LEAP Submission
for
Third DIHARD Diarization Challenge

Prachi Singh, Rajat Varma, Venkat Krishnamohan,
Srikanth Raj Chetupalli, Sriram Ganapathy,

Learning and Extraction of Acoustic Patterns (LEAP) Lab,
Electrical Eng., Indian Institute of Science, Bangalore.

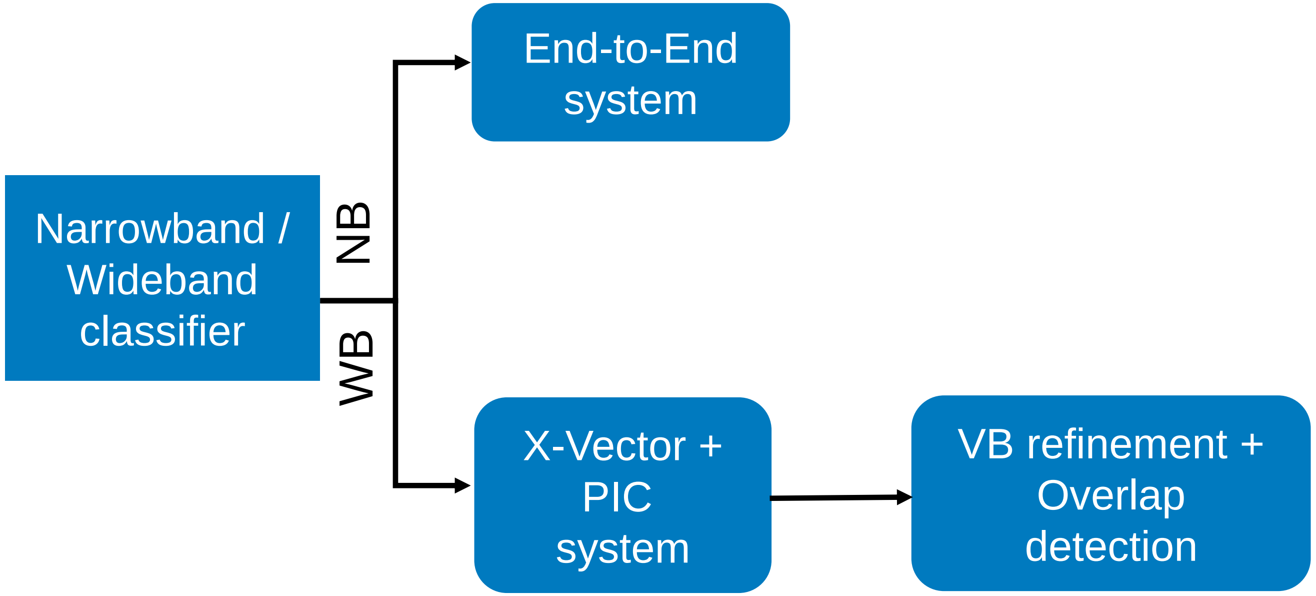
Outline

- Introduction
- LEAP systems Description
  - Narrowband-Wideband Classifier
  - Wideband PIC system
  - Narrowband End-to-End system
- Experiments & Results
- Conclusion

Introduction

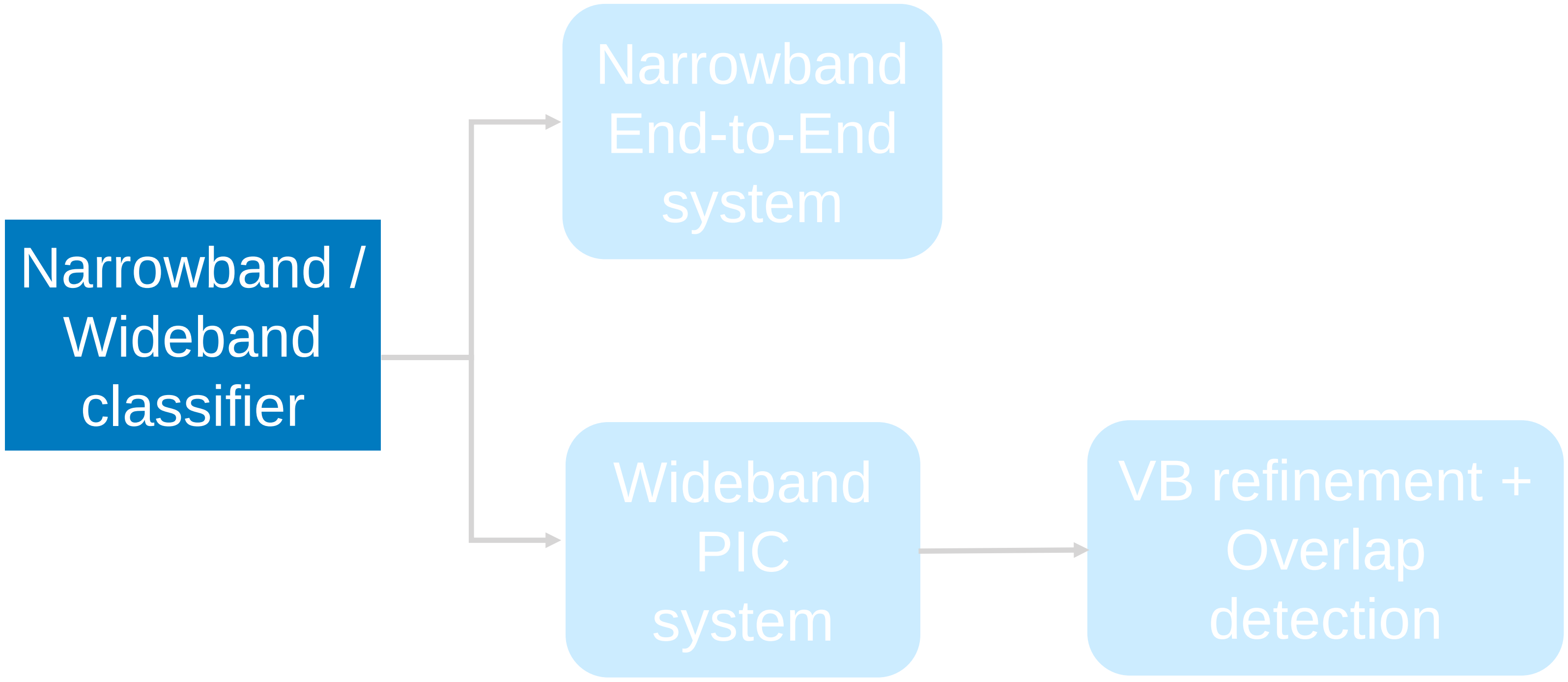
Introduction

- DIHARD III dataset has a mix of narrowband and wideband speech recordings
- In the dev set, 24% of the recordings are narrowband with two speakers per recording
- Proposal:
  - Classify recordings based on bandwidth: narrowband vs wideband
  - Combine models optimized for each band



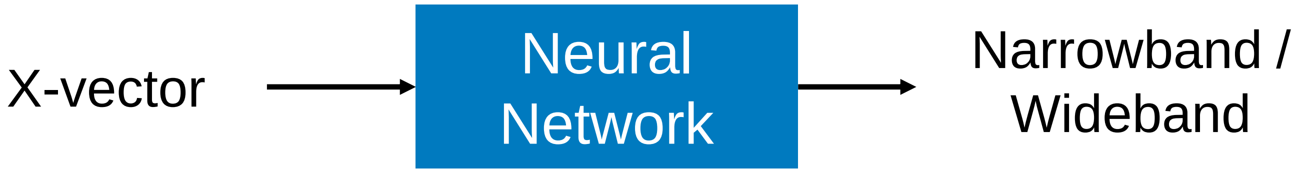
LEAP systems Description

Overall scheme

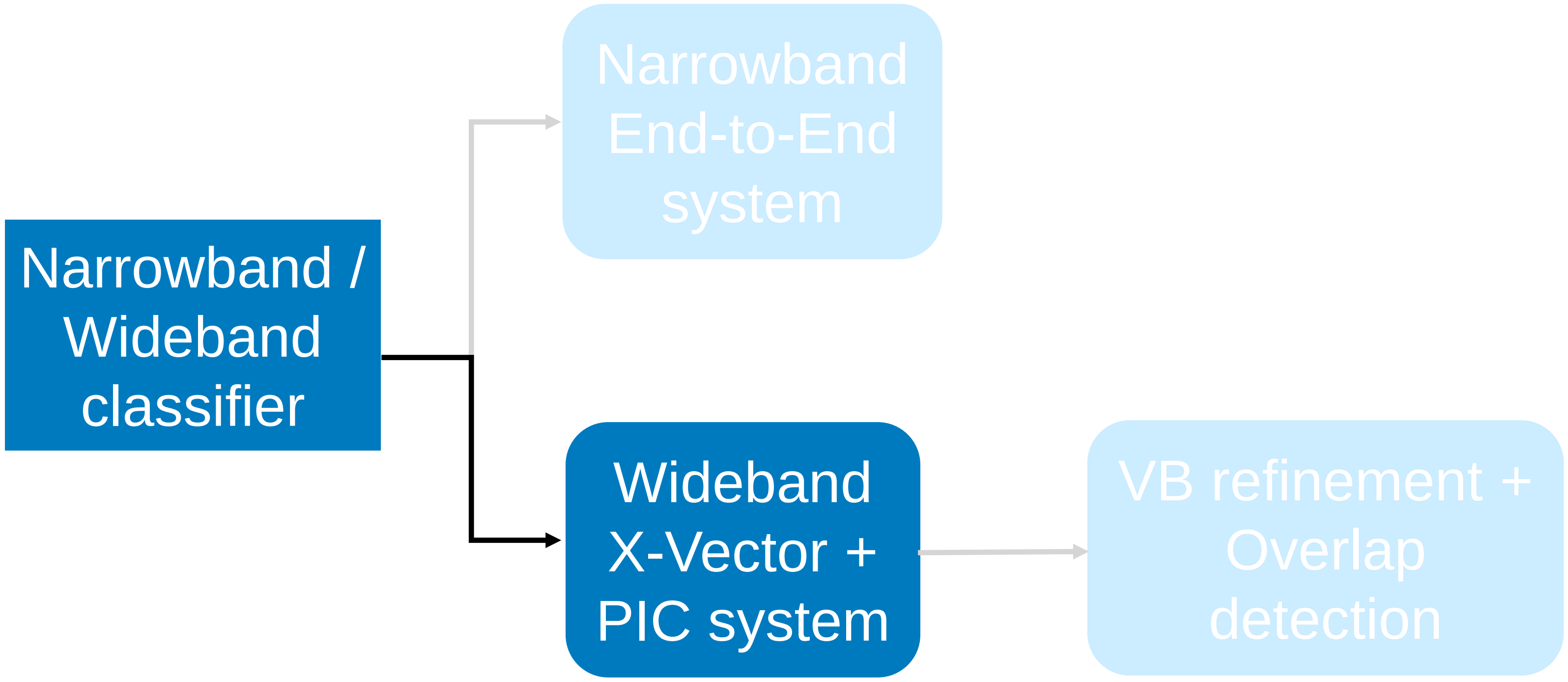


Narrowband-Wideband Classifier

- 2-layer NN with 512-d X-vectors, extracted every 5s using segments of duration 10s, as input features
- Output of the network classify between two bands using **majority voting** of the **segment-wise prediction**

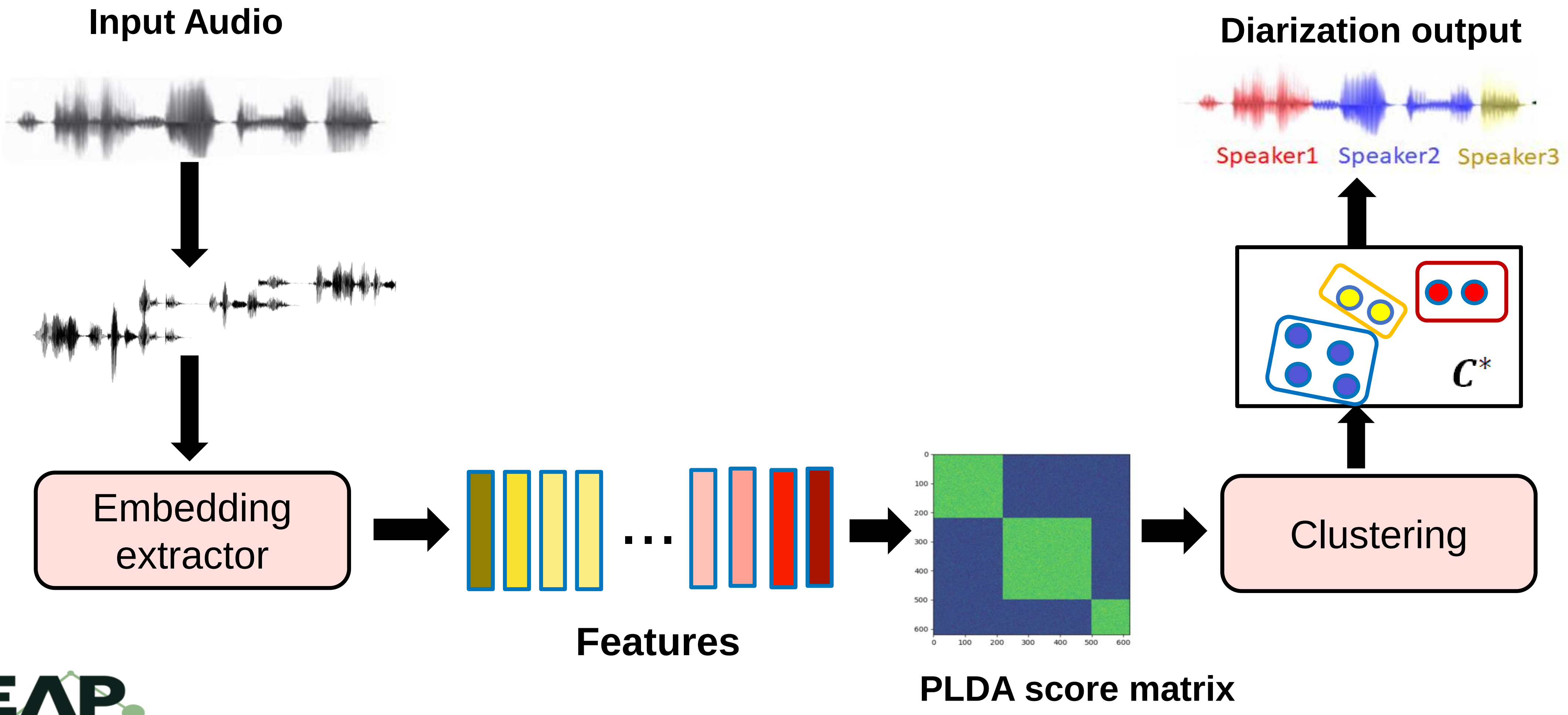


Overall scheme



Wideband x-vector PIC system

- Inspired by multi-stage baseline system shown below.



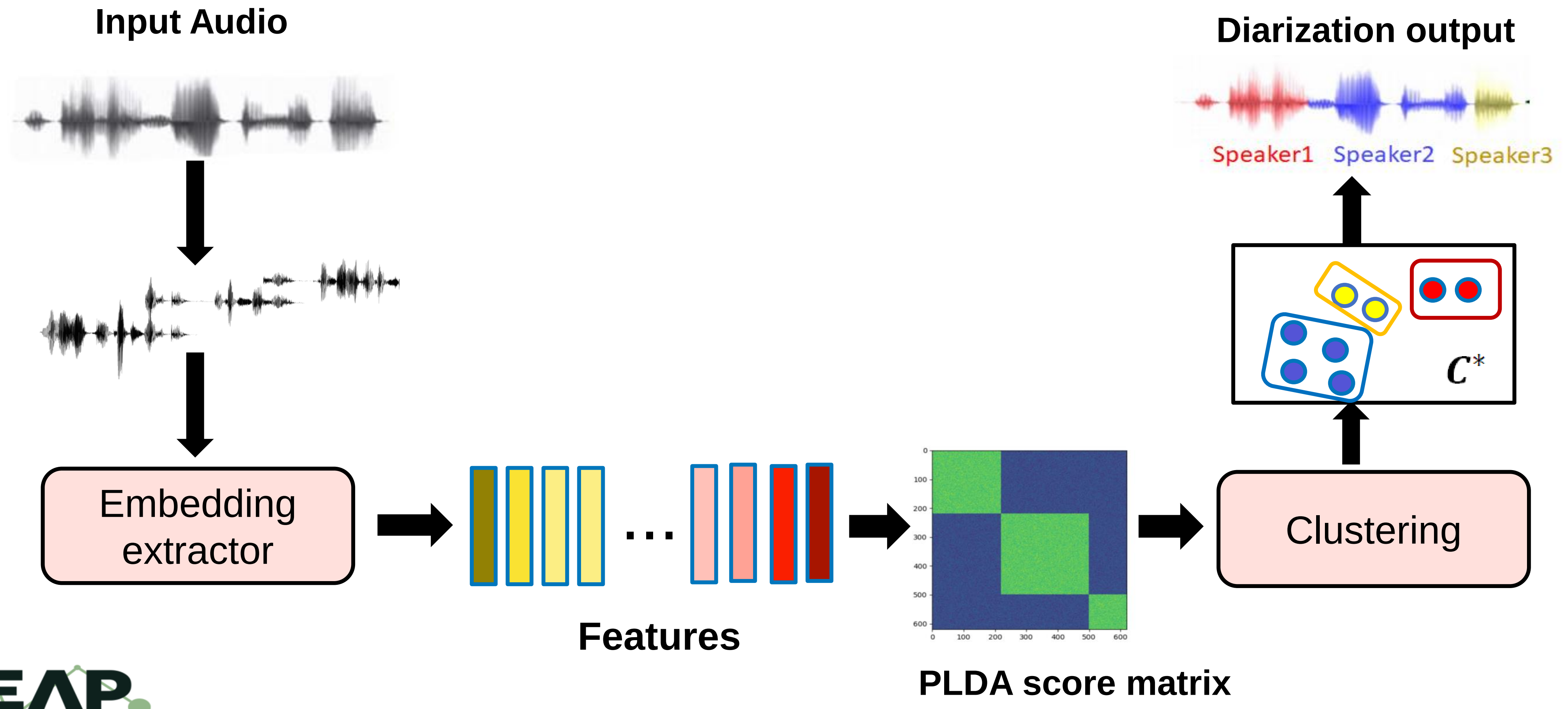
Embedding Extractors

ETDNN[1]	FTDNN[2]
• Extended-TDNN architecture has 13 layers • +/- 11 temporal context • 11th layer 512-d affine output are the x-vectors	• Factorized-TDNN architecture has 14 layers • +/-16 temporal context • replaced the pre-pooling layers by a factorized TDNN • 12th layer 512-d affine output are the x-vectors

[1]Snyder et. al., “Speaker recognition for multi-speaker conversations using x-vectors, ICASSP 2019
[2]Povey et. al., “Semi-orthogonal low-rank matrix factorization for deep neural networks”, INTERSPEECH 2018

Wideband x-vector PIC system

- Inspired by multi-stage baseline system shown below.



Path Integral Clustering (PIC)

- Graph-structural based agglomerative clustering algorithm where graph encodes the structure of the embedding space.

- Given a set of vectors $X = \{x_1, x_2, .., x_n\}$, it involves creation of directed graph G= (V,E)

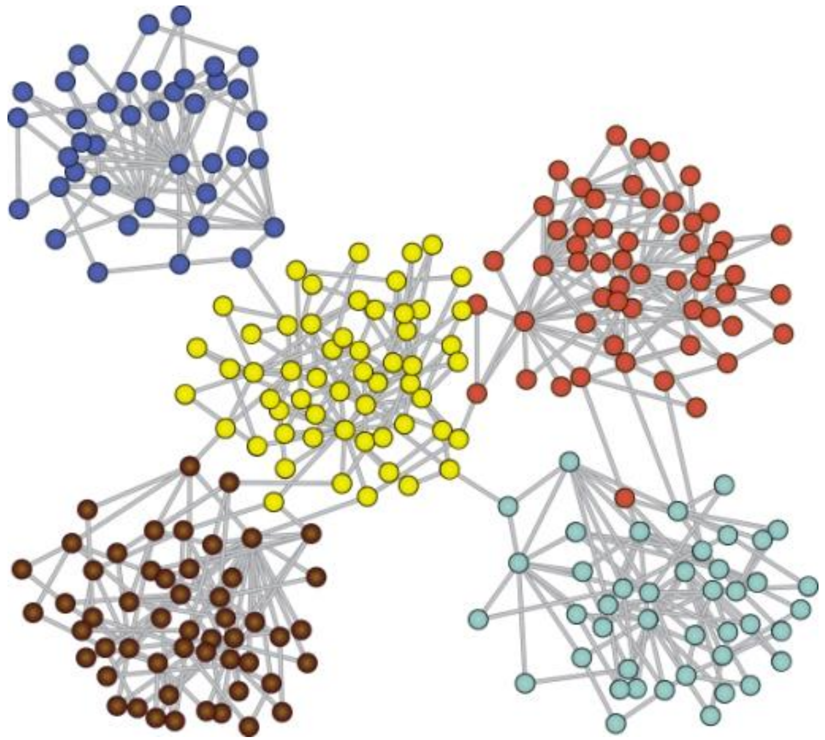
  - V is the set of vertices corresponding to the samples in X

  - E is the set of edges connecting vertices

  - Weighted Graph Adjacency matrix (W) given as,

$$w_{ij} = S(i,j) \; if \; x_j \, \epsilon N_i^K$$
$$= \quad 0 \quad otherwise$$

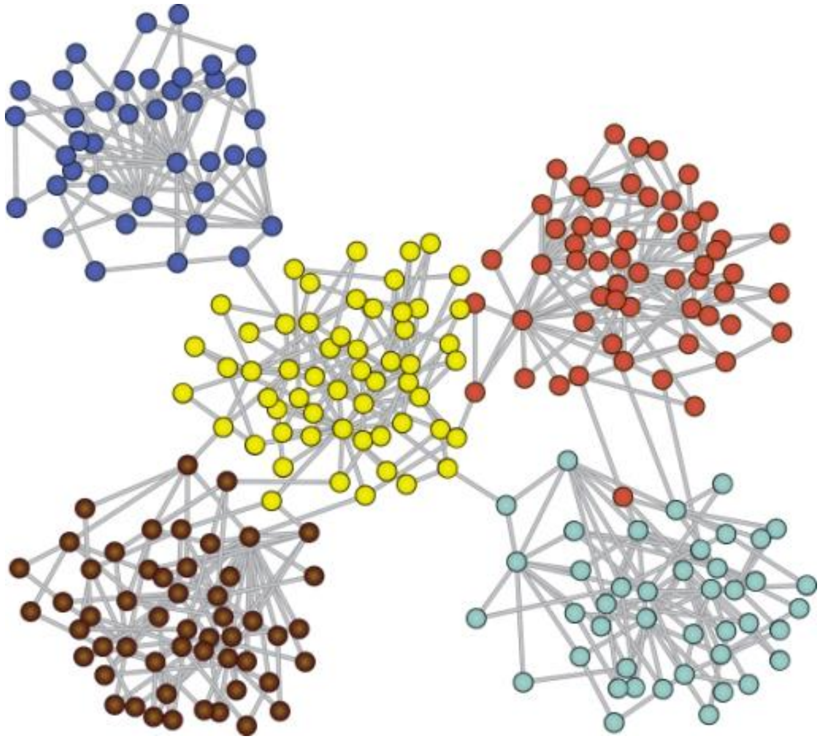
where, S(i,j) is the pairwise similarity between x_i and x_j, N_i^K is the set of K nearest neighbour of x_i



Zhang et. Al.. Agglomerative clustering via maximum incremental path integral. Pattern Recognition

Path Integral Clustering (PIC)

- Uses path integral as a structural descriptor of clusters
- Path integral is the sum of all possible paths of all possible lengths within each cluster
- High path integral indicates more stable cluster
- Encourages merging of cluster towards higher stability



Path Integral Clustering (PIC)

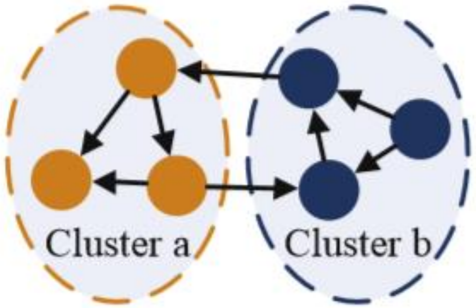
- Merges two clusters at each time step based on maximum affinity
- Affinity is computed as:

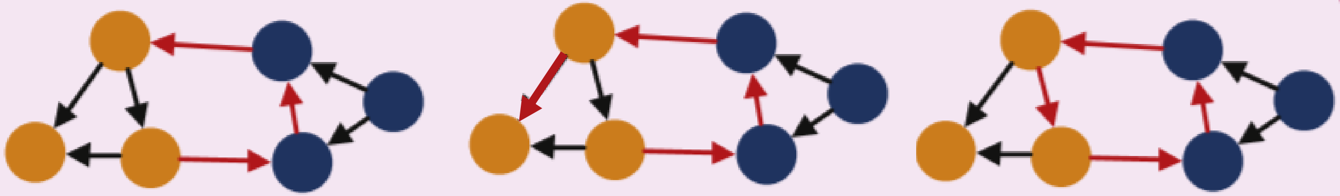
$$\mathcal{A}_{\mathcal{C}_a,\mathcal{C}_b} = (\mathcal{S}_{\mathcal{C}_a|\mathcal{C}_a\cup\mathcal{C}_b} - \mathcal{S}_{\mathcal{C}_a}) + (\mathcal{S}_{\mathcal{C}_b|\mathcal{C}_a\cup\mathcal{C}_b} - \mathcal{S}_{\mathcal{C}_b}).$$

  - $\boldsymbol{S_{C_a}}$: Path integral of cluster $\boldsymbol{C}_a$
  - $\boldsymbol{S_{C_a|C_aUC_b}}$: Conditional path integral of cluster $\boldsymbol{C}_a$ (Sum of all possible paths in $\boldsymbol{C}_a U \boldsymbol{C}_b$ such that starting and ending vertices must be within $\boldsymbol{C}_a$)
  - $\boldsymbol{S_{C_a|C_aUC_b}} - \boldsymbol{S_{C_a}}$: Incremental Path integral of $\boldsymbol{C}_a$

PIC illustration

$$\mathcal{A}_{\mathcal{C}_a,\mathcal{C}_b} = (\mathcal{S}_{\mathcal{C}_a|\mathcal{C}_a \cup \mathcal{C}_b} - \mathcal{S}_{\mathcal{C}_a}) + (\mathcal{S}_{\mathcal{C}_b|\mathcal{C}_a \cup \mathcal{C}_b} - \mathcal{S}_{\mathcal{C}_b}).$$

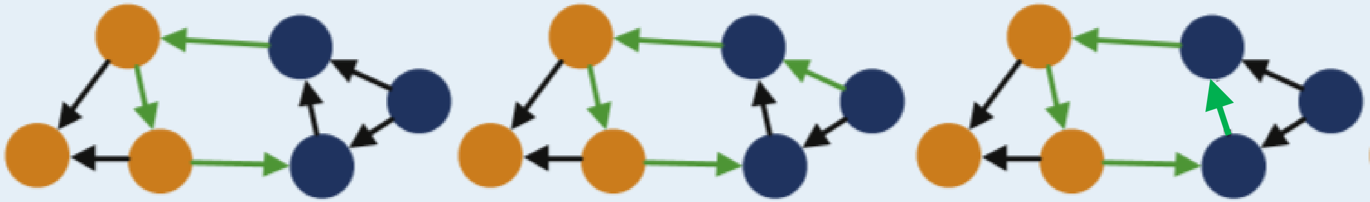




$= \boldsymbol{S_{C_a|C_a U C_b} - S_{C_a}}$

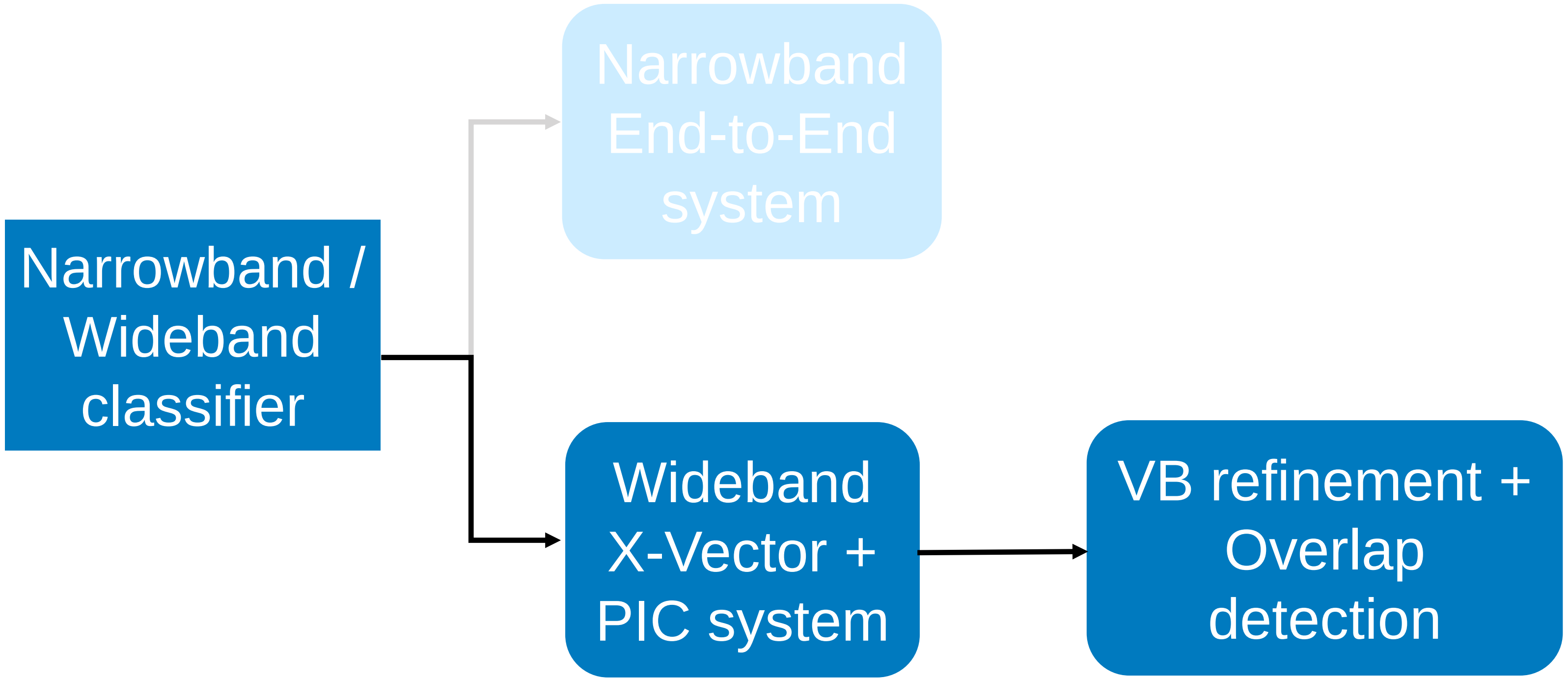
Incremental paths in Cluster a

$\boldsymbol{S_{C_b|C_a U C_b} - S_{C_b}} =$



Incremental paths in Cluster b

Overall scheme



VB refinement and Overlap detection (VB-overlap)

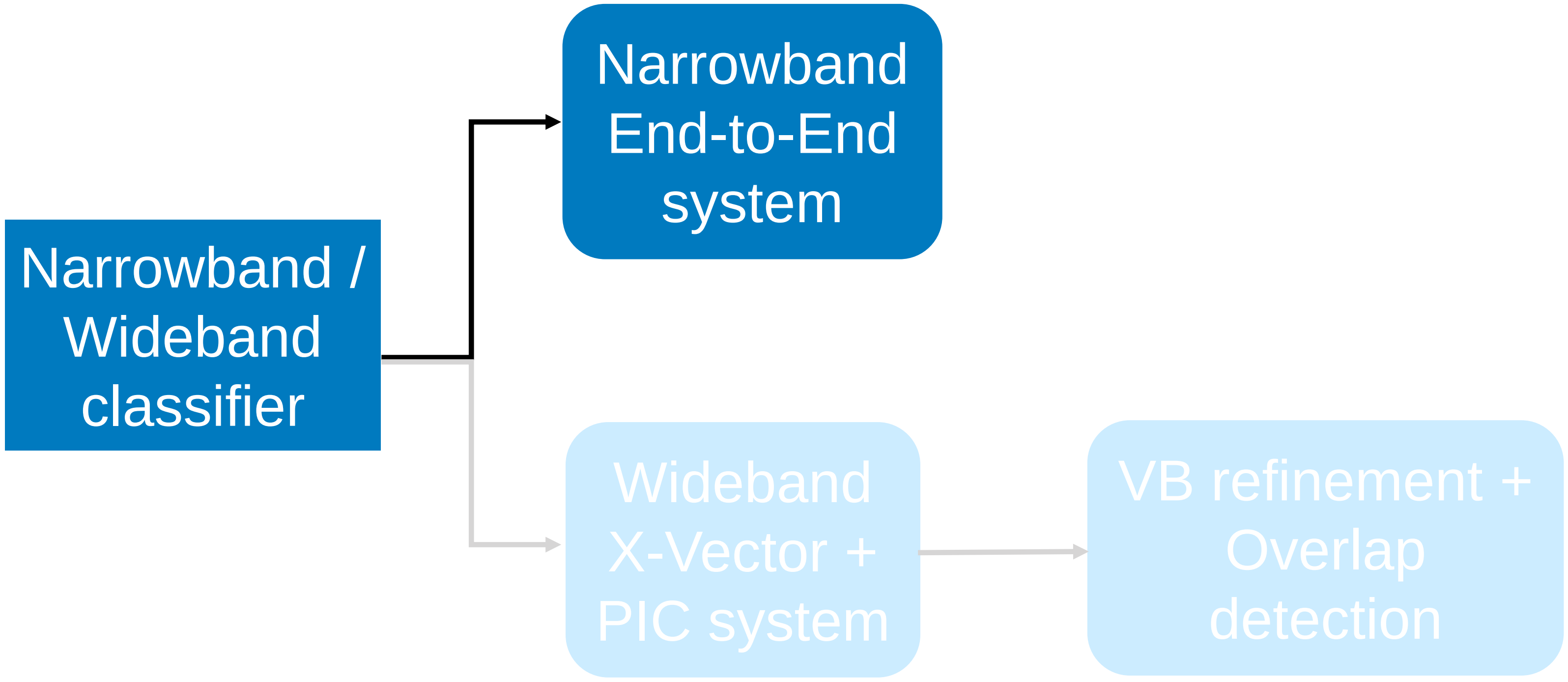
- Refinement of segment boundaries using **Variational Bayes Hidden Markov Model (VB-HMM)[1] with posterior scaling[2]**

- Overlap detection is done using module in **pyannote.audio python toolkit[3]**

- The segments identified as overlap by the detector are then used to refine the segments obtained after VB-HMM based on posteriors

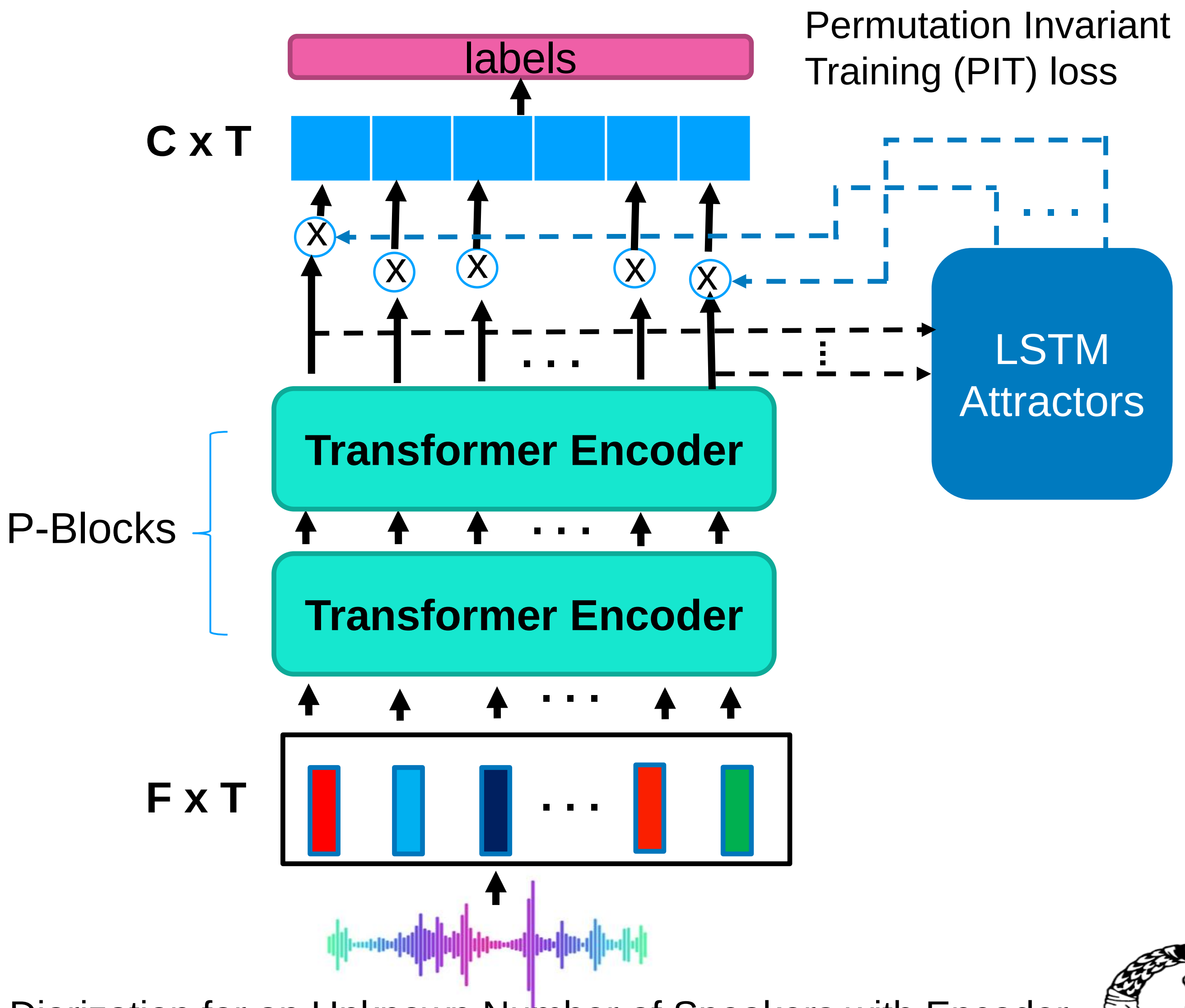
[1]Diez et. al., "Speaker diarization based onBayesian HMM with eigenvoice priors," Odyssey, 2018
[2]Singh et. al., "LEAP diarization system for the second DIHARD challenge," INTERSPEECH 2019
[3]Bredin et. al., "pyannote.audio: neural building blocks for speaker diarization," ICASSP 2020

Overall scheme



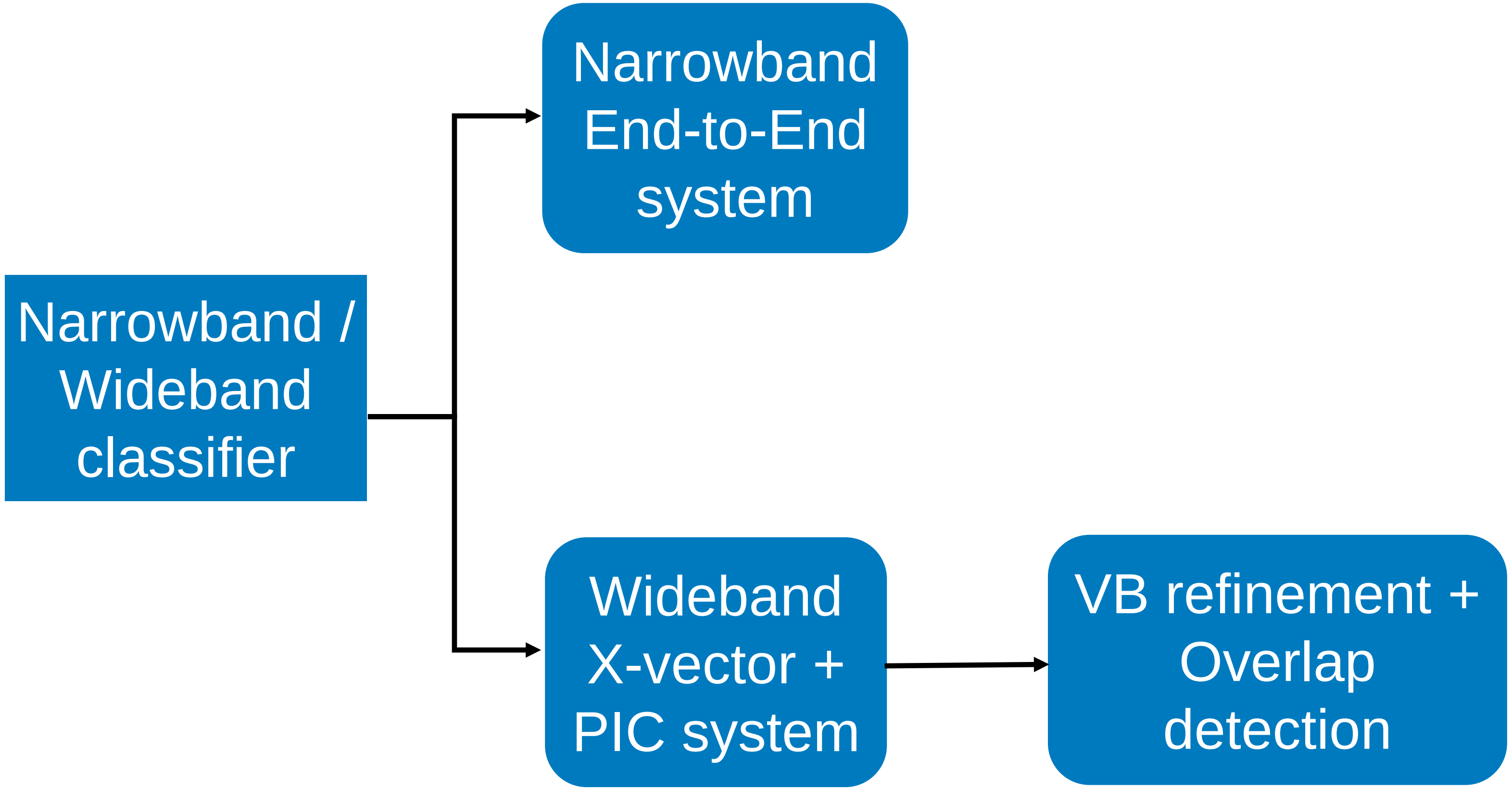
Narrowband End-to-End system

- The architecture of the model is like the **SA-EEND** with the encoder-decoder based **attractor** calculation **(EDA)**[1]
- Input is **23-d log-Mel-filterbank features** with a context of **+/- 7 frames**
- The model uses **4 stacked Transformer encoder blocks**; each block consists of **256 attention units with 4 attention heads**
- Trained using **permutation-invariant** (PIT) loss



[1]Horiguchi et. al.,"End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors

Overall scheme



Experiments & Results

Training

- **Embedding extractors:** ETDNN and FTDNN both are trained using Voxceleb1 and Voxceleb2 datasets for speaker classification of 7,323 speakers.
  - A separate PLDA model is trained for ETDNN and FTDNN using subset of x-vector training set.
- **SA-EEND:** Simulated 100,000 two-speaker mixtures from Switchboard-2 (Phase I, II, III), Switchboard Cellular (Part 1, Part 2) and NIST SRE datasets 2004-2008
  - Trained for 100 epochs using permutation-invariant (PIT) training criterion
  - As narrowband contains 2-speaker telephone recordings from fisher dataset, we adapted the model on CALLHOME subset containing 2-speaker files

Evaluation

- **Wideband x-vector PIC system (WPS):** For both tracks we perform following steps:
  - Extract x-vectors from 1.5s segments with 0.25s of shift.
  - We consider (i) cosine scores (ii) PLDA scores, to compute similarity between segments.
  - To compute the scores, we follow the same pre-processing steps (whitening transform + length norm + recording level PCA) as used in baseline setup.
  - We experiment with different clustering techniques like AHC, path integral clustering (PIC).
  - The number of speakers is generated using the PLDA+AHC threshold fine tuned over the dev set.
  - For Track2, we use the pre-trained SAD model from baseline setup to generate speech segments.

Evaluation

- **Narrowband End-to-End system (NES):**
  - We down sample audio to 8KHz and pass it to model to generate frame wise posteriors.
  - We subsample the frame level features by different factors to avoid abrupt speaker change and reduce memory computation.
  - Predict at least one speaker based on maximum posterior probability.
  - Apply threshold on posteriors to detect presence of more than one speaker.
  - Remove the silence frames using the ground truth SAD for track1 and pre-trained SAD for track2.

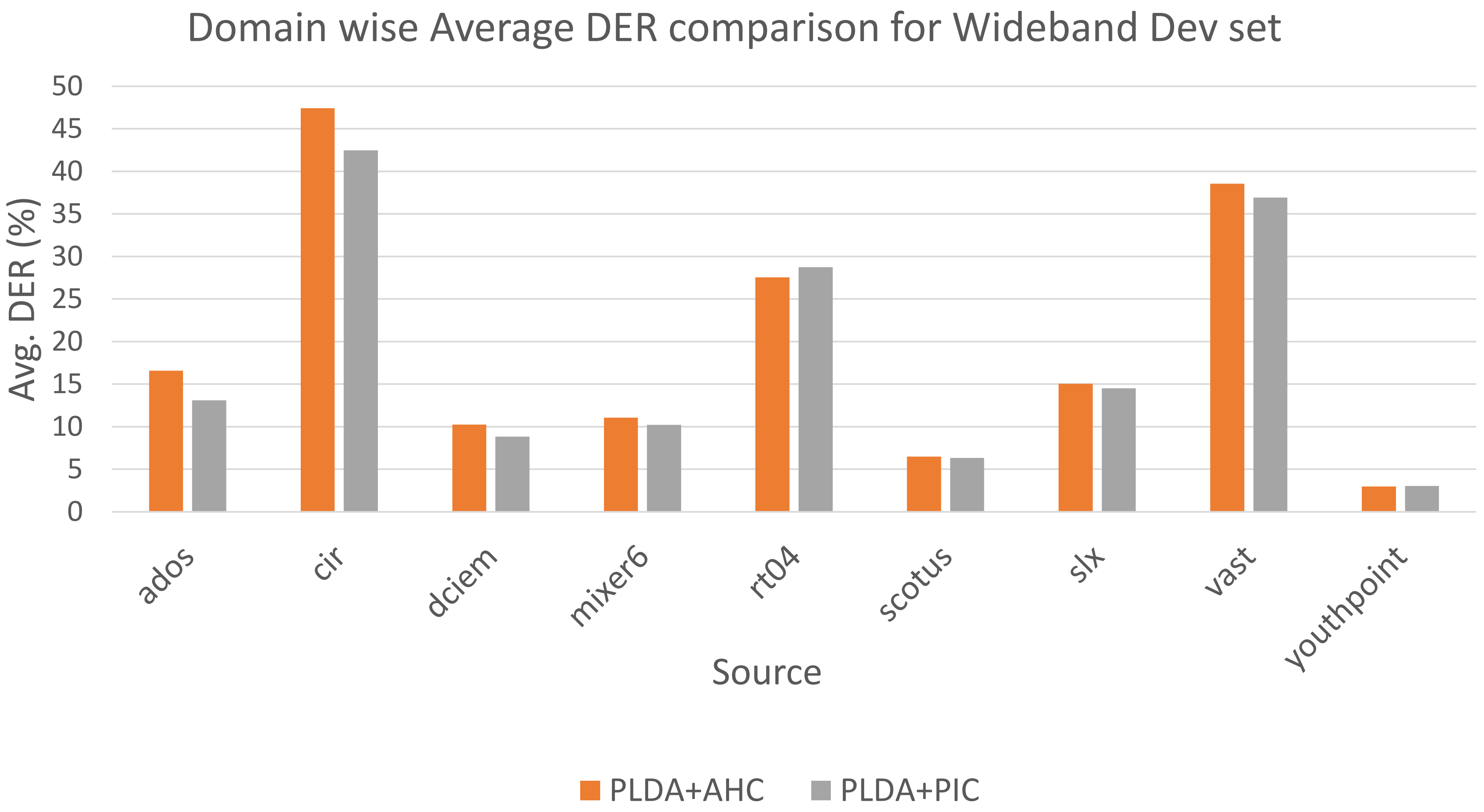
Results – Track1 Dev

Wideband ETDNN System config.	Dev DER (JER)
PLDA + AHC (S1)	20.09 (43.86)
PLDA + PIC (S2)	**19.06 (42.44)**
Cosine + PIC	19.78 (43.61)
Baseline (with VB)	19.95 (44.94)
S1 + VB-overlap	17.70 (42.93)
S2 + VB-overlap	**17.03 (41.92)**

Narrowband System config.	Dev DER (JER)
Baseline*	16.03 (20.21)
SA-EEND V1	9.84 (12.00)
SA-EEND V2	**9.34 (11.19)**

*baseline with oracle number of speakers, V1 = subsampling by 10, V2= subsampling by 5

AHC vs PIC (Domain wise)



Results – Track1 Systems

System config.	Set	Dev DER (JER)	Eval DER (JER)
Baseline[1]	Full	19.10 (41.10)	19.68 (44.32)
	Core	19.97 (45.52)	21.35 (48.89)
WPS (ETDNN) + NES	Full	**14.45 (37.09)**	**14.93 (37.09)**
	Core	16.43 (42.45)	18.2 (43.28)
WPS (FTDNN) + NES	Full	**14.34 (37.31)**	**14.88 (36.73)**
	core	16.26 (42.75)	18.07 (42.82)

WPS=Wideband PIC system, NES= Narrowband End-to-End system

[1]Ryant et. al., "The Third DIHARD Diarization Challenge," 2020

Results – Track2 Systems

System config.	Set	Dev DER (JER)	Eval DER (JER)
Baseline[1]	Full	21.35 (42.97)	25.76 (47.64)
	Core	22.31 (47.28)	28.31 (52.44)
WPS (ETDNN) + NES	Full	**16.77 (37.15)**	**21.04 (39.68)**
	Core	18.64 (41.93)	24.92 (45.32)
WPS (FTDNN) + NES	Full	**16.53 (38.50)**	**21.09 (39.54)**
	core	18.34 (43.62)	24.99 (45.13)

WPS=Wideband PIC system, NES= Narrowband End-to-End system

[1]Ryant et. al., "The Third DIHARD Diarization Challenge," 2020

Conclusion

- Proposed a combination of narrowband and wideband diarization system
- End-to-End system is found to perform better for two speaker narrowband recordings.
- The baseline framework is optimized for wideband recordings, using better speaker space embeddings (ETDNN, FTDNN), and novel path integral clustering scheme.
- ETDNN and FTDNN are found to have similar performance, and system combination at the score level improves the overall DER only marginally.

Thank you !

Questions ?

email: prachisingh@iisc.ac.in