p_{tt'} &= p(\text{ID}(t) = \text{ID}(t') \mid d_{tt'}) \\
 &= \frac{1}{1 + \frac{\pi_{\neq}}{\pi_{=}} \cdot \frac{p(d_{tt'} \mid \text{ID}(t) \neq \text{ID}(t'))}{p(d_{tt'} \mid \text{ID}(t) = \text{ID}(t'))}}
 \end{aligned}$$



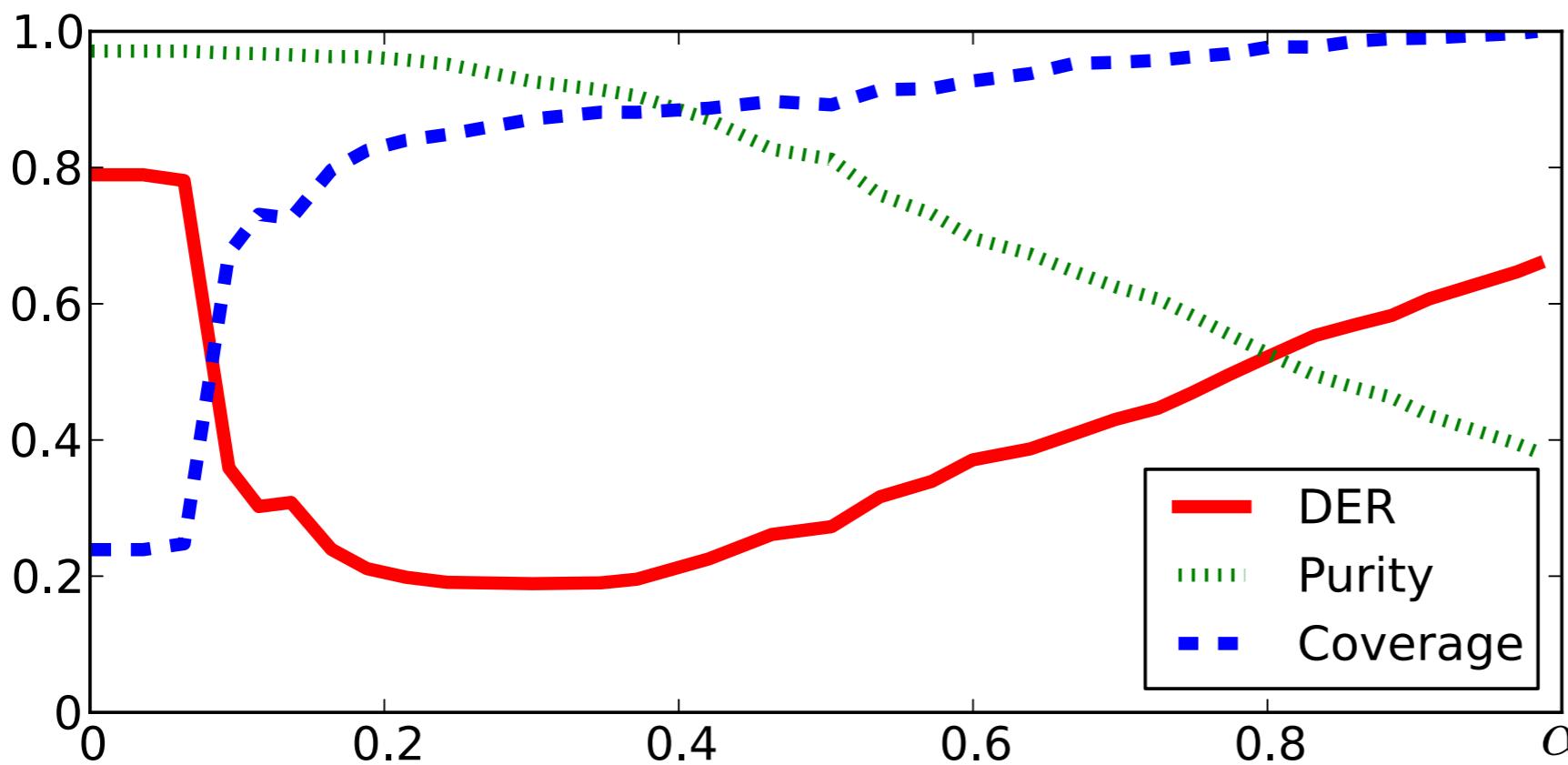
# effect of $\alpha$

$$\mathcal{L}^\alpha(\delta, \mathcal{E}, p) = |\mathcal{E}|^{-1} \left[ \underbrace{\alpha \cdot \sum_{(v,v') \in \mathcal{E}} \delta_{vv'} \cdot p_{vv'}}_{\text{intra-cluster similarity}} + (1 - \alpha) \cdot \underbrace{\sum_{(v,v') \in \mathcal{E}} (1 - \delta_{vv'}) \cdot (1 - p_{vv'})}_{\text{inter-cluster dissimilarity}} \right]$$

# effect of $\alpha$

$$\mathcal{L}^\alpha(\delta, \mathcal{E}, p) = |\mathcal{E}|^{-1} \left[ \underbrace{\alpha \cdot \sum_{(v,v') \in \mathcal{E}} \delta_{vv'} \cdot p_{vv'}}_{\text{intra-cluster similarity}} + (1 - \alpha) \cdot \underbrace{\sum_{(v,v') \in \mathcal{E}} (1 - \delta_{vv'}) \cdot (1 - p_{vv'})}_{\text{inter-cluster dissimilarity}} \right]$$

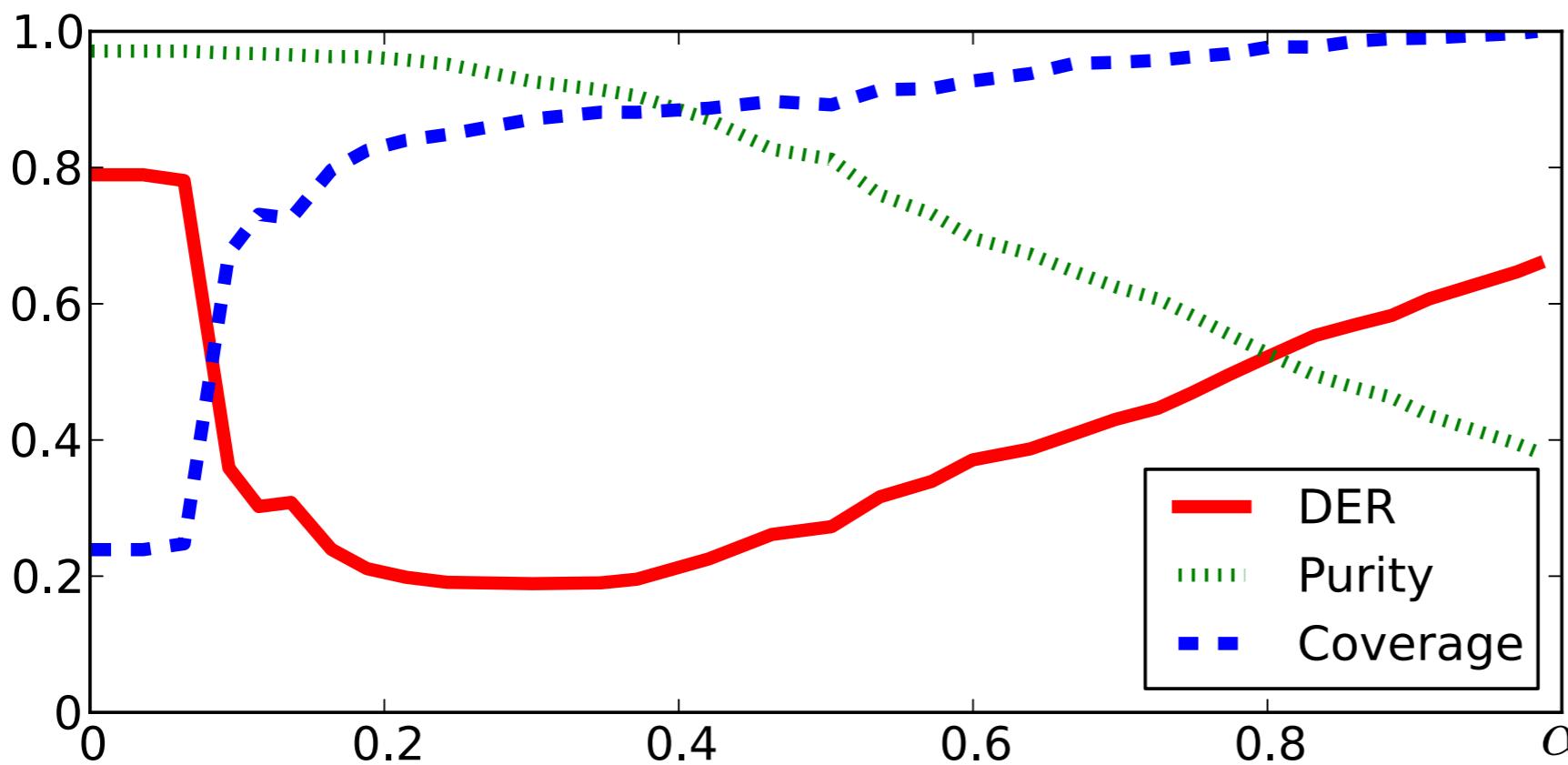
- cluster **purity** vs. cluster **coverage**



# effect of $\alpha$

$$\mathcal{L}^\alpha(\delta, \mathcal{E}, p) = |\mathcal{E}|^{-1} \left[ \underbrace{\alpha \cdot \sum_{(v,v') \in \mathcal{E}} \delta_{vv'} \cdot p_{vv'}}_{\text{intra-cluster similarity}} + \underbrace{(1 - \alpha) \cdot \sum_{(v,v') \in \mathcal{E}} (1 - \delta_{vv'}) \cdot (1 - p_{vv'})}_{\text{inter-cluster dissimilarity}} \right]$$

- cluster **purity** vs. cluster **coverage**



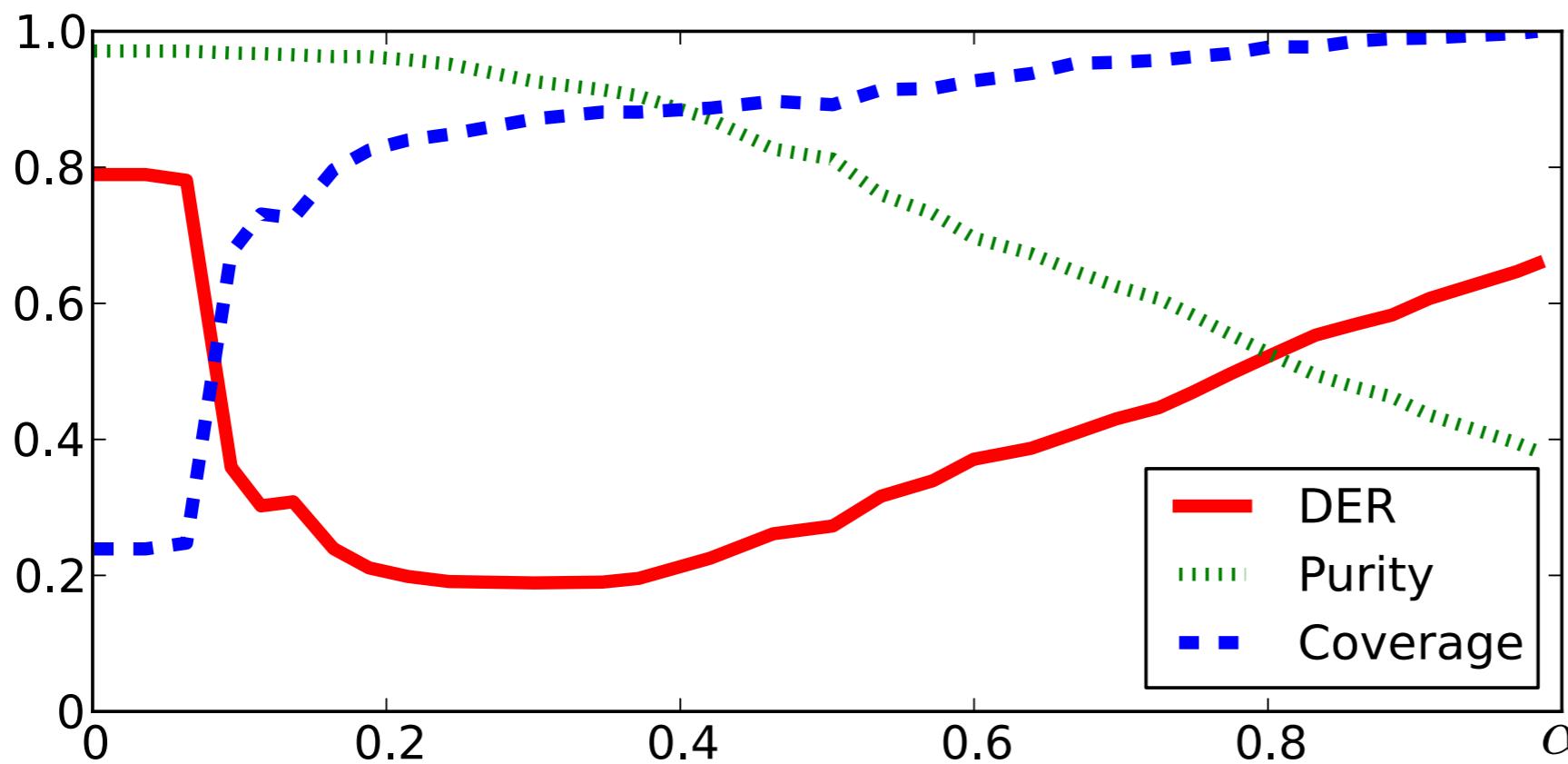
not nearly as good as standard BIC clustering



# effect of $\alpha$

$$\mathcal{L}^\alpha(\delta, \mathcal{E}, p) = |\mathcal{E}|^{-1} \left[ \underbrace{\alpha \cdot \sum_{(v,v') \in \mathcal{E}} \delta_{vv'} \cdot p_{vv'}}_{\text{intra-cluster similarity}} + \underbrace{(1 - \alpha) \cdot \sum_{(v,v') \in \mathcal{E}} (1 - \delta_{vv'}) \cdot (1 - p_{vv'})}_{\text{inter-cluster dissimilarity}} \right]$$

- cluster **purity** vs. cluster **coverage**



not nearly as good as standard BIC clustering



any **question** so far?

# **building** person instance graphs

# person instance graph

- weighted undirected graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, p)$$

- instance vertices

speech turns  $t \in \mathcal{T}$

written names  $w \in \mathcal{W}$

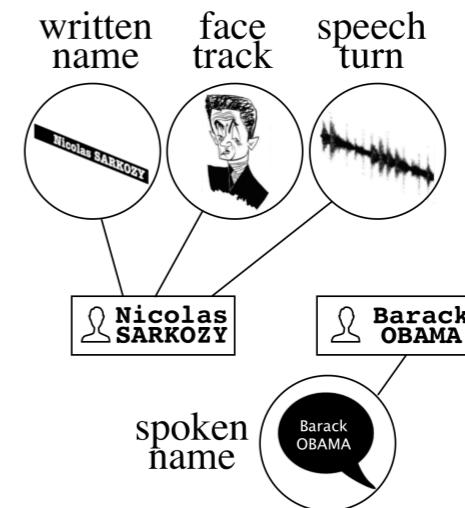
spoken names  $s \in \mathcal{S}$

~~face tracks~~

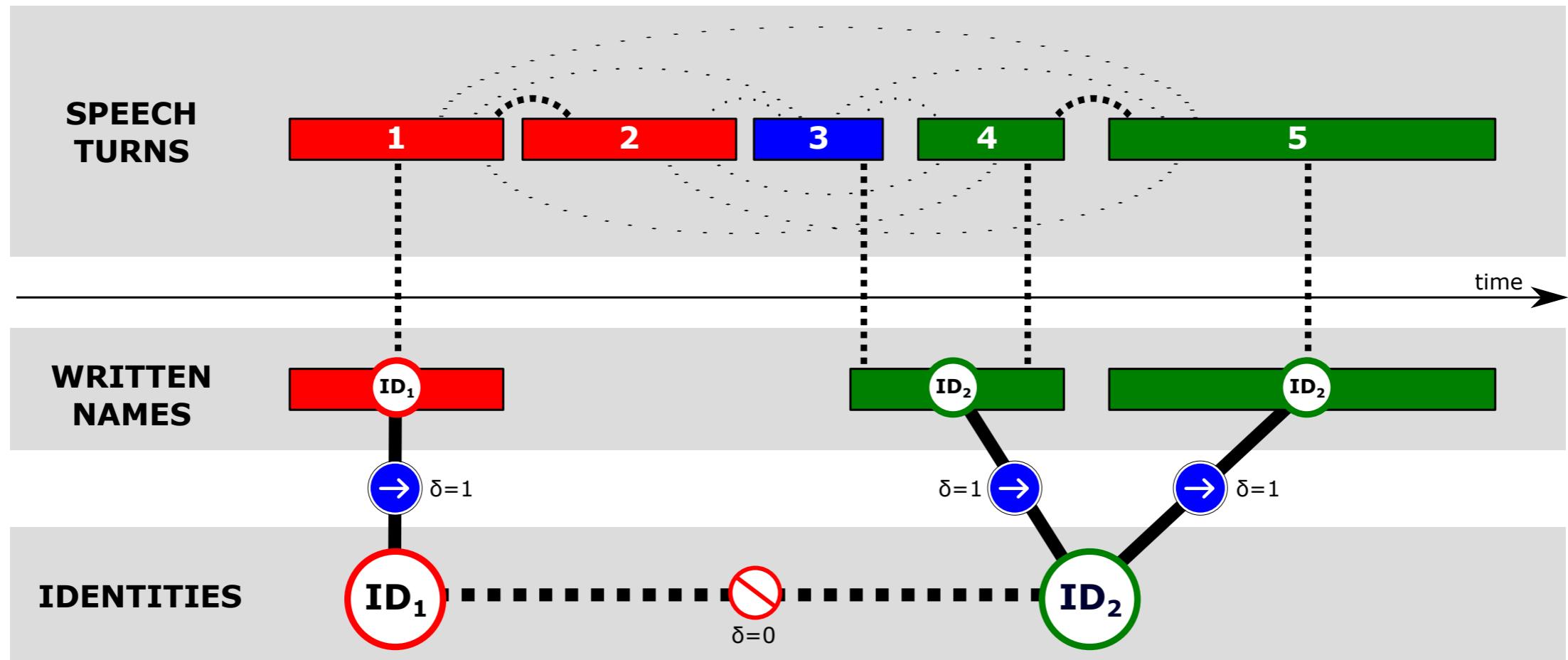
- identity vertices

- edges  $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$

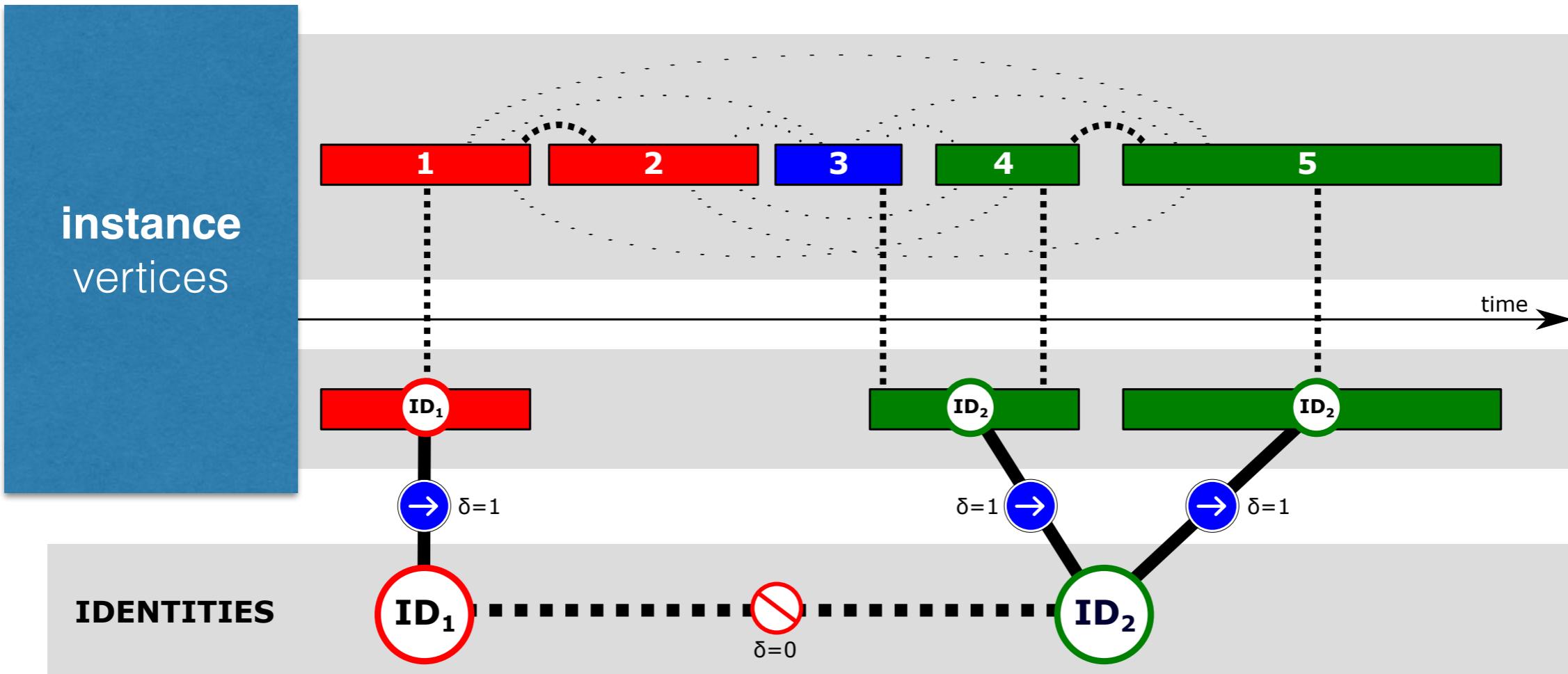
weighted by  $p \in [0, 1]^{\mathcal{E}}$



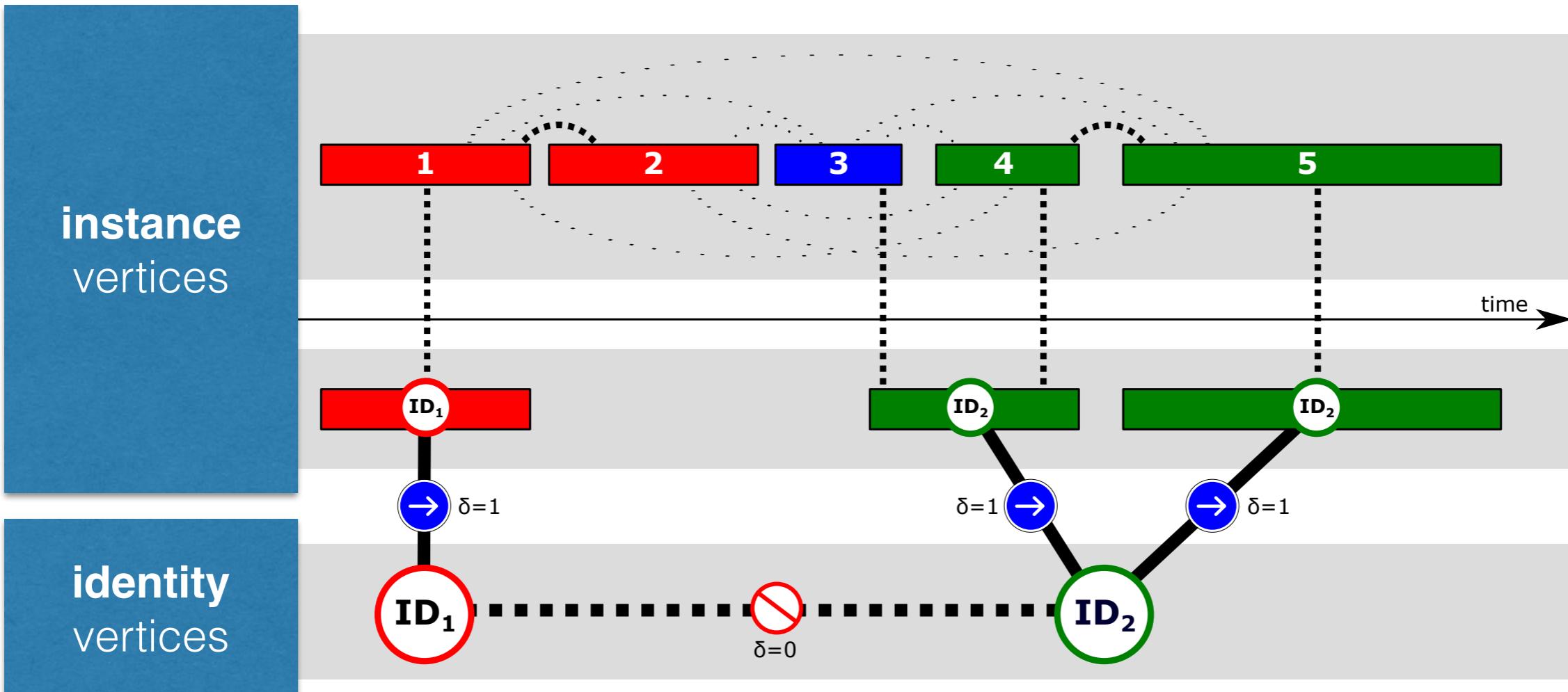
# vertices



# vertices

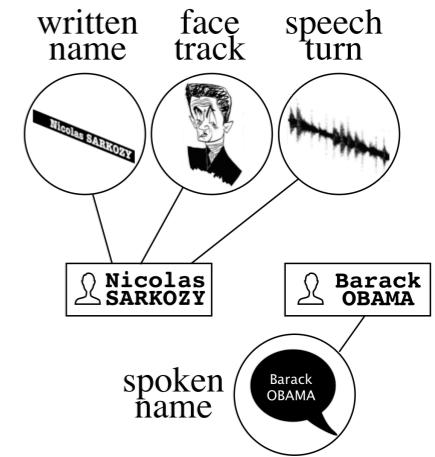


# vertices





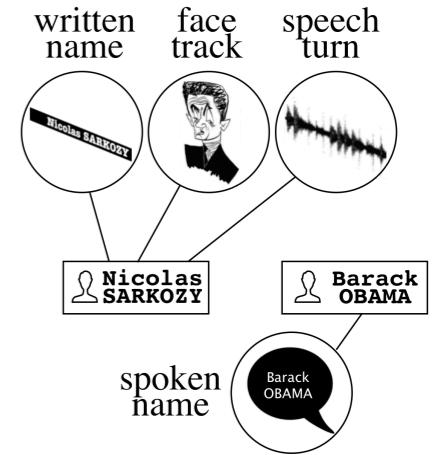
# vertices



- 4 sources of **instance** vertices
  - **speech turns**  $t \in \mathcal{T}$   
speech activity detection  
divergence-based segmentation
  - **face tracks**  
not used here
  - **written names**  $w \in \mathcal{W}$   
video optical character  
recognition  
named entity detection
  - **spoken names**  $s \in \mathcal{S}$   
manual speech transcription  
manual named entity detection

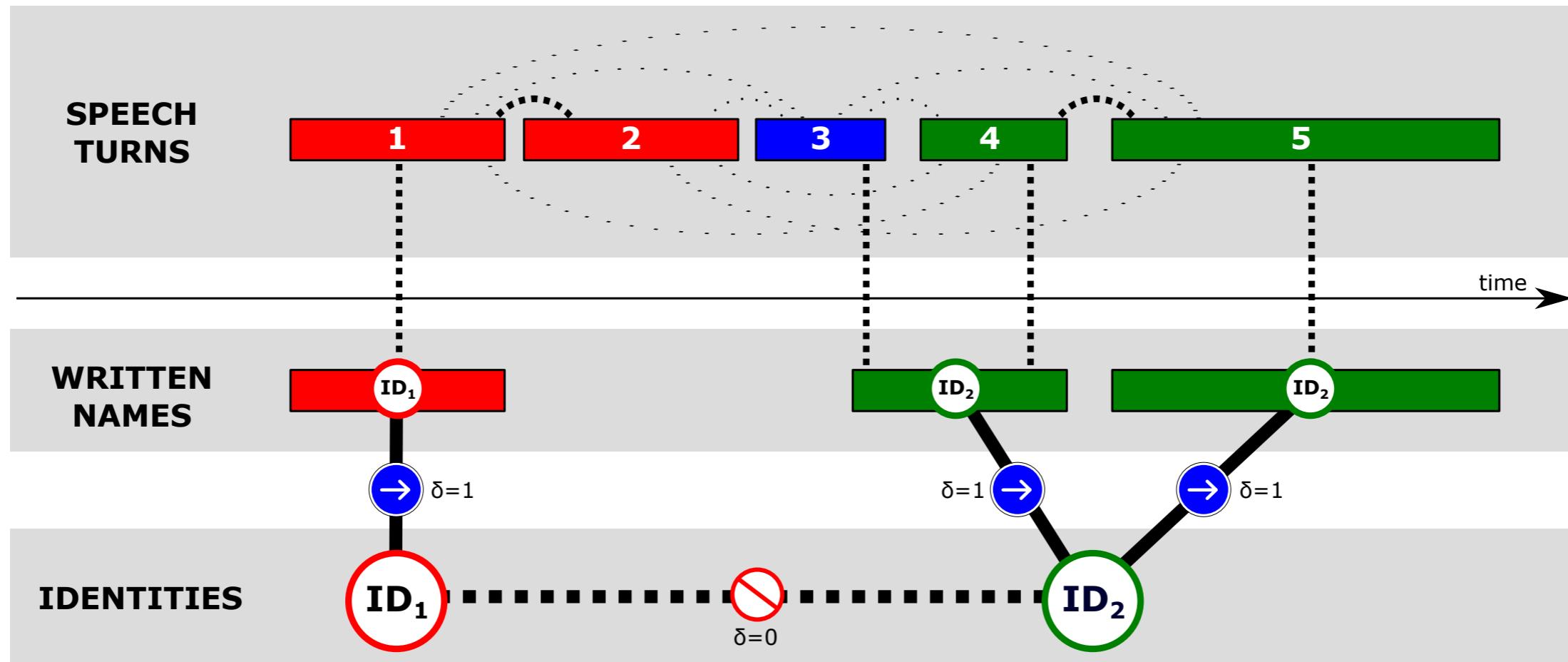


# vertices

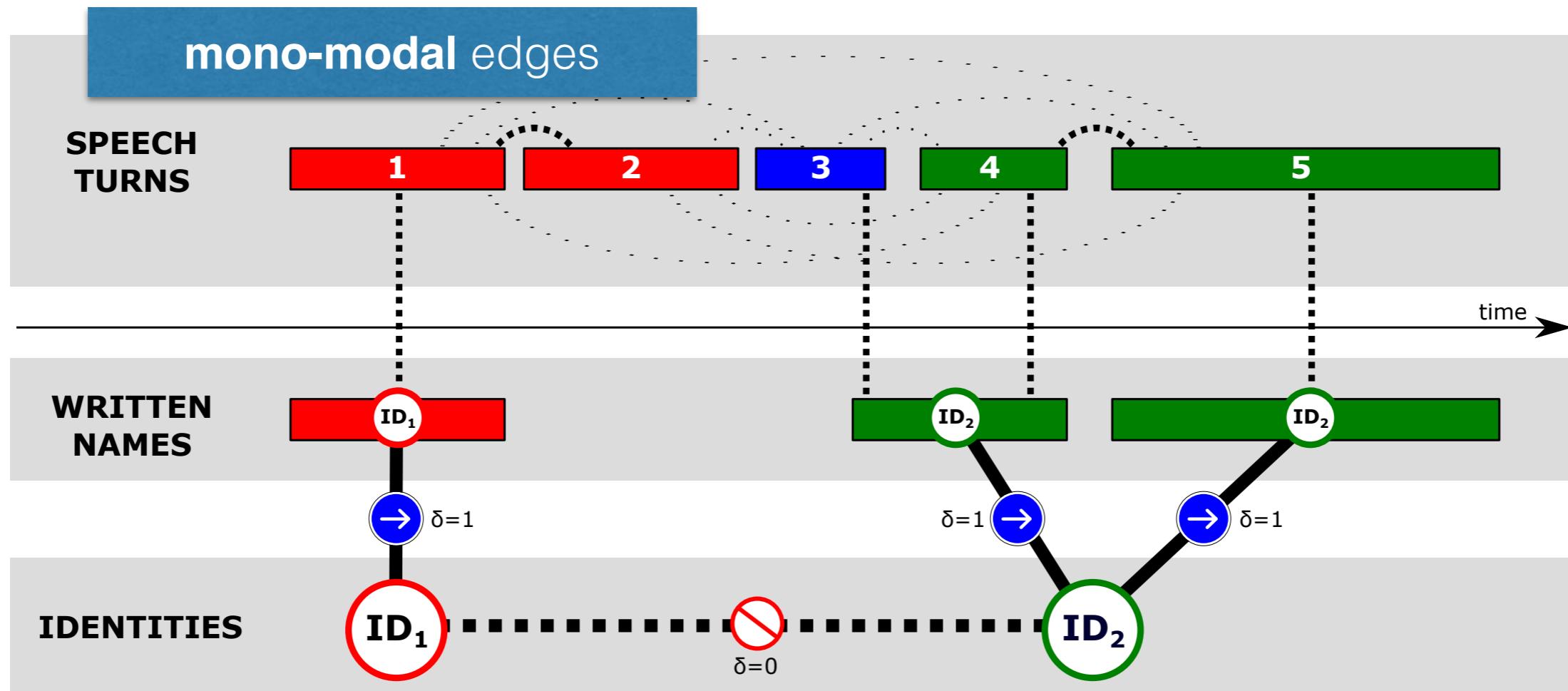


- 4 sources of **instance** vertices
  - **speech turns**  $t \in \mathcal{T}$   
speech activity detection  
divergence-based segmentation
  - **face tracks**  
not used here
  - **written names**  $w \in \mathcal{W}$   
video optical character  
recognition  
named entity detection
  - **spoken names**  $s \in \mathcal{S}$   
manual speech transcription  
manual named entity detection
- 4 sources of **identity** vertices
  - a unique identifier per person  
Nicolas Sarkozy  $\Rightarrow$  Nicolas\_SARKOZY
  - **supervised** identity vertices  $\mathcal{I}^*$ 
    - from acoustic **speaker models** (around 350 vertices)
    - from visual **face models**
  - **unsupervised** identity vertices
    - from **written names**  $\mathcal{I}_{\mathcal{W}}$
    - from **spoken names**  $\mathcal{I}_{\mathcal{S}}$

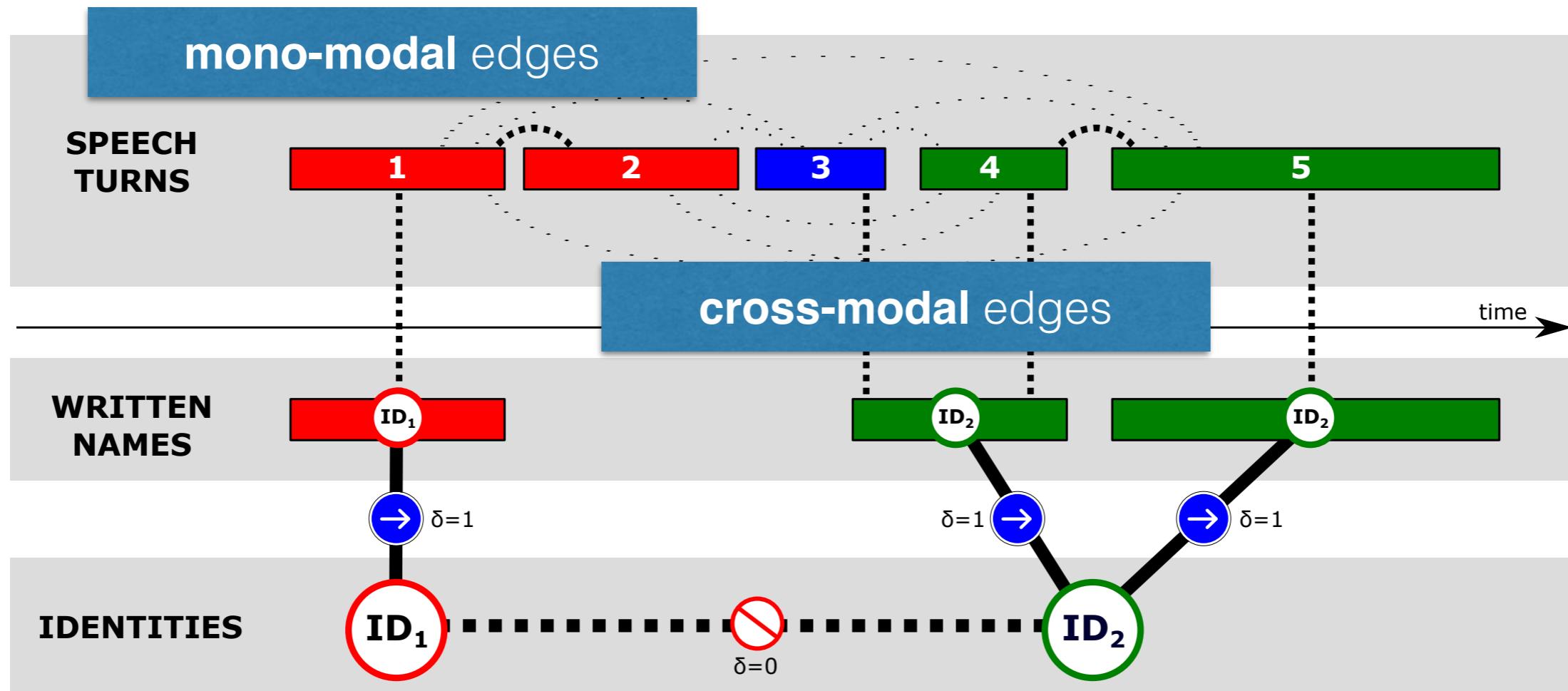
# edges



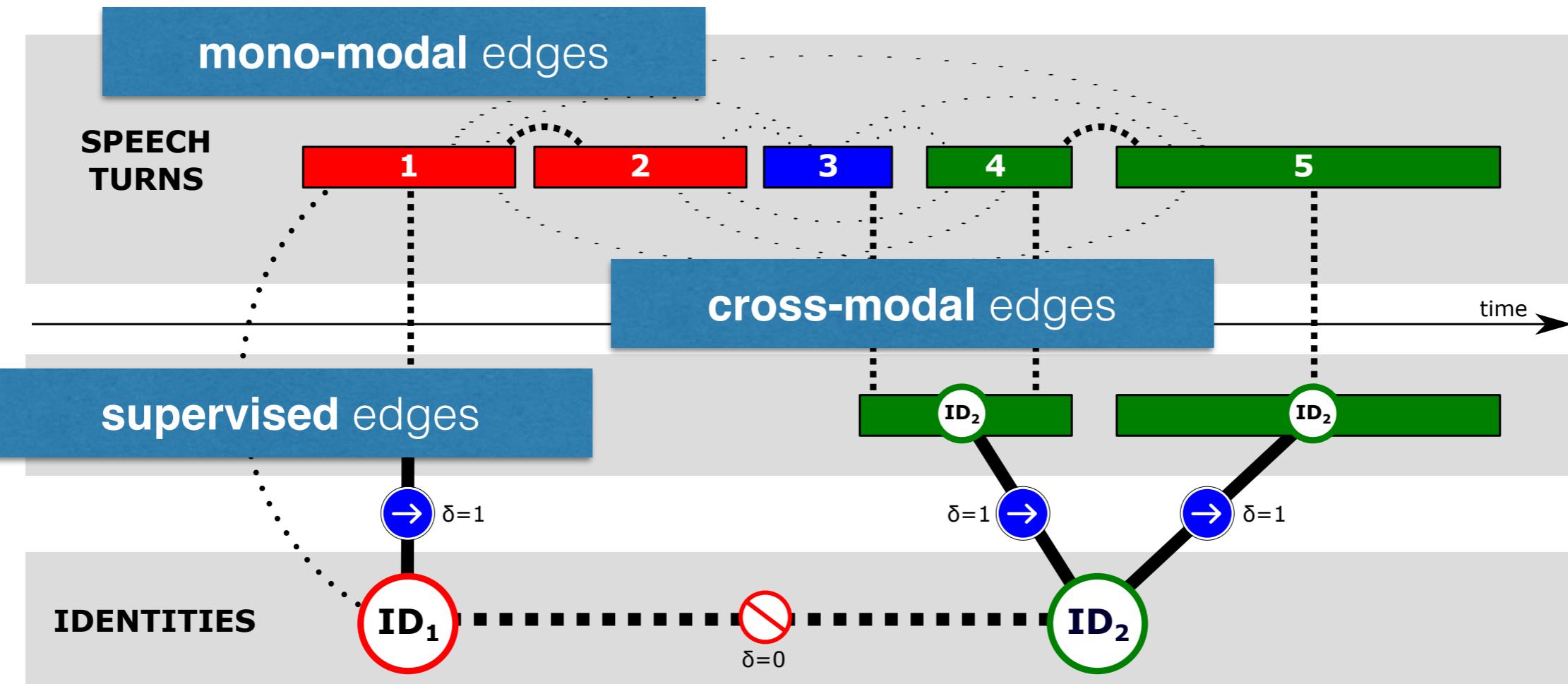
# edges



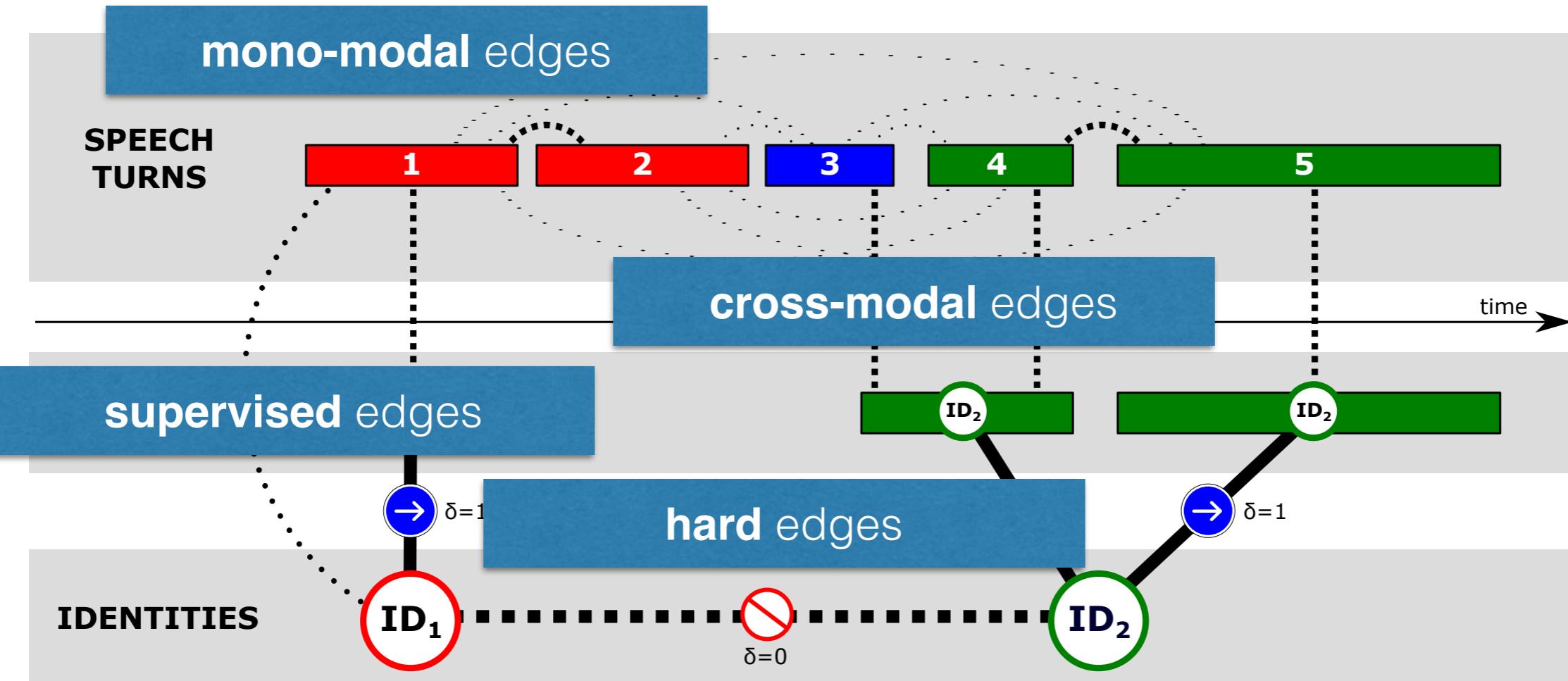
# edges



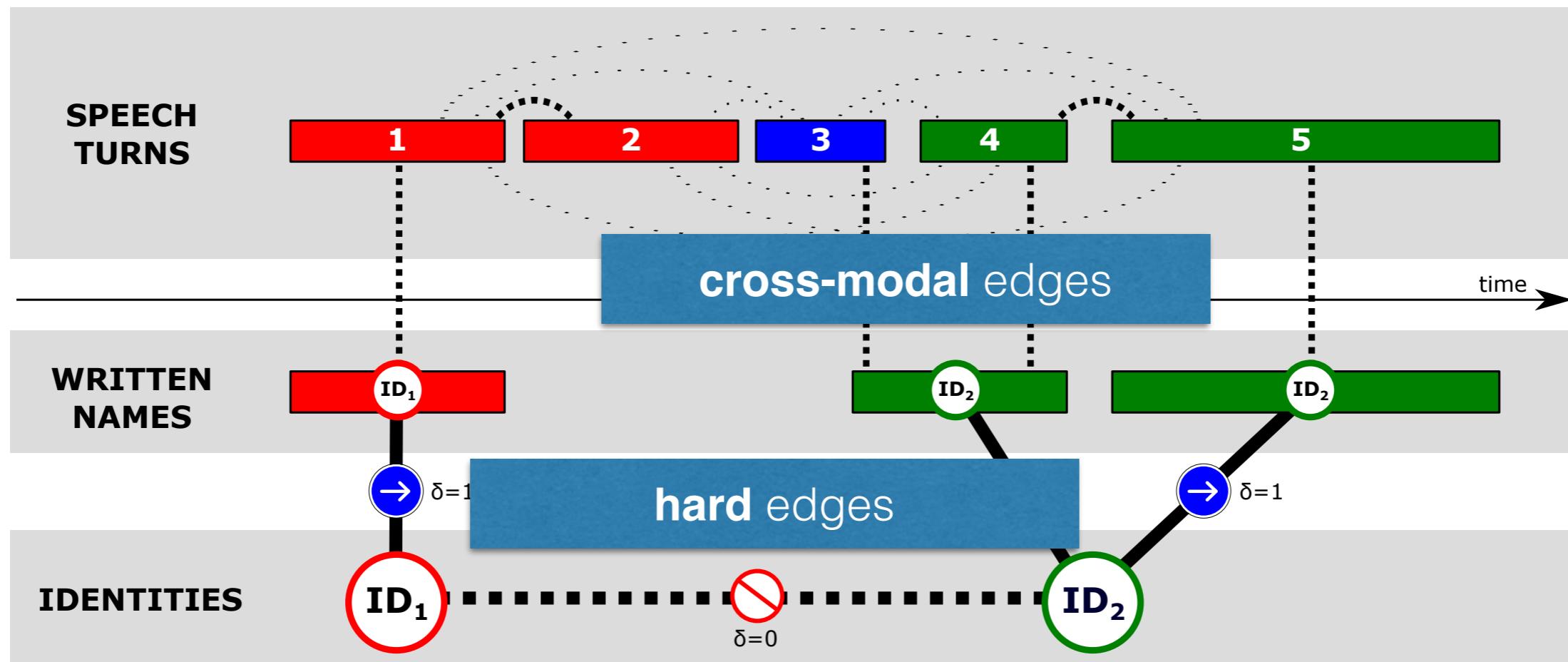
# edges



# edges



# edges



# cross-modal edges

$$t \leftrightarrow w$$

probability that a name written at time t  
is the name of current speaker



$$p_{tw} = 0.956$$



$$p_{tw} = 0.996$$

$$w \Leftrightarrow i_w$$

written instance vertices are connected  
to their corresponding identity vertices

$$p_{wi_w} = 1$$

# cross-modal edges

$$t \leftrightarrow w$$

probability that a name written at time  $t$  is the name of current speaker



$$p_{tw} = 0.956$$



$$p_{tw} = 0.996$$

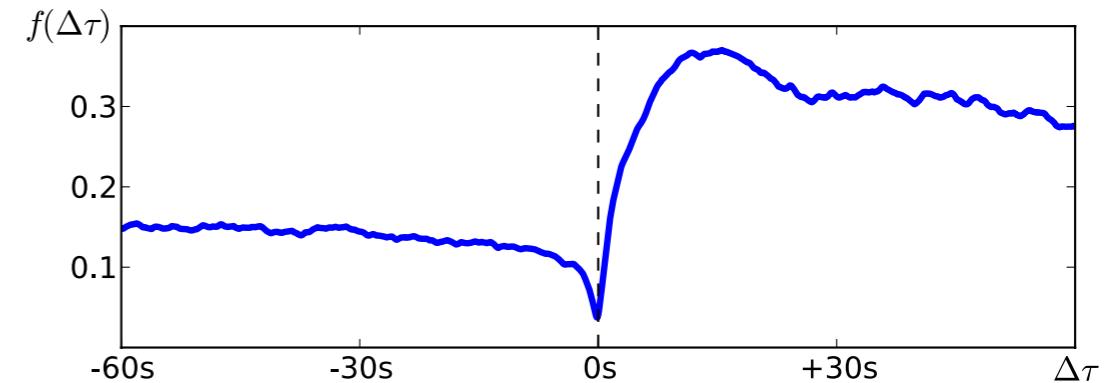
$$w \Leftrightarrow i_w$$

written instance vertices are connected to their corresponding identity vertices

$$p_{wi_w} = 1$$

$$t \leftrightarrow s$$

probability that a name pronounced at time  $\tau$  is the name of speaker at time  $\tau + \Delta\tau$



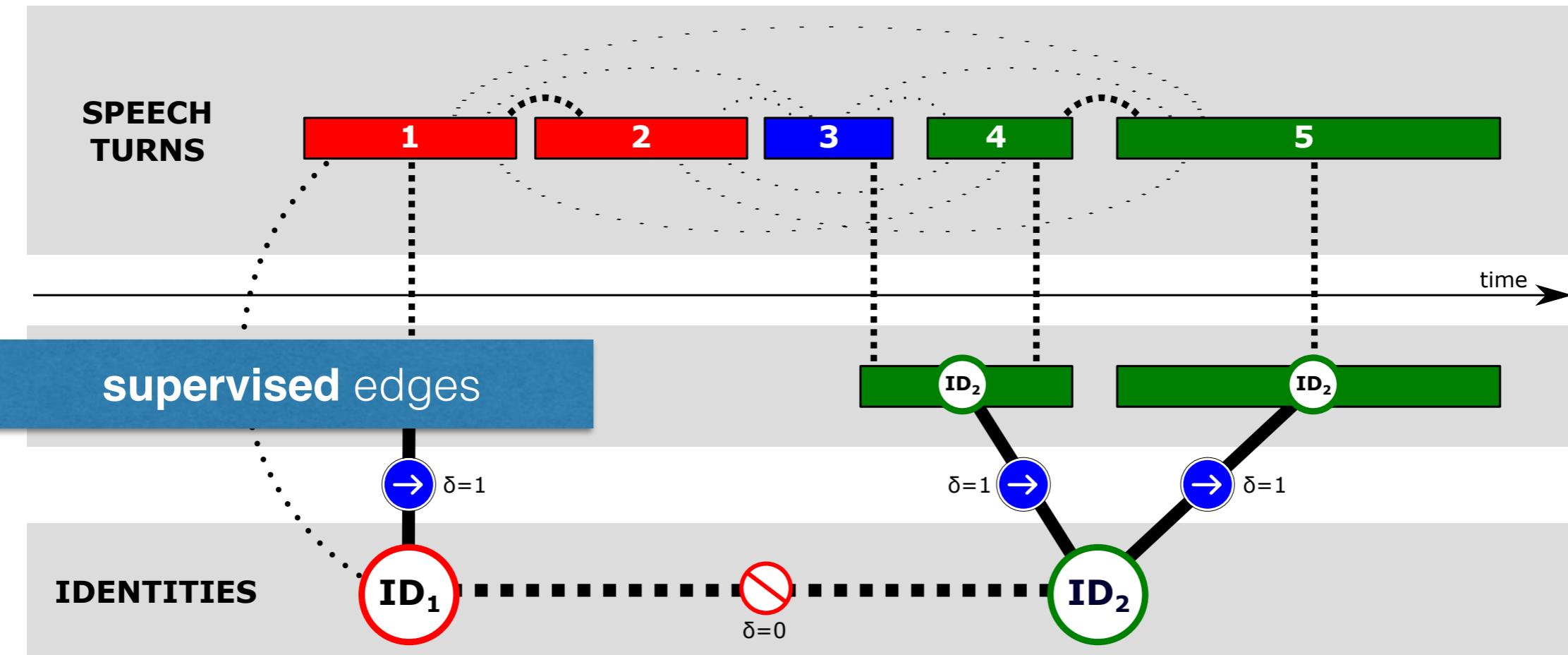
$$\begin{aligned} p_{ts} &= p(\text{ID}(t) = \text{ID}(s) \mid t, s) \\ &= \frac{1}{|T(t)|} \int_{\tau \in T(t)} f(\tau - \tau_s) \, d\tau \end{aligned}$$

$$s \Leftrightarrow i_s$$

spoken instance vertices are connected to their corresponding identity vertices

$$p_{si_s} = 1$$

# edges



# supervised edges

- GMM / UBM speaker identification

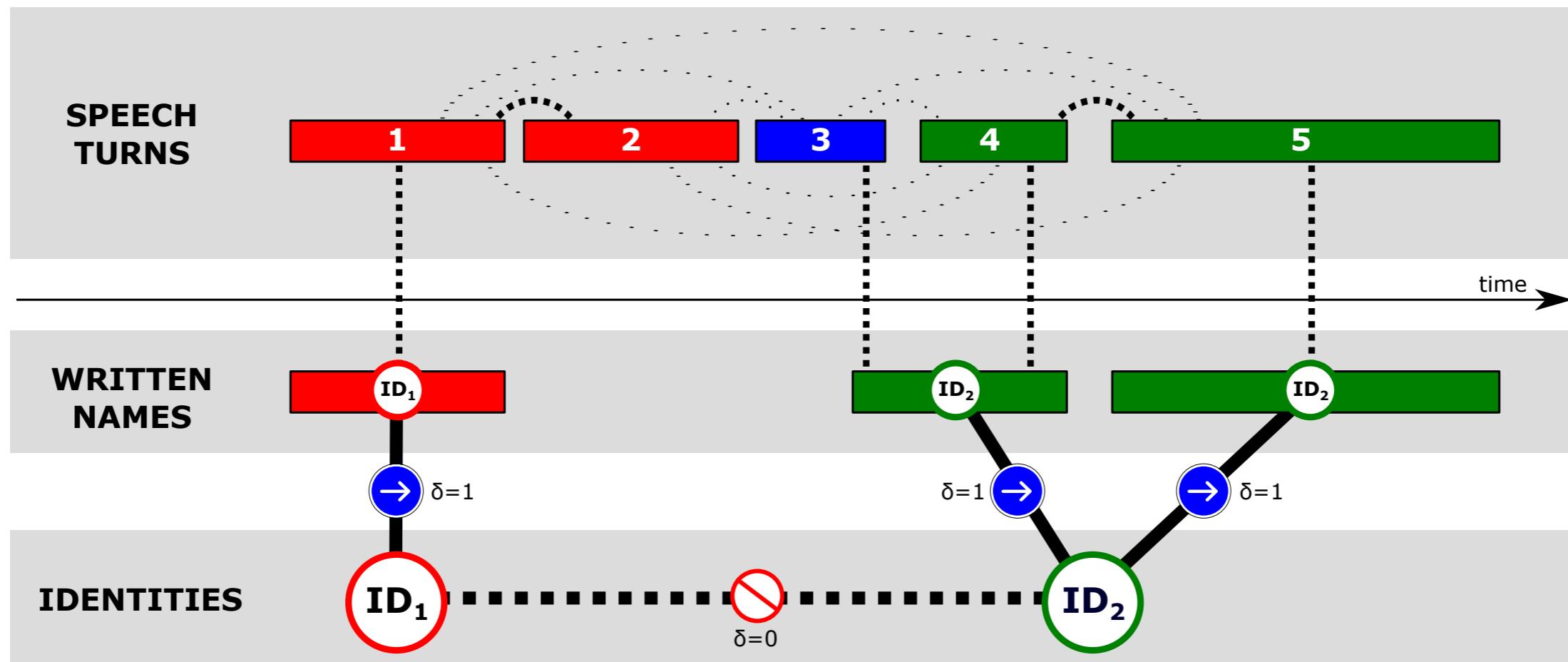
$$d_{ti} = \frac{1}{n_t} \left[ \log \prod_{x \in X_t} f(x | \lambda_i) - \log \prod_{x \in X_t} f(x | \lambda_\Omega) \right]$$

$$\begin{aligned} p_{ti} &= p(\text{ID}(t) = i \mid d_{ti}) \\ &= \frac{\pi_i \cdot \frac{p(d_{ti} \mid \text{ID}(t) = i)}{p(d_{ti} \mid \text{ID}(t) \neq i)}}{\pi_{\text{?}} + \sum_{i' \in \mathcal{I}^*} \pi_{i'} \cdot \frac{p(d_{ti'} \mid \text{ID}(t) = i')}{p(d_{ti'} \mid \text{ID}(t) \neq i')}} \end{aligned}$$

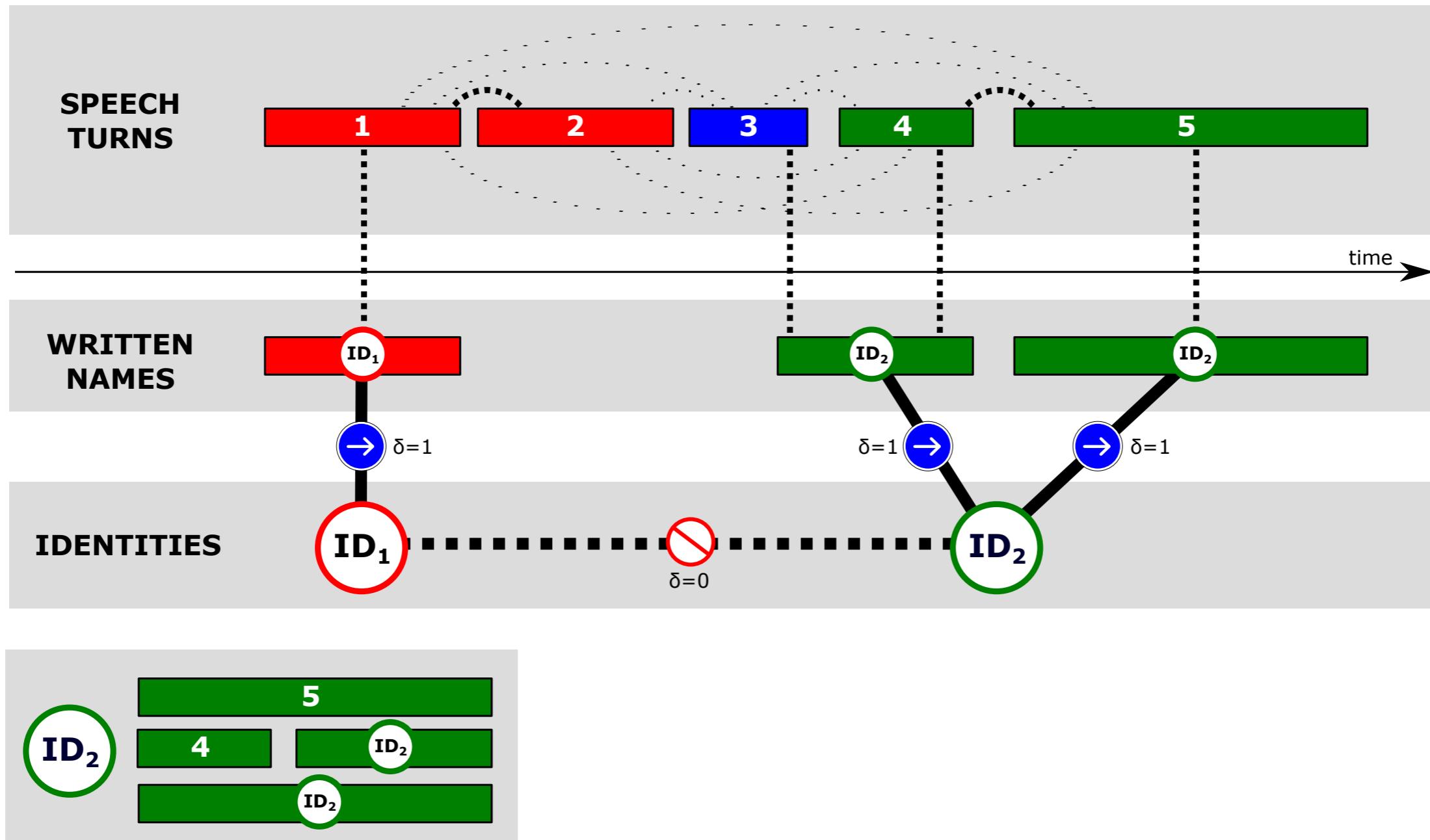
- every speech turn is connected  $t \leftrightarrow i$  to all ( $\approx 350$ ) identity vertices

# mining person instance graphs

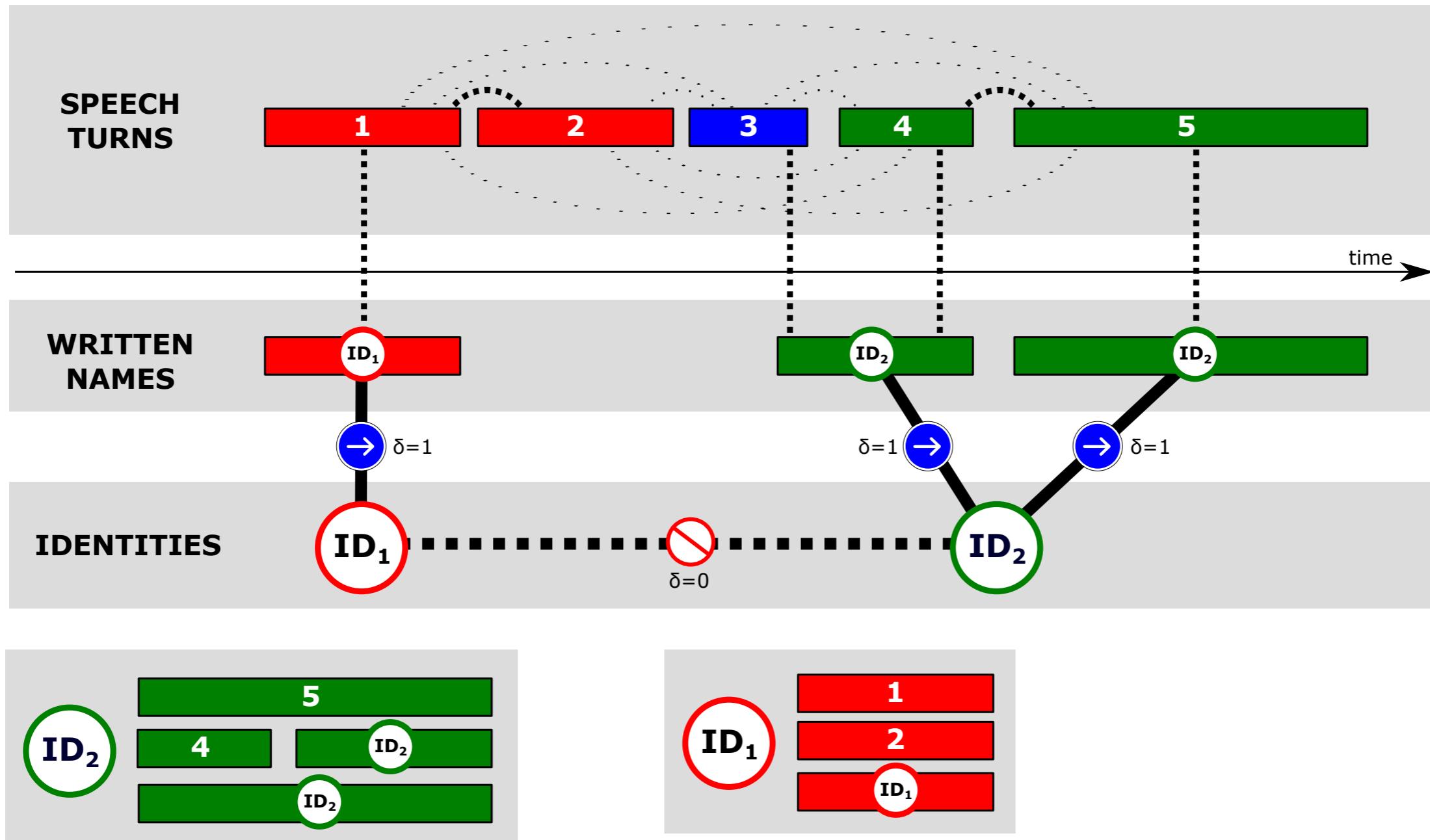
# expected results



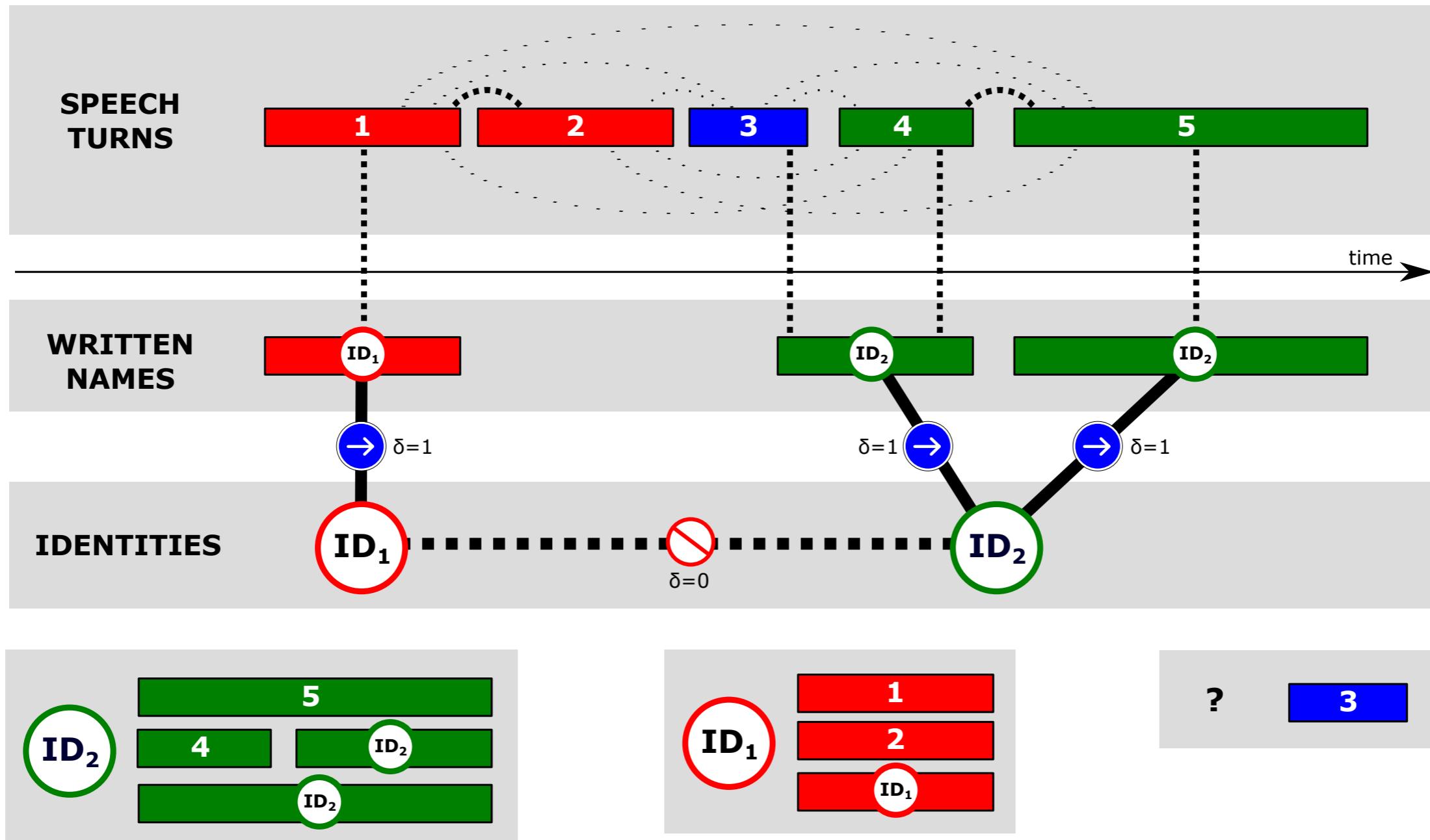
# expected results



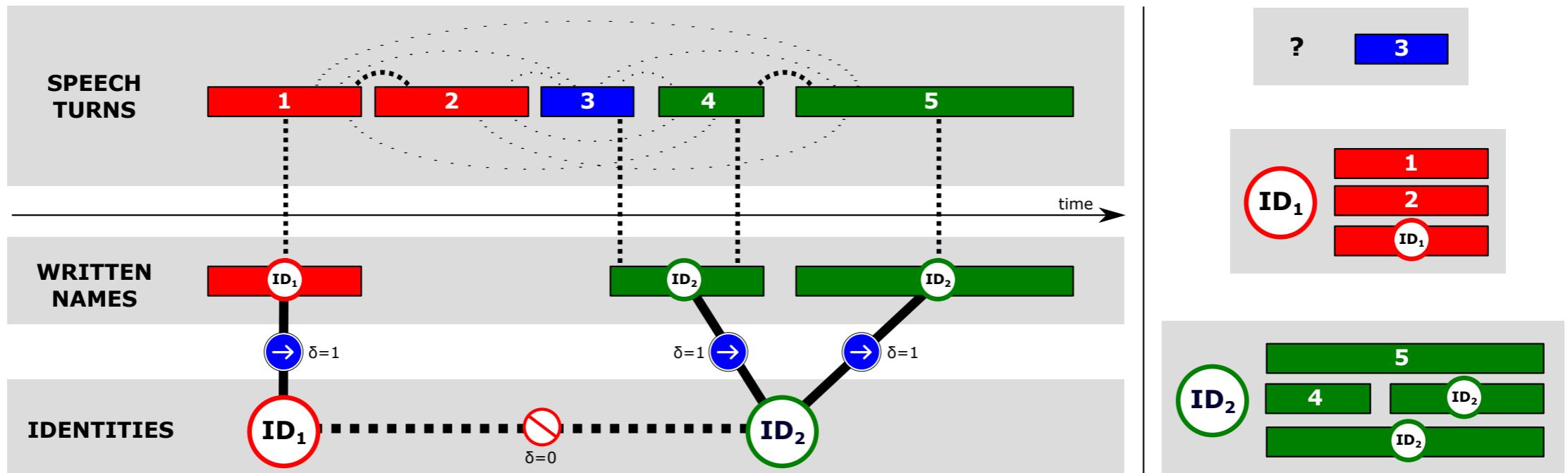
# expected results



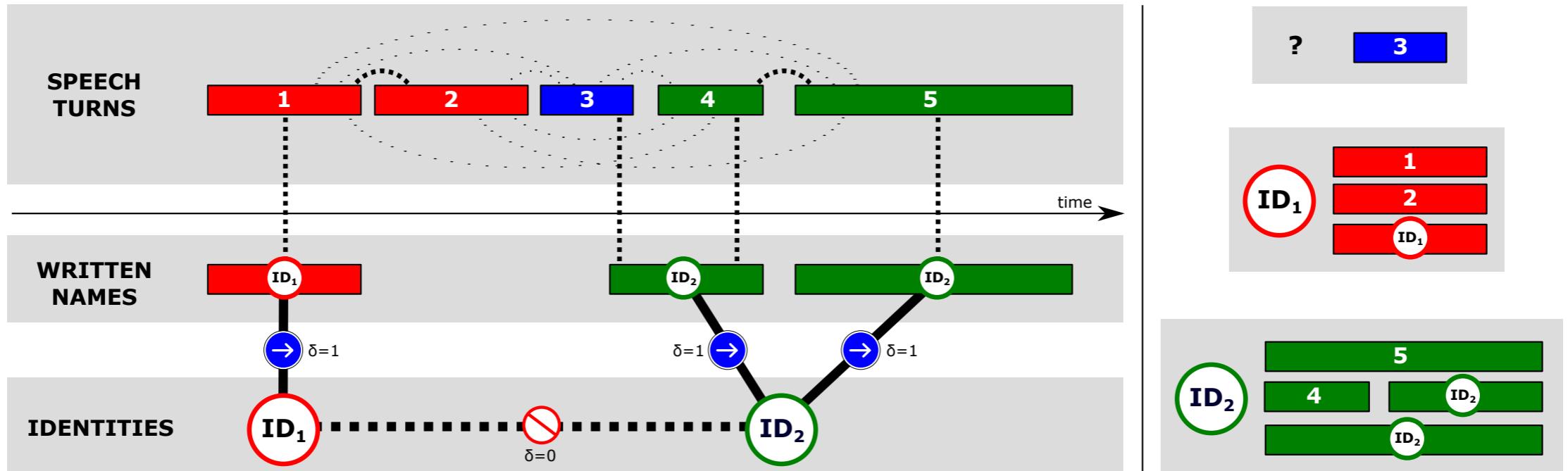
# expected results



# additional constraints



# additional constraints



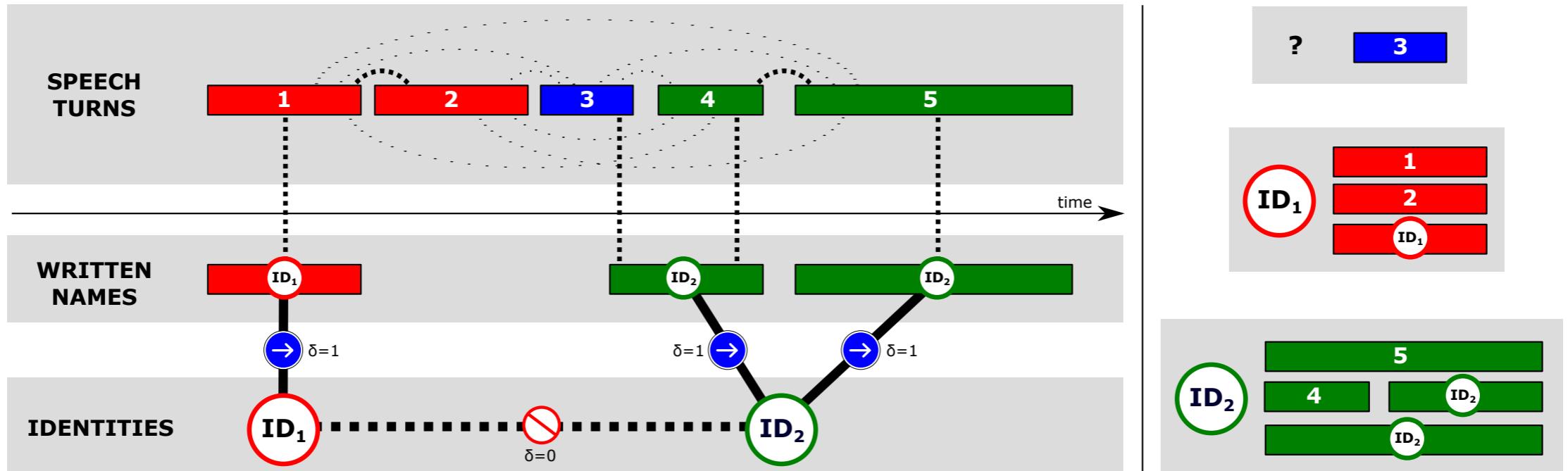
two identity vertices  
cannot end up  
in the same cluster

$$\forall (i, i') \in \mathcal{I}^2, i \neq i' \implies \delta_{ii'} = 0$$

any instance has at most one identity

$$\forall v \in \mathcal{V}, \sum_{i \in \mathcal{I}} \delta_{vi} \leq 1$$

# additional constraints



two identity vertices  
cannot end up  
in the same cluster

$$\forall (i, i') \in \mathcal{I}^2, i \neq i' \implies \delta_{ii'} = 0$$

any instance has at most one identity

$$\forall v \in \mathcal{V}, \sum_{i \in \mathcal{I}} \delta_{vi} \leq 1$$

any written or spoken instance  
must be in the same cluster  
as their corresponding identity

$$\forall w \in \mathcal{W}, \delta_{wi_w} = 1$$

$$\forall s \in \mathcal{S}, \delta_{si_s} = 1$$

# multimodal objective function

$$\mathcal{L}^\alpha(\delta, \mathcal{E}, p) = |\mathcal{E}|^{-1} \left[ \underbrace{\alpha \cdot \sum_{(v,v') \in \mathcal{E}} \delta_{vv'} \cdot p_{vv'}}_{\text{intra-cluster similarity}} + (1 - \alpha) \cdot \underbrace{\sum_{(v,v') \in \mathcal{E}} (1 - \delta_{vv'}) \cdot (1 - p_{vv'})}_{\text{inter-cluster dissimilarity}} \right]$$

would result in giving **more weight** to **dominant type** of edges

# multimodal objective function

$$\mathcal{L}^\alpha(\delta, \mathcal{E}, p) = |\mathcal{E}|^{-1} \left[ \underbrace{\alpha \cdot \sum_{(v,v') \in \mathcal{E}} \delta_{vv'} \cdot p_{vv'}}_{\text{intra-cluster similarity}} + (1 - \alpha) \cdot \underbrace{\sum_{(v,v') \in \mathcal{E}} (1 - \delta_{vv'}) \cdot (1 - p_{vv'})}_{\text{inter-cluster dissimilarity}} \right]$$

would result in giving **more weight** to **dominant type** of edges

$$\mathcal{L}_\beta^\alpha(\delta, \mathcal{E}, p) = \sum_{\substack{x \in \{\mathcal{T}, \mathcal{W}, \mathcal{S}, \mathcal{I}\} \\ y \in \{\mathcal{T}, \mathcal{W}, \mathcal{S}, \mathcal{I}\}}} \beta_{xy} \cdot \mathcal{L}^{\alpha_{xy}}(\delta, \mathcal{E} \cap (x \times y), p) \quad \alpha_{xy} \in [0, 1] \quad \beta_{xy} \in [0, 1] \quad \sum_{x,y} \beta_{xy} = 1$$

**one objective function per edge type subgraph**

allows different weights  $\beta$  for different edge types

allows fine tuning of  $\alpha$  for each edge type

# multimodal objective function

$$\mathcal{L}^\alpha(\delta, \mathcal{E}, p) = |\mathcal{E}|^{-1} \left[ \underbrace{\alpha \cdot \sum_{(v,v') \in \mathcal{E}} \delta_{vv'} \cdot p_{vv'}}_{\text{intra-cluster similarity}} + (1 - \alpha) \cdot \underbrace{\sum_{(v,v') \in \mathcal{E}} (1 - \delta_{vv'}) \cdot (1 - p_{vv'})}_{\text{inter-cluster dissimilarity}} \right]$$

would result in giving **more weight** to **dominant type** of edges

$$\mathcal{L}_\beta^\alpha(\delta, \mathcal{E}, p) = \sum_{\substack{x \in \{\mathcal{T}, \mathcal{W}, \mathcal{S}, \mathcal{I}\} \\ y \in \{\mathcal{T}, \mathcal{W}, \mathcal{S}, \mathcal{I}\}}} \beta_{xy} \cdot \mathcal{L}^{\alpha_{xy}}(\delta, \mathcal{E} \cap (x \times y), p) \quad \alpha_{xy} \in [0, 1] \quad \beta_{xy} \in [0, 1] \quad \sum_{x,y} \beta_{xy} = 1$$

**one objective function per edge type subgraph**

allows different weights  $\beta$  for different edge types

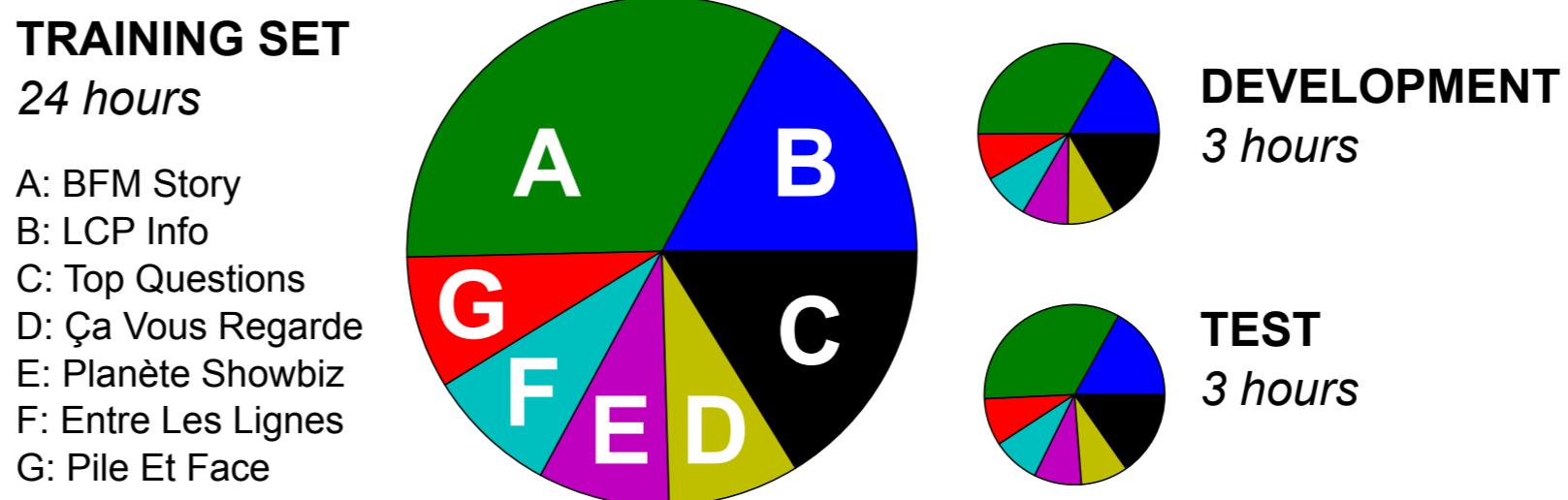
allows fine tuning of  $\alpha$  for each edge type

$$\delta^* = \underset{\delta \in \Delta_V}{\operatorname{argmax}} \mathcal{L}_\beta^\alpha(\delta, \mathcal{E}, p)$$

$\alpha$  and  $\beta$  are optimized by **random search**

# experimental protocol

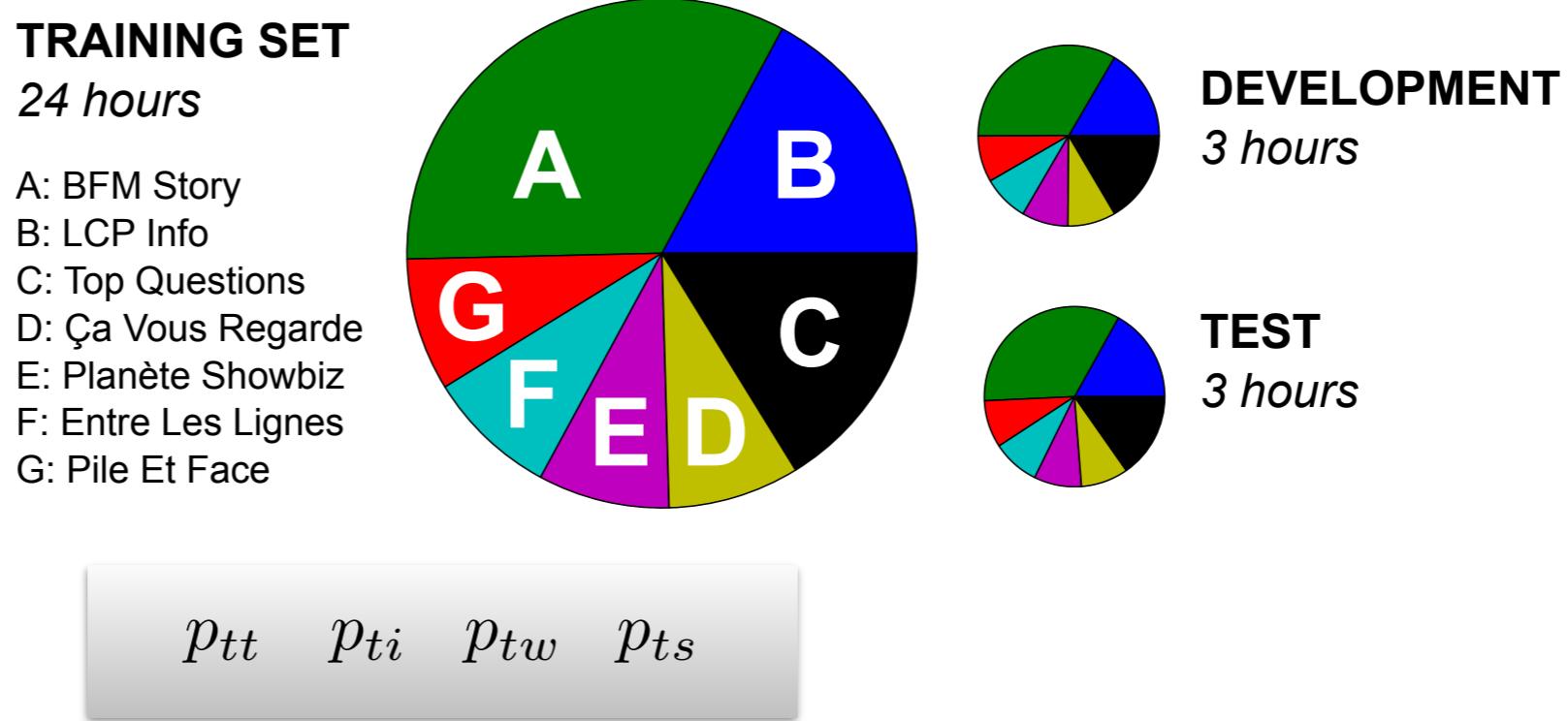
- corpora



- evaluation metrics
  - diarization** error rate
  - identification** error rate / **precision** / **recall**

# experimental protocol

- corpora



- evaluation metrics
  - diarization** error rate
  - identification** error rate / **precision** / **recall**

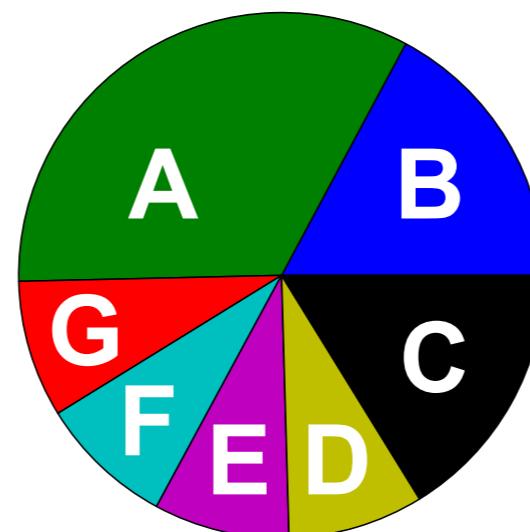
# experimental protocol

- corpora

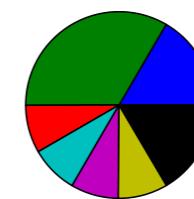
**TRAINING SET**

*24 hours*

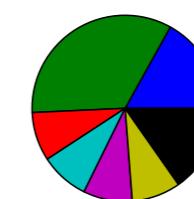
A: BFM Story  
B: LCP Info  
C: Top Questions  
D: Ça Vous Regarde  
E: Planète Showbiz  
F: Entre Les Lignes  
G: Pile Et Face



**DEVELOPMENT**  
*3 hours*



**TEST**  
*3 hours*



$p_{tt}$     $p_{ti}$     $p_{tw}$     $p_{ts}$

- evaluation metrics

**diarization** error rate

**identification** error rate / **precision** / **recall**

$XER \in \{\text{IER}, \text{DER}\}$

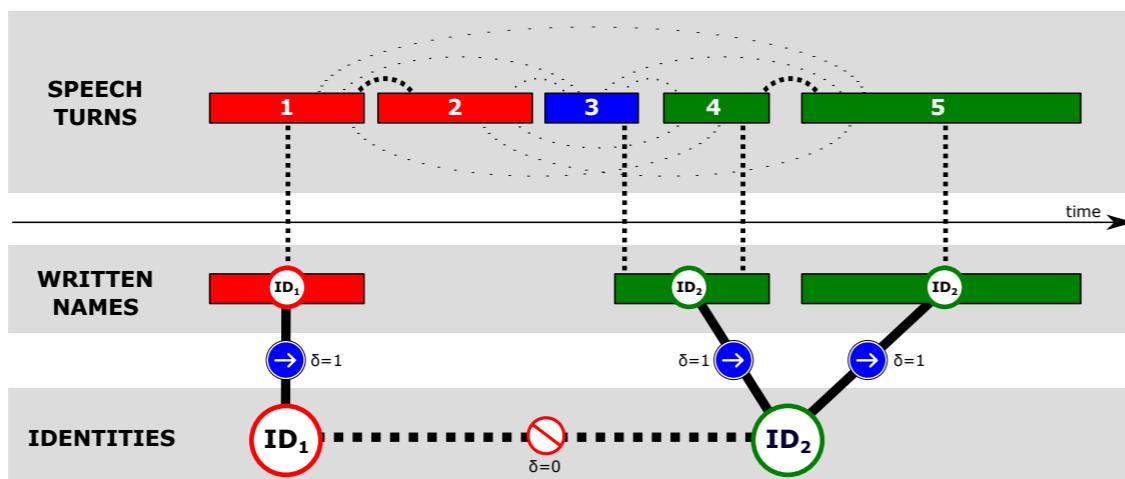
$$(\alpha^*, \beta^*) = \underset{\alpha, \beta}{\operatorname{argmin}} \mathbb{E}_{\text{dev}} [\text{XER}(r, h)]$$

# different graph configuration for different **tasks**

- speaker diarization  
speech turn vertices only
- speaker identification
  - **supervised mono-modal** speaker identification  
no written or spoken vertices
  - **unsupervised cross-modal** speaker identification  
no supervised edges  
only written or spoken vertices
  - **multi-modal** speaker identification  
the best of both worlds

# results speaker diarization

Vertices	Edges	DER	Purity	Coverage
$\mathcal{T}$	$t \leftrightarrow t'$	<b>21.1%</b>	<b>94.7%</b>	83.6%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_W$	$t \leftrightarrow t' \leftrightarrow w \leftrightarrow i_w$	18.3%	93.4%	85.8%
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	18.2%	94.0%	86.1%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_W \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \leftrightarrow i_w$	<b>17.8%</b>	93.9%	85.7%
BIC clustering baseline [1]		<b>19.8%</b>	<b>92.1%</b>	86.8%



written names do help 😊



supervision does help 😊



complementarity 😃



# results mono-modal speaker identification

Vertices	Edges	IER	P.	R.
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t$	49.4%	54.7%	54.3%
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	<b>47.9%</b>	55.9%	<b>55.8%</b>
GMM-UBM baseline		<b>48.9%</b>	57.5%	54.3%



- the first configuration is equivalent to usual open-set speaker identification

$$\text{ID}(t) = \begin{cases} i^* = \underset{i \in \mathcal{I}^*}{\operatorname{argmax}} p_{ti} & \text{if } p_{ti^*} > (1 - \alpha_{\mathcal{T}\mathcal{I}^*}) \\ \textcircled{?} & \text{otherwise.} \end{cases}$$

- GMM-UBM baseline relies on preliminary speaker diarization step and performs identification at cluster level

- adding  $t \leftrightarrow t'$  edges addresses this limitation

$$\beta_{\mathcal{T}\mathcal{T}} = 0.55 \quad \beta_{\mathcal{T}\mathcal{I}} = 0.45 \quad \alpha_{\mathcal{T}\mathcal{T}} = 0.19$$

# results mono-modal speaker identification

Vertices	Edges	IER	P.	R.
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t$	49.4%	54.7%	54.3%
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	<b>47.9%</b>	55.9%	<b>55.8%</b>
GMM-UBM baseline		<b>48.9%</b>	57.5%	54.3%



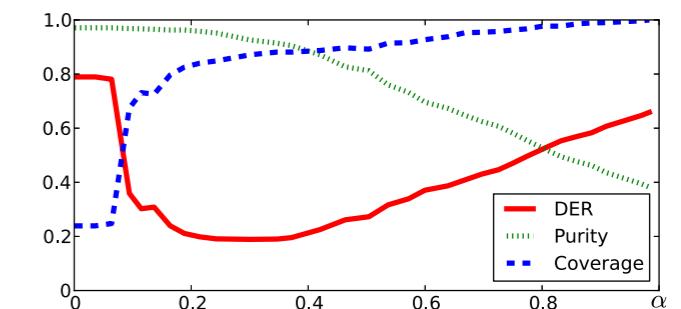
- the first configuration is equivalent to usual open-set speaker identification

$$\text{ID}(t) = \begin{cases} i^* = \underset{i \in \mathcal{I}^*}{\operatorname{argmax}} p_{ti} & \text{if } p_{ti^*} > (1 - \alpha_{\mathcal{T}\mathcal{I}^*}) \\ \text{?} & \text{otherwise.} \end{cases}$$

- GMM-UBM baseline relies on preliminary speaker diarization step and performs identification at cluster level

- adding  $t \leftrightarrow t'$  edges addresses this limitation

$$\beta_{\mathcal{T}\mathcal{T}} = 0.55 \quad \beta_{\mathcal{T}\mathcal{I}} = 0.45 \quad \alpha_{\mathcal{T}\mathcal{T}} = 0.19$$



# results cross-modal speaker identification

Vertices	Edges	IER	Precision	Recall
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	<b>46.5%</b>	66.8%	56.9%
$\mathcal{T} \cup \mathcal{S} \cup \mathcal{I}_{\mathcal{S}}$	$t \leftrightarrow t' \leftrightarrow s \Leftrightarrow i_s$	81.8%	21.5%	21.4%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}} \cup \mathcal{S} \cup \mathcal{I}_{\mathcal{S}}$	$i_w \Leftrightarrow w \leftrightarrow t \leftrightarrow t' \leftrightarrow s \Leftrightarrow i_s$	<b>45.6%</b>	62.7%	58.2%



# results cross-modal speaker identification

Vertices	Edges	IER	Precision	Recall
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_W$	$t \leftrightarrow t' \leftrightarrow w \leftrightarrow i_w$	<b>46.5%</b>	66.8%	56.9%
$\mathcal{T} \cup \mathcal{S} \cup \mathcal{I}_S$	$t \leftrightarrow t' \leftrightarrow s \leftrightarrow i_s$	81.8%	21.5%	21.4%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_W \cup \mathcal{S} \cup \mathcal{I}_S$	$i_w \Leftrightarrow w \Leftrightarrow t \Leftrightarrow t' \Leftrightarrow s \Leftrightarrow i_s$	<b>45.6%</b>	62.7%	58.2%

- to be compared with *oracle* performance

# results cross-modal speaker identification

Vertices	Edges	IER	Precision	Recall
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_W$	$t \leftrightarrow t' \leftrightarrow w \leftrightarrow i_w$	<b>46.5%</b>	66.8%	56.9%
$\mathcal{T} \cup \mathcal{S} \cup \mathcal{I}_S$	$t \leftrightarrow t' \leftrightarrow s \leftrightarrow i_s$	81.8%	21.5%	21.4%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_W \cup \mathcal{S} \cup \mathcal{I}_S$	$i_w \Leftrightarrow w \Leftrightarrow t \Leftrightarrow t' \Leftrightarrow s \Leftrightarrow i_s$	<b>45.6%</b>	62.7%	58.2%

- to be compared with *oracle* performance

Vertices	IER	Precision	Recall
$\mathcal{I}_W$	<b>39.3%</b>	100.0%	<b>63.8%</b>
$\mathcal{I}_S$	35.8%	100.0%	67.3%
$\mathcal{I}_W \cup \mathcal{I}_S$	21.5%	100.0%	81.9%

# results multi-modal speaker identification

Vertices	Edges	All			Anchors			All but anchors		
		IER	P.	R.	IER	P.	R.	IER	P.	R.
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	47.9%	55.9%	55.8%	<b>20.3%</b>	<b>86.6%</b>	<b>79.7%</b>	51.8%	50.8%	48.3%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	46.5%	66.8%	56.8%	79.4%	31.8%	20.6%	34.4%	76.9%	65.8%
$\mathcal{T} \cup \mathcal{I}^* \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	<b>25.3%</b>	<b>79.4%</b>	<b>78.6%</b>	23.9%	82.9%	76.1%	<b>22.4%</b>	<b>82.5%</b>	<b>77.9%</b>

# results multi-modal speaker identification

Vertices	Edges	All			Anchors			All but anchors		
		IER	P.	R.	IER	P.	R.	IER	P.	R.
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	47.9%	55.9%	55.8%	<b>20.3%</b>	<b>86.6%</b>	<b>79.7%</b>	51.8%	50.8%	48.3%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	46.5%	66.8%	56.8%	79.4%	31.8%	20.6%	34.4%	76.9%	65.8%
$\mathcal{T} \cup \mathcal{I}^* \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	<b>25.3%</b>	<b>79.4%</b>	<b>78.6%</b>	23.9%	82.9%	76.1%	<b>22.4%</b>	<b>82.5%</b>	<b>77.9%</b>

astonished  
face 

# results multi-modal speaker identification

Vertices	Edges	All			Anchors			All but anchors		
		IER	P.	R.	IER	P.	R.	IER	P.	R.
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	47.9%	55.9%	55.8%	<b>20.3%</b>	<b>86.6%</b>	<b>79.7%</b>	51.8%	50.8%	48.3%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_W$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	46.5%	66.8%	56.8%	79.4%	31.8%	20.6%	34.4%	76.9%	65.8%
$\mathcal{T} \cup \mathcal{I}^* \cup \mathcal{W} \cup \mathcal{I}_W$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	<b>25.3%</b>	<b>79.4%</b>	<b>78.6%</b>	23.9%	82.9%	76.1%	<b>22.4%</b>	<b>82.5%</b>	<b>77.9%</b>

astonished  
face



grinning  
face



# results multi-modal speaker identification

Vertices	Edges	All			Anchors			All but anchors		
		IER	P.	R.	IER	P.	R.	IER	P.	R.
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	47.9%	55.9%	55.8%	<b>20.3%</b>	<b>86.6%</b>	<b>79.7%</b>	51.8%	50.8%	48.3%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_W$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	46.5%	66.8%	56.8%	79.4%	31.8%	20.6%	34.4%	76.9%	65.8%
$\mathcal{T} \cup \mathcal{I}^* \cup \mathcal{W} \cup \mathcal{I}_W$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	<b>25.3%</b>	<b>79.4%</b>	<b>78.6%</b>	23.9%	82.9%	76.1%	<b>22.4%</b>	<b>82.5%</b>	<b>77.9%</b>

astonished  
face



grinning  
face



- **the best of both worlds**

- supervised for anchors  
lots of training data
- cross-modal for guests or reporters  
often introduced by overlaid names
- **com-ple-men-ta-ry** approaches

# results multi-modal speaker identification

Vertices	Edges	All			Anchors			All but anchors		
		IER	P.	R.	IER	P.	R.	IER	P.	R.
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	47.9%	55.9%	55.8%	<b>20.3%</b>	<b>86.6%</b>	<b>79.7%</b>	51.8%	50.8%	48.3%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_W$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	46.5%	66.8%	56.8%	79.4%	31.8%	20.6%	34.4%	76.9%	65.8%
$\mathcal{T} \cup \mathcal{I}^* \cup \mathcal{W} \cup \mathcal{I}_W$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	<b>25.3%</b>	<b>79.4%</b>	<b>78.6%</b>	23.9%	82.9%	76.1%	<b>22.4%</b>	<b>82.5%</b>	<b>77.9%</b>

- **the best of both worlds**

- supervised for anchors  
lots of training data
- cross-modal for guests or reporters  
often introduced by overlaid names
- **com-ple-men-ta-ry** approaches

astonished  
face



grinning  
face



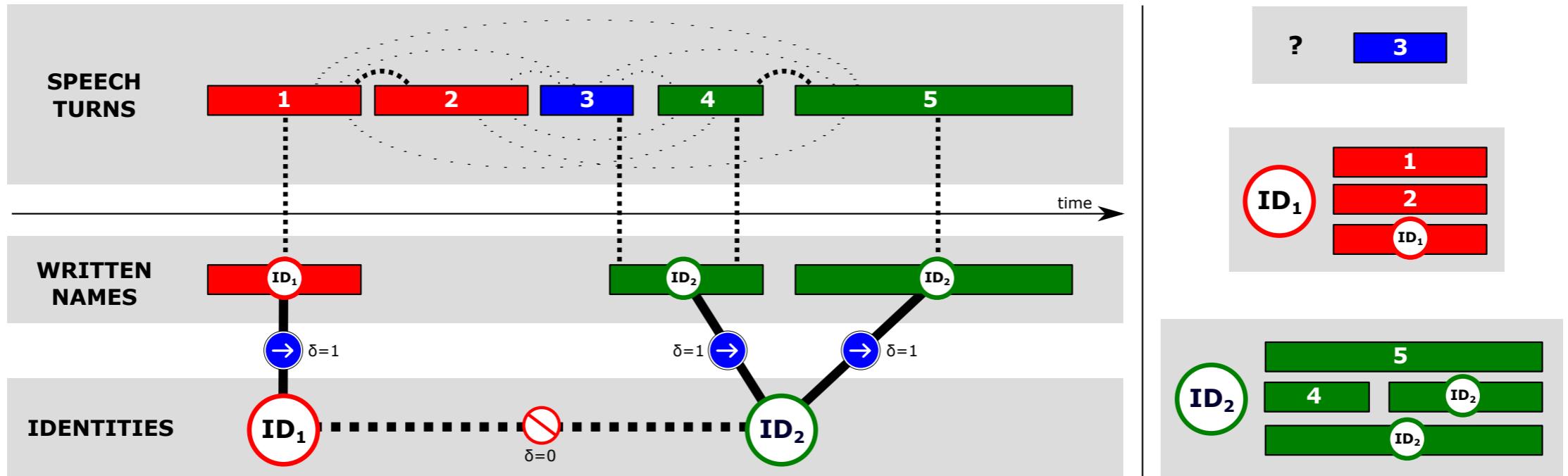
confused  
face



some room for  
improvement  
about anchors



# transitivity constraints **relaxation**



**strict** transitivity constraints

$$\forall (v, v', v'') \in \mathcal{V}^3,$$

$$\delta_{vv'} = 1 \wedge \delta_{v'v''} = 1 \implies \delta_{vv''} = 1$$

**loose** transitivity constraints

$$\forall (v, v', i) \in (\mathcal{V} \setminus \mathcal{I})^2 \times \mathcal{I},$$

$$\delta_{vi} = 1 \wedge \delta_{v'i} = 1 \not\implies \delta_{vv'} = 1$$

# results

## transitivity constraints relaxation

**strict** transitivity constraints

$$\forall (v, v', v'') \in \mathcal{V}^3,$$

$$\delta_{vv'} = 1 \wedge \delta_{v'v''} = 1 \implies \delta_{vv''} = 1$$

**loose** transitivity constraints

$$\forall (v, v', i) \in (\mathcal{V} \setminus \mathcal{I})^2 \times \mathcal{I},$$

$$\delta_{vi} = 1 \wedge \delta_{v'i} = 1 \not\implies \delta_{vv'} = 1$$

Constraints	Strict	Loose
DER	17.8%	20.0%
Purity	93.9%	88.2%
Coverage	85.7%	91.1%
IER	27.6%	25.3%
Precision	85.3%	79.4%
Recall	75.9%	78.6%

# conclusion

- a **unified** framework for
  - mono-
  - cross- modal person recognition
  - multi-

- what's **next**?
  - alternative objective functions

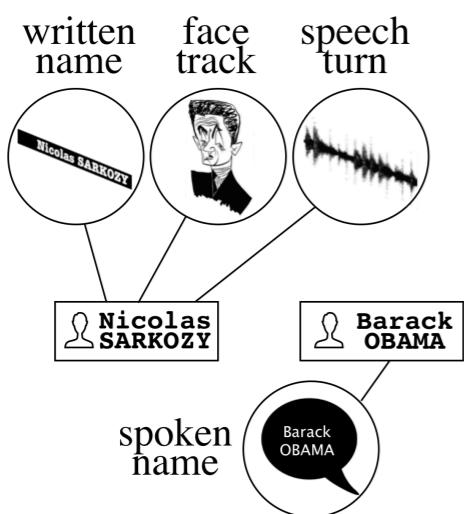
replace sum of probabilities by their product  
take instance durations into account

- extension to **face** modality
  - $f \leftrightarrow f'$  face clustering
  - $f \leftrightarrow i$  supervised face recognition
  - $t \leftrightarrow f$  audio-based face recognition

demo

# QCOMPÈRE

BY  Quaero



person instance **graphs**  
for person **recognition**  
in **multimedia** data

application to **speaker** identification in TV broadcast