“Sheldon speaking, bonjour !”

Leveraging **multilingual** tracks for **(weakly) supervised speaker identification**

Hervé Bredin
Anindya Roy
Nicolas Pécheux
Alexandre Allauzen

CNRS // LIMSI // France
bredin@limsi.fr // herve.niderb.fr

motivations

TV series are very popular…

motivations

TV series are very popular… *« Bazinga »*

motivations

TV series are very popular…

« Bazinga »
« Winter is coming »

motivations

TV series are very popular…

« Bazinga »
« Winter is coming »
« So say we all »

motivations

TV series are very popular…

« Bazinga »
« Winter is coming »
« So say we all »

… but TV series are not (easily) searchable.

find all video sequences where
Sheldon (honestly) apologizes to Leonard

motivations

TV series are very popular…

« Bazinga »
« Winter is coming »
« So say we all »

… but TV series are not (easily) searchable.

find all video sequences where
Sheldon (honestly) apologizes to Leonard

speech transcription

I'm really sorry, Leonard.

motivations

TV series are very popular…

« Bazinga »
« Winter is coming »
« So say we all »

… but TV series are not (easily) searchable.

find all video sequences where
Sheldon (honestly) apologizes to Leonard

speech transcription — ***I'm really sorry, Leonard.***

speaker identification — ***SHELDON:*** *«* ***I'm really sorry, Leonard.*** *»*

subtitles + transcripts= ❤️

subtitles + transcripts= ❤️

DVD subtitles

00:11.27 → 00:13.83
-...it will not've gone
through both slits.
-Agreed.

00:13.99 → 00:15.06
What's your point?

00:15.23 → 00:18.07
There's no point. I just
think it's a good idea
for a T-shirt

subtitles + transcripts= ❤️

DVD subtitles

00:11.27 → 00:13.83
-...it will not've gone
through both slits.
-Agreed.

00:13.99 → 00:15.06
What's your point?

00:15.23 → 00:18.07
There's no point. I just
think it's a good idea
for a T-shirt

fan transcripts

Sheldon:
it will not have gone
through both slits.

Leonard:
Agreed, what's your point?

Sheldon:
There's no point, I just
think it's a good idea
for a tee-shirt.

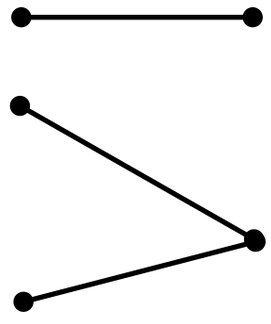
subtitles + transcripts= ❤️

DVD subtitles

00:11.27 → 00:13.83
-...it will not've gone
through both slits.
-Agreed.

00:13.99 → 00:15.06
What's your point?

00:15.23 → 00:18.07
There's no point. I just
think it's a good idea
for a T-shirt



Sheldon:
it will not have gone
through both slits.

Leonard:
Agreed, what's your point?

Sheldon:
There's no point, I just
think it's a good idea
for a tee-shirt.

fan transcripts

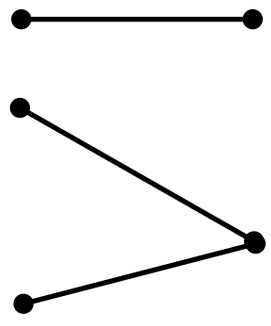
subtitles + transcripts= ❤️

DVD subtitles

00:11.27 → 00:13.83
-...it will not've gone
through both slits.
-Agreed.

00:13.99 → 00:15.06
What's your point?

00:15.23 → 00:18.07
There's no point. I just
think it's a good idea
for a T-shirt



Sheldon:
it will not have gone
through both slits.

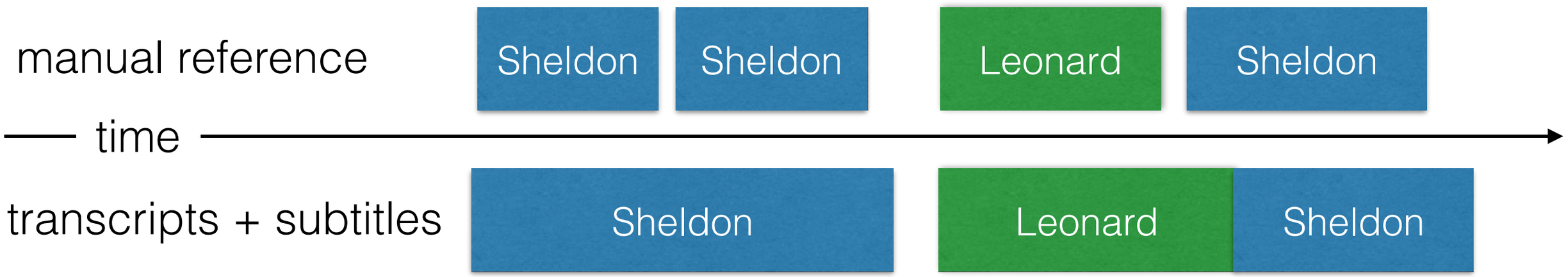
Leonard:
Agreed, what's your point?

Sheldon:
There's no point, I just
think it's a good idea
for a tee-shirt.

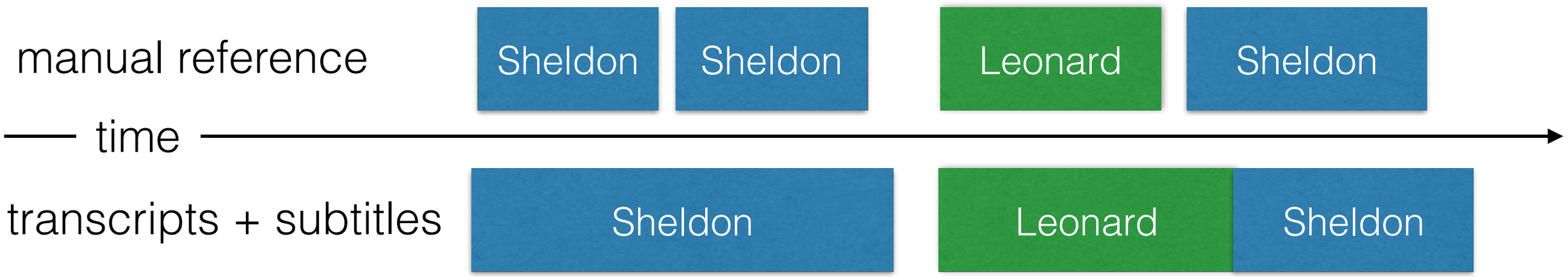
fan transcripts

00:11.27 → 00:13.37	Sheldon
00:13.37 → 00:13.83	Leonard
00:13.99 → 00:15.06	Leonard
00:15.23 → 00:18.07	Sheldon

motivations (cont.)

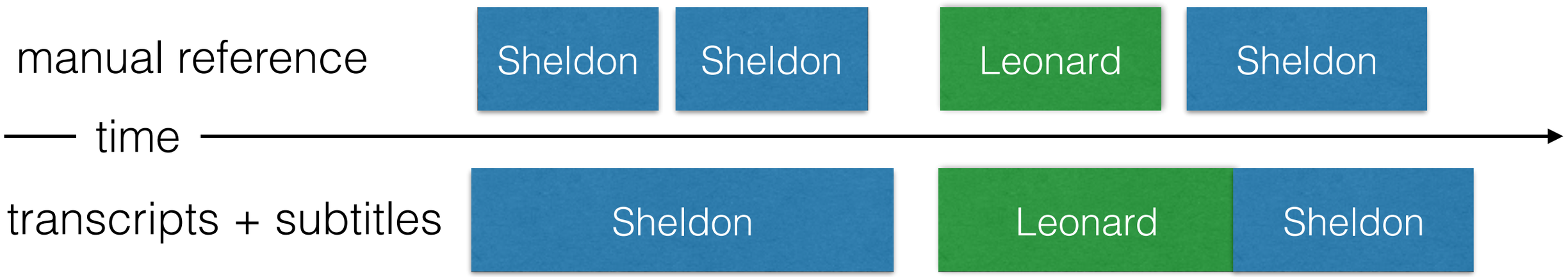


motivations (cont.)



- when both transcripts and subtitles are available
  no need to go any further (<10% error)

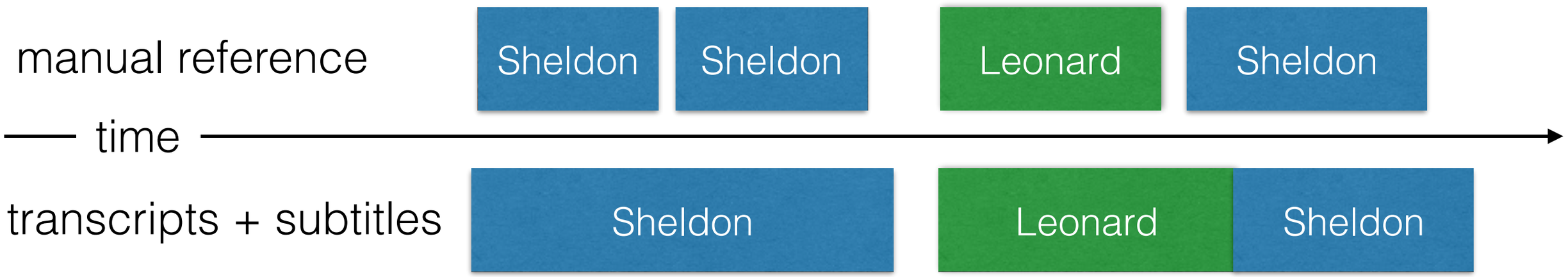
motivations (cont.)



- when both transcripts and subtitles are available
  no need to go any further (<10% error)

- what about episodes for which they are not yet available?

motivations (cont.)



- when both transcripts and subtitles are available
  no need to go any further (<10% error)

- what about episodes for which they are not yet available?

✨ automatic speaker identification

outline

- weakly supervised speaker identification
- bilingual speaker identification
- conclusion

speaker identification in TV series

speaker identification in TV series

- 👍🏻 (relatively) small set of characters

speaker identification in TV series

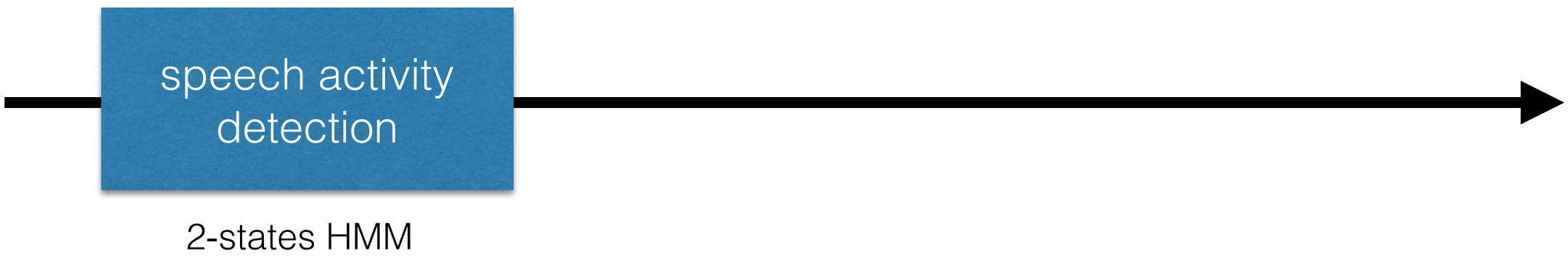
- 👍🏻 (relatively) small set of characters
- 👎🏻 spontaneous speech
  short speech turns, fast interactions

speaker identification in TV series

- 👍🏻 (relatively) small set of characters
- 👎🏻 spontaneous speech
  short speech turns, fast interactions

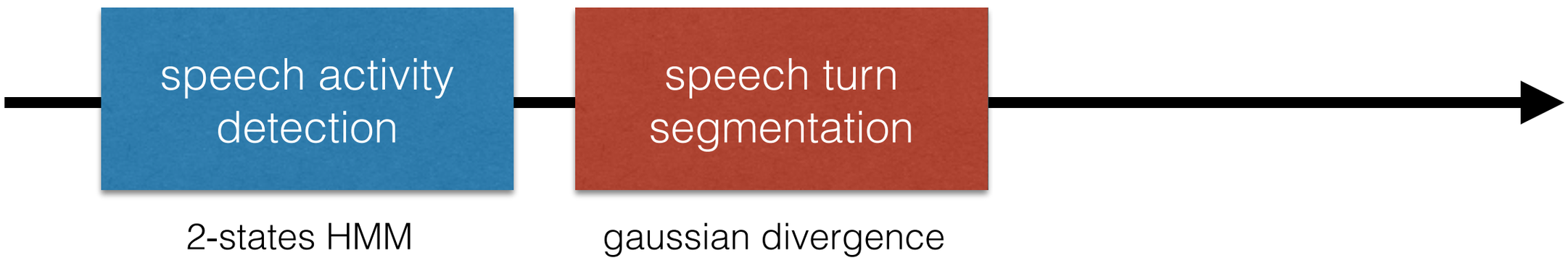
speaker identification in TV series

- 👍🏻 (relatively) small set of characters
- 👎🏻 spontaneous speech
  short speech turns, fast interactions



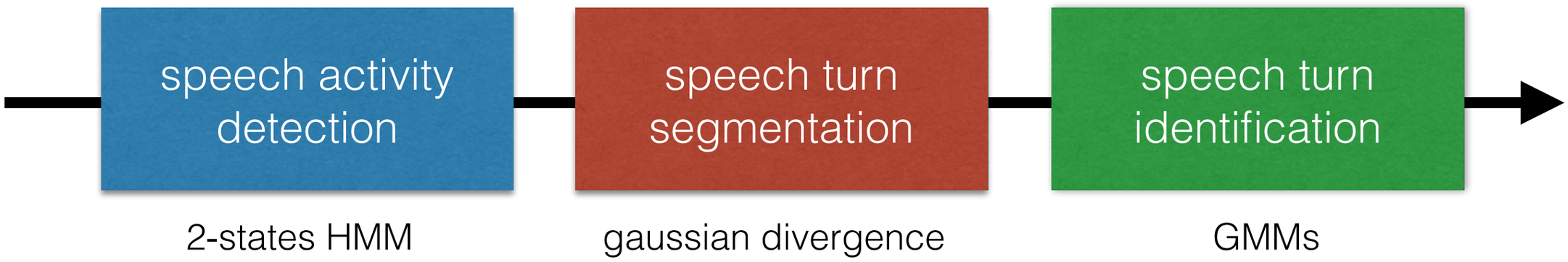
speaker identification in TV series

- 👍🏻 (relatively) small set of characters
- 👎🏻 spontaneous speech
  short speech turns, fast interactions



speaker identification in TV series

- 👍🏻 (relatively) small set of characters
- 👎🏻 spontaneous speech
  short speech turns, fast interactions



experimental setup

episodes 1 to 6 (**2 hours**)

leave-one-episode-out **cross-validation**

manual reference: 6 labels
(Sheldon, Leonard, Penny, Howard, Raj, other)

audio tracks & **subtitles** (English & French)
transcripts (English)

detection error rate $\frac{\text{miss} + \text{fa}}{\text{speech}}$

identification error rate $\frac{\text{miss} + \text{fa} + \text{confusion}}{\text{speech}}$

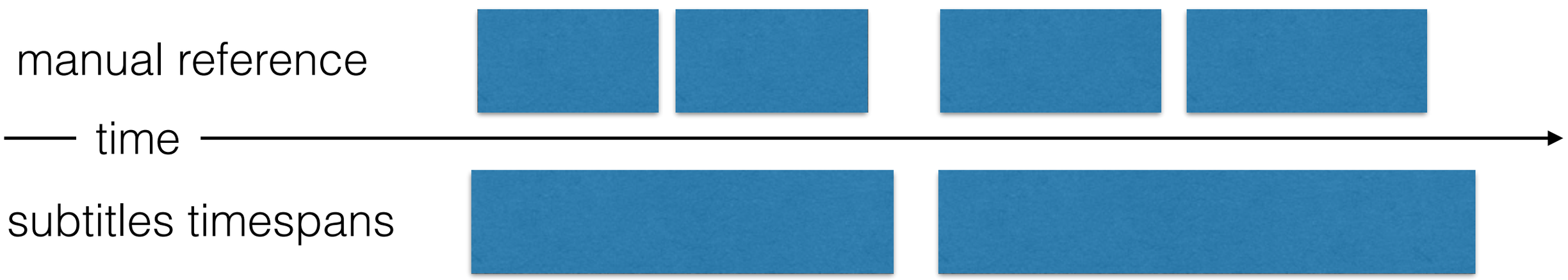
weakly supervised

- 2-class (speech vs. non-speech) hidden Markov model
  based on standard MFCC coefficients

weakly supervised

- 2-class (speech vs. non-speech) hidden Markov model
  based on standard MFCC coefficients

- ✨ use (free) subtitle timespans for training
  instead of (costly) manual reference



results speech activity detection

Approach	Subtitles	**Fully supervised**	**Weakly supervised**
DER	19.8%	**7.8%**	**8.1%**
Precision	74.6%	94.8%	91.2%
Recall	95.6%	90.4%	94.1%

DER // Detection error rate $\frac{\text{miss} + \text{fa}}{\text{speech}}$

results speech activity detection

Approach	Subtitles	**Fully supervised**	**Weakly supervised**
DER	19.8%	**7.8%**	**8.1%**
Precision	74.6%	94.8%	91.2%
Recall	95.6%	90.4%	94.1%

DER // Detection error rate $\frac{\text{miss} + \text{fa}}{\text{speech}}$

- subtitles timespans contain **25% non-speech**

results speech activity detection

Approach	Subtitles	Fully supervised	Weakly supervised
DER	19.8%	**7.8%**	**8.1%**
Precision	74.6%	94.8%	91.2%
Recall	95.6%	90.4%	94.1%

DER // Detection error rate $\frac{\text{miss} + \text{fa}}{\text{speech}}$

- subtitles timespans contain **25% non-speech**

- **precision** is better for **fully** supervised
  recall is better fo **weakly** supervised

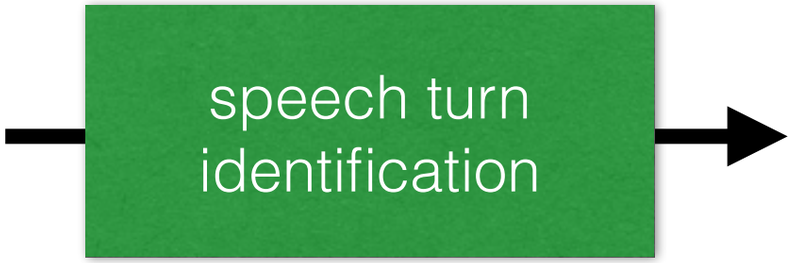
results speech activity detection

Approach	Subtitles	**Fully supervised**	**Weakly supervised**
DER	19.8%	**7.8%**	**8.1%**
Precision	74.6%	94.8%	91.2%
Recall	95.6%	90.4%	94.1%

DER // Detection error rate $\frac{\text{miss} + \text{fa}}{\text{speech}}$

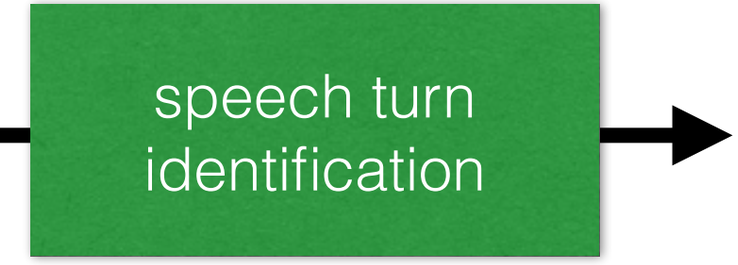
- subtitles timespans contain **25% non-speech**
- **precision** is better for **fully** supervised
  recall is better fo **weakly** supervised
- **weakly** supervised is almost **on par with fully** supervised

weakly supervised



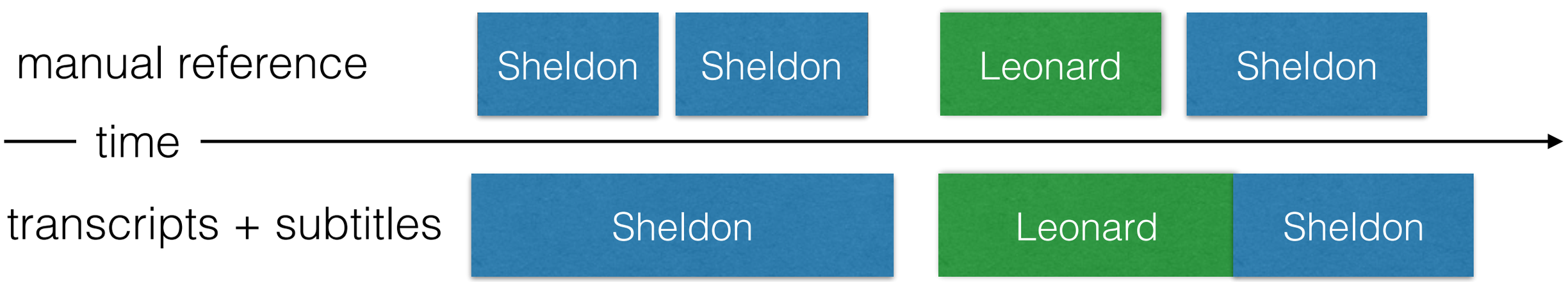
- 6-class Gaussian Mixture Models classification
  5 main characters vs. all others

weakly supervised



- 6-class Gaussian Mixture Models classification
  5 main characters vs. all others

- ✨ use (free) subtitle + transcripts for training
  instead of (costly) manual reference



results speaker identification

Speaker identification approach	Speech activity detection		
	Reference	Fully supervised	Weakly supervised
Oracle	10.0%	24.5%	25.4%
Labeled subtitles	12.8%	27.0%	28.2%
Fully supervised	18.6%	35.9%	37.9%
Weakly supervised	18.5%	35.6%	37.8%

results speaker identification

Speaker identification approach	Speech activity detection		
	Reference	Fully supervised	Weakly supervised
Oracle	10.0%	24.5%	25.4%
Labeled subtitles	12.8%	27.0%	28.2%
Fully supervised	18.6%	35.9%	37.9%
Weakly supervised	18.5%	35.6%	37.8%

- **secondary characters** represent 10% of speech duration

results speaker identification

Speaker identification approach	Speech activity detection		
	Reference	Fully supervised	Weakly supervised
Oracle	10.0%	24.5%	25.4%
Labeled subtitles	12.8%	27.0%	28.2%
Fully supervised	18.6%	35.9%	37.9%
Weakly supervised	18.5%	35.6%	37.8%

- **secondary characters** represent 10% of speech duration
- we should **use subtitles and transcripts when available**

results speaker identification

Speaker identification approach	Speech activity detection		
	Reference	Fully supervised	Weakly supervised
Oracle	10.0%	24.5%	25.4%
Labeled subtitles	12.8%	27.0%	28.2%
Fully supervised	18.6%	35.9%	37.9%
Weakly supervised	18.5%	35.6%	37.8%

- **secondary characters** represent 10% of speech duration
- we should **use subtitles and transcripts when available**
- **weakly** supervised is almost **on par with fully** supervised

bilingual speaker identification

bilingual speaker identification

bilingual speaker identification

- some characters may be easier to distinguish using their French-dubbed voices than with their original American English voice

bilingual speaker identification

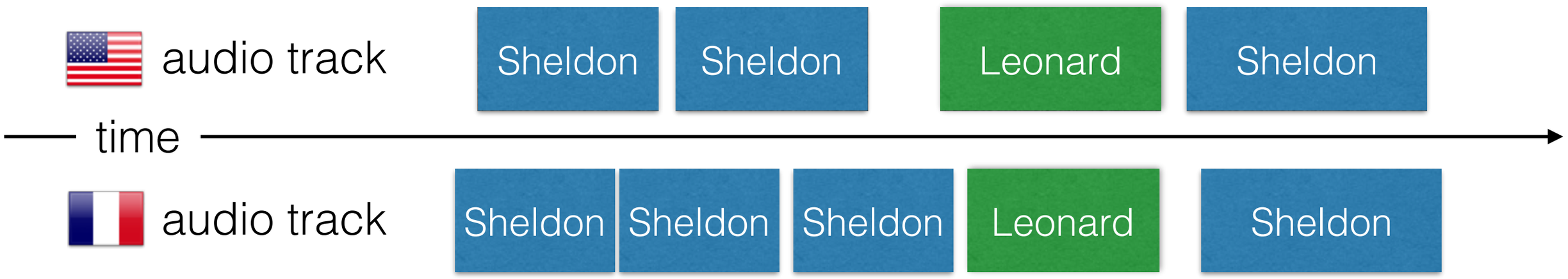
- some characters may be easier to distinguish using their French-dubbed voices than with their original American English voice
- ✨ train two speaker identification systems (🇺🇸 and 🇫🇷) and apply score fusion to improve the results

$$\rho_{ti} = \alpha \cdot \rho_{ti}^{\text{🇺🇸}} + (1 - \alpha) \cdot \rho_{ti}^{\text{🇫🇷}}$$

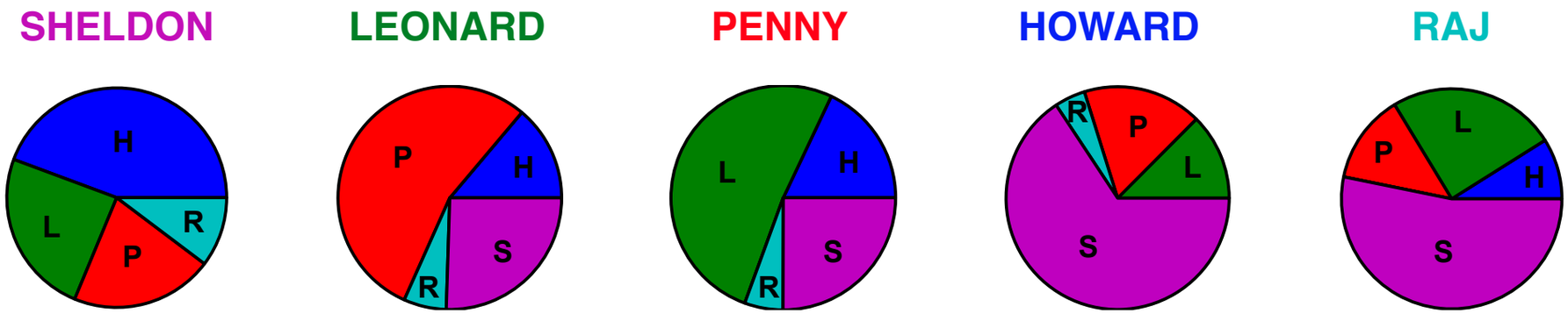
bilingual speaker identification

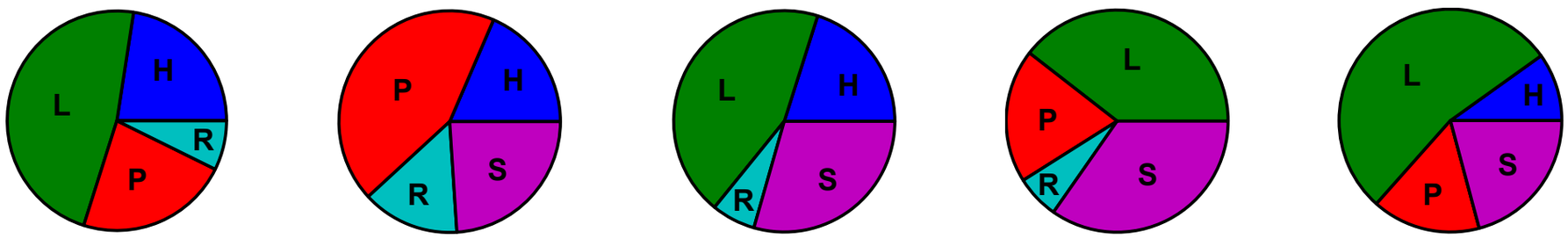
- some characters may be easier to distinguish using their French-dubbed voices than with their original American English voice
- ✨ train two speaker identification systems (🇺🇸 and 🇫🇷) and apply score fusion to improve the results

$$\rho_{ti} = \alpha \cdot \rho_{ti}^{\text{🇺🇸}} + (1 - \alpha) \cdot \rho_{ti}^{\text{🇫🇷}}$$

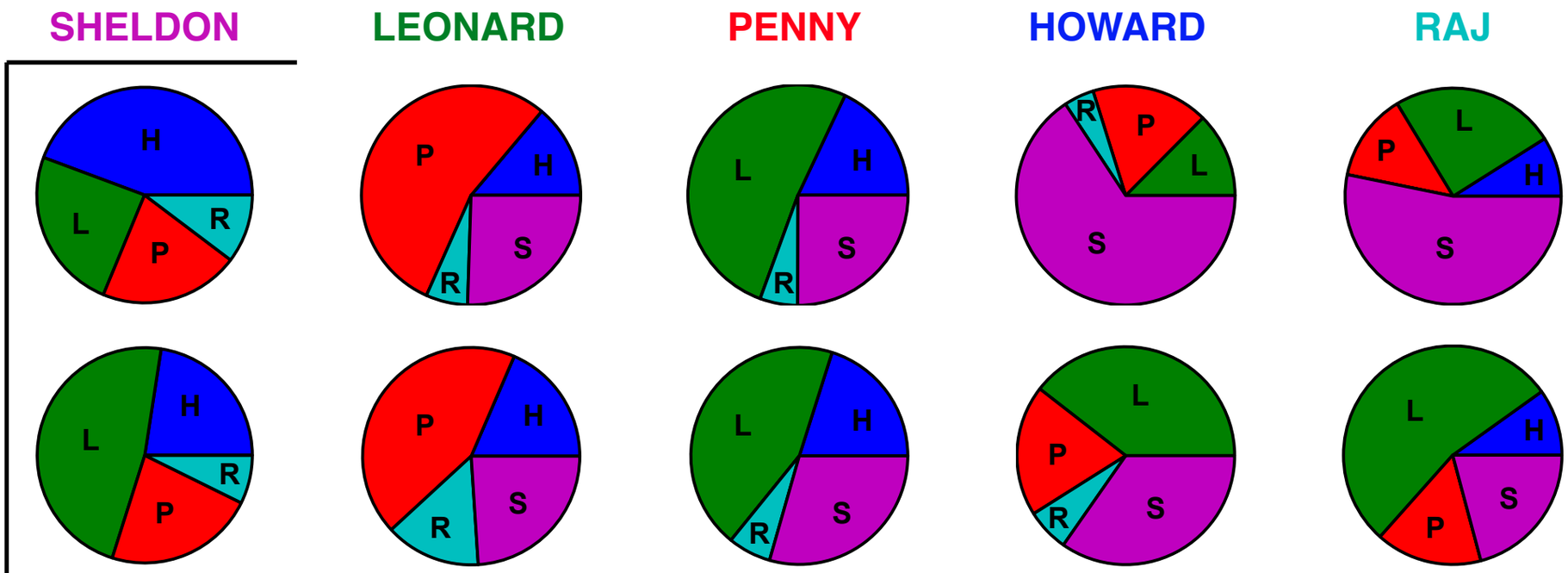


bilingual confusion matrix



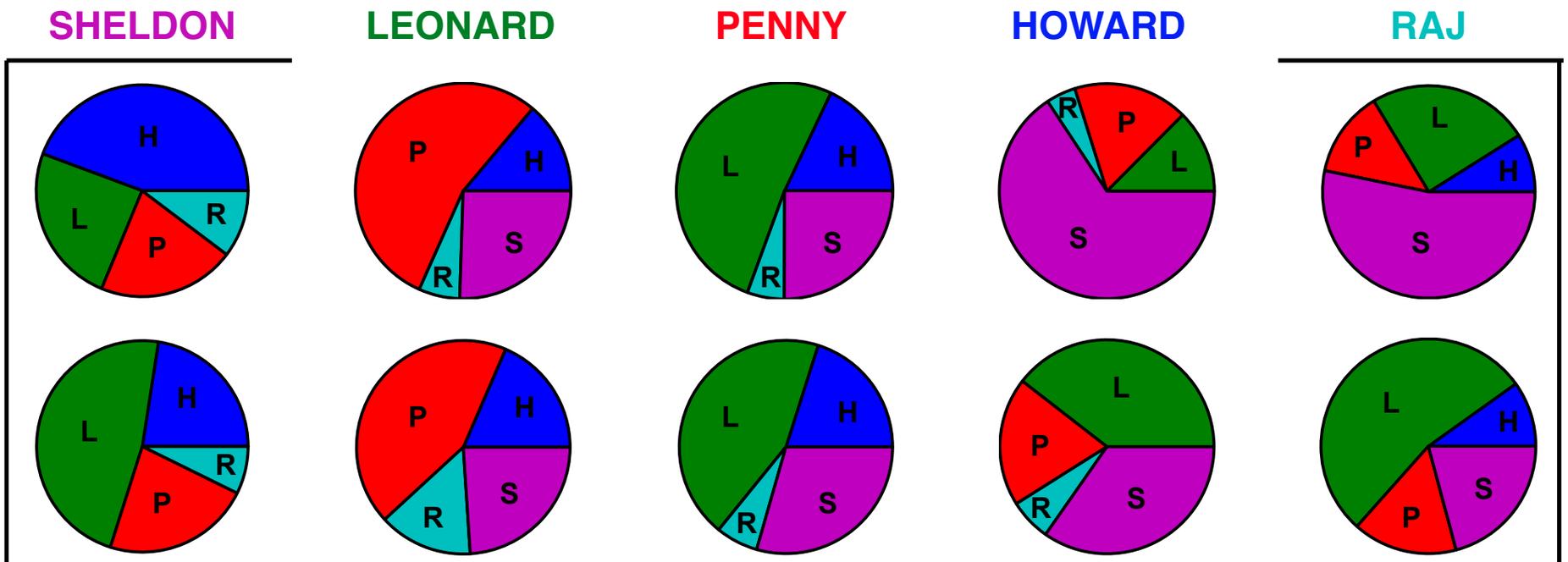


bilingual confusion matrix



SHELDON is mostly confused with **HOWARD**
SHELDON is mostly confused with **LEONARD**

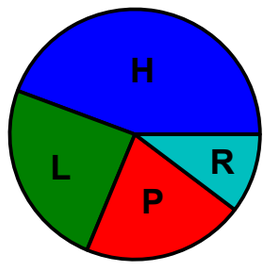
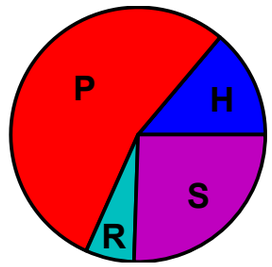
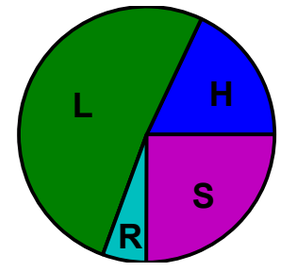
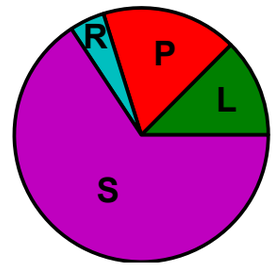
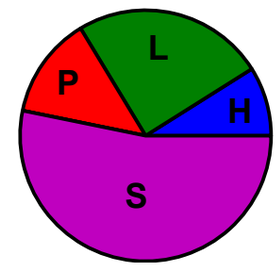
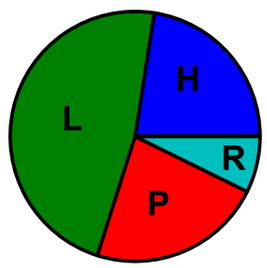
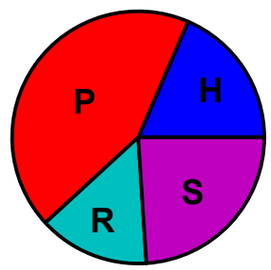
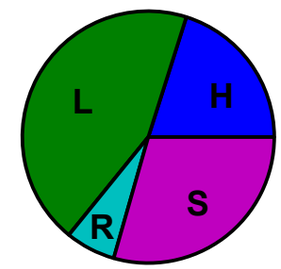
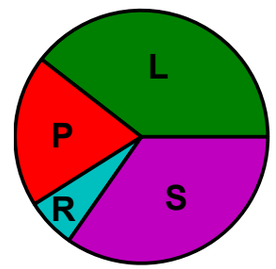
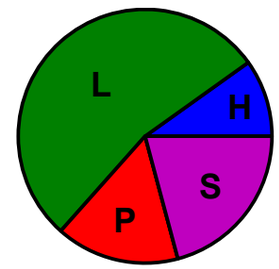
bilingual confusion matrix



SHELDON is mostly confused with **HOWARD**
SHELDON is mostly confused with **LEONARD**

RAJ is mostly confused with **SHELDON**
RAJ is mostly confused with **LEONARD**

bilingual confusion matrix

	SHELDON	LEONARD	PENNY	HOWARD	RAJ
					
				

SHELDON is mostly confused with **HOWARD**
SHELDON is mostly confused with **LEONARD**

RAJ is mostly confused with **SHELDON** 
RAJ is mostly confused with **LEONARD** 

language has no impact for **PENNY**
(she is the only female character)

results bilingual speaker identification

	IER	
🇺🇸	37.8 %	
🇺🇸 + 🇫🇷	32.8 %	
	-13 %	

- bilingual beats monolingual

results bilingual speaker identification

	IER	**confusion 5 main characters**
🇺🇸	37.8 %	7.1 %
🇺🇸 + 🇫🇷	32.8 %	2.1 %
	-13 %	-70 %

- bilingual beats monolingual
- bilingual is almost perfect for 5 main characters

results bilingual speaker identification

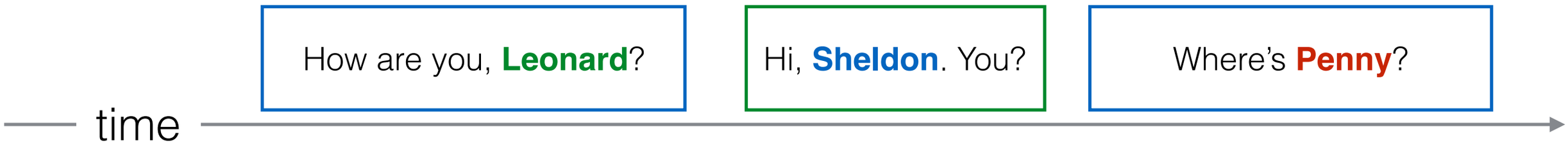
	IER	**confusion 5 main characters**
🇺🇸	37.8 %	7.1 %
🇺🇸 + 🇫🇷	32.8 %	2.1 %
	-13 %	-70 %

- bilingual beats monolingual
- bilingual is almost perfect for 5 main characters

- remaining errors
  speech turn segmentation: 15%
  secondary characters: 15%

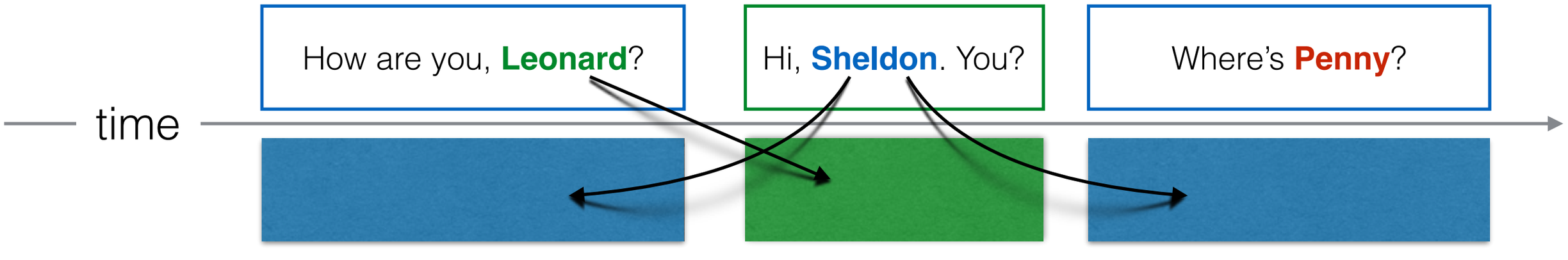
what's next?

- completely unsupervised speaker identification



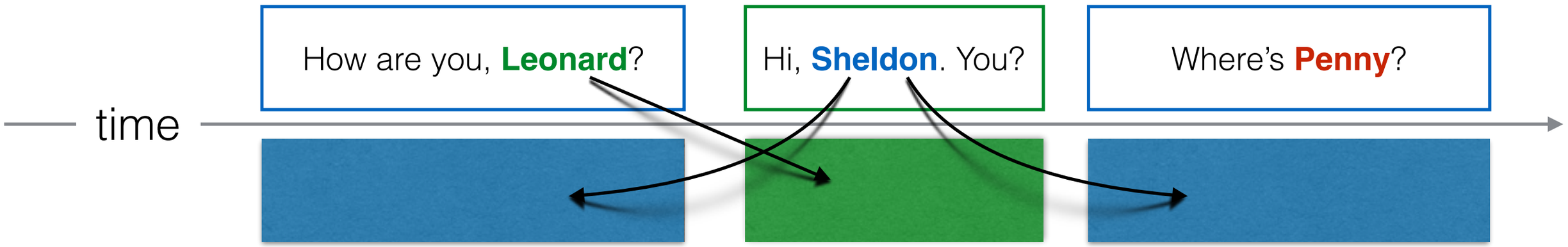
what's next?

- completely unsupervised speaker identification



what's next?

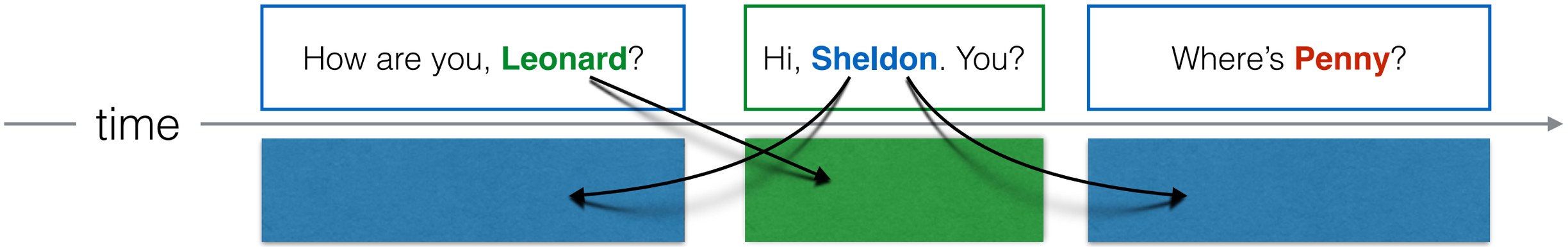
- completely unsupervised speaker identification



- audio-visual character recognition

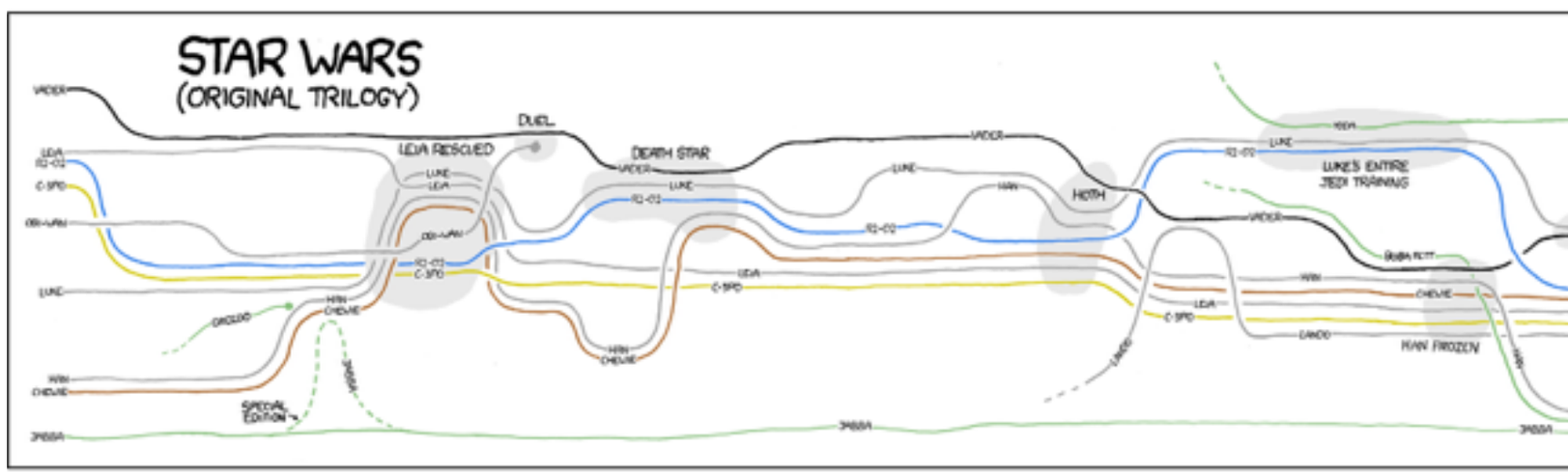
what's next?

- completely unsupervised speaker identification



- audio-visual character recognition

- *StoryGraph*
  Tapaswi et al.
  CVPR 2014



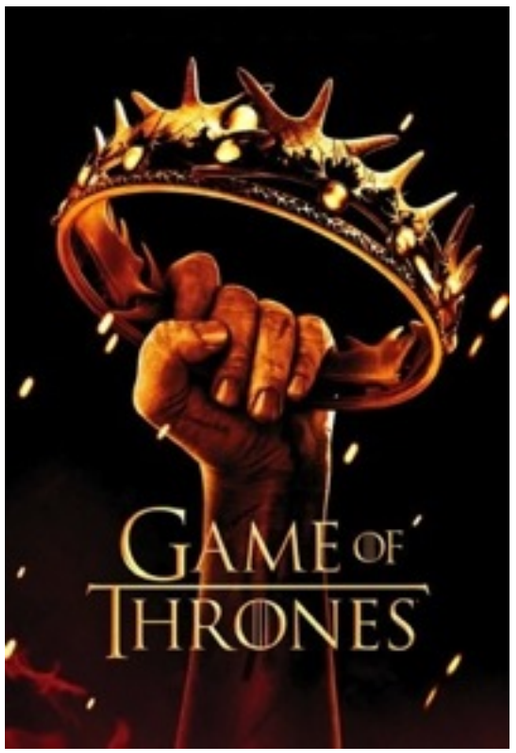
xkcd.com/657/

.niderb.fr

"bring your own DVDs" reproducible corpus

provides tool to automatically extract video, audio tracks and subtitles

speaker labels
episode outlines
episode transcript
forced alignment



episode outlines
episode transcript
forced alignment
scene segmentation

your
TVD
plugin
goes
here

QUESTIONS?

reproducible corpus

tvd.niderb.fr/corpus

reproducible research

tvd.niderb.fr/research

open-source Python libraries
(incl. speech processing modules)

github.com/pyannote

contact us at

metadatv.limsi.fr

Q&A slides

automatic alignment

1. Pre-processing

« if a photon is directed through a plane with two slits in it »

photon
direct
plane
slit

TF-IDF

automatic alignment

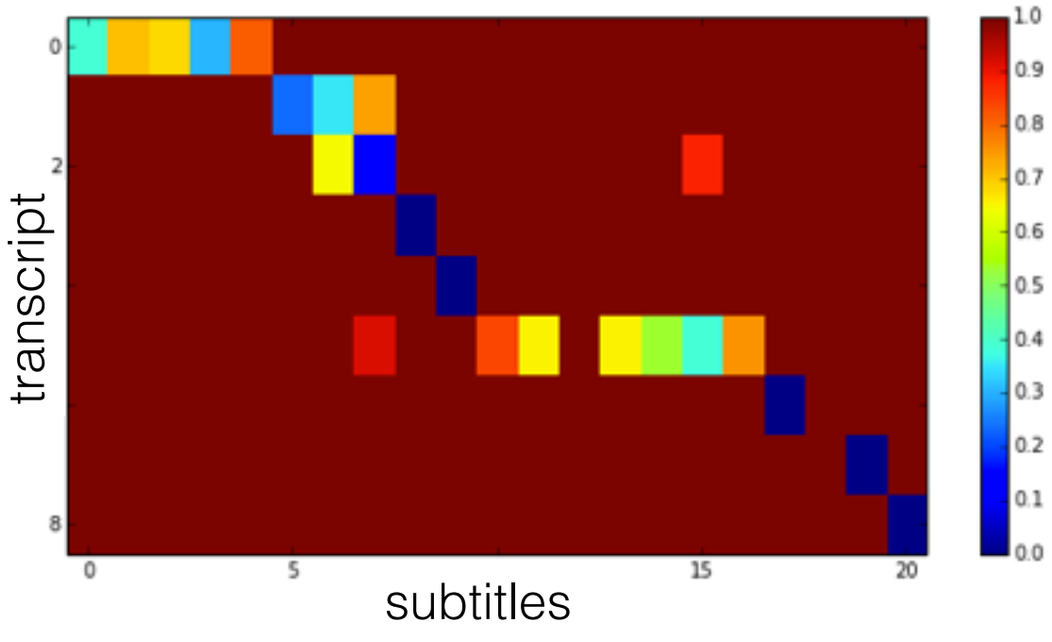
1. Pre-processing

« if a photon is directed through a plane with two slits in it »

photon
direct
plane
slit

TF-IDF

2. Cosine distance matrix



automatic alignment

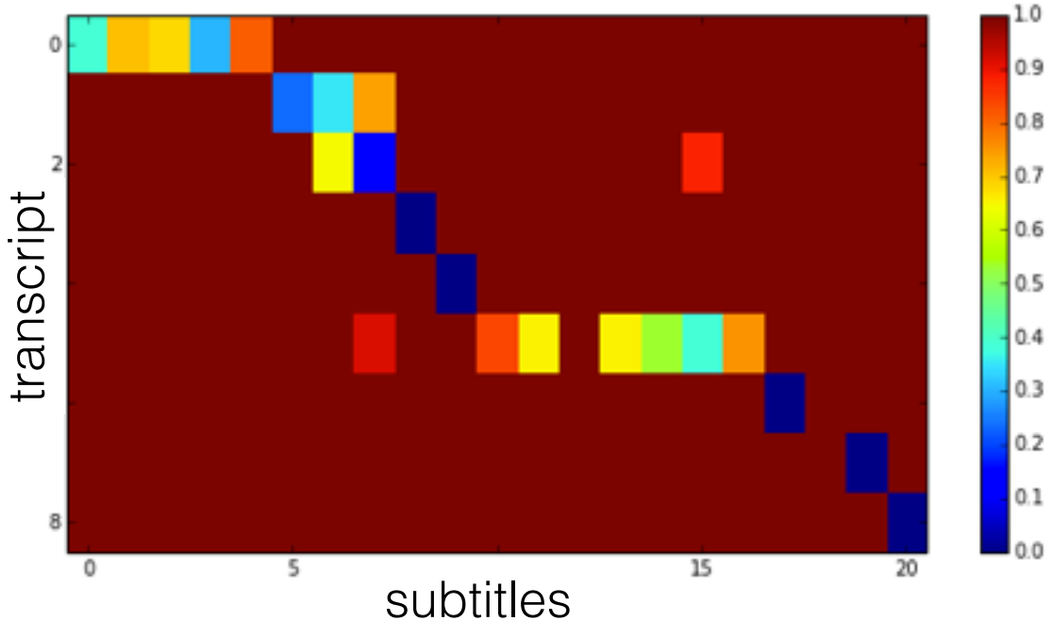
1. Pre-processing

« if a photon is directed through a plane with two slits in it »

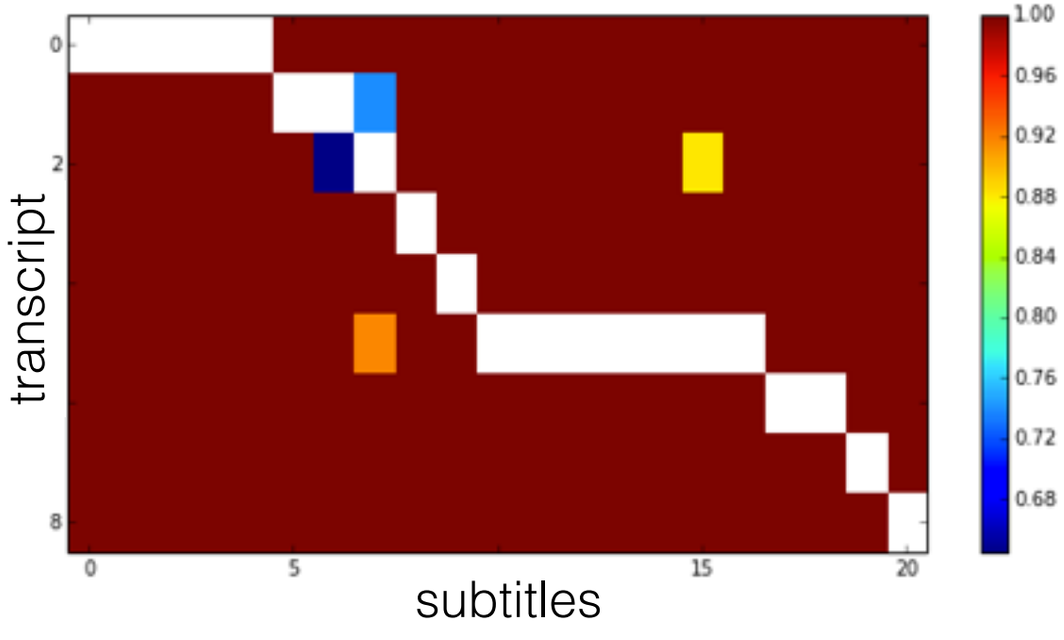
photon
direct
plane
slit

TF-IDF

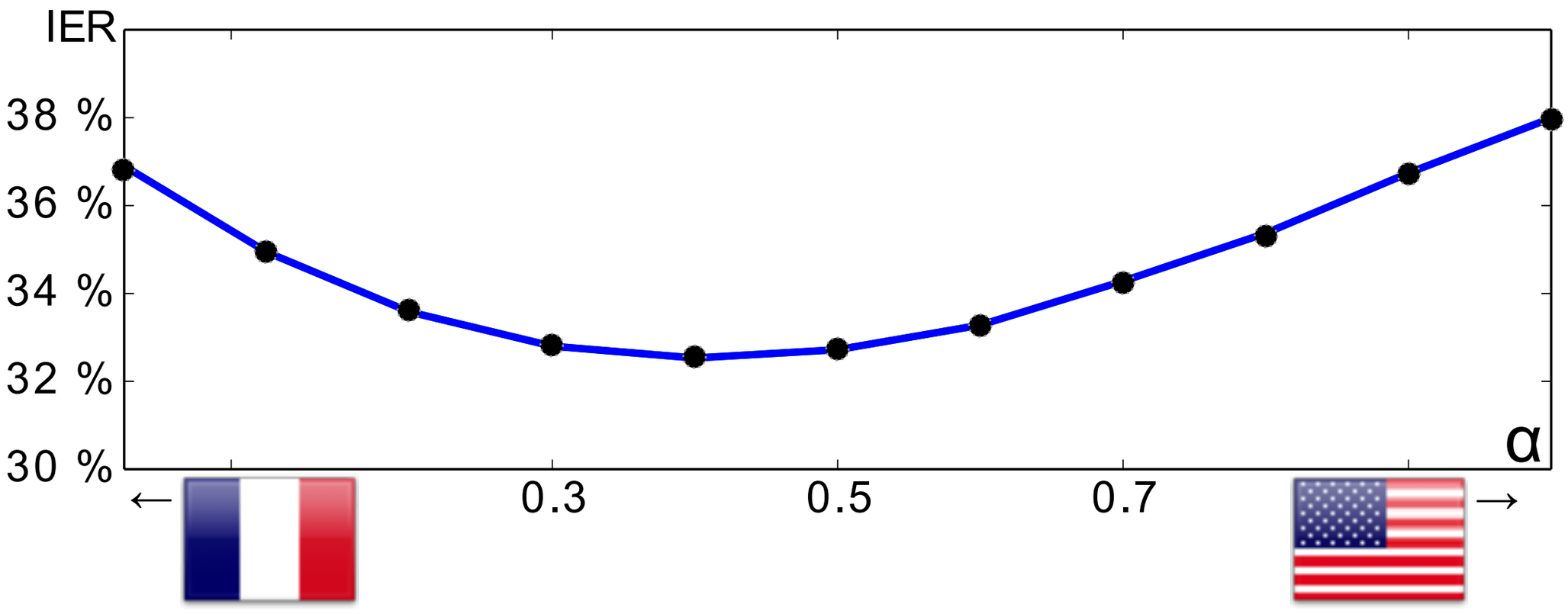
2. Cosine distance matrix



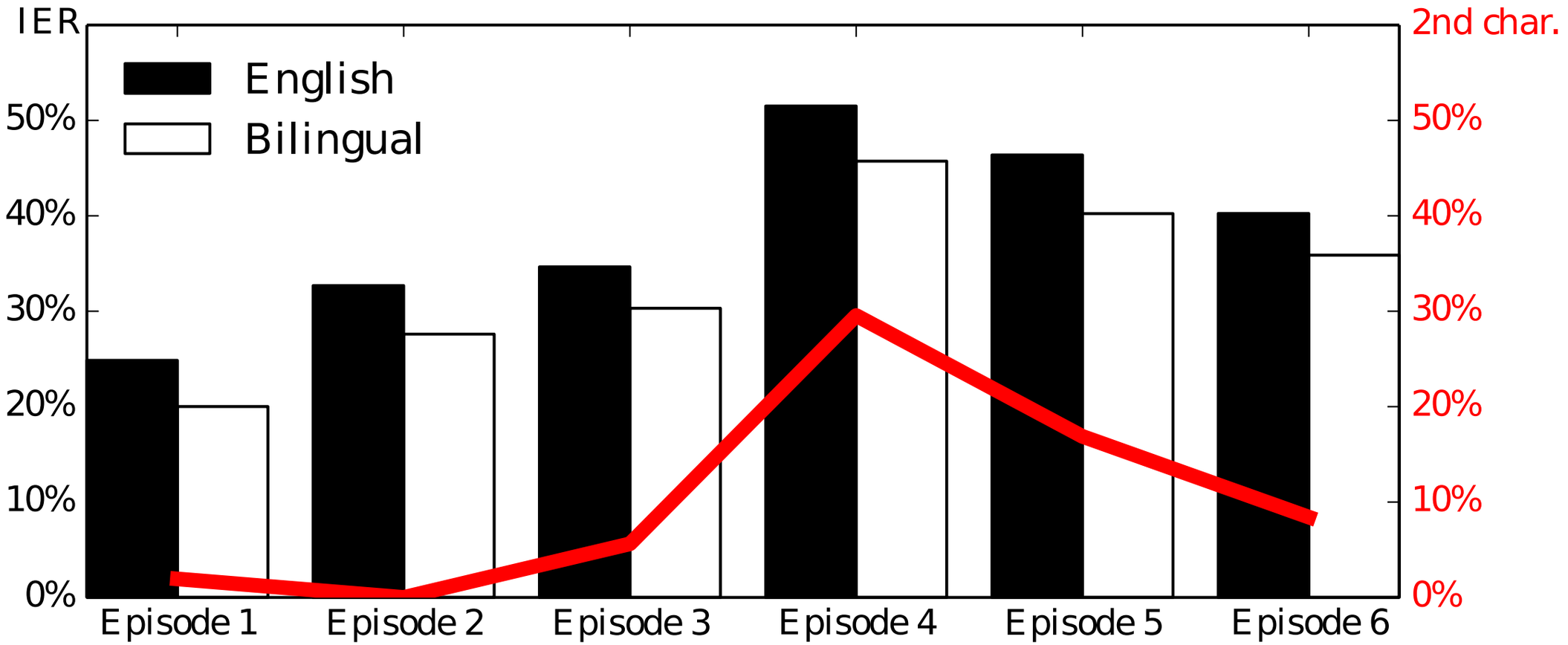
3. Dynamic time warping



influence of α



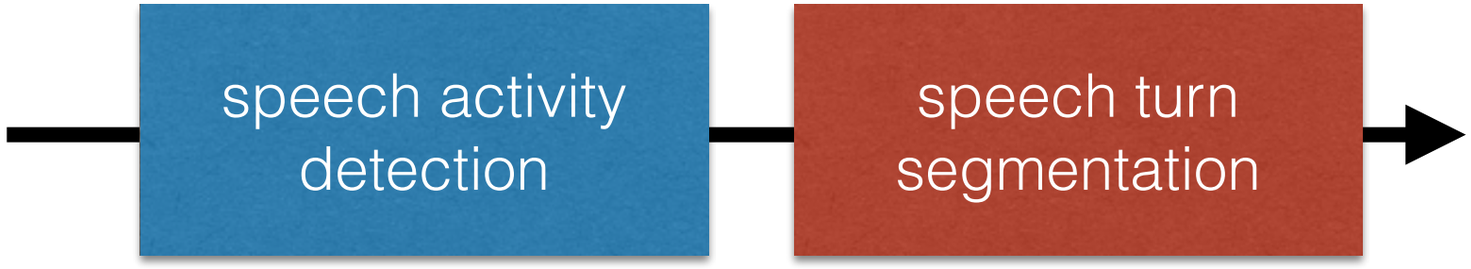
detailed performance



results speaker identification

Speaker identification approach	Segmentation		
	Reference	Fully supervised	Weakly supervised
Oracle	10.0%	24.5%	25.4%
Labeled subtitles	12.8%	27.0%	28.2%
Fully supervised	18.6%	35.9%	37.9%
Weakly supervised	18.5%	35.6%	37.8%

the **segmentation** step is the **main source of errors**



metadatv.limsi.fr

- Why reverse-engineer when we could have access to this information from the start?

- **Problem:** most of the metadata do get lost at one point or another of the production pipeline

- **Solution:** make every actor of the production pipeline realize that their metadata could be very useful down the road