CHiME WORKSHOP

Diarization: the missing link in Speech Technologies

Leibny Paola Garcia Perera

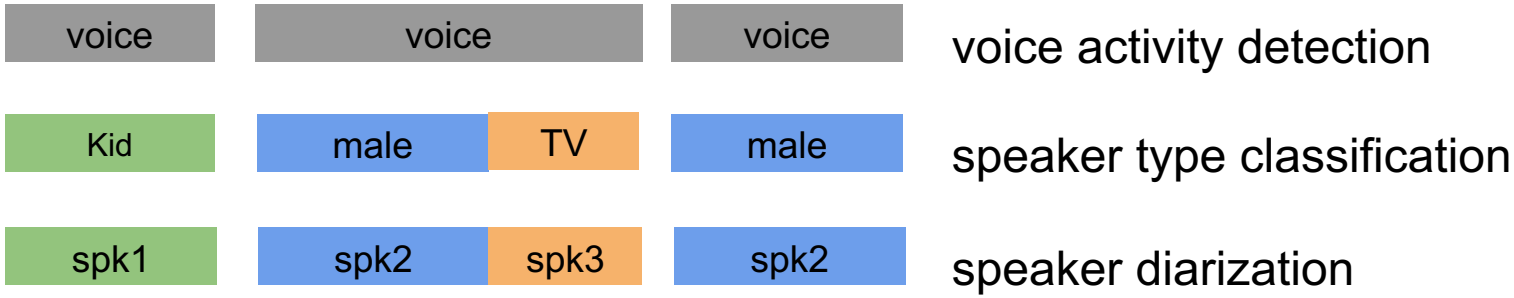
Special thanks to: Latane Bullock, Zili Huang, Jiamin Xie, Fei Wu, Herve Bredin, Shinji Watanabe, Jesus Villalba, Dan Povey, Sanjeev Khudanpur

ALL THE TEAM at JSALT workshop

CHiME5 and CHiME6 JHU TEAMs

What data do we have out there?

What is diarization?

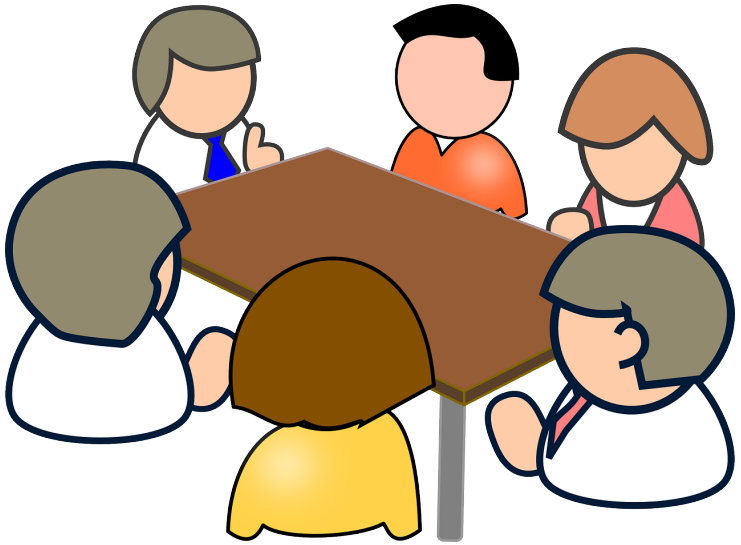


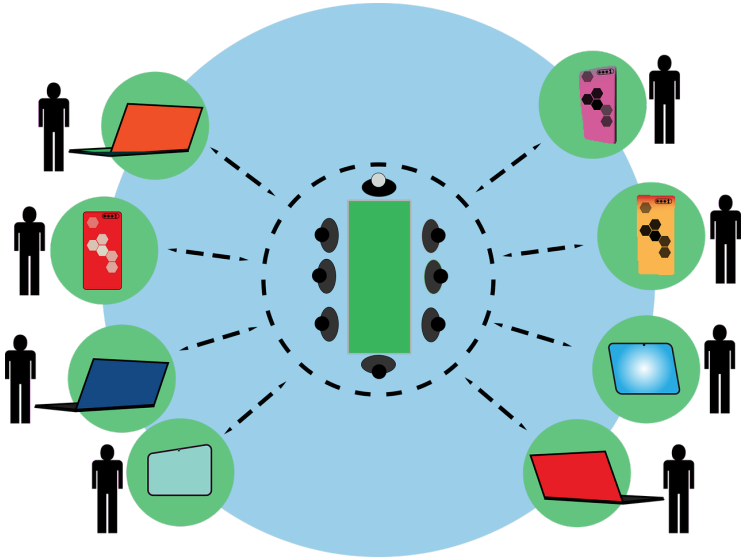
Who spoke when?

What for?

- For transcribing a meeting

S1: If we want to address the next diarization problems..
S2: You should need to first breakdown those results.
S3: Ok, I will put them in a Table or graph.





What for?

- For transcribing patient/doctor

What for?

- For segmenting long-day recordings

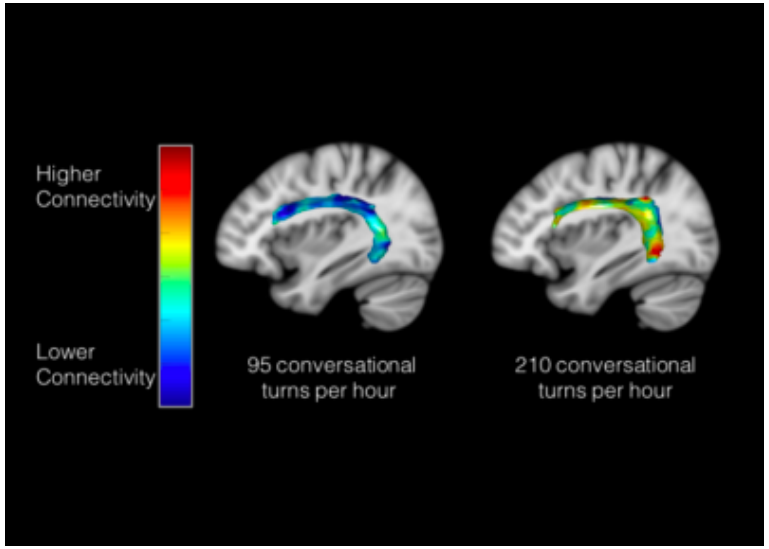
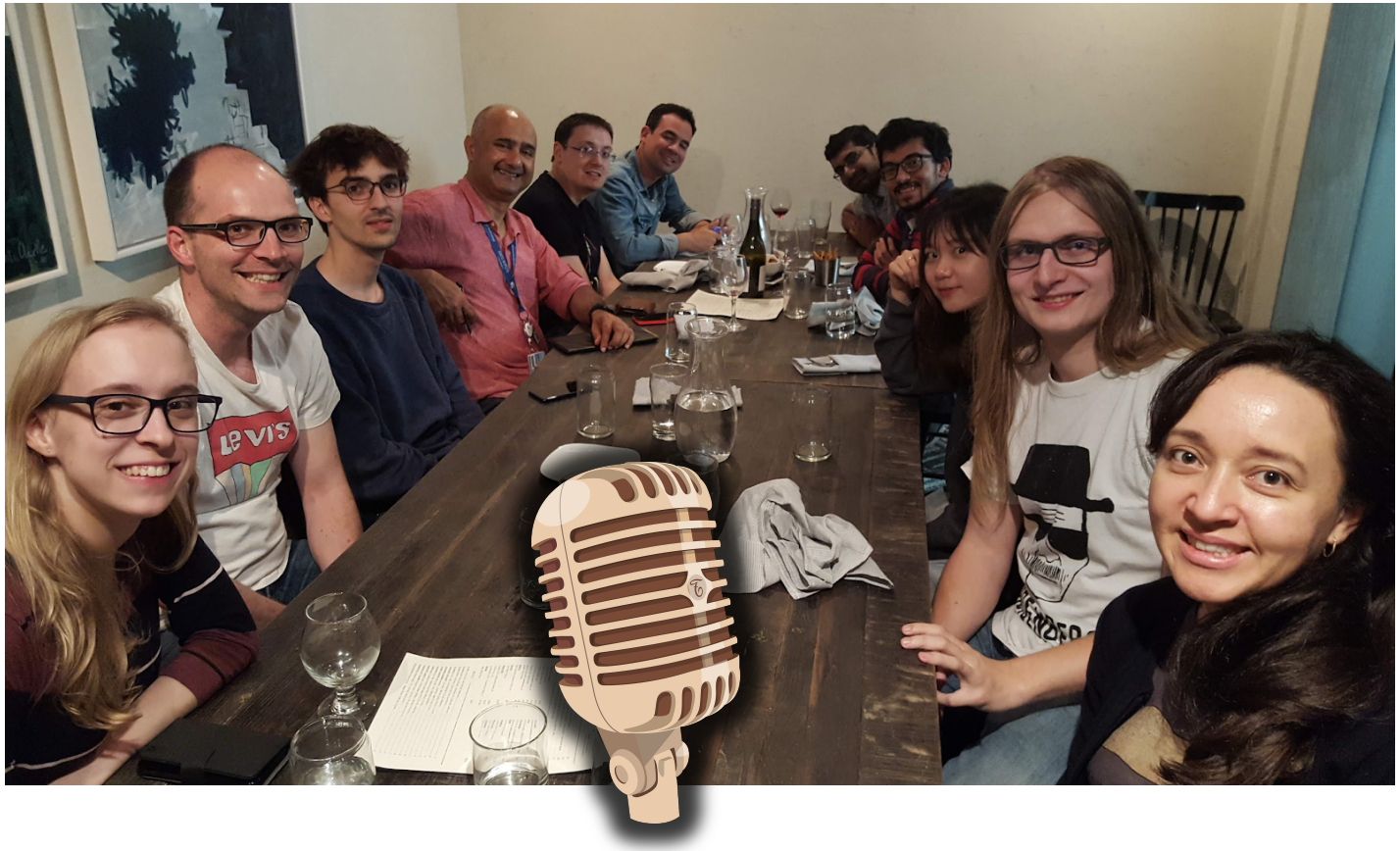


Figure from LENA , paper Language Exposure Relates to Structural Neural Connectivity in Childhood, DARCLE

What for?

- For fun ☺, tag your friends…



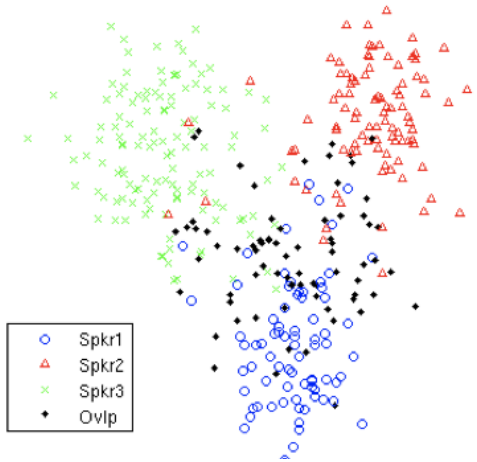
In the beginning…

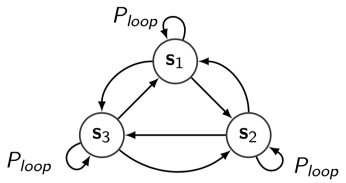
- In the beginning diarization needed help from ASR.
  - First, ran the ASR and then diarization.

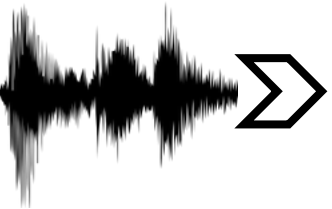
  - Now things changed… diarization first, then ASR

The story about diarization

- State of the art pipeline (ideal case)







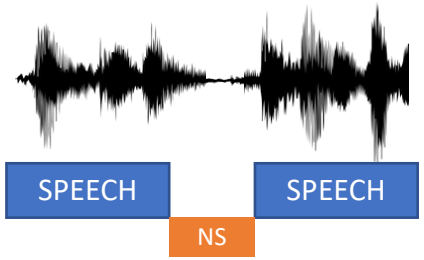
VAD Segmentation

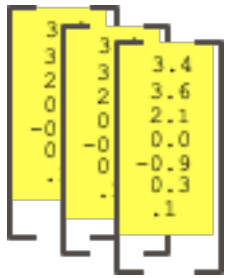
Embedding extractor

Scoring

CLustering

Resegmentation

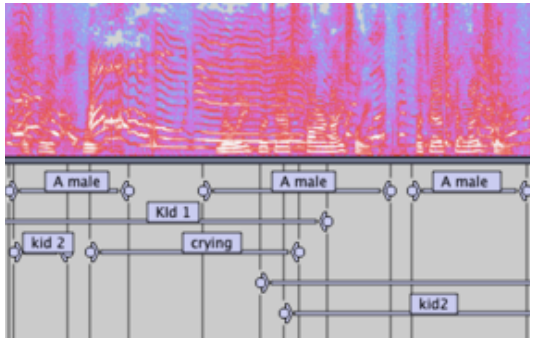




$$X = \begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & x_{1m} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & x_{nm} \end{bmatrix}$$

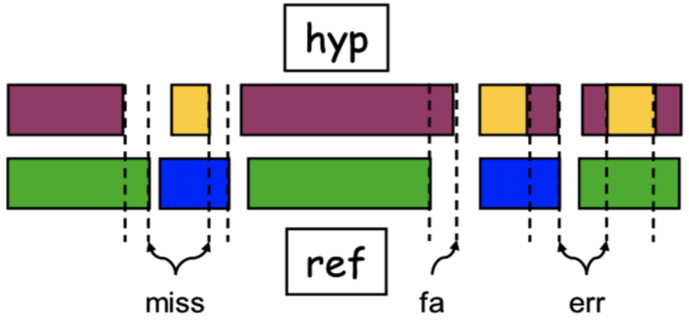
rttm

rttm



The story about diarization

- Metrics

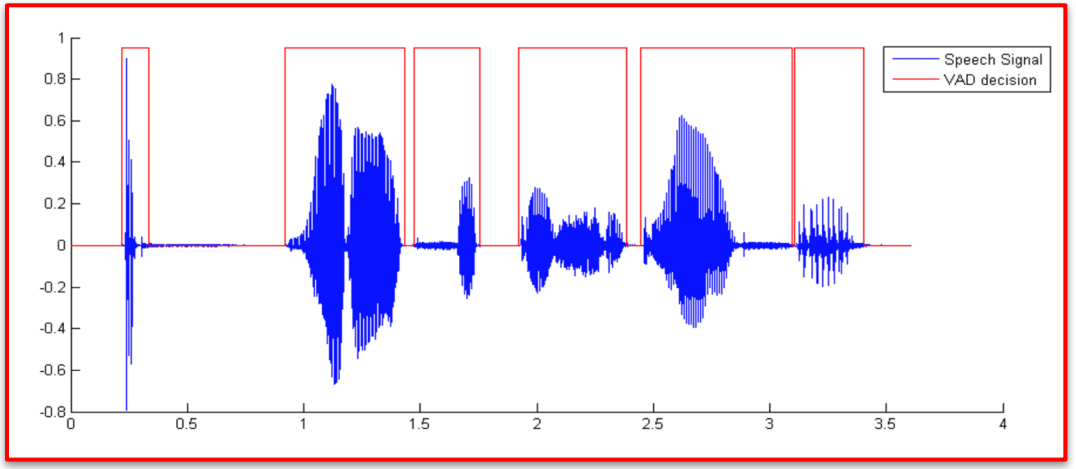


Diarization Error Rate (DER)

$$\text{DER} = \frac{\text{false alarm} + \text{missed detection} + \text{confusion}}{\text{total}}$$

The story about diarization

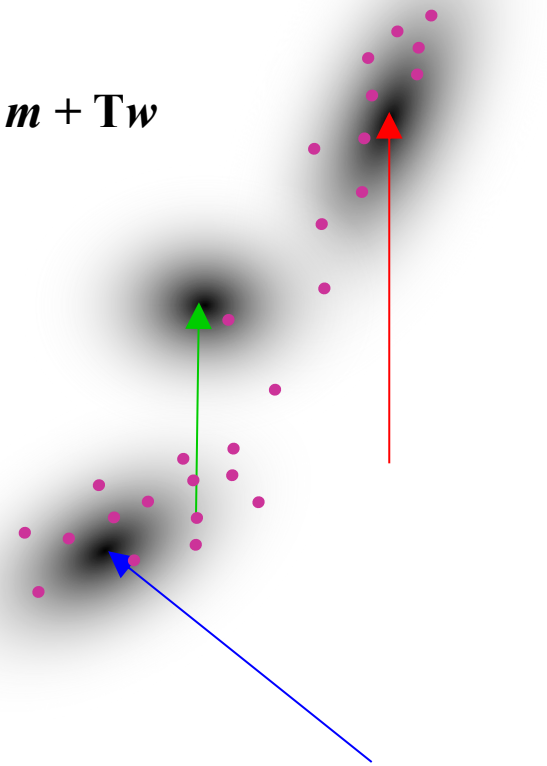
- Voice activity detector (VAD)
  - Energy VAD (seemed to work well)



The story about diarization

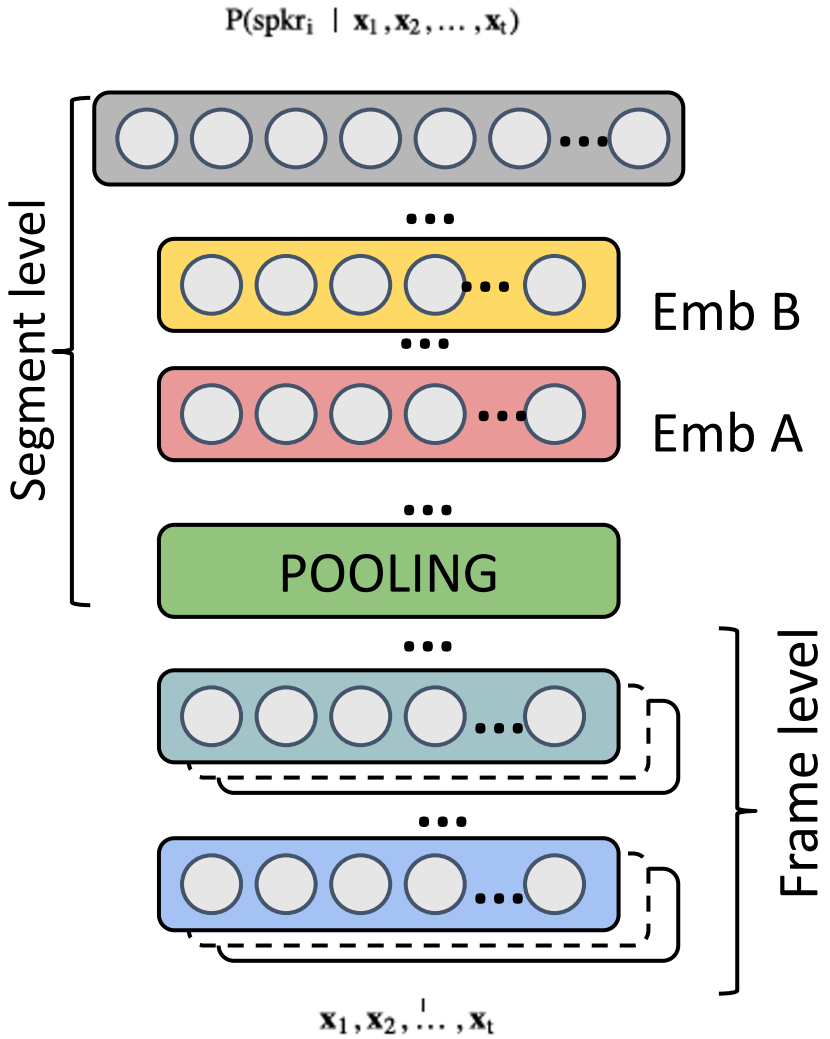
- Embeddings

$$\mathbf{M} = \boldsymbol{m} + \mathbf{T}\boldsymbol{w}$$



i-vector

x-vector

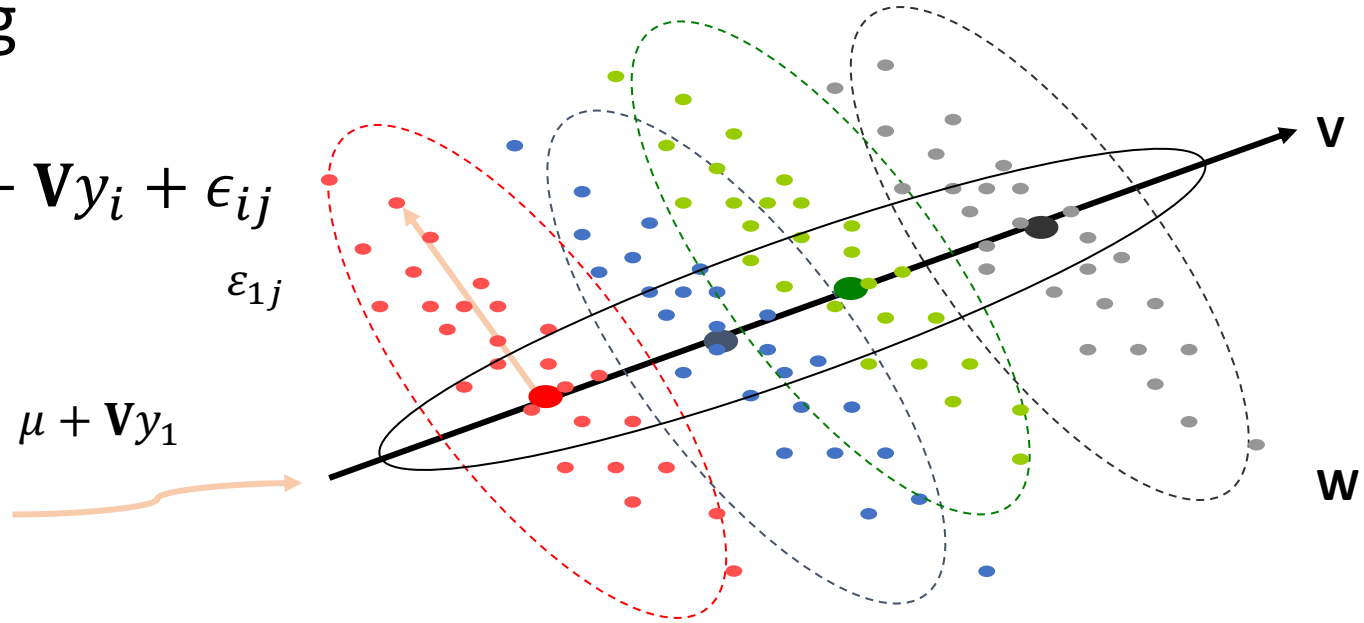


Work from Najim Dehak, et.al., 2010,
David Snyder, et.al., 2016-2018.

The story about diarization

- PLDA Scoring

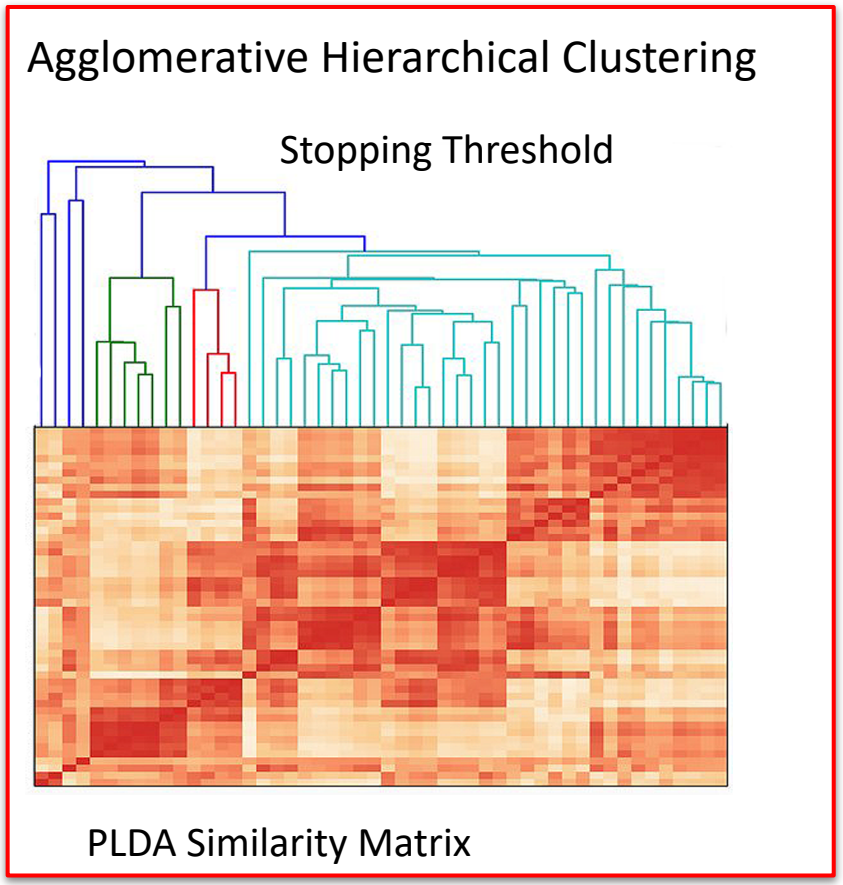
$$w_{ij} = \mu + \mathbf{V}y_i + \epsilon_{ij}$$



$$\mathrm{LLR} = \log \frac{P(w_1, w_2|same)}{P(w_1, w_2|diff)} = w_1^T A w_2 + \mathrm{w_1^T B w_1} + \mathrm{w_2^T B w_2} + \mathrm{C^T w_1} + \mathrm{C^T w_2} + \mathrm{D}$$

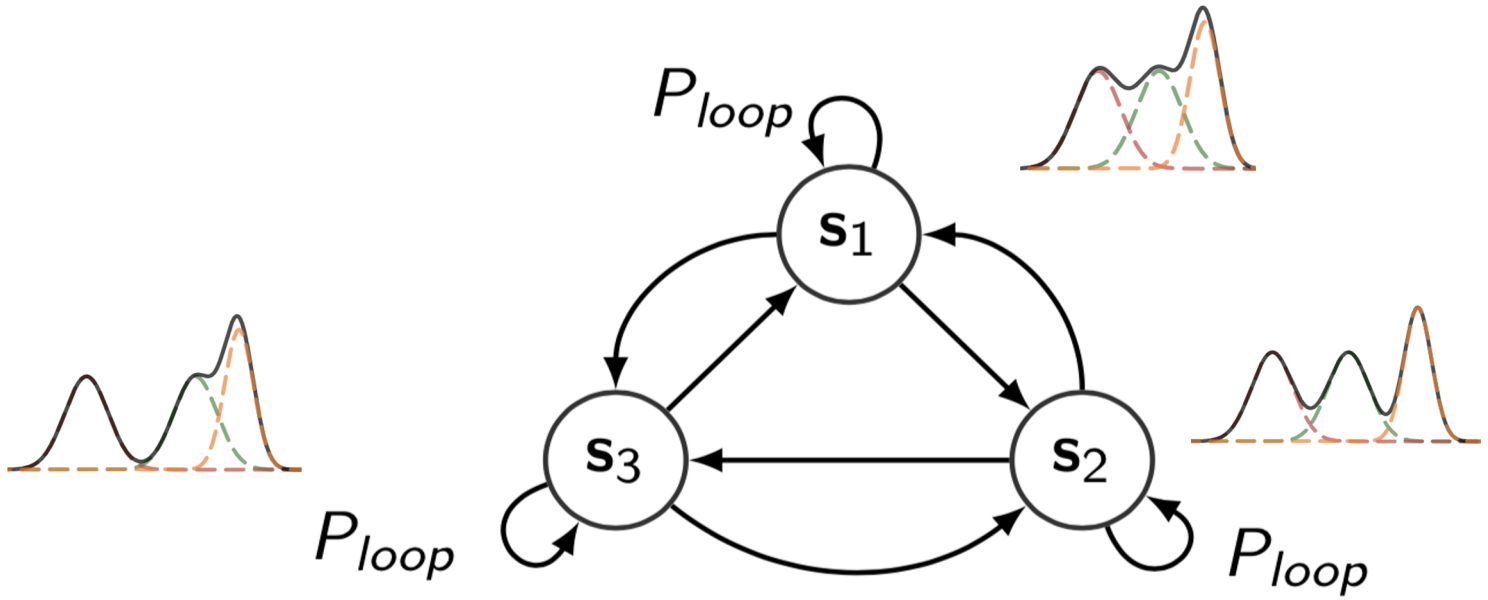
The story about diarization

- Clustering



The story about diarization

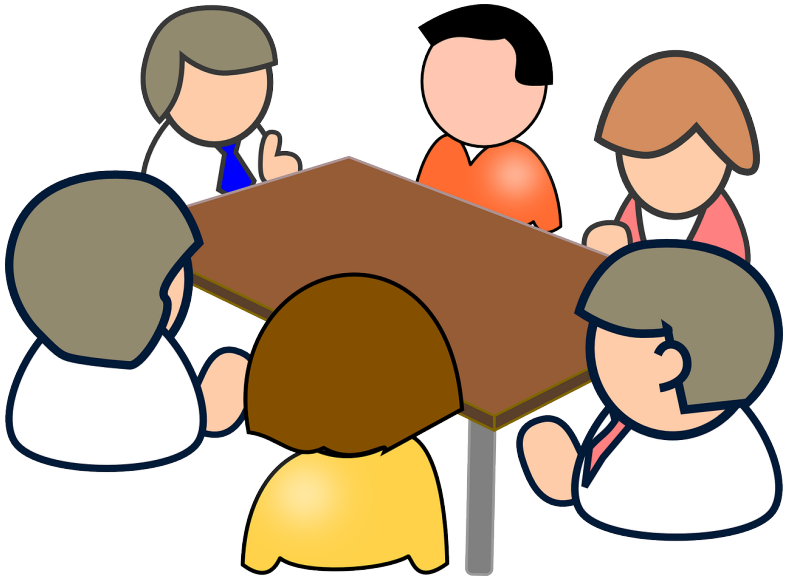
- VB-HMM Re-segmentation



Work from Mieria Diez, et.al., 2018

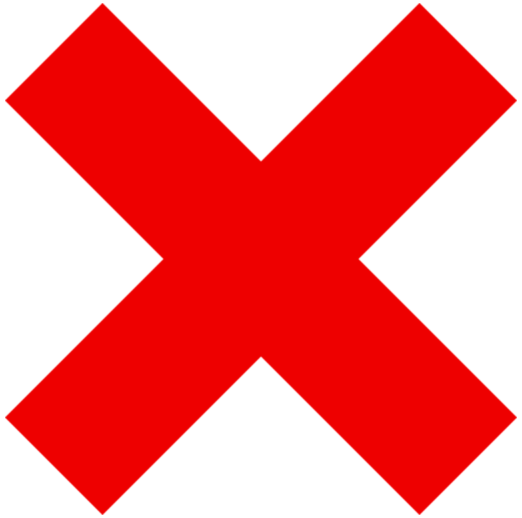
Does it work fine?

It does when we have collaborative speakers and almost no noise.



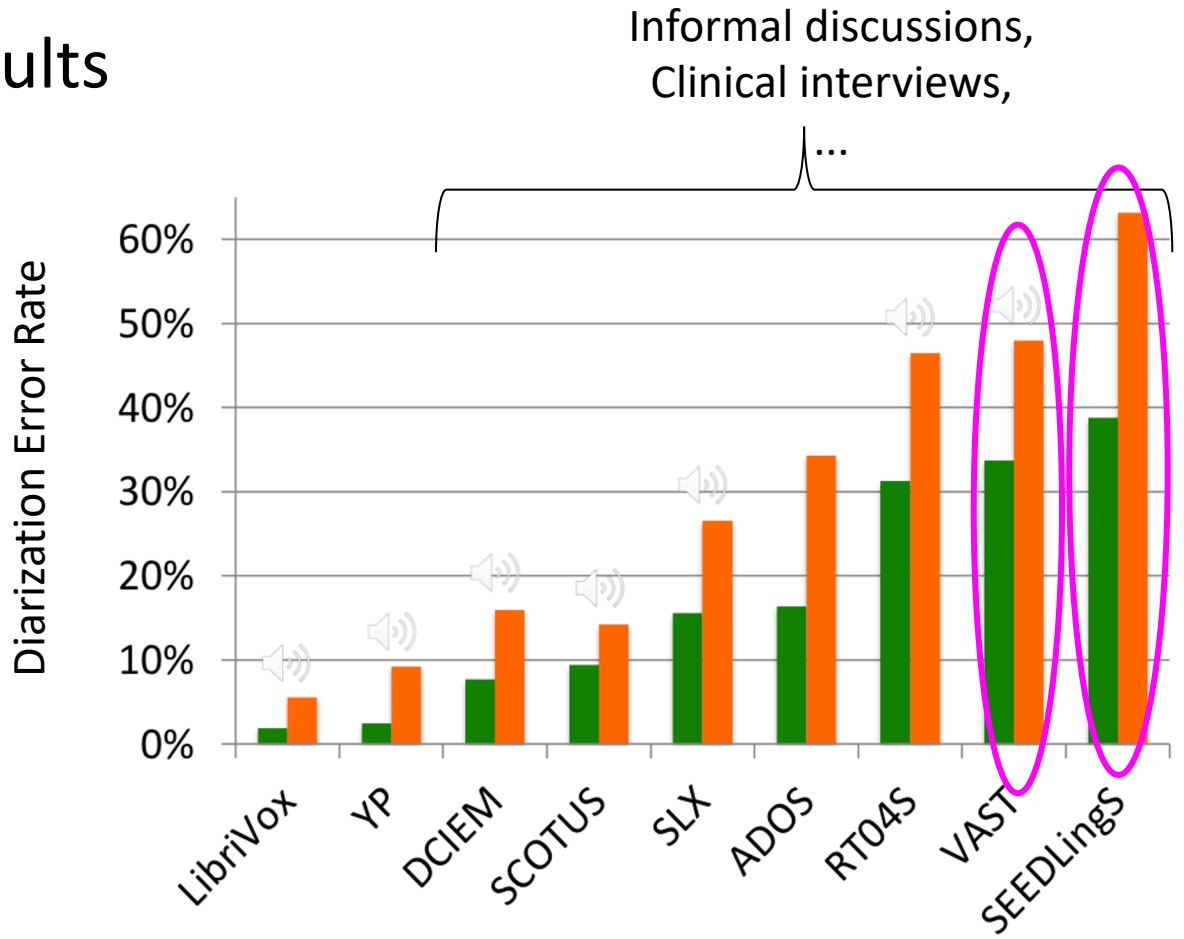
Does it work fine?

What about other scenarios?



Let’s see some numbers

- Dihard I results

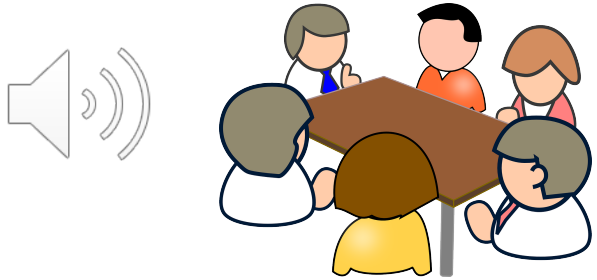


Work by Greg Sell, et.al., 2018

Shortcuts from JSALT

What are the characteristics of your data?

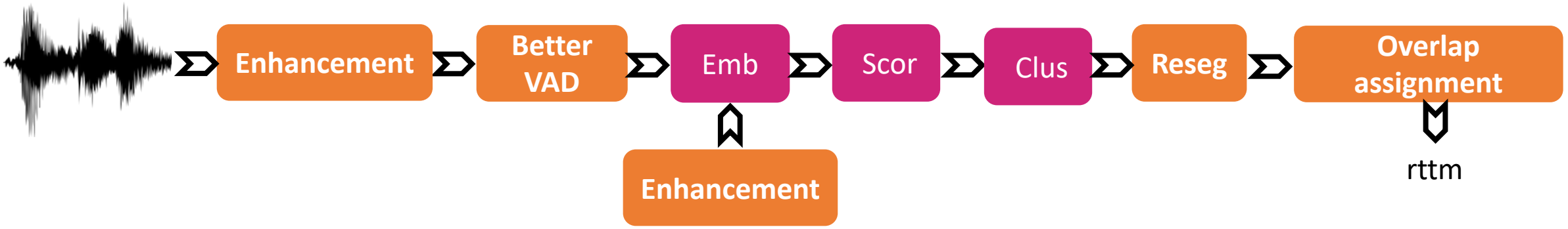
Is it like this one?



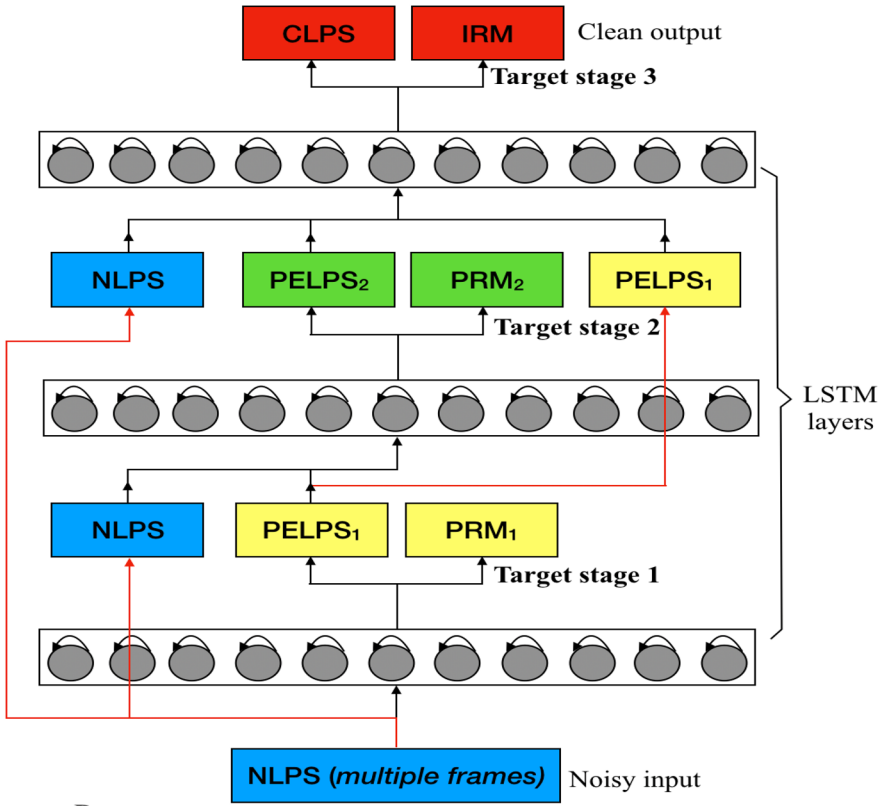
Or like this one?

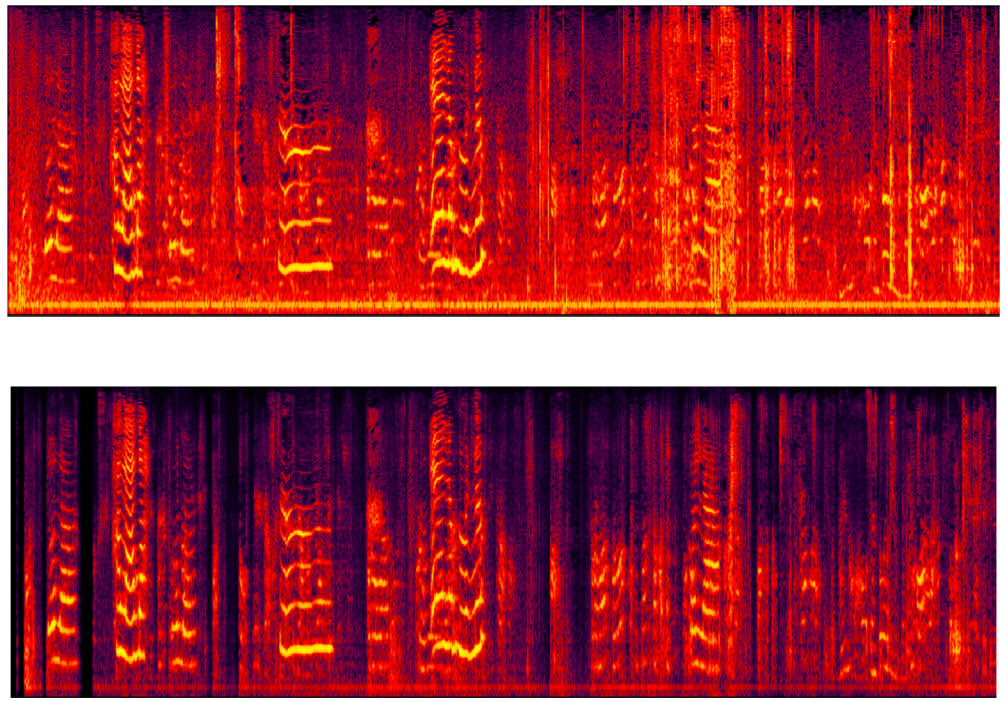
Shortcuts from JSALT

Will talk about them one by one, here the big picture



Enhancement



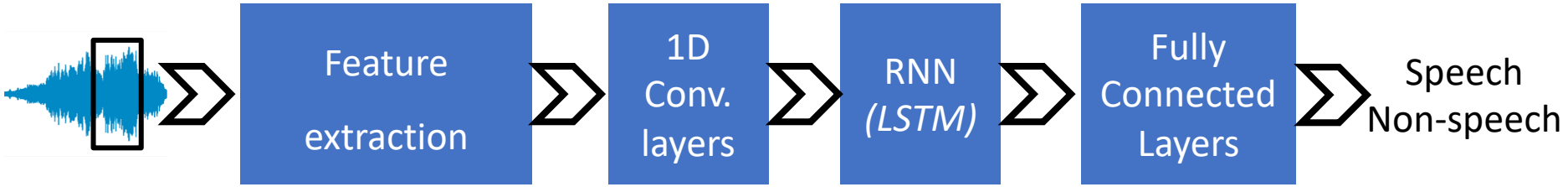


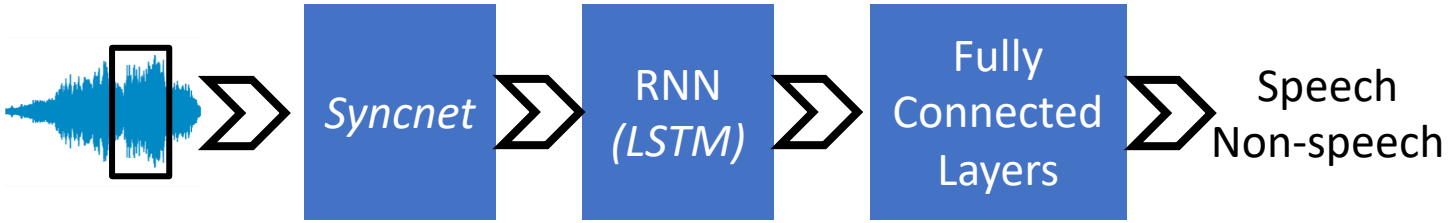
Work by Lei Sun, et.al., at JSALT 2019 workshop

SuperVAD

- Neural network VAD

pyannote

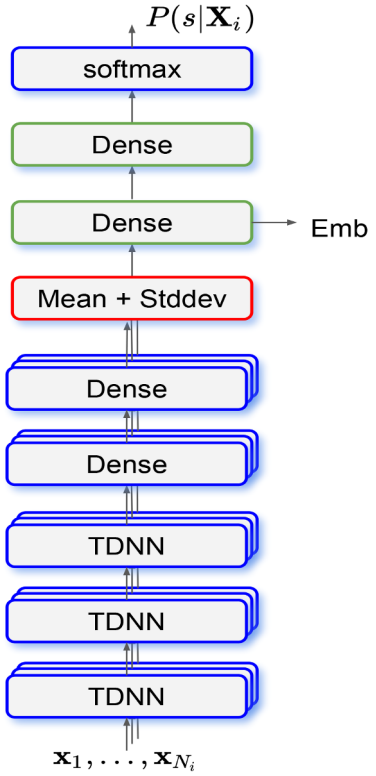




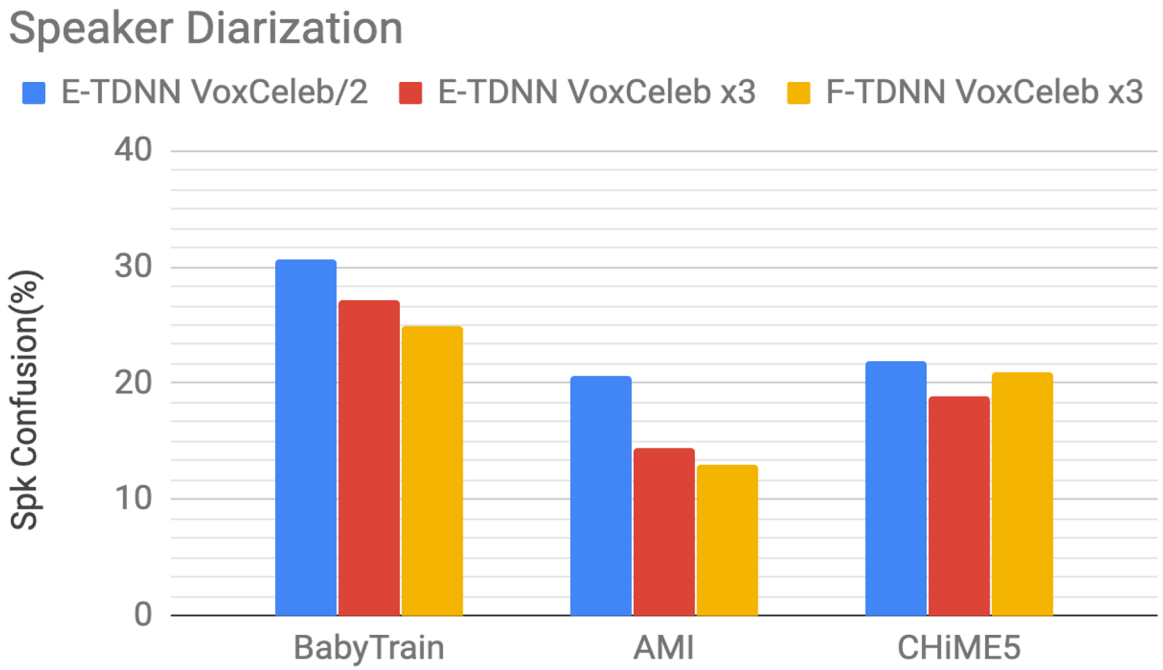
Work by Marvin Levechin, et.al., at JSALT 2019 workshop

Embeddings

- Enrollment x-vector
- N Test x-vectors: 1 per diarization cluster

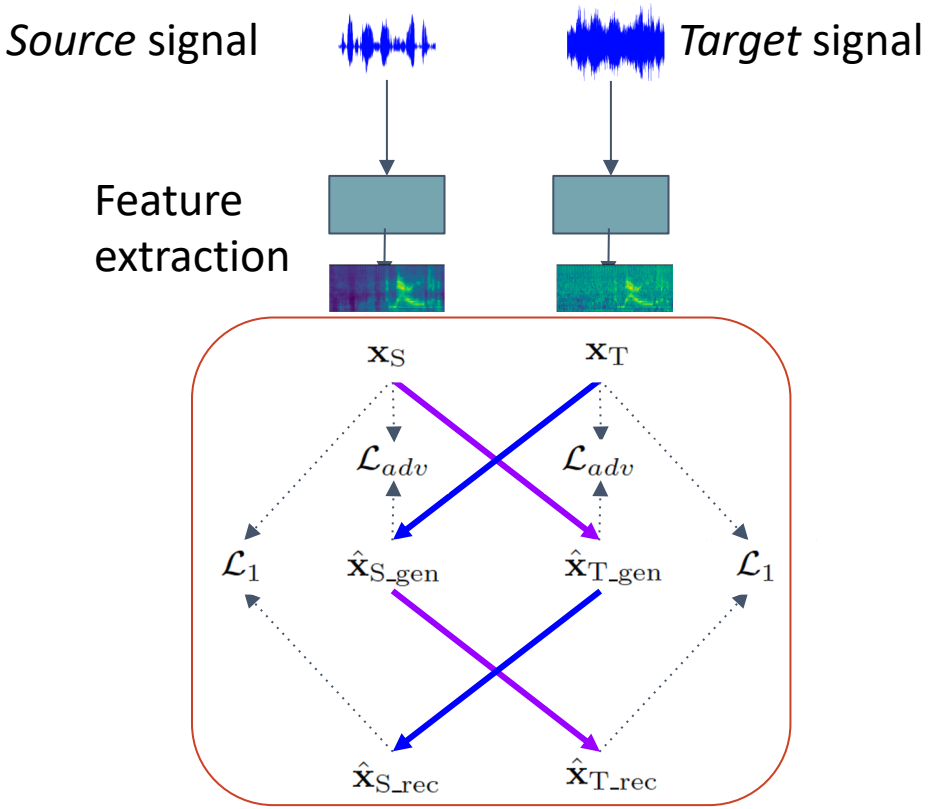


* **F- TDNN** plus **augmentation** showed the best result



Work by Jesus Villalba, et.al., at JSALT 2019 workshop

Feature enhancement

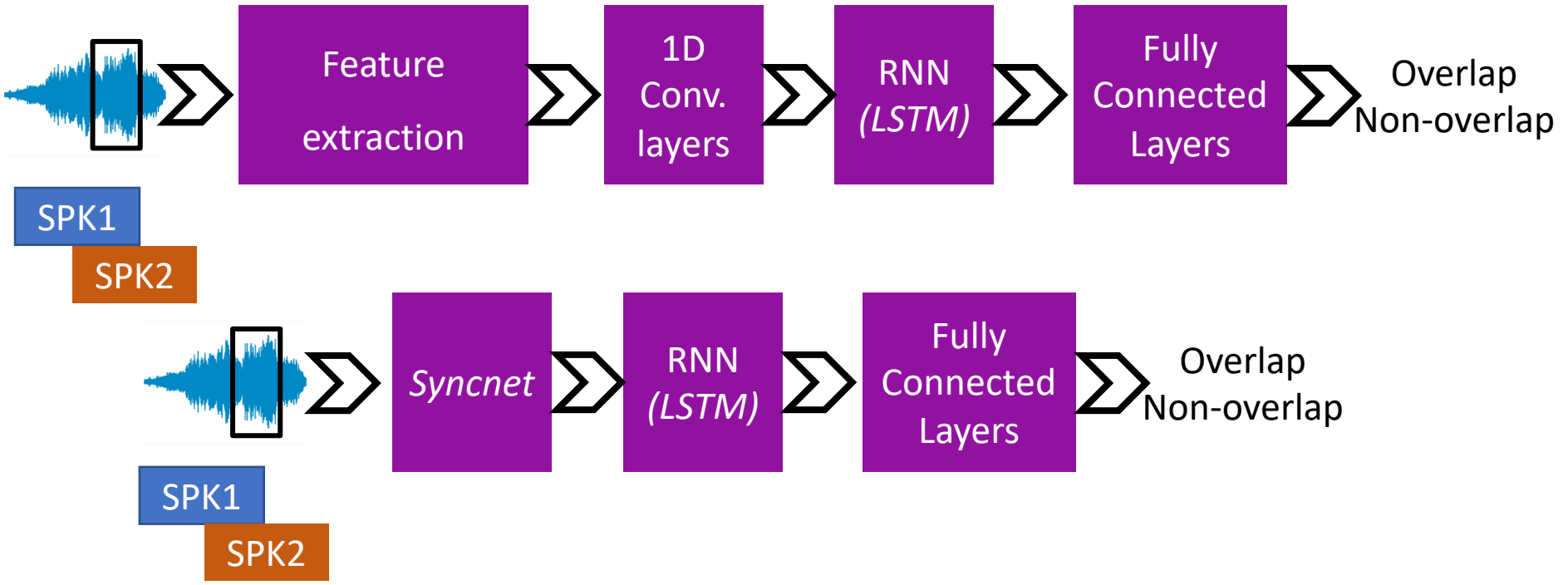


Work by Phani Nidadavolu et.al., at JSALT2019 workshop

Overlap Assignment

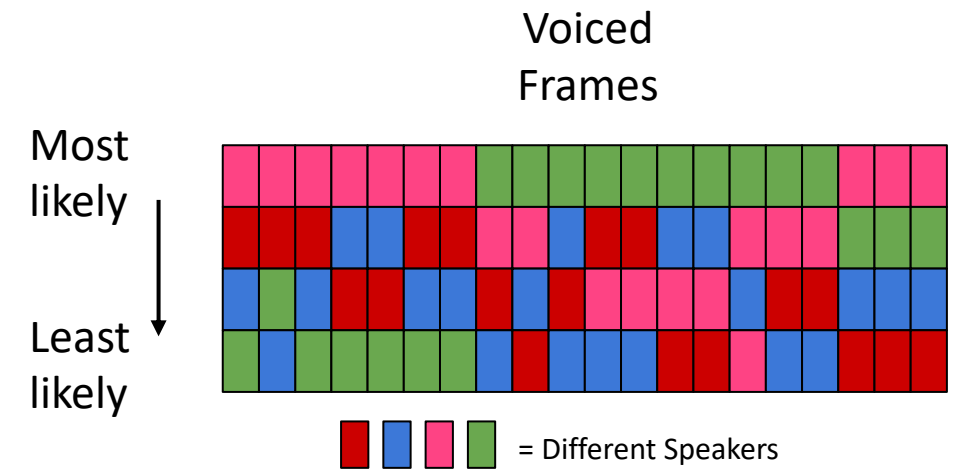
- Neural network Overlap detector

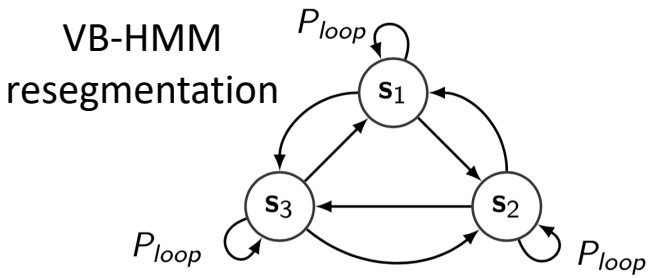
pyannote

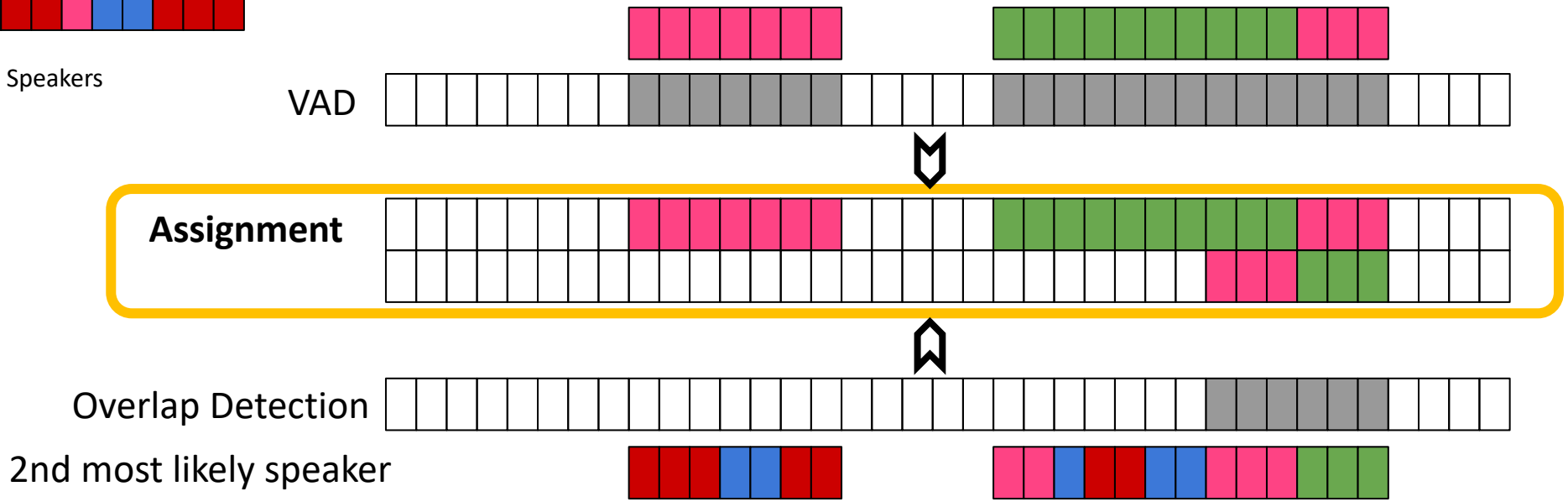


Work by Latane Bullock, et.al., at ICASSP 2020

Overlap Assignment





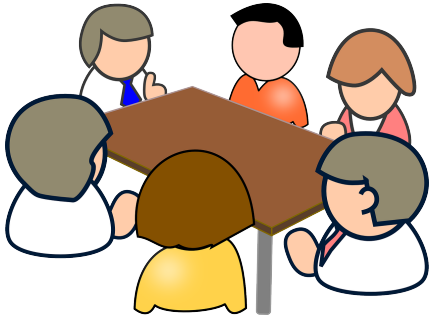


Work by Latane Bullock, et.al., at ICASSP 2020

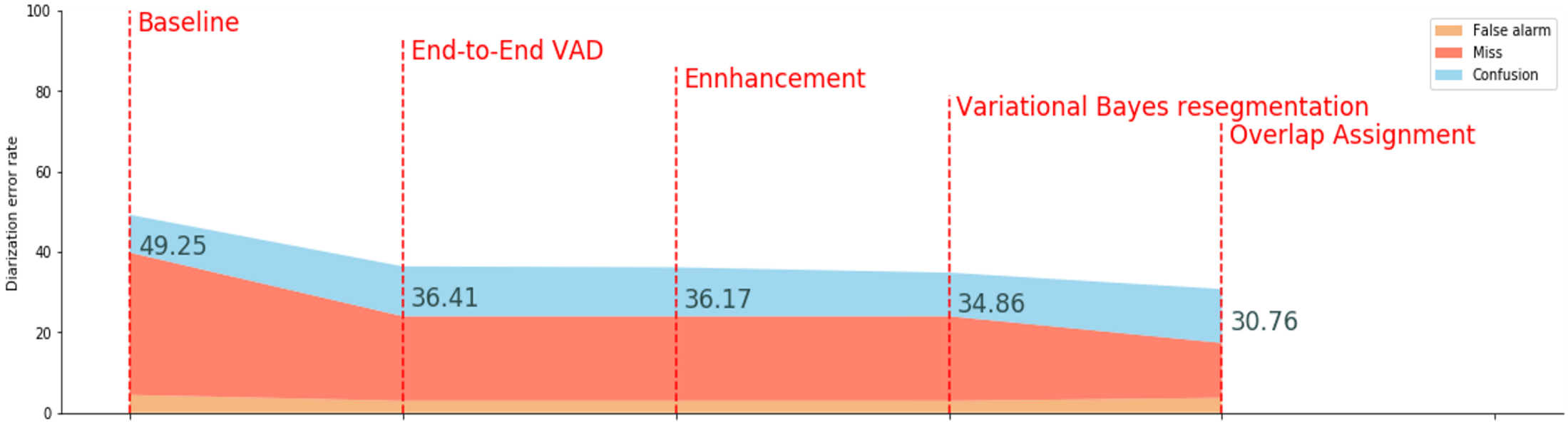
Some results on real datasets

Domain	Setting	N speakers/ sessions	Time	Provider
Meeting	3 different meeting rooms	180 speakers x 3.5 sessions per speaker (sps)	98 h	AMI
Indoor	(kitchen, dining, living)	80 speakers	50 h	Chime5
Wild	*uncontrolled*	450 recurrent speakers, up to 40 sps (longitudinal)	225 h	BabyTrain

In the end what we got?

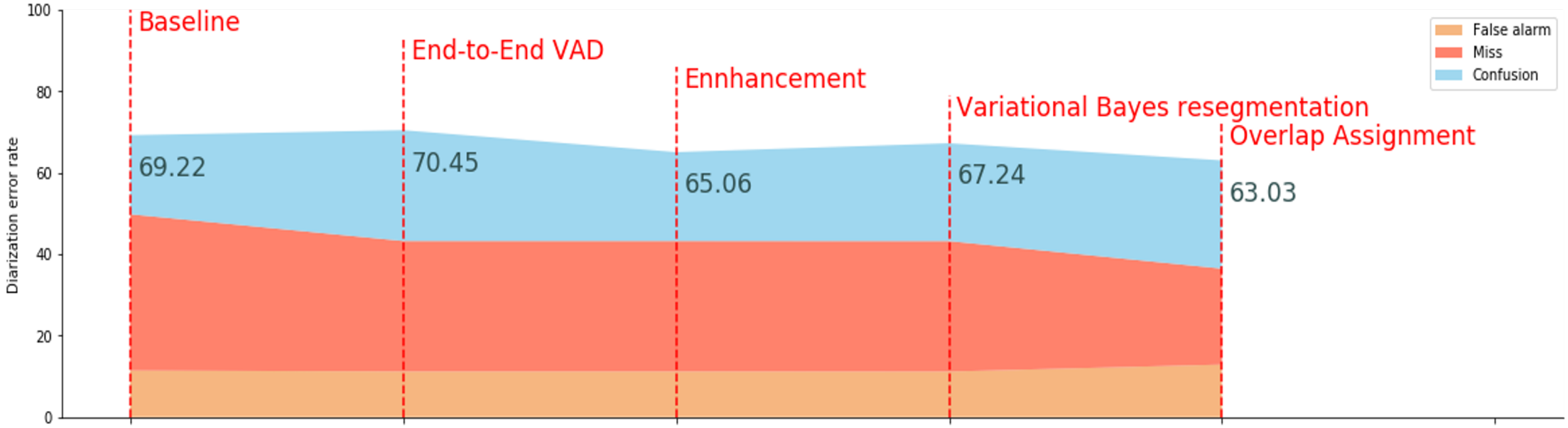


- AMI



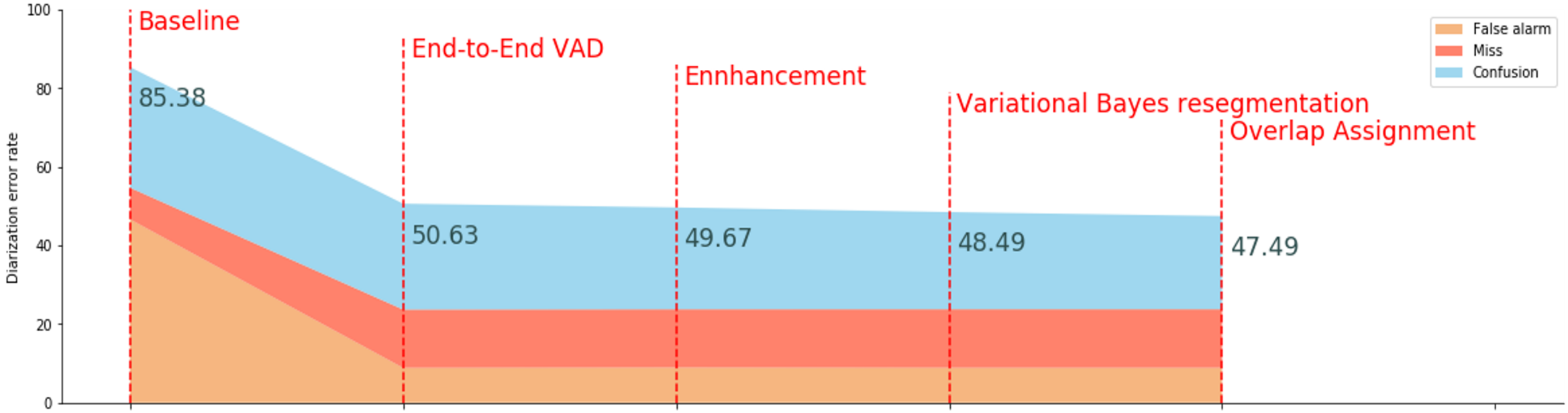
In the end what we got?

- CHiME5

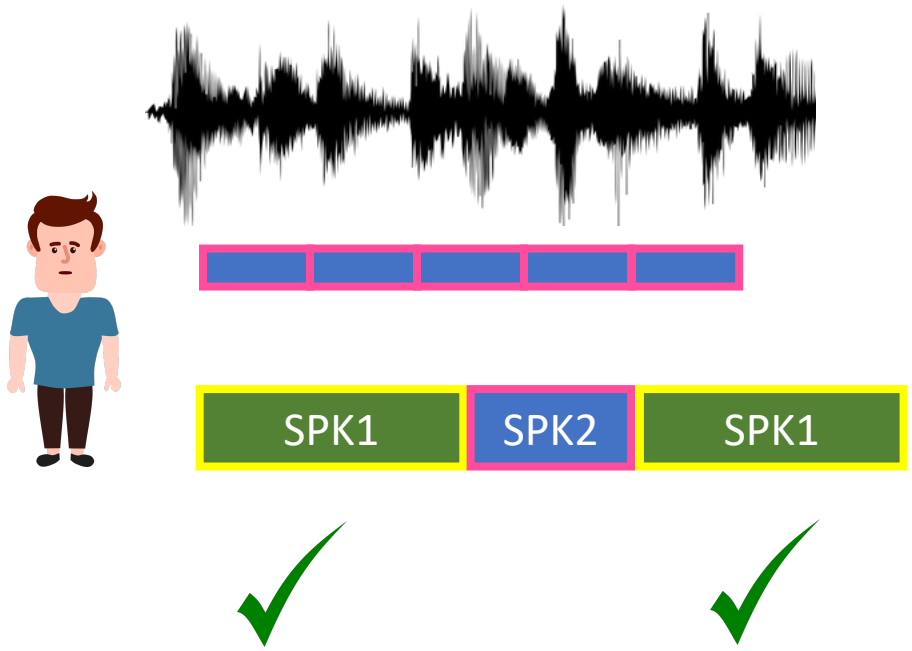


In the end what we got?

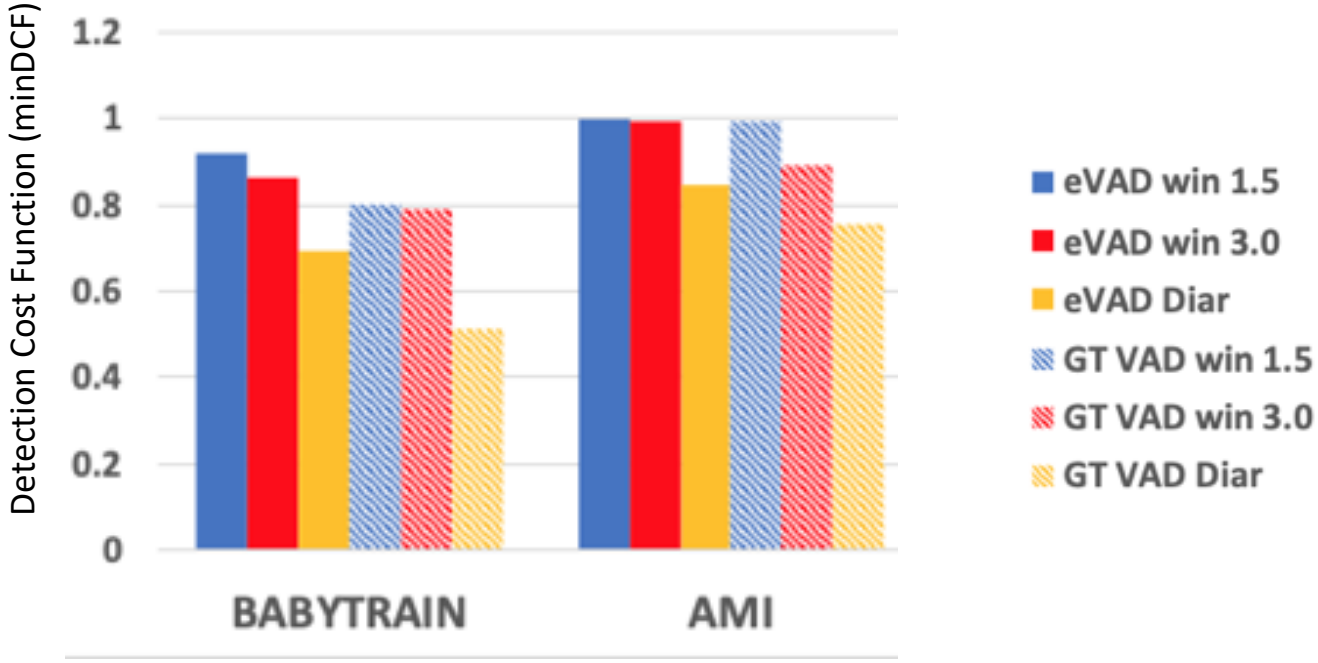
- BabyTrain



Side effect of the diarization on Speaker Tracking



Windowing vs Diarization with Energy VAD and ground truth VAD



Work by Diego Castan et.al., at JSALT 2019 workshop

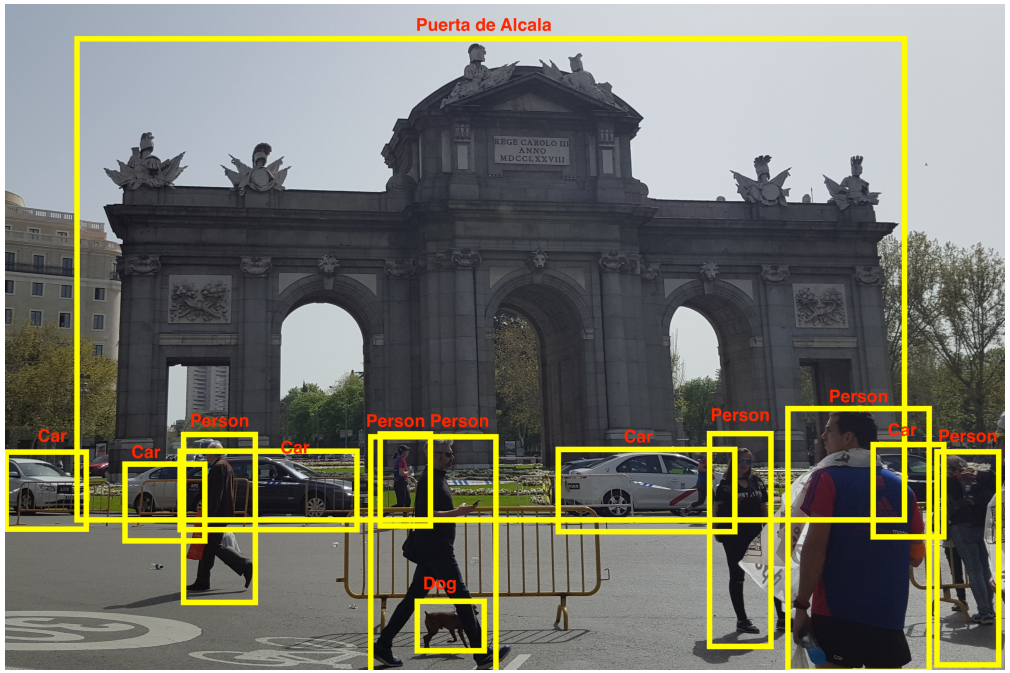
Two questions arise now…

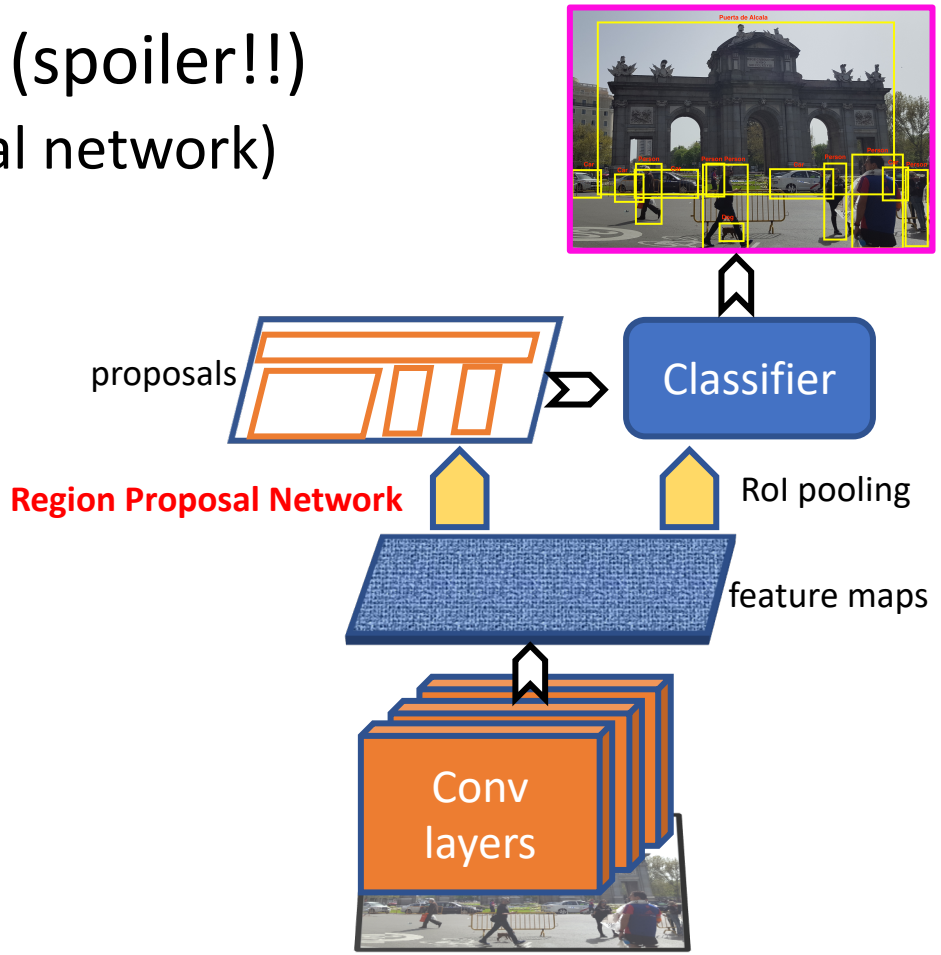
- Are there End-to-end or neural network diarization systems?
- How to use this knowledge can be used in multiple-array ASR?

~~What is the present?~~ What is the future?

Faster RCNN

- One of the first attempts on using NNs (spoiler!!)
  - Called RPNSD (inspired by Region proposal network)



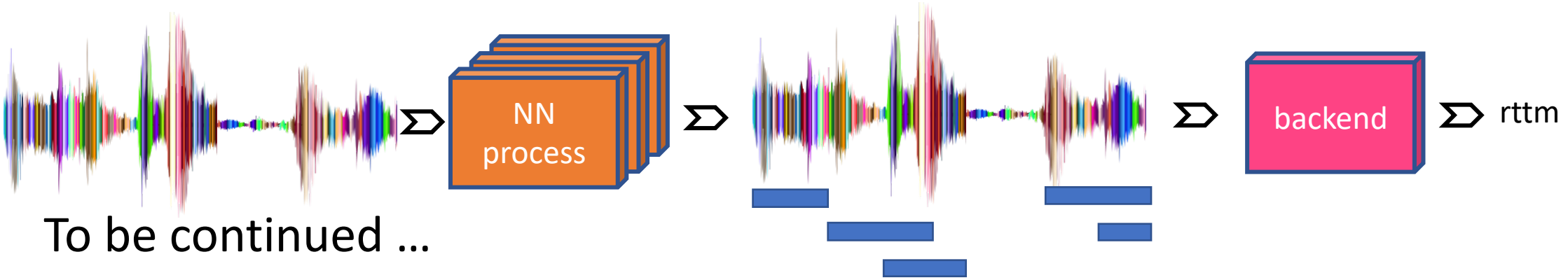


Work by Zili Huang, et.al., at ICASSP 2020

Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems*. 2015.

What is the present?

- Same idea but how to do it with speech ☺
- A good contribution so far and moving forward.

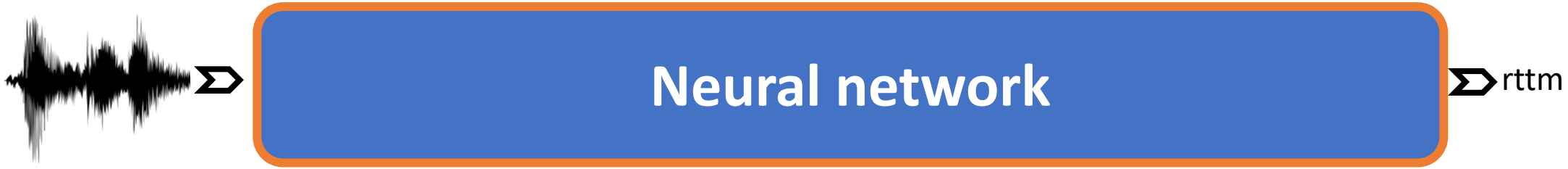


To be continued …

Work by Zili Huang, et.al., at ICASSP 2020

What is the present?

- End-to-end Neural Diarization (EEND)

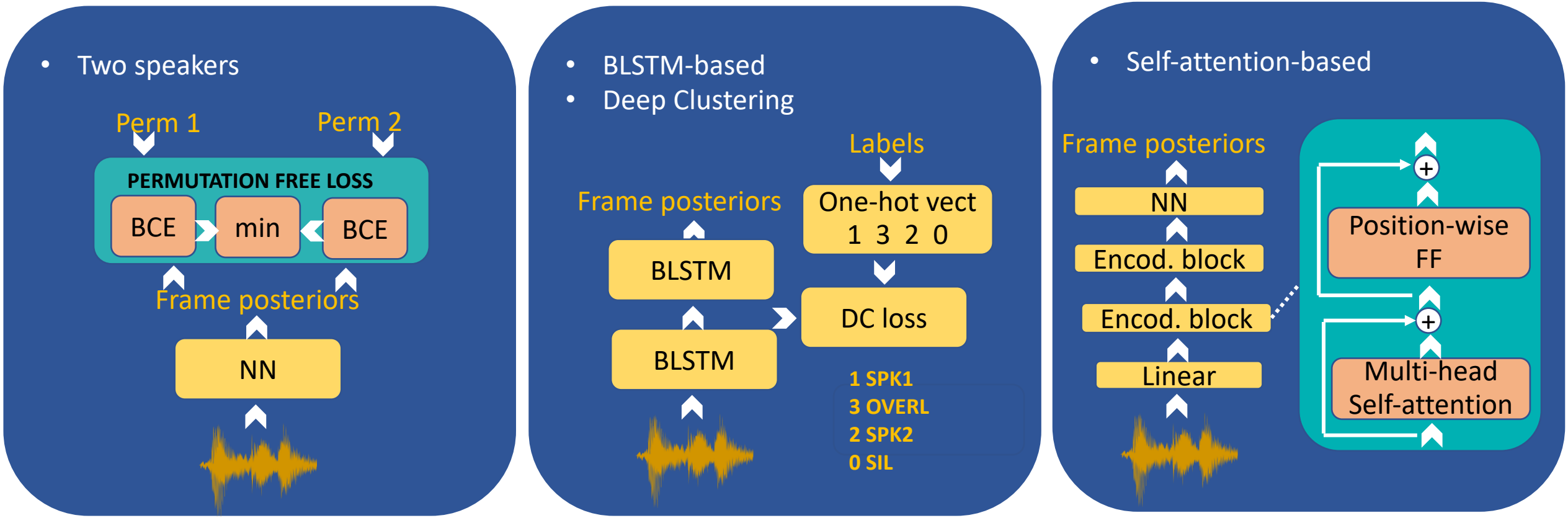


Method	Callhome DER (%)
X-vector	11.53
SA-EEND	12.66
SA-EEND adapted	10.76

Work by Yusuke Fujita, et.al., 2019 at https://arxiv.org/pdf/2003.02966.pdf

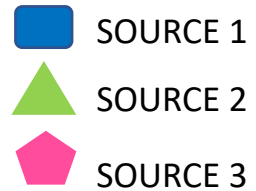
What is the present?

- EEND evolution

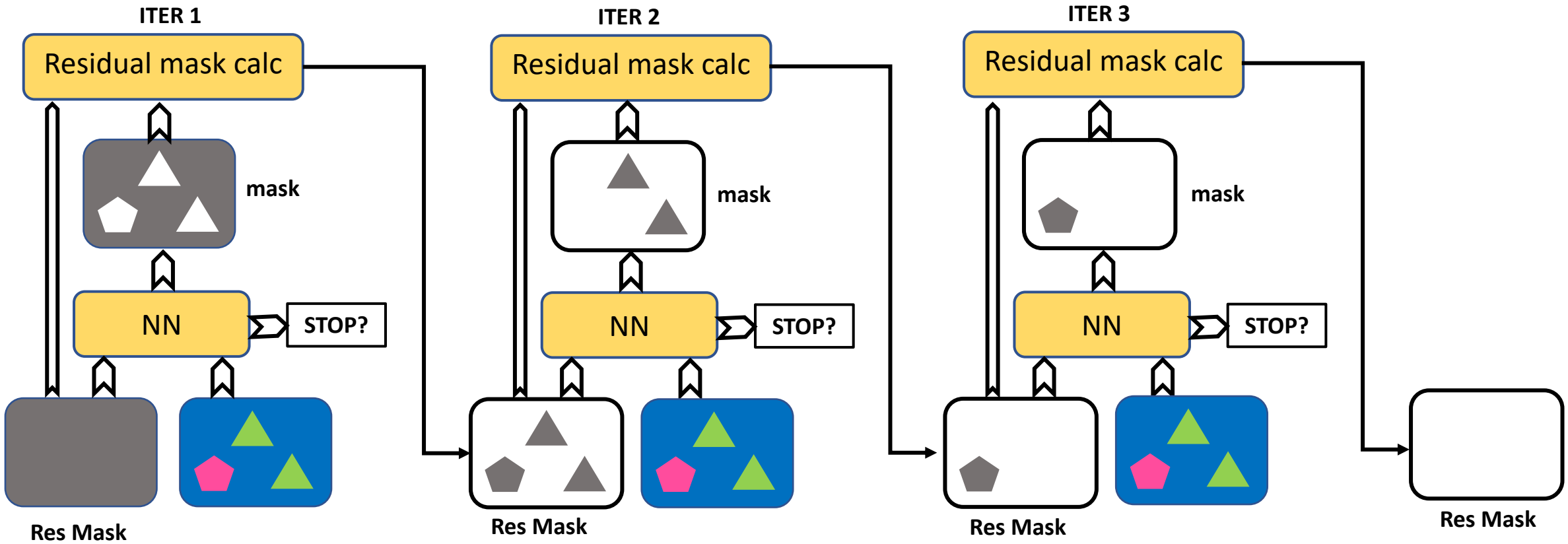


Work by Yusuke Fujita, et.al. 2019, at https://arxiv.org/pdf/2003.02966.pdf

What is the present?



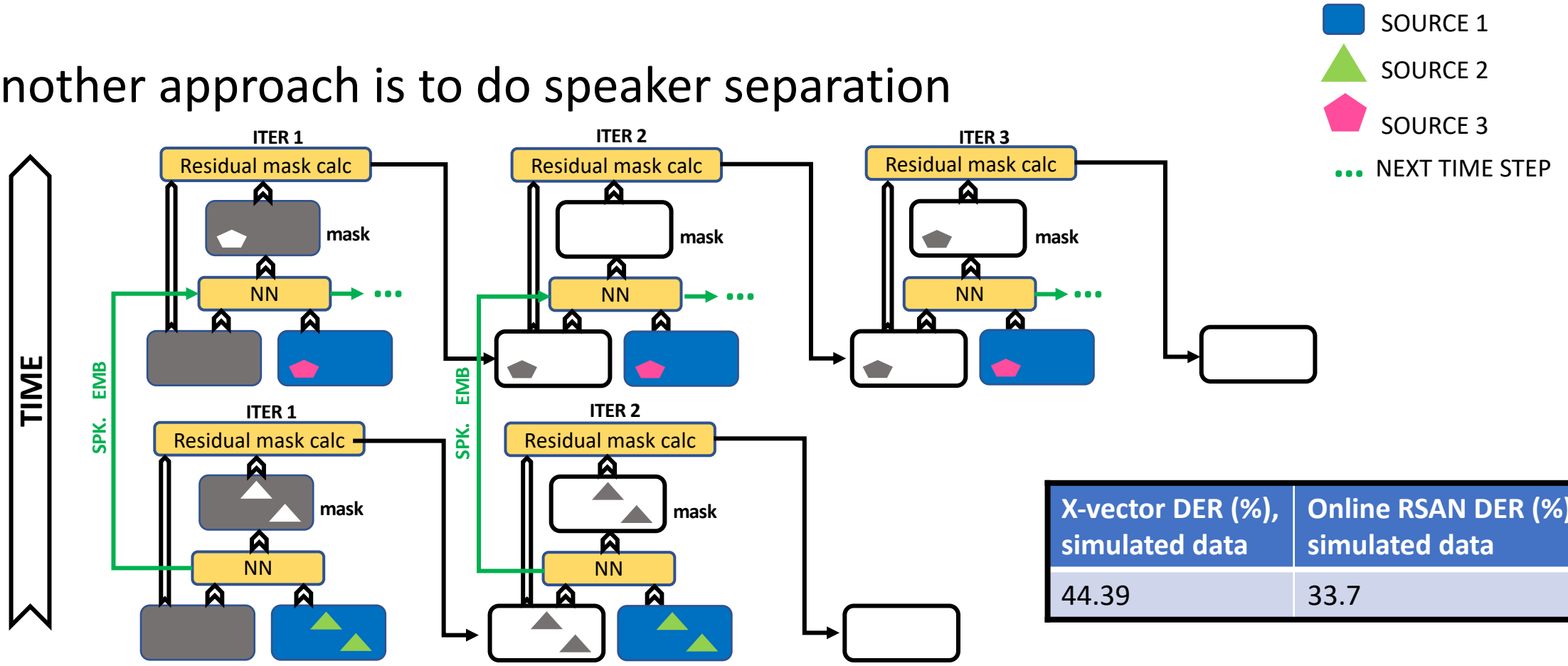
- RSAN (Recurrent selective attention network)



Work by Keisuke Kinoshita, et.al., at ICASSP 2020 https://arxiv.org/pdf/2003.03987.pdf

What is the present?

- Another approach is to do speaker separation



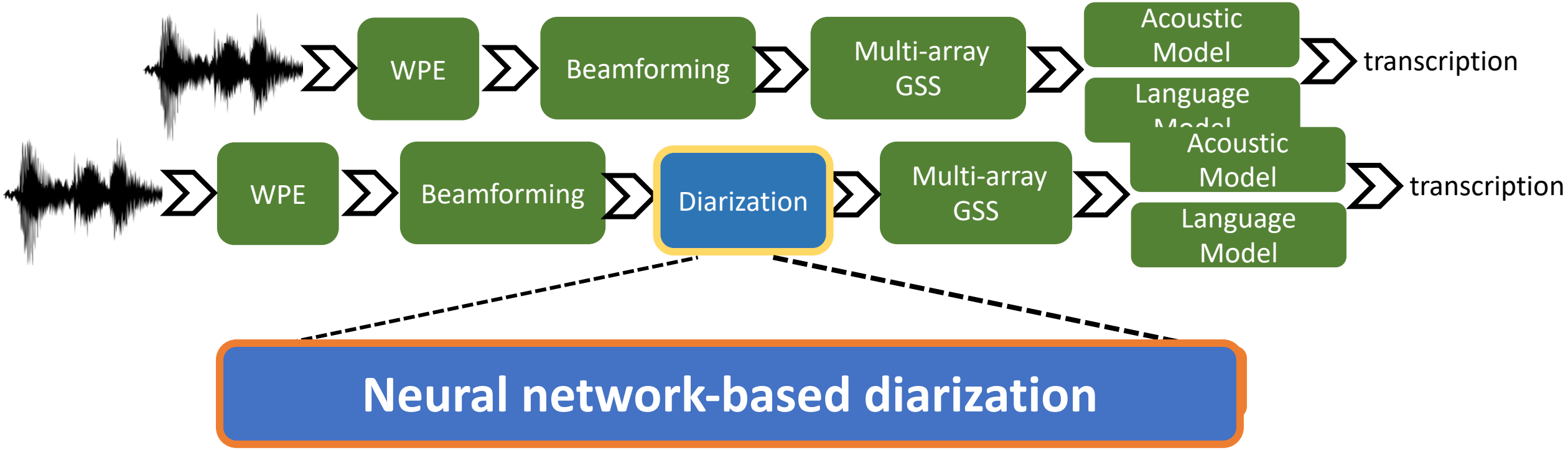
X-vector DER (%), simulated data	Online RSAN DER (%), simulated data
44.39	33.7

Work by Keisuke Kinoshita, et.al., at ICASSP 2020 https://arxiv.org/pdf/2003.03987.pdf

How to use this knowledge when having multiple arrays?

You can include diarization techniques as part of your ASR!

ASR and multi-microphones



Takeaways

- Diarization modules needed to have a competitive system:
  - Good enhancement
  - Good VAD
  - Good Embeddings
  - Overlap detector and assignment
- New approaches (NN-based) are also becoming very good!

Future work

- It is still not easy to estimate the number of speakers
- VAD
  - How to handle domain mismatch
- Re-think clustering
  - Unsupervised adaptation (take overlap into a account during clustering)
- Overlap assignment refinement
- How to do effective diarization on long recordings?
- Is it possible to train an ASR and a diarization system jointly?

Thank you!

Questions?

Questions?