Graph Clustering approaches for Speaker Diarization of Conversational Speech

PhD Thesis Defense
6th February 2024

Prachi Singh
PhD Student,
Learning and Extraction of Acoustic Patterns (LEAP) Lab,
Electrical Engineering, Indian Institute of Science, Bangalore.

Advisor: Dr. Sriram Ganapathy

LEAP

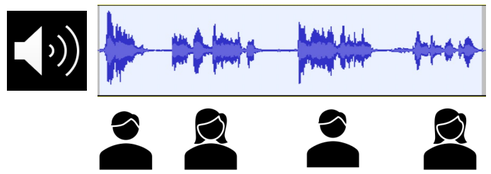
Outline

- Introduction
  - Motivation
  - Methodology
  - Contributions
- Background study
  - Related work
  - Performance metrics
  - Datasets
- Proposed Graph Clustering approaches
- Conclusion and Future Directions

Introduction

Motivation

Transcribing audio into text using speaker information generates much meaningful text



Hello

Hello. How are you Nitin?

I am doing great. How are you Meenu?

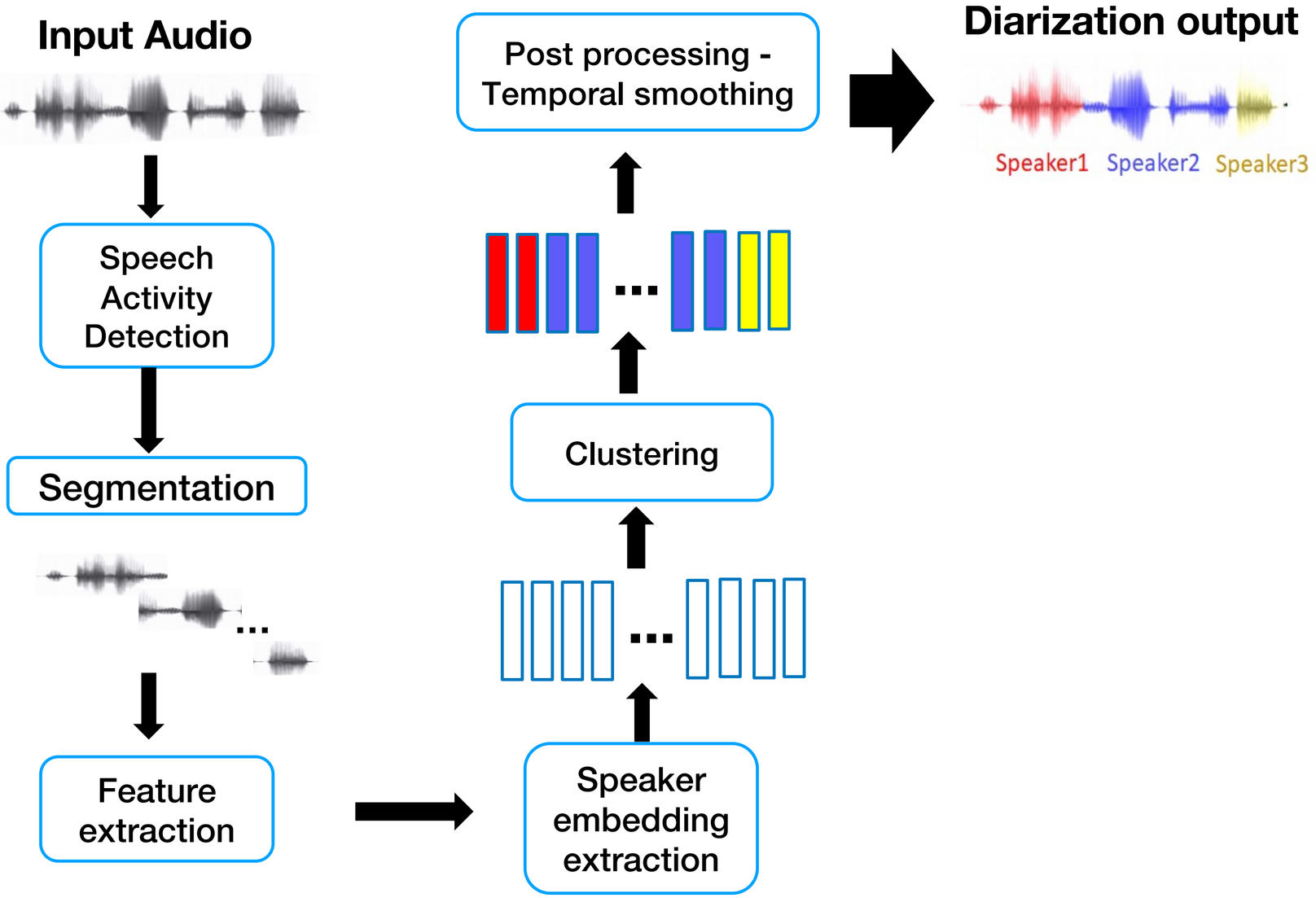
I am doing also great.

The task of finding "**who spoke when**" is called **Speaker Diarization**.

Transcribing meeting

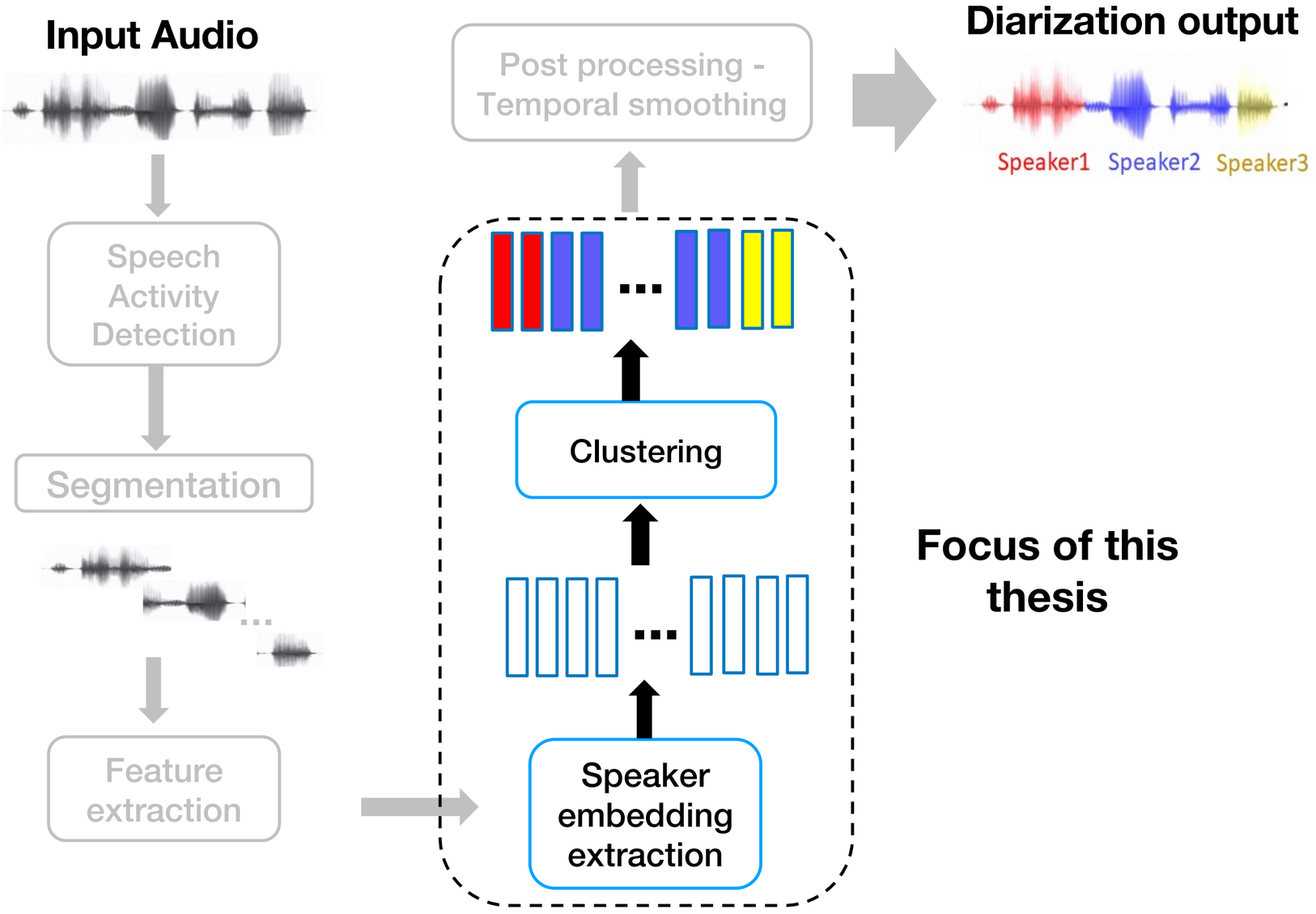
Call center interactions Analysis

Methodology



Sell et al., Diarization is hard: some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge, 2018.

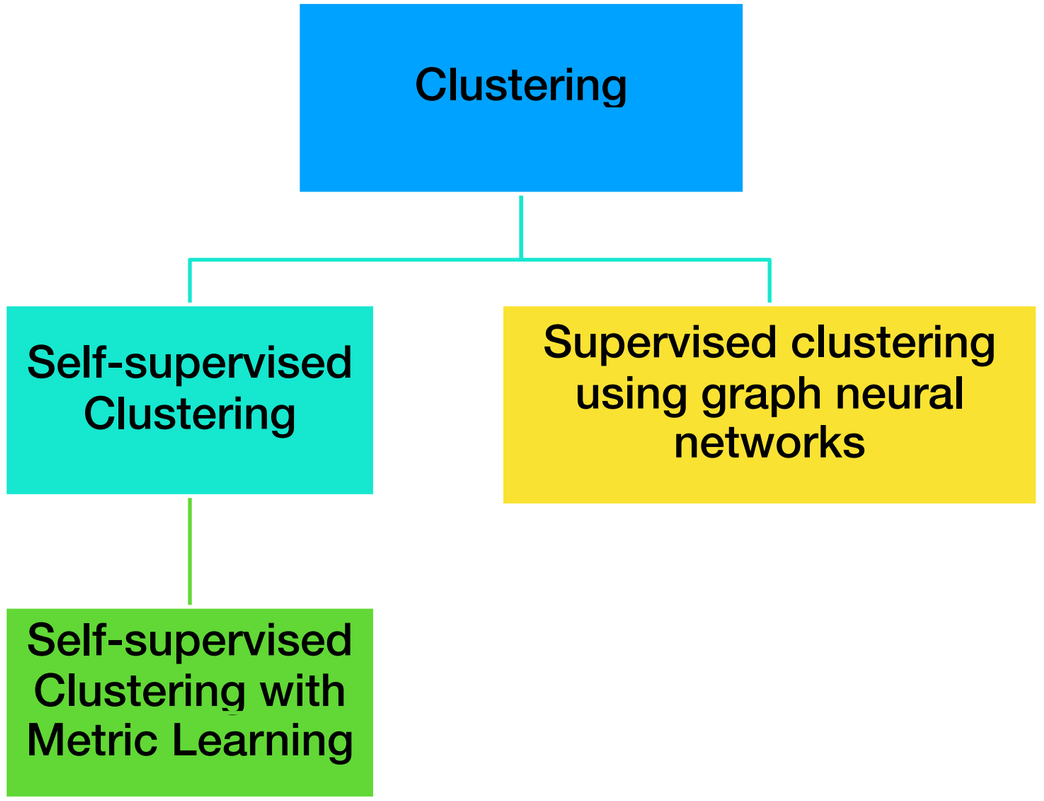
Contributions outline



Contributions outline

Application of **graph models to temporal segmentation of speech is the first of its kind.**

- Novel hierarchical graph clustering
- Self-supervised metric learning to generate similarity for clustering
- Supervised hierarchical graph clustering



Background study

Related work

Unsupervised Clustering approaches

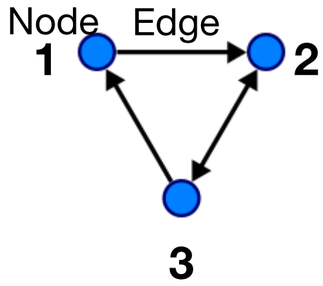
Forming groups based on hidden patterns in the unlabeled data

- Hierarchical clustering - Agglomerative Hierarchical Clustering (AHC)
- Graph Clustering

Clustering approaches

Graph

- A graph *G* can be well described by the set of **vertices** *V* and **edges** *E* it contains. G=(V,E)
- The **vertices** are often called **nodes.**



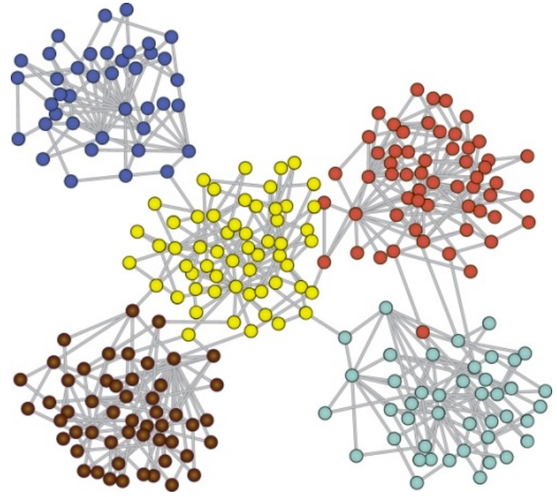
- **Adjacency matrix (A)** captures connections between nodes,
- $A_{ij} = 1, if\ Node\ i\ is\ connected\ \ to\ j\ by\ an\ edge$
- $A_{ij} = 0, if\ Node\ i\ and\ j\ are\ not\ connected$
- A with real weights to the edges is called as weighted adjacency matrix.

$$A=\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

Graph clustering

Clustering the nodes such that **many edges** are present **within each cluster** and **fewer edges between the clusters**.

Example: Spectral Clustering (SC)



Recent Graph Approaches

- **Graph attentional/convolution encoder (GAE)[1]** based approach for metric learning followed by spectral clustering.

- **Graph Attention- Based Deep Embedded Clustering (GADEC)[2] :** Graph attention-based clustering using multi-objective training.

[3]Wang et al., ICASSP, 2020
[2]Wei et al., Speech Communications, 2023

Related work

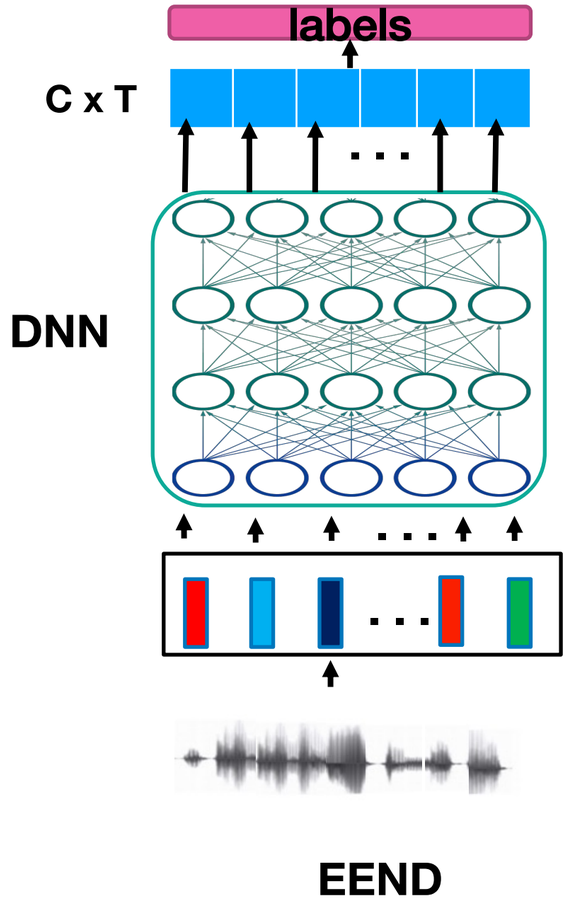
- Speaker embeddings/representations
  - i-vector[1] – statistical model
  - d-vector[2] – Deep Neural Network
  - **x-vector[3]** –Time delay Neural Network (widely used)
- Similarity measure
  - Cosine[4]
  - **PLDA[5]** (widely used)

[1]Dehak et al., 2011, [2]Variani et al., 2014, [3]Snyder et.al.,2018
[4]Senoussaoui et al., 2014, [5]Sell and Garcia-Romero, 2014, [5]Sell et al., 2018

Related work

End to end neural diarization (EEND)[1]

- Transformer is used to perform speaker activity detection
- Takes input as F-dimensional audio features and generate C speaker labels



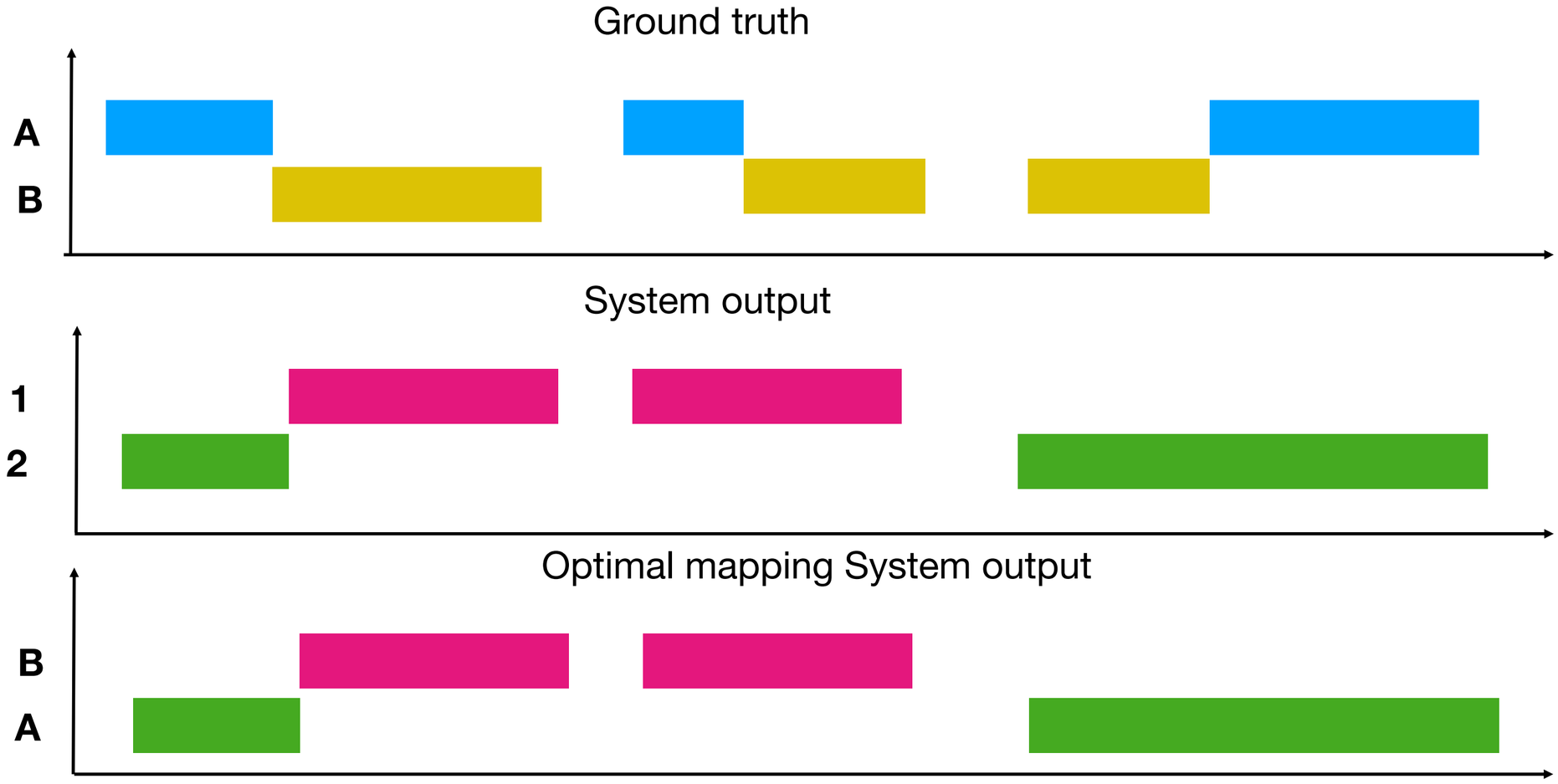
Cons:

- Requires huge amount of labelled data for training.
- Difficult to generalize for higher number of speakers.
- Cannot handle long duration recording at a time.

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

Performance metric

Optimal mapping: $\text{argmax}(A \cap 1, A \cap 2)$, $\text{argmax}(B \cap 1, B \cap 2)$



Performance metric

- **Diarization error rate (DER)** is the standard metric for evaluating and comparing speaker diarization systems.

- It is defined as follows:

$$DER = \frac{false\ alarm + miss\ detection + \boldsymbol{speaker\ confusion}}{total\ speakers\ duration}$$

  - $false\ alarm$ - duration of non-speech predicted as speech
  - $miss\ detection$ - duration of speech of a speaker predicted as non-speech
  - $\boldsymbol{speaker\ confusion}$ – duration of a reference speaker predicted as another speaker in system output after optimal mapping
  - $total\ speakers\ duration$ – total duration of all the speakers present

Test Datasets

Narrowband (sampling rate: 8kHz)

Wideband (sampling rate: 16kHz)

CALLHOME (CH) [1]

- Telephone data
- #Recordings - 500
- Duration: 2-5 mins
- 2-7 speakers

AMI [2]

- Meeting dataset
- #Recordings - Dev set: 18, Eval set: 16
- Duration: 20-60mins
- 3-5 speakers

DIHARD III [3]

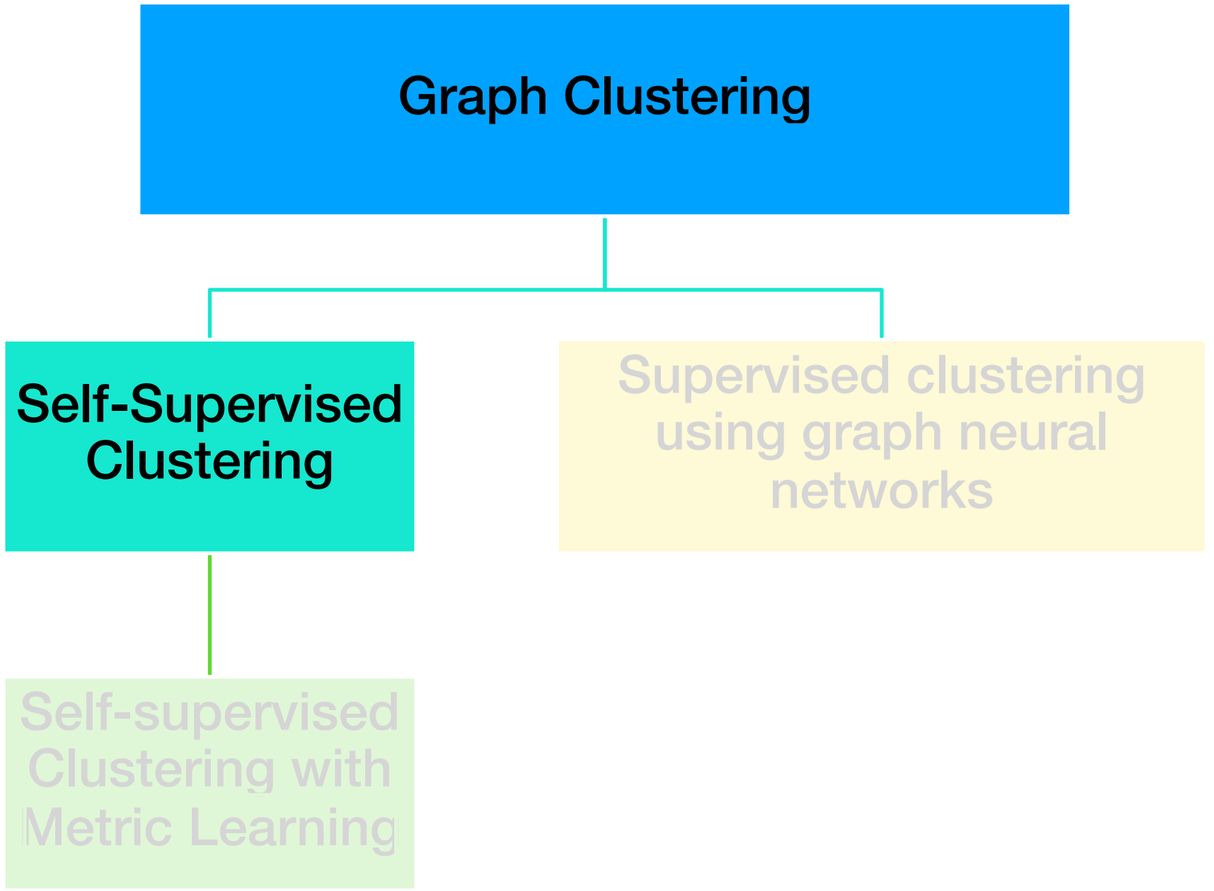
- Speech diarization challenge data
- 9-11 domains
- #Recordings – Dev set:254, Evalset:259
- Duration: 0.5-10 mins
- 1-10 speakers

Voxconverse [4]

-]extracted from YouTube videos.
- #Recordings - dev set: 216 and eval set: 232
- Duration: 22s - 20mins.
- 1-21 speakers.

[1] Mark et al., 2000 NIST Speaker Recognition Evaluation
[2] McCowan et al., The AMI meeting corpus, 2005
[3] Ryant et al., The Third DIHARD Diarization Challenge, 2020
[4] Chung et al., Spot the Conversation: Speaker Diarisation in the Wild, 2020

- Introduced **self-supervised learning using NN.**
- Introduced **hierarchical graph clustering.**



Proposed Approach 1

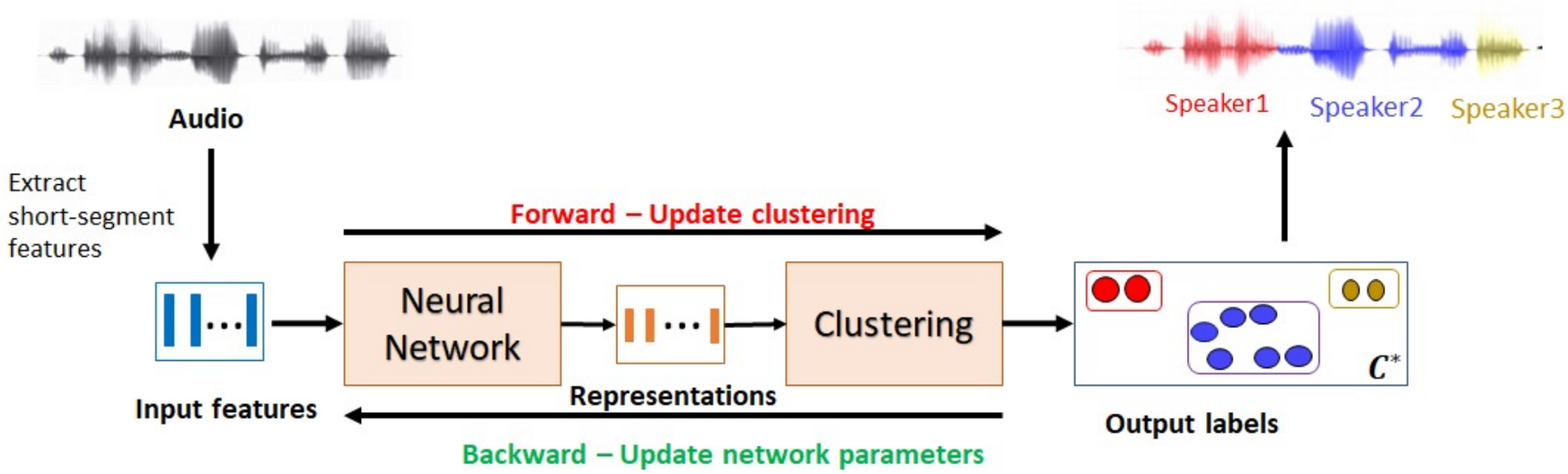
Motivation

- Each stage (embedding extraction and clustering) is optimized independently.
- The test set will contain unseen domains and speakers.
- Can clustering provide self-supervisory targets for representation learning[1]?
- Can we improve clustering using the succinct speaker representations ?

[1]Yang et. al. , “Joint unsupervised learning of deep representations and image clusters,” in CVPR, 2016

Self-supervised clustering

Self-Supervised Clustering alternates between **merging the clusters** for a fixed embedding representation and **learning the representations** using the given cluster labels, till we reach the **required number of clusters/speakers.**



Prachi Singh, Sriram Ganapathy, 'Deep self-supervised hierarchical clustering for speaker diarization', INTERSPEECH 2020.

SSC Algorithm

Variables:

$\boldsymbol{X} = \{x_1, \ldots, x_{N_r}\} \epsilon R^D$: X-vectors sequence of recording r

$\boldsymbol{Y} = \{y_1, \ldots, y_{N_r}\} \epsilon R^d$: lower dimensional representations

$\mathbf{z} = \{z_1, \ldots, z_{N_r}\} \epsilon R$: segment labels

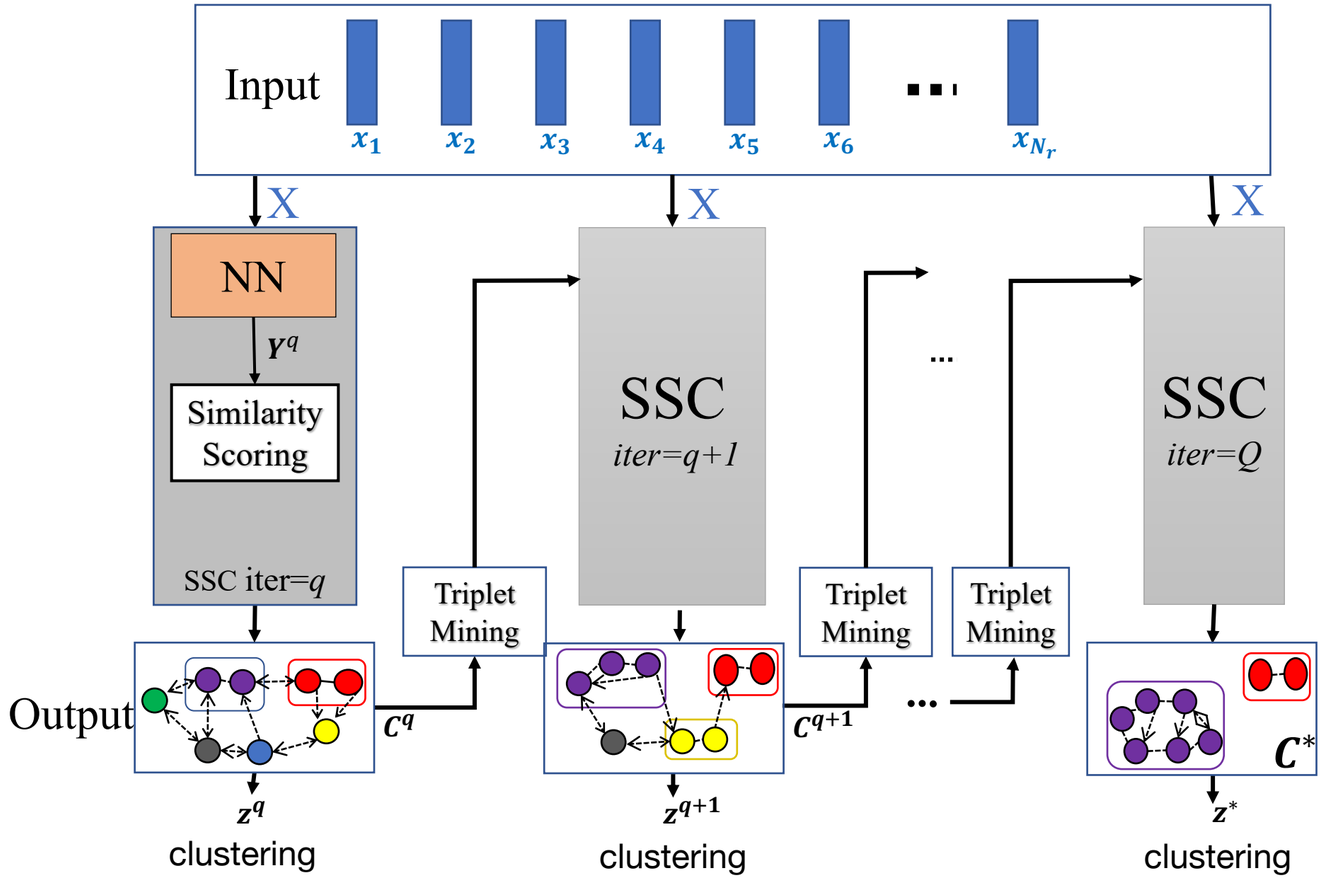
$\boldsymbol{\theta}$: DNN parameters

$(\boldsymbol{Y}^q, \boldsymbol{z}^q, \boldsymbol{\theta}^q)$: refer to variables at iteration q

N^q: Number of clusters at iteration q

N^*: target number of clusters

For NN training at iteration q, use clustering results from q-1 to sample positive and negative pairs of triplets.



P. Singh and S. Ganapathy, "Self-supervised Representation Learning With Path Integral Clustering For Speaker Diarization", IEEE TASLP (2021).

NN training–Triplet loss

- For each cluster C_i^q, pick two elements as anchor and positive $\{y_i, y_j\}$.
- For negative pair, element (y_l) is selected randomly from any other cluster.
- Triplet loss:

$$\theta^q = argmax_\theta \sum_{i,j,l} [s(i,j) - \alpha(s(i,l) + s(j,l))]$$

$s(i,j)$ – similarity score ; $0 < \alpha \leq 1$: weighting factor

Agglomerative clustering

AHC

Merging Criterion:

In an AHC algorithm, the merging criterion for merging two clusters C_a^q and C_b^q where q is the iteration index is given as

$$\{C_a^q, C_b^q\} = arg \max_{C_i, C_j \in C, i \neq j} A(C_i, C_j)$$

(where, A denote the affinity measure between two clusters.)

Agglomerative clustering

Path integral clustering (PIC)

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

1. **Measures the affinity of clusters based on the neighborhood graph** hence is **more robust to noisy distances**.
2. Uses robust **graph structural merging strategy for noisy links.**
3. It **does not assume anything on the underlying data distributions** and **only need the pairwise similarities** of samples.

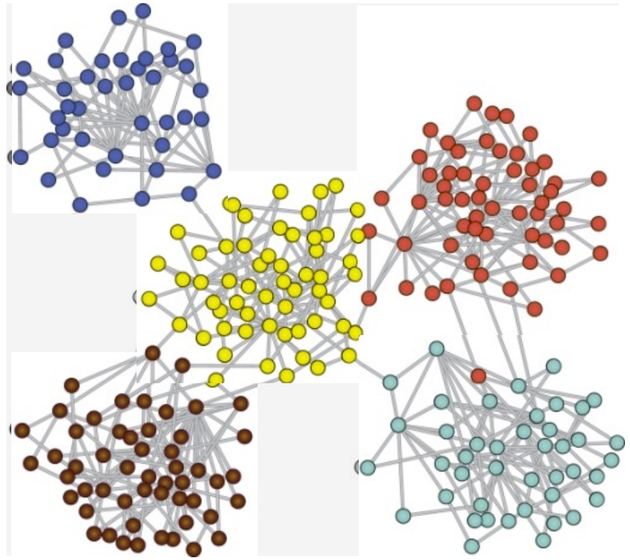
Path Integral Clustering (PIC)

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

  - 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

Baselines[12]

Step	Parameter	CH	AMI
-	Sampling rate	8kHz	16kHz
Segmentation	Window size	1.5s, 0.75s shift	1.5s, 0.75s shift
Embedding extraction (x-vector) extraction	Architecture	7-layers TDNN	7-layers TDNN
	Train set	SWBD, SRE	Voxceleb 1,2
	Train #speakers	4,285	7,323
	Input features	23D MFCCs	30D MFCCs
	x-vector dimension	128	512
Similarity score	type	PLDA	PLDA
Clustering	type	AHC	AHC

Implementation details

config	CH	AMI
x-vectors/recording	50-700	1000-4000
2-layer DNN	128x10	512X30
Learning rate	0.001	0.001
Annealing	No	Yes
Batch	Full	Mini-batch
epochs	5-10	5-10

Initialization

- **Weight initialization** and **training** are **file specific.**
- Uses **processing steps from baseline system.**
- **First layer** is initialized using **global PCA** computed using held out set followed by **length norm**.
- **Second layer** is initialized using **file-level PCA.**
- Affinity measure : **Cosine** similarity.

CH Results

- Performance metric: Diarization Error Rate (DER) (%)
- Considering only non-overlapping speech regions with tolerance collar (0.25s).

System	Known N*	Unknown N*
x-vec + cosine + AHC	8.9	10.0
x-vec + cosine + SC	9.4	11.9
x-vec + PLDA + AHC	7.0	8.0
x-vec + cosine + PIC	7.7	9.3
SSC-AHC	6.4	8.3
SSC-PIC	6.4	7.5
+ Temp. cont.	**6.3**	**7.0**

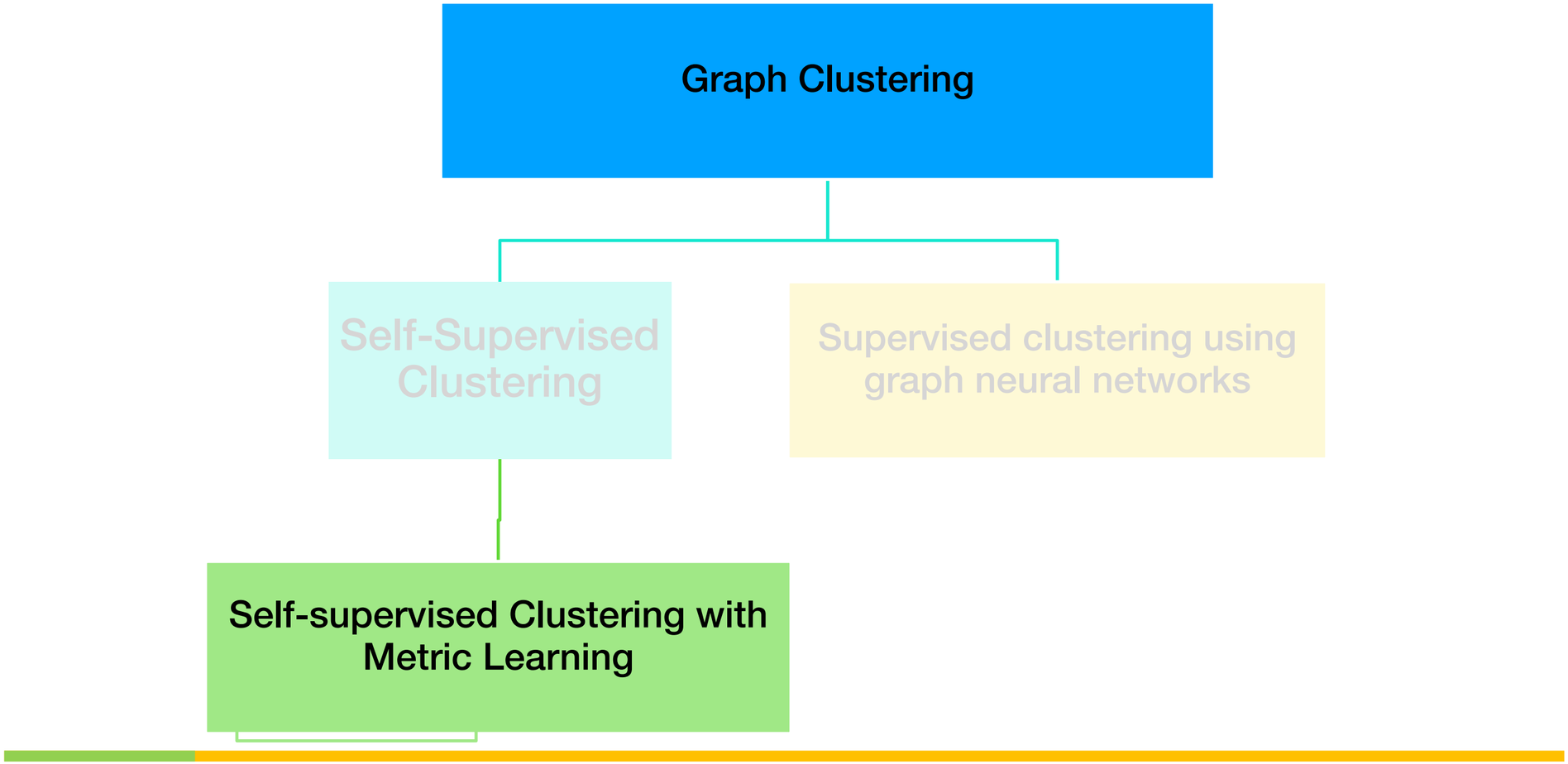
AMI Results

System	Known N*		Unknown N*	
	Dev.	Eval.	Dev.	Eval.
x-vec + cosine + AHC	34.6	30.2	18.2	15.5
x-vec + cosine +SC	30.2	25.5	40.0	31.1
x-vec + PLDA + AHC (Baseline)	15.7	16.0	13.7	16.3
SSC-PLDA-AHC	9.4	11.1	10.7	11.6
x-vec + PLDA + PIC	9.4	9.3	9.8	10.4
x-vec + cosine + PIC	8.9	7.3	9.0	7.3
SSC-PIC	7.3	7.2	8.1	7.6
+ Temp. cont.	**6.2**	**6.4**	**6.4**	**6.7**

Prachi Singh and Sriram Ganapathy, “Self-supervised Representation Learning with Path Integral Clustering for Speaker Diarization”, IEEE Transactions on Audio Speech and Language Processing,2021.

Summary

- Proposed **self-supervised clustering algorithm using DNN** which iteratively updates representation learning and clustering.
- Introduced **path integral clustering – hierarchical graph clustering for first time for diarization.**
- Helped to **increase separation between representations** of different speakers.
- Showed improvements on AMI and CALLHOME dataset.



Proposed Approach 2

Motivation

- **SSC uses cosine similarity** to perform clustering.
- **Prior work on clustering performs better with PLDA score than cosine.**
- PLDA[1] is a parametric model which is trained to learn speaker distributions.
- Can we train the SSC with learnable scoring/metric function?

[1]Sergey Ioffe, Probabilistic linear discriminant analysis, 2006

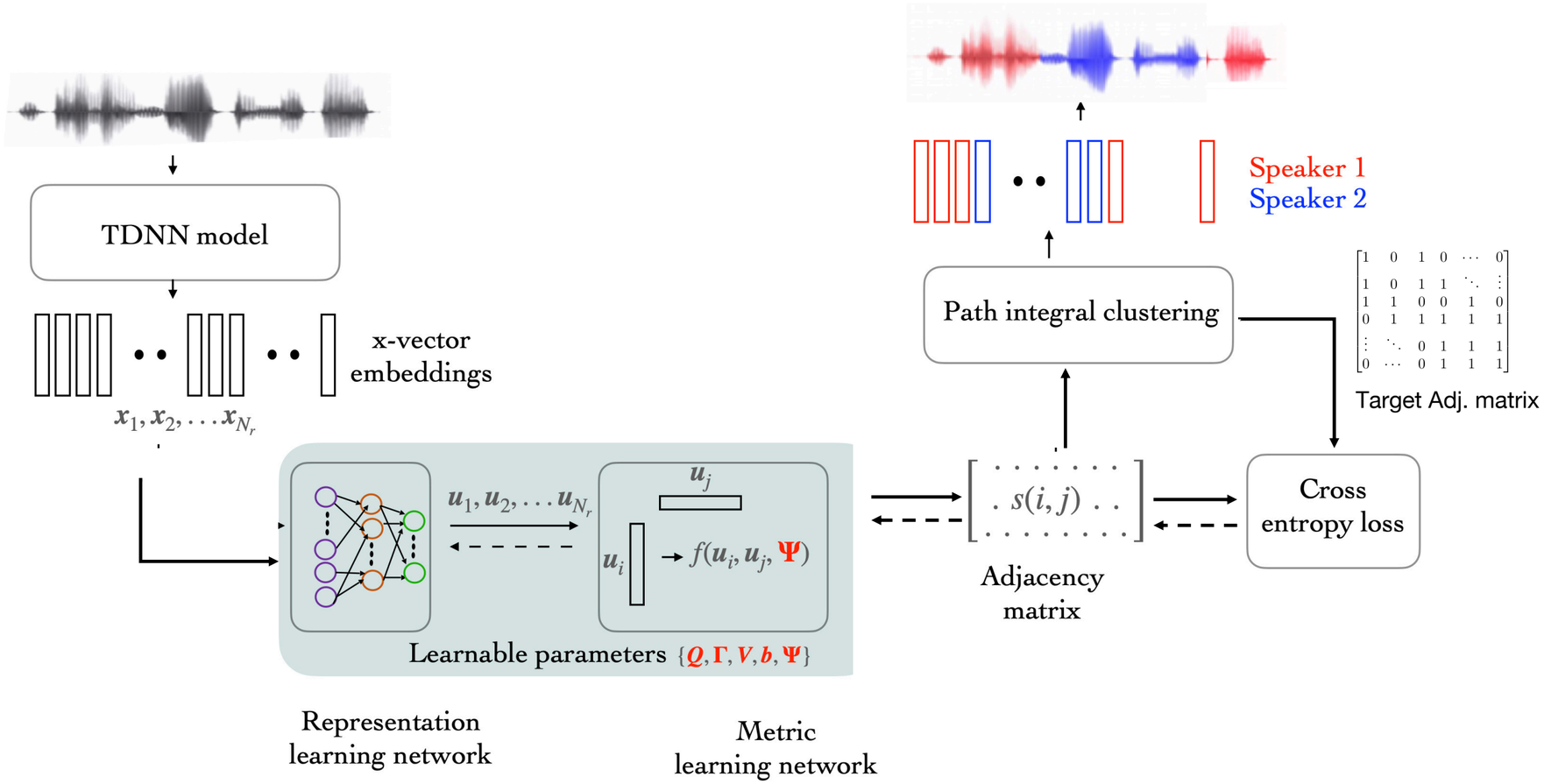
SelfSup-PLDA-PIC

- **Self-supervised metric learning** with **graph-based clustering** algorithm (**SelfSup-PLDA-PIC**) jointly **performs representation learning and metric learning** using the initial clustering results.

- **Propose** a **neural version of PLDA** to incorporate **deep learning** of the **PLDA model parameters**.

P. Singh and S. Ganapathy, “Self-Supervised Metric Learning with Graph Clustering for Speaker Diarization”, IEEE ASRU 2021.

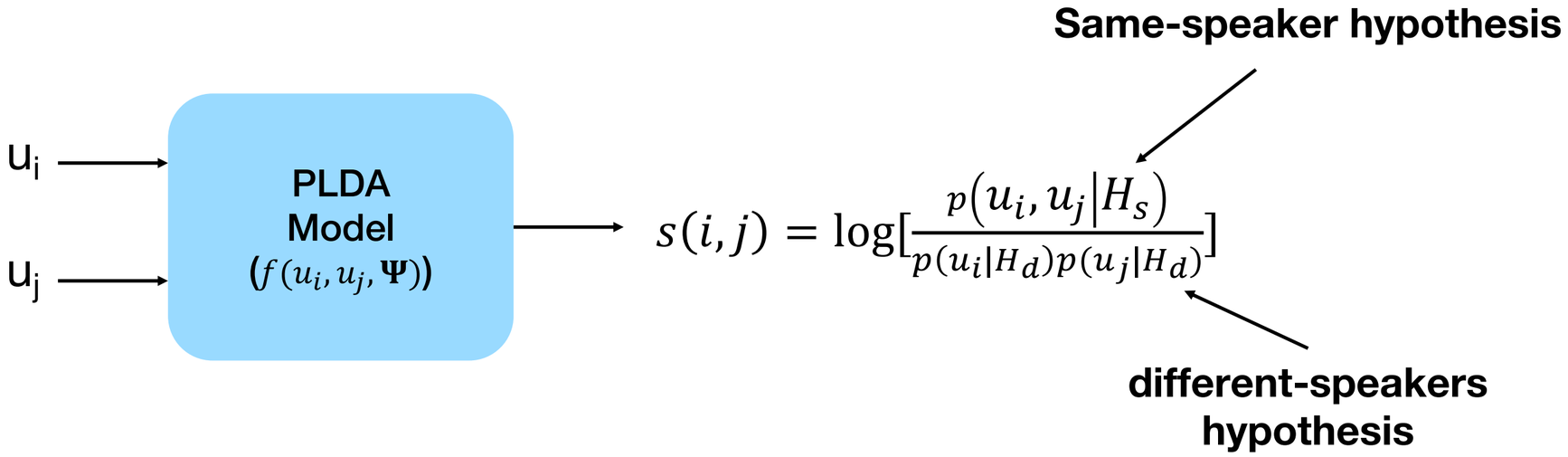
Block diagram: SelfSup-PLDA-PIC



Metric Learning using PLDA model

- Probabilistic Linear Discriminant Analysis (PLDA)[1] is a supervised generative model trained to learn distributions of different speakers.

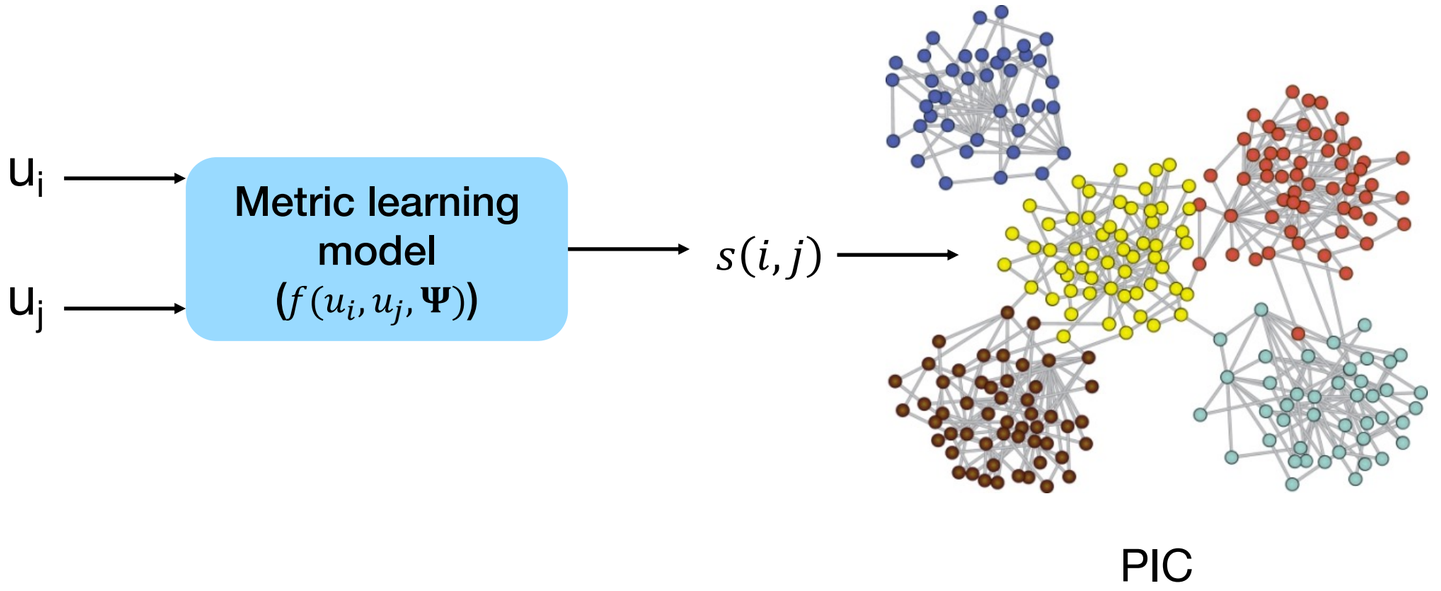
- It can be used to find pairwise similarity score between embeddings from unseen speakers as follows



[1]Sergey Ioffe, Probabilistic linear discriminant analysis, 2006

Metric Learning using PLDA model

- Replacing PLDA model with a learnable parametric model with parameter $\boldsymbol{\Psi}$



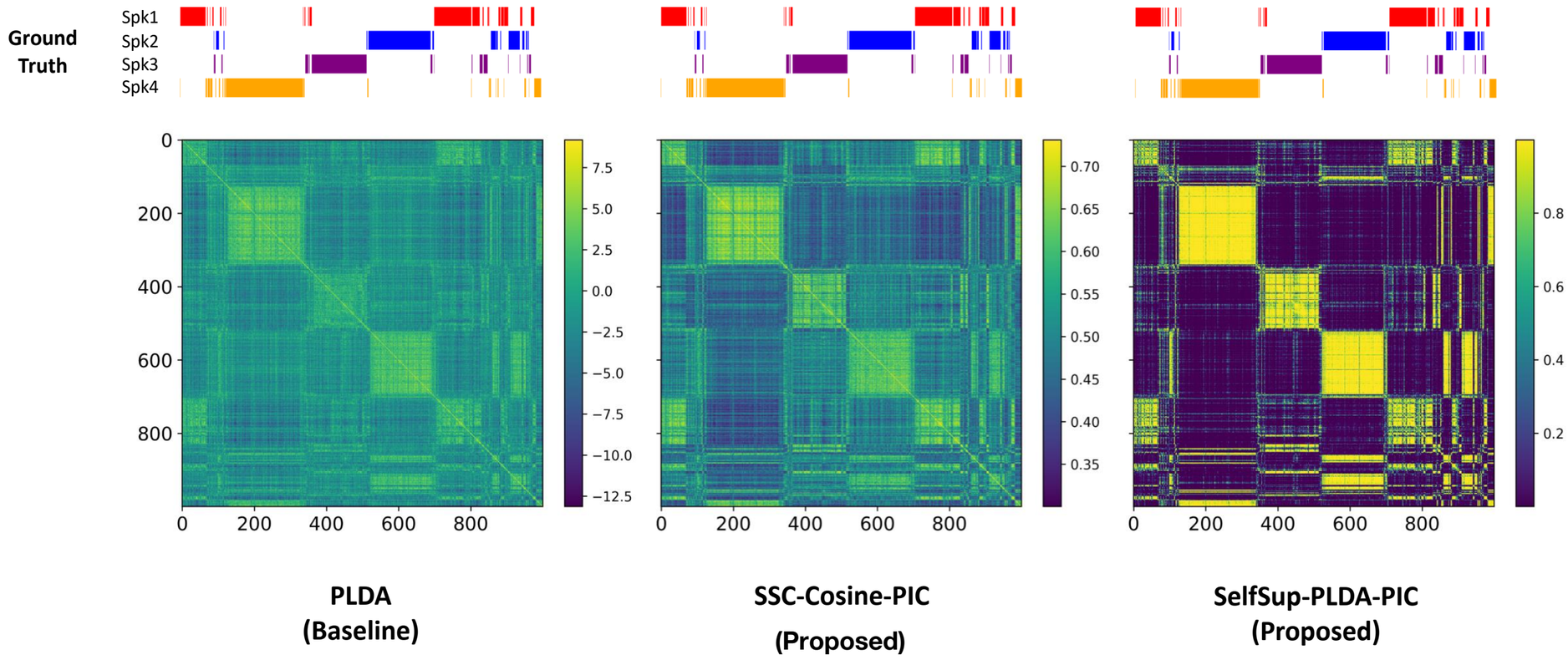
AMI Results

AMI DER (%) Results – Ignoring overlaps and with collar 0.25s

System	Known N*		Unknown N*	
	Dev.	Eval.	Dev.	Eval.
x-vec + PLDA + AHC	**15.9**	**12.2**	**13.1**	**12.3**
x-vec + PLDA + PIC	5.1	10.2	5.8	11.4
SSC-Cosine-PIC	**5.3**	**6.2**	**6.5**	**8.4**
SelfSup-PLDA-AHC	7.9	7.3	7.7	9.4
SelfSup-PLDA-PIC[1]	4.2	6.2	4.4	6.9
_ + Temporal continuity	4.2	4.2	**4.4**	**4.9**
SelfSup-PLDA-PIC + VBx[2]	-	-	**2.9**	**4.2**

[1]P. Singh and S. Ganapathy, "Self-Supervised Metric Learning with Graph Clustering for Speaker Diarization", IEEE ASRU 2021.
[2]Diez et al., Bayesian HMM based x-vector clustering for speaker diarization, 2019

AMI Visualization



Similarity score matrices comparison for 4-speaker recording from AMI development set

DIHARD Results

Average DER (%) on the DIHARD dataset considering overlapping regions with no tolerance collar.

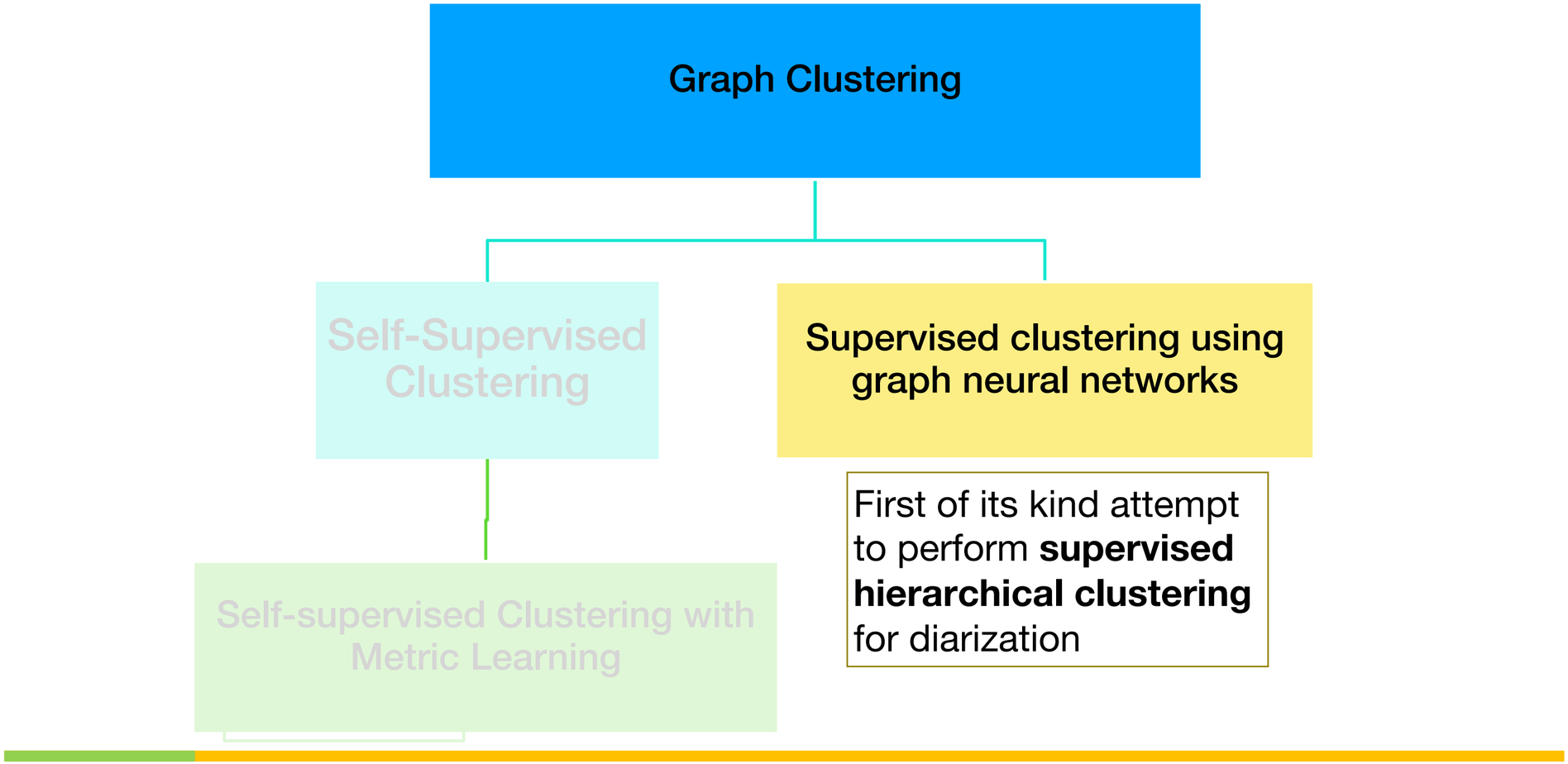
For recordings with ≤ 7 speakers and > 7 speakers.

System	≤ 7 speakers		> 7 speakers	
	Dev.	Eval.	Dev.	Eval.
X-vec + PLDA + AHC	18.0	19.3	36.6	27.1
X-vec + PLDA + PIC	17.7	17.8	36.5	24.0
SelfSup-PLDA-PIC	17.0	17.2	39.5	28.1

Performance degraded as number of speakers increases as initial clustering becomes unreliable.

Summary

- Proposed **self-supervised metric learning** approach using **PLDA**.
- Adapted similarity scores for each test recording.



Proposed Approach 3

Motivation

- Self-supervised clustering is less reliable when recording **contains higher number of speakers (>7)**.

- The end goal is **to minimize the clustering errors** to improve performance

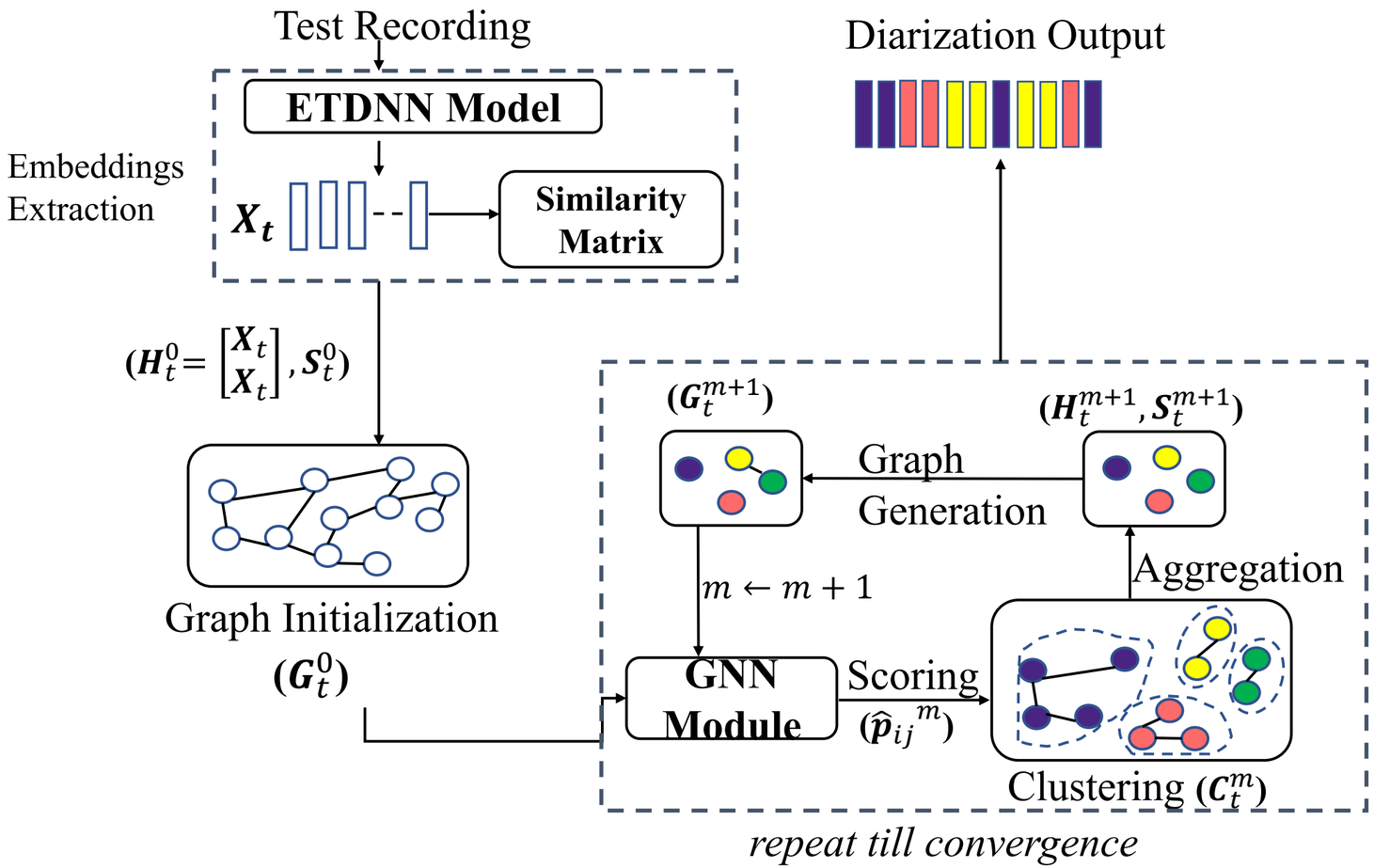
- **Can we train a supervised model with the clustering objective**?

Supervised HierArchical GRaph Clustering (SHARC)

- Performs **supervised clustering** using **Graph Neural Networks (GNN).**
- Represents the speaker embeddings using graph.
- Clustering loss is used to update edges of the graph.
- Generates node labels based on clustering performed on updated edges at **each level of hierarchy**.
- **E-SHARC** : Joint learning of embedding extractor and GNN

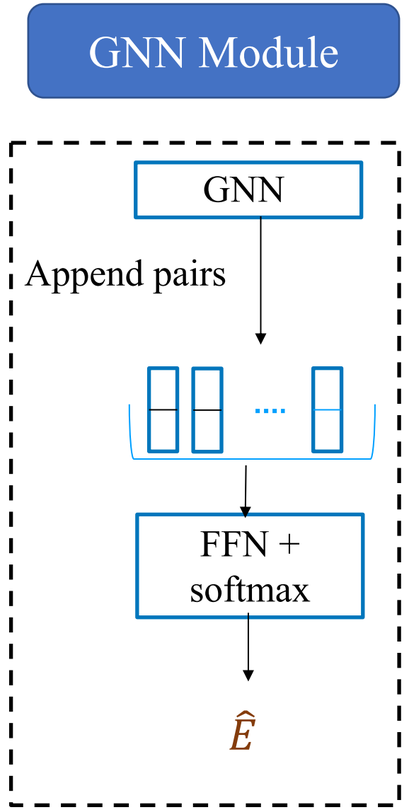
Yifan Xing et. al, “Learning hierarchical graph neural networks for image clustering,” in Proc. IEEE ICCV, 2021.

Block diagram: E-SHARC Inference



GNN scoring

- GNN scoring function Ψ - a learnable GNN module designed for supervised clustering.



- Output: edge prediction probability p_{ij} between node i and j.
- N_i^k- k-nearest neighbors of node vi,

$$\hat{e}_{ij} = 2p_{ij} - 1 \in [-1, 1] \forall j \in N_i^k$$

- Density of node i :

$$d_i = \frac{1}{k} \sum_{j \in N_i^k} e_{ij} \boldsymbol{S}_r(i, j)$$

- Ground truth:
- Predicted:

$$\hat{d}_i = \frac{1}{k} \sum_{j \in N_i^k} \hat{e}_{ij} \boldsymbol{S}_r(i, j)$$

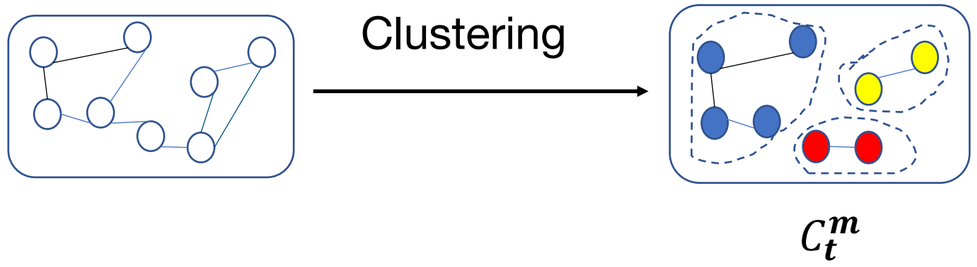
Clustering

- At each level of hierarchy m, it creates a candidate edge set ε(i)

$$\varepsilon(i) = \{j | (v_i, v_j) \in E_m, \quad \hat{d}_i \leq \hat{d}_j \quad \text{and} \quad p_{ij} \geq p_\tau\}$$

- For any i, if ε(i) is not empty, we pick $j = \text{argmax}_{j \in \varepsilon(i)} \hat{e}_{ij}$

- A set of connected components C_t^m , forms clusters for the next level (m + 1).



Training loss

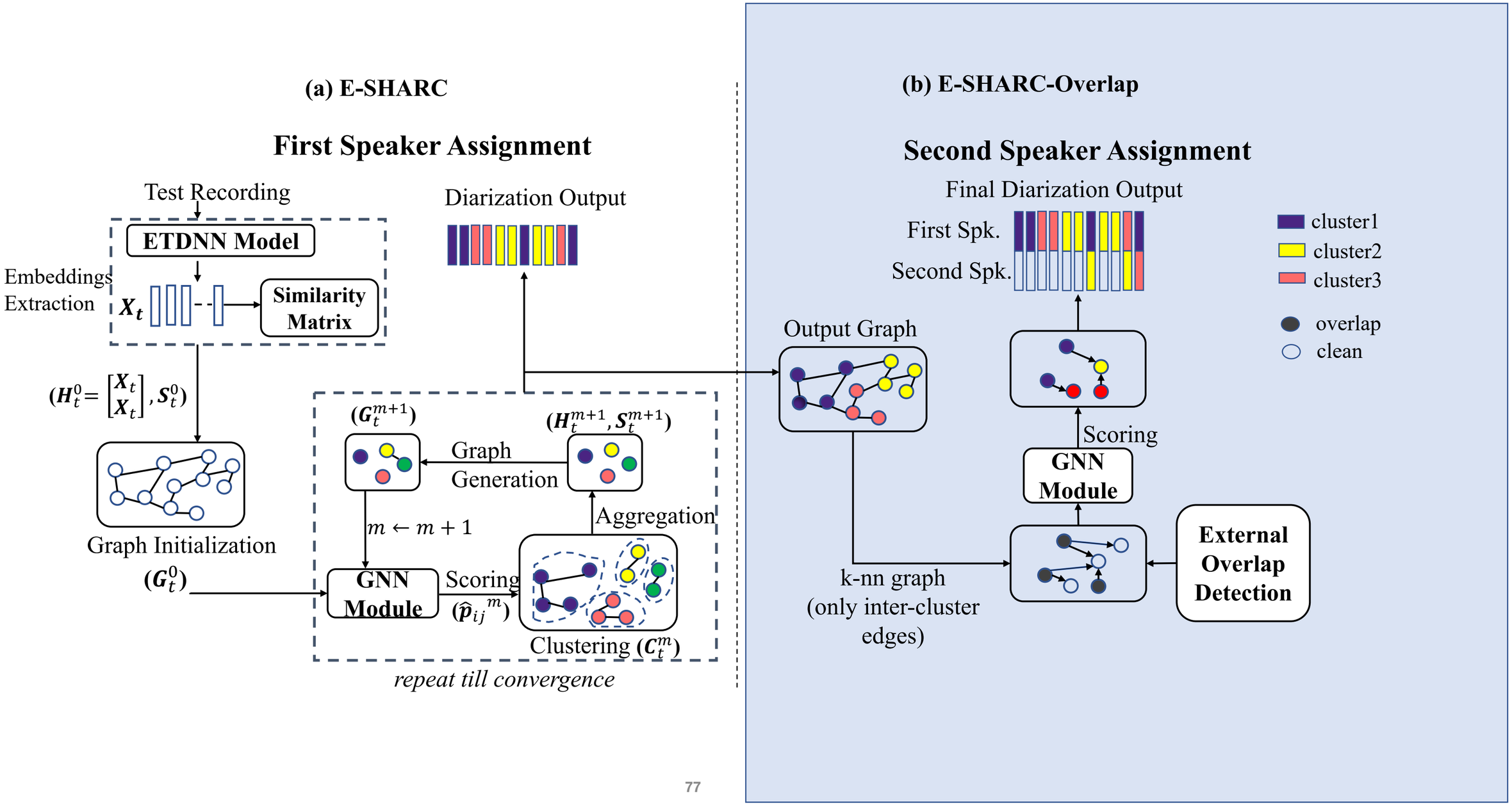
- Loss: $L = L_{conn} + L_{den}$
  - $L_{conn} = \frac{1}{|E|}\sum_{i,j \in E} p_{ij} log\, \hat{p}_{ij} + (1 - p_{ij}) log\, (1 - \hat{p}_{ij})$

    p_{ij}- Ground truth edge labels,$\hat{p}_{ij}$- predicted edge labels

  - $L_{den} = \frac{1}{|V|}\sum_{i=1}^{|V|} ||d_i - \hat{d}_i||_2^2$ ∀i ∈ {1, ..., |V |}, where |V| is the cardinality of V
- d_i : ground truth node density, $\hat{d}_i$: predicted node density

- What about overlapping speech ?

Block diagram: E-SHARC-Overlap Inference



Experiments

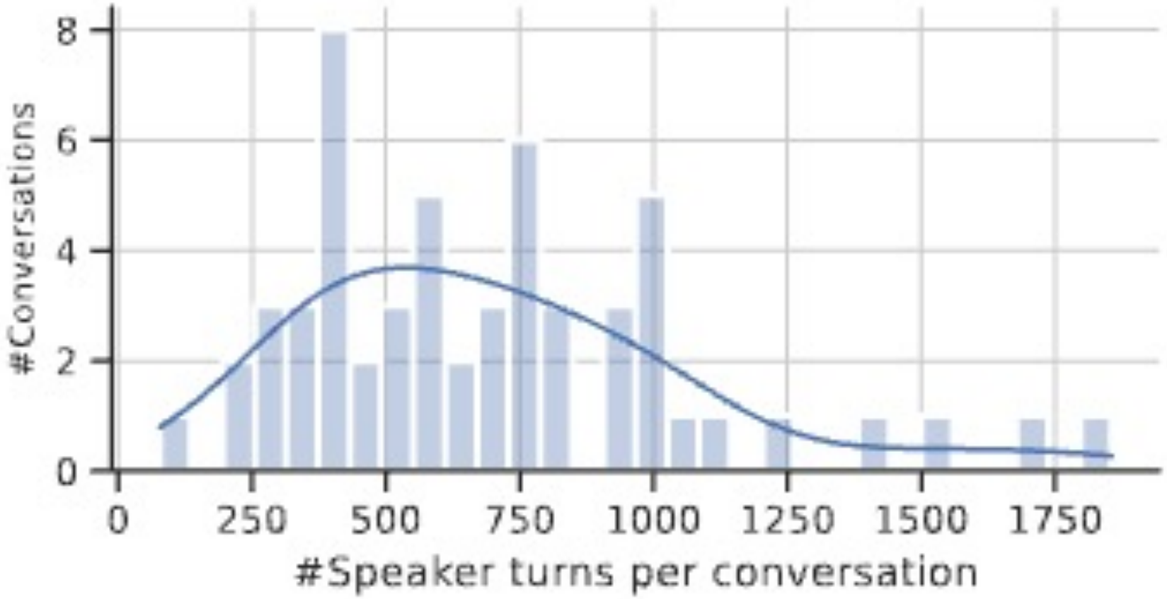
Datasets

AMI : Meeting dataset

Voxconverse: Youtube videos

DISPLACE 2023 dataset[1]: Natural multilingual, multi-speaker speech recordings.

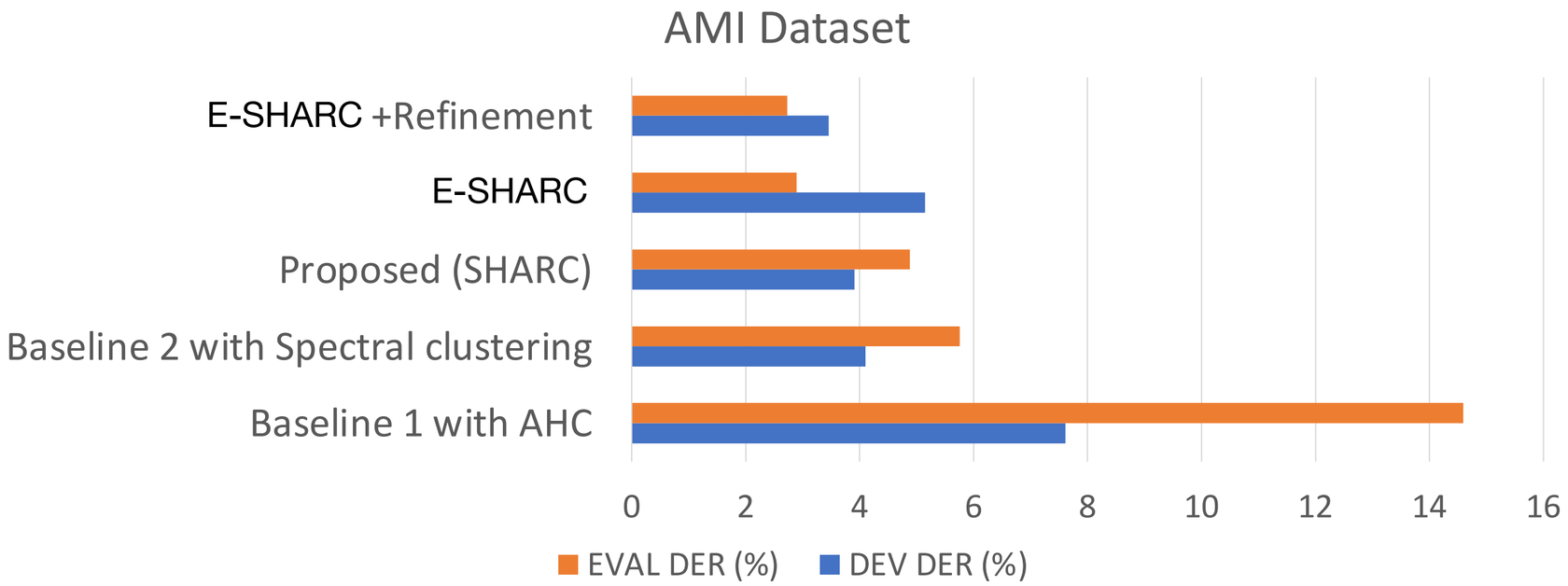
- #Recordings- dev set: 27 and eval set: 29.
- Duration: 30-60 mins
- #speakers varies from 3-5, and #languages varies from 1-3.



[1]Baghel et al., Interspeech 2023

Results

- Performance : DER (%) (lower the better)

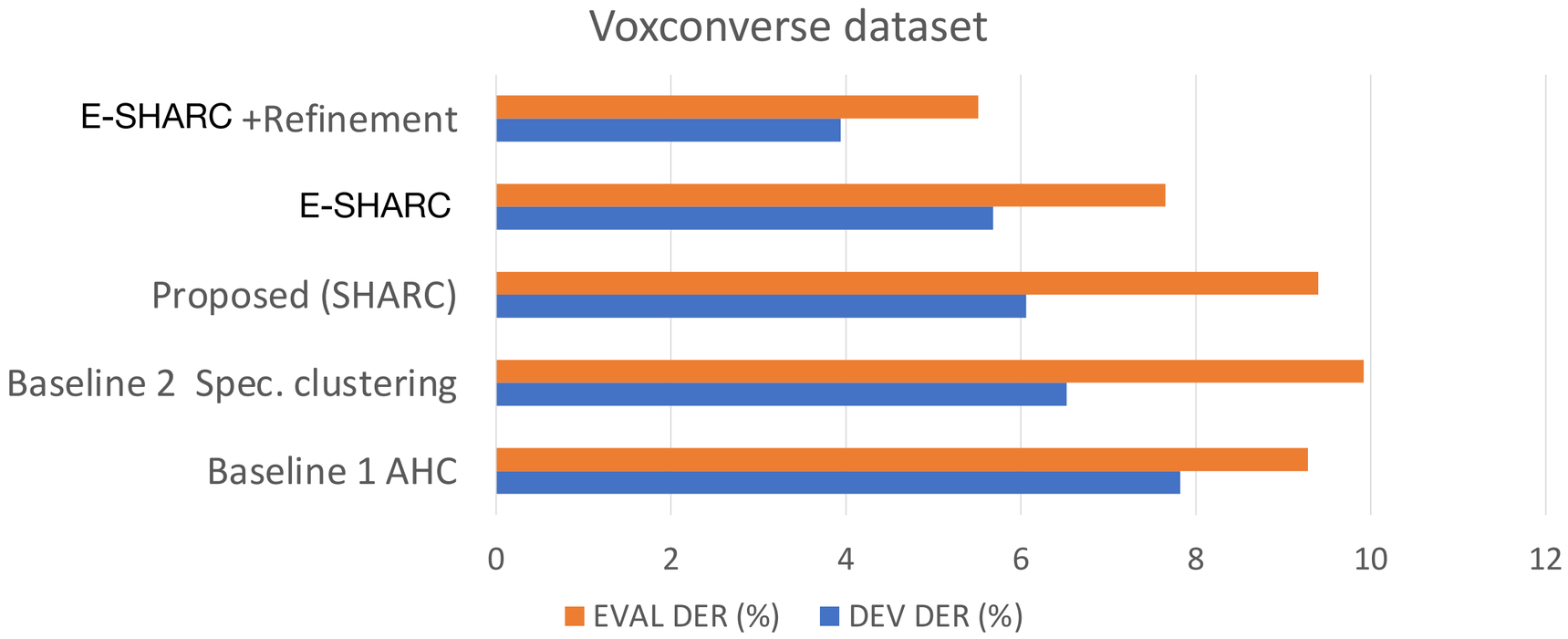


53% relative improvement over best baseline

P. Singh, A. Kaul and S. Ganapathy, “Supervised Hierarchical Clustering using Graph Neural Networks for Speaker Diarization”,IEEE ICASSP 2023.

Results

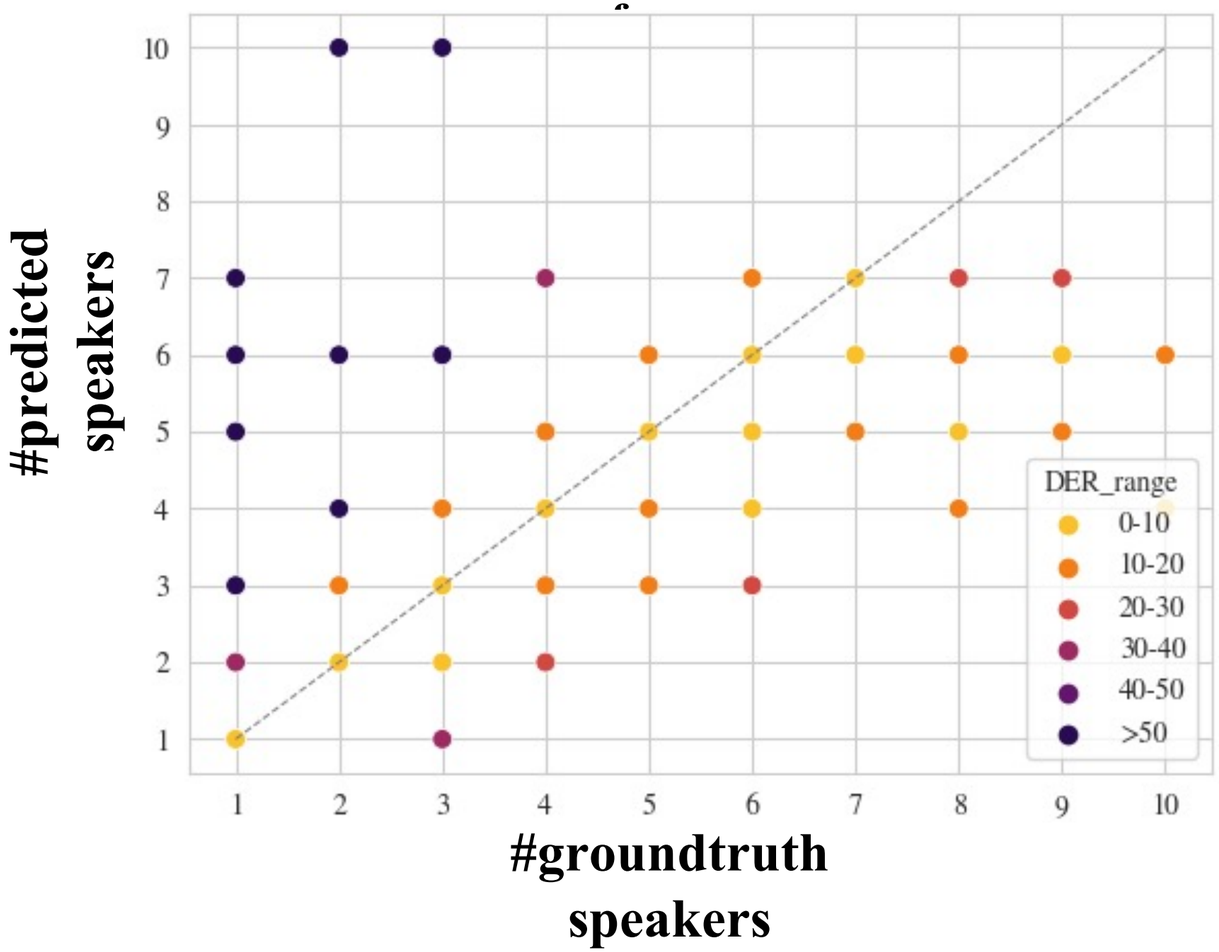
- Performance : DER (%) (lower the better)



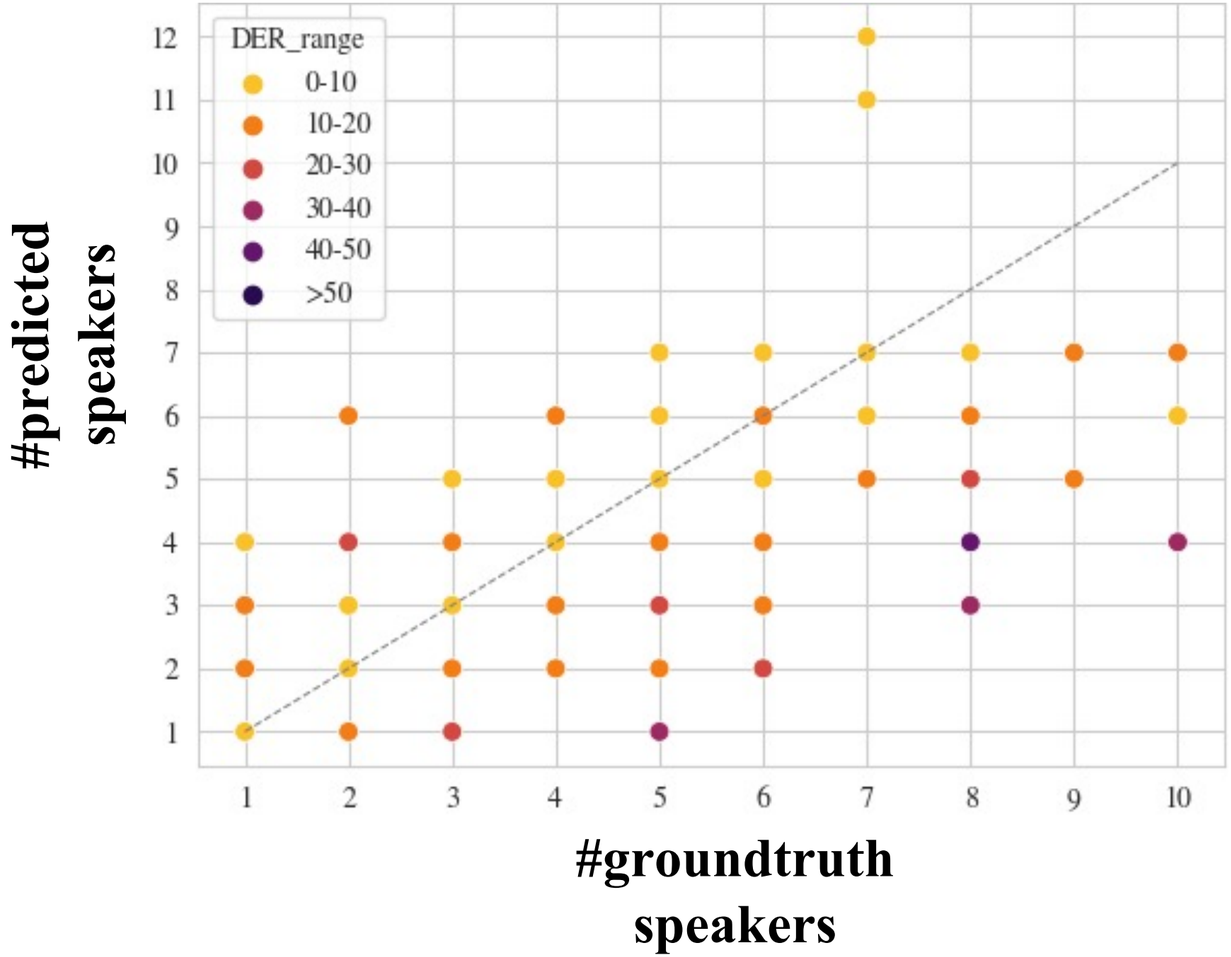
41% relative improvement over best baseline

Voxconverse Dev set

(a) SC Average DER (%)



(b) ESHARC Average DER (%) performance



Overlap Results

DER* - with overlap + no collar

AMI	Eval DER* (%)
AHC + overlap	26.67
SC + overlap	20.36
SHARC + overlap	19.50
E-SHARC + overlap	17.99
Voxconverse	**Eval DER* (%)**
AHC + overlap	12.05
SC + overlap	13.73
SHARC + overlap	12.56
E-SHARC + overlap	11.42
DISPLACE	**Eval DER* (%)**
AHC + overlap	40.47
SC + overlap	40.65
SHARC + overlap	32.73
E-SHARC + overlap	32.45

Overlap detector: Bredin et al., pyannote.audio: neural building blocks for speaker diarization, 2020

Recent works Results

- DER – without overlap + 0.25s collar
- DER* - with overlap + no collar

AMI SDM System	DER*	DER
Pyannote [1]	29.1	-
x-vec+AHC+VBx [2]	27.4	12.6
SelfSup-PLDA-PIC +VBx [3]	23.8	5.5
Raj et al. [4]	23.7	-
Plaquet et al. [5]	22.9	-
GAE-based+ SC [6]	-	5.5
GADEC-based [6]	-	4.2
E-SHARC (proposed)	19.83	2.9
E-SHARC-Ovp +VBx (prop.)	**17.2**	2.6

[1] Bredin et al., Interspeech, 2021
[2] Landini et al., 2020
[3] Singh et al., ASRU, 2021
[4] Raj et al., arxiv, 2022
[5] Plaquet et al., Interspeech, 2023
[6] Wei et al., Speech Communications, 2023

Recent works Results

- DER – without overlap + 0.25s collar
- DER* - with overlap + no collar

Voxconverse System	DER*	DER
Pyannote [1]	11.9	-
Plaquet et al. [2]	10.4	-
GAE-based+ SC [3]	-	8.0
GADEC-based [3]	-	7.6
E-SHARC (prop.)	11.68	7.6
E-SHARC-Ovp +VBx (proposed)	**10.1**	6.3
DISPLACE System		
DISPLACE Baseline [4]	32.2	14.6
E-SHARC-Ovp +VBx (prop.) + Baseline SAD	31.4	13.0
Winning system [4]	**27.8**	7.3

[1] Bredin, Interspeech, 2021
[2] Plaquet et al., Interspeech, 2023
[3] Wei et al., Speech Communications, 2023
[4] Baghel et al., 2023

Summary

- Introduced **supervised hierarchical clustering for speaker diarization for the first time**.
- Designed an **end-to-end approach** to perform speaker diarization using **Graph Neural Networks**.
- Introduced **overlapped speaker prediction**.
- Achieved state-of-the-art performance on benchmark datasets.

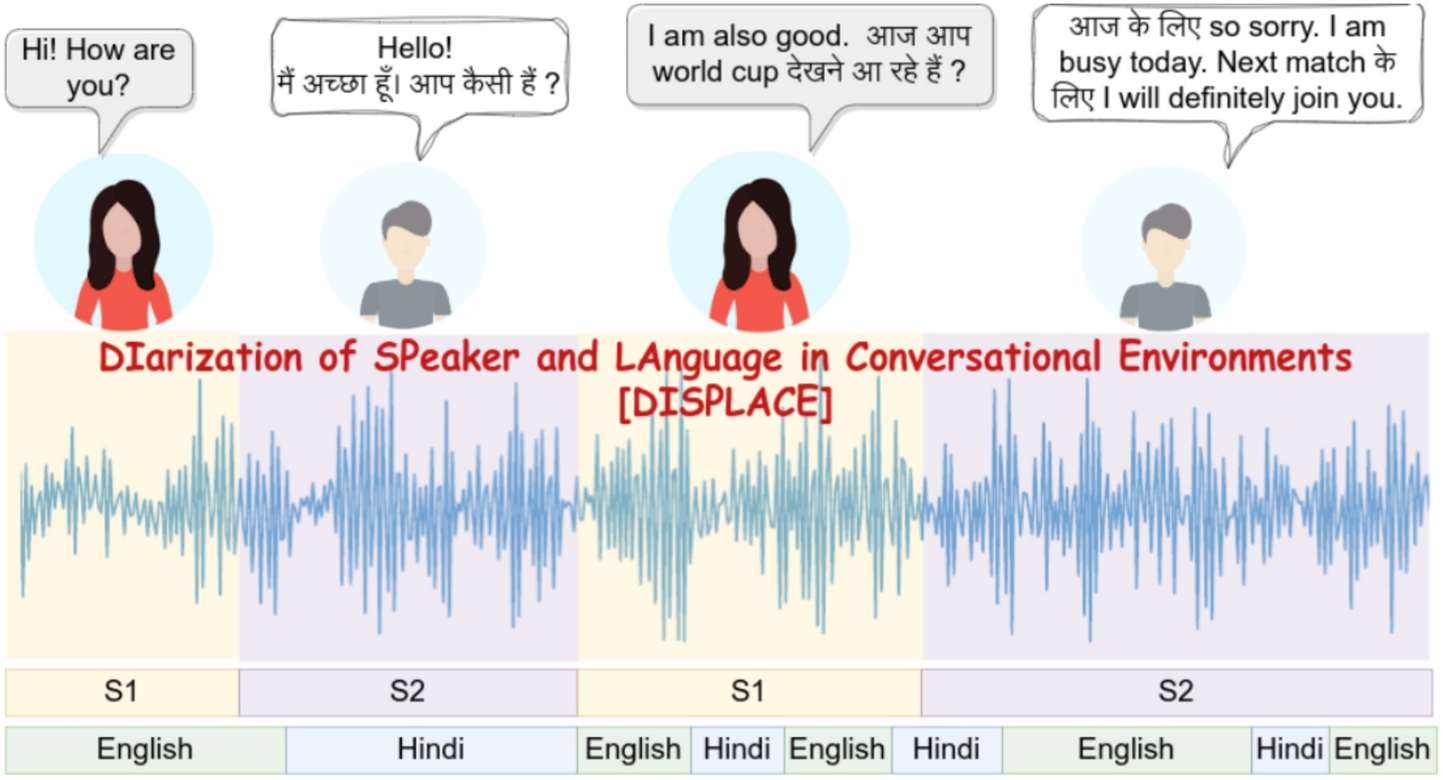
Conclusion and Future Directions

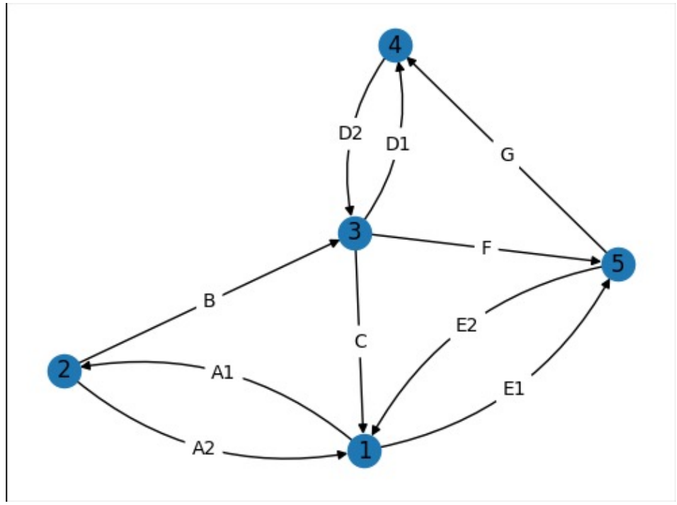
Concluding remarks

Proposed Approaches	Novelties	Limitations
SSC	• Introduced **self-supervised clustering using DNN** • Introduced **PIC graph clustering** for the first time to improve diarization.	• Similarity scoring is not learnable (cosine)
SelfSup-PLDA-PIC	• Introduced self-supervised **metric learning**	• Performance depends on initial clustering • Degrades with higher number of speakers
SHARC	• **First time** performed **supervised hierarchical clustering** for diarization	• Increased training time • Require domain specific training • Not purely end-to-end

Future Directions

Multilingual conversation Diarization



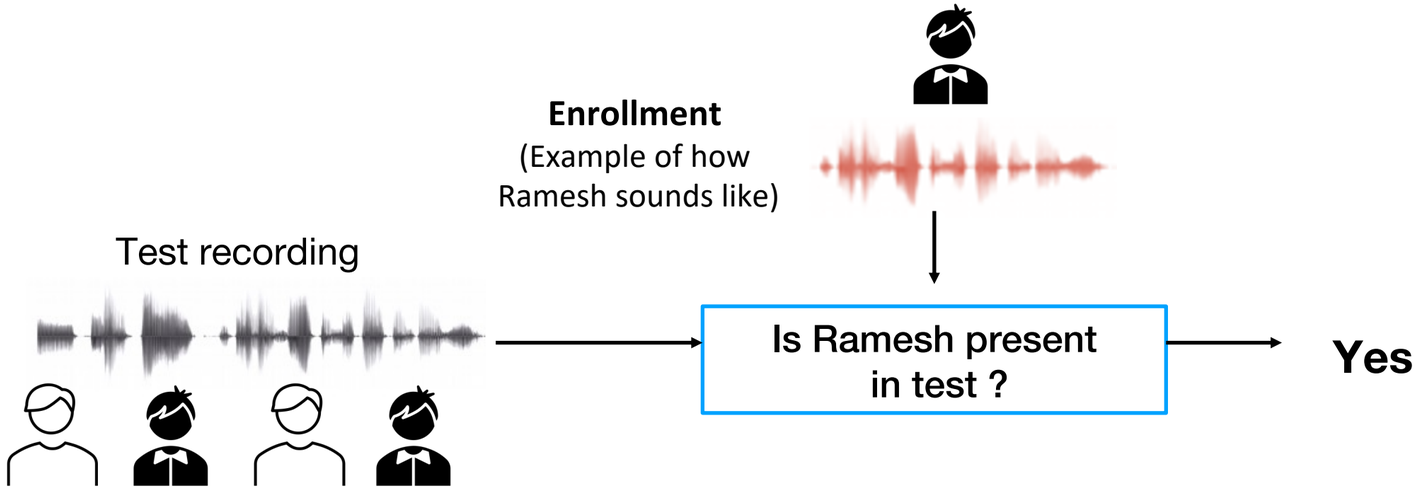


Use Multi-edge graph to perform multi-task learning

Source: https://displace2023.github.io/

Future Directions

Target speaker identification in conversational speech



Possible direction

- Need to handle channel mismatch
- Avoid clustering within target speaker recording

Publications based on the thesis

- **Peer-reviewed Journals:**
  1. **P. Singh** and S. Ganapathy, “Self-supervised Representation Learning With Path Integral Clustering For Speaker Diarization”, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2021).
  2. **P. Singh** and S. Ganapathy, “Speaker Diarization with Graph Based Supervised Hierarchical Clustering” (under review).

- **Peer-reviewed Conferences:**
  1. **P. Singh**, A. Kaul and S. Ganapathy, “Supervised Hierarchical Clustering using Graph Neural Networks for Speaker Diarization”, **IEEE ICASSP 2023**.
  2. **P. Singh** and S. Ganapathy, “Self-Supervised Metric Learning with Graph Clustering for Speaker Diarization”, **IEEE ASRU 2021**.
  3. **P. Singh**, R. Varma, V. Krishnamohan, S. R. Chetupalli, and S. Ganapathy. “LEAP Submission for the Third DIHARD Diarization Challenge”, **INTERSPEECH 2021**.
  4. **P. Singh** and S. Ganapathy, “Deep Self-Supervised Hierarchical Clustering for Speaker Diarization”, **INTERSPEECH 2020**.
  5. **P. Singh**, Harsha Vardhan MA, S. Ganapathy, A. Kanagasundaram, “LEAP Diarization System for the Second DIHARD Challenge”, **INTERSPEECH 2019**.

Indian Institute of Science

"Since 1909, when it came to be,

thousands have drunk in its glory;

Getting in is tough but

leaving it is more rough,

such is the charm of IISc."

- Anonymous

Thank you for your attention !