The Yellow Brick road of diarization, challenges and other neural paths

Leibny Paola García Perera

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

CENTER FOR LANGUAGE
AND SPEECH PROCESSING

Collaborators

- Thanks to each one of them!
- JHU: Zili Huang, Desh Raj, Matthew Maciejewski, Jesus Villalba, Sanjeev Khudanpur
- Hitachi: Yusuke Fujita, Shota Horiguchi, Yawen Xue, Yuki Takashima, Nelson Yalta
- CMU: Shinji Watanabe
- NTT: Marc Delcroix, Keisuke Kinoshita
- NTU: Suzy Styles, Fei Ting Woon, Justin Dawels, Victoria Chua, Hexin Liu
- Some other collaborators: Latané Bullock, Hervé Bredin, Marvin Lavenchin, Carlos Castillo

Disclaimer

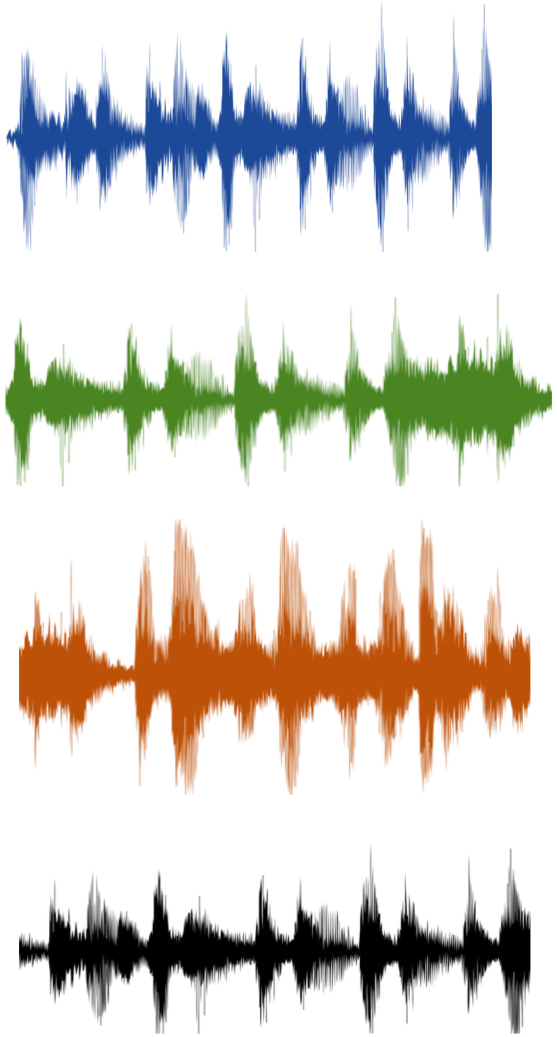
There are a lot of fascinating works that are not in this presentation, apologies in advance.

I chose the ones in which somehow I have been involved or have closely followed the research

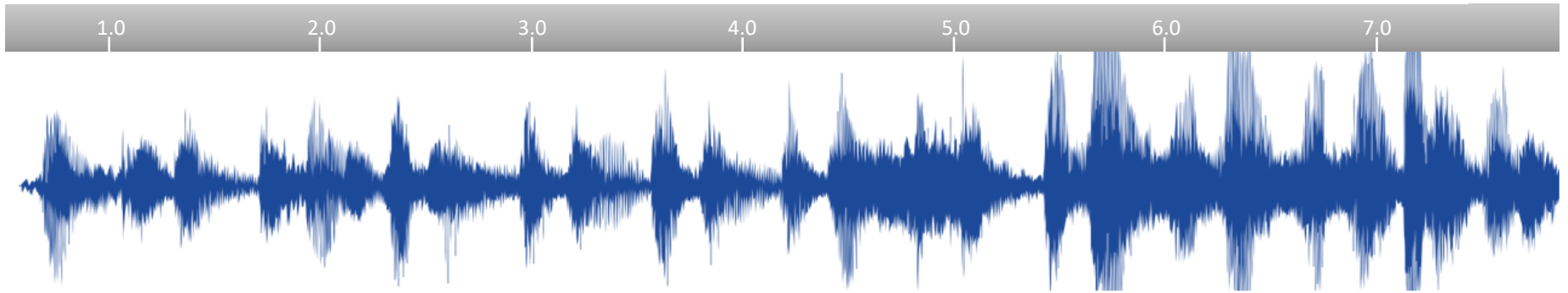
Technology is like a tornado

What data do we have?

What is the goal?



What is the goal?



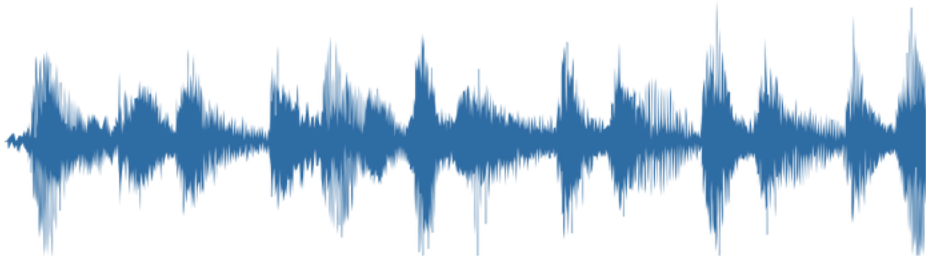
SPK1 SPK1 SPK2 SPK2 SPK3

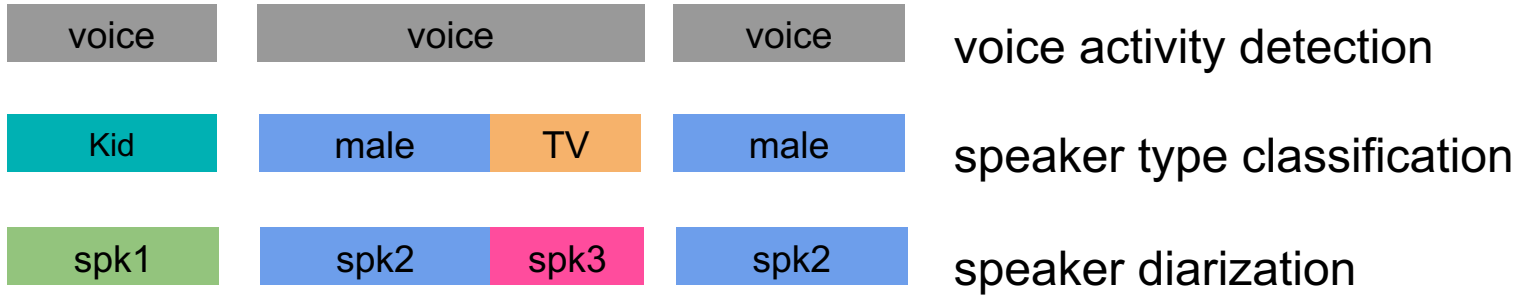
Spk1: If we want to address the next diarization problems..

Spk2: You should need to first breakdown those results.

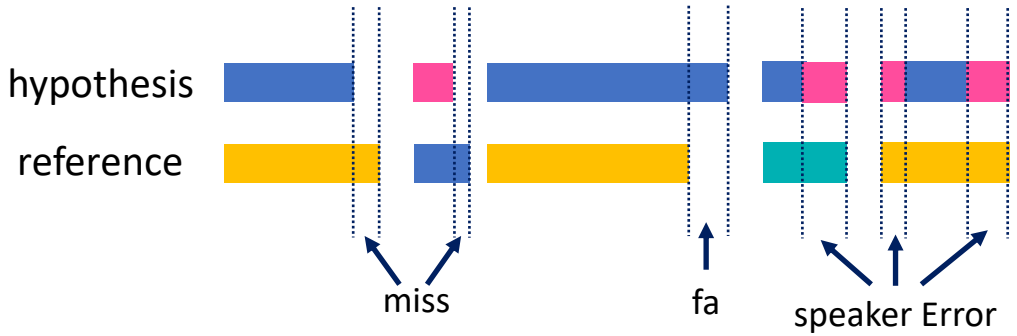
Spk3: Ok, I will put them in a Table or graph. I will also detail the algorithm…

Who spoke when?





How do we know how good we are?



Diarization Error Rate

$$\mathrm{DER} = \frac{\text{false alarm} \ + \ \text{missed detection} \ + \text{speaker error}}{\text{total}}$$

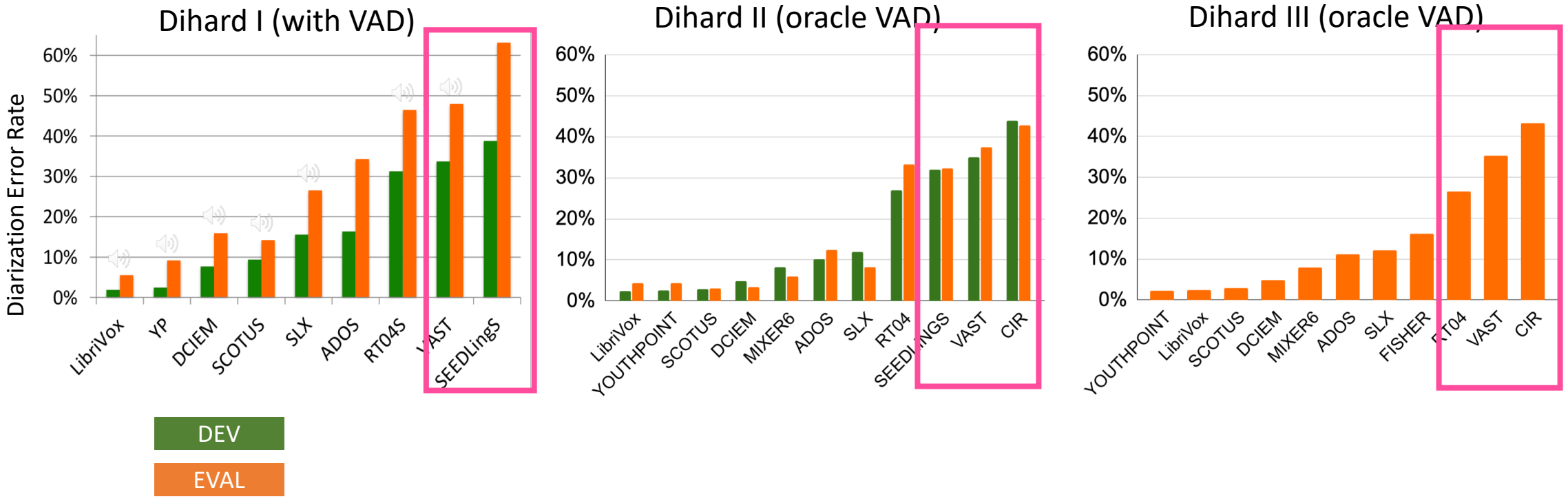
Jaccard Error Rate

For each reference speaker s_i, where $i = 1, \ldots, N$

$$\mathrm{JER}_{s_i} = \frac{\text{false alarm}_{s_i} \ + \ \text{missed detection}_{s_i}}{\text{total}_{s_i \cup s_{i_{\text{hypothesis}}}}}$$

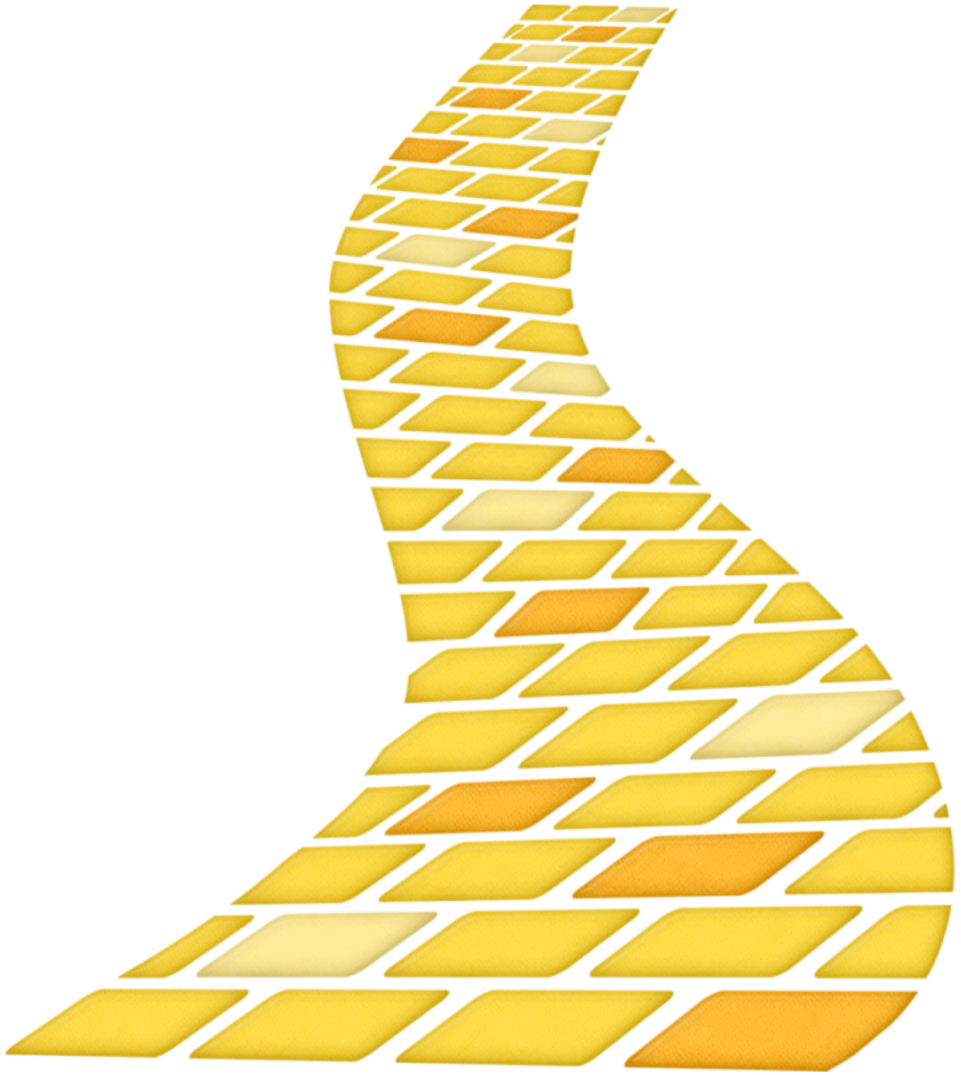
$$\mathrm{JER} = \frac{1}{N}\sum_{1}^{N} \mathrm{JER}_{s_i}$$

Where are we?

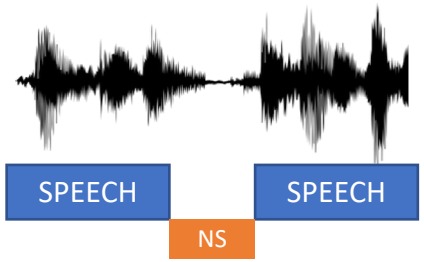


Greg Sell, et.al., 2018, Mireia Diez,et.al., 2019, Dihard III Hitachi-JHU, 2020

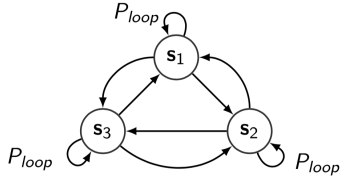
The Yellow Brick road



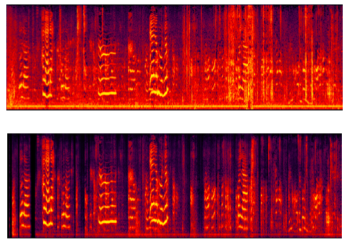
The Yellow Brick road

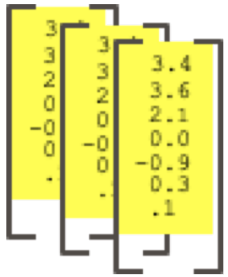


$$X = \begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & x_{1m} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & x_{nm} \end{bmatrix}$$

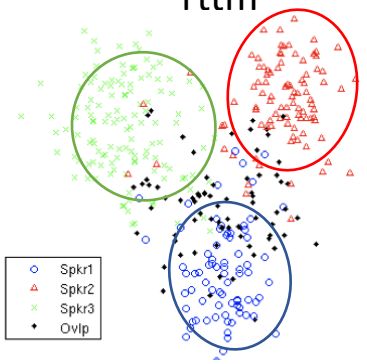


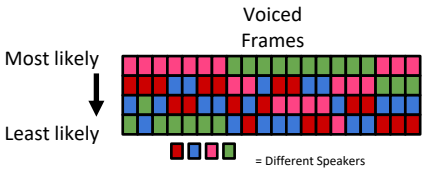
Enhan → VAD → Embedding → Scoring → Clustering → Reseg → Overlap assign





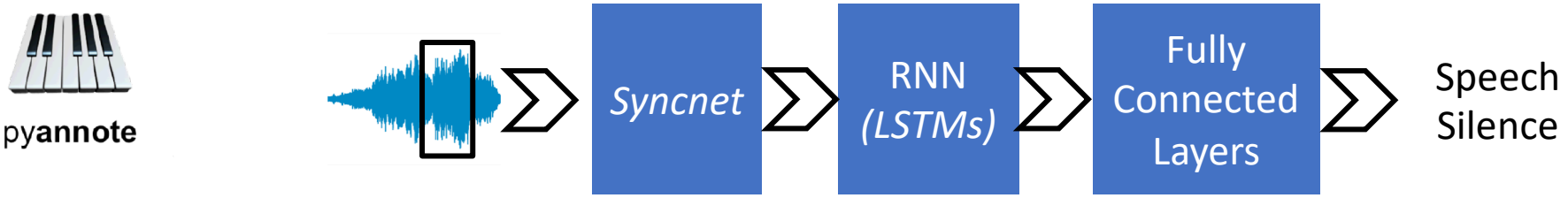
rttm rttm rttm



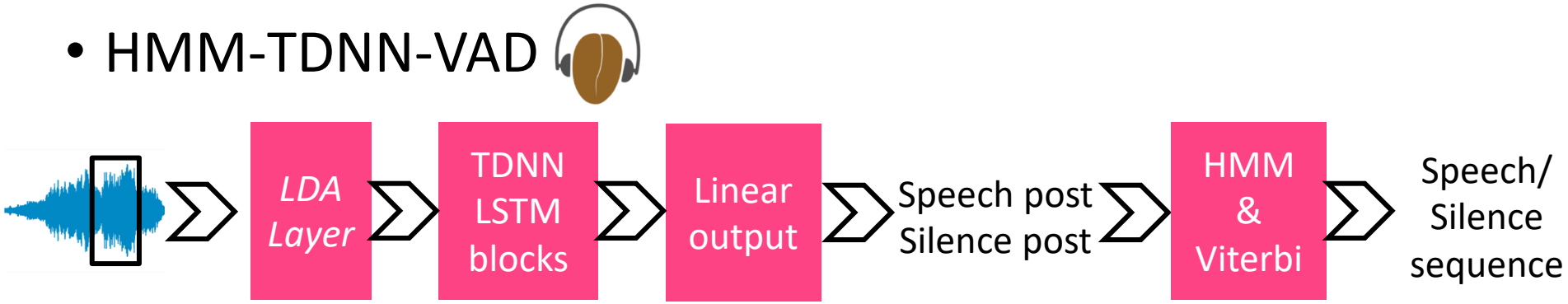


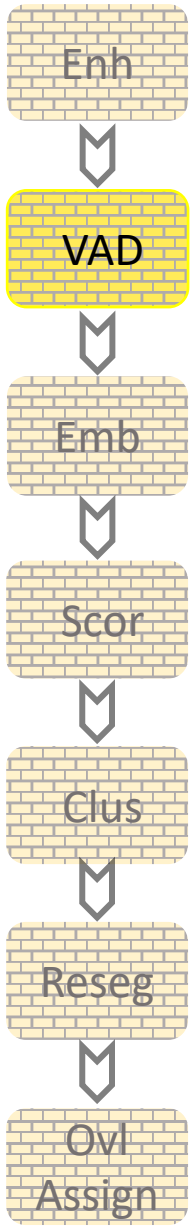
VAD

- Neural network VAD



- HMM-TDNN-VAD

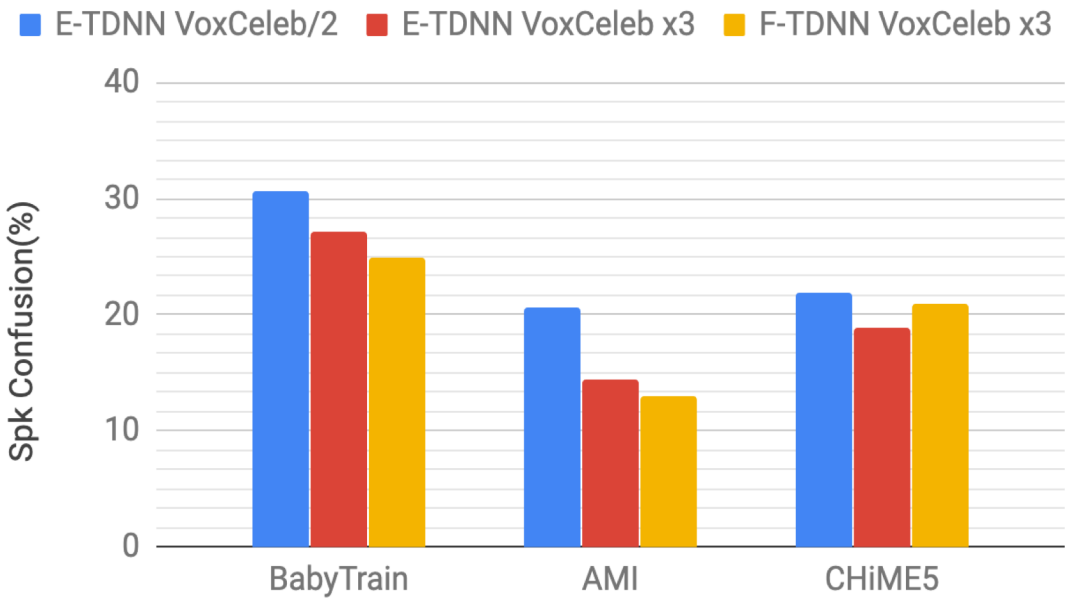




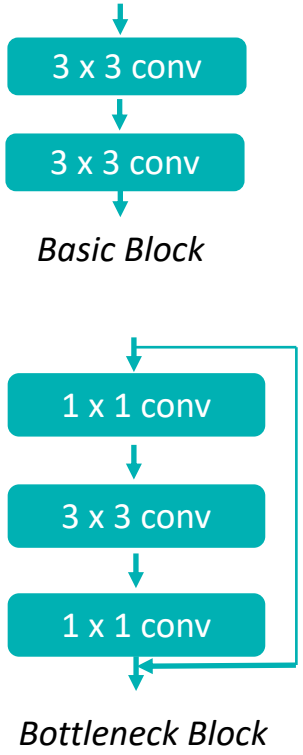
Marvin Lavechin, et.al., at JSALT 2019 workshop, ICASSP 2020, TDNN-VAD by JHU Speech team

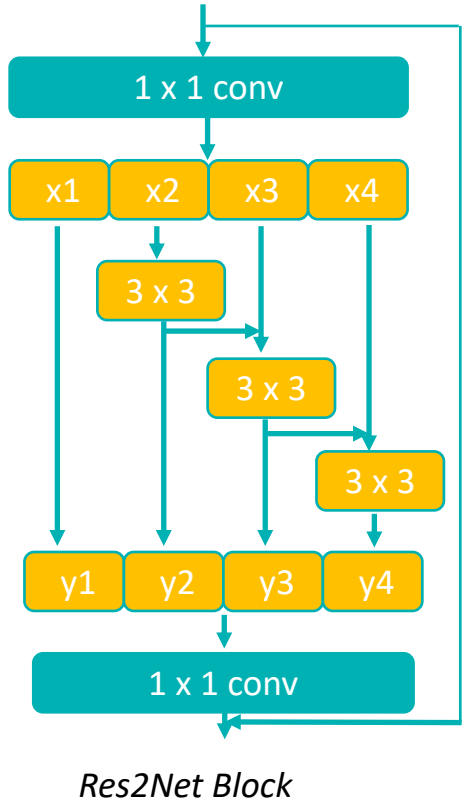
Embeddings

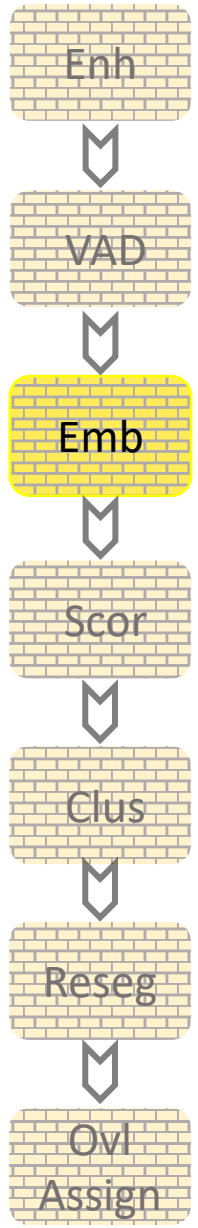
TDNN-xvector



Res2net







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

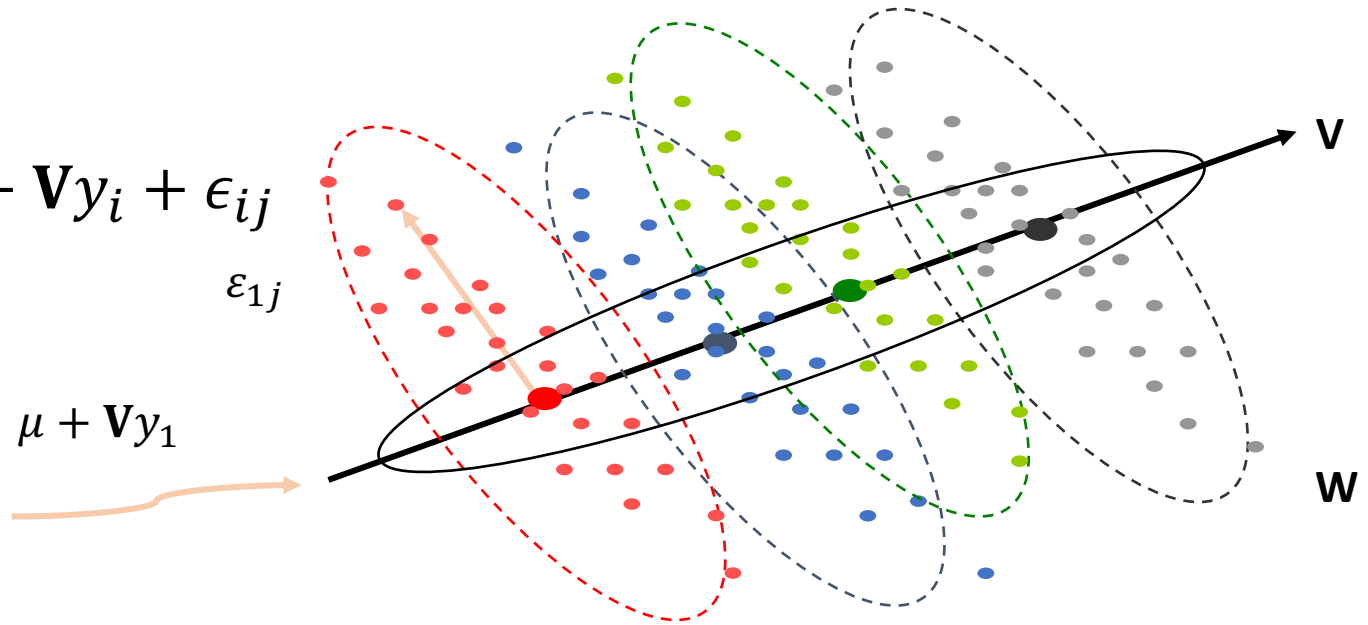
Shang-Hua Gao, et.al., Res2net: A new multi-scale backbone architecture

Xiong, Xiao, et. Al., Microsoft Speaker Diarization system for Voxceleb speaker recognition challenge 2020

Scoring

- PLDA

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

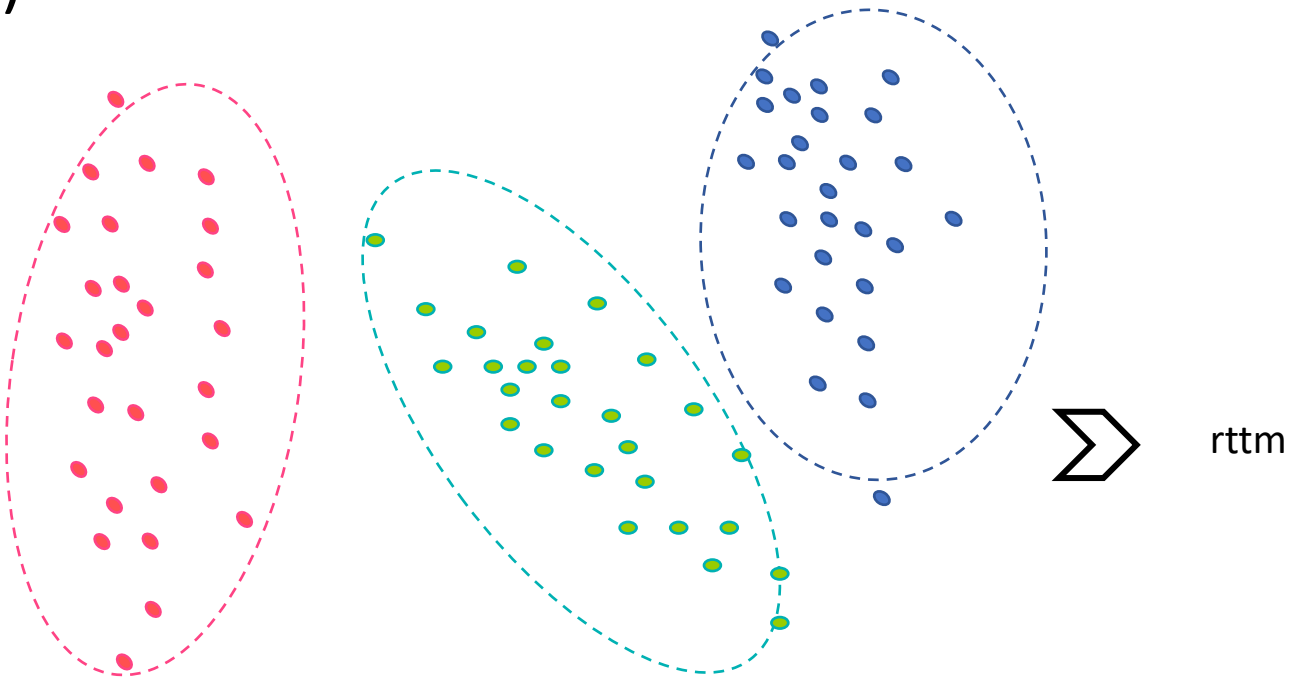


$$\text{LLR} = \log \frac{P(w_1, w_2|same)}{P(w_1, w_2|diff)} = w_1^T A w_2 + \mathrm{w_1^T B w_1} + \mathrm{w_2^T B w_2} + \mathrm{C^T w_1} + \mathrm{C^T w_2} + \mathrm{D}$$

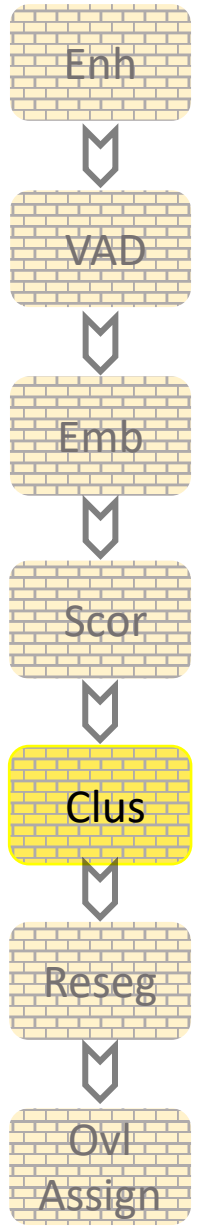
Clustering

- Clustering (AHC)

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



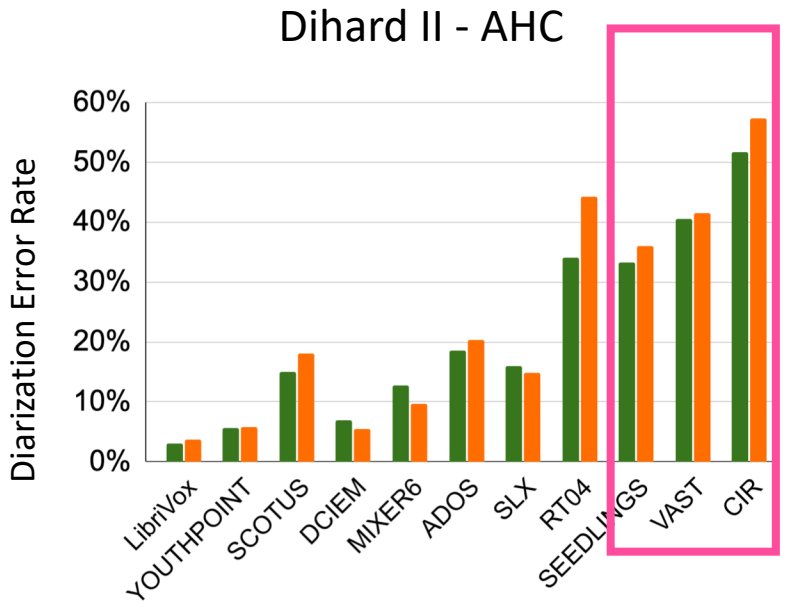
rttm

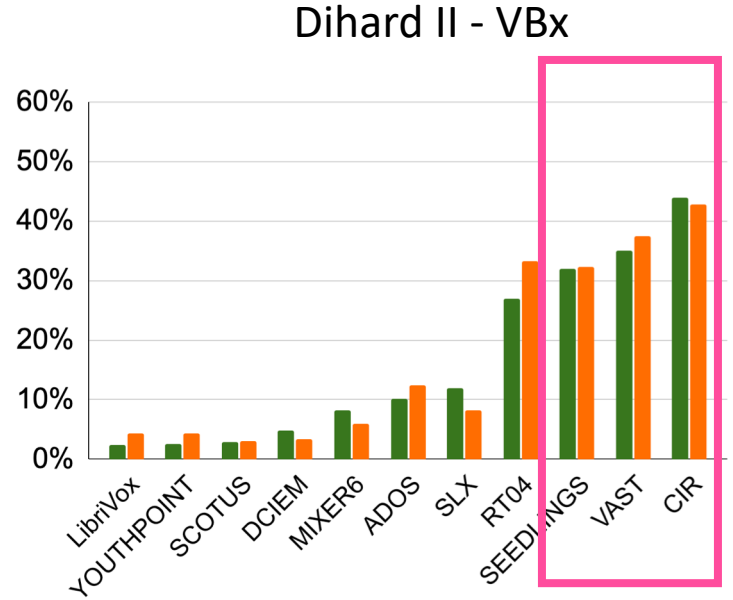


Mieria Diez, et.al., Speaker Diarization based on Bayesian HMM with Eigenvoice Priors 2018

Clustering

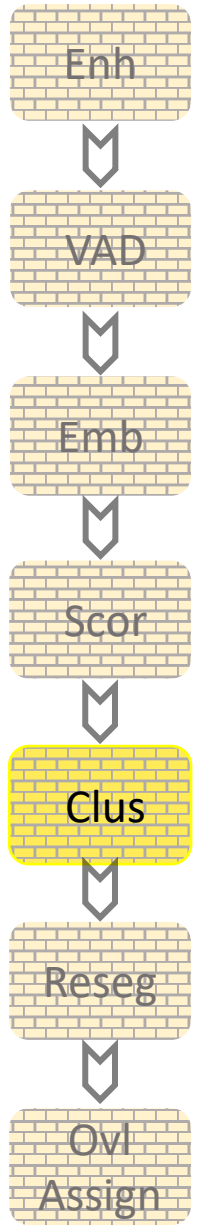
- VB-HMM Clustering





DEV

EVAL

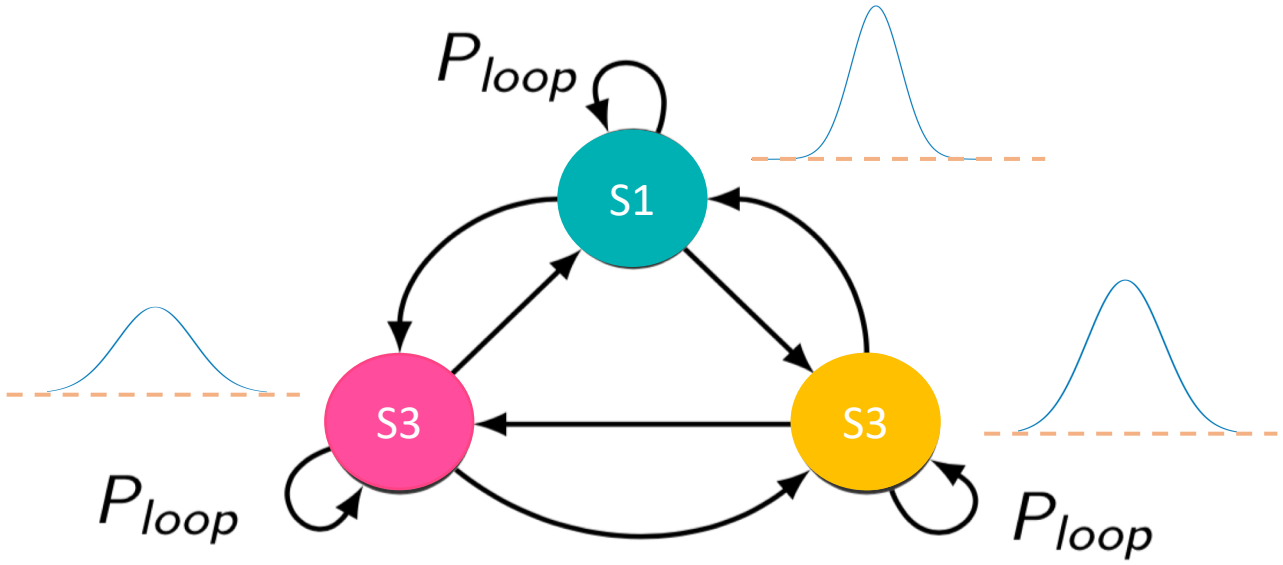


Mieria Diez, et.al., Speaker Diarization based on Bayesian HMM with Eigenvoice Priors 2018

Greg, Sell, et. Al., Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge

Clustering

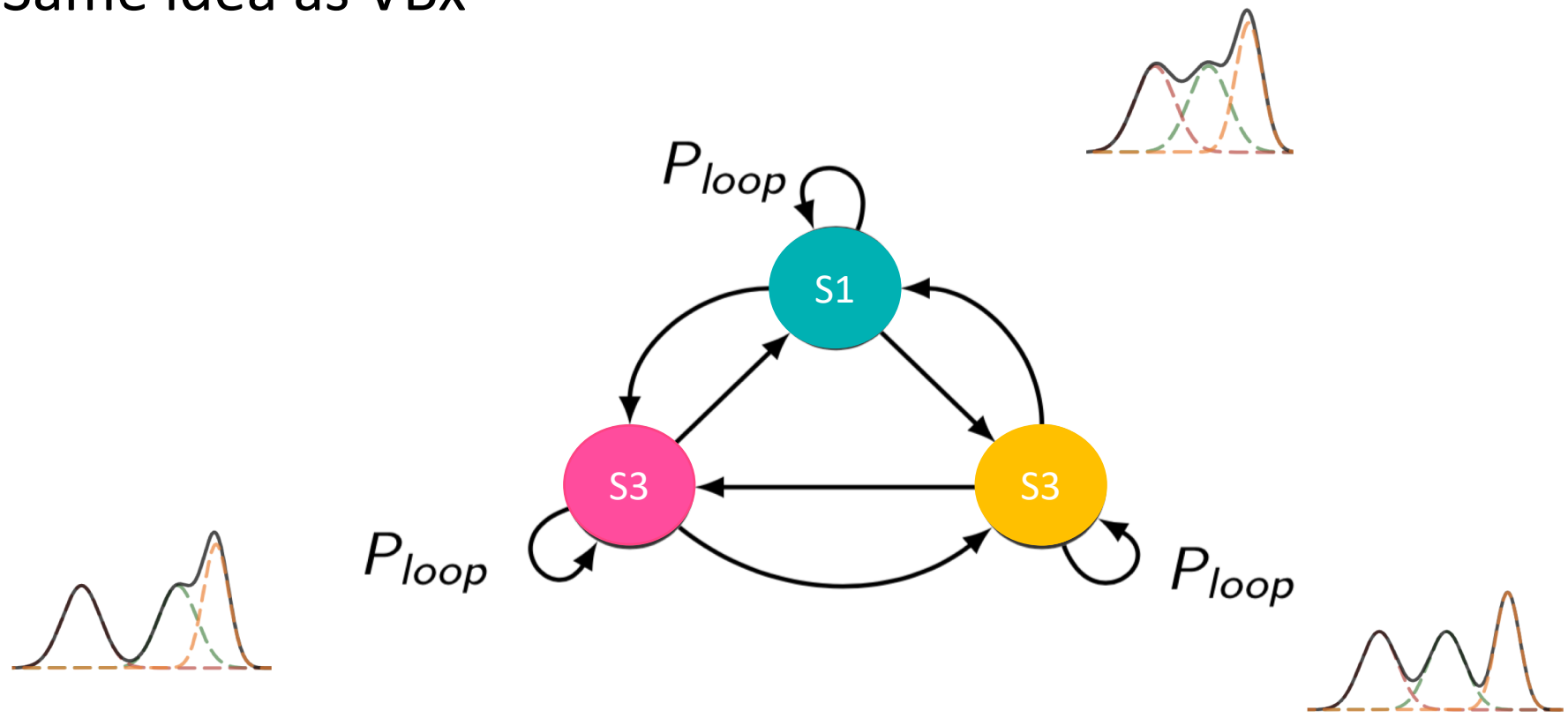
- VB-HMM Clustering - VBx

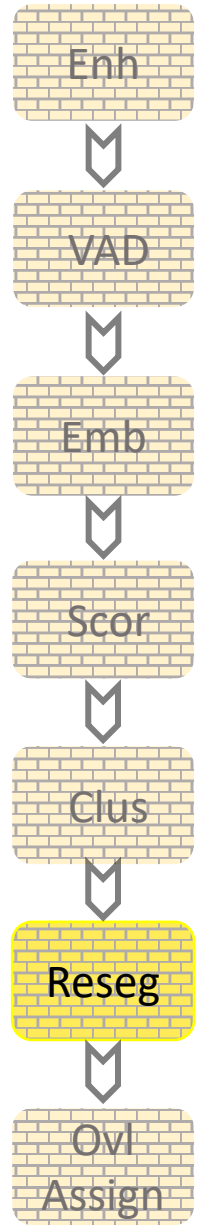


Mieria Diez, et.al., Speaker Diarization based on Bayesian HMM with Eigenvoice Priors 2018

Resegmentation

- Same idea as VBx

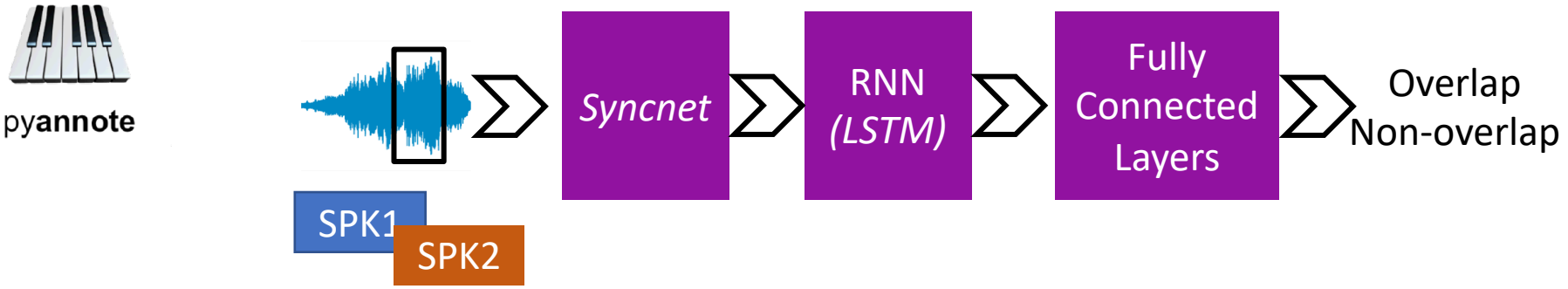




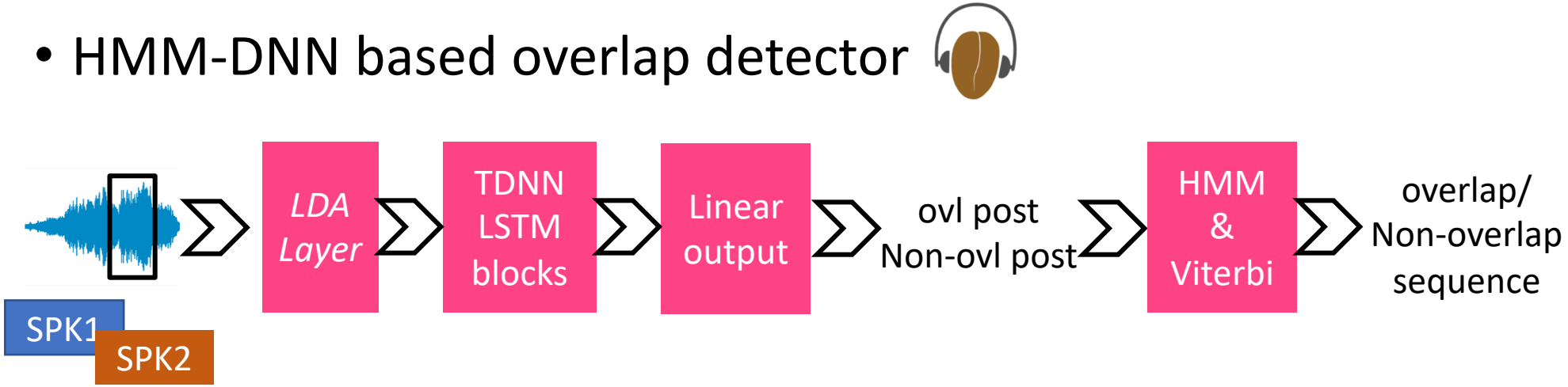
Latane Bullock, et.al., Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection, 2020
Desh Raj, et.al., Multi-class Spectral Clustering with Overlaps for Speaker Diarization, 2020

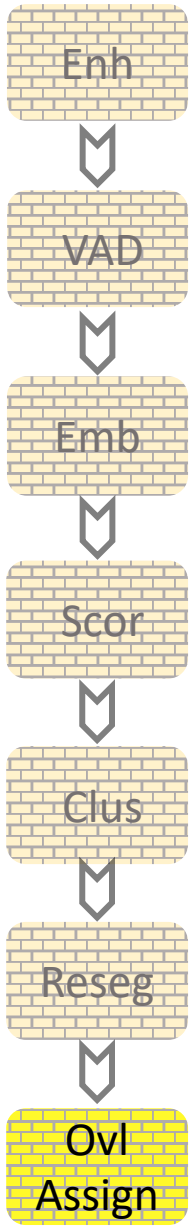
Overlap Assignment

- Neural network Overlap detector



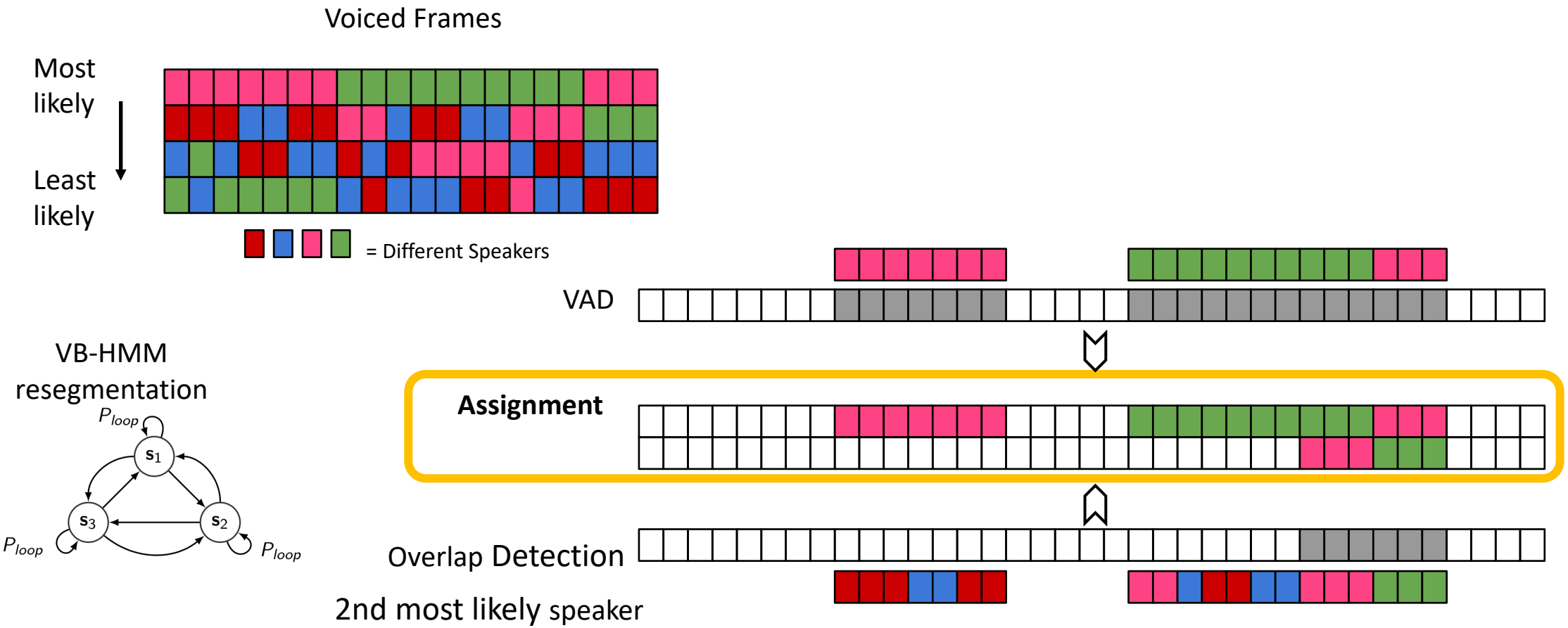
- HMM-DNN based overlap detector

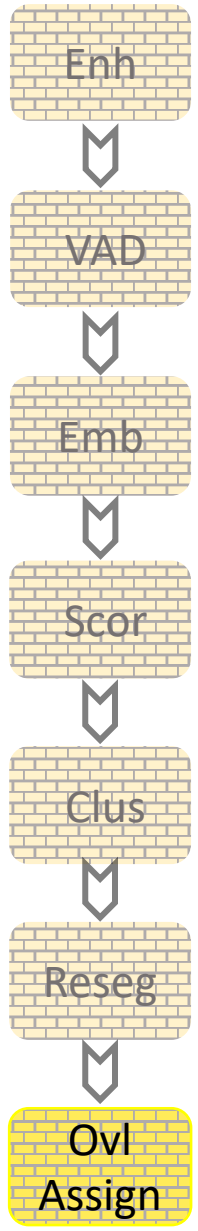




Latane Bullock, et.al., Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection, 2020
Desh Raj, et.al., Multi-class Spectral Clustering with Overlaps for Speaker Diarization, 2020

Overlap Assignment



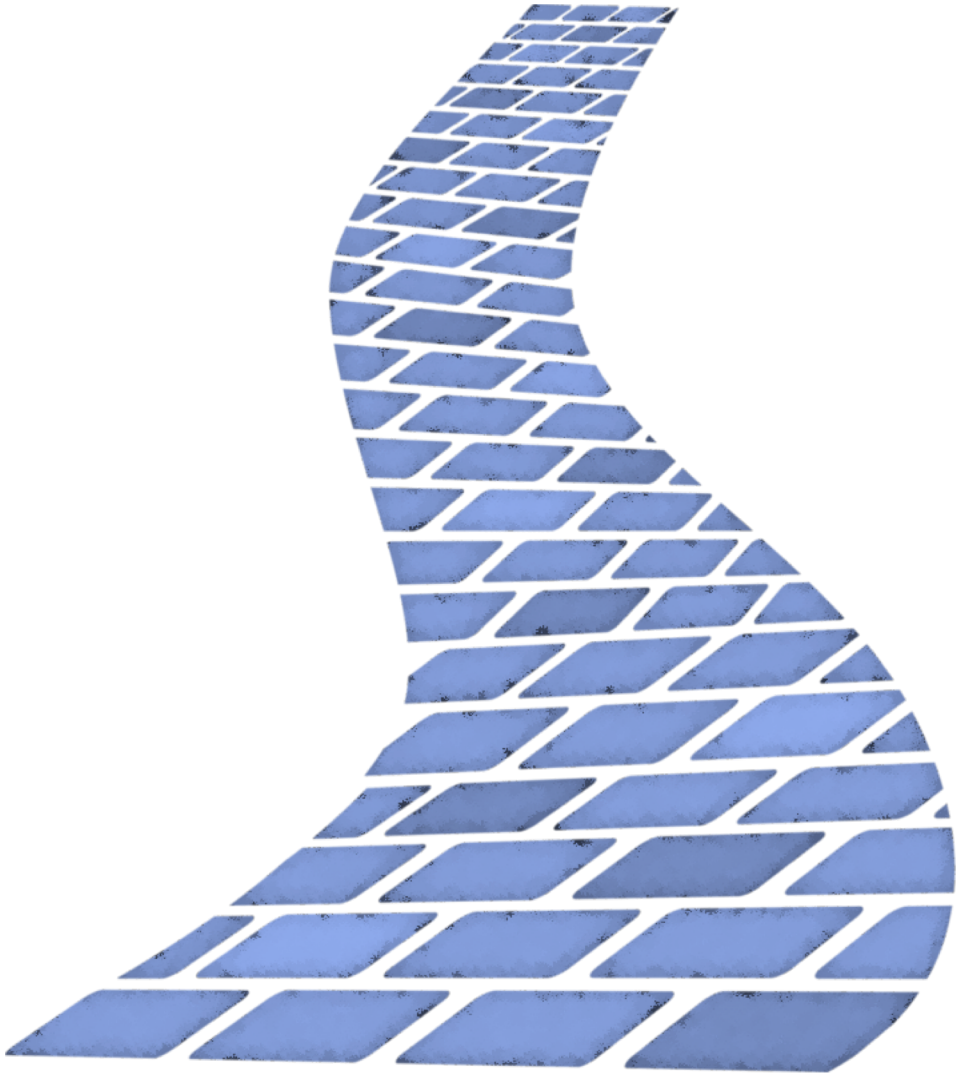


Latane Bullock, et.al., Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection, 2020

Yellow brick road issues

- Our yellow brick road has some issues:
  - We are still trying handle overlapping speakers
  - The system is not designed to minimize the diarization error but optimizes every module separately.

So we look for solutions!



What about a fully neural path?

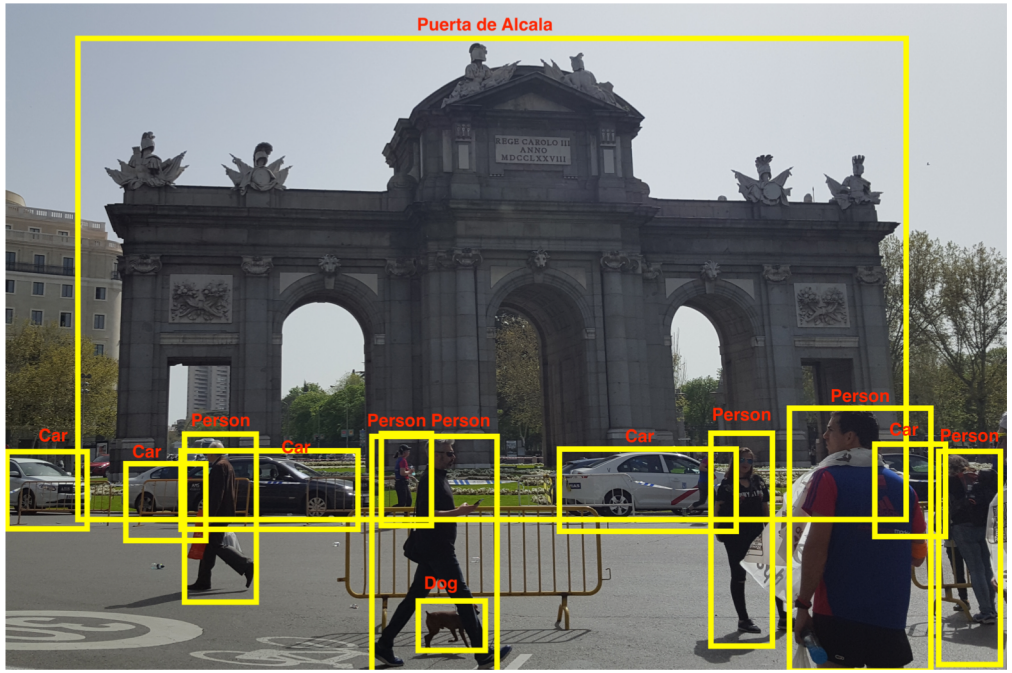
A single DNN with a single optimization stage where the input is *speech* (or speech features) and the output are the *labels* (or speaker posteriors)?

Something in between?

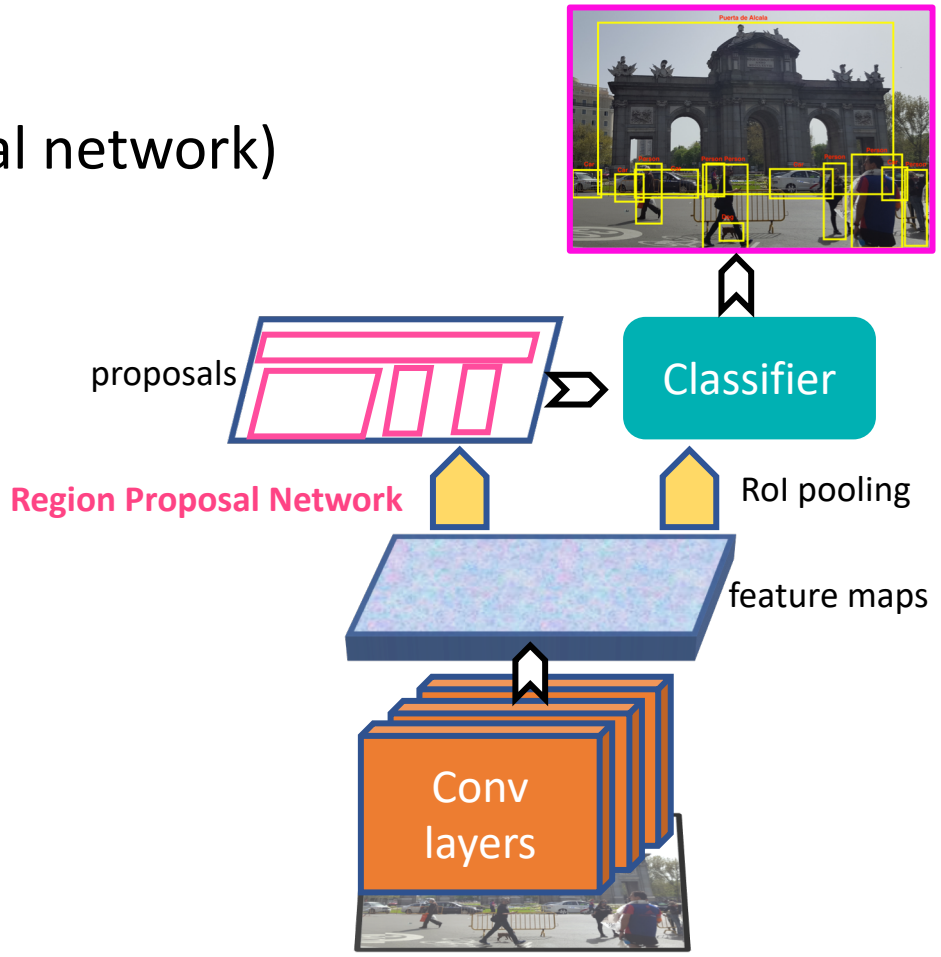
- Too risky, let’s first try something in between
- Trial towards end-to end

Region Proposal Network

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



Faster RCNN

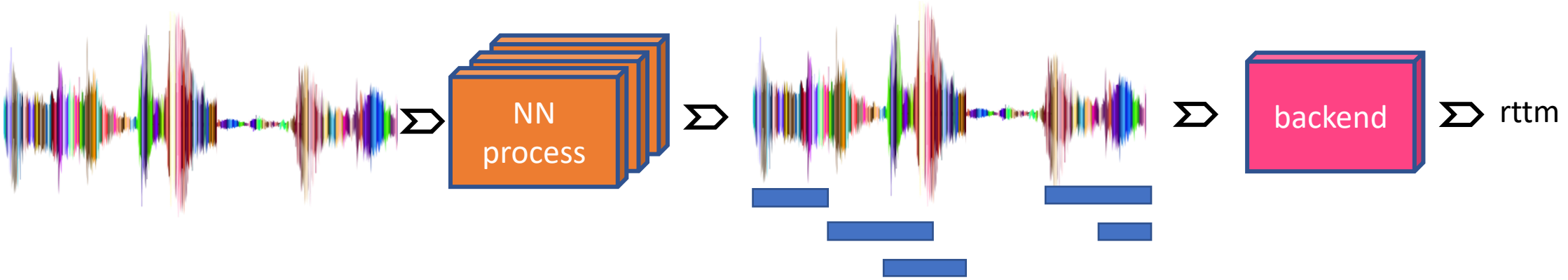


Zili Huang, et.al., Speaker Diarization with Region Proposal Network, 2020

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

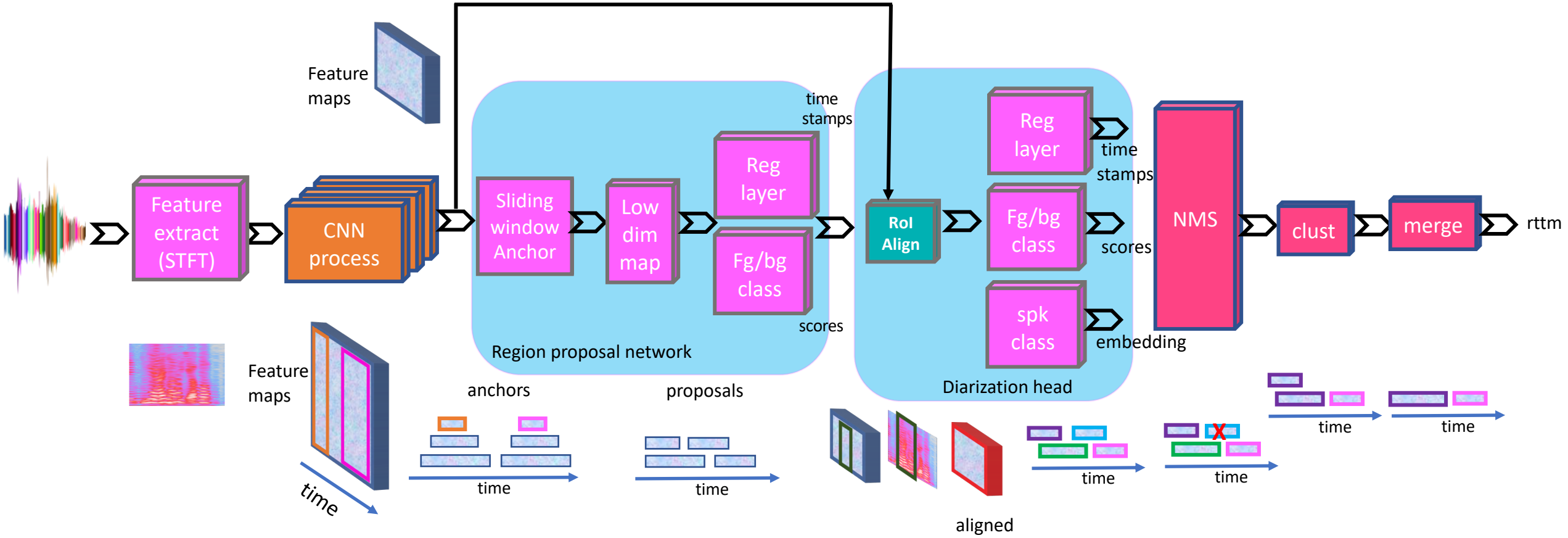
Region Proposal Network for Speaker Diarization

- Same idea but how to do it with speech ☺



Zili Huang, et.al., Speaker Diarization with Region Proposal Network, 2020

Region Proposal Network for Speaker Diarization



Zili Huang, et.al., Speaker Diarization with Region Proposal Network, 2020

Region Proposal Network for Speaker Diarization

- How to control this pipeline?

$$L = L_{\mathrm{RPN}_{\mathrm{cls}}} + L_{\mathrm{RPN}_{\mathrm{reg}}} + L_{\mathrm{RCNN}_{\mathrm{cls}}} + L_{\mathrm{RCNN}_{\mathrm{reg}}} + \alpha L_{\mathrm{spk}_{\mathrm{cls}}}$$

  - L_{cls} (classification loss): classifies whether a speech segment is foreground or background
  - L_{reg} (regression loss): smooth L1 loss to regress the center point and the length of the speech segments
  - $L_{\mathrm{spk}_{\mathrm{cls}}}$ (speaker classification loss): classifies the speaker identity of the speech segments

System	DER (%)	JER (%)
DIHARD baseline	40.86	66.60
DIHARD best VBx	27.11	49.07
RPNSD #oracle num spk	33.12	49.69

Zili Huang, et.al., Speaker Diarization with Region Proposal Network, 2020

Now that we feel more comfortable…

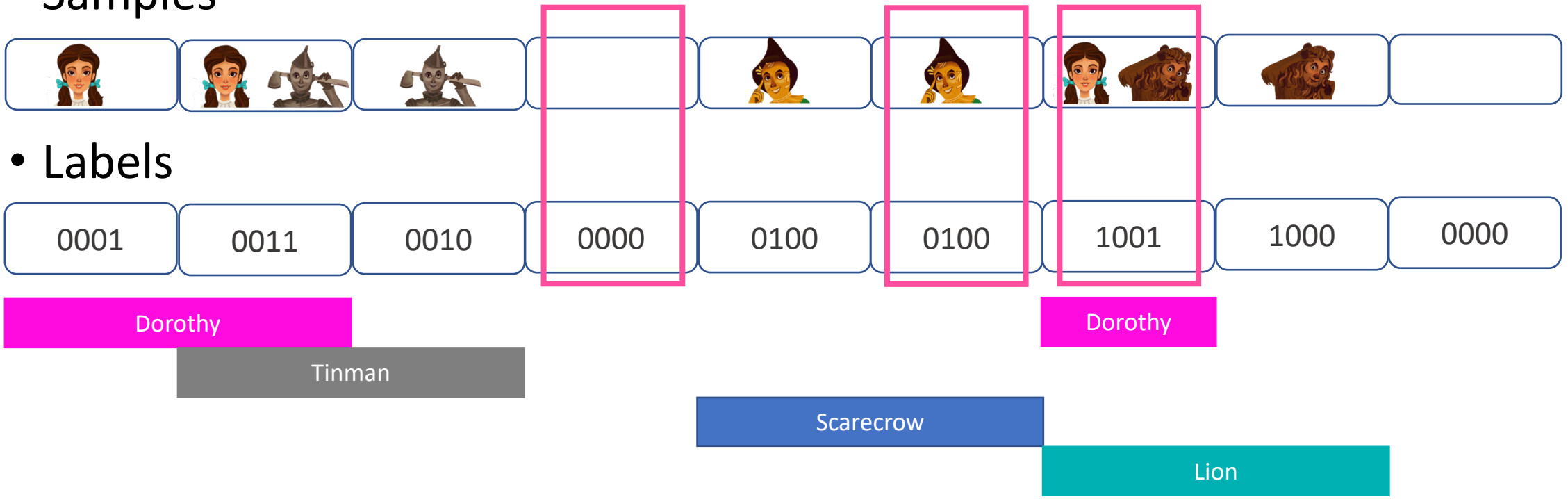
Let’s take a look at the End-to-End approaches

Neural diarization

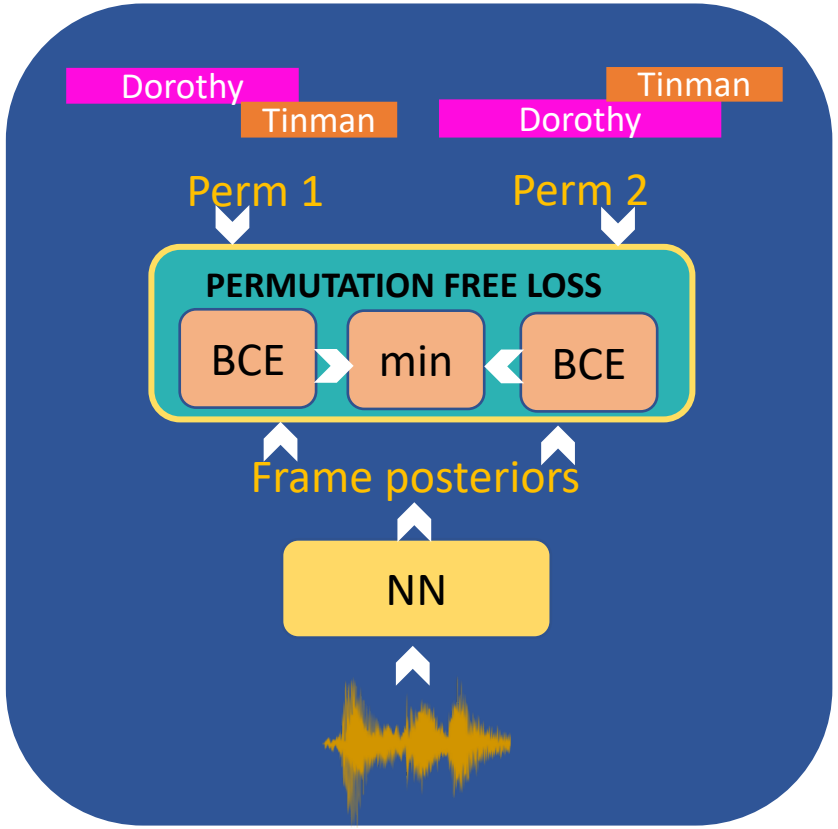
Yusuke Fujita, et.al., End-to-End Neural Diarization: Reformulating Speaker Diarization as Simple Multi-label Classification,2019

Diarization as a multi-label classification

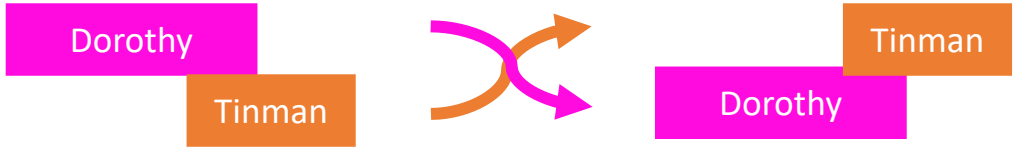
- Samples
- Labels



EEND

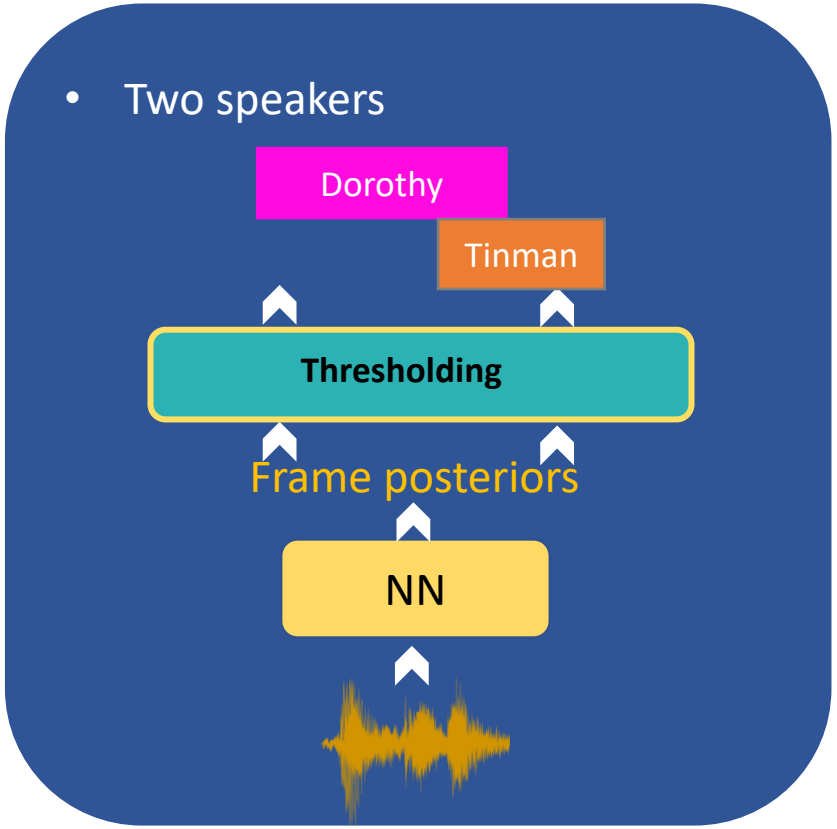


- End-to-end neural diarization
- Single network, supervised
- Two speaker case, proof of concept
- Handles overlapping speech!
- Training uses permutation invariant training (PIT) to prevent the labeling ambiguity.



Yusuke Fujita, et.al., End-to-End Neural Diarization: Reformulating Speaker Diarization as Simple Multi-label Classification,2019

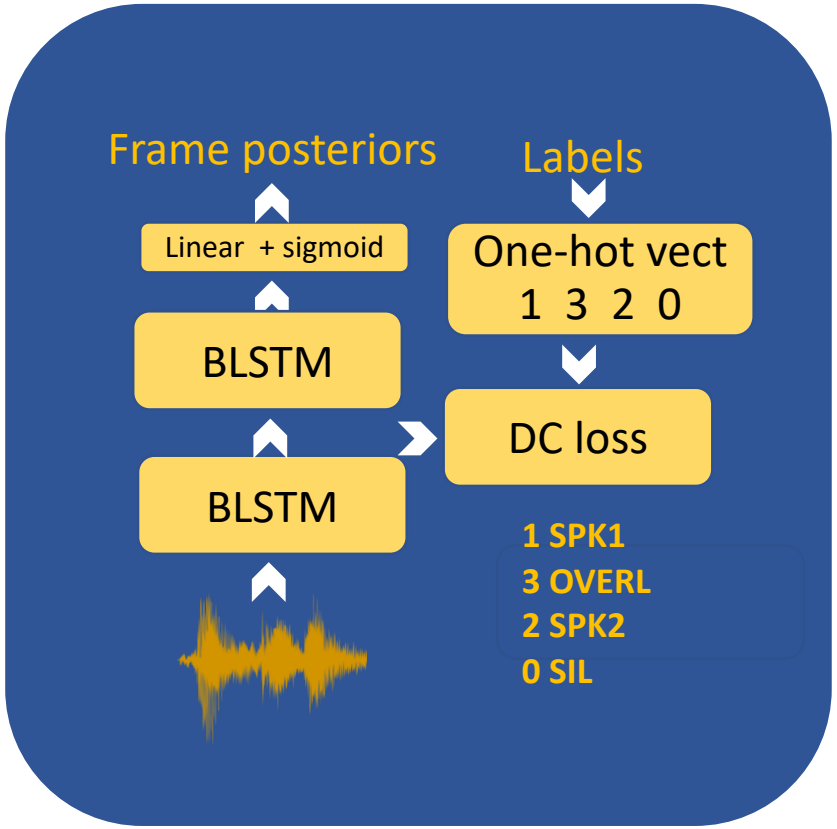
EEND



- During test
  - Frame posteriors
  - Threshold
  - Speaker labels

Yusuke Fujita, et.al., End-to-End Neural Diarization: Reformulating Speaker Diarization as Simple Multi-label Classification,2019

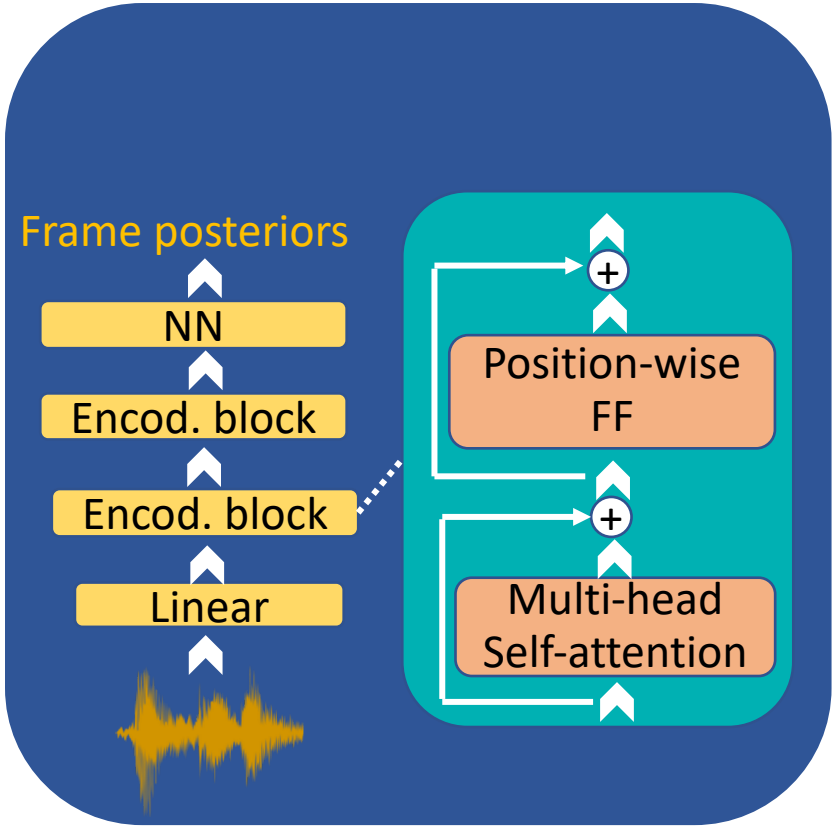
EEND



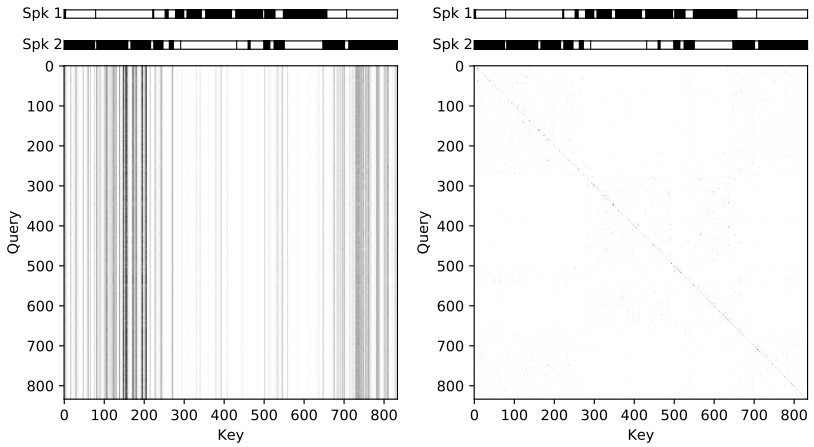
- BLSTM
- Embeddings from lower layers
  - Speaker training criterion on middle layer activations
- Deep Clustering to partition the embedding into:
  - Speaker dependent-clusters
  - Overlap
  - silence

Yusuke Fujita, et.al., End-to-End Neural Diarization: Reformulating Speaker Diarization as Simple Multi-label Classification,2019

EEND-SA

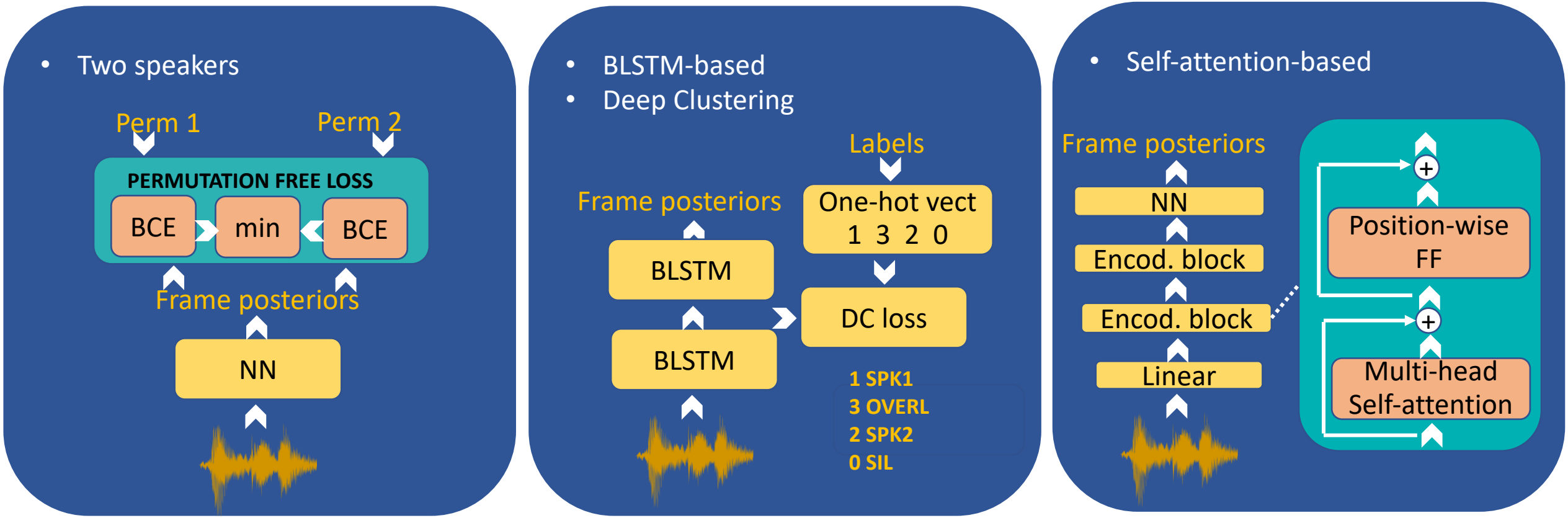


- BLSTM captures local temporal dynamics
- Self attention captures long context
  - The heads capture different characteristics



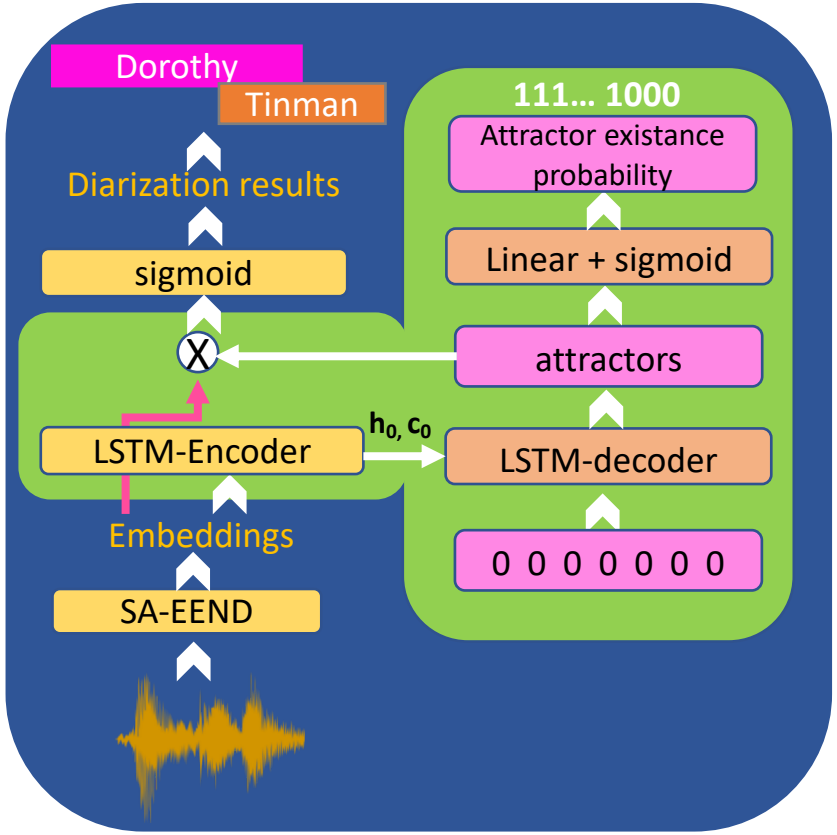
Yusuke Fujita, et.al., End-to-End Neural Diarization: Reformulating Speaker Diarization as Simple Multi-label Classification,2019

EEND evolution



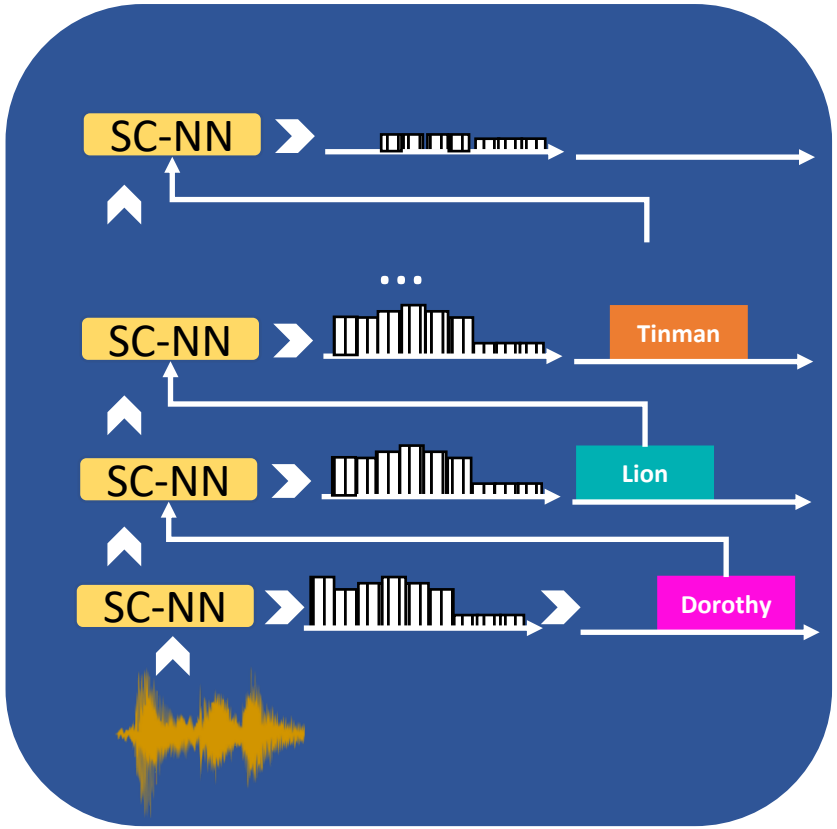
Yusuke Fujita, et.al., End-to-End Neural Diarization: Reformulating Speaker Diarization as Simple Multi-label Classification, 2019

EEND-EDA



- Encoder decoder attractor for variable number of speakers.
- Starting point is the SA-EEND and the embeddings.
- LSTM-decoder produce attractors
- Dot product between attractors and embeddings produce the diarization results.

SC-EEND



- Speaker wise conditional EEND
- Deals with variable number of speakers
- Fully conditional model
- Decode speaker-wise sequentially, conditioned on previous speech activities
- Uses teacher–forcing in the training with a modification that takes of the appropriate permutation.

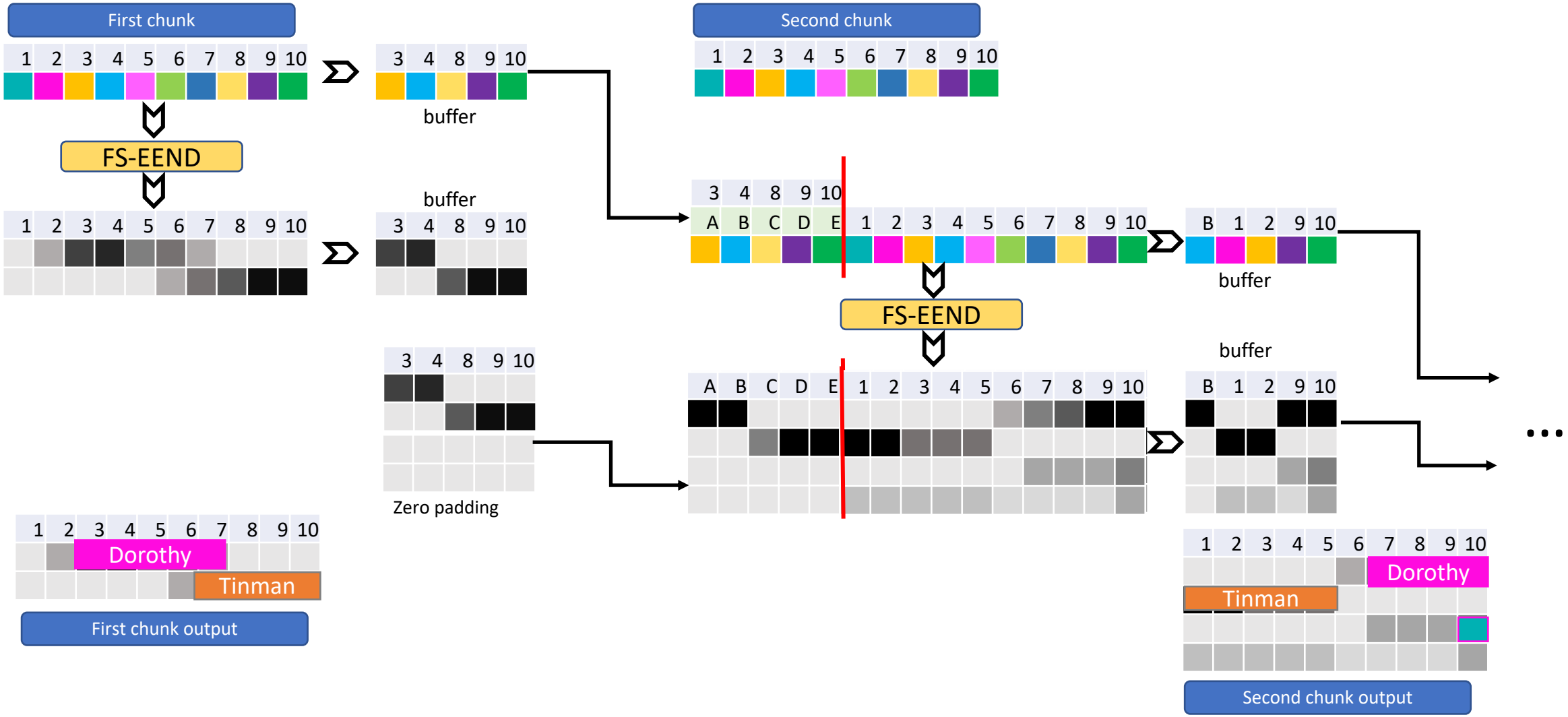
Yusuke Fujita, et.al., Neural Speaker Diarization with Speaker-Wise Chain Rule,2020

Some results

DIHARD III (track 1- oracle SAD)

	DEV		EVAL	
System	full	core	full	core
Baseline	19.41	20.25	19.25	20.65
TDNN+VBx+Ovlassign	13.87	14.88	15.65	18.20
EEND-EDA	**12.92**	13.95	**13.95**	**17.28**
SC-EEND	13.13	**13.13**	15.16	19.14

Online diarization EEND



Yawen Xue, et.al., Online End-to-End Neural Diarization with Speaker Tracing Buffer, 2021

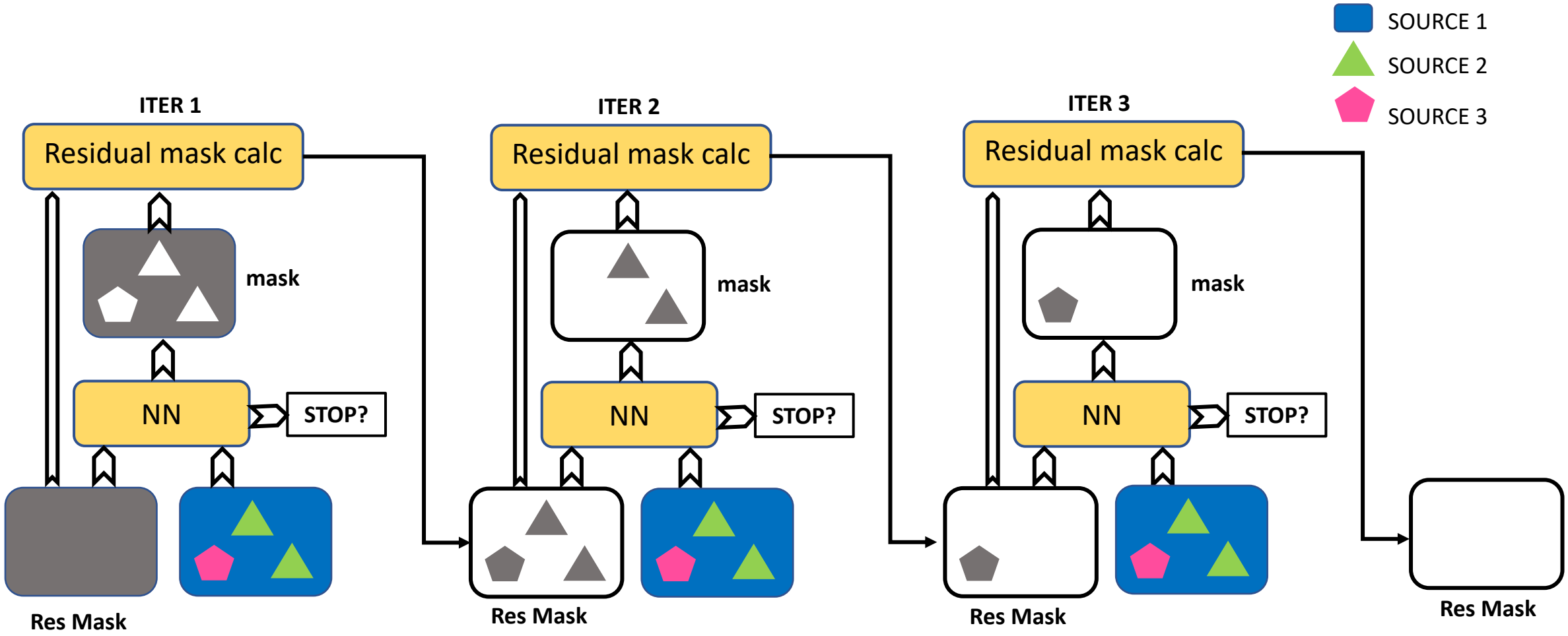
More results

DIHARD II (track 1- oracle SAD)

System	DER (%)
Baseline (offline)	26.0
UIS-RNN-SML*	27.3
EEND-EDA w/STB	**25.9**
SC-EEND w/STB	**25.3**

*Enrico Fini, et.al., Supervised online diarization with sample mean loss for multi-domain data, 2020
Yawen Xue, et.al., Online End-to-End Neural Diarization with Speaker Tracing Buffer with speaker-tracing buffer, 2021

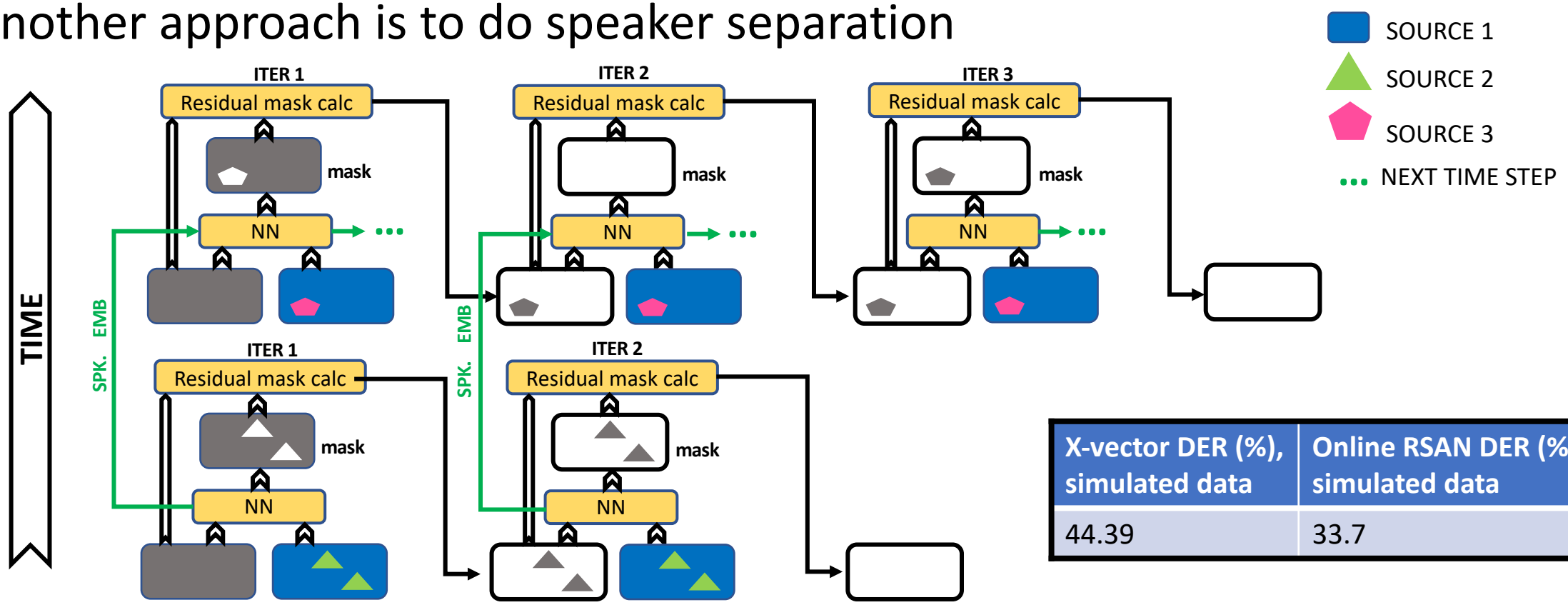
RSAN (Recurrent selective attention network)



Keisuke Kinoshita, et.al., Tackling Real Noisy Reverberant Meetings with All-Neural Source Separation, Counting, and Diarization System, 2020

RSAN (Recurrent selective attention network)

- Another approach is to do speaker separation

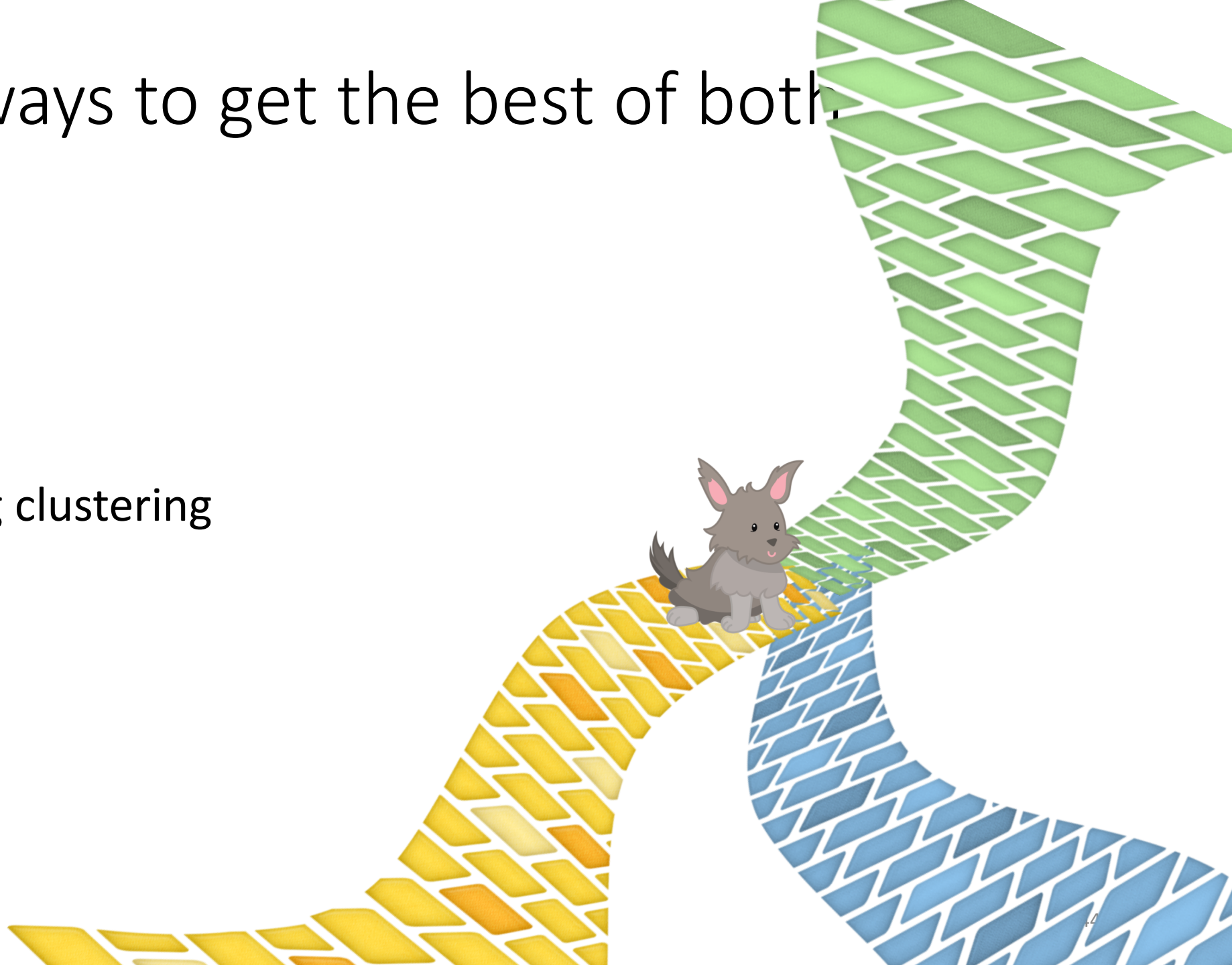


X-vector DER (%), simulated data	Online RSAN DER (%), simulated data
44.39	33.7

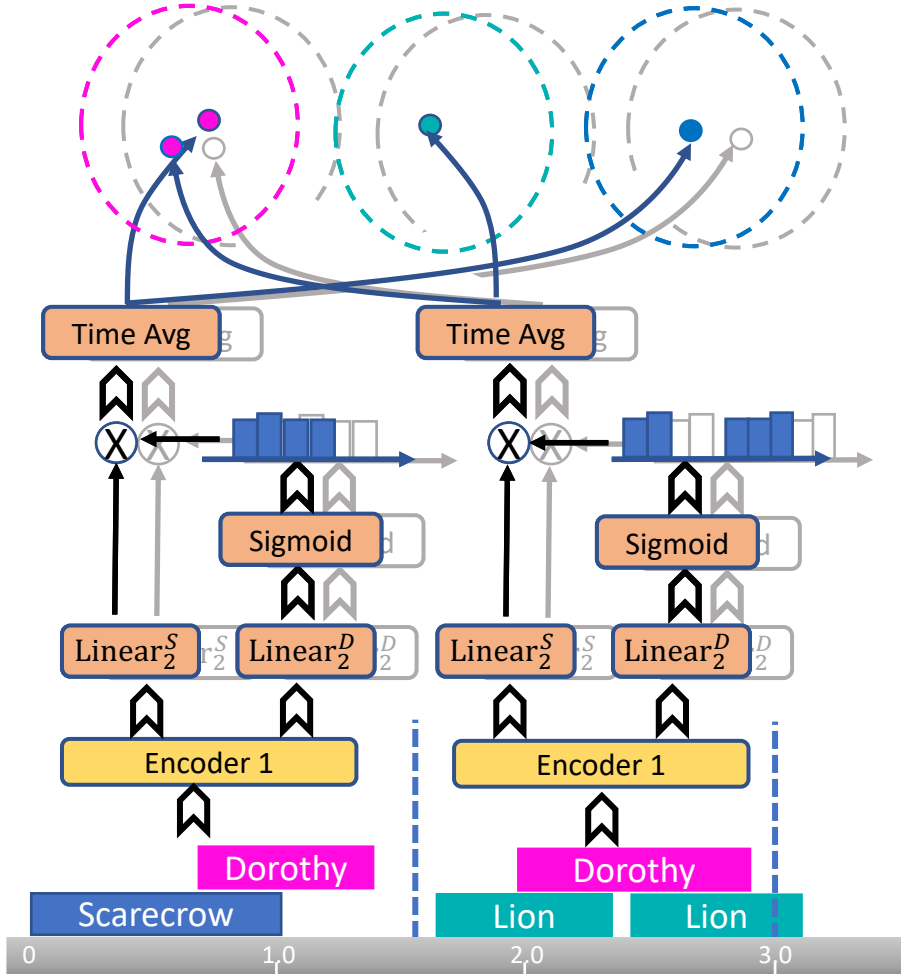
Keisuke Kinoshita, et.al., Tackling Real Noisy Reverberant Meetings with All-Neural Source Separation, Counting, and Diarization System, 2020

Are there ways to get the best of both worlds?

- Part End-to-end
- Part embedding clustering



Other approaches



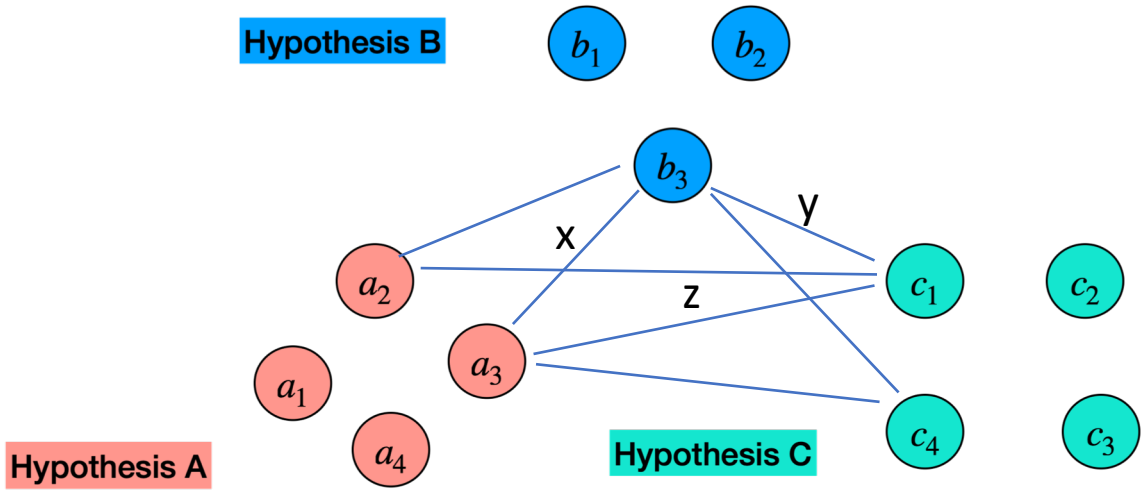
- EEND-vector clustering
- Hybrid system for overlapped speech, long recordings and different number of speakers

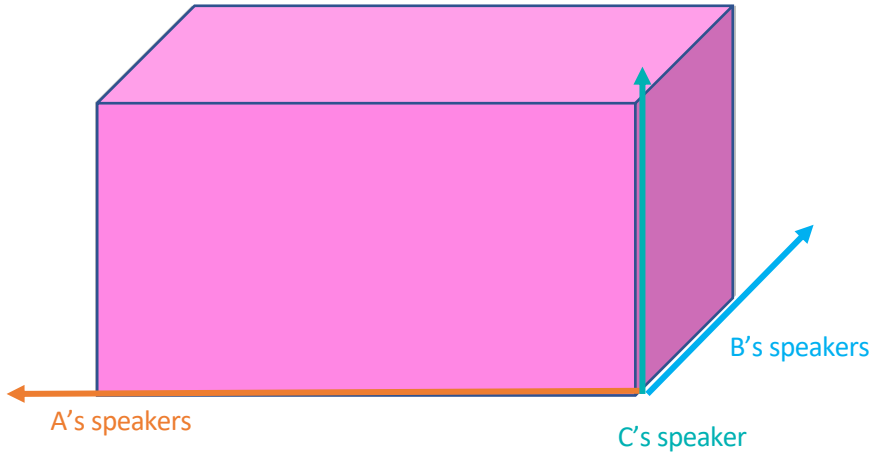
	Test duration (min)			
System	3	5	10	20
EEND	7.9	8.8	9.2	N/A
Propos+chunking+clust	9.1	8.2	7.9	7.7

Keisuke Kinoshita, et.al., Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds, 2020

DOVER-Lap

- Dover with overlap handling
- Two stages:
  - Label mapping

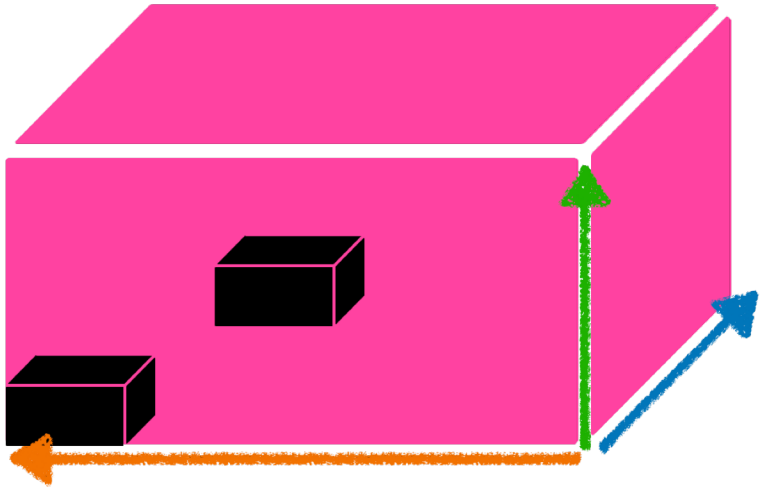


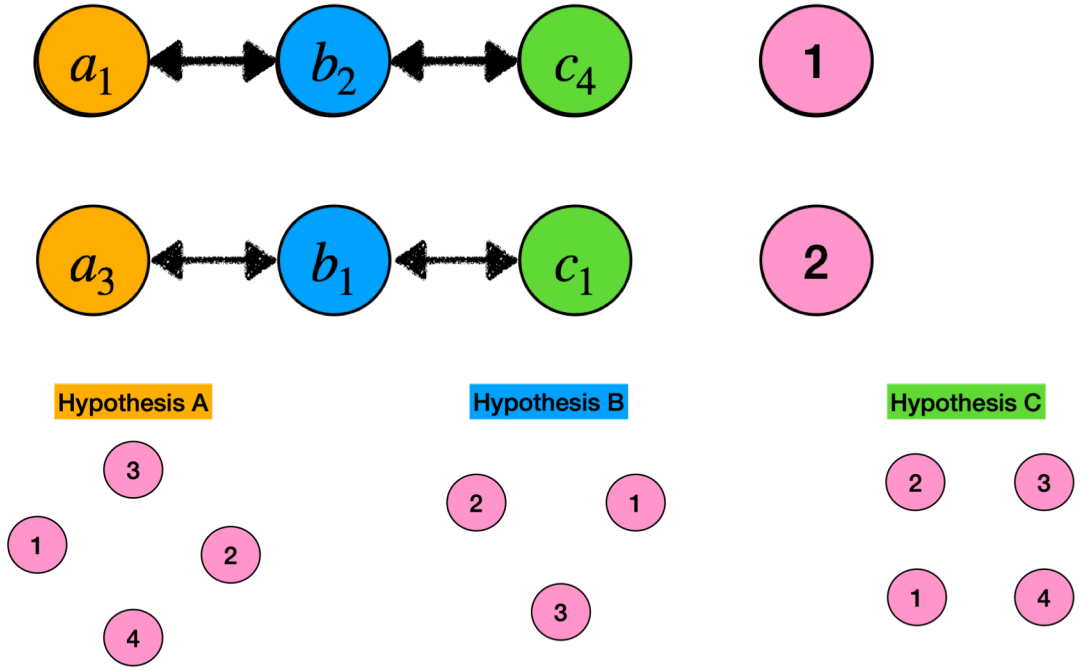


Desh Raj, et.al., DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs, 2021

DOVER-Lap

- Tuple with lowest cost and assign the same label

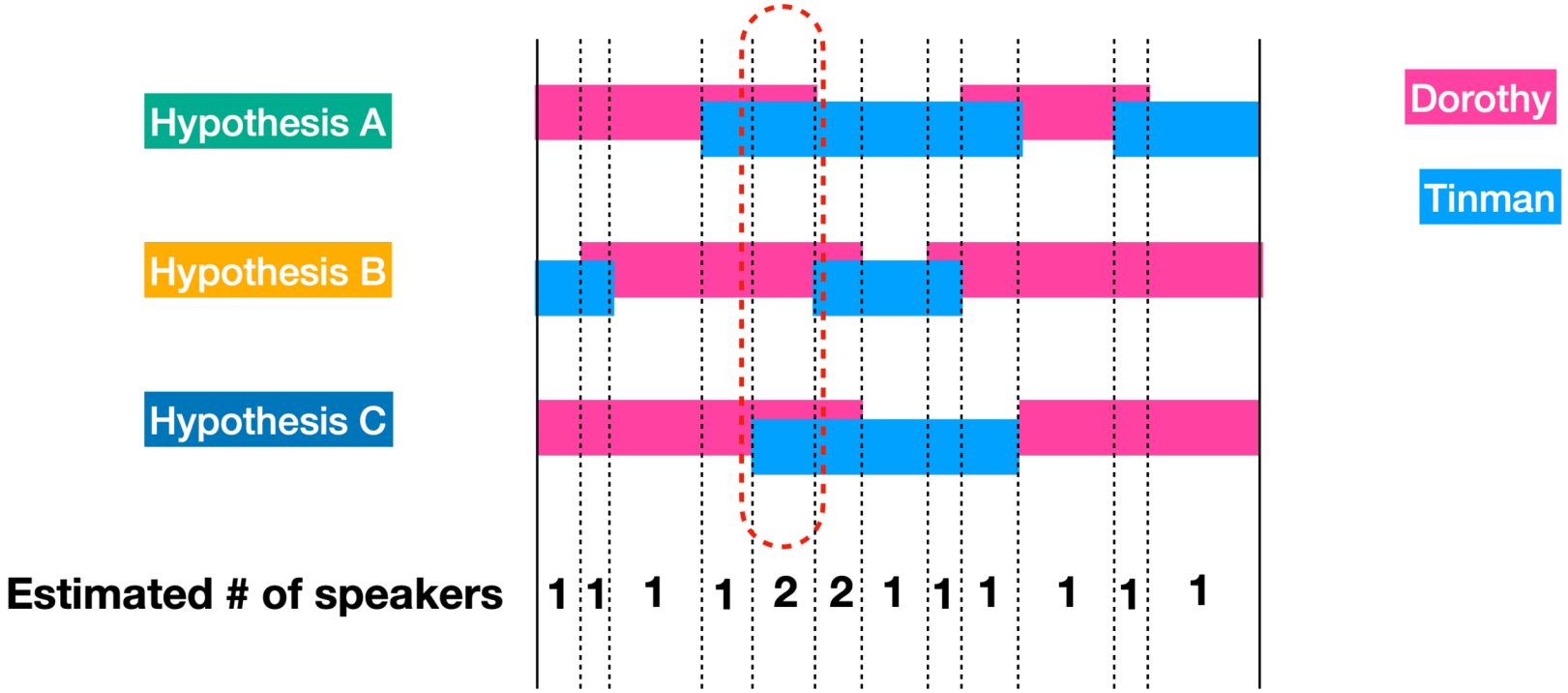




Desh Raj, et.al., DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs, 2021

Dover-Lap

- Label voting

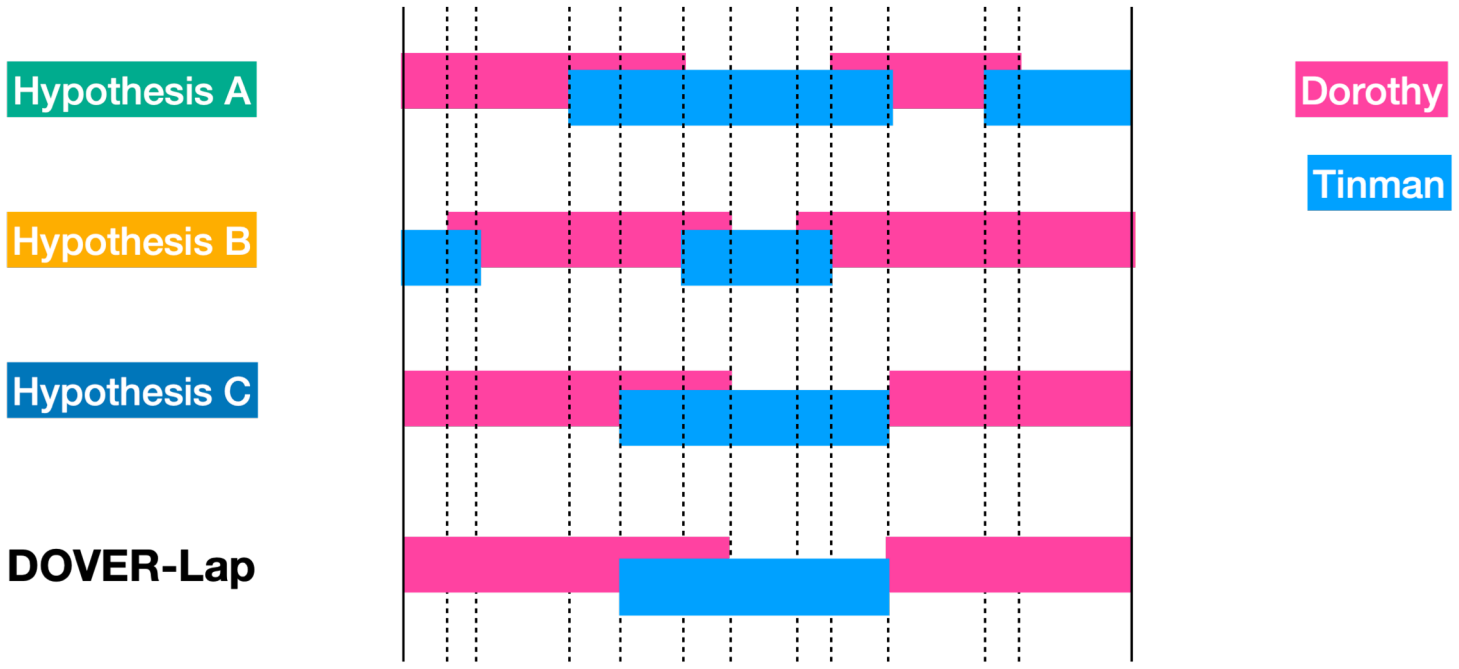


 # speakers = weighted mean of # speakers in hypotheses

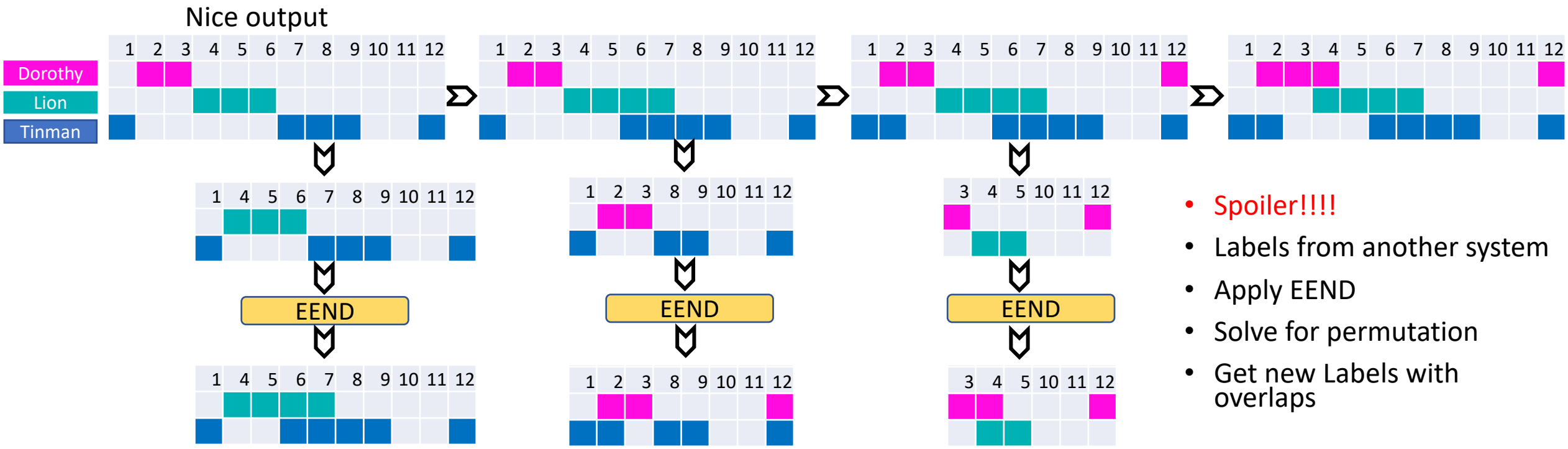
Weights -> obtained by ranking hypotheses by **total cost**

Desh Raj, et.al., DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs, 2021

Dover-Lap


Hypothesis A
Hypothesis B
Hypothesis C
DOVER-Lap
Dorothy
Tinman

EENDasP



- Spoiler!!!!
- Labels from another system
- Apply EEND
- Solve for permutation
- Get new Labels with overlaps

More results

- Dihard III (track 1 – oracle SAD)

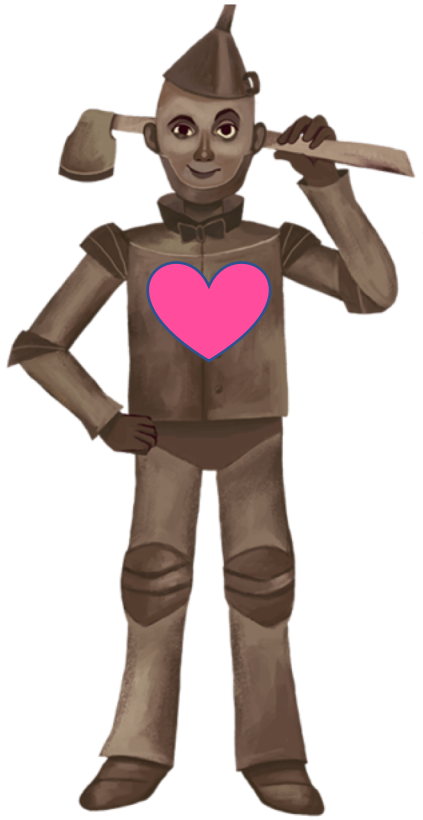
	DEV		EVAL	
System	full	core	full	core
Baseline	19.41	20.25	19.25	20.65
TDNN+VBx+Ovlassign	13.87	14.88	15.65	18.20
EEND-EDA	12.92	13.95	13.95	17.28
SC-EEND	13.13	13.13	15.16	19.14
TDNN+VBx+EENDasP	12.63	14.61	13.30	15.92
DOVER-Lap ( )	**10.73**	**12.56**	**11.83**	**14.41**

What is beyond Oz (diarization)

- Child centered data

System (Seedlings)	DER (%)
Baseline AHC	63
VBx+	61.49
EEND-EDA	62.57
Oracle VAD VBx	32.33

System (BLIP)	DER (%)
Baseline AHC	86.26
VBx	65.81
Oracle VAD VBx	41.01



- Stay tuned!

An example

- Day-long recordings

- In real life we *don't have dev data*, we *only have eval data* ☺

- Diarization error rates for all systems drop dramatically.

VanDam, Mark (2018). VanDam Public Daylong HomeBank Corpus. doi:10.21415/T5388S
https://media.talkbank.org/homebank/Public/VanDam-5minute/CI40/
Victoria Chua, Suzy Styles, et.al., Blip audio private collection provided by NTU

Takeaways

- Lots of flavors to choose from ☺
- We have huge improvements, but we are not yet there.
- Neural diarization is becoming as good as embedding clustering methods
- Overlap detection is still an ongoing research
- Diarization to help downstream tasks (like ASR)
- Day-long recordings, cocktail party scenarios need diarization solutions
- Self-supervised learning for diarization as a new direction

Thank you!