Diarization in a world full of DNNs (special tutorial)

Leibny Paola Garcia Perera

Johns Hopkins
Whiting School of Engineering

Center for Language and Speech Processing

Collaborators

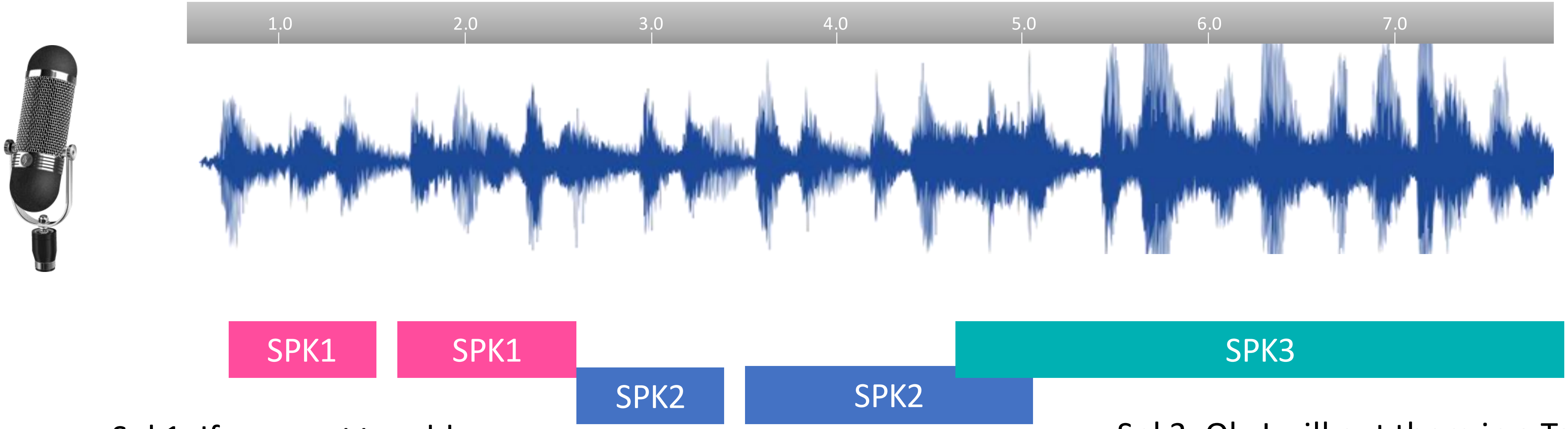
- Thanks to each one of them!
- JHU: Zili Huang, Desh Raj, Matthew Maciejewski, Jesus Villalba, Sanjeev Khudanpur
- NTU: Hexin Xie, Victoria Chua, Suzy Styles, Justin Dauwels
- CMU: Shinji Watanabe
- Some other collaborators: Latane Bullock, Herve Bredin, Marvin Lavenchin
- From Hitachi: Yusuke Fujita (now at Line), Shota Horiguchi, Yawen Xue, Yuki Takashima, Nelson Yalta
- CCWD, CDS project (lots of people behind this amazing project)

Disclaimer

There a lot of fascinating works out there.

These are just a few ones in which somehow I have been involved.

What is the goal?

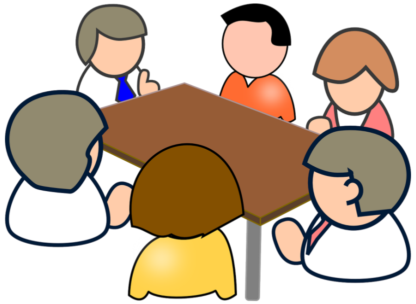


Spk1: If we want to address the next diarization problems..

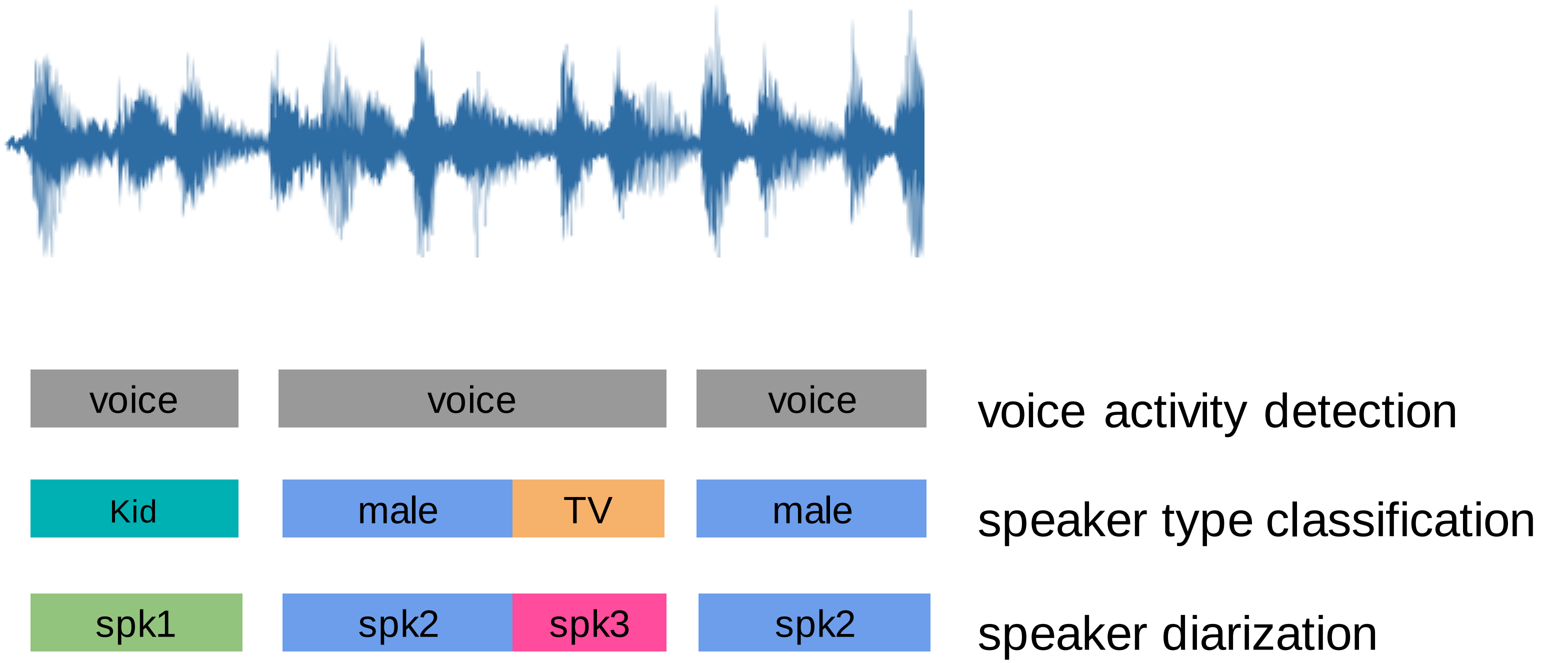
Spk2: You should need to first breakdown those results.

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

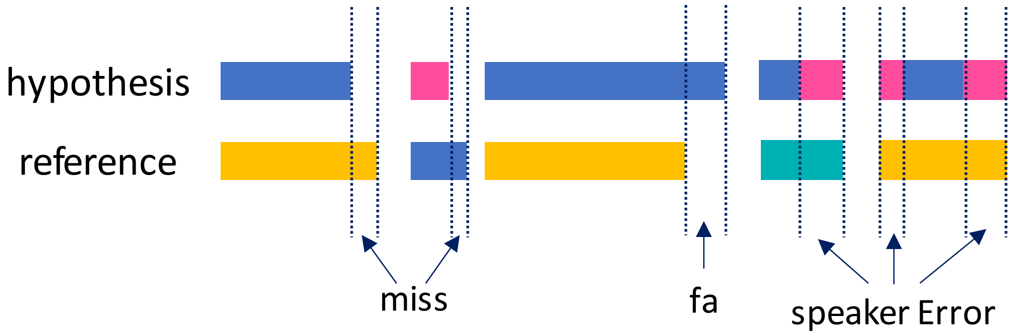
Where is diarization in the speech world?



Who spoke when?



How good the diarization is?



Diarization Error Rate

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

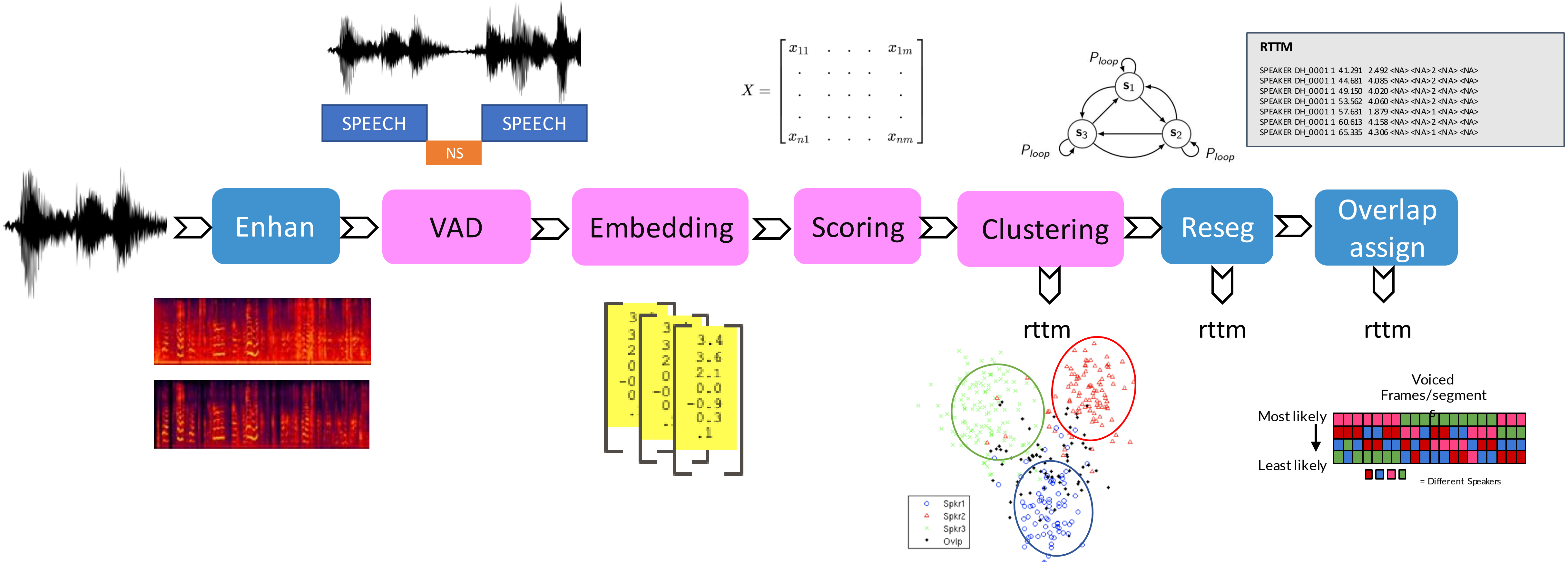
Jaccard Error Rate

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

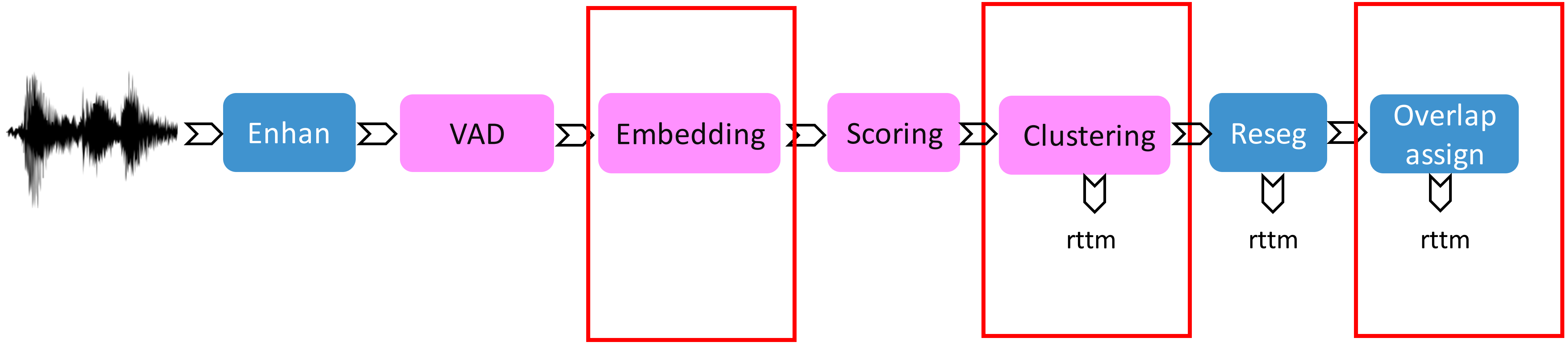
$$\mathrm{JER}_{s_i} = \frac{\text{false alarm}_{s_i} + \text{missed detection}_{s_i}}{\text{total}_{s_i}}$$

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

Traditional diarization approach

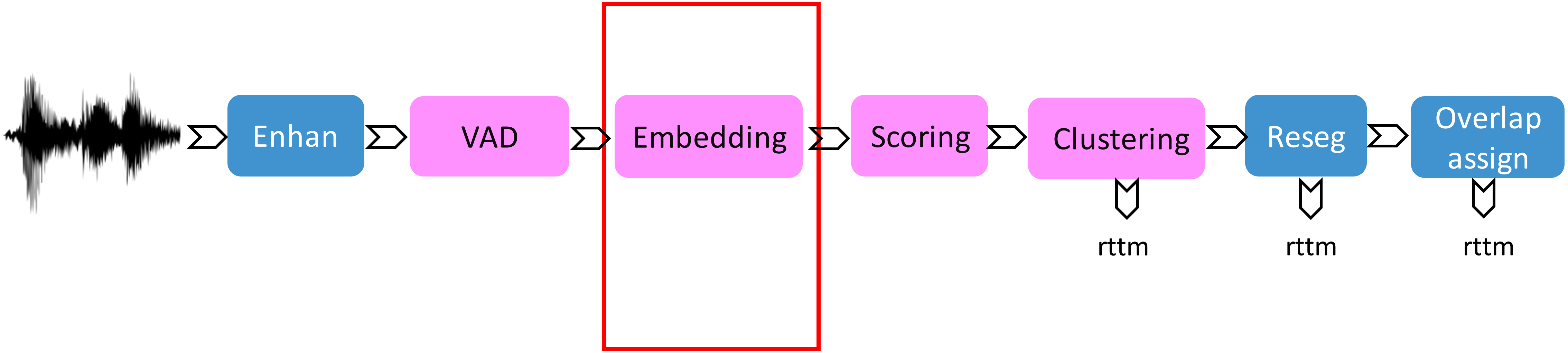


Key ideas



Enhan
VAD
Embedding
Scoring
Clustering
Reseg
Overlap assign
rttm
rttm
rttm

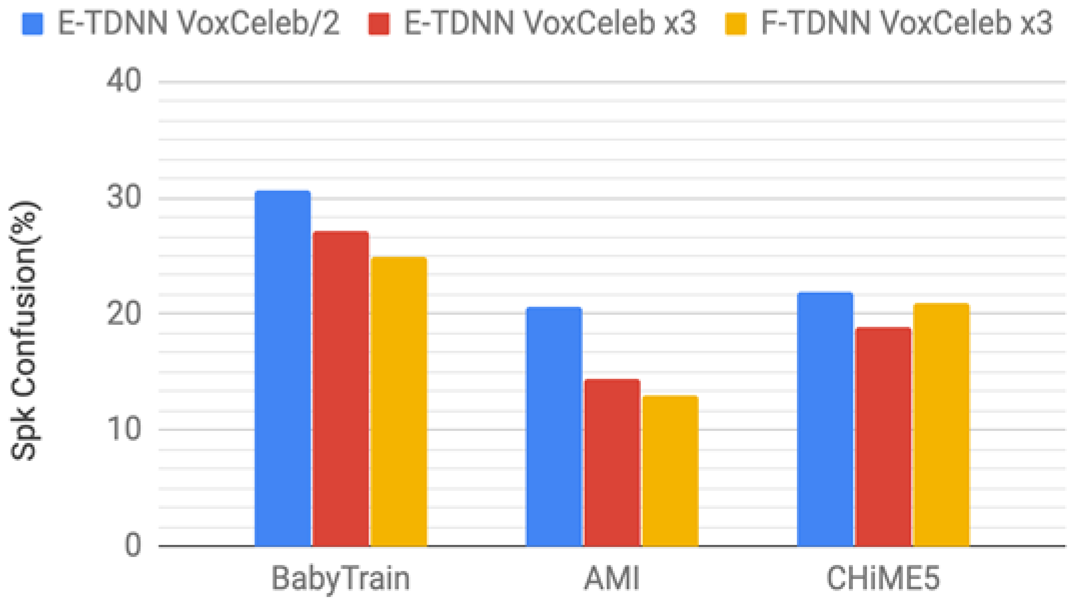
Key ideas



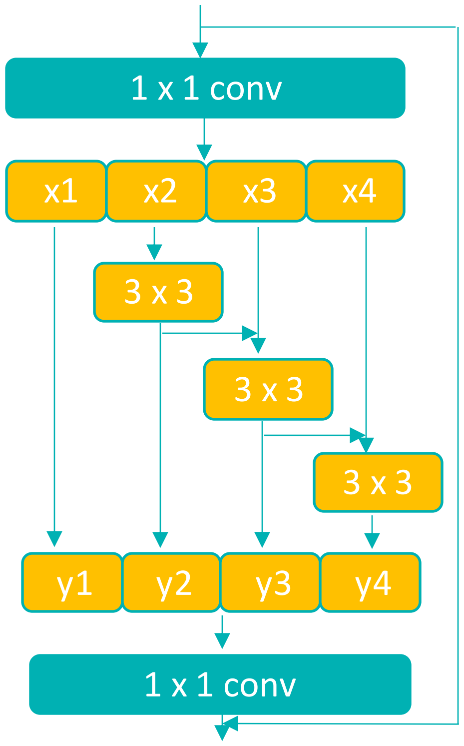
Enhan
VAD
Embedding
Scoring
Clustering
Reseg
Overlap assign
rttm
rttm
rttm

Embeddings

TDNN-xvector

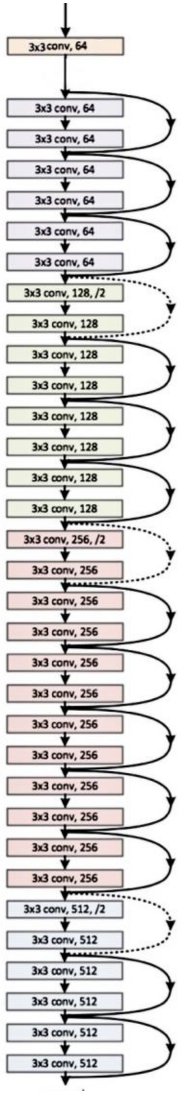


Res2net



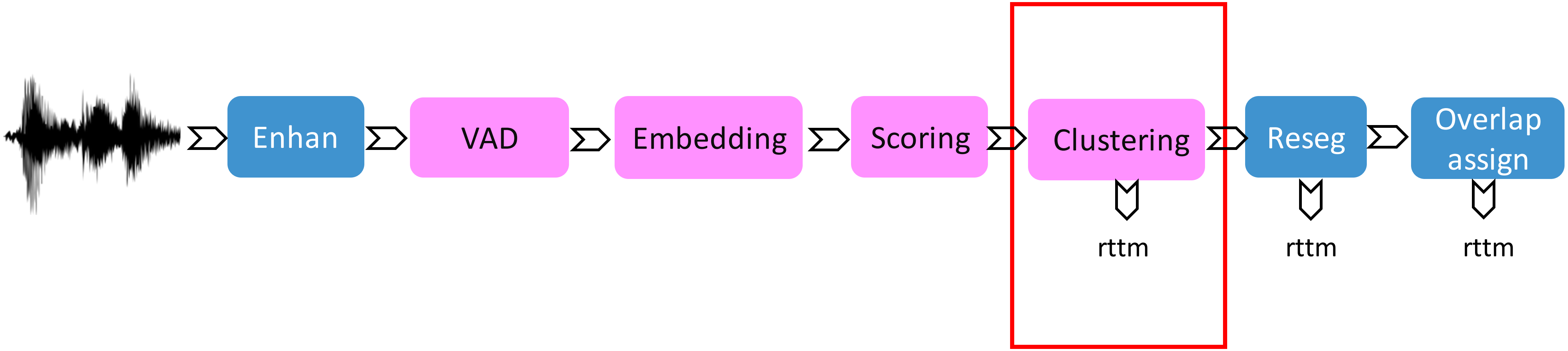
Res2Net Block

ResNet34



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
Wang, Weiking, et.al., The DKU-DukeECE-Lenovo Diarization System for VoxSRC 2021

Key ideas

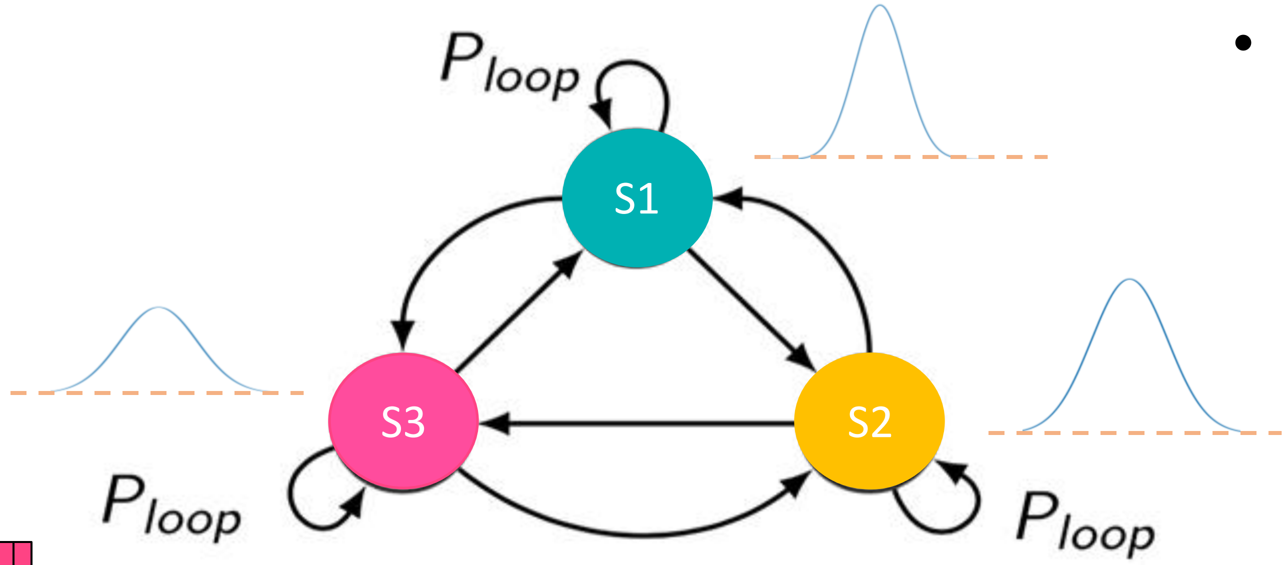


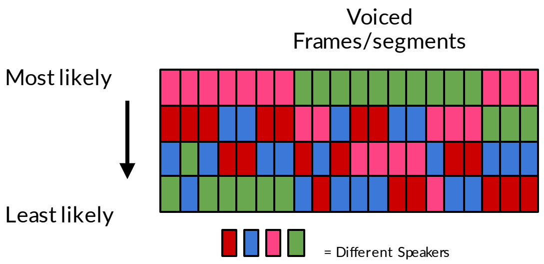
Enhan
VAD
Embedding
Scoring
Clustering
Reseg
Overlap assign
rttm
rttm
rttm

Clustering

- VB-HMM Clustering - VBx

- Uses x-vectors
- Same model as the Bayesian HMM
- A single Gaussian per state
- Parameters were initialized from pre-trained **PLDA** model.

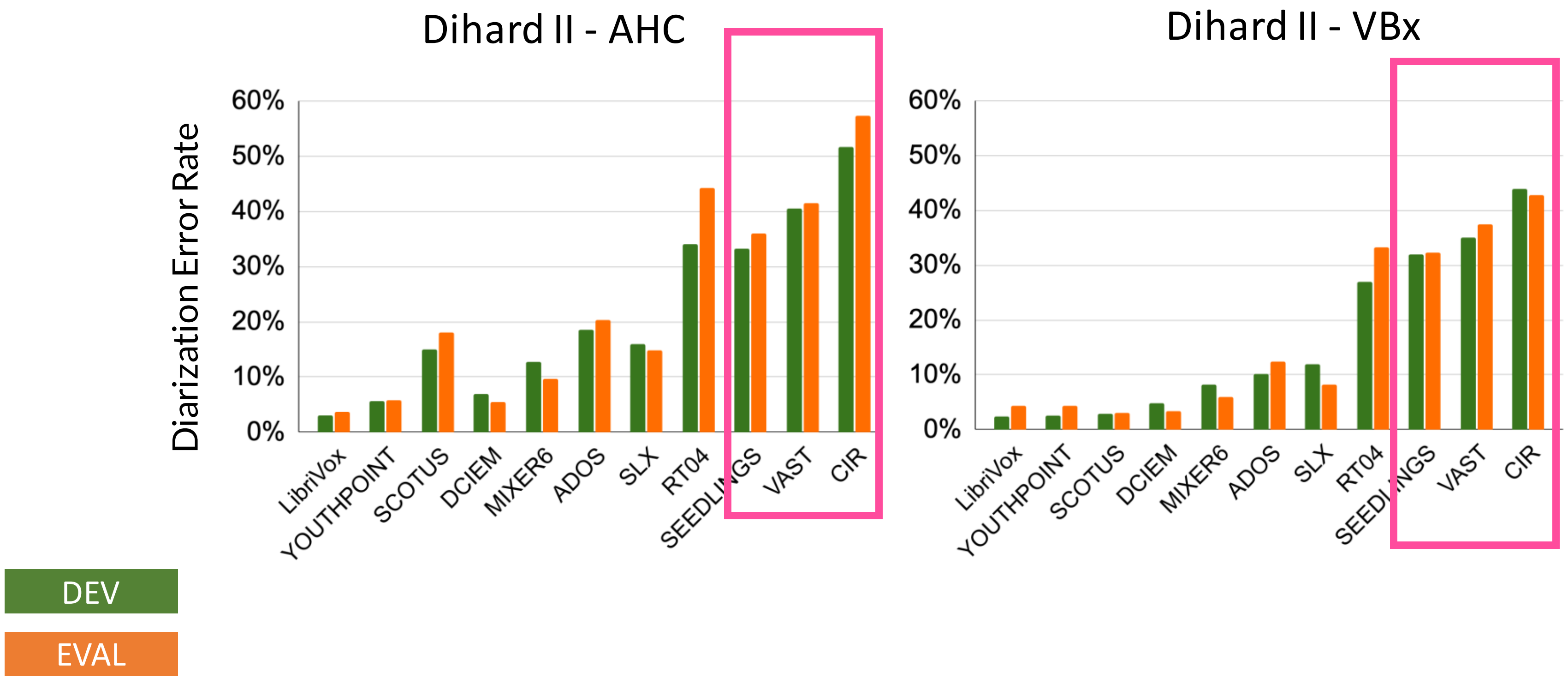




Mireia Diez, et.al., Speaker Diarization based on Bayesian HMM with Eigenvoice Priors, 2018.

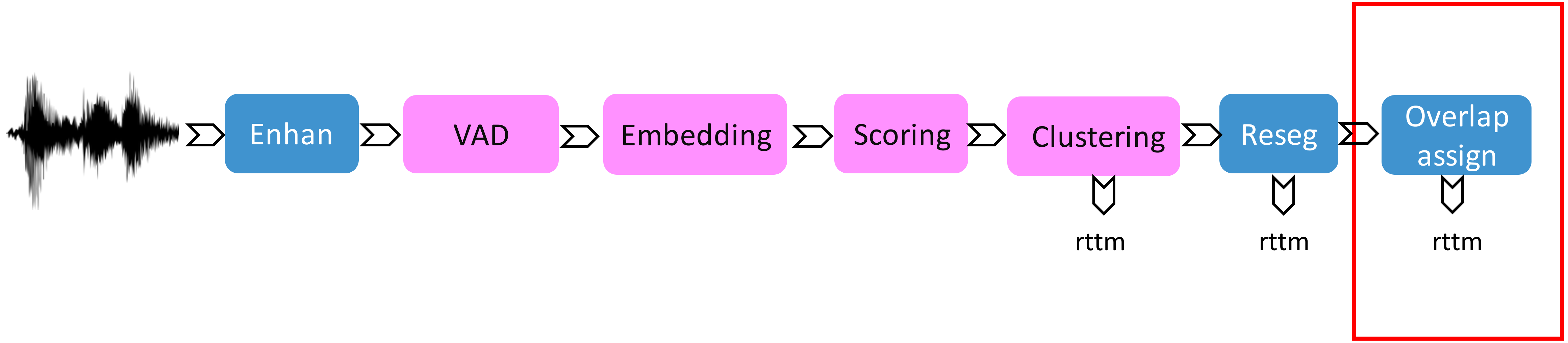
Clustering

- VB-HMM Clustering



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, 2017.

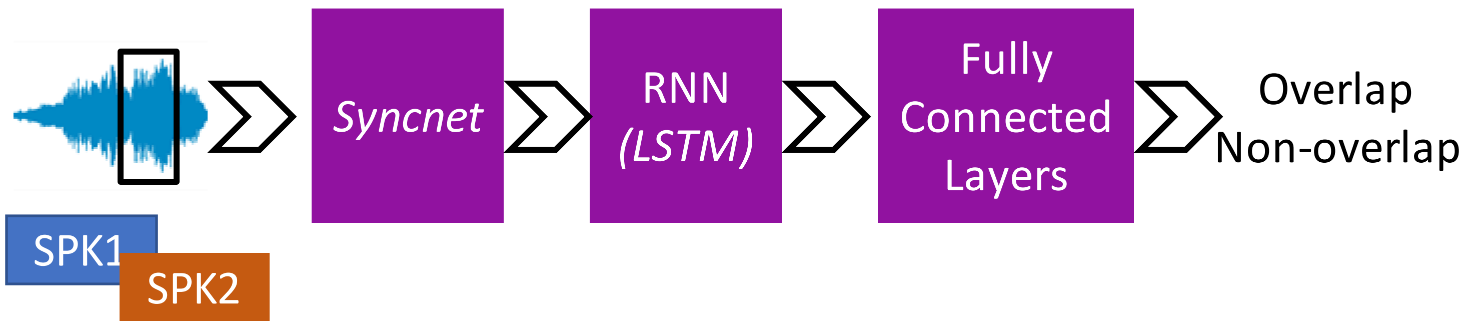
Key ideas


Enhan
VAD
Embedding
Scoring
Clustering
Reseg
Overlap assign
rttm
rttm
rttm

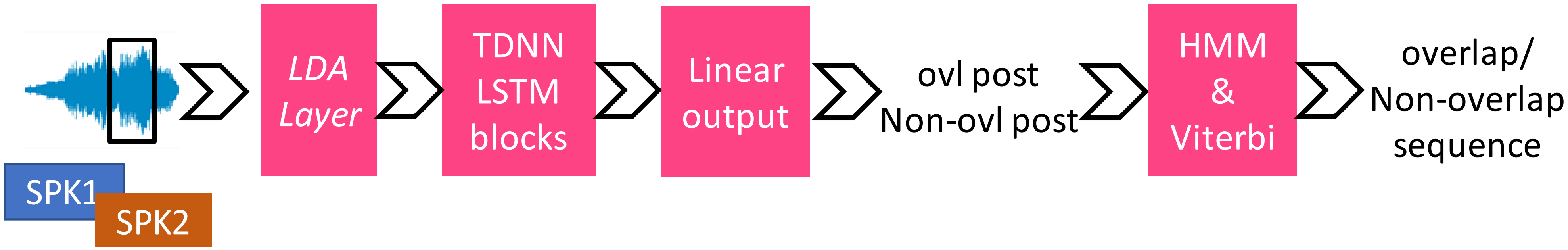
Overlap Assignment

- Neural network Overlap detector

pyannote

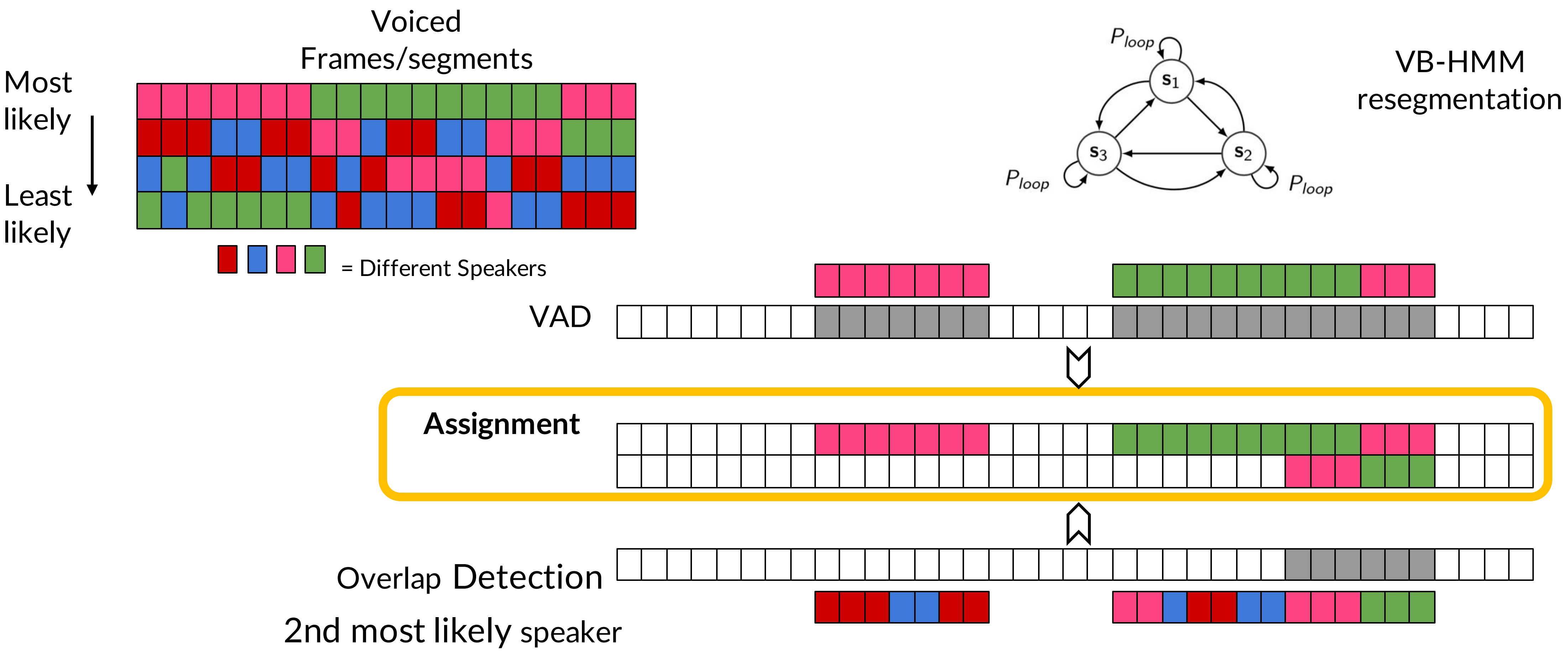


- HMM-DNN based overlap detector



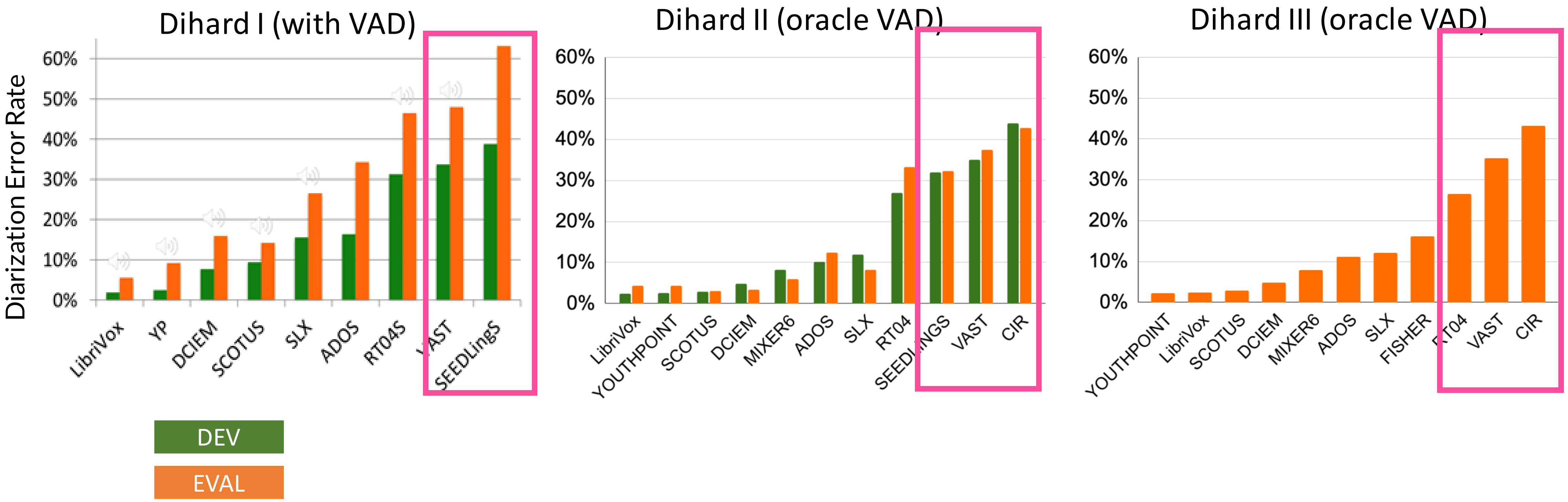
Latane Bullock, et.al., Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection, 2020.
Desh Raj, et.al., Multi-class Spectral Clustering with Overlaps for Speaker Diarization, 2020.

Overlap Assignment



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

Where are we?



Mireia 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, 2017.
Hitachi-JHU diarization system, DIHARD III, 2020.

S.O.S! We have some issues

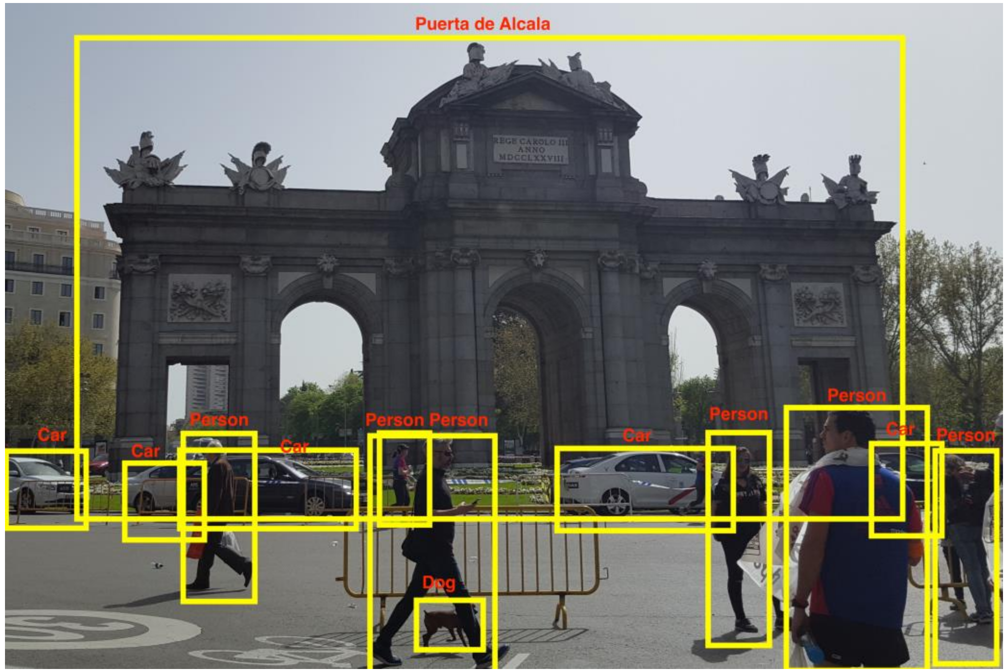
- We are still trying to handle overlapping speakers

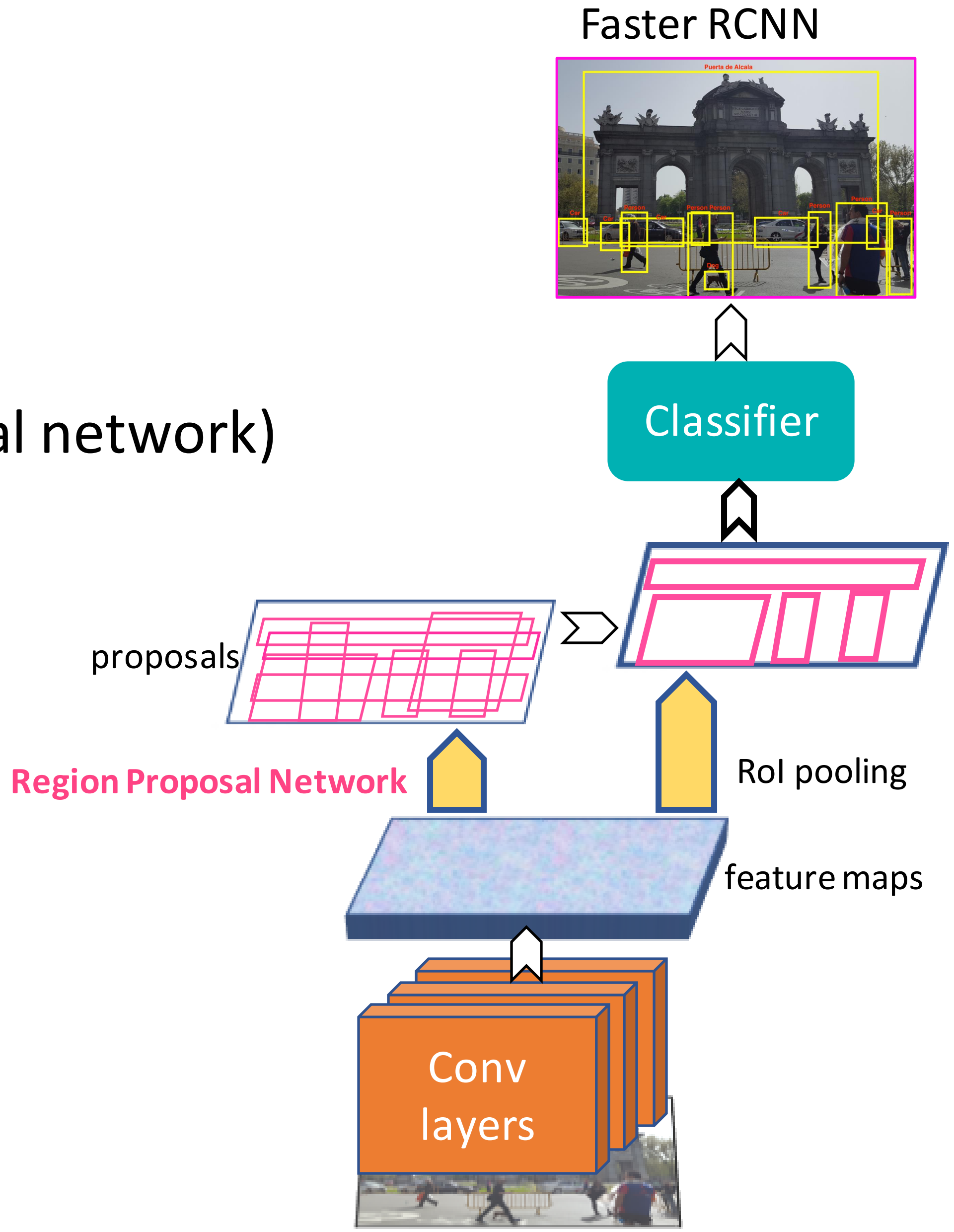
- The system is not designed to minimize the diarization error but optimizes every module separately.

So we look for solutions!

Region Proposal Network

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

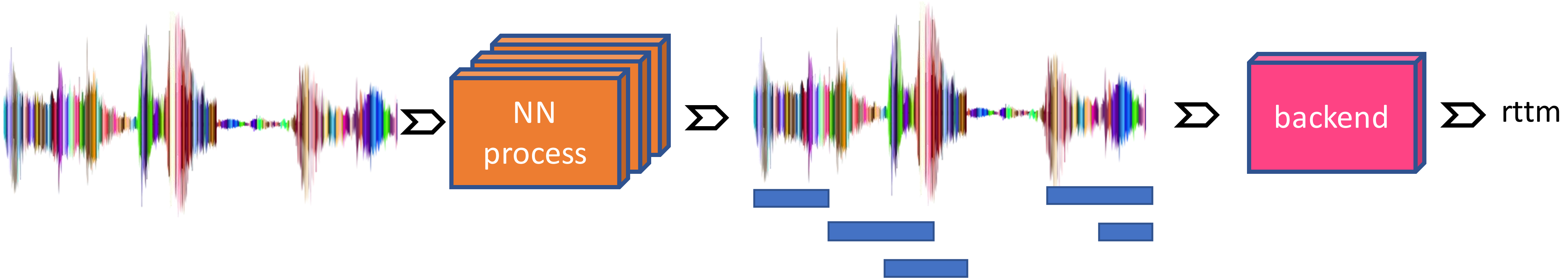




Zili Huang, et.al., Speaker Diarization with Region Proposal Network, 2020.
Ren, Shaoqing, et al. "Faster R-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems,* 2015.

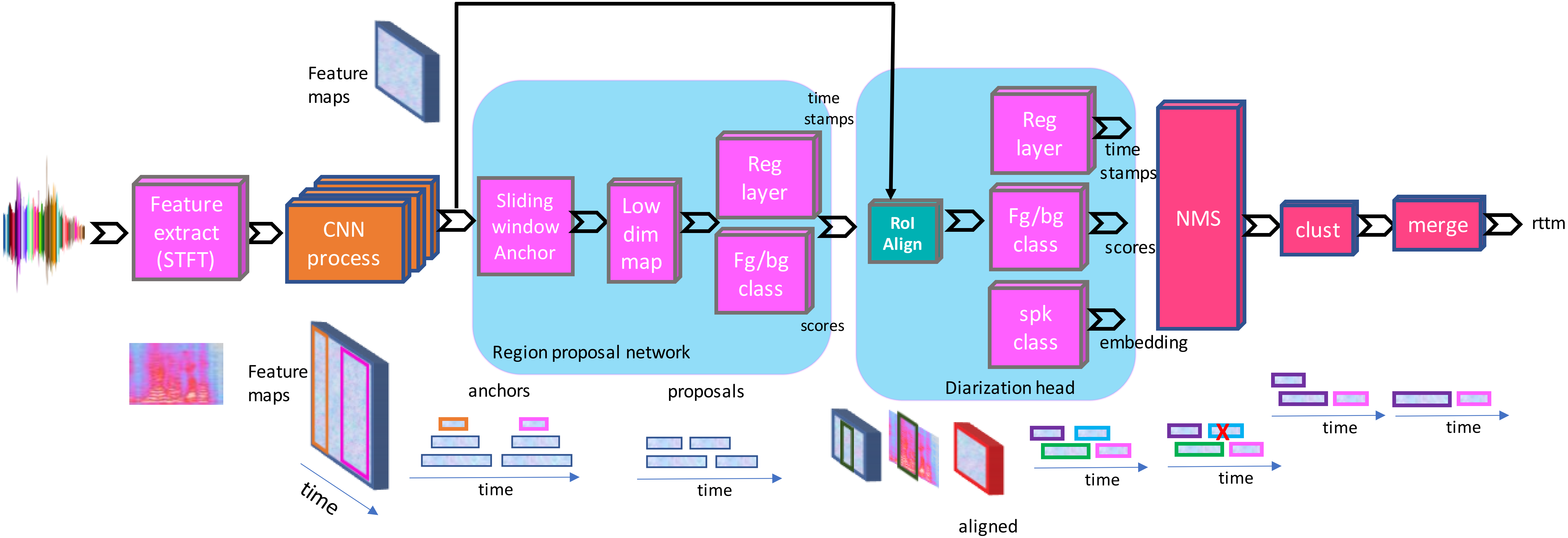
Region Proposal Network for Speaker Diarization

- Same idea as in image processing but 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?

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

  - 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.

We still need to know:

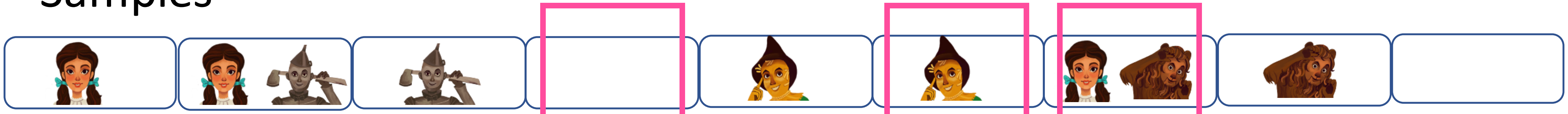
- How to handle overlapping speakers?
- How to design the system in such a way that the diarization error is minimized?
- Using DNNs in a world full of DNNs

Neural diarization

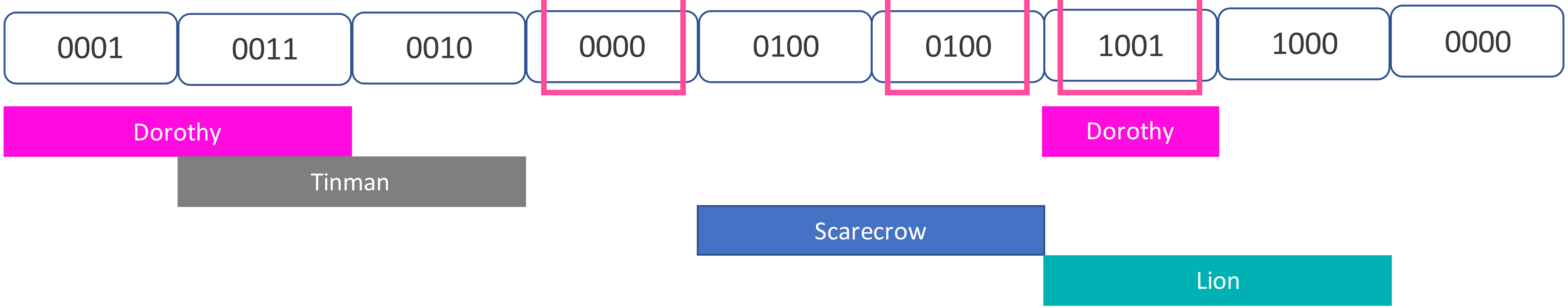
Yusuke Fujita, et.al., End-to-End Neural Diarization: Reformulating Speaker Diarization as Simple Multi-label Classification,2019.

Diarization as a multi-label classification

- Samples

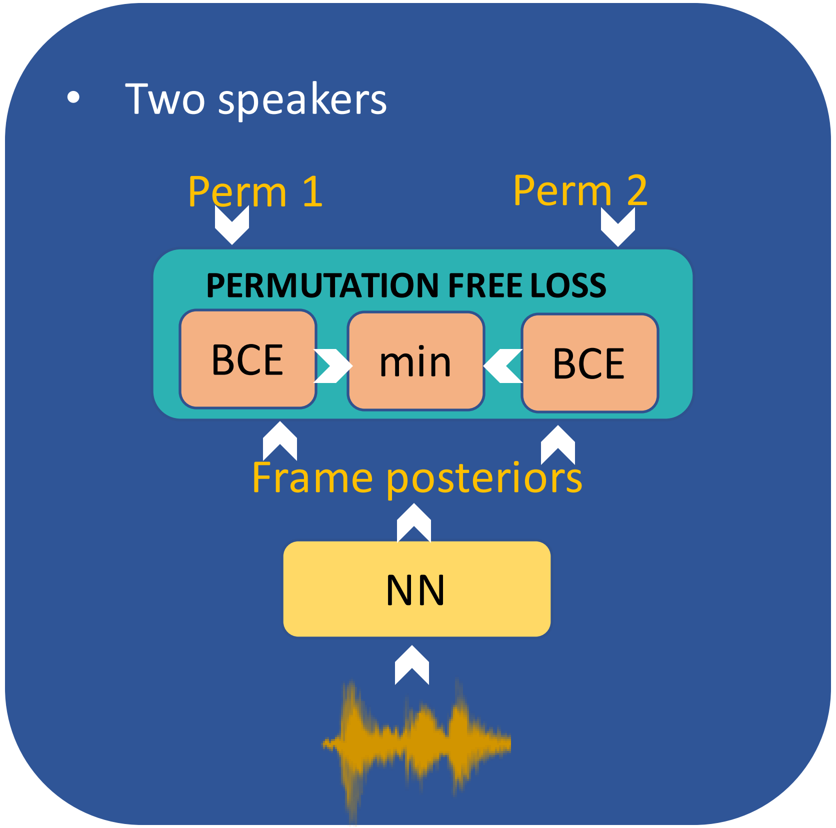


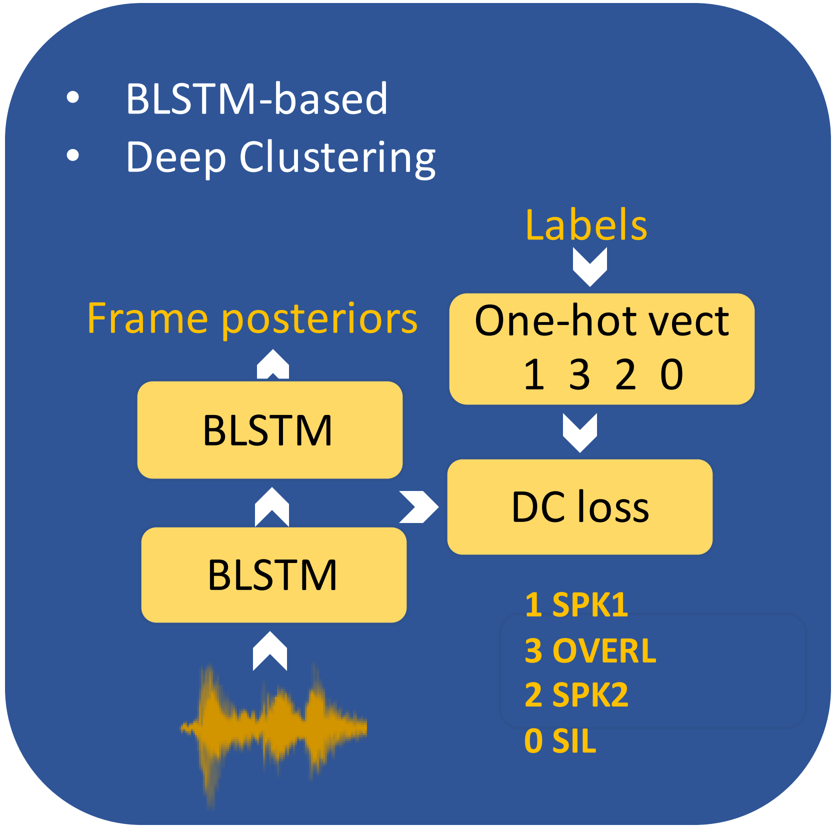
- Labels

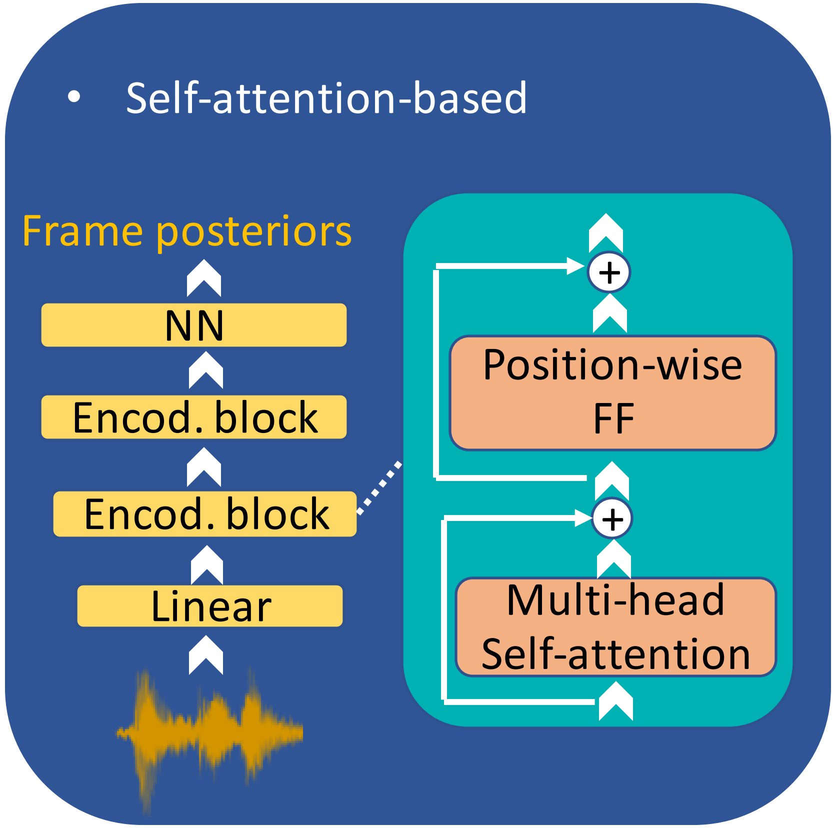


EEND evolution

$$\mathcal{L}_{diar} = \frac{1}{ST} \min_{\phi \in \mathrm{perm}(S)} \sum_{t=1}^{T} \mathrm{BCE}(\hat{y}_t, y_t^{\phi})$$

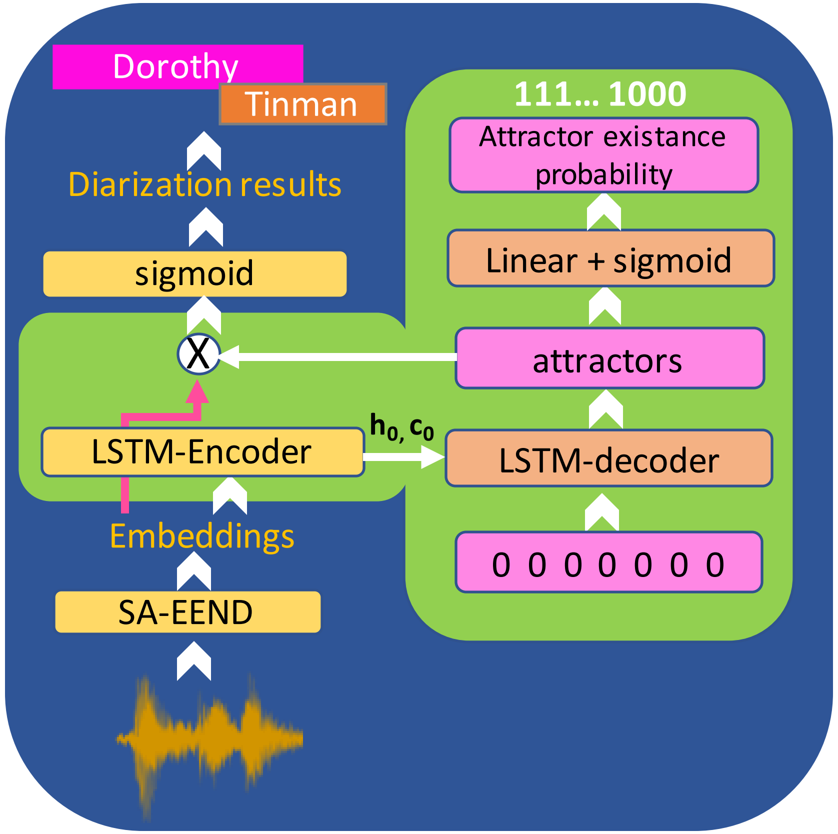






Yusuke Fujita, et.al., End-to-End Neural Diarization: Reformulating Speaker Diarization as Simple Multi-label Classification, 2019.

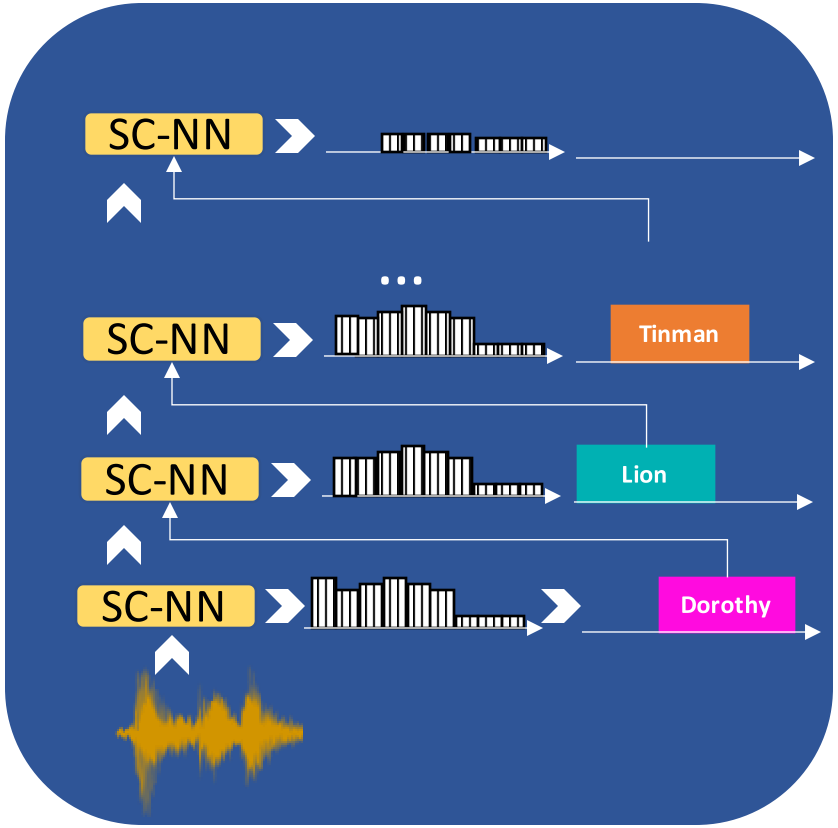
EEND-EDA



- 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.

Shota Horiguchi, et.al., End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors, 2020.

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 the appropriate permutation.

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

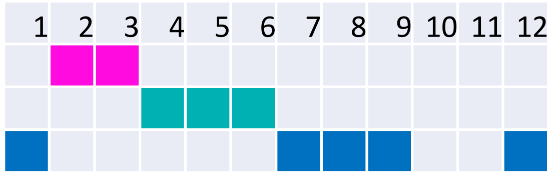
Some results

DIHARD III (track 1- oracle SAD)

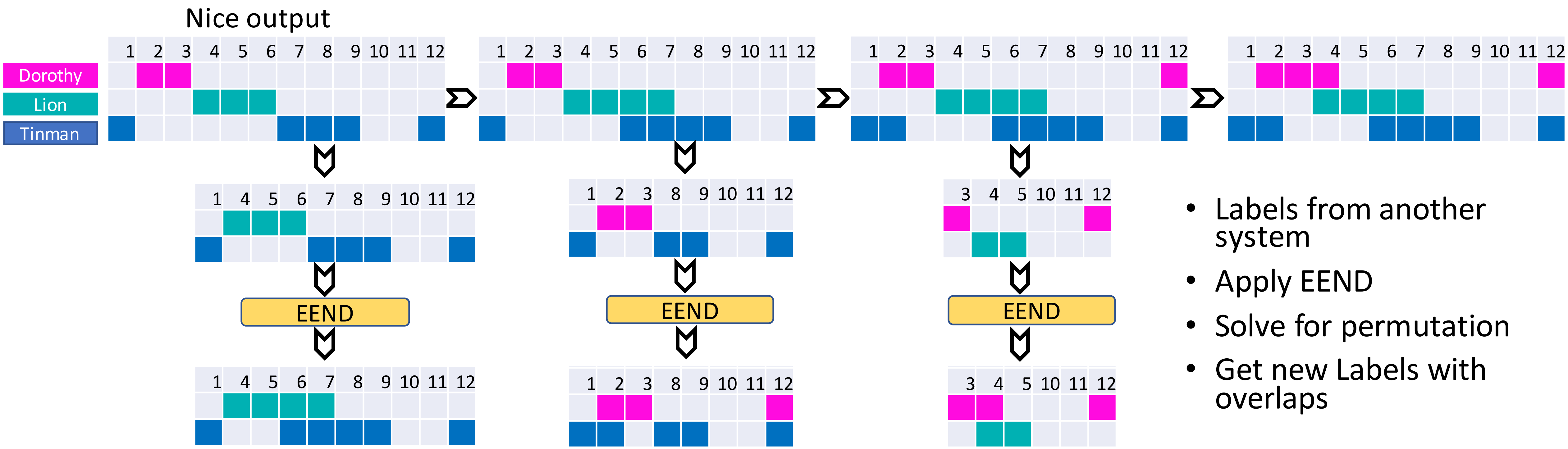
	DEV (DER%)		EVAL (DER%)	
System	full	core	full	core
Baseline	19.41	20.25	19.25	20.65
TDNN+VBx+Ovlassign	13.87	14.88	15.65	18.20
EEND-EDA	**12.92**	13.95	**13.95**	**17.28**
SC-EEND	13.13	**13.13**	15.16	19.14

What if we have a system that does not consider overlap, is it possible to fix it?

Nice output



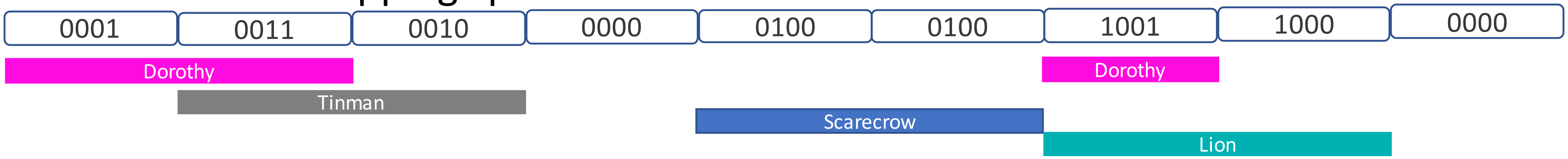
EENDasP



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

Shota Horiguchi, et.al., End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors, 2021.

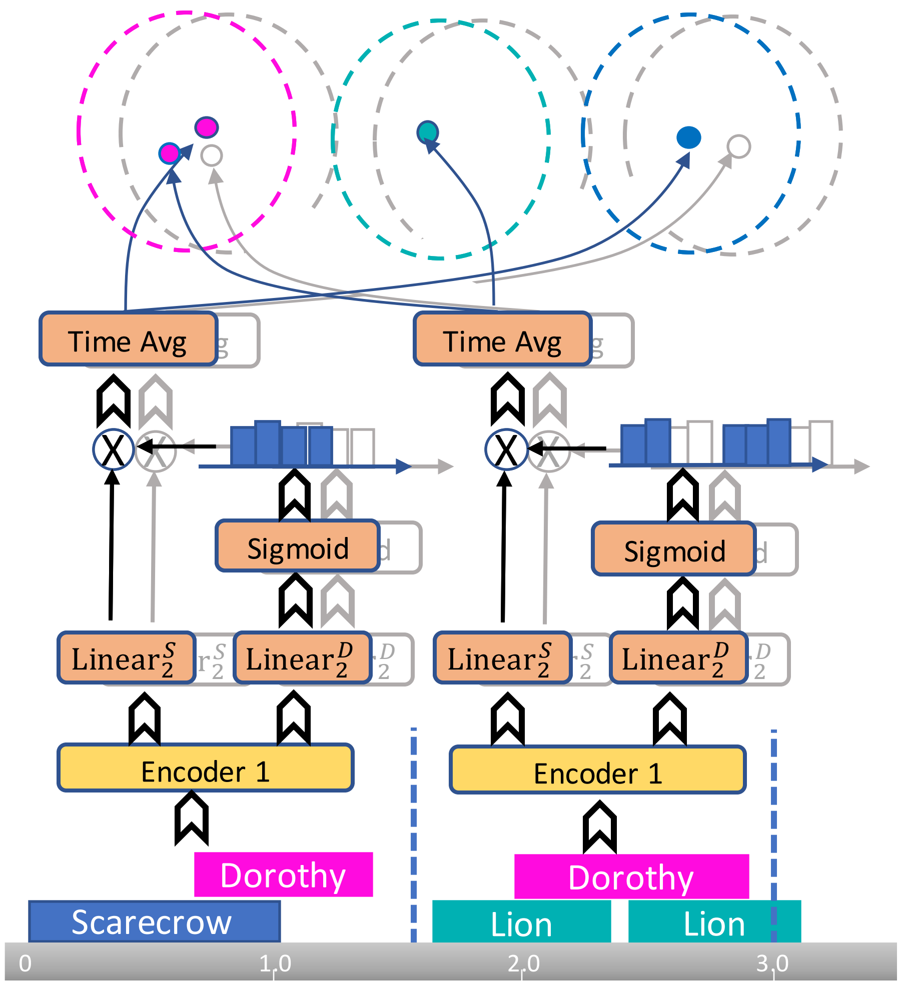
- We are considering the two problems now:
  - Handle overlapping speakers



  - Designed to minimize the diarization error.

$$\mathcal{L}_{diar} = \frac{1}{ST} \min_{\phi \in \mathrm{perm}(S)} \sum_{t=1}^{T} \mathrm{BCE}(\hat{y}_t, y_t^{\phi})$$

EEND-vector clustering

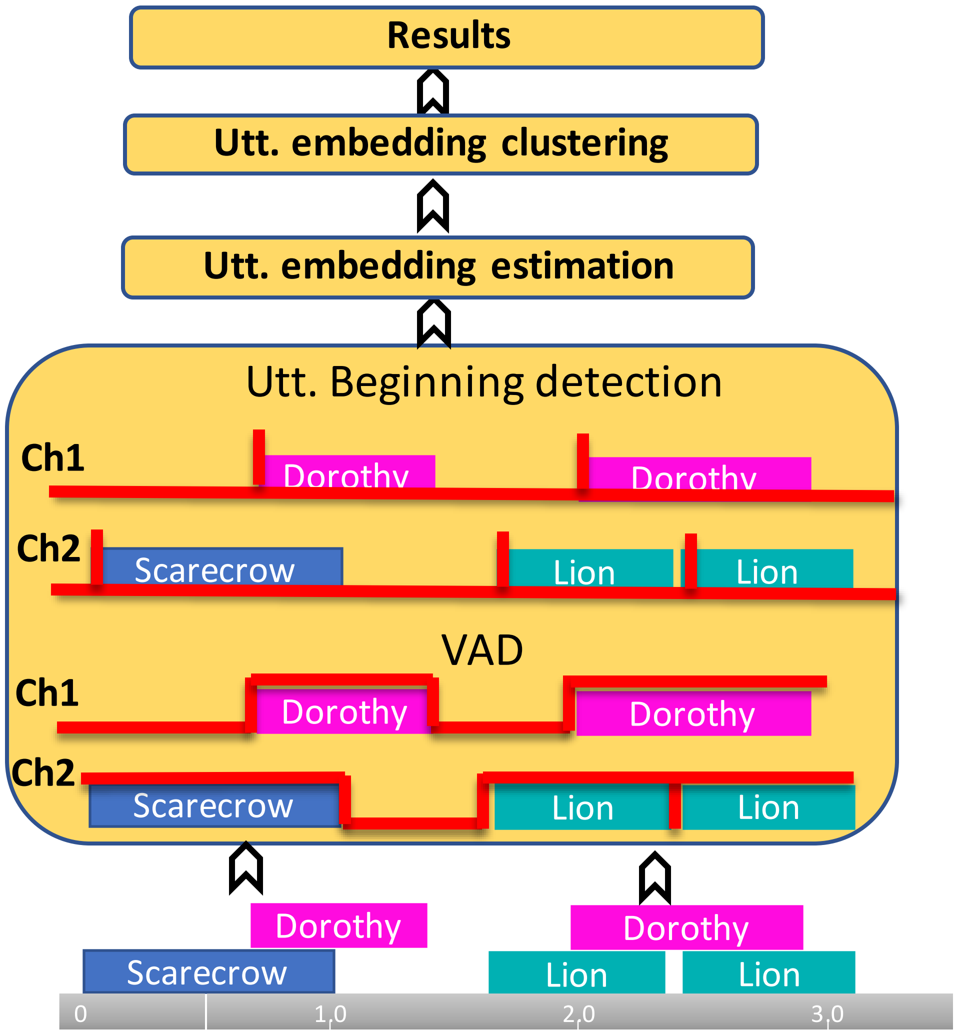


- EEND-vector clustering
- Hybrid system for overlapped speech, long recordings and different number of speakers

	Test duration (min)			
System	3	5	10	20
EEND	8.0	8.7	9.3	N/A
Propos+chunking+clust	7.4	6.5	5.9	5.5

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

EEND (utterance by utterance)

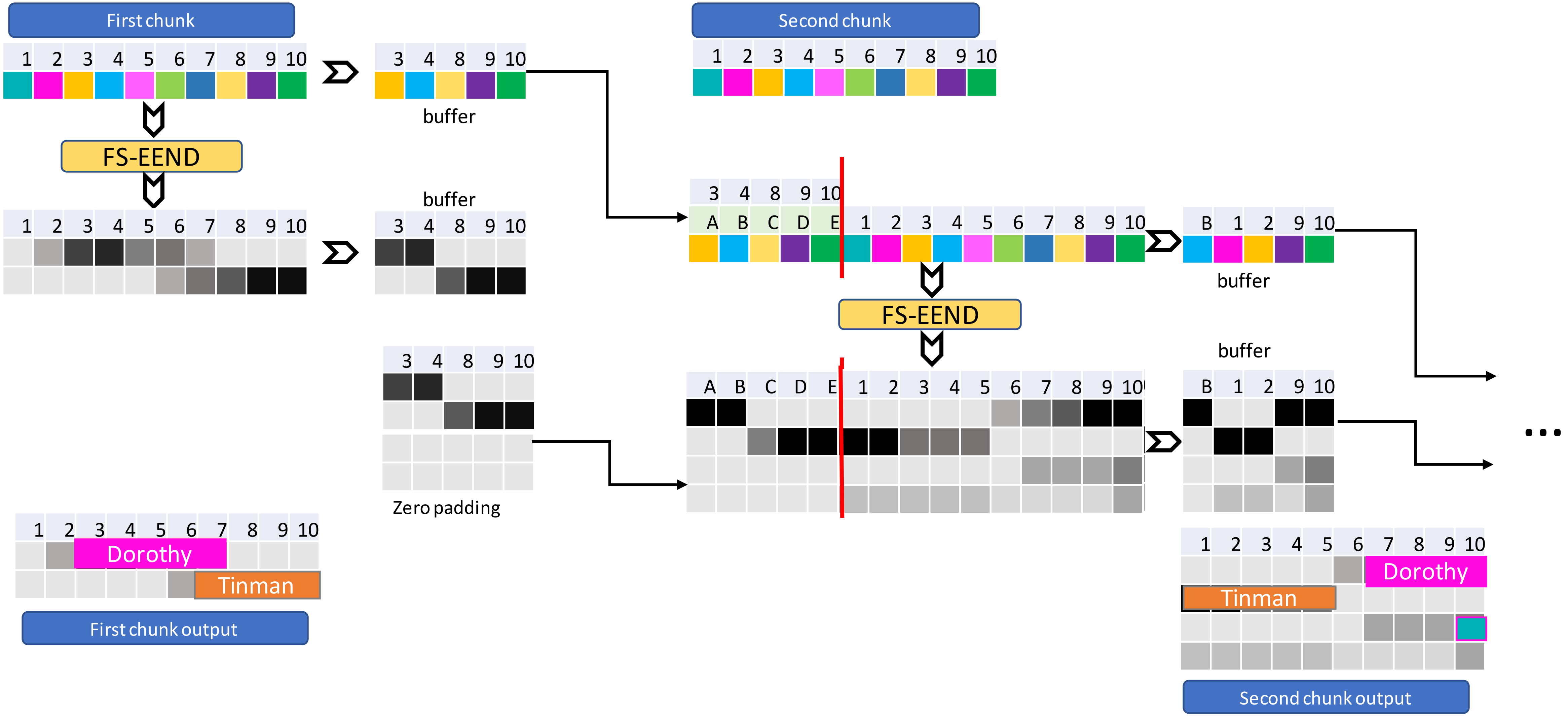


- Graph-PIT –based VAD
- Segmentation is not longer a limitation
- Utterance by utterance diarization
- Callhome

Method	Number of speakers					
	2	3	4	5	6	Avg
EEND-VC-5s	7.0	14.2	16.7	31.6	29.9	13.7
Graph-PIT-EEND-VC	7.1	12.6	18.3	31.1	30.7	13.5

Keisuke Kinoshita, et al., Utterance-by-utterance overlap-aware neural diarization with Graph-PIT

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**

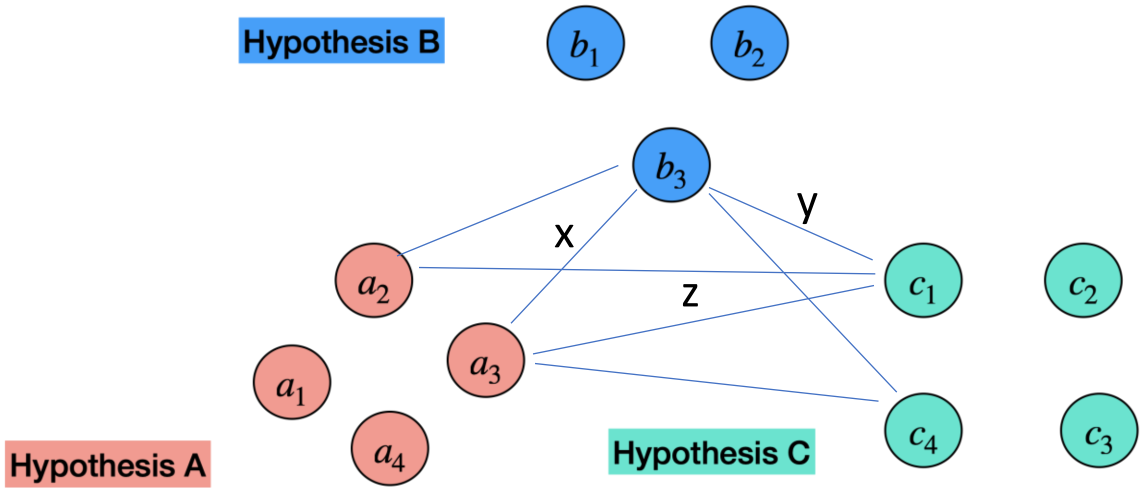
*Enrico Fini, et.al., Supervised online diarization with sample mean loss for multi-domain data, 2020.
Yawen Xue, et.al., Online End-to-End Neural Diarization with Speaker Tracing Buffer, 2021.

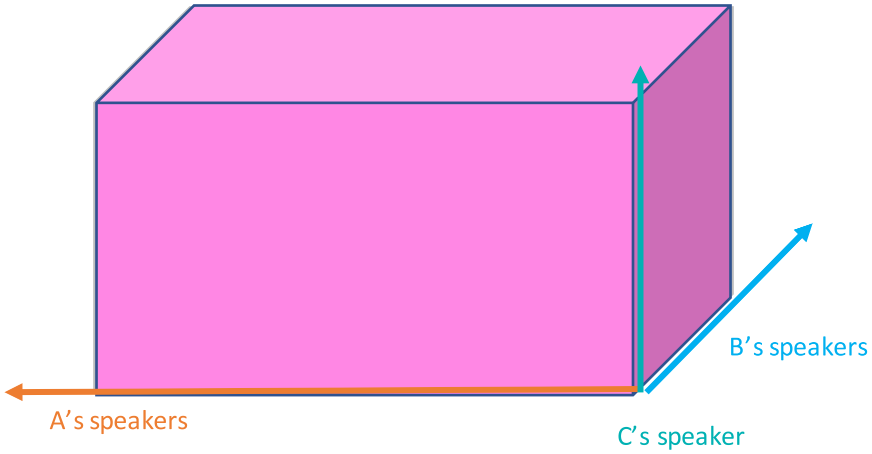
Let’s say that we have four or five different systems with different scores for overlapping regions. Is there a way to deal with them?

DOVER-lap

DOVER-Lap

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

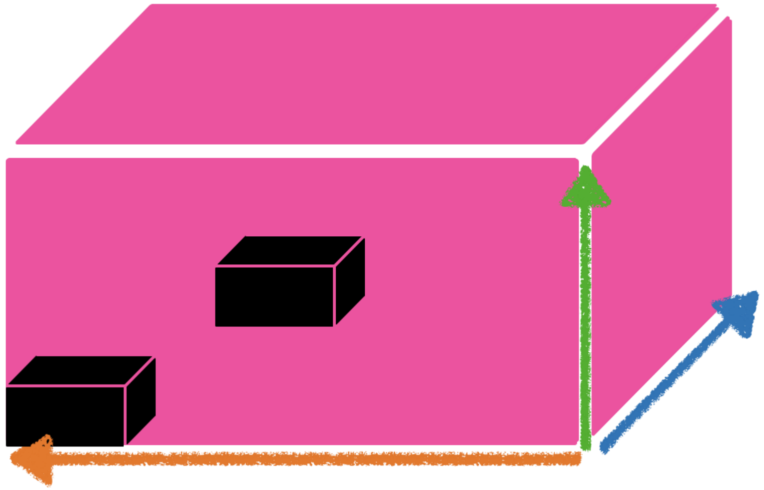


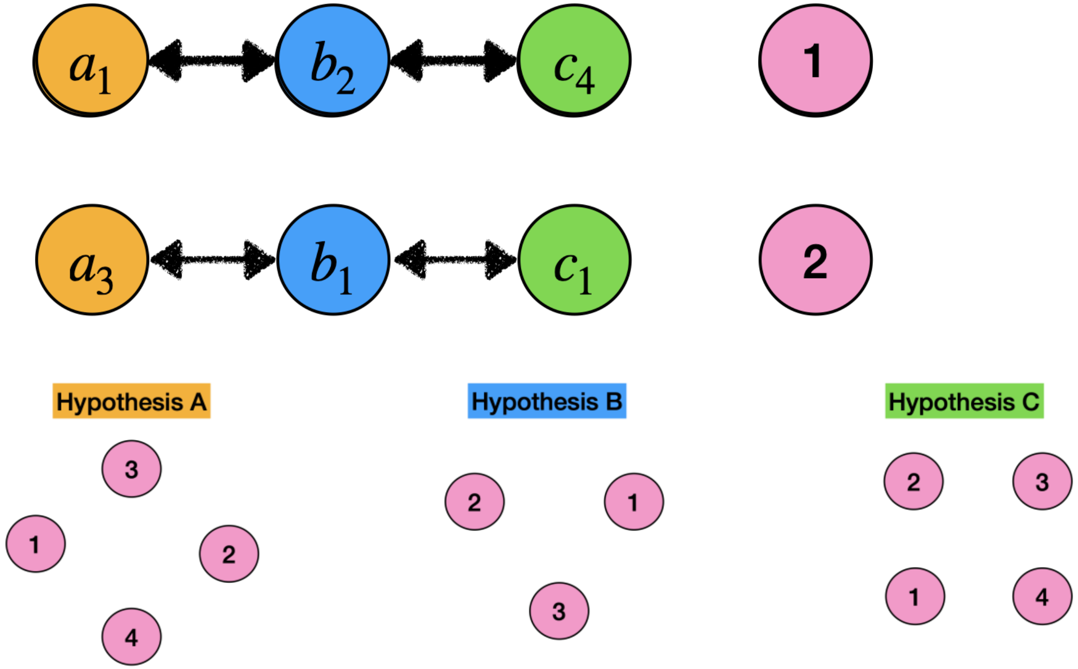


Desh Raj, et.al., DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs, 2021.

DOVER-Lap

- Tuple with lowest cost and assign the same label

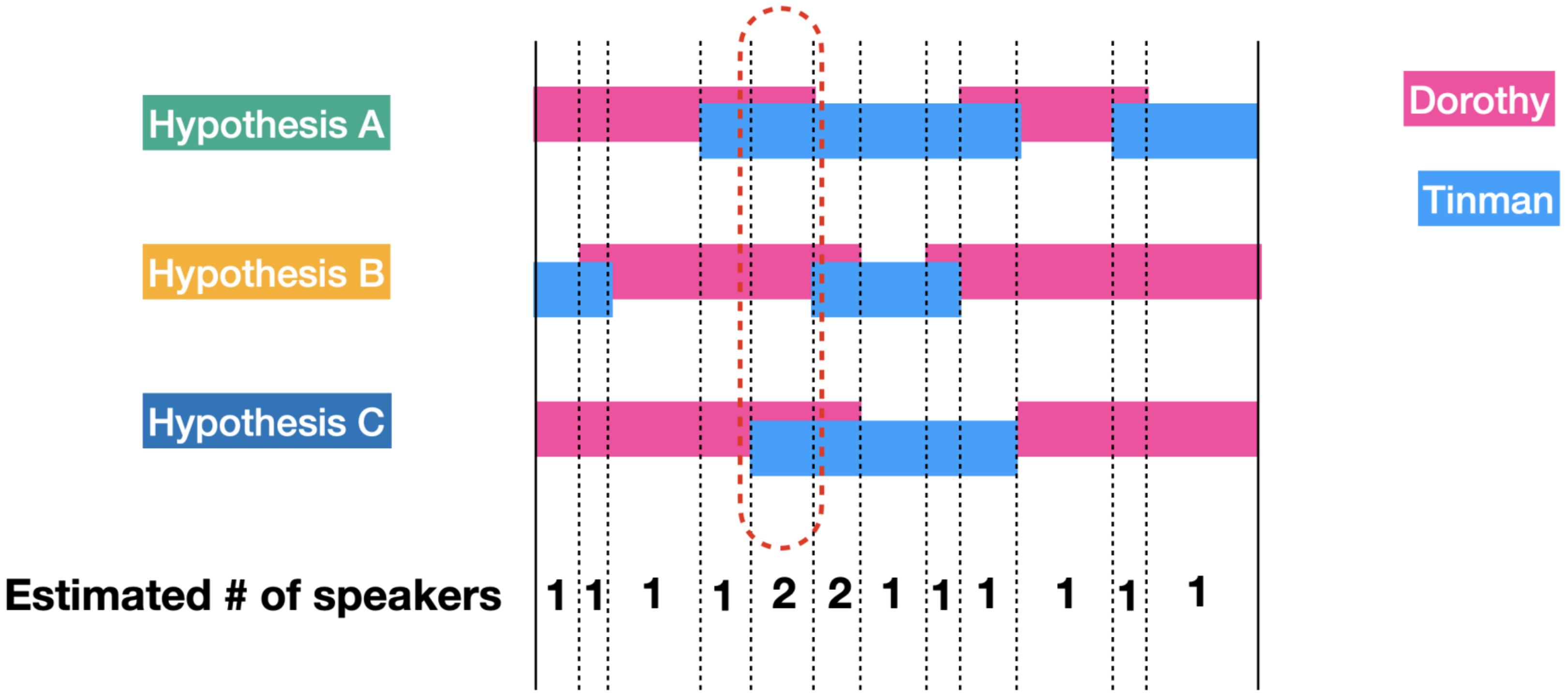




Desh Raj, et.al., DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs, 2021.

Dover-Lap

- Label voting

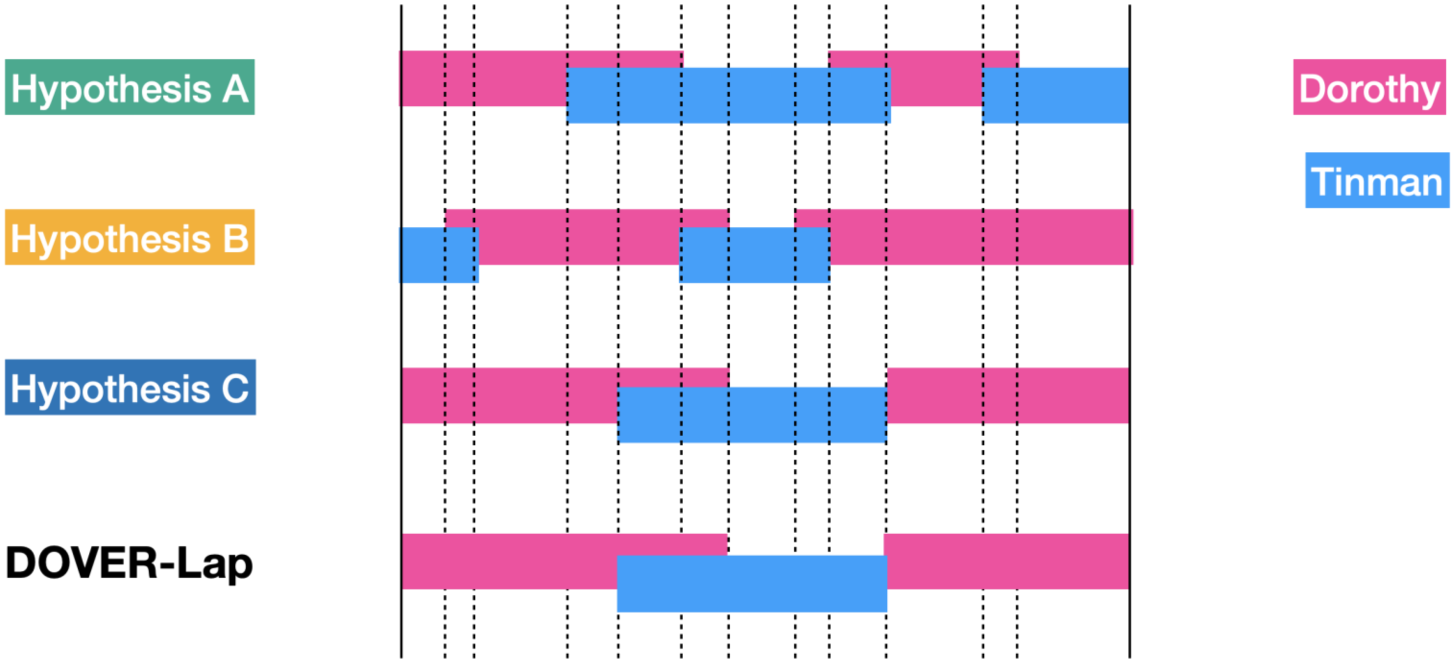


 # speakers = weighted mean of # speakers in hypotheses

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



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

Some results

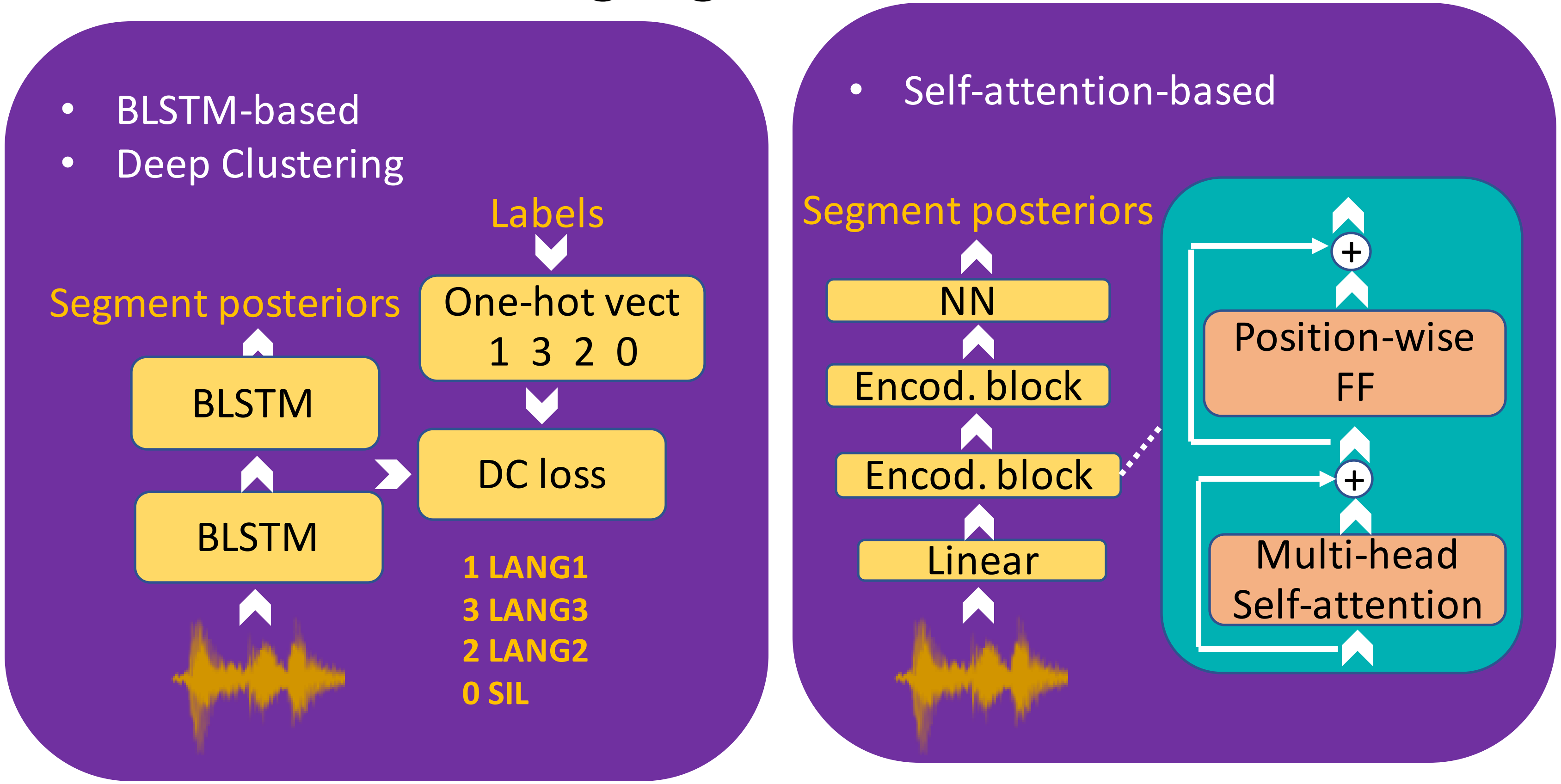
- Dihard III (track 1 – oracle SAD)

	DEV (DER%)		EVAL (DER%)	
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**

Can this idea be extended to other tasks?

Language diarization

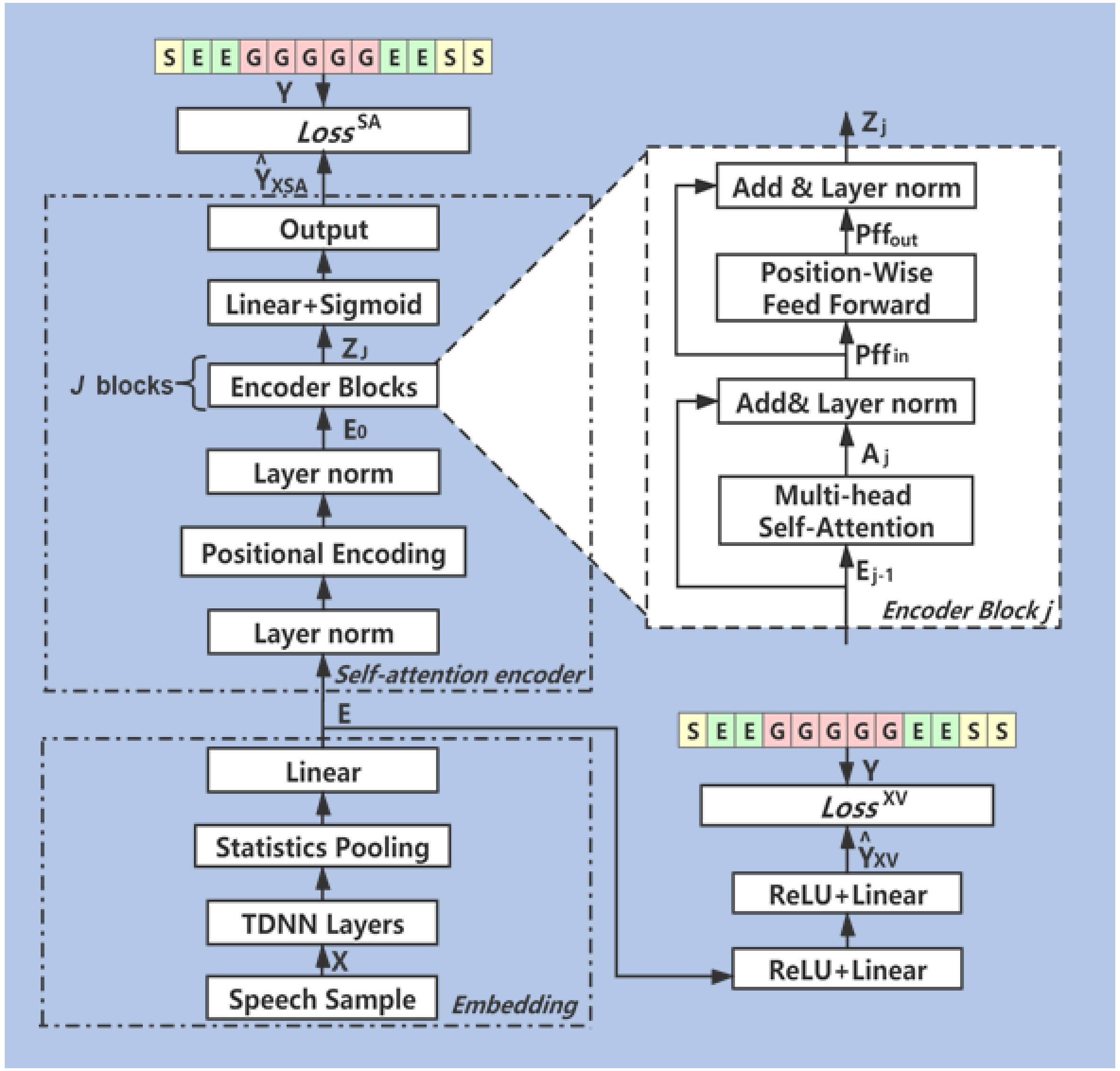
- End-to-end extension for Language diarization



Hexin Liu, et.al., End to End Language Diarization for Bilingual codeswitching data , 2021.

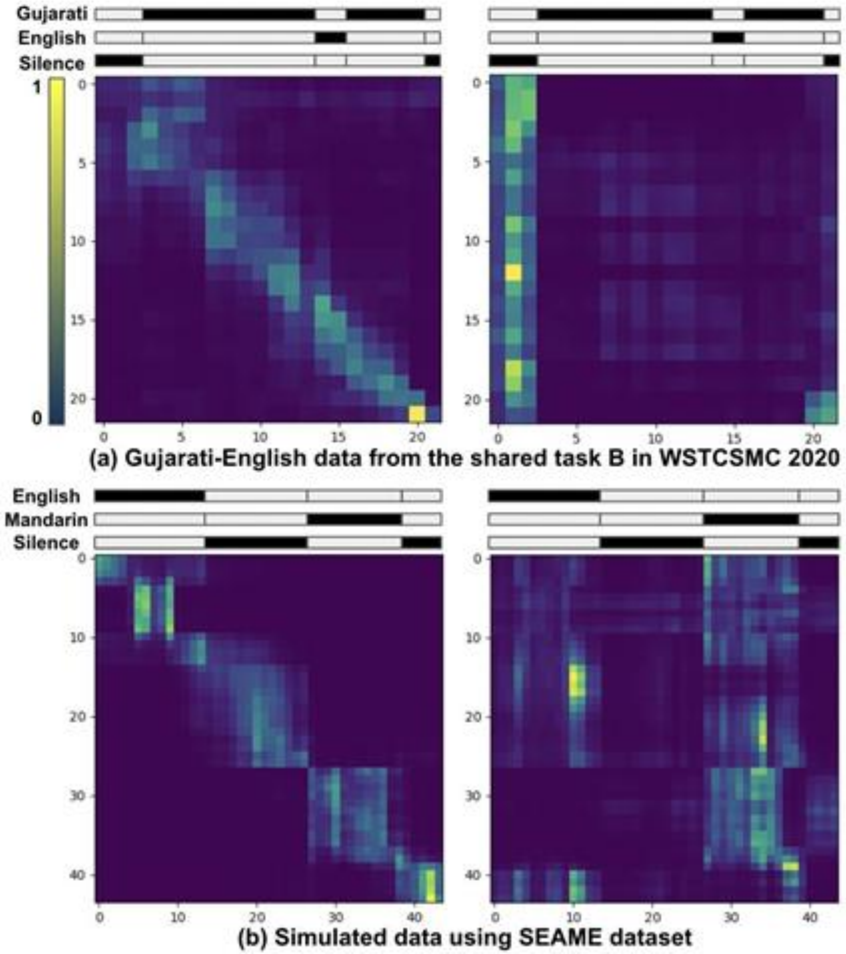
Language diarization

- Similar idea as EEND
    - **Bilingual code-switching speech**
- ○ End-to-end model with joint training
- ○ The **speech activity detection,** defining silences as a label
- ○ Hierarchical processing:
    - ▪ Segment-level (200ms) local language information: x-vector approach
    - ▪ Global dependency: self-attention transformer encoder



Hexin Liu, et.al., End to End Language Diarization for Bilingual codeswitching data , 2021.

Language diarization



What are the heads doing?

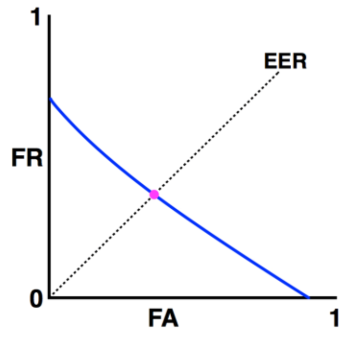
- The left head shows a linear transformation
- The right head highlights the different classes.

Hexin Liu, et.al., End to End Language Diarization for Bilingual codeswitching data , 2021.

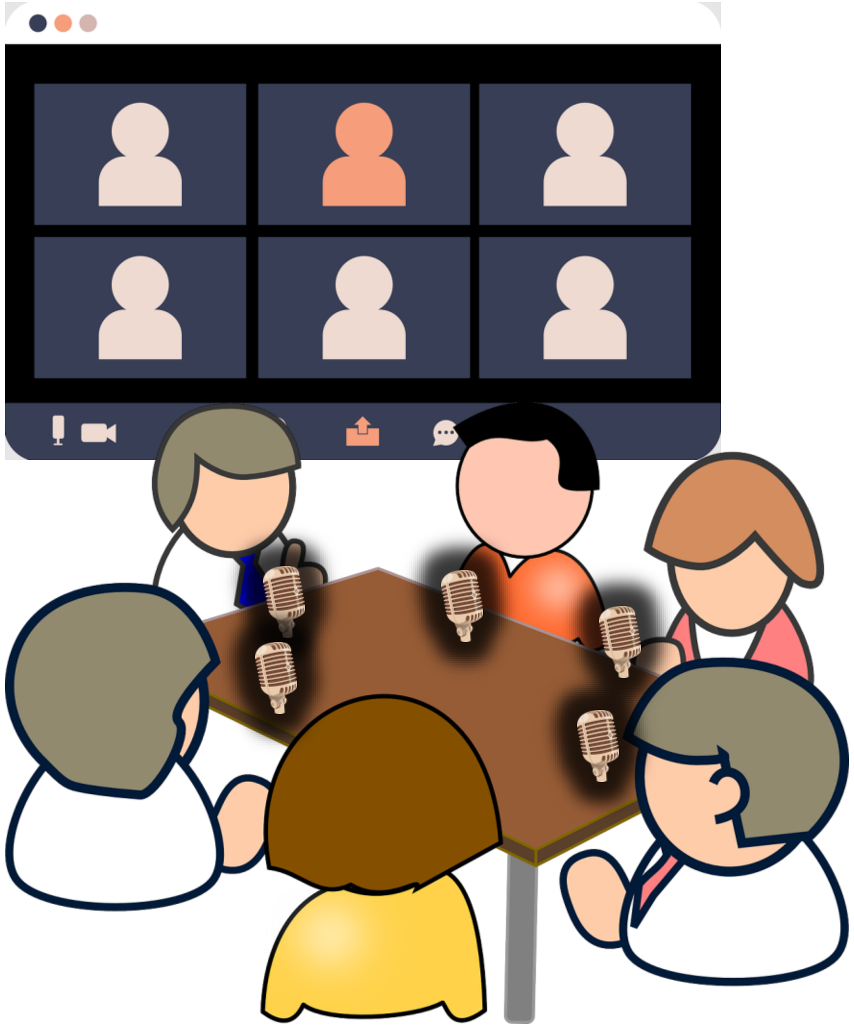
Some results on Language diarization

- Comparison of our approaches on 3-language-pair code-switching data in WSTCSMC 2020 (First Workshop on Speech Technologies for Code-switching in Multilingual Communities 2020)

	EER (%)					
Method	en	gu	ta	te	silence	Accuracy(%)
BLSTM-E2E	6.27	3.94	3.55	3.52	2.97	80.15
SA-E2E	6.33	3.59	3.73	3.65	3.49	79.21
XSA-E2E	5.99	2.98	3.21	3.05	3.56	81.20



Multi-channel end-to-end Neural Diarization



- EEND using multi-channel signals from distributed microphones
- Transformer encoders in EEND replaced to process a multichannel input:
  - Spatio-temporal encoders
  - co-attention encoders
- Model adaptation method using only single-channel recordings.
- The method works on multi-channel inputs, such as in hybrid meetings

Shota Huriguchi, et.al., Multi-channel End-to-end Neural Diarization with Distributed Microphones, 2021.

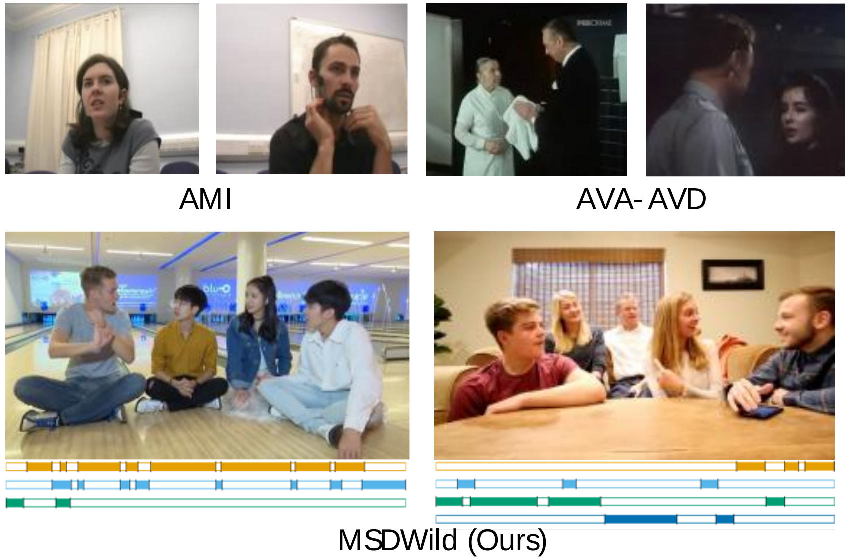
Why is this effort important?

Because we can use it as part of downstream tasks

- ASR
- Multi-microphone ASR
- Virtual assistants
- Broadcast transcriptions
- Emotion recognition
- …

Multimodal

Audio-Visual (dataset)



- MSDWILD: MULTI-MODAL SPEAKER DIARIZATION DATASET IN THE WILD
- Benchmark dataset
- Rich real world scenarios
- Different languages
- No overediting
- Overlap speech
- Cocktail party research
- Video and audio released

Tao Liu, et al., MSDWILD: MULTI-MODAL SPEAKER DIARIZATION DATASET IN THE WILD

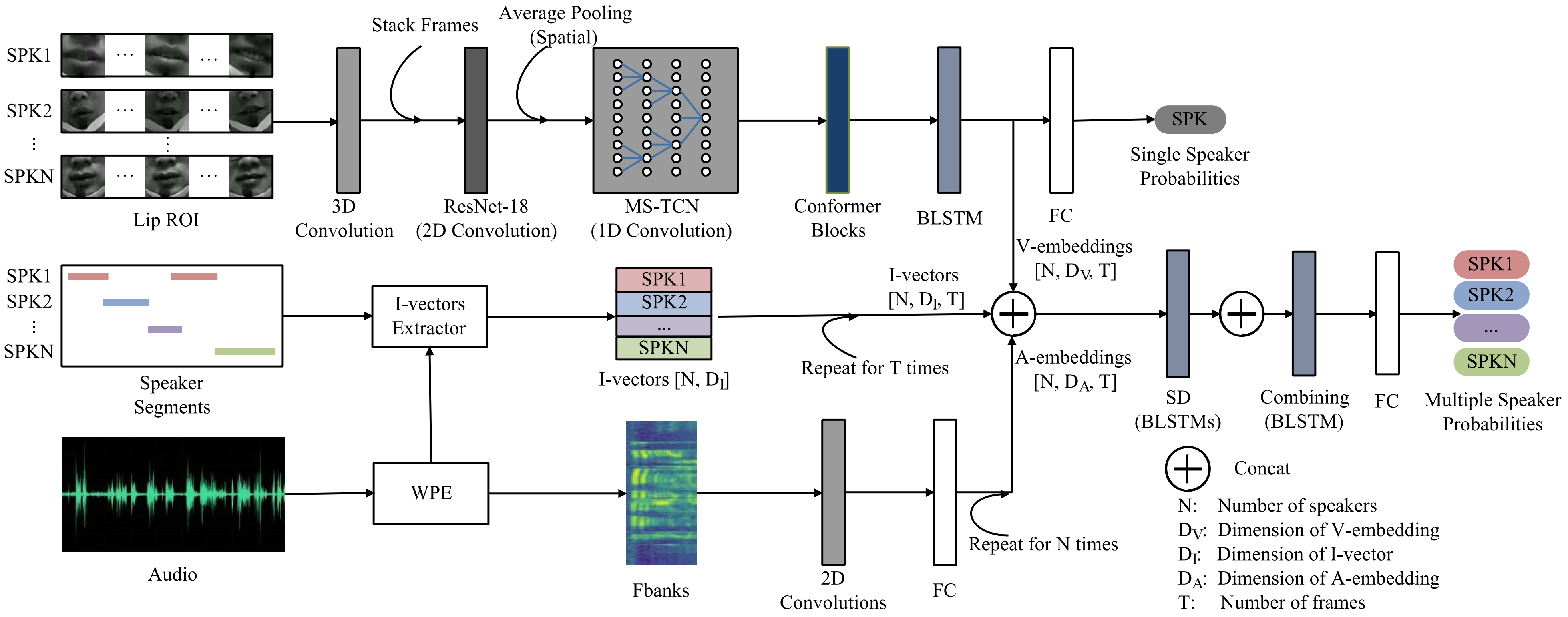
Audio-Visual (dataset)

Dataset	source	#videos	duration	Speech %	overlapped	#spk	#SC	noise	continuos	language
Voxconverse	TV-show	448	63h 50min	90.7	3.6	1/5.6/21	3.28	no	yes	en
MSDWild	Daily conversation	3143	80h 3min	91.29	14.01	2/2.7/10	11.8	yes	yes	multi

Method	DER	JER
Audio-only	43.15	84.28
Visual-only	53.71	62.71
Audio-visual (two-stream)	52.6	63.7
Audio visual (fused)	25.86	54. 79

Tao Liu, et al., MSDWILD: MULTI-MODAL SPEAKER DIARIZATION DATASET IN THE WILD

Audio-Visual (end-to-end)



Mao-kui He, et al., End-to-End Audio-Visual Neural Speaker Diarization

Audio-Visual (end-to-end)

- End-to-end Audio-visual diarization:
  - audio features, multi-speaker lip (ROI), ivectors.
- Classification output layers produce labels.
- Handle:
  - Overlap, speech vs non-speech
- I-vectors used for alignment
- MISP (eval)

Model	DER (w -VAD)	DER (w/o -VAD)
TS-VAD	28.95	-
VSD	13.07	19.64
Audio-visual AVSD w/i-vector	10.05	
Audio-visual AVSD i-vector	10.1	
Audio-visual AVSD i-vector+Joint training	9.49	10.99
Doverlap	8.85	-

Mao-kui He, et al., End-to-End Audio-Visual Neural Speaker Diarization

Takeaways

- EEND, VBx, TS-VAD are still good solutions.
- Unsupervised/self-supervised methods were proposed.
- Some methods deal with long recordings (podcasts).
- New methods are out of the traditional (sparse optimization, role labels)
- Online diarization is becoming feasible with good results.
- Some approaches are still on simulated data, some others are on real data

Research directions

- Multi-modal diarization (text, video, audio)
- Speaker imbalance
- Online diarization
- Real world data
- Long recordings (like children’s speech)
- We should consider speech separation algorithms that combined with diarization can improve the overall performance
- Diarization is only part of more complex scenarios, eg., using diarization for ASR

Why does it really matter?

Bilingualism in child-centered speech

- 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, VanDam Public Daylong HomeBank Corpus. doi:10.21415/T5388S,2018. https://media.talkbank.org/homebank/Public/VanDam-5minute/CI40/
Victoria Chua, Suzy Styles, et.al., Blip audio private collection provided by NTU, 2020.

Child centered data

Speaker Diarization

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

Language diarization

System (BLIP)	ACC(%)	EER(%)
xvector-LD	80	20
XSA-LD	89	11

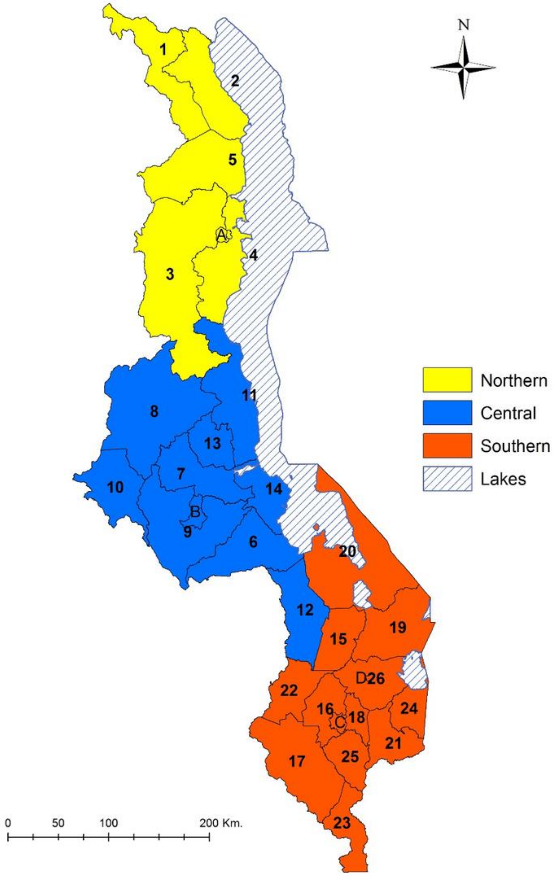
VanDam, Mark, VanDam Public Daylong HomeBank Corpus. doi:10.21415/T5388S,2018.
https://media.talkbank.org/homebank/Public/VanDam-5minute/CI40/
Victoria Chua, Suzy Styles, et.al., Blip audio private collection provided by NTU, 2020.
Liu, Hexin, BLIP data results, internal report, 2021.

Child development study

- The loop: ”*child and adolescent under-development, as existing programs struggle to deliver the right intervention at the right time.*” ☹
- Breaking the loop: ”*high-frequency data on child and youth develop-ment aided by customized tech-nologies to inform timely respons-es, tailored to the needs of each child and adolescent.*”

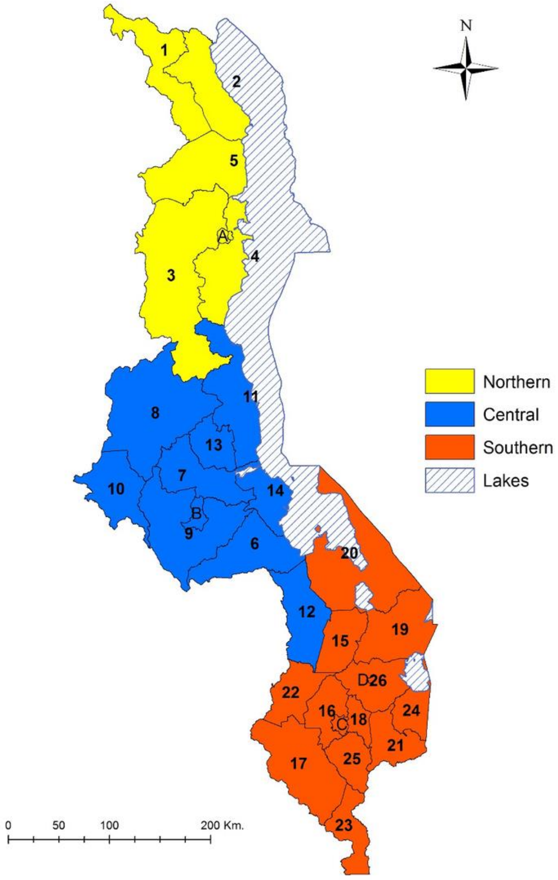
CCWD webpage: https://www.ccwd.uzh.ch/projects/2018_06_01-child_development_study.html

What are we doing?



- Do you know this country?
- Do you know the language?

What are we doing?



- Location: Malawi
- Language: Chichewa
- From the recordings, we plan to analyze aspects of adult-child and child-child communication such as:

diarization

  - Number and type (adult vs child) of speakers/interactants
  - Amount of adult speech
  - Amount of child speech
  - Number of conversational turns between the target child and other speakers (adults and children)
  - Timing of conversational turns

ASR

  - Number of different words (requires transcription)
  - Complexity of utterances (requires transcription)
  - Types of questions (requires transcription)

Takeaways

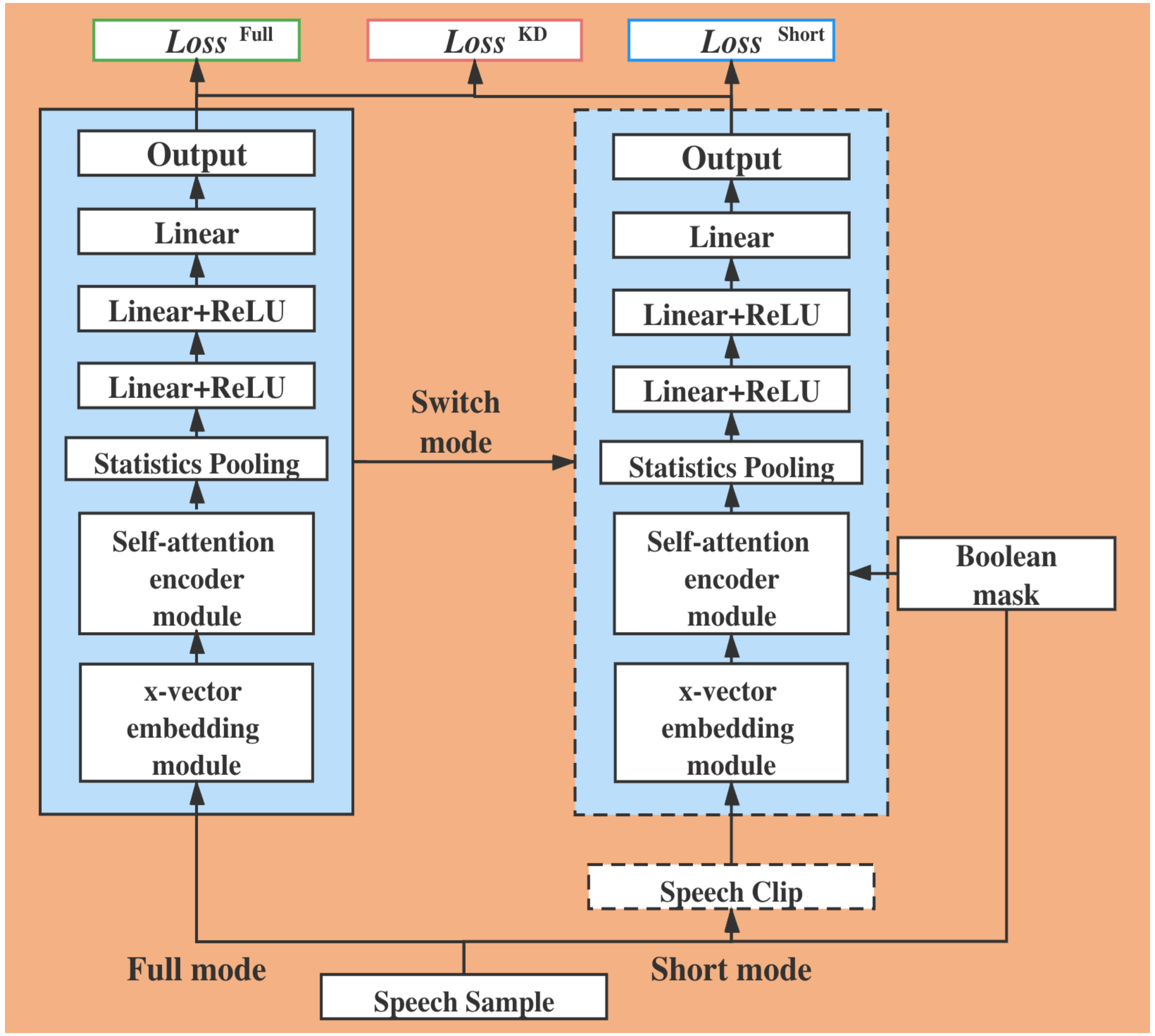
- Lots of flavors to choose from ☺
- We have huge improvements, but we are not yet there
- Neural diarization is becoming as good as embedding clustering methods
- Overlap detection is still an ongoing research
- Speaker imbalance is also an ongoing research
- Diarization to help downstream tasks (like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

Thank you!

Questions?

- Complementary slides if needed.

Dual Mode Language Identification



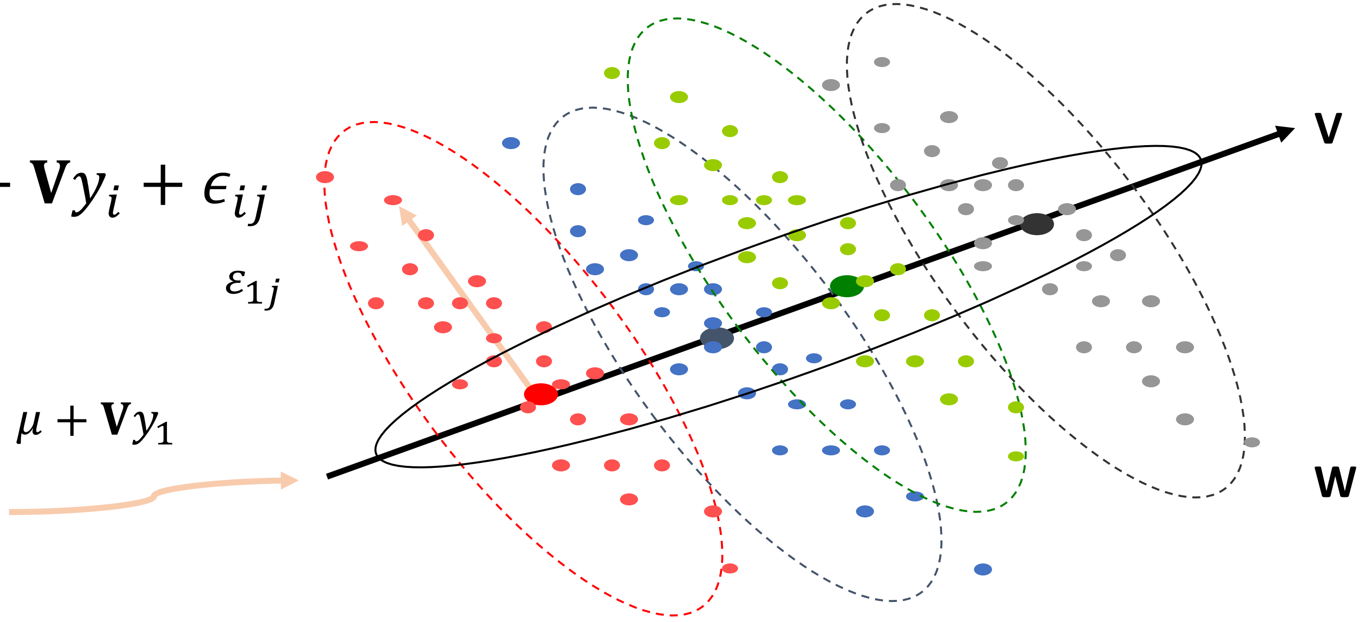
- Addressing short utterances
- XSA-LID model jointly optimizing
  - full-length speech
  - short clip (extracted by a specific Boolean mask)
- We apply knowledge distillation (KD) to boost the performance on short utterances.
- We investigate the impact of clip-wise linguistic variability and lexical integrity for LID.

Hexin Liu, et.al., Dual Mode Language Identification with Knowledge Distillation for Short Utterances , 2021.

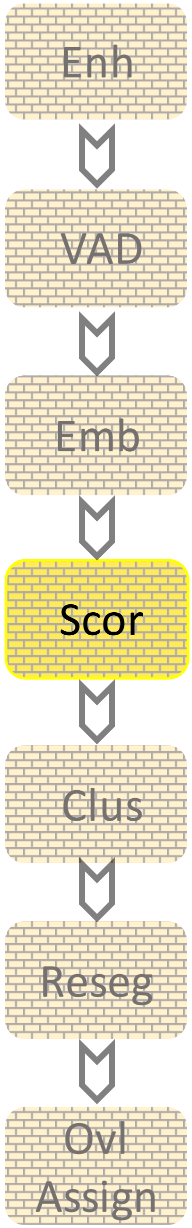
Scoring

- PLDA

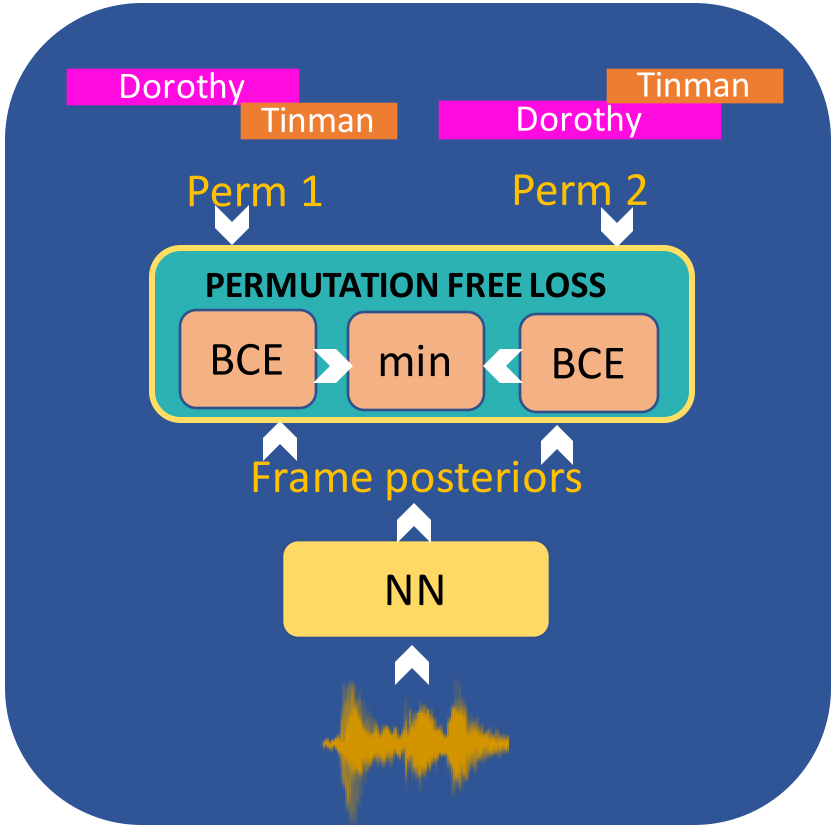
$$w_{ij} = \mu + \mathbf{V}y_i + \epsilon_{ij}$$



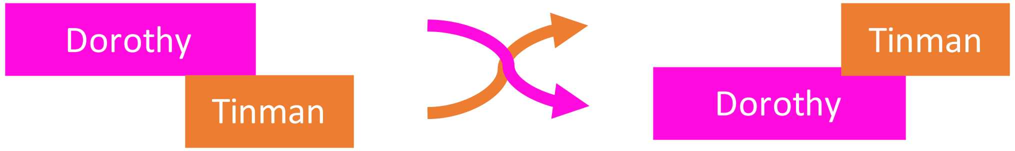
$$\mathrm{LLR} = \log \frac{P(w_1, w_2|same)}{P(w_1, w_2|diff)} = w_1^T A w_2 + \mathrm{w}_1^{\mathrm{T}}\mathrm{B}\mathrm{w}_1 + \mathrm{w}_2^{\mathrm{T}}\mathrm{B}\mathrm{w}_2 + \mathrm{C}^{\mathrm{T}}\mathrm{w}_1 + \mathrm{C}^{\mathrm{T}}\mathrm{w}_2 + \mathrm{D}$$



EEND

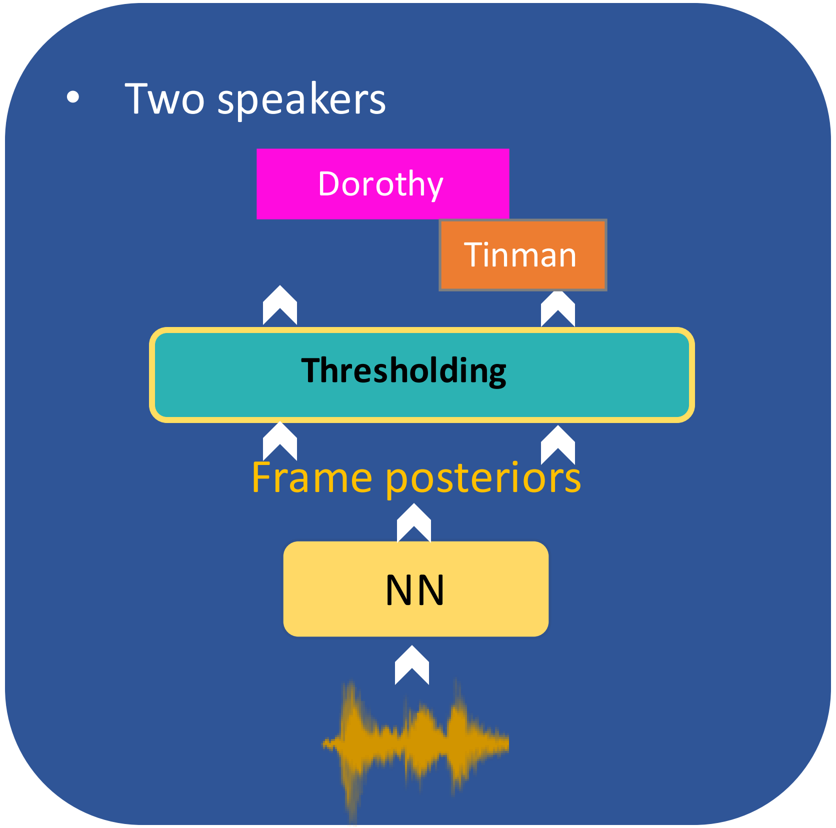


- End-to-end neural diarization
- Single network, supervised
- Two speaker case, proof of concept
- Handles overlapping speech!
- Training uses permutation invariant training (PIT) to prevent the labeling ambiguity.



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

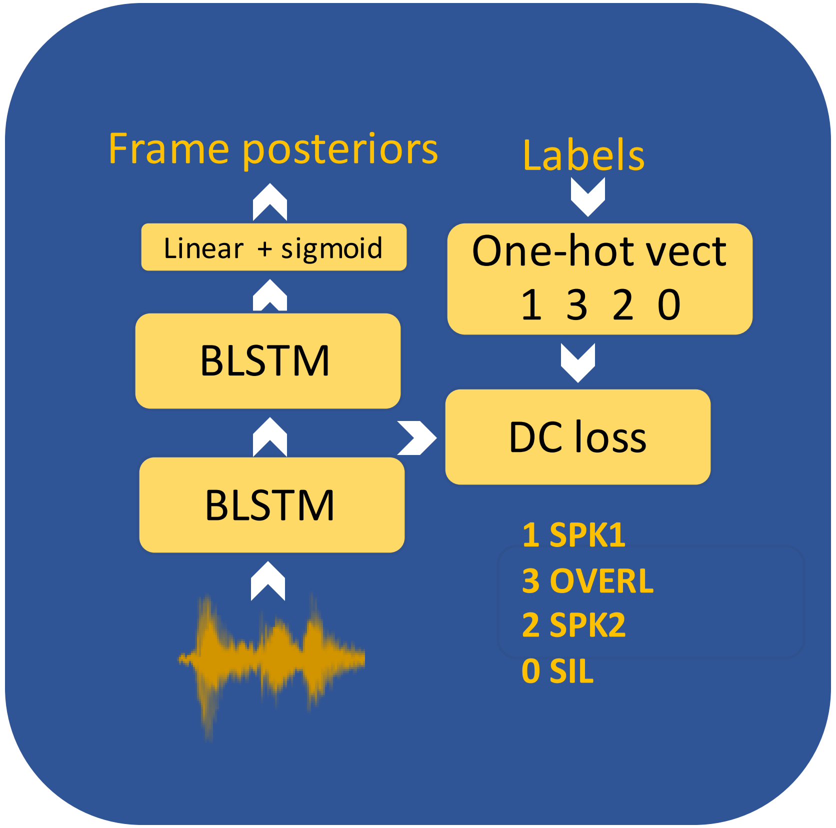
EEND



- 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

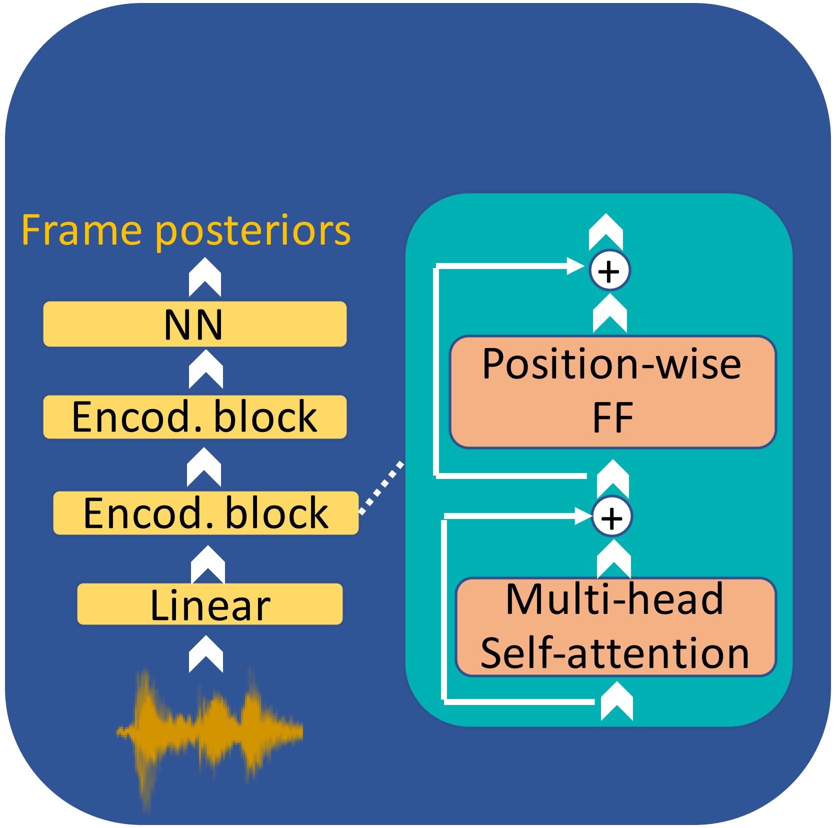
EEND



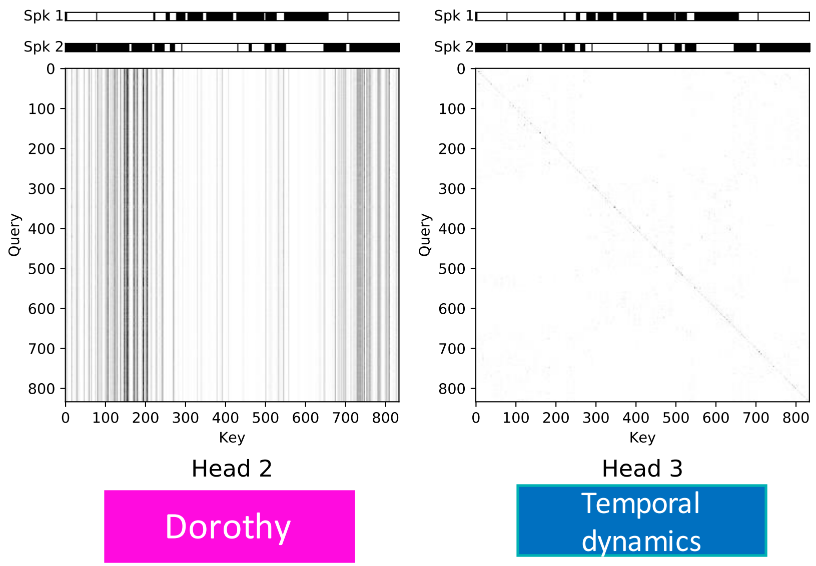
- BLSTM
- Embeddings from lower layers
  - Speaker training criterion on middle layer activations
- Deep Clustering to partition the embedding into:
  - Speaker dependent-clusters
  - Overlap
  - silence

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

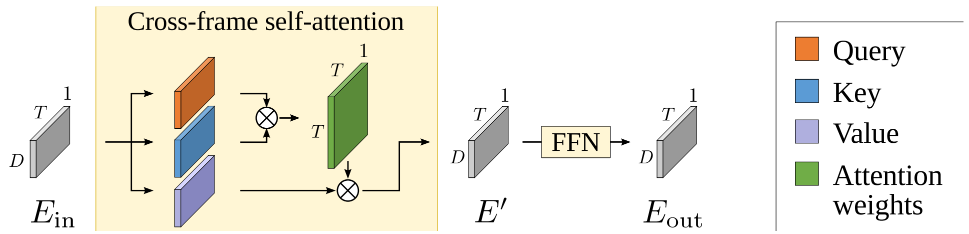
EEND-SA

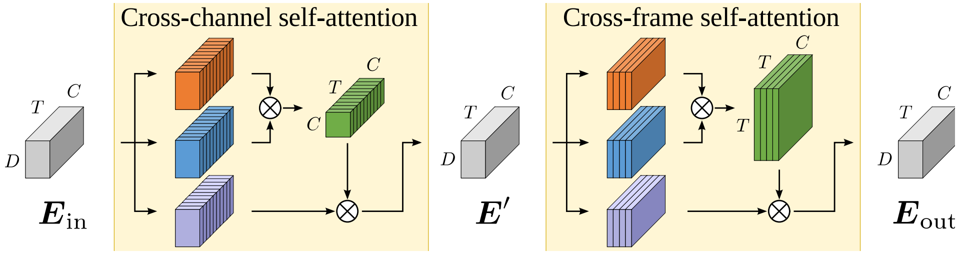


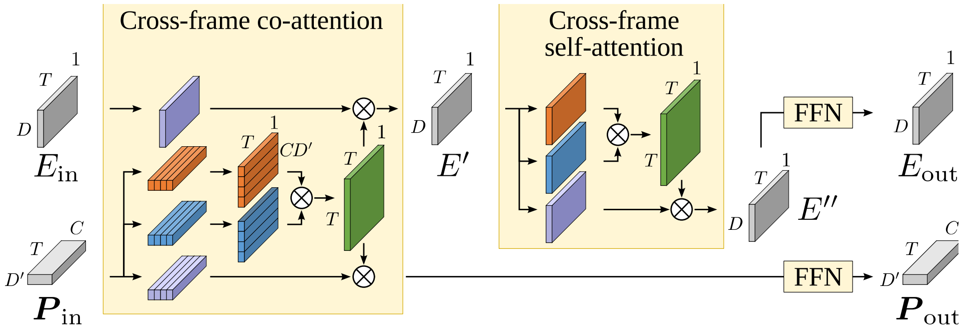
- BLSTM captures local temporal dynamics
- Self attention captures long context
  - The heads capture different characteristics



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


Cross-frame self-attention
FFN
Query
Key
Value
Attention weights
E_{in}
E'
E_{out}


Cross-channel self-attention
Cross-frame self-attention
$\boldsymbol{E}_{\mathrm{in}}$
$\boldsymbol{E}'$
$\boldsymbol{E}_{\mathrm{out}}$


Cross-frame co-attention
Cross-frame self-attention
FFN
FFN
E_{in}
$\boldsymbol{P}_{\mathrm{in}}$
E'
E''
E_{out}
$\boldsymbol{P}_{\mathrm{out}}$