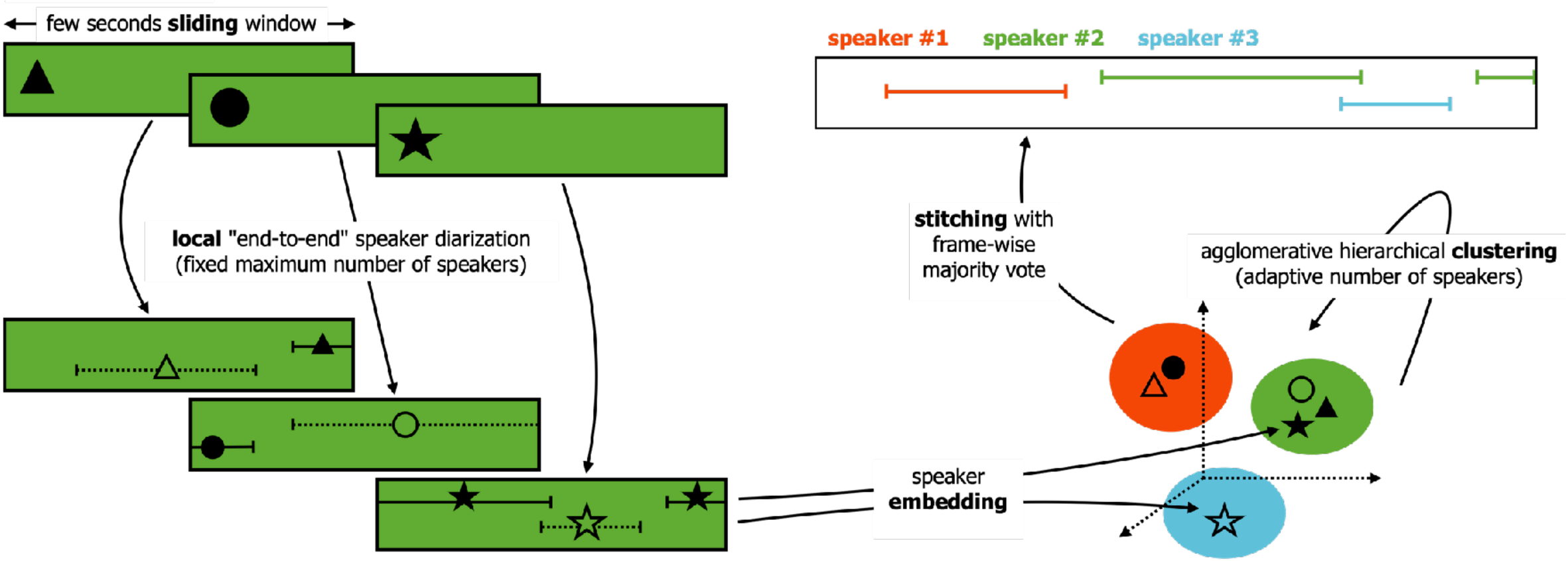
at VoxSRC 2023

Séverin Baroudi
Hervé Bredin
Alexis Plaquet
Thomas Pellegrini

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Main principle

the best of both worlds

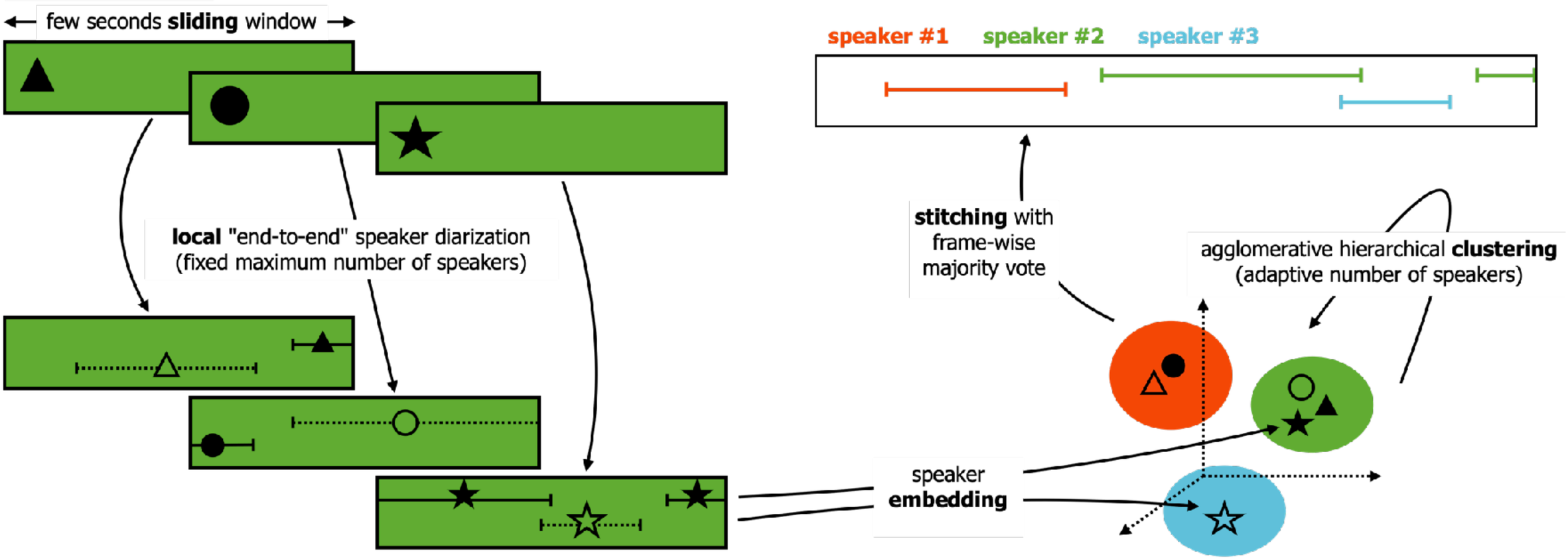


Hervé Bredin

INTERSPEECH 2023

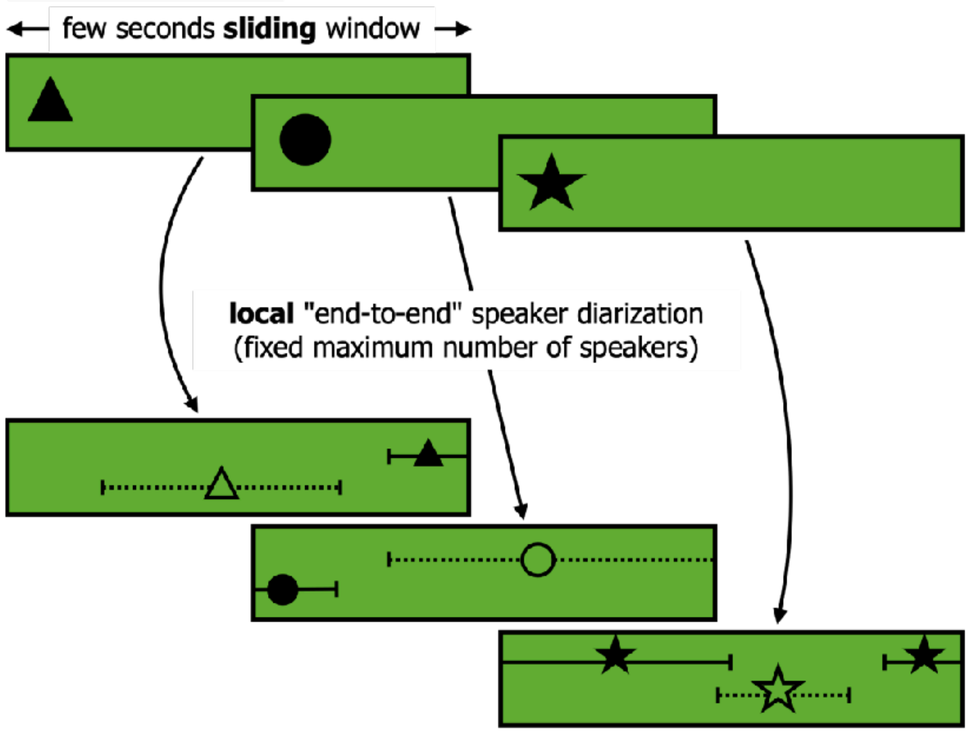
Main principle

the best of both worlds



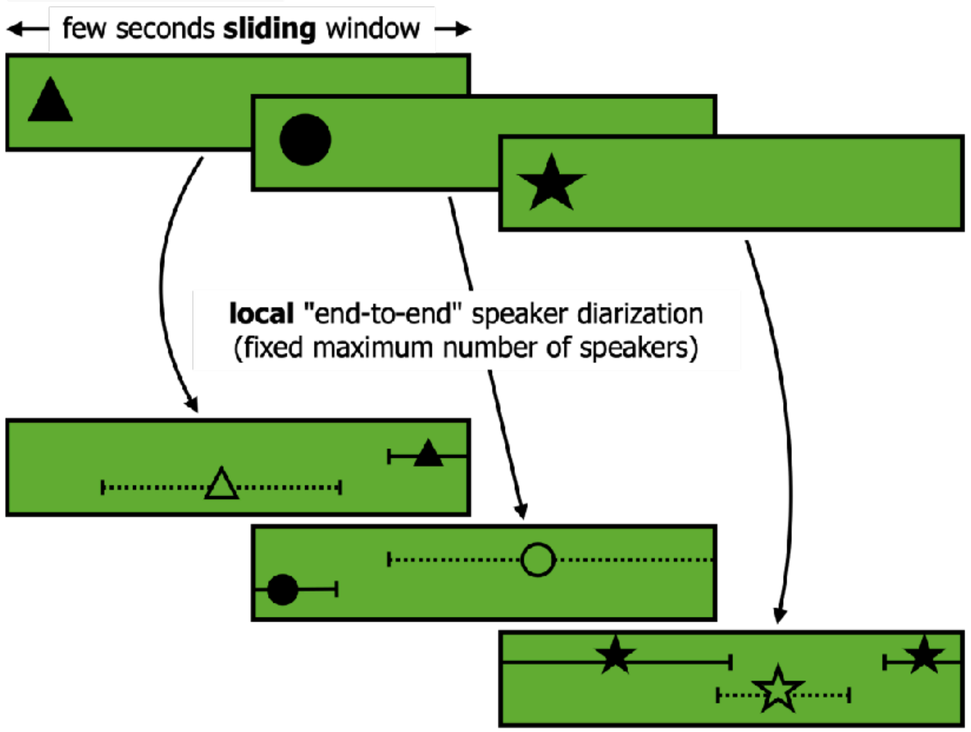
Local "end-to-end" speaker diarization

Improvements since last year



Local "end-to-end" speaker diarization

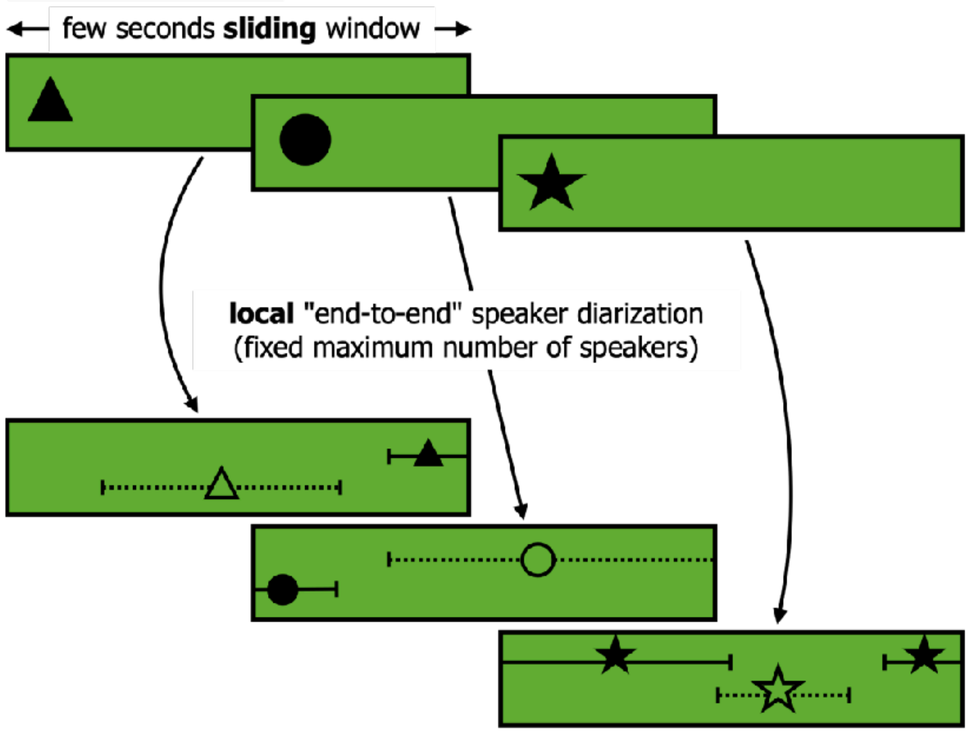
Improvements since last year



- Larger training set
  AMI, DIHARD, VoxConverse
  + AISHELL4+ AliMeeting
  + AVA-AVD + Ego4D
  + MSDWild + REPERE

Local "end-to-end" speaker diarization

Improvements since last year

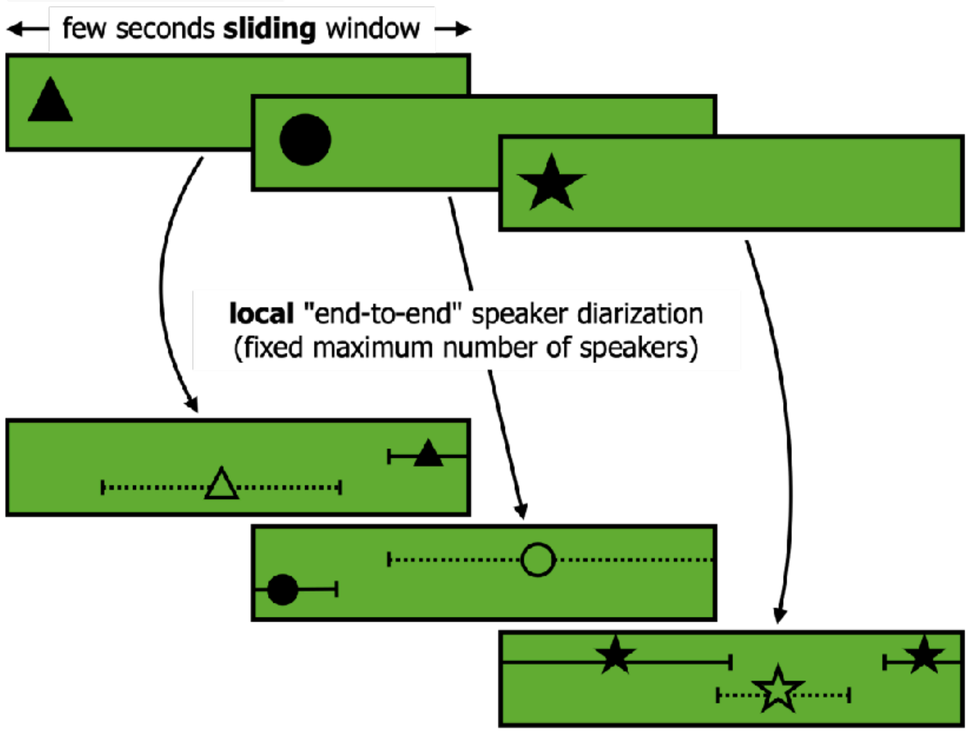


- Larger training set
  AMI, DIHARD, VoxConverse
  + AISHELL4+ AliMeeting
  + AVA-AVD + Ego4D
  + MSDWild + REPERE

- Longer windows
  from 5s to 10s

Local "end-to-end" speaker diarization

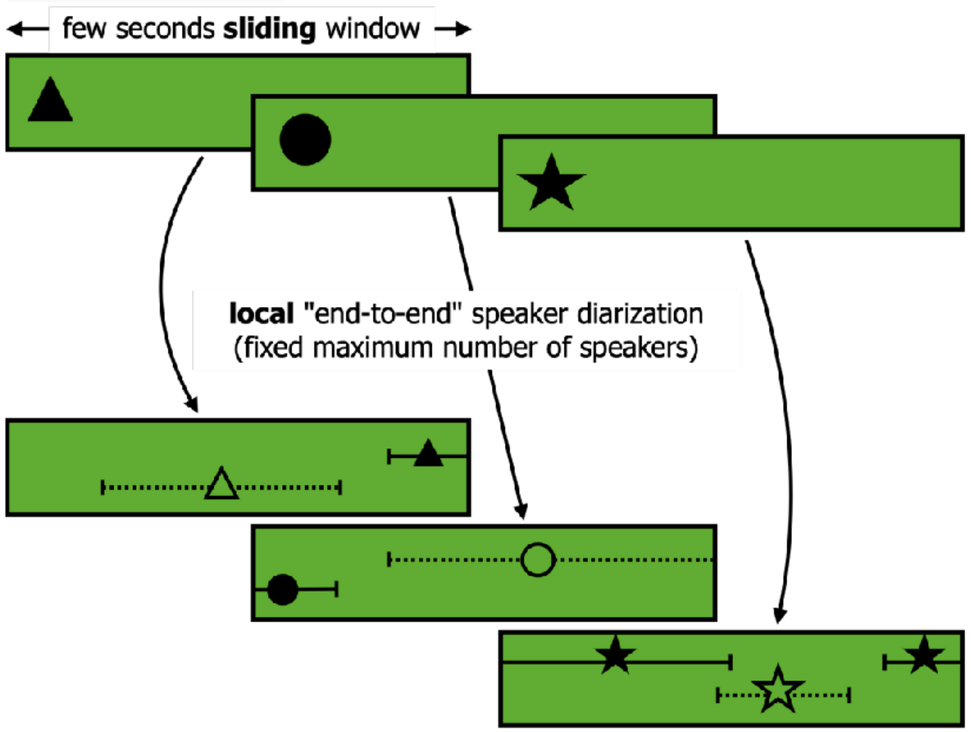
Improvements since last year



- Larger training set
  AMI, DIHARD, VoxConverse
  + AISHELL4+ AliMeeting
  + AVA-AVD + Ego4D
  + MSDWild + REPERE
- Longer windows
  from 5s to 10s
- ***Powerset* encoding**

Local "end-to-end" speaker diarization

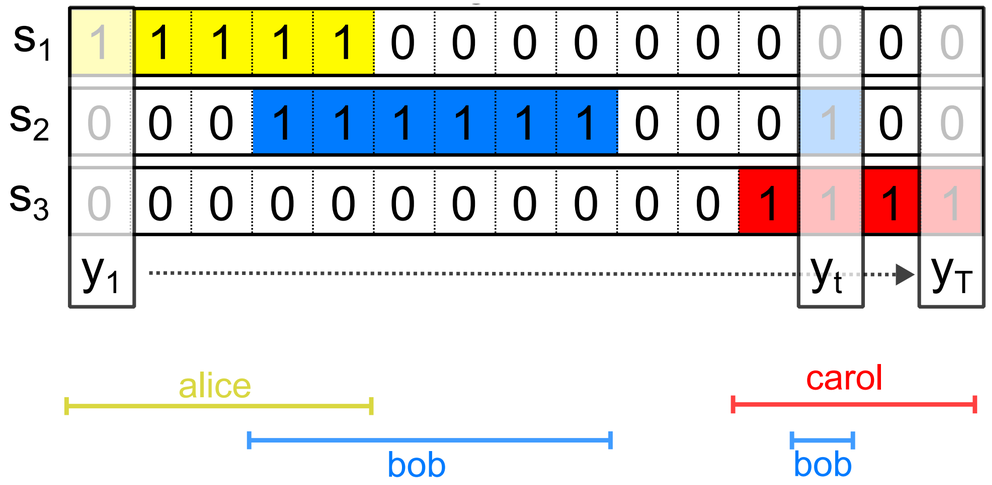
Improvements since last year

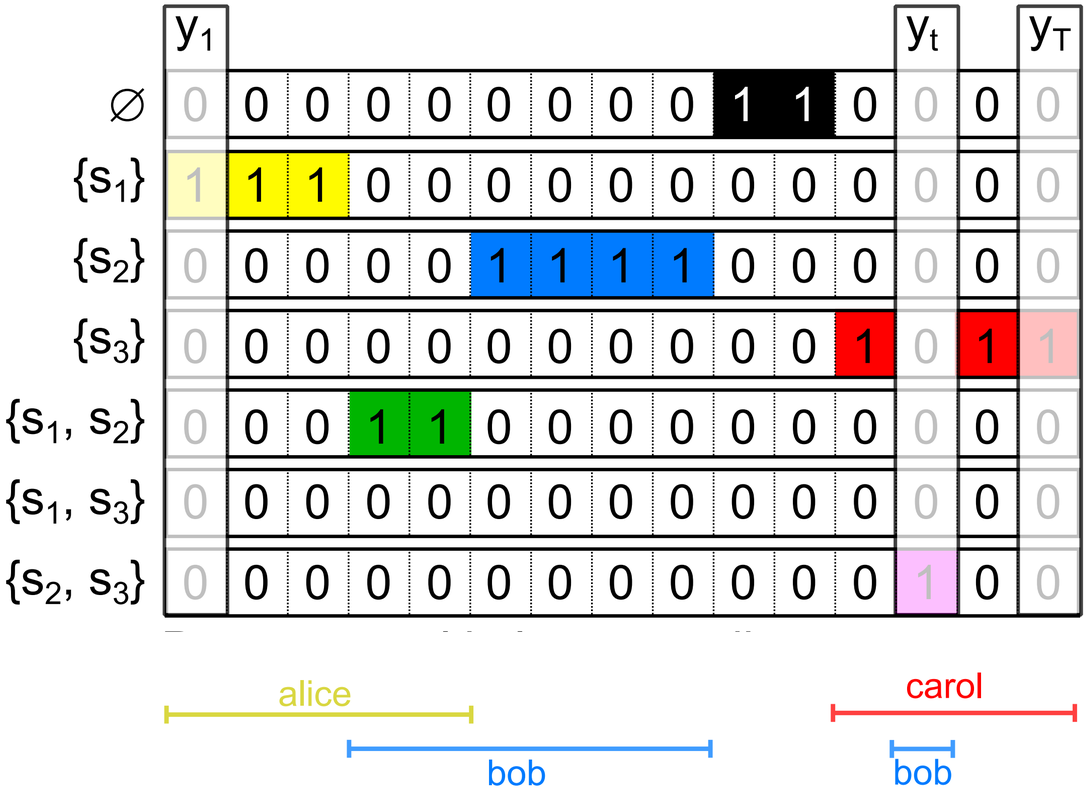


- Larger training set
  AMI, DIHARD, VoxConverse
  + AISHELL4+ AliMeeting
  + AVA-AVD + Ego4D
  + MSDWild + REPERE
- Longer windows
  from 5s to 10s
- ***Powerset*** **encoding**
- ***Self-supervised*** **feature extraction**

Powerset encoding

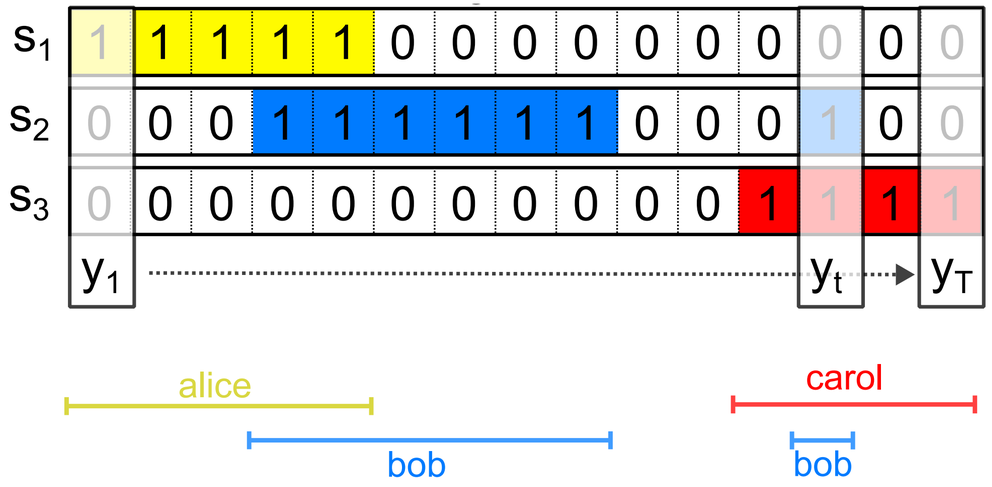
from multi-label (+ threshold) to multi-class (+ argmax)

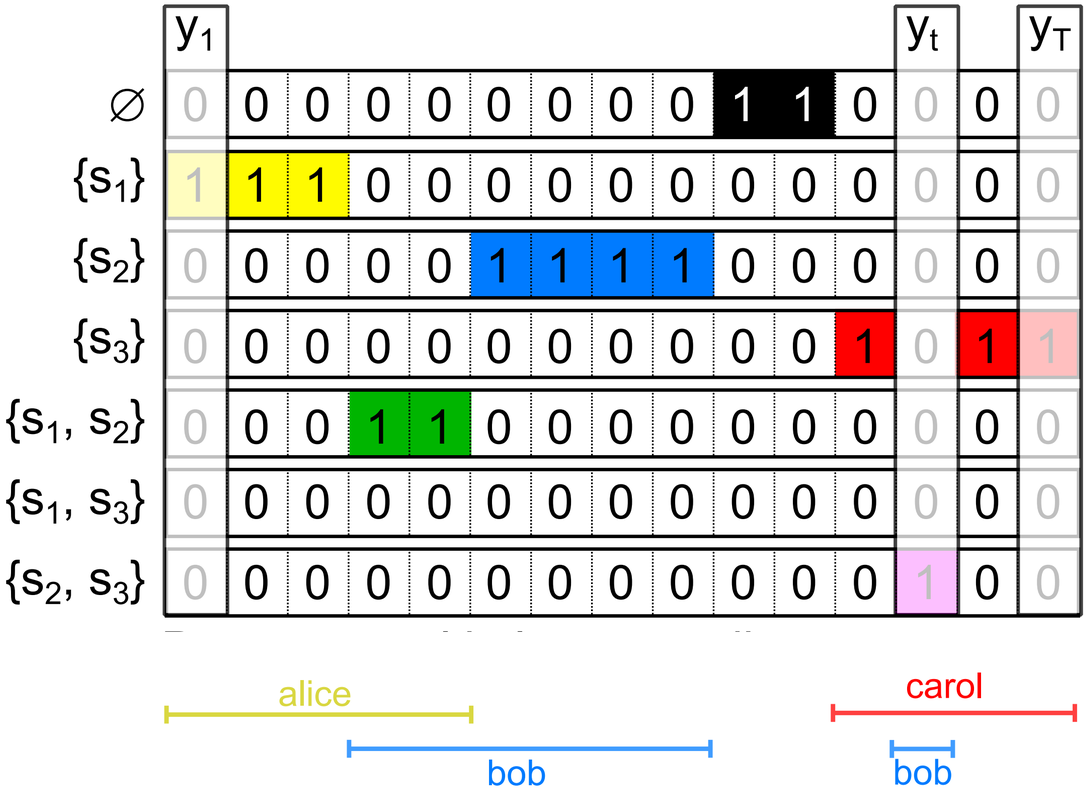




Powerset encoding

from multi-label (+ threshold) to multi-class (+ argmax)





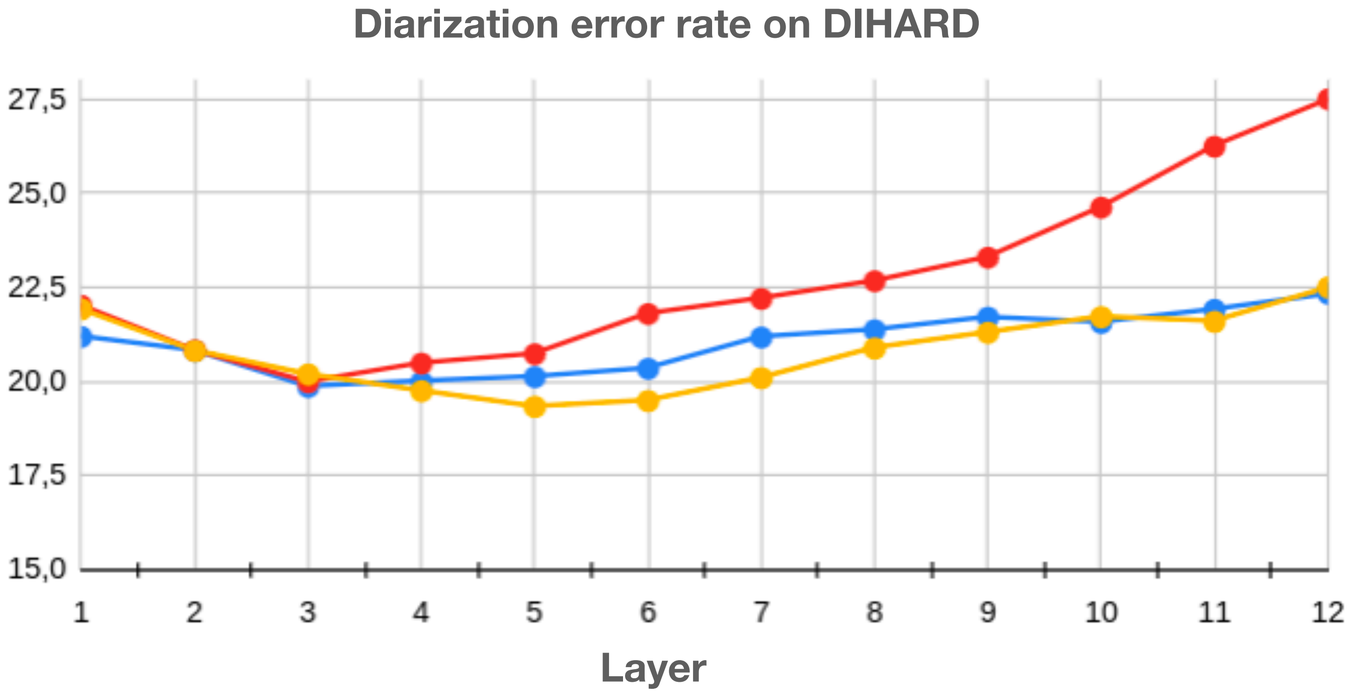
Powerset multi-class cross entropy loss for neural speaker diarization

Alexis Plaquet & Hervé Bredin

INTERSPEECH 2023

Self-supervised feature extraction

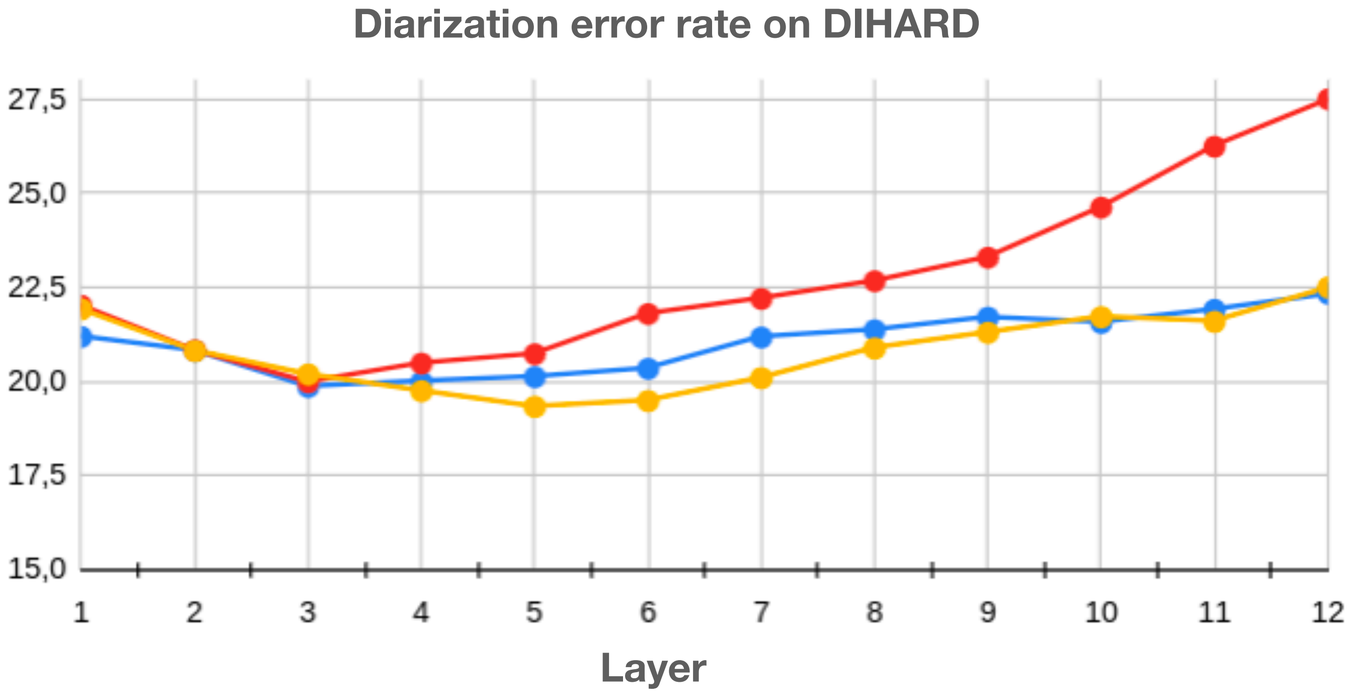
wav2vec 2.0 vs *HuBERT* vs *WavLM*



WavLM pretrained
on Librispeech

Self-supervised feature extraction

wav2vec 2.0 vs *HuBERT* vs *WavLM*

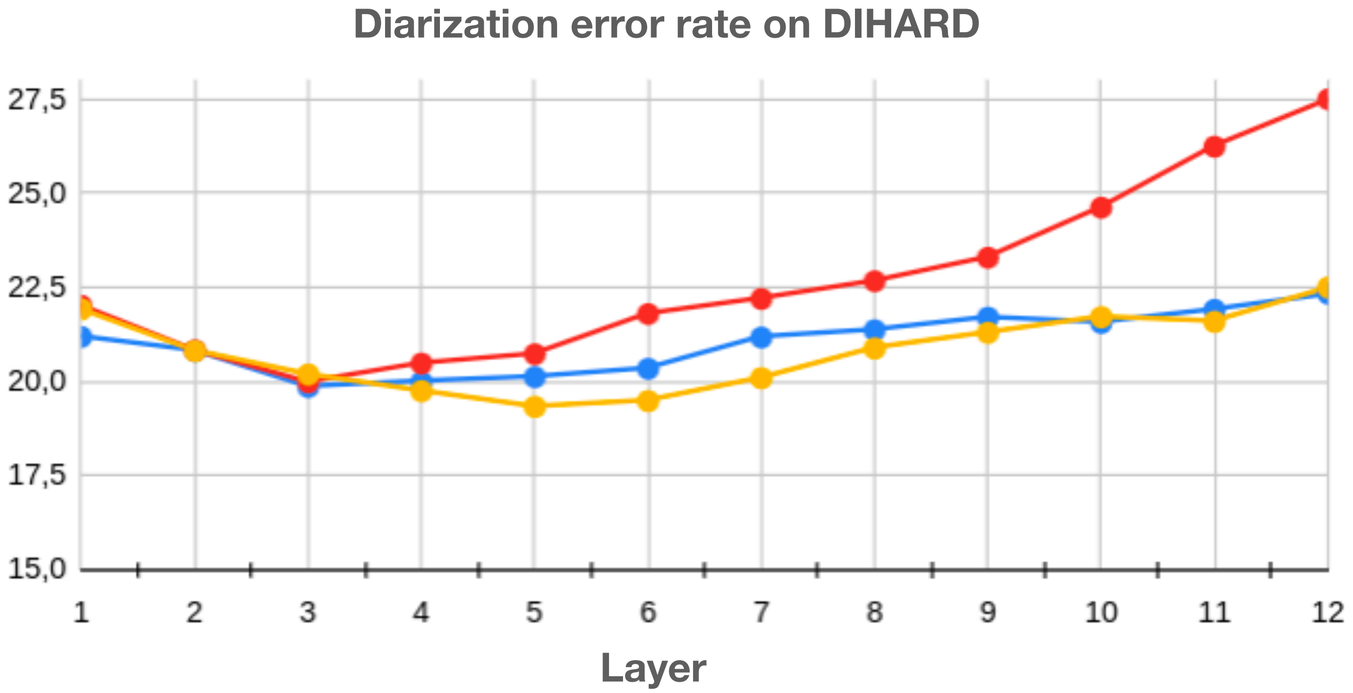


WavLM pretrained on Librispeech

WavLM pretrained on diarization datasets

Self-supervised feature extraction

wav2vec 2.0* vs *HuBERT* vs *WavLM



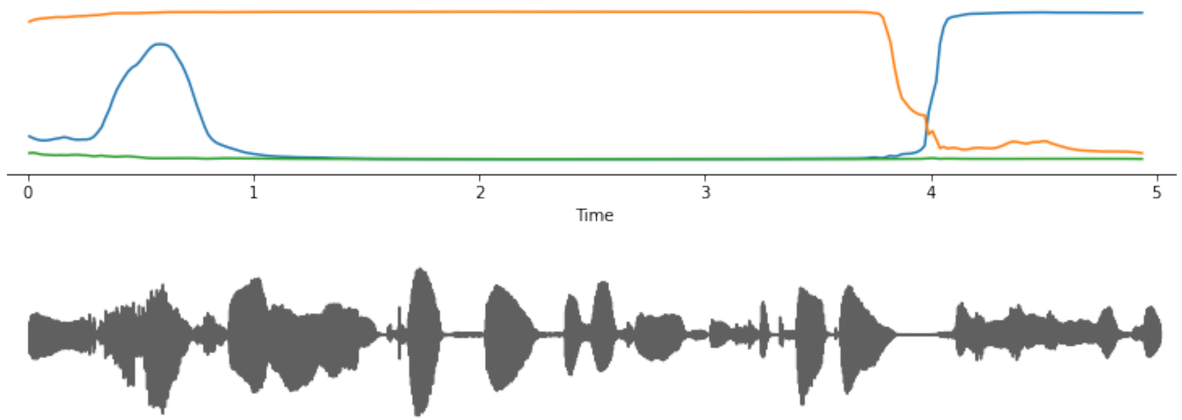
WavLM pretrained on Librispeech

21% relative decrease in diarization error rate

WavLM pretrained on diarization datasets

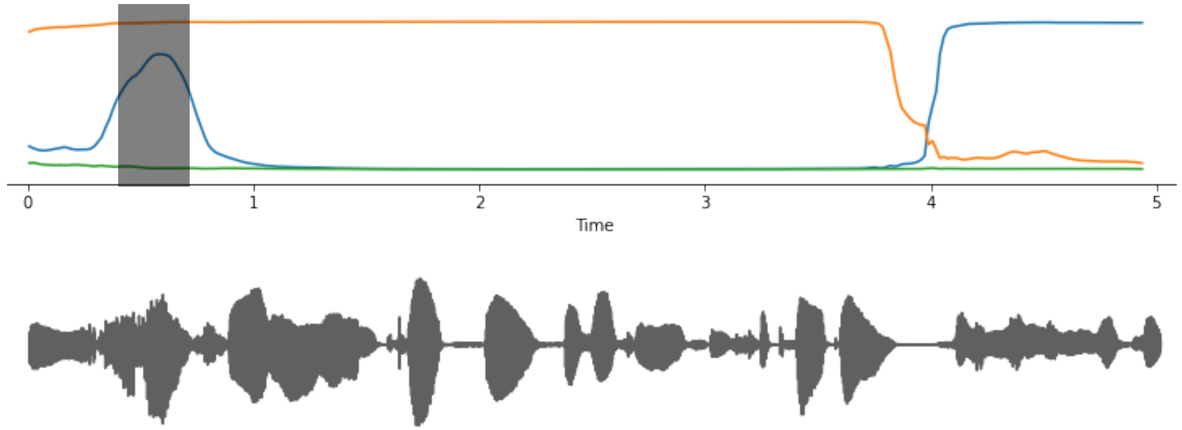
Speaker embedding

from SpeechBrain ECAPA-TDNN to WeSpeaker ResNet34



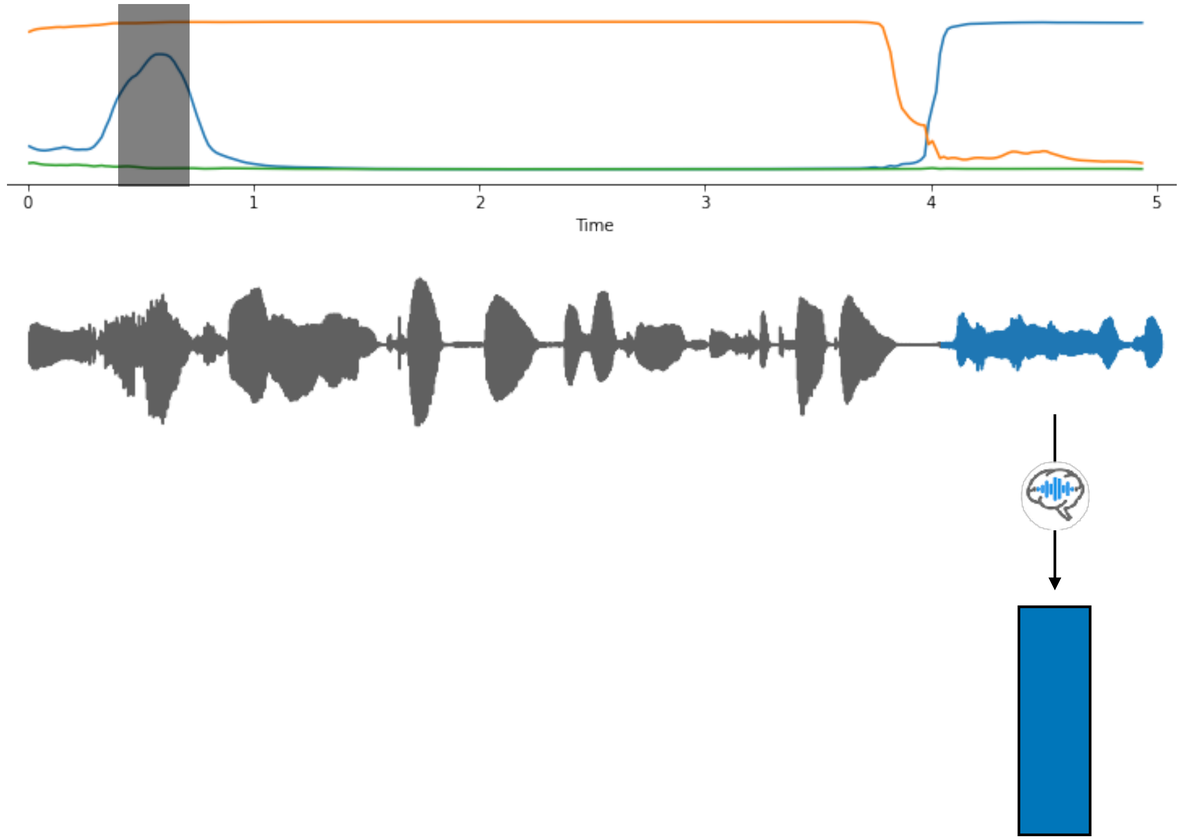
Speaker embedding

from SpeechBrain ECAPA-TDNN to WeSpeaker ResNet34



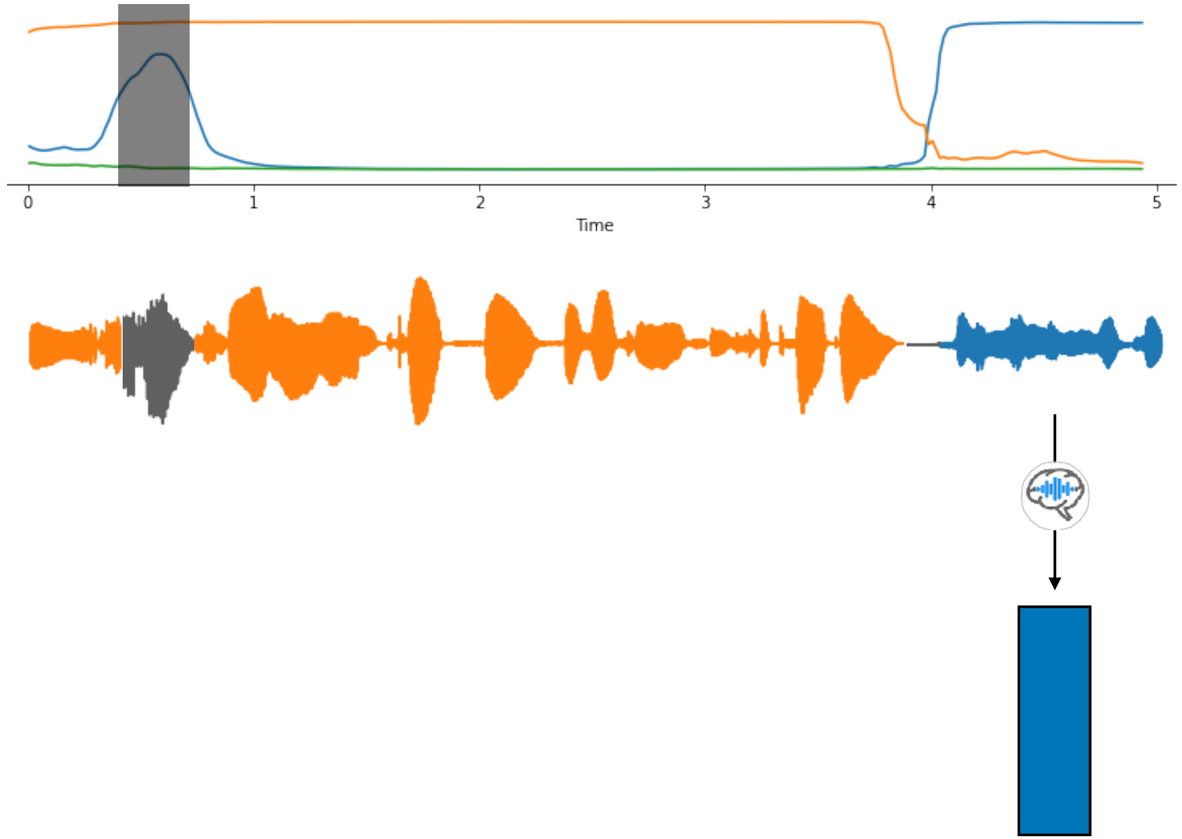
Speaker embedding

from SpeechBrain ECAPA-TDNN to WeSpeaker ResNet34



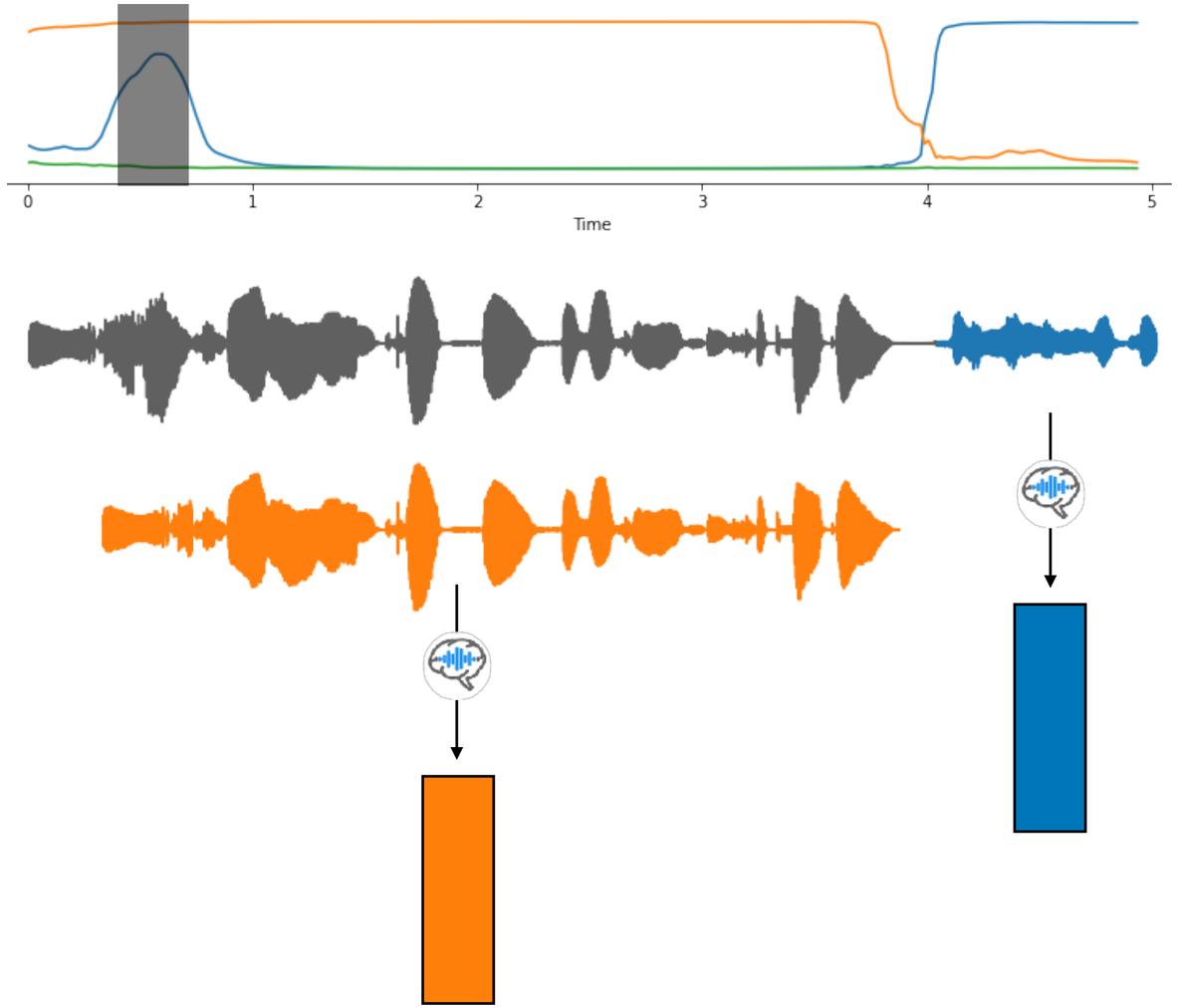
Speaker embedding

from SpeechBrain ECAPA-TDNN to WeSpeaker ResNet34



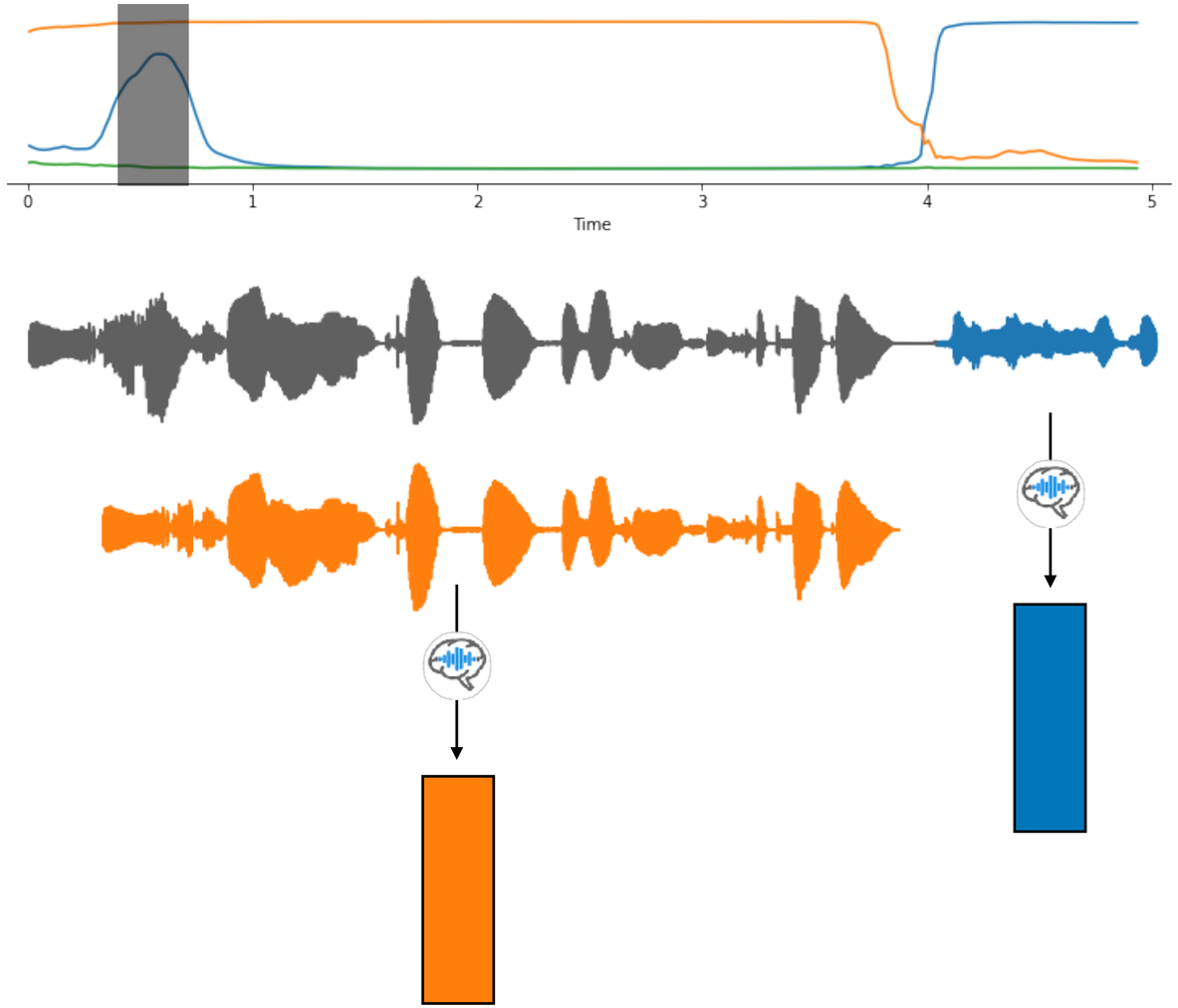
Speaker embedding

from SpeechBrain ECAPA-TDNN to WeSpeaker ResNet34



Speaker embedding

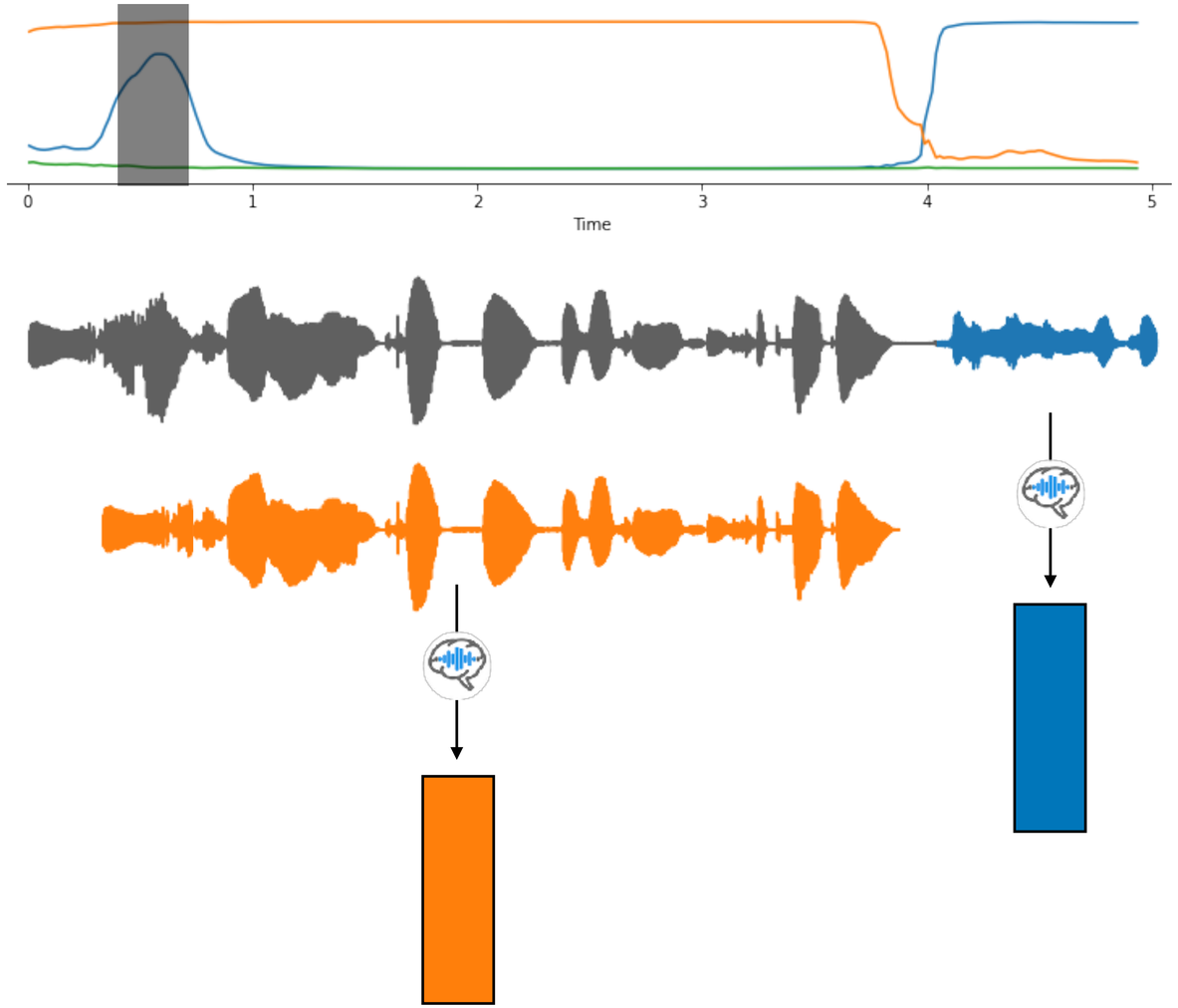
from SpeechBrain ECAPA-TDNN to WeSpeaker ResNet34



SpeechBrain
ECAPA-TDNN

Speaker embedding

from SpeechBrain ECAPA-TDNN to WeSpeaker ResNet34

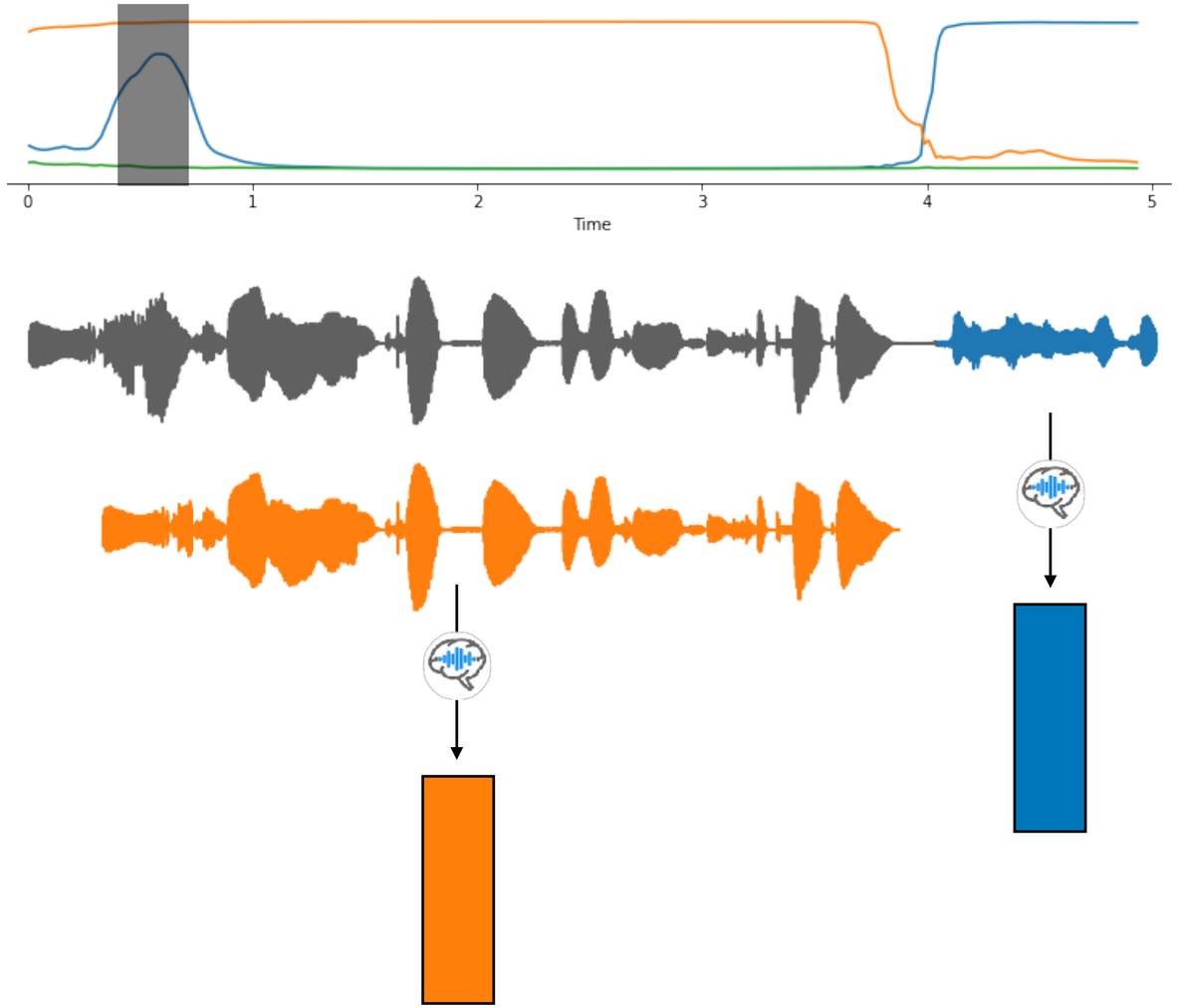


SpeechBrain
ECAPA-TDNN

WeSpeaker
ResNet34

Speaker embedding

from SpeechBrain ECAPA-TDNN to WeSpeaker ResNet34



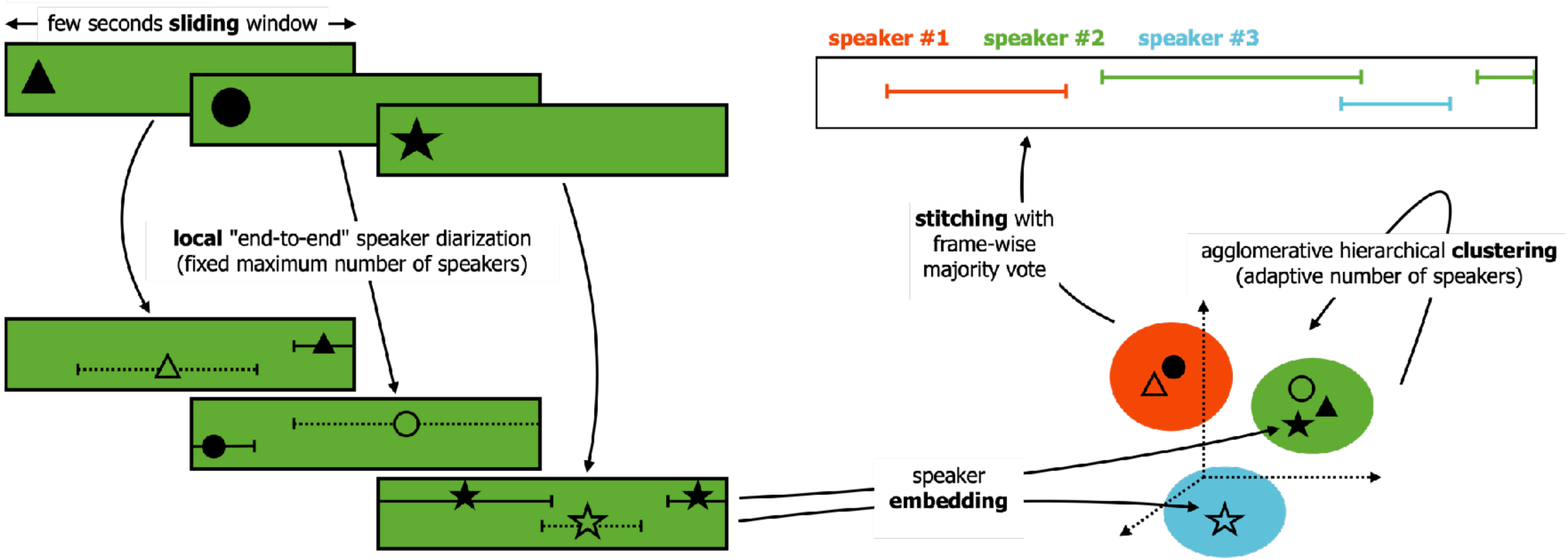
SpeechBrain
ECAPA-TDNN

14% relative decrease
in speaker confusion

WeSpeaker
ResNet34

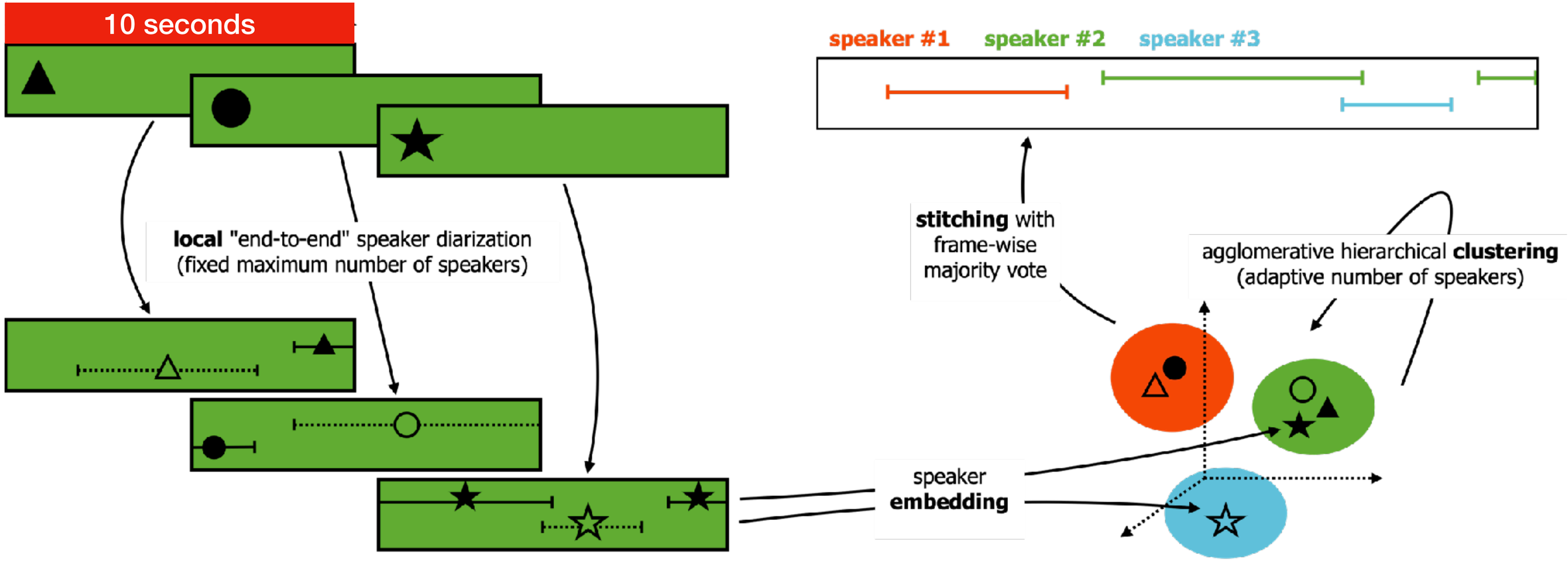
Final run

Single system (no ensembling)



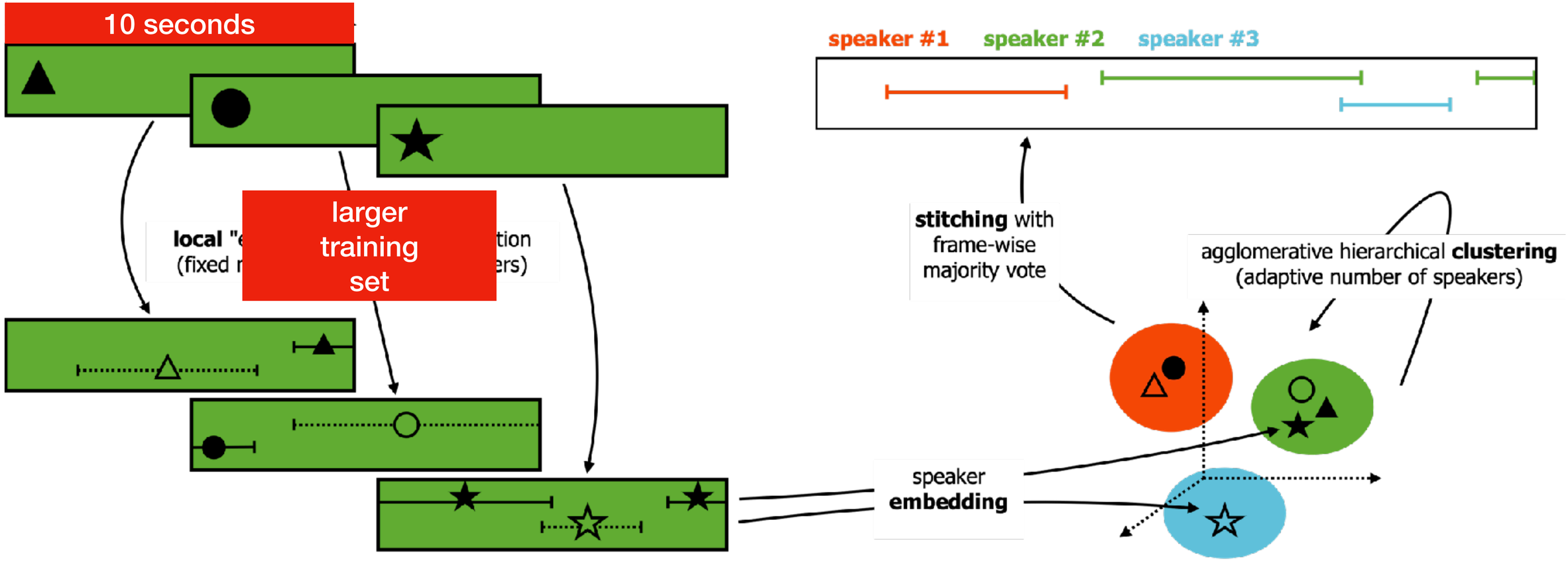
Final run

Single system (no ensembling)



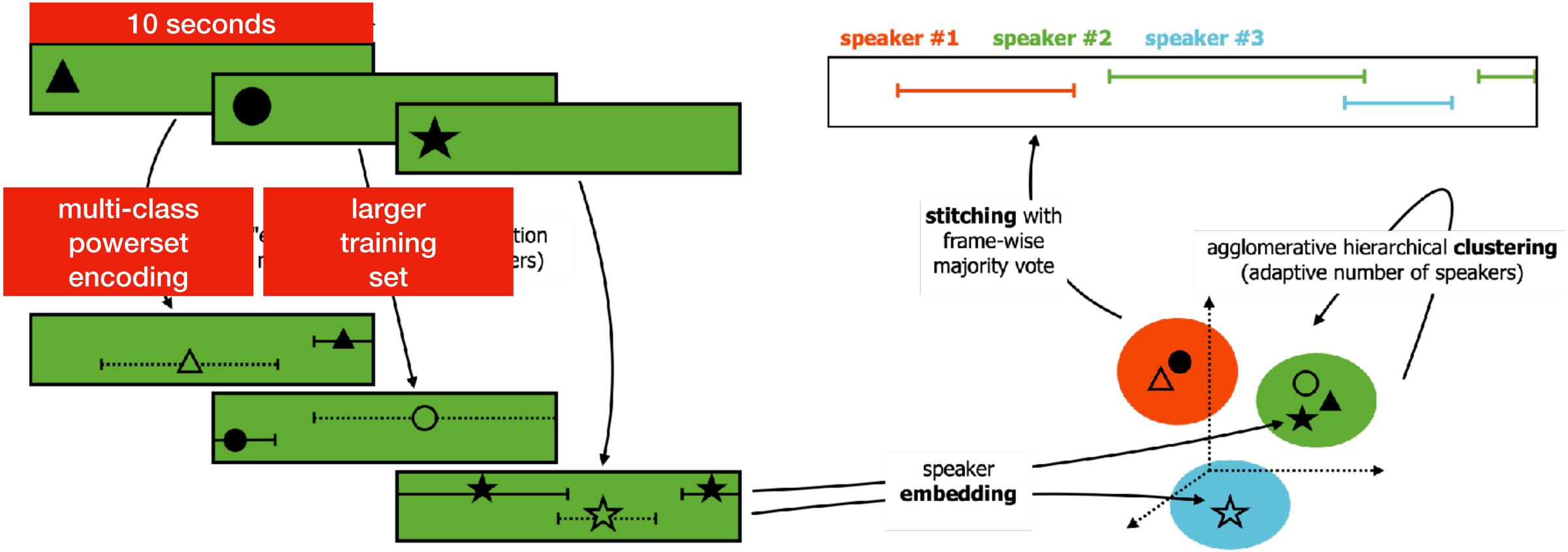
Final run

Single system (no ensembling)



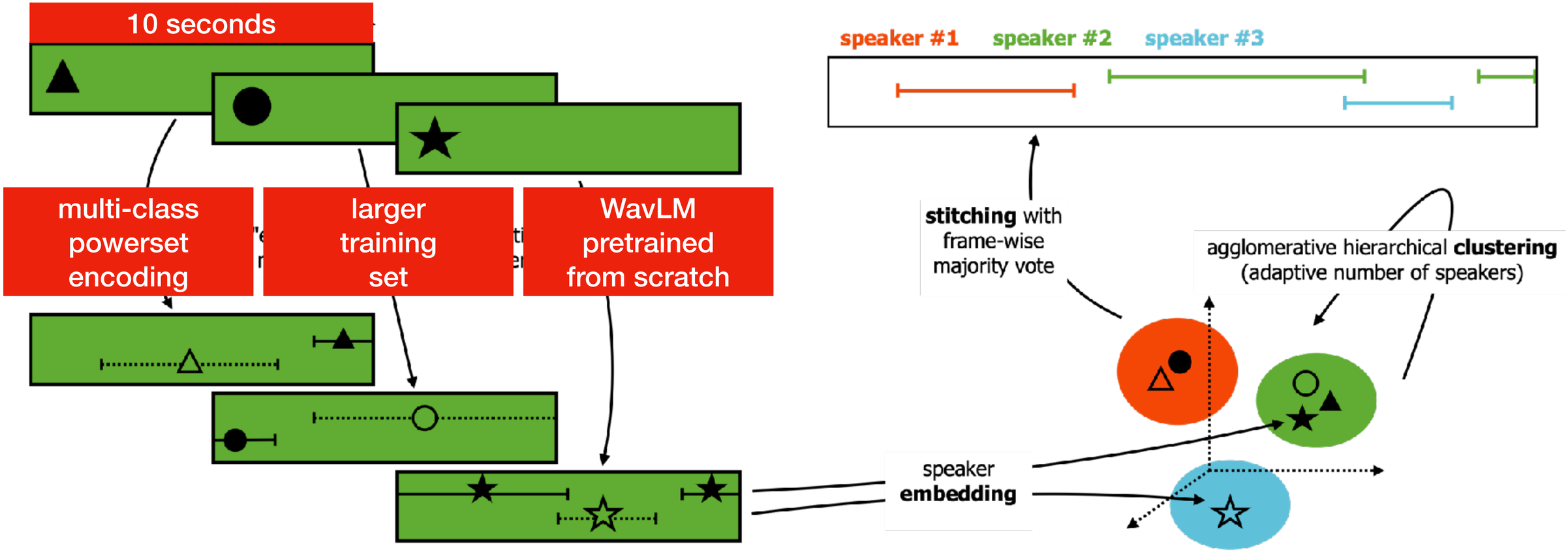
Final run

Single system (no ensembling)



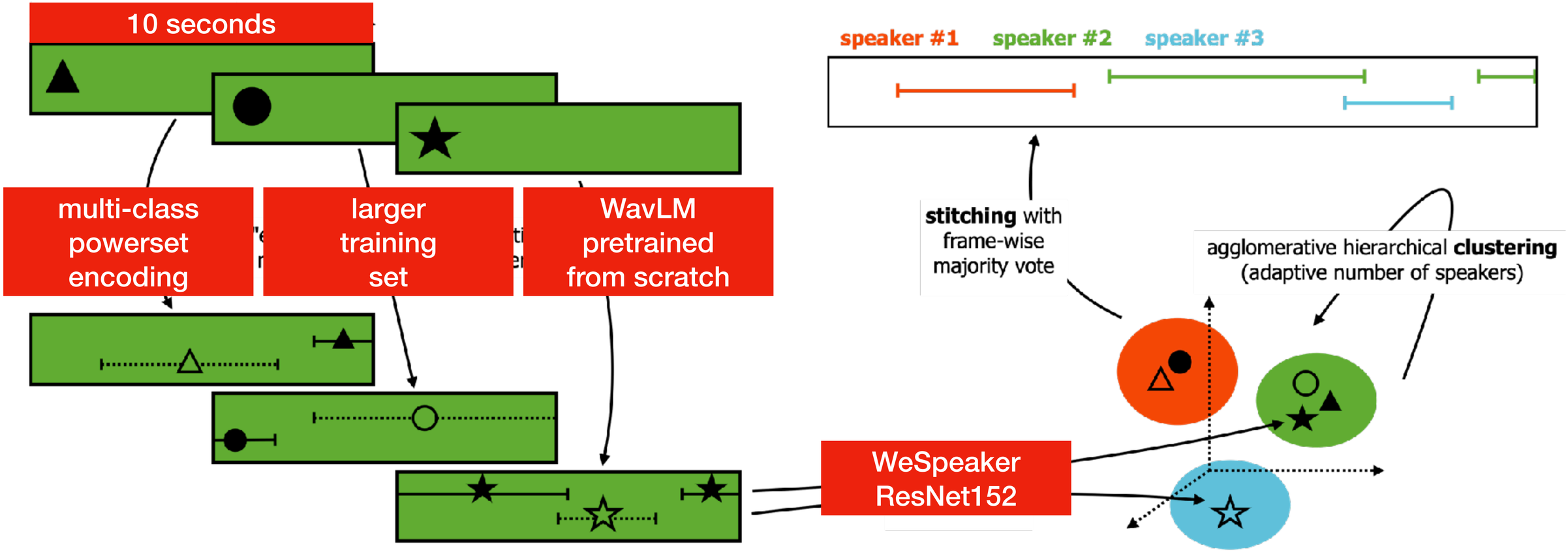
Final run

Single system (no ensembling)



Final run

Single system (no ensembling)



Acknowledgments

GENCI

"Jean Zay" supercomputer

Get your stickers!

Hiring!

pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe
Hervé Bredin

Powerset multi-class cross entropy loss for neural speaker diarization
Alexis Plaquet & Hervé Bredin