OVERLAP-AWARE DIARIZATION:
RESEGMENTATION USING NEURAL END-TO-END OVERLAPPED SPEECH DETECTION

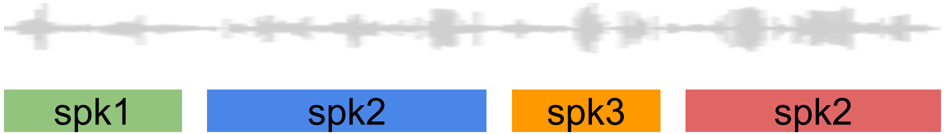
ICASSP2020 Barcelona

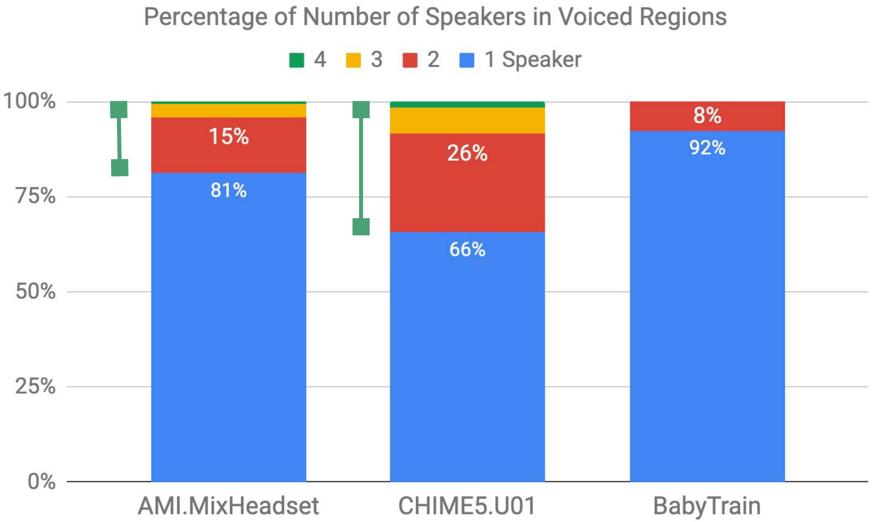
Latané Bullock
Undergraduate
Rice University

Hervé Bredin
Researcher
LIMSI, CNRS France

Leibny Paola Garcia-Perera
Assistant Research Scientist
CLSP, Johns Hopkins University

Diarization - “Who spoke when?”

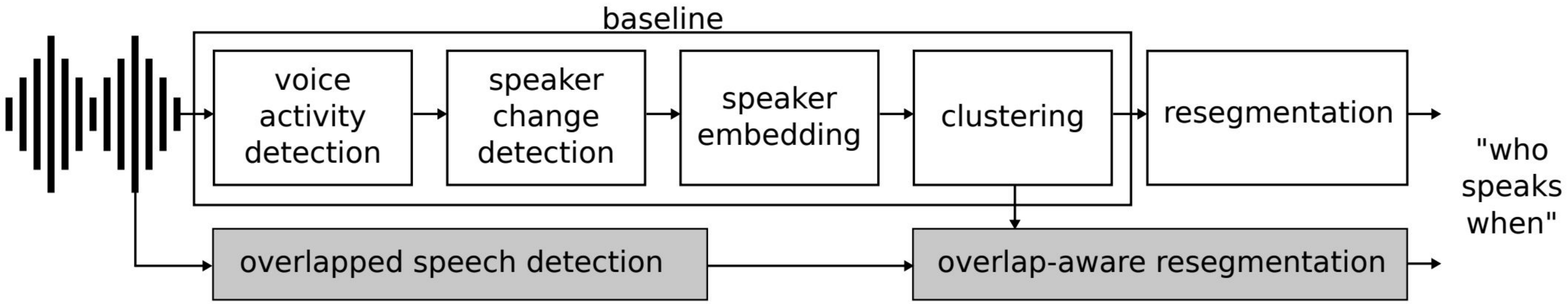




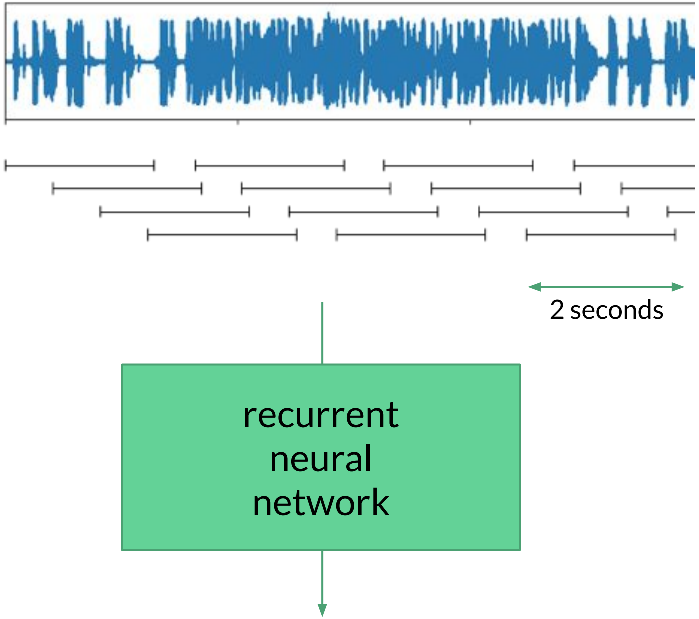
In adverse audio recordings

- Large proportion of overlapped speech
  across all datasets
- Leads to high
  missed detection rate
- May lead to high
  speaker confusion rate

Contributions: Overlap Detection and Overlap-Aware Resegmentation



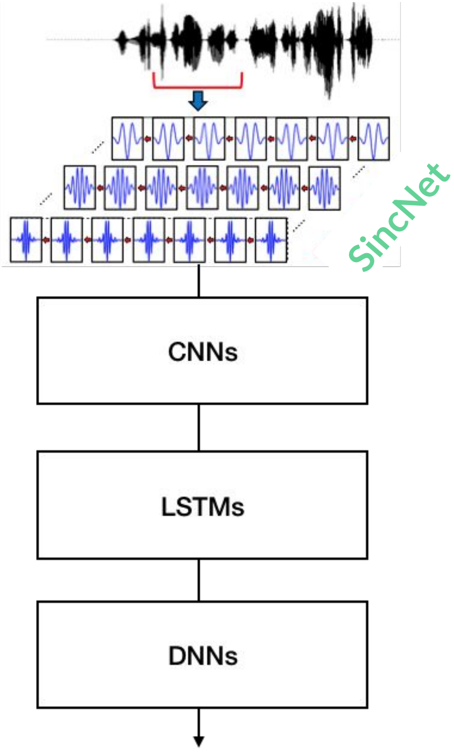
Overlap Detection with pyannote



2 seconds

recurrent neural network

1 = overlap
0 = no overlap



SincNet

CNNs

LSTMs

DNNs

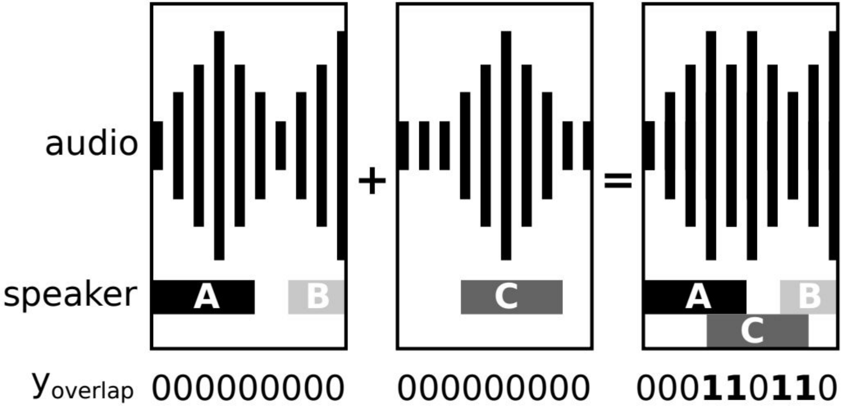
Framewise scores

Data augmentation

- Training a network directly does not work
  - Class imbalance
  - Lack of variability
- Two types of training samples
  - **50% regular**

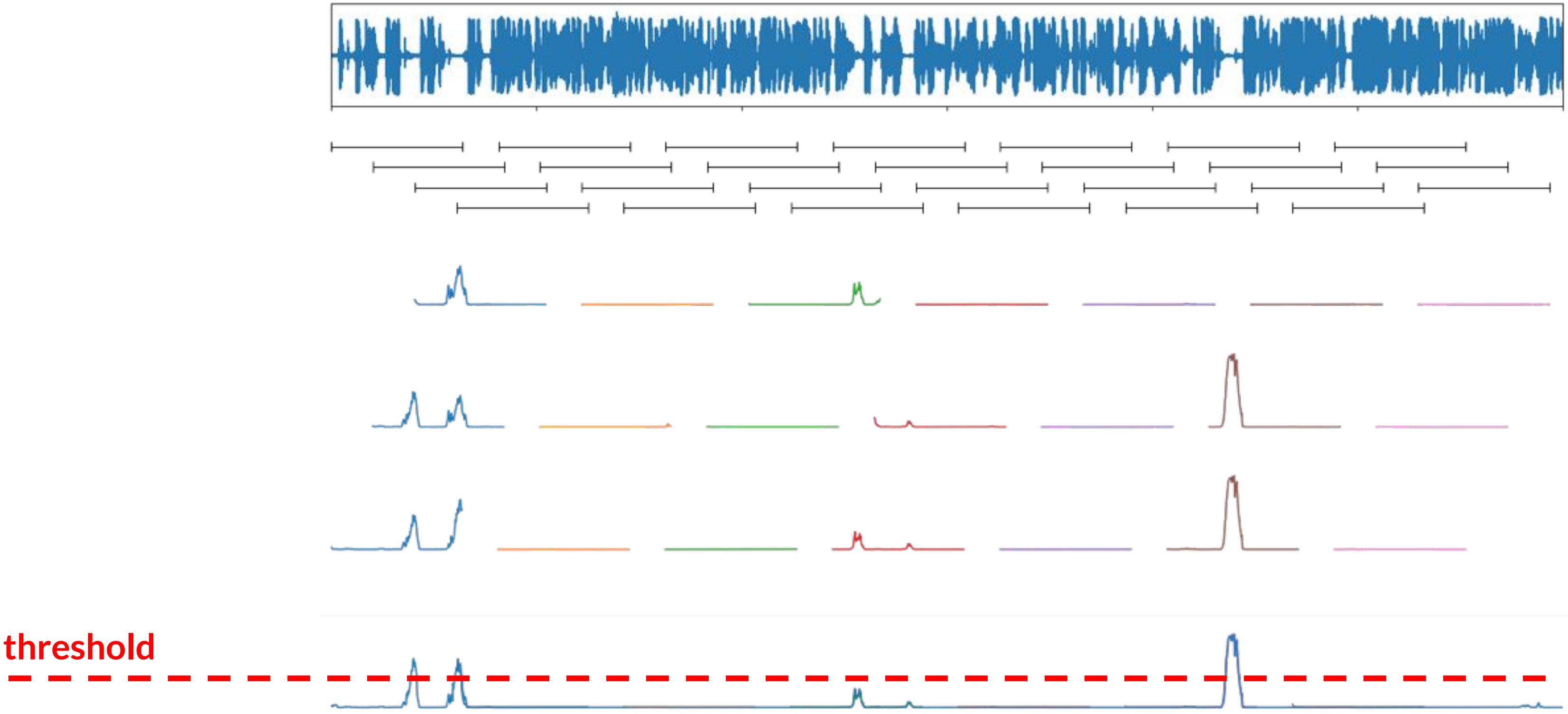
    2-second chunks extracted from the training set randomly
  - **50% made-up**

    weighted sum of two chunks

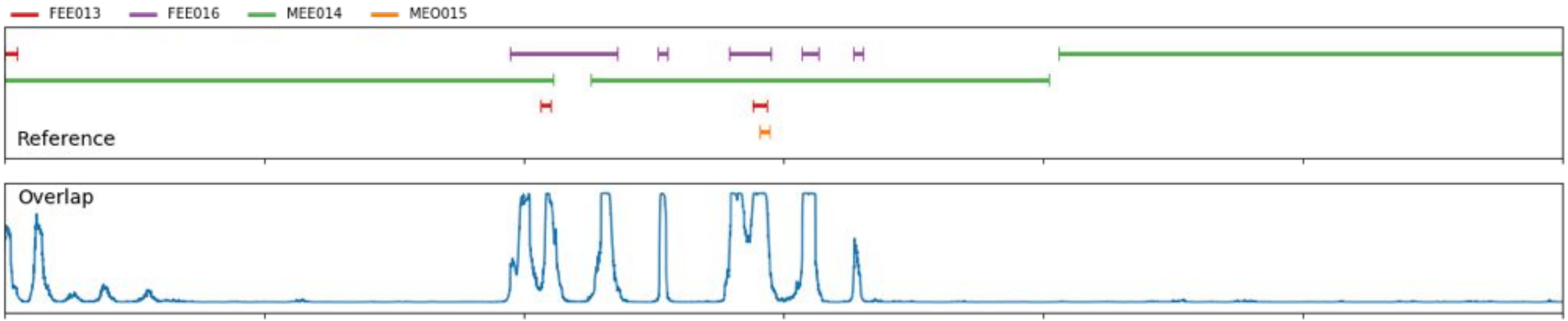


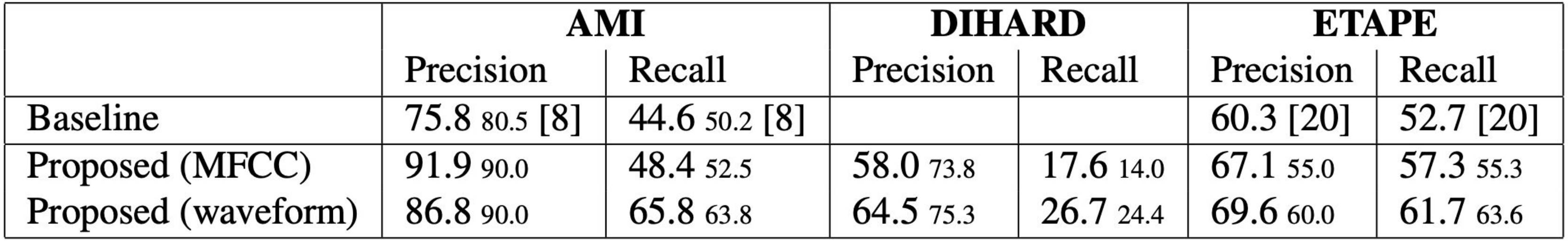
To increase the number of positive training samples for overlapped speech detection, artificial audio chunks are created by summing two random audio chunks

At test time...



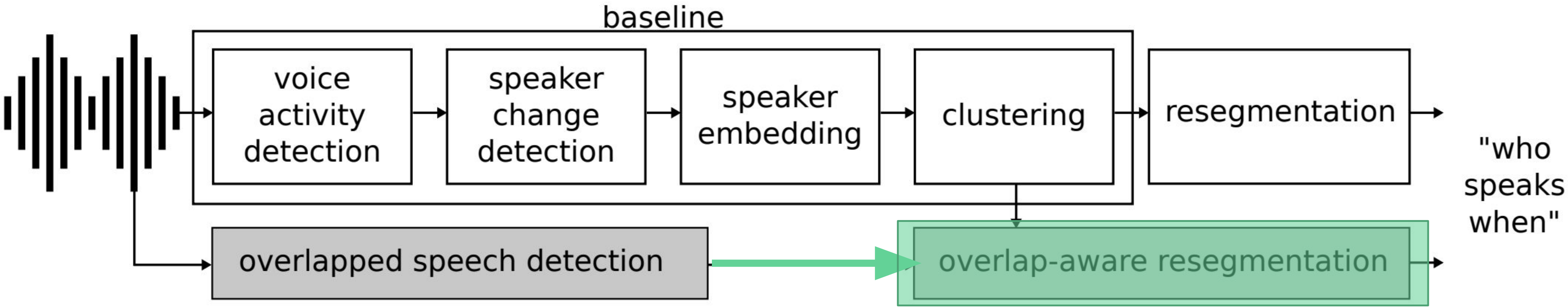
Overlap Detection: Results





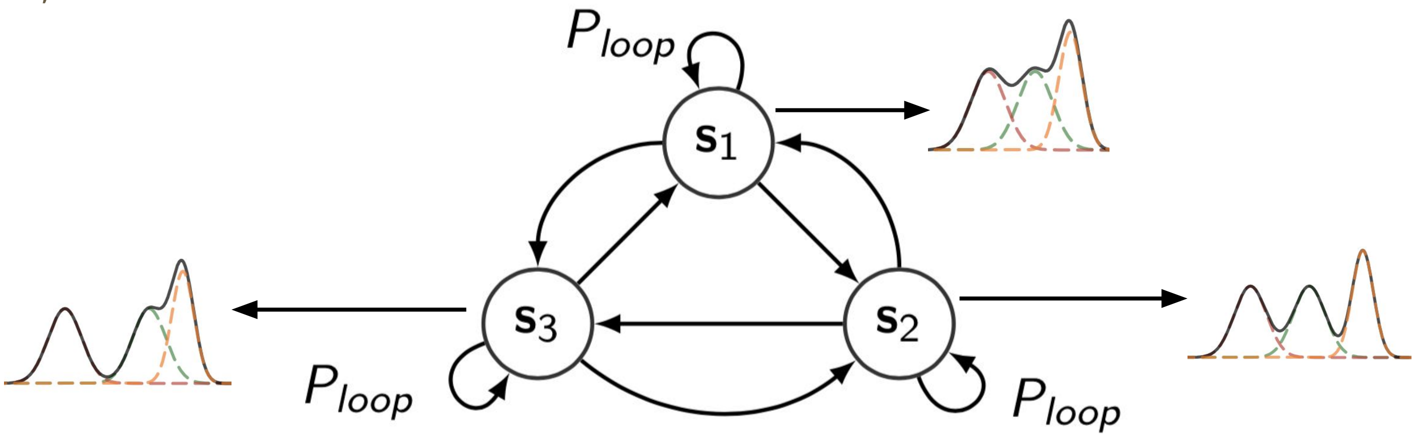
	AMI		DIHARD		ETAPE	
	Precision	Recall	Precision	Recall	Precision	Recall
Baseline	75.8 80.5 [8]	44.6 50.2 [8]			60.3 [20]	52.7 [20]
Proposed (MFCC)	91.9 90.0	48.4 52.5	58.0 73.8	17.6 14.0	67.1 55.0	57.3 55.3
Proposed (waveform)	86.8 90.0	65.8 63.8	64.5 75.3	26.7 24.4	69.6 60.0	61.7 63.6

Overlap-Aware Resegmentation



Variational Bayes HMM-GMM Resegmentation

Diez et al, 2018



Hidden Markov model where:

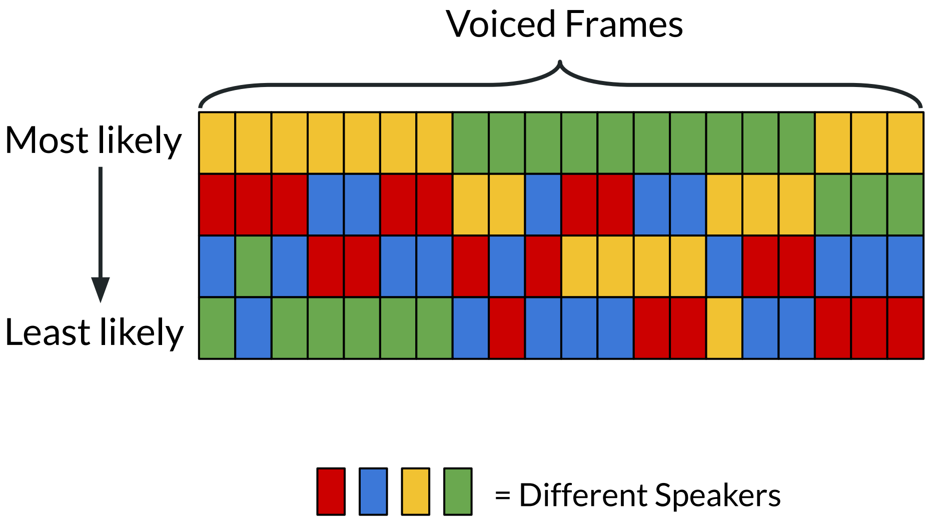
- state represents a speaker
- state distributions are GMMs constrained by eigenvoice priors

Single model (theoretically) infers:

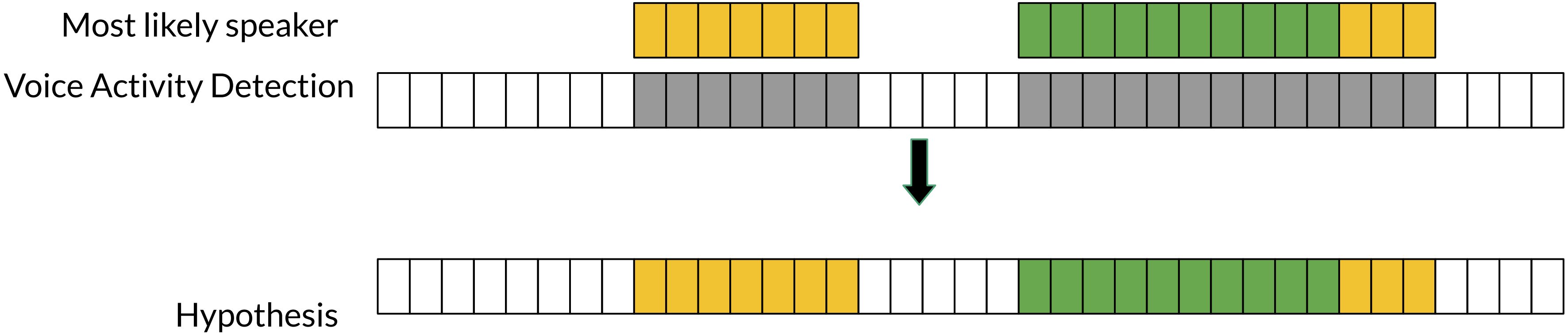
- Speaker distributions
- **Number of speakers**
- **Speaker sequence**

… with the Variational Bayes

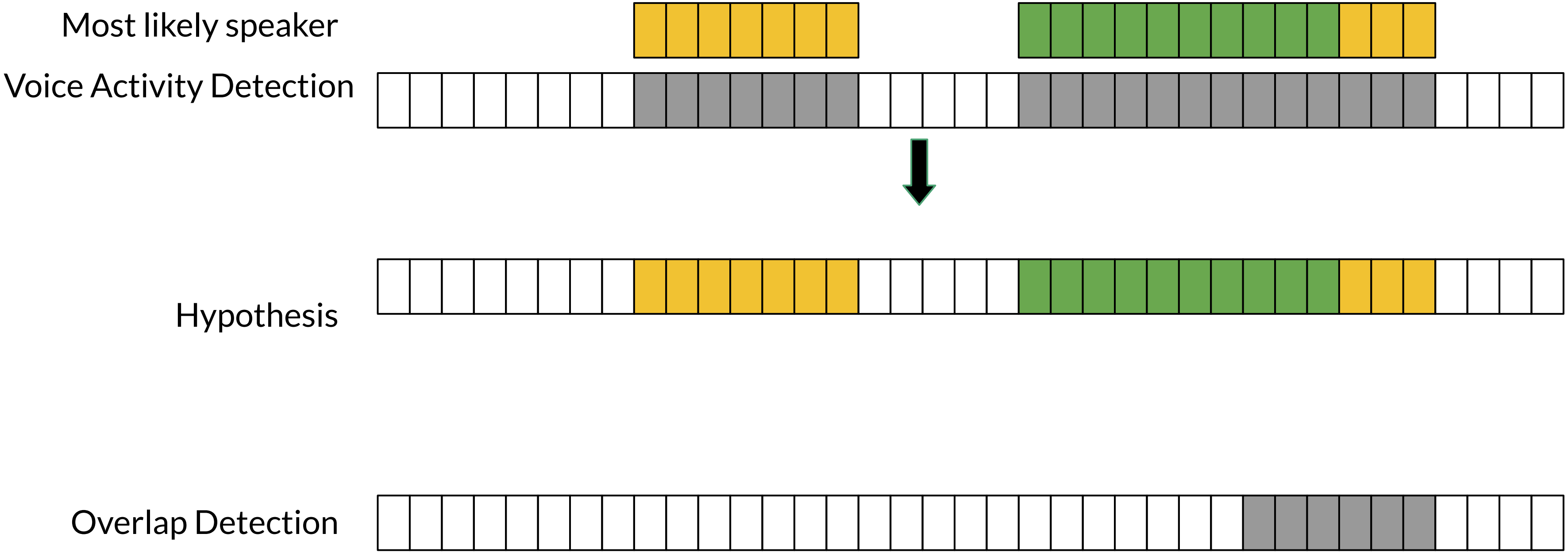
Remnant of VB-HMM resegmentation, the **speaker attribution matrix**:



Two-speaker Assignment in Overlap Regions



Two-speaker Assignment in Overlap Regions



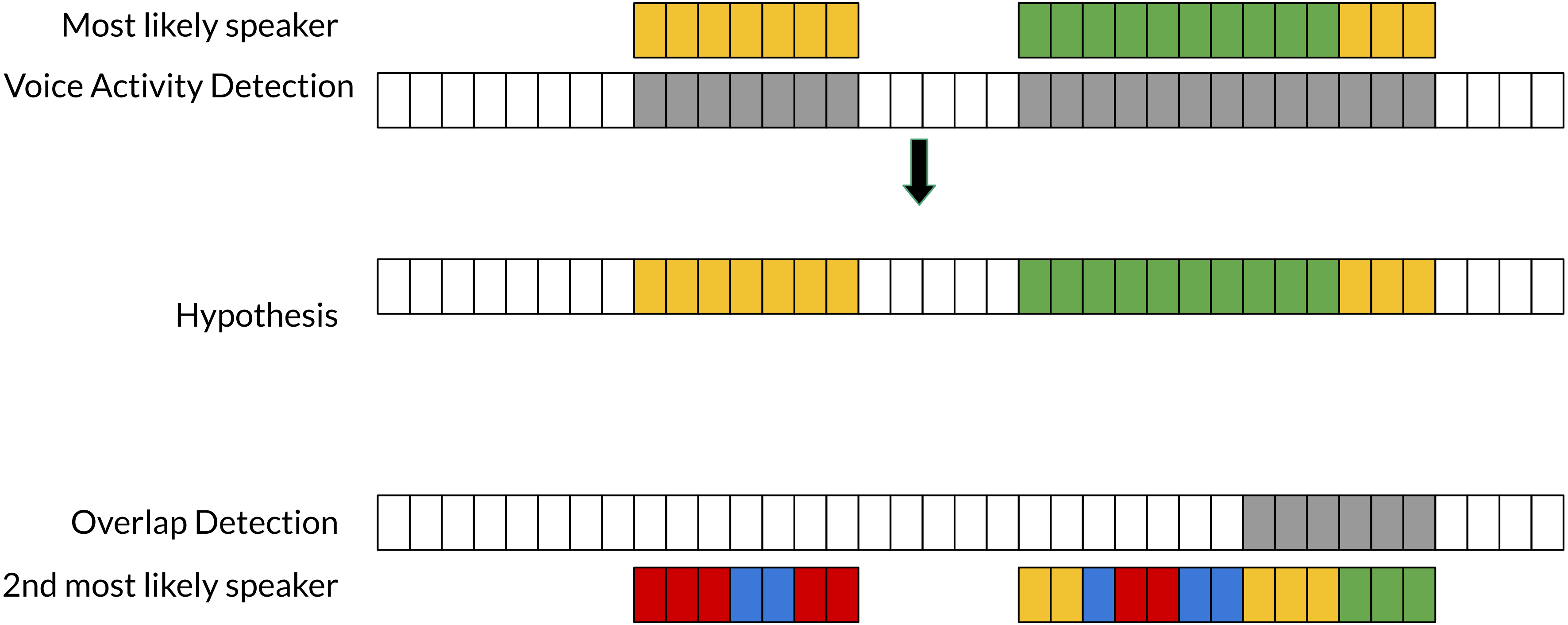
Most likely speaker

Voice Activity Detection

Hypothesis

Overlap Detection

Two-speaker Assignment in Overlap Regions



Most likely speaker

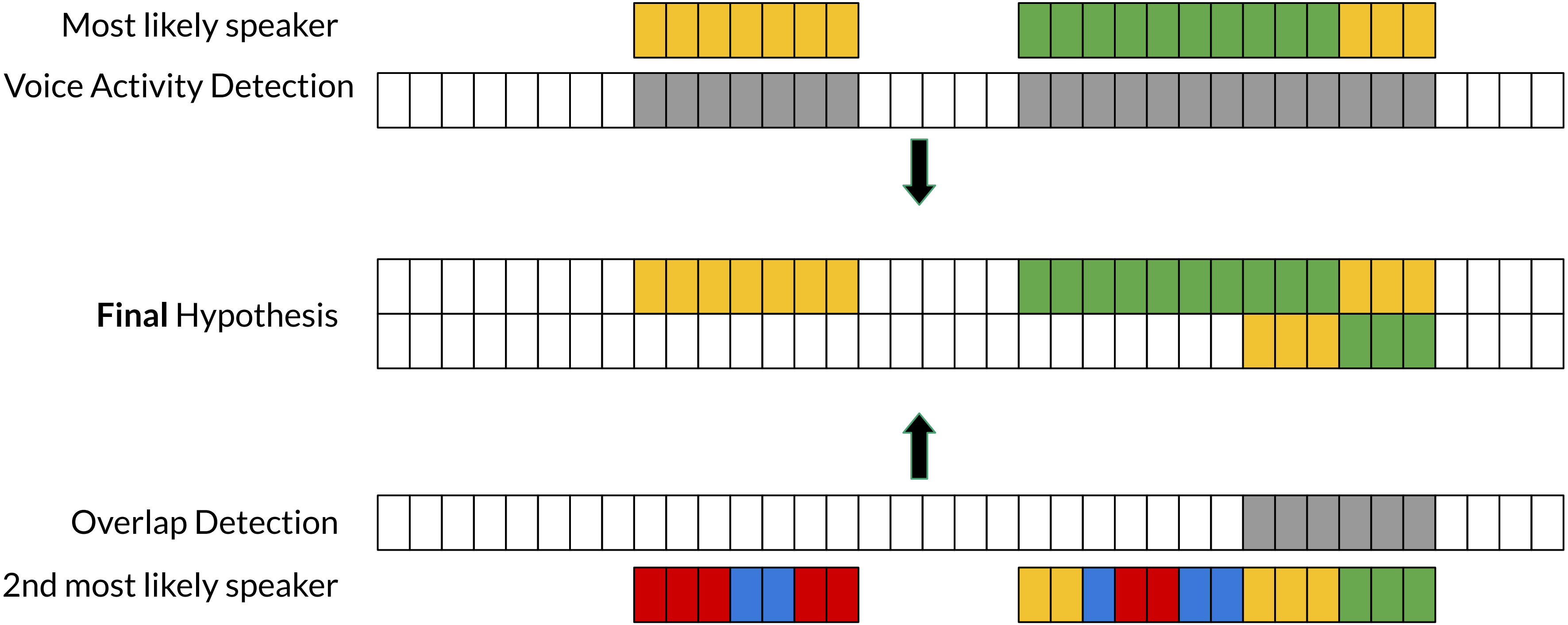
Voice Activity Detection

Hypothesis

Overlap Detection

2nd most likely speaker

Two-speaker Assignment in Overlap Regions



Most likely speaker

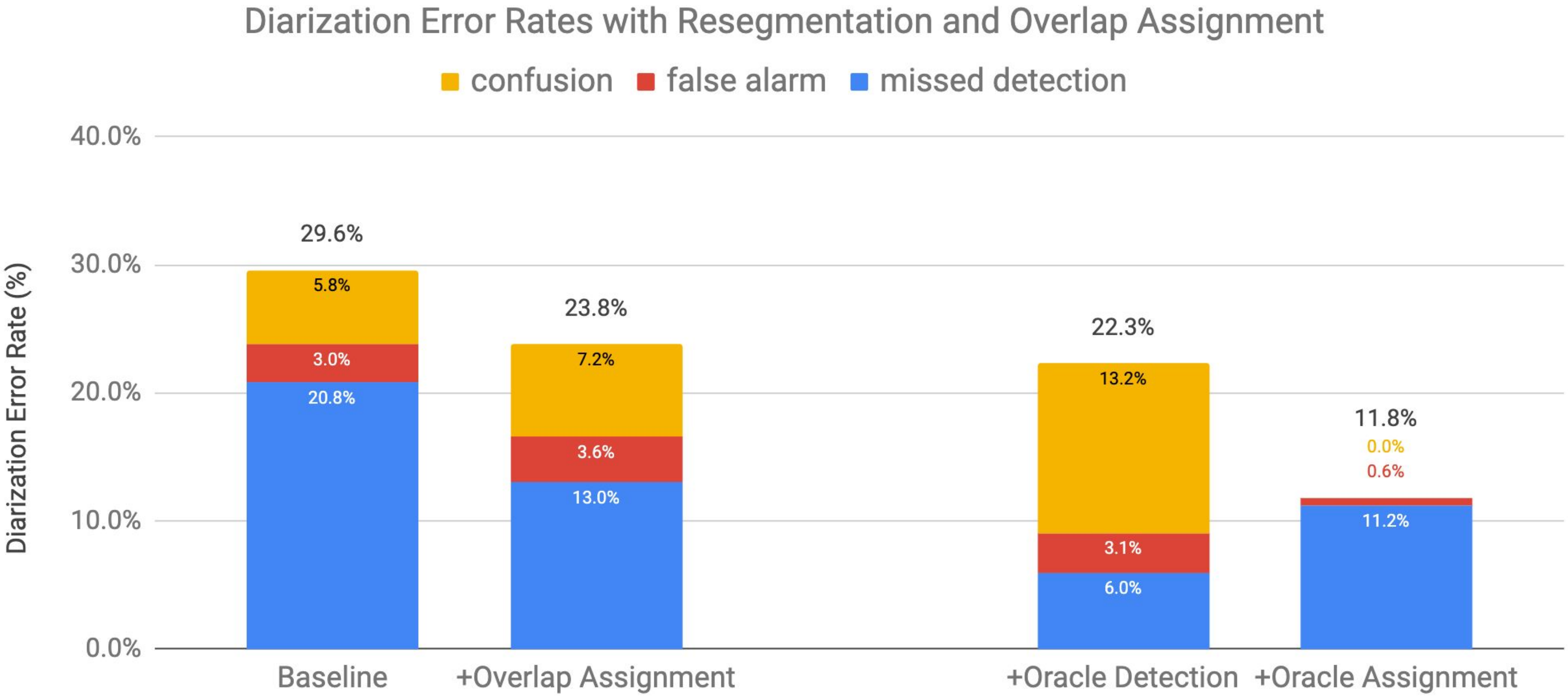
Voice Activity Detection

Final Hypothesis

Overlap Detection

2nd most likely speaker

Results: Overlap Assignment on AMI Headset Mix



Conclusions and Takeaways

Overlap detector

- State-of-the-art performance on AMI and ETAPE, sets standard for future comparison on DIHARD II
- Primary gains from decreased missed detection

Overlap-aware Resegmentation

- Results in large decreases in DER
- BUT at the cost of increases in confusion error

Acknowledgements

2019 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies (JSALT)

L'École de Technologie Supérieure (Montreal, Canada)

Johns Hopkins University with unrestricted gifts from Amazon, Facebook, Google, and Microsoft. It was also supported by the French National Research Agency (ANR) via the funding of the PLUMCOT project (ANR-16- CE92-0025).