Graph Clustering approaches for Speaker Diarization of Conversational Speech

PhD Thesis Colloquium
21st July 2023

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

Advisor: Dr. Sriram Ganapathy

Outline

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

Introduction

Motivation

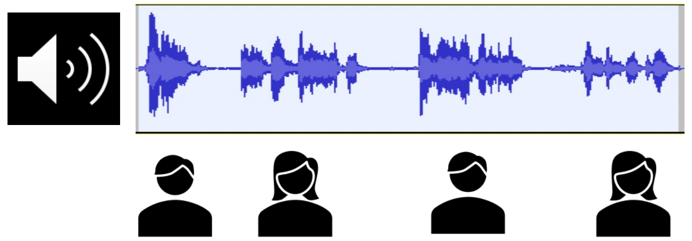
What is a conversational speech ?

Conversational audio contains multiple speakers engaged in a conversation.

Modelling of such audio requires understanding speakers' characteristics and content

Motivation

Transcribing audio into text using speaker information generates much meaningful text



Hello

Hello. How are you Nitin?

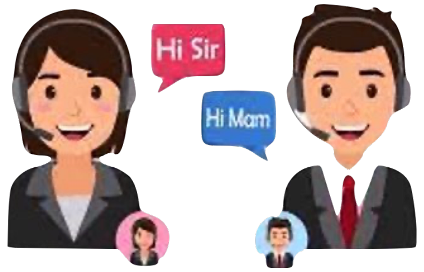
I am doing great. How are you Meenu?

I am doing also great.

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

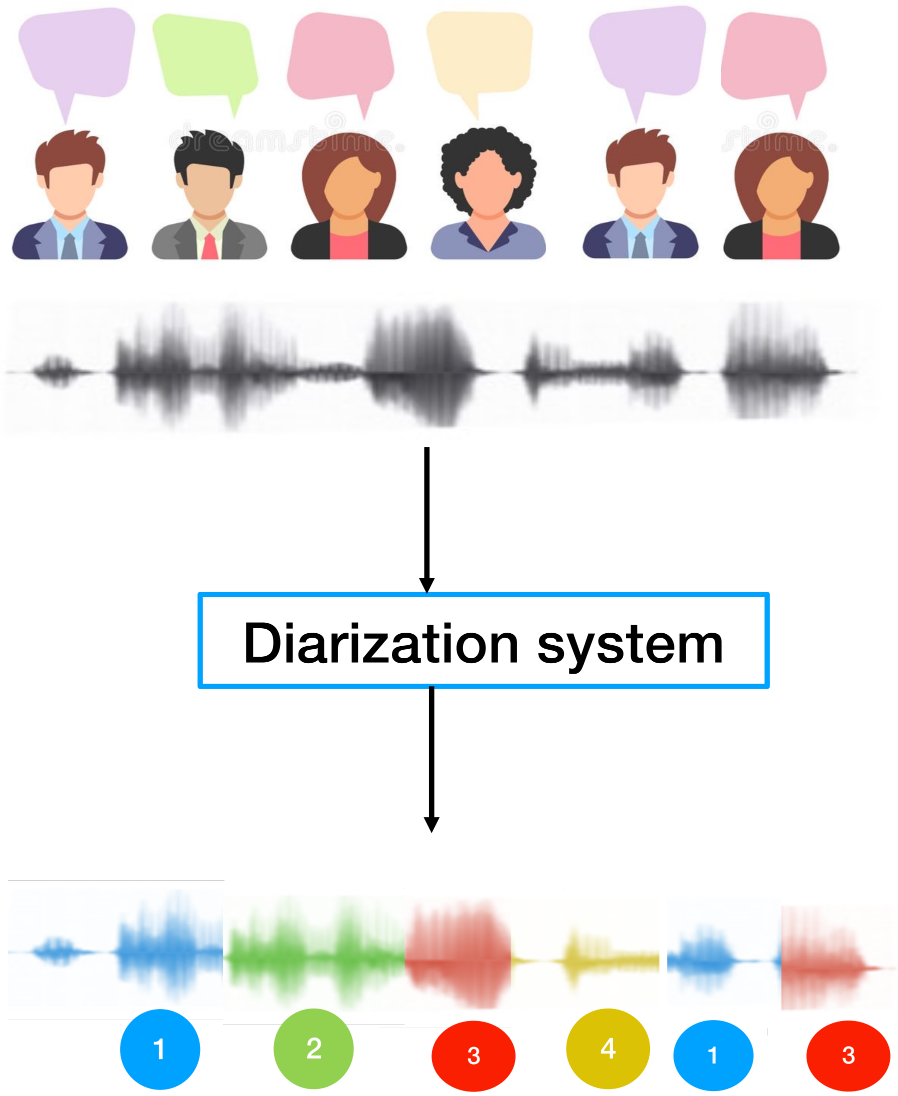
Transcribing meeting

Call center interactions Analysis

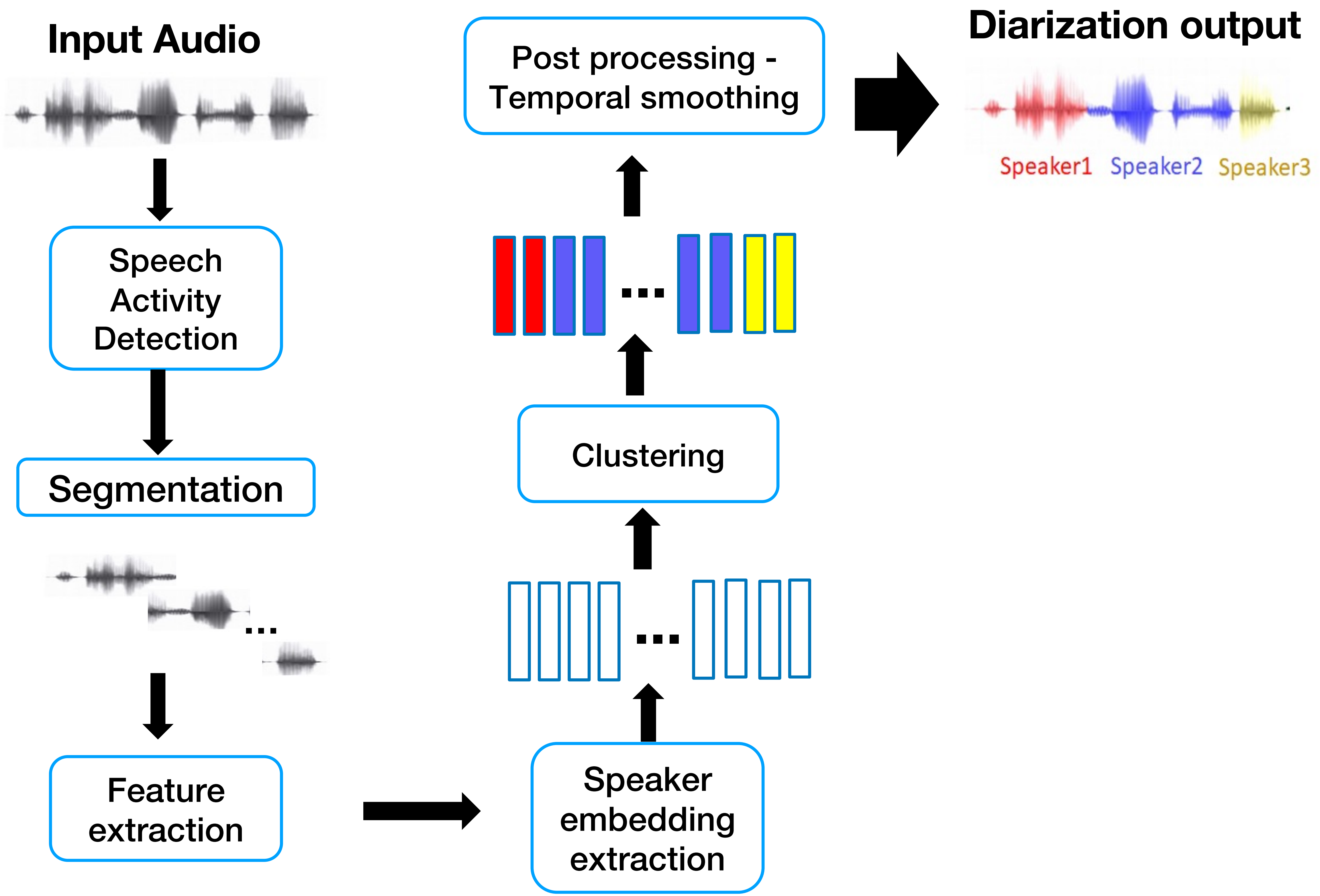


Definition

Speaker diarization is the task of partitioning an input audio recording into segments based on speakers and assign relative speaker labels.

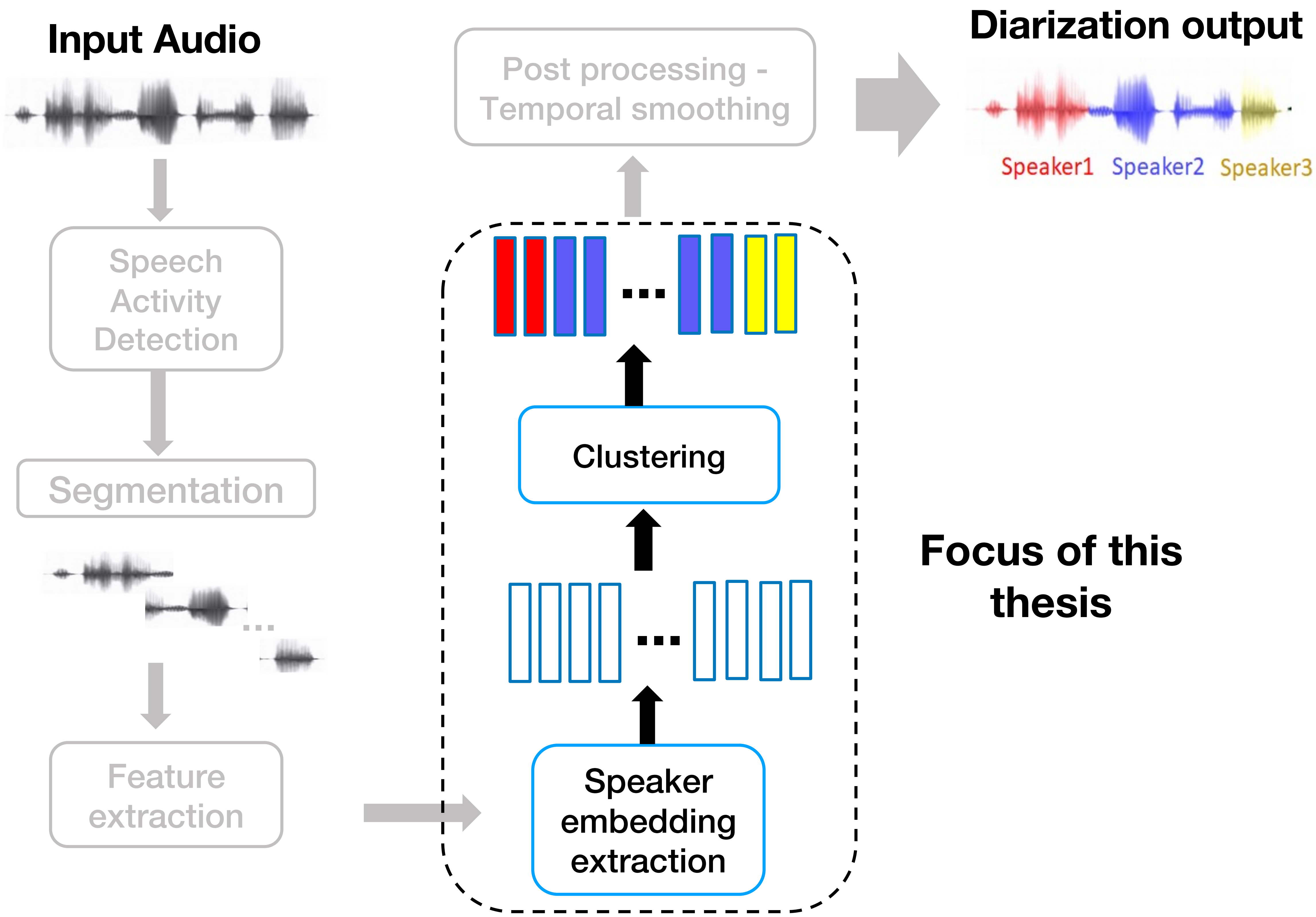


Methodology



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

Contributions outline



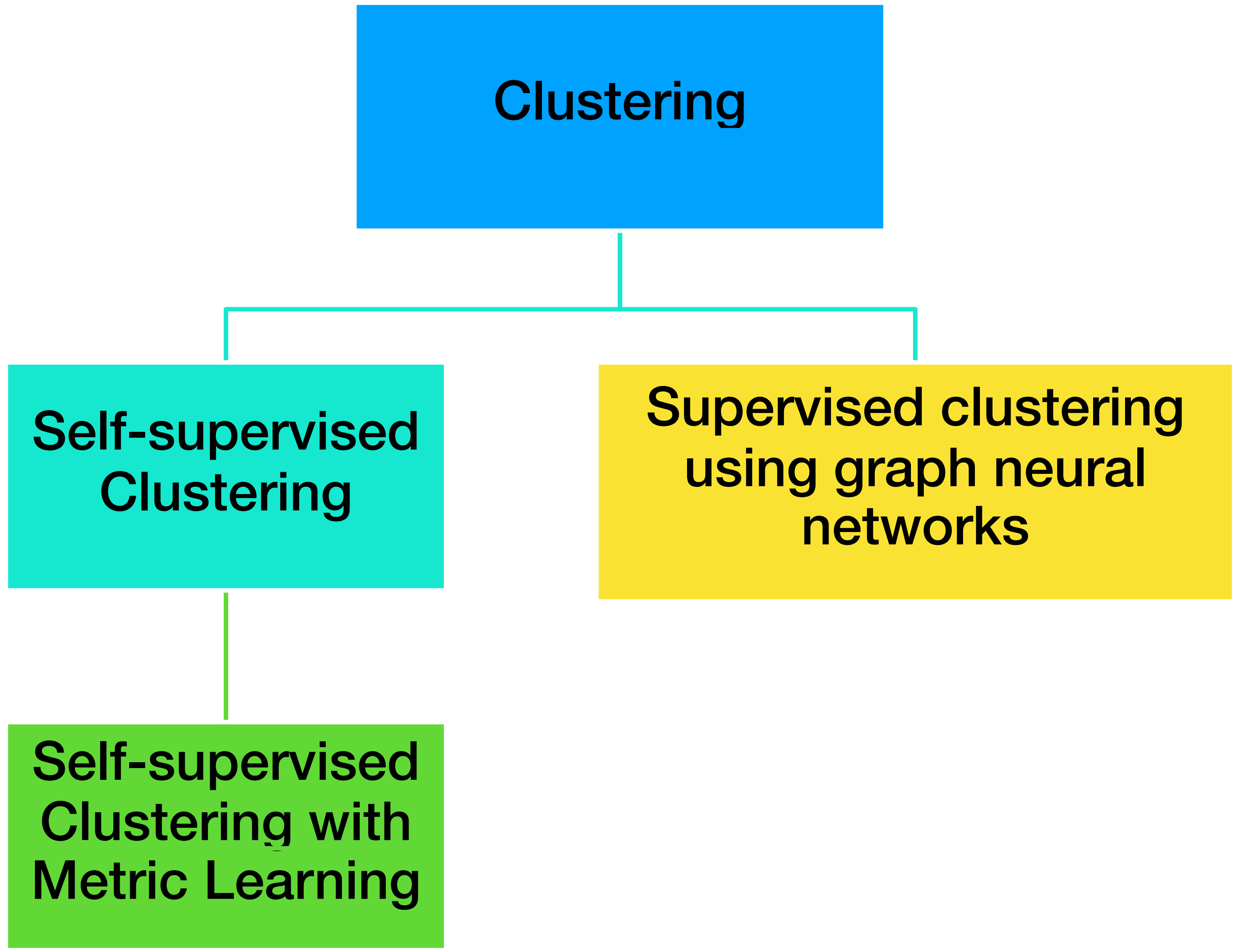
Contributions outline

- Clustering is a crucial step in speaker diarization as it enables
  - Accurate speaker segmentation
  - Turn-taking detection
  - Speaker model creation
  - Speaker adaptation, and evaluation
- Improving speaker embeddings can help improve clustering

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

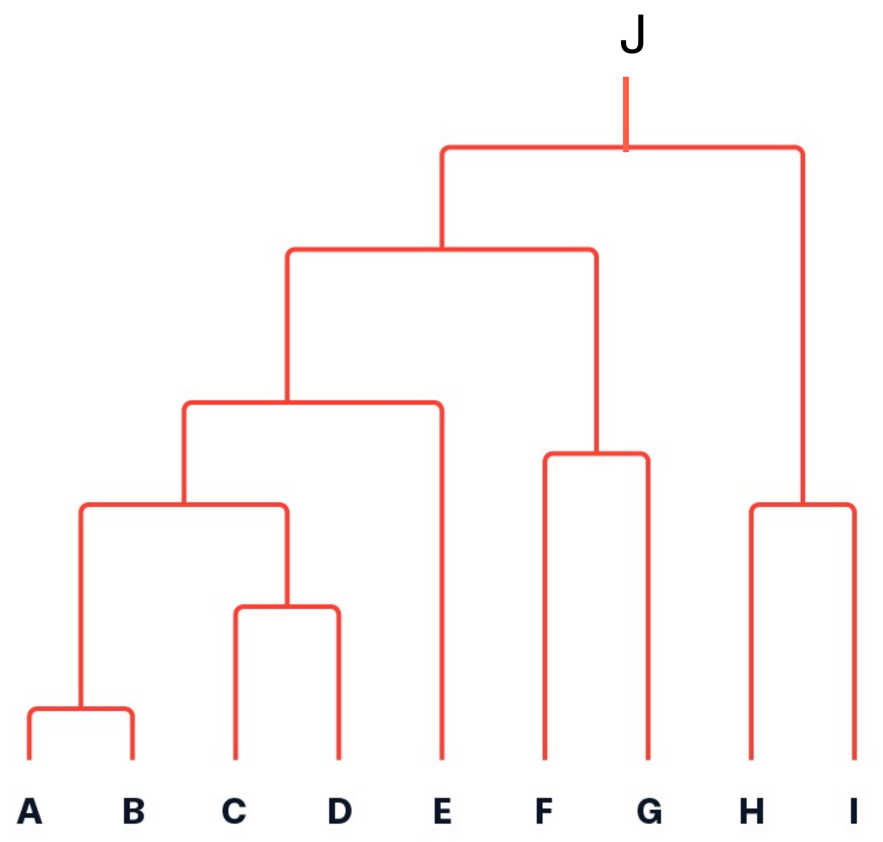
- Hierarchical clustering
- Graph Clustering

Unsupervised Clustering approaches

Hierarchical clustering

- Clusters are visually represented in a hierarchical tree called a dendrogram.

Agglomerative
(Bottom-up)



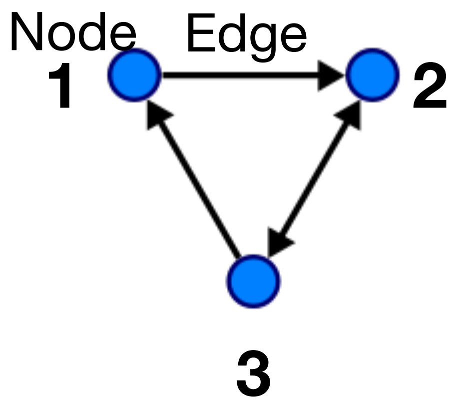
Divisive
(Top-down)

Example: Agglomerative Hierarchical clustering (AHC)

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**.



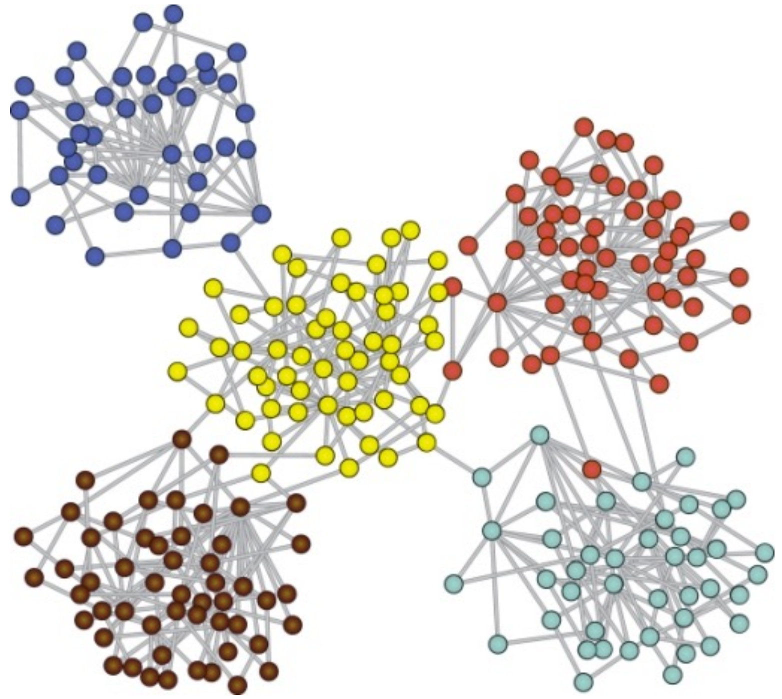
- **Adjacency matrix (A)** captures connections between nodes,
- $A_{ij} = 1, if\ Node\ i\ is\ connected\ \ to\ j\ by\ an\ edge$
- $A_{ij} = 0, if\ Node\ i\ and\ j\ are\ not\ connected$
- A with real weights to the edges is called as weighted adjacency matrix.

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

Graph clustering

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

Example: Spectral Clustering (SC)



Can we combine the two ?

- Why not !
  - Hierarchical Graph Clustering

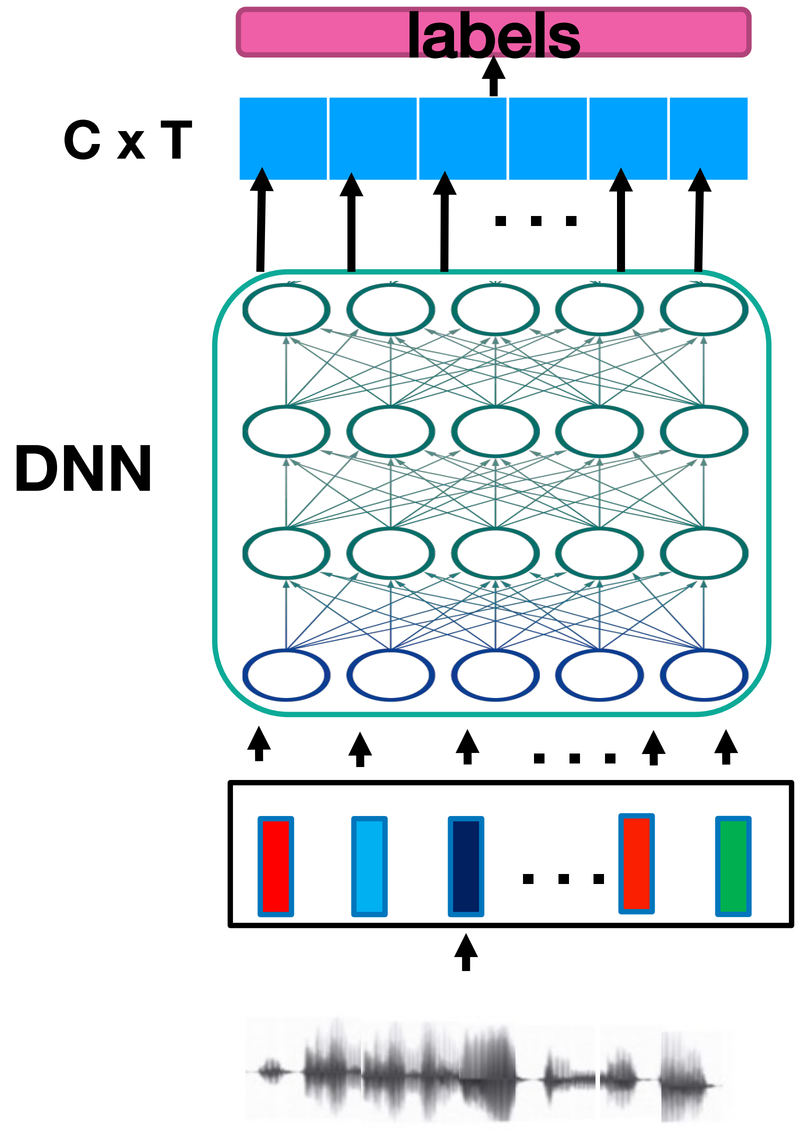
Related work

- 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)

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



EEND

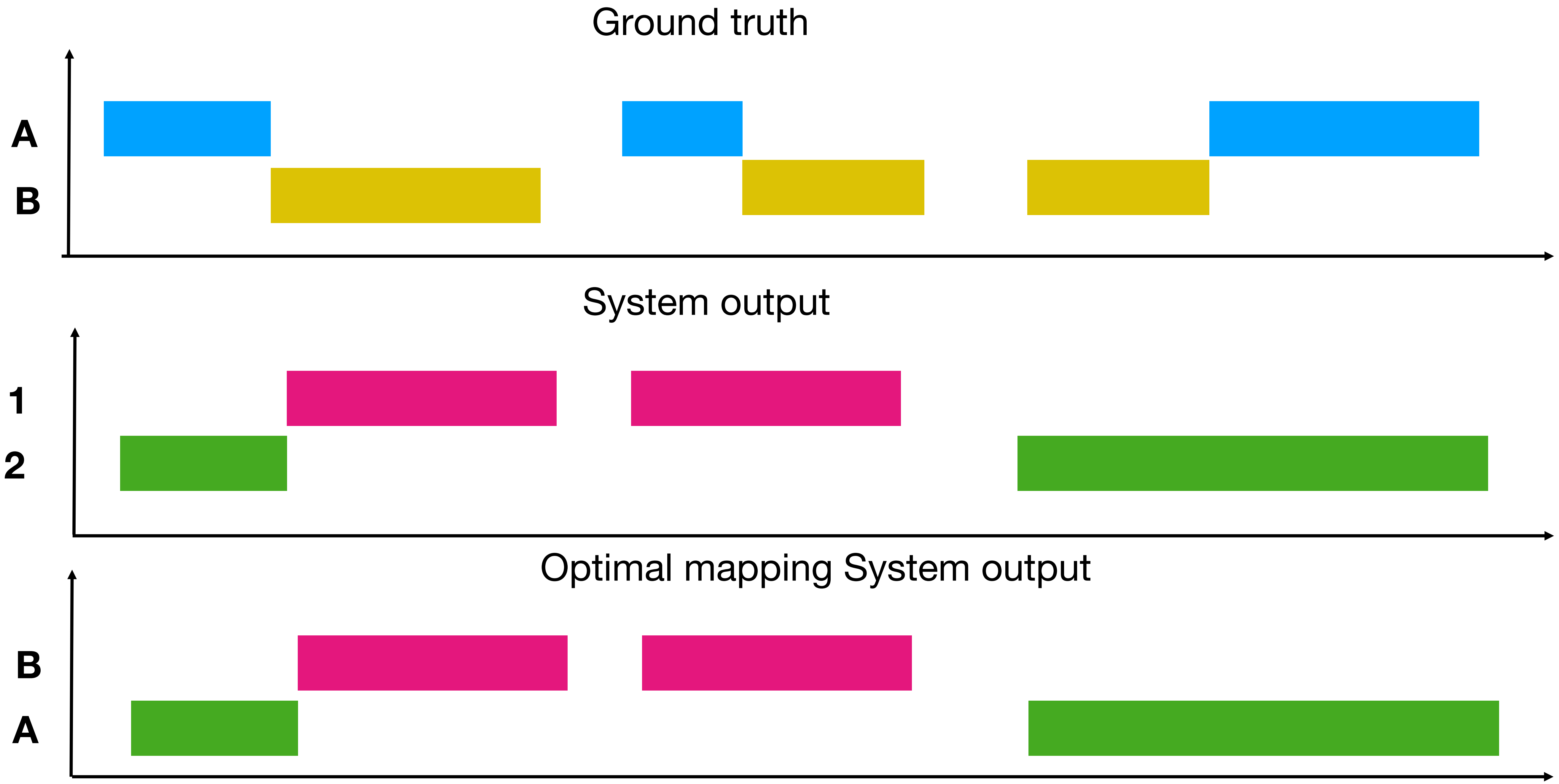
Cons:

- 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: $\mathrm{argmax}(A \cap 1, A \cap 2)$, $\mathrm{argmax}(B \cap 1, B \cap 2)$



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

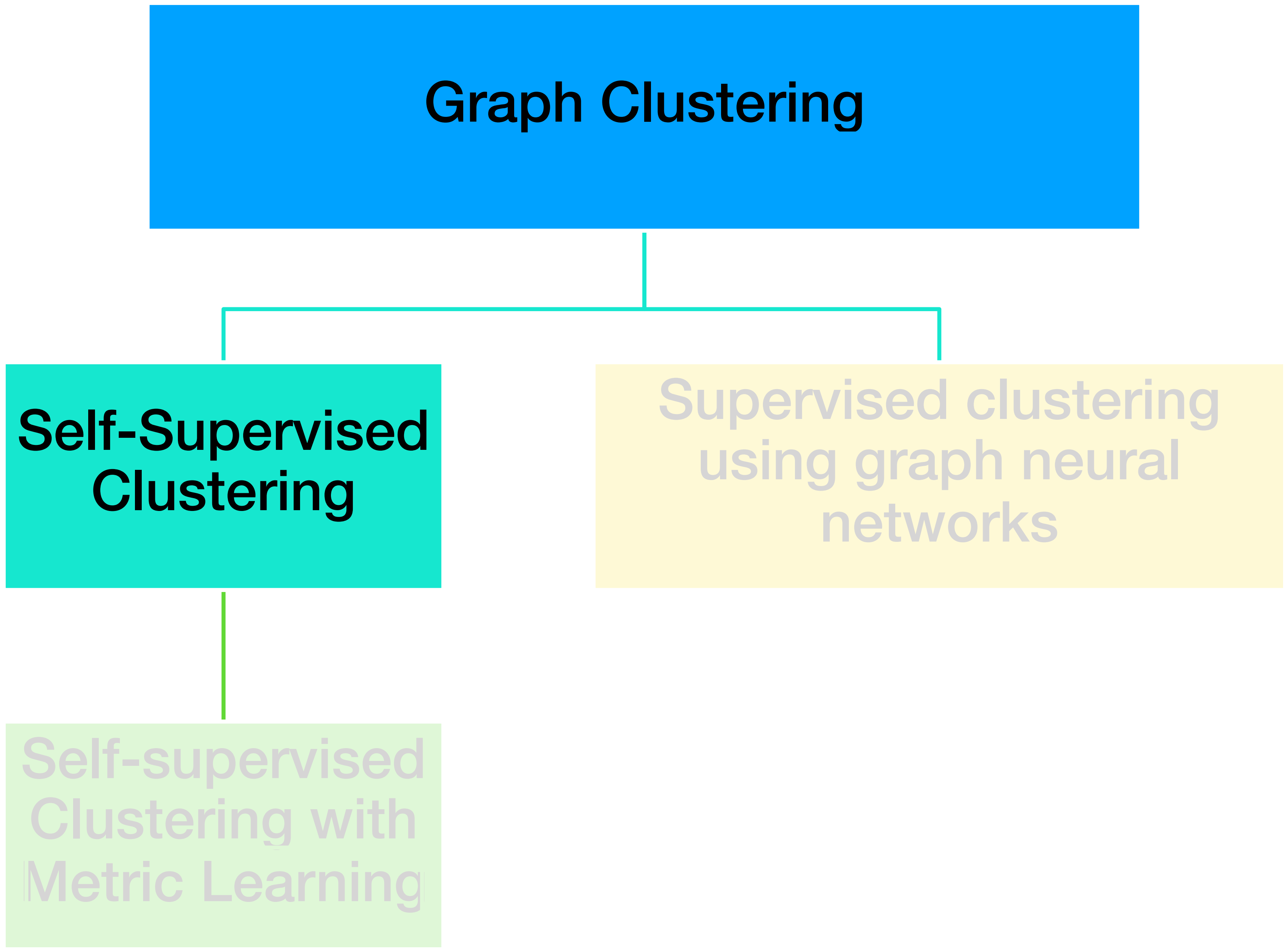
Narrowband (sampling rate: 8kHz)

Wideband (sampling rate: 16kHz)

CALLHOME (CH) [1]	AMI [2]	DIHARD III [3]	Voxconverse [4]
○ Multi-lingual telephone data ○ Recordings - 500, CH1 – 250, CH2- 250, ○ 2-5 mins ○ 2-7 speakers	○ Augmented Multi-party Interactions ○ Recorded at four different sites (Edinburgh, Idiap, TNO, Brno) ○ Recordings - Dev set: 18, Eval set: 16 ○ 20-60mins ○ 3-4 speakers	○ Speech diarization challenge data ○ 9-11 domains e.g, audiobooks, telephone recording, meetings, web videos. ○ Recordings – Dev set:254, Evalset:259 ○ 0.5-10 mins ○ 1-10 speakers	○ Voxconverse challenge data ○ Conversational dataset extracted from YouTube videos. ○ dev set: 216 and eval set:232 ○ 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 DNN**.
- Introduced **hierarchical graph clustering**.



Proposed Approach 1

Motivation

- The clustering approaches extract short-segment speaker embeddings from a pre-trained network (x-vectors) and perform unsupervised clustering.

- Each stage (embedding extraction and clustering) is optimized independently.

- The test set will contain unseen domains and speakers.

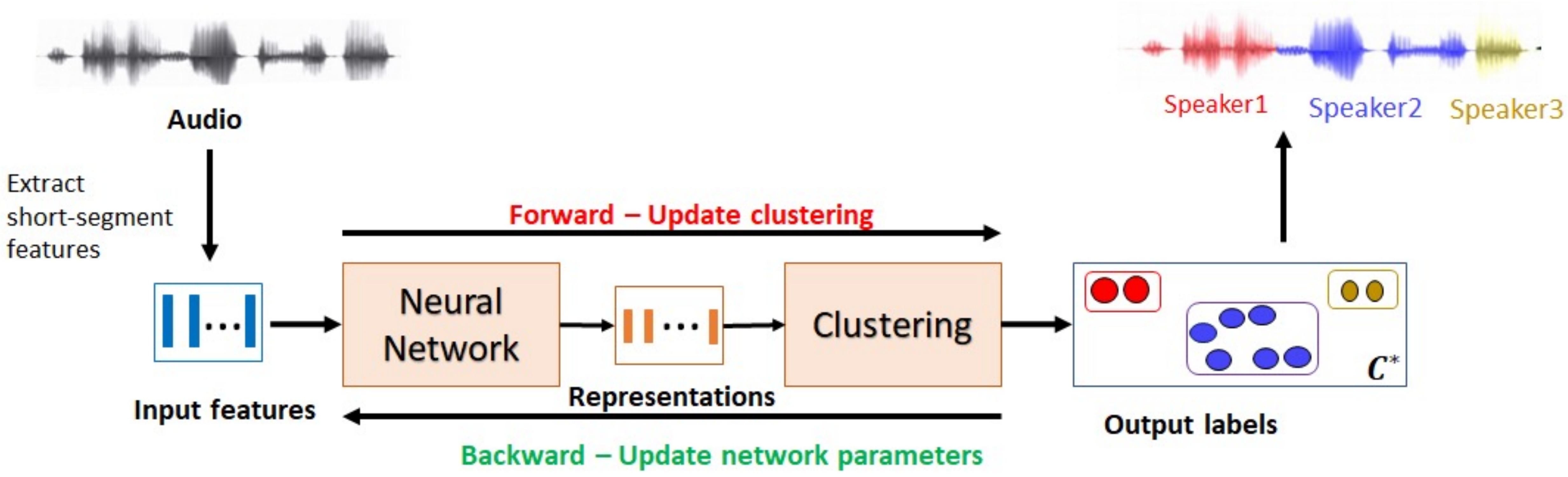
Motivation

- Succinct speaker representations are beneficial for clustering while clustering results can provide self-supervisory targets for representation learning[1].
- Creating a feedback loop from the output of the clustering algorithm to the input can help improve the representations used for clustering.
- This is referred to as self-supervised clustering (SSC).
- The data itself provides supervision labels for model learning

[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:

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

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

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

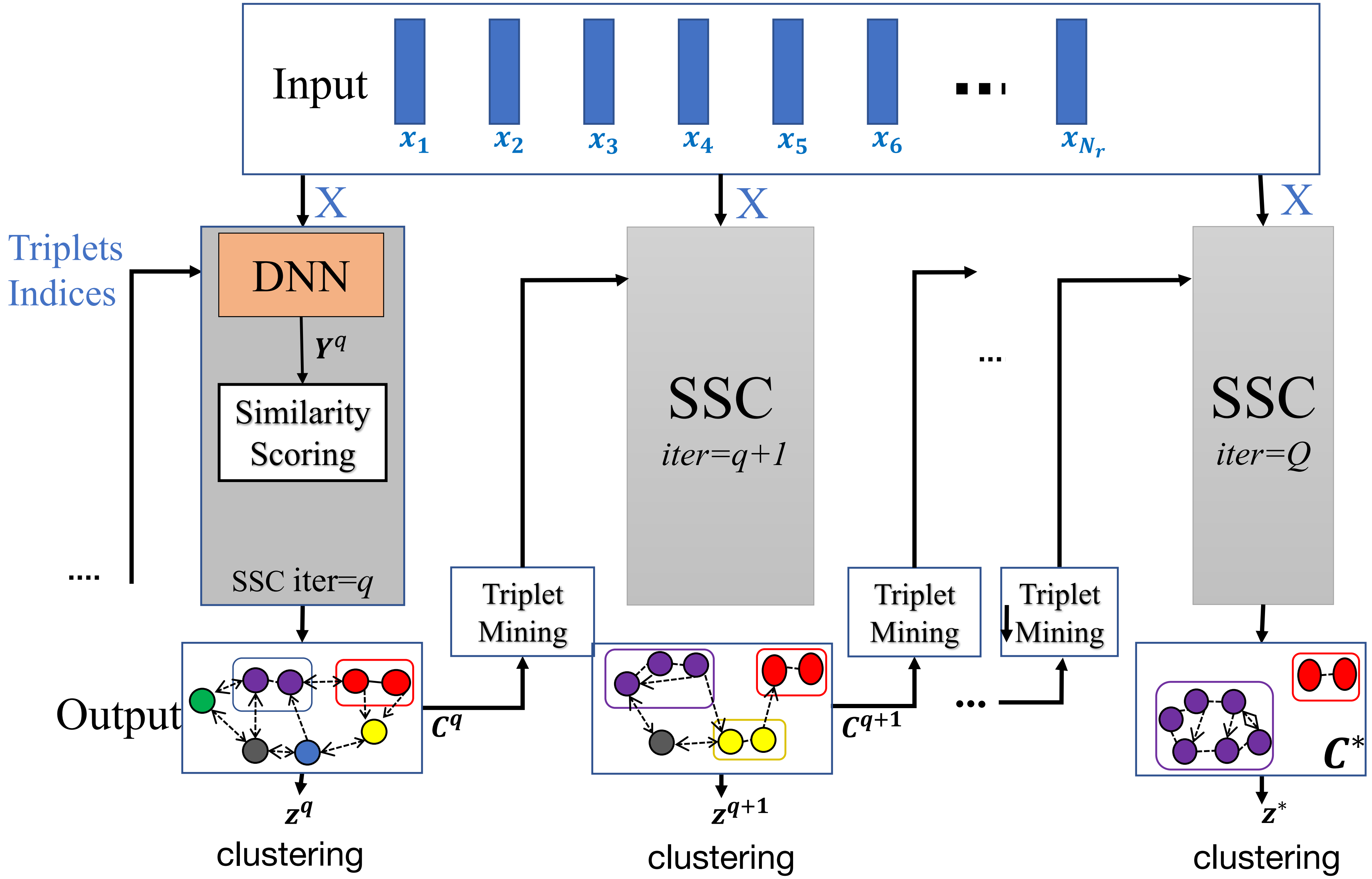
$\boldsymbol{\theta}$: DNN parameters

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

N^q: Number of clusters at iteration q

N^*: target number of clusters

For DNN 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).

DNN 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

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

(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.

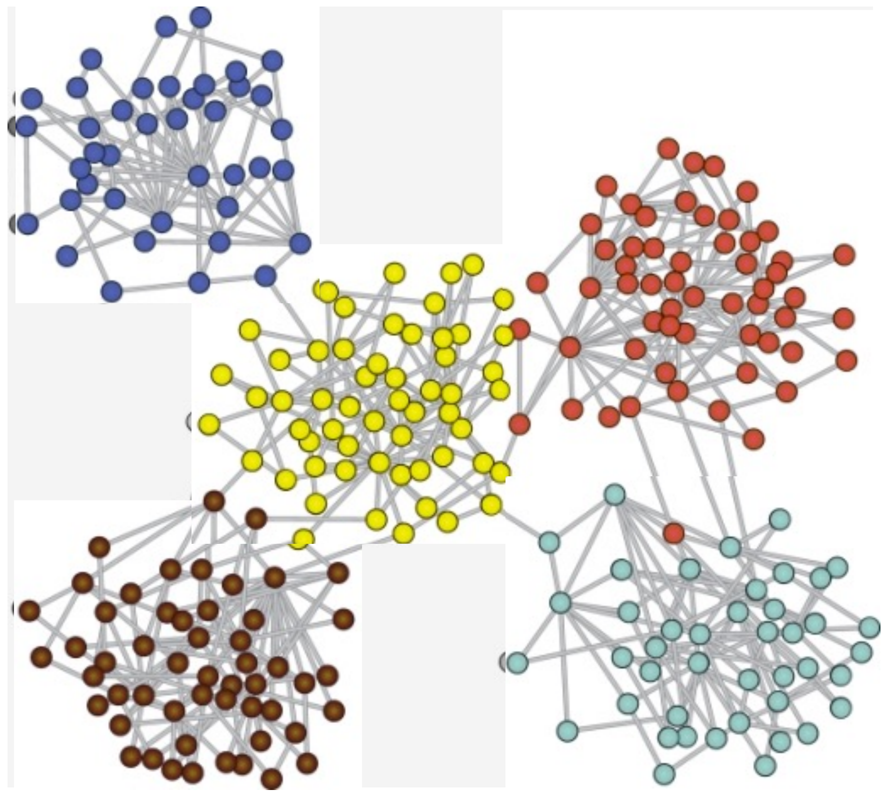
Path Integral Clustering (PIC)

- 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,

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

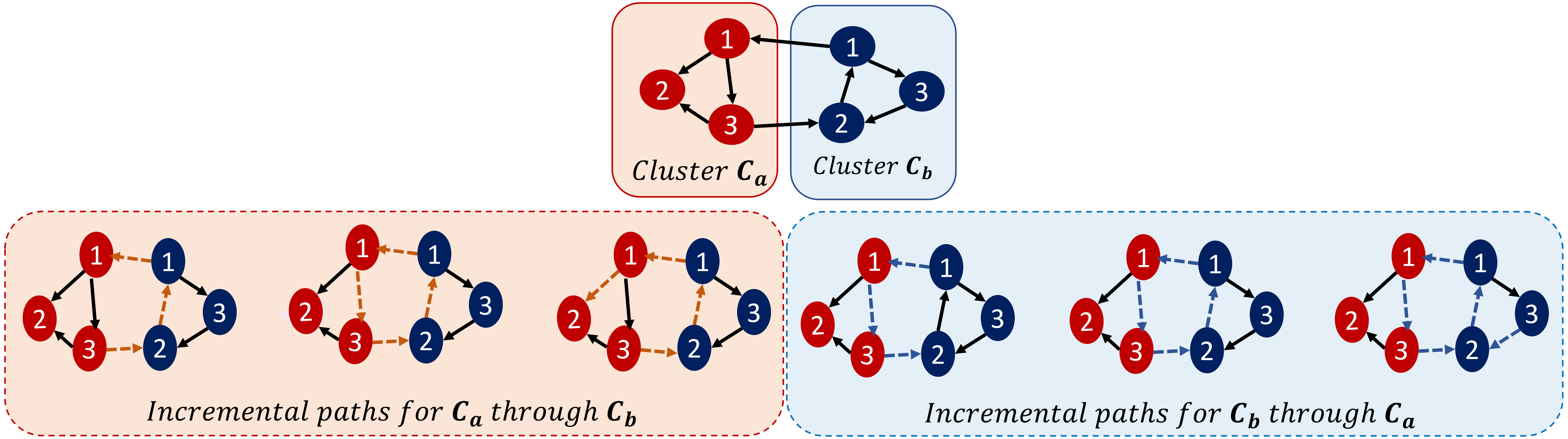
where, S(i,j) is the pairwise similarity between x_i and x_j, N_i^K is the set of K nearest neighbour of x_i



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

PIC illustration

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



$S_{C_a|C_aUC_b} - S_{C_a}$

$S_{C_b|C_aUC_b} - S_{C_b}$

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 for PLDA scoring
- 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

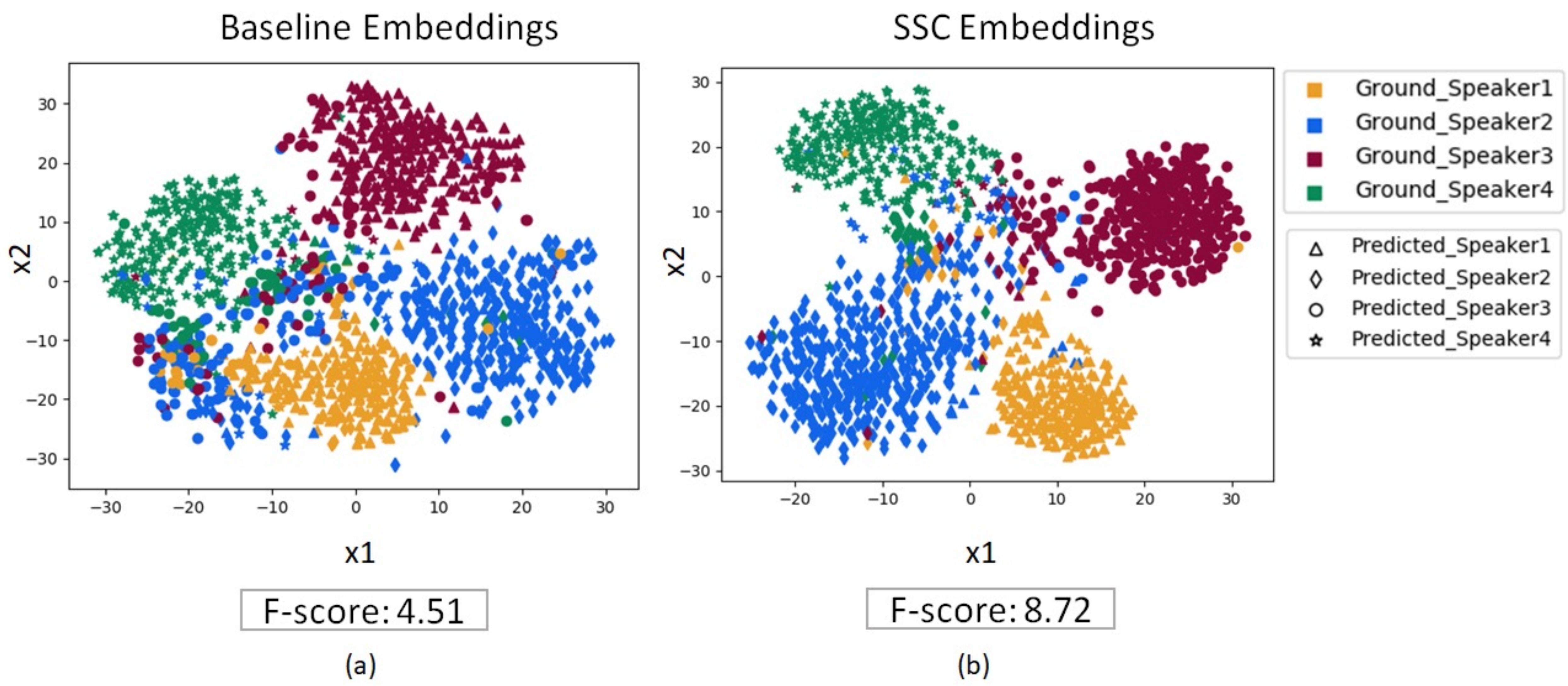
System	Known N*		Unknown N*	
	Dev.	Eval.	Dev.	Eval.
x-vec + cosine + AHC	34.6	30.2	18.2	15.5
x-vec + cosine +SC	30.2	25.5	40.0	31.1
x-vec + PLDA + AHC (Baseline)	15.7	16.0	13.7	16.3
SSC-PLDA-AHC	9.4	11.1	10.7	11.6
x-vec + PLDA + PIC	9.4	9.3	9.8	10.4
x-vec + cosine + PIC	8.9	7.3	9.0	7.3
SSC-PIC	7.3	7.2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

LEAP

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

AMI Visualization



t-SNE based visualization of embeddings extracted on 1.5s audio segments from the meeting dataset.

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.

AMI Results

DER comparison with other published works

System	Known N*		Unknown N*	
	Dev.	Eval.	Dev.	Eval.
Semi-sup learning[1]	-	-	17.5	22.0
Incremental[2] learning	-	15.6	-	20.0
GAN clustering[3]	**10.2**	**10.1**	**11.0**	**11.3**
2D self-attention[4]	-	-	12.2	13.0
Baseline	14.4	16.5	12.9	13.6
SSC-PIC	**4.6**	**6.5**	**5.2**	**5.4**

[1]Pal et al., A study of semi-supervised speaker diarization system using GAN mixture mode, 2019
[2]Dawalatabad et al., Incremental Transfer Learning in Two-pass Information Bottleneck Based Speaker Diarization System for Meetings, 2019
[3]Pal et al.,Speaker diarization using latent space clustering in generative adversarial network, 2020
[4]Sun et al., Speaker diarisation using {2D} self-attentive combination of embeddings, 2019

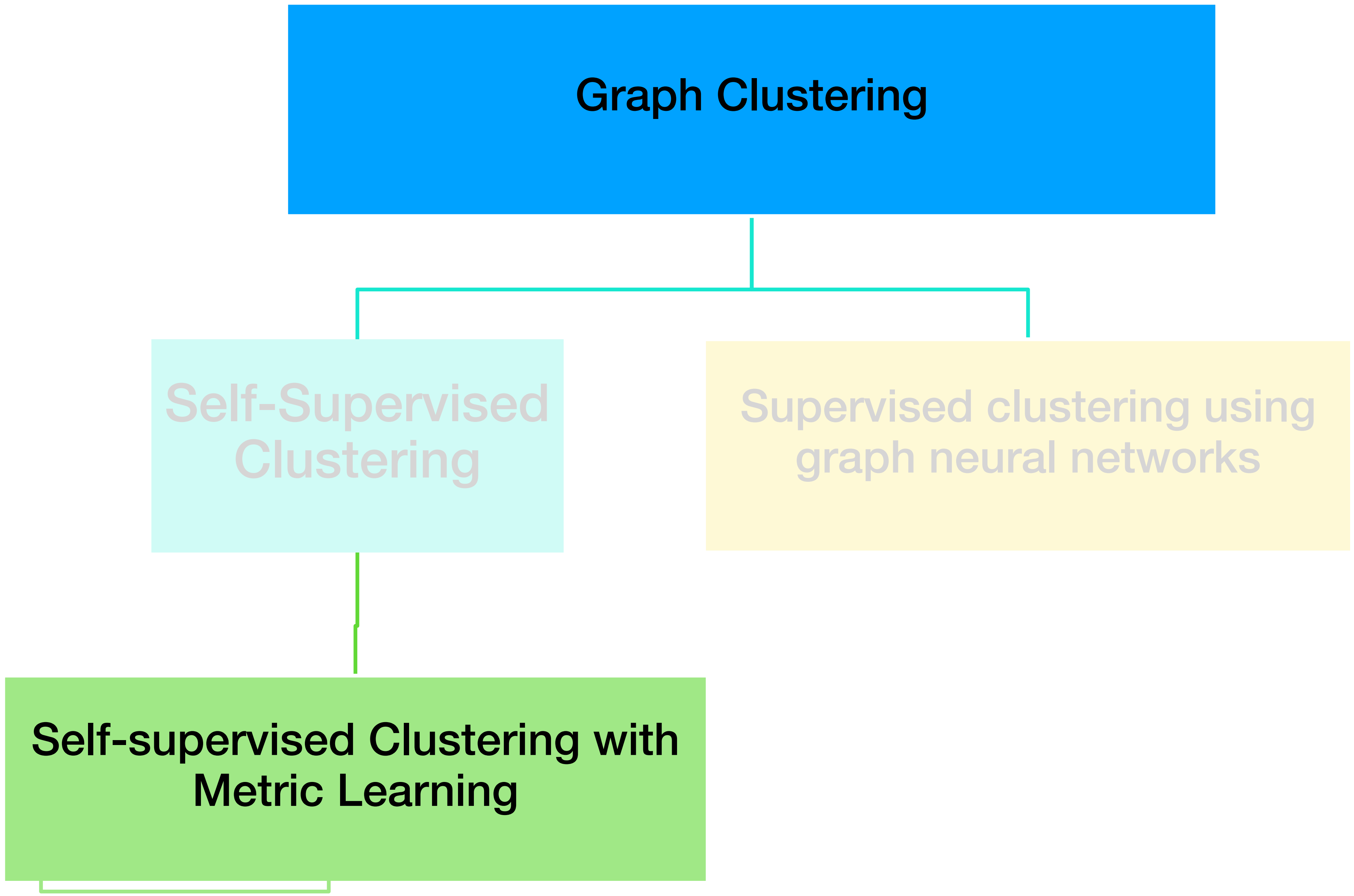
Summary

- 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.**

- Encourages separation between representations of different speakers.

- 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 using Expectation Maximization (EM).
- Can we train the SSC with learnable scoring/metric function?
  - Yes. SelfSup-PLDA-PIC.

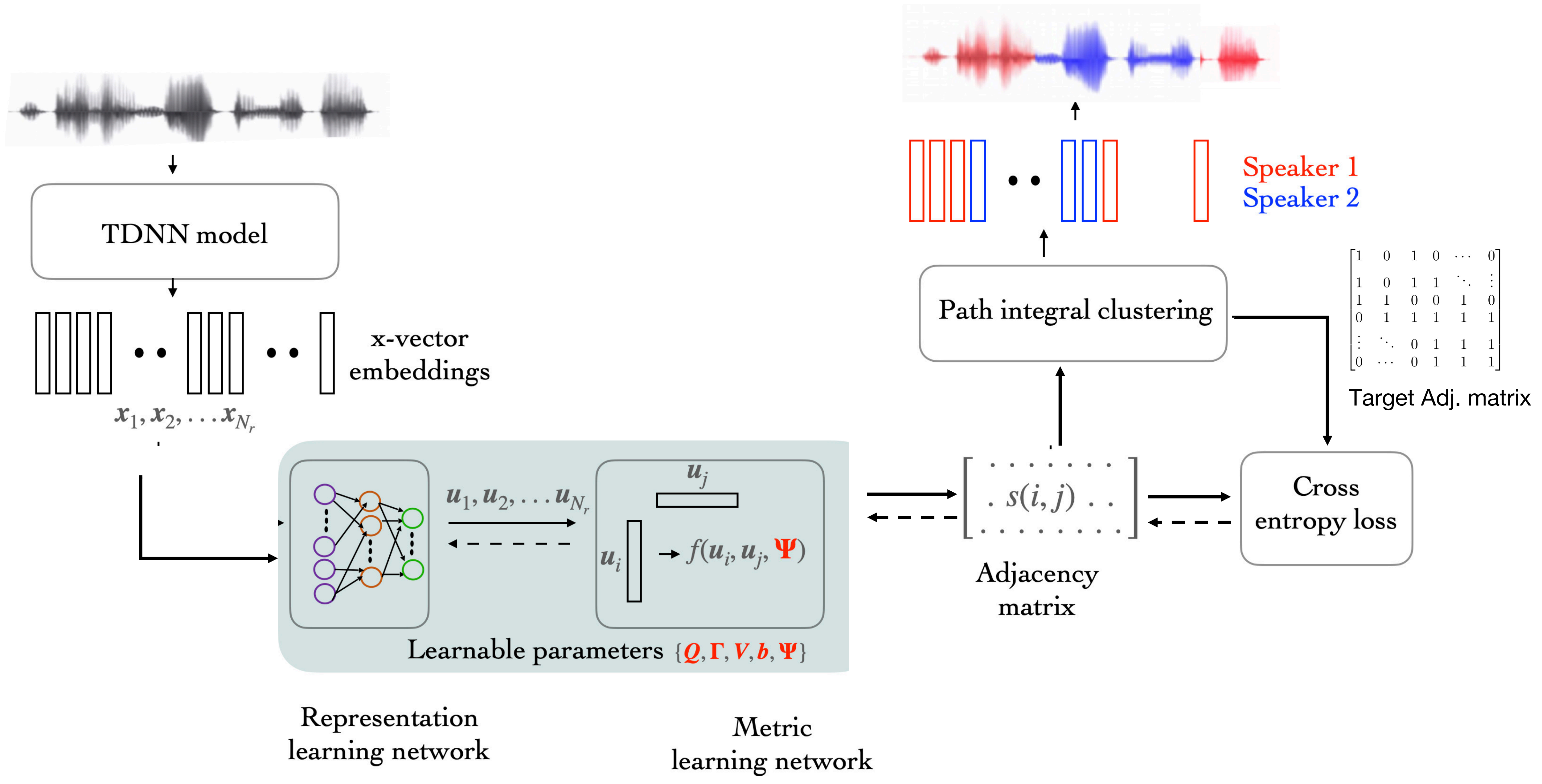
SelfSup-PLDA-PIC

- **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

u_i, u_j → PLDA Model ($f(u_i, u_j, \boldsymbol{\Psi})$) →

$$s(i,j) = \log\left[\frac{p(u_i, u_j | H_s)}{p(u_i|H_d)\,p(u_j|H_d)}\right]$$

Same-speaker hypothesis

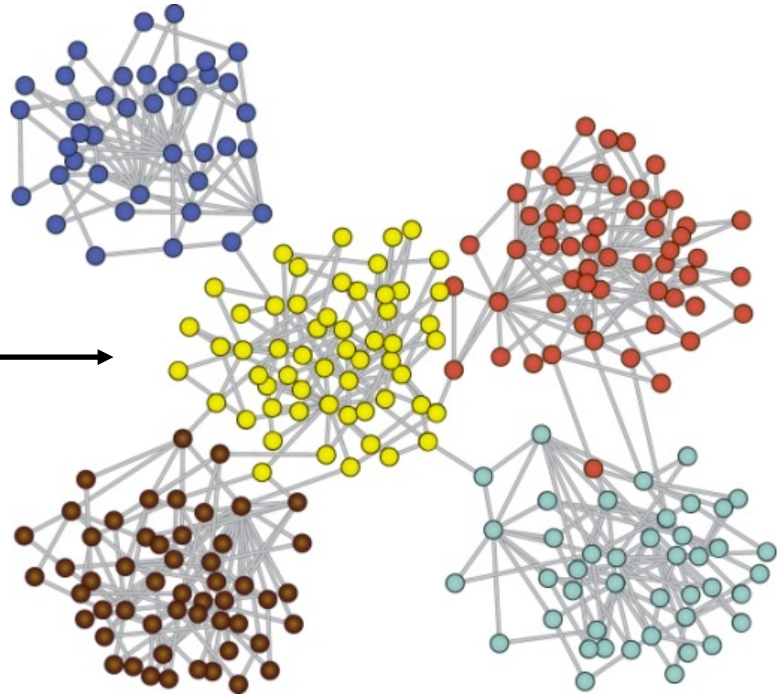
different-speakers hypothesis

Metric Learning using PLDA model

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

$u_i \rightarrow$ **Metric learning model** $(f(u_i, u_j, \boldsymbol{\Psi}))$ $\leftarrow u_j$

$\rightarrow s(i, j) \rightarrow$



PIC

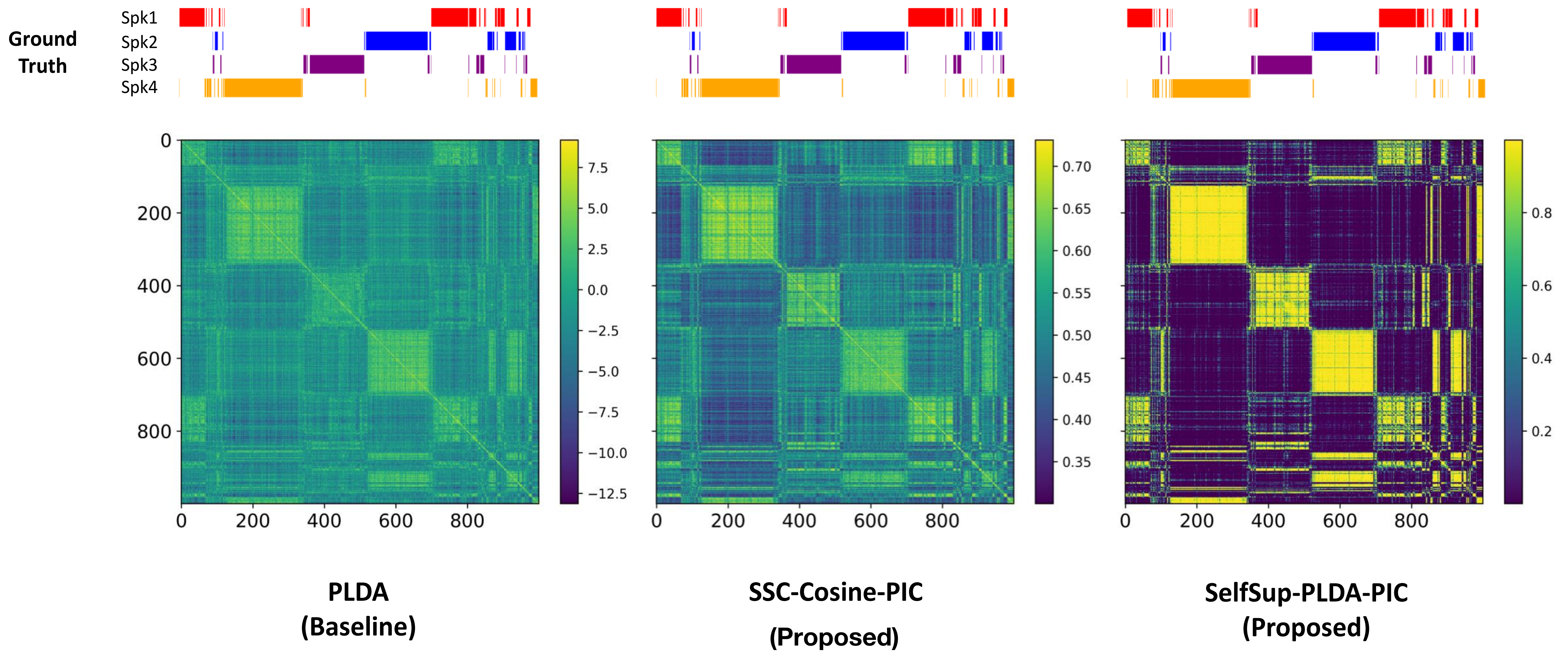
AMI Results

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

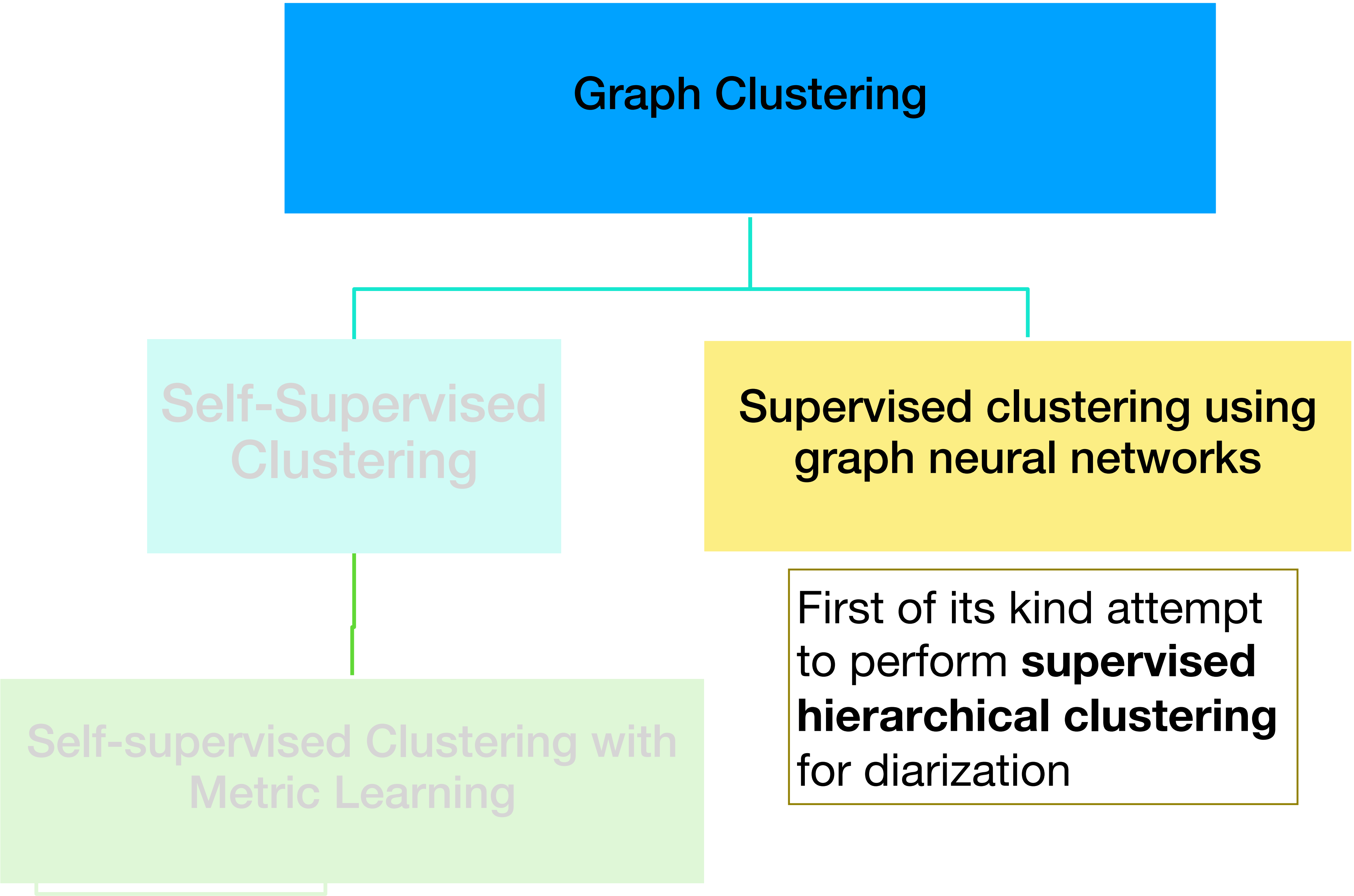
Average DER (%) on the DIHARD dataset considering overlapping regions with no tolerance collar.

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

Summary

- Proposed **self-supervised metric learning** approach using **PLDA**.
- Increases inter-speaker distance and decreases intra-speaker distance.
- Performance degrades as number of speakers increases as initial clustering becomes unreliable.



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 model with the clustering objective**?

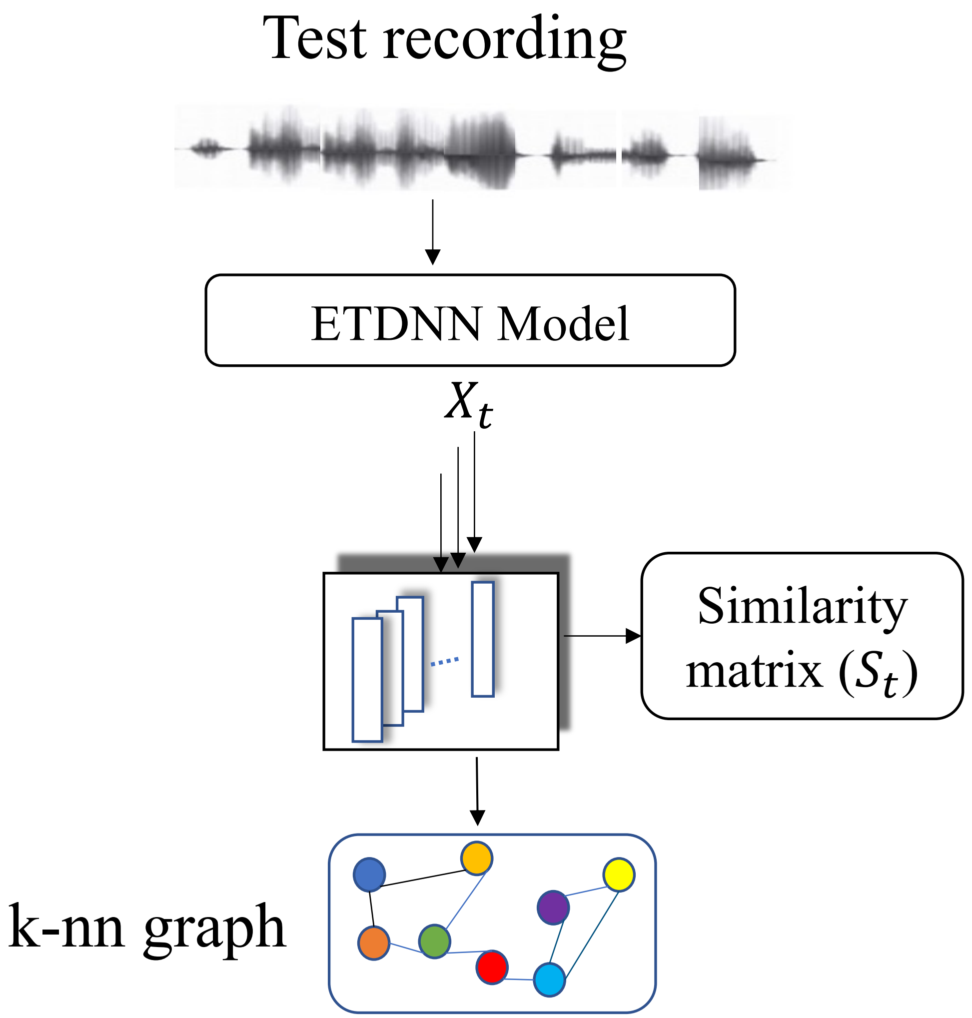
Supervised HierArchical GRaph Clustering (SHARC)

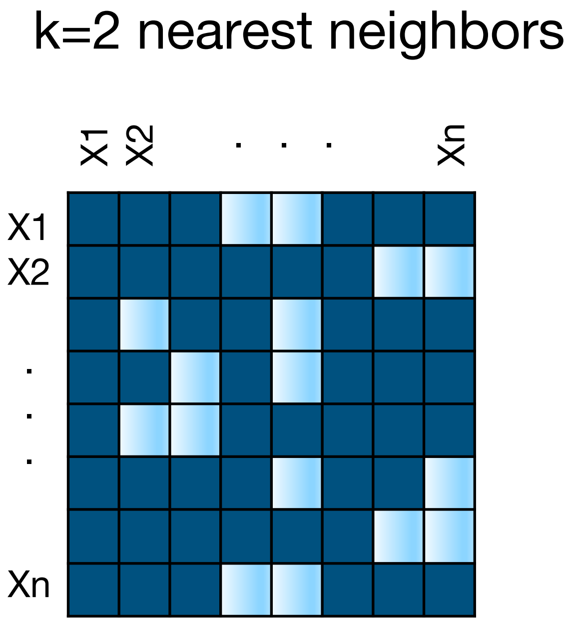
- Performs **supervised clustering** using **Graph Neural Networks (GNN).**
- Represents the speaker embeddings using graph.
- Clustering loss is used to update edges of the graph.
- Generates node labels based on clustering performed on updated edges at **each level of hierarchy**.

SHARC components

- Graph Generation
- GNN scoring
- Clustering
- Aggregation

Graph generation





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,

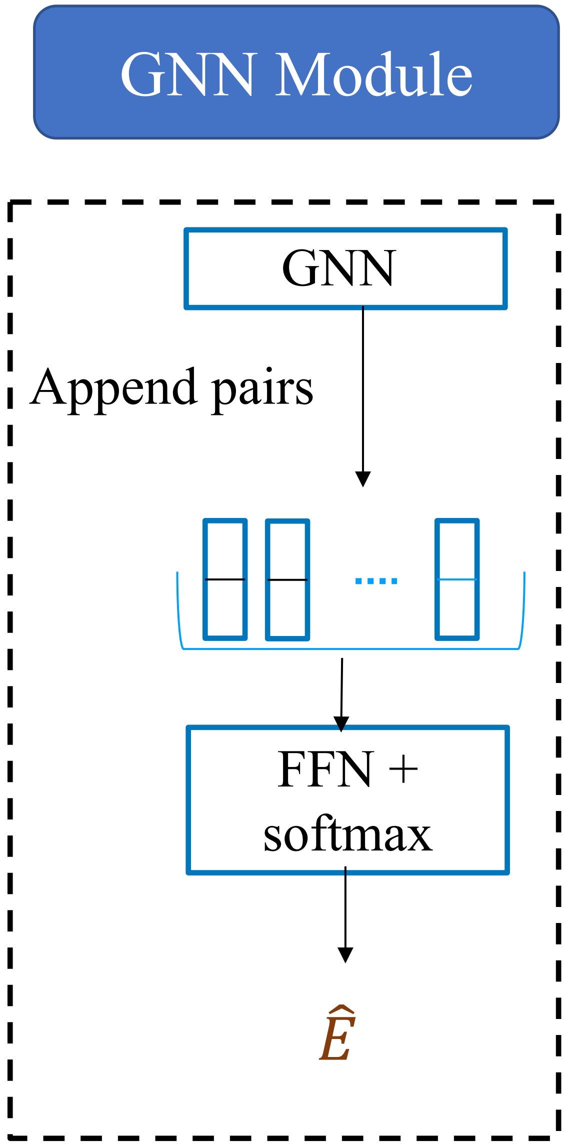
$$\hat{e}_{ij} = 2p_{ij} - 1 \in [-1, 1] \forall j \in N_i^k$$

- Density of node i :
- Ground truth:

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

- 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)

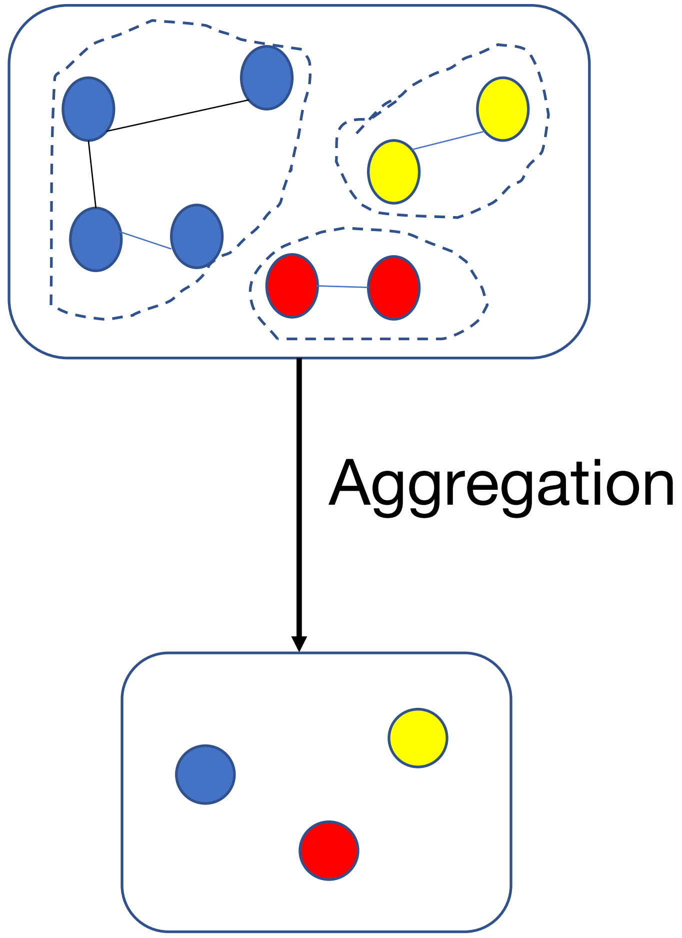
$$\varepsilon(i) = \{j|(v_i, v_j) \in E_m, \quad \hat{d}_i \leq \hat{d}_j \quad \text{and} \quad p_{ij} \geq p_\tau\}$$

- For any i, if ε(i) is not empty, we pick $j = \text{argmax}_{j \in \varepsilon(i)} \hat{e}_{ij}$ and add (vi,vj) to E'_m

- A set of connected components C'_m , forms clusters for the next level (m + 1).

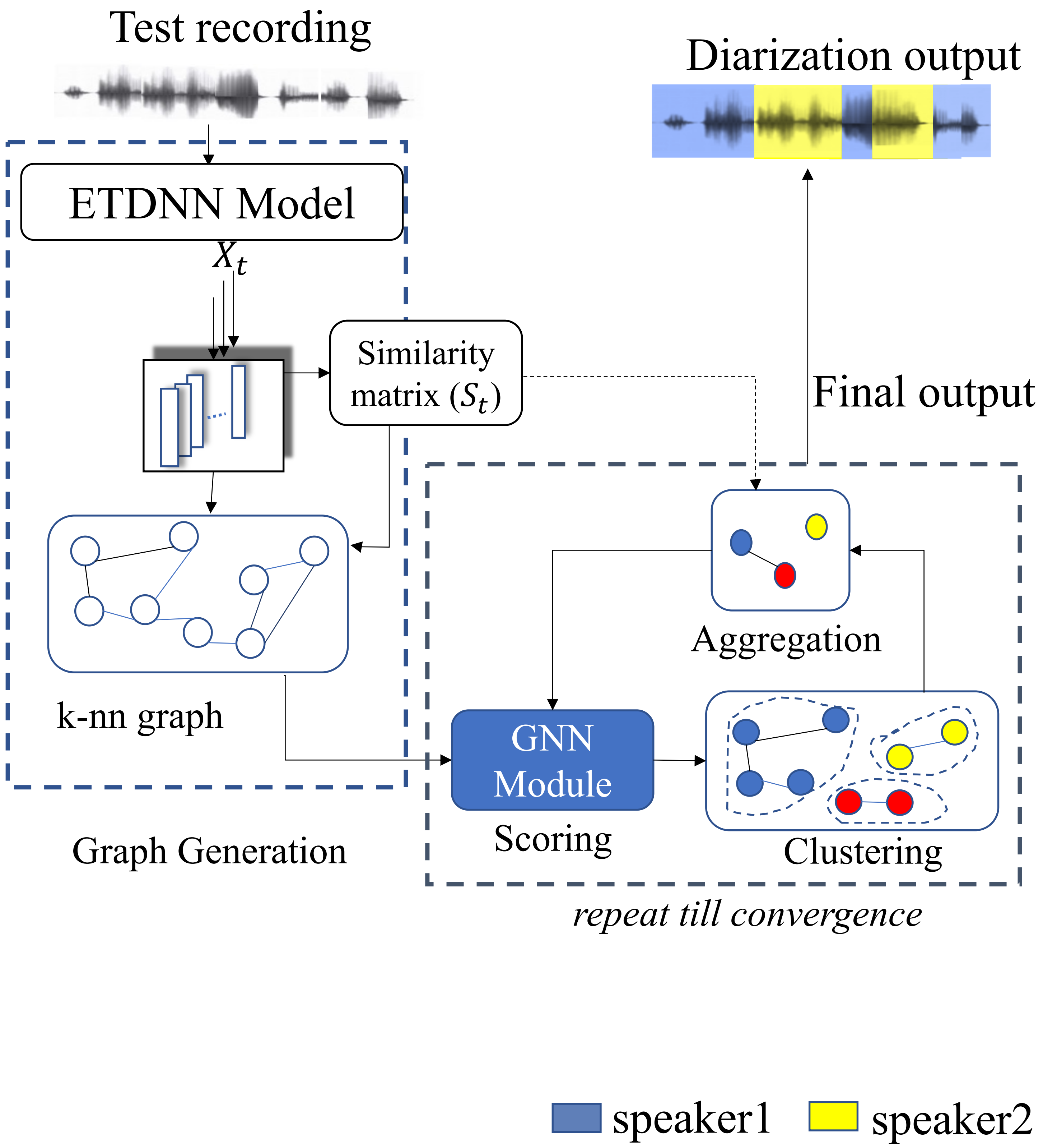
Feature Aggregation

The aggregation function Ψ - obtain node representations for next level.



[1]Prachi Singh, Amrit Kaul, Sriram Ganapathy, "Supervised Hierarchical Clustering Using Graph Neural Networks For Speaker Diarization", accepted in ICASSP 2023

Block diagram: SHARC Inference



SHARC Components

1. Graph Generation
2. GNN scoring
3. Clustering
4. Aggregation

Training loss

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

  q_{ij}- Ground truth edge labels, 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

Experiments

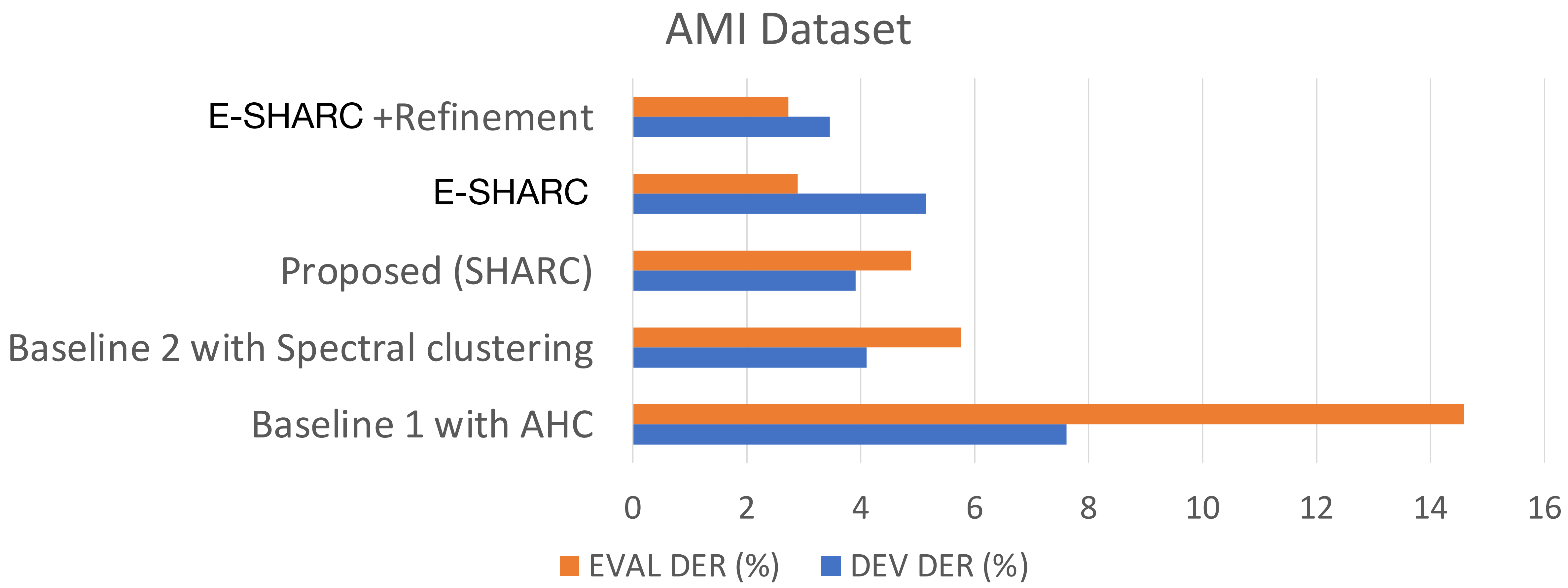
Datasets

AMI : Meeting dataset

Voxconverse: Youtube videos

Results

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

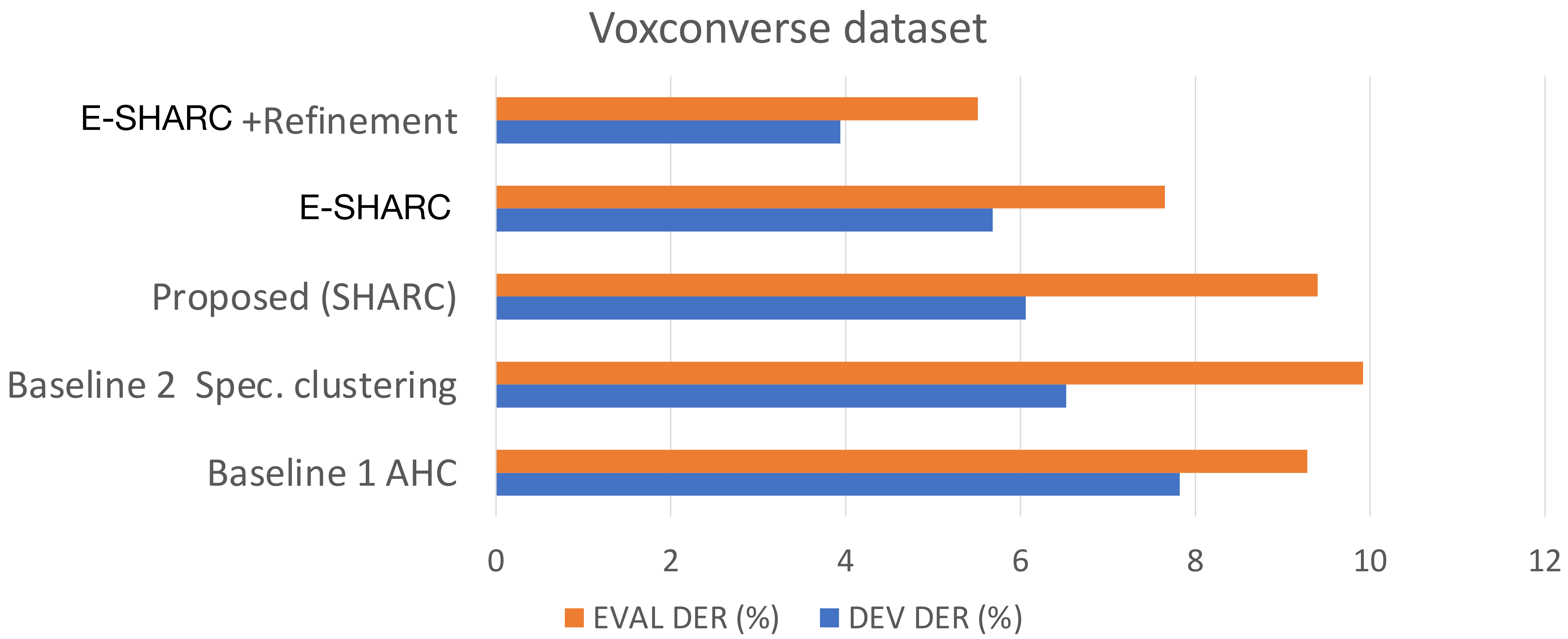


53% relative improvement over best baseline

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

Cluster Purity and Coverage

Purity: The percentage of segments from predicted speaker belong to one speaker in ground truth

Coverage: The percentage of segments from ground truth speaker is covered by predicted speaker.

Voxconverse

Method	Purity	Coverage
Baseline with AHC	93.5	89.5
Baseline with SC	92.0	92.3
SHARC	93.0	92.4
E-SHARC	93.0	92.9

Results

Results with pyannote overlap detection[1]

AMI	Eval DER (%)
AHC + overlap	26.30
SC + overlap	18.10
SHARC + overlap	19.32
E-SHARC + overlap	17.95
Voxconverse	**Eval DER (%)**
AHC + overlap	11.66
SC + overlap	10.73
SHARC + overlap	10.89
E-SHARC + overlap	10.44

LEAP

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

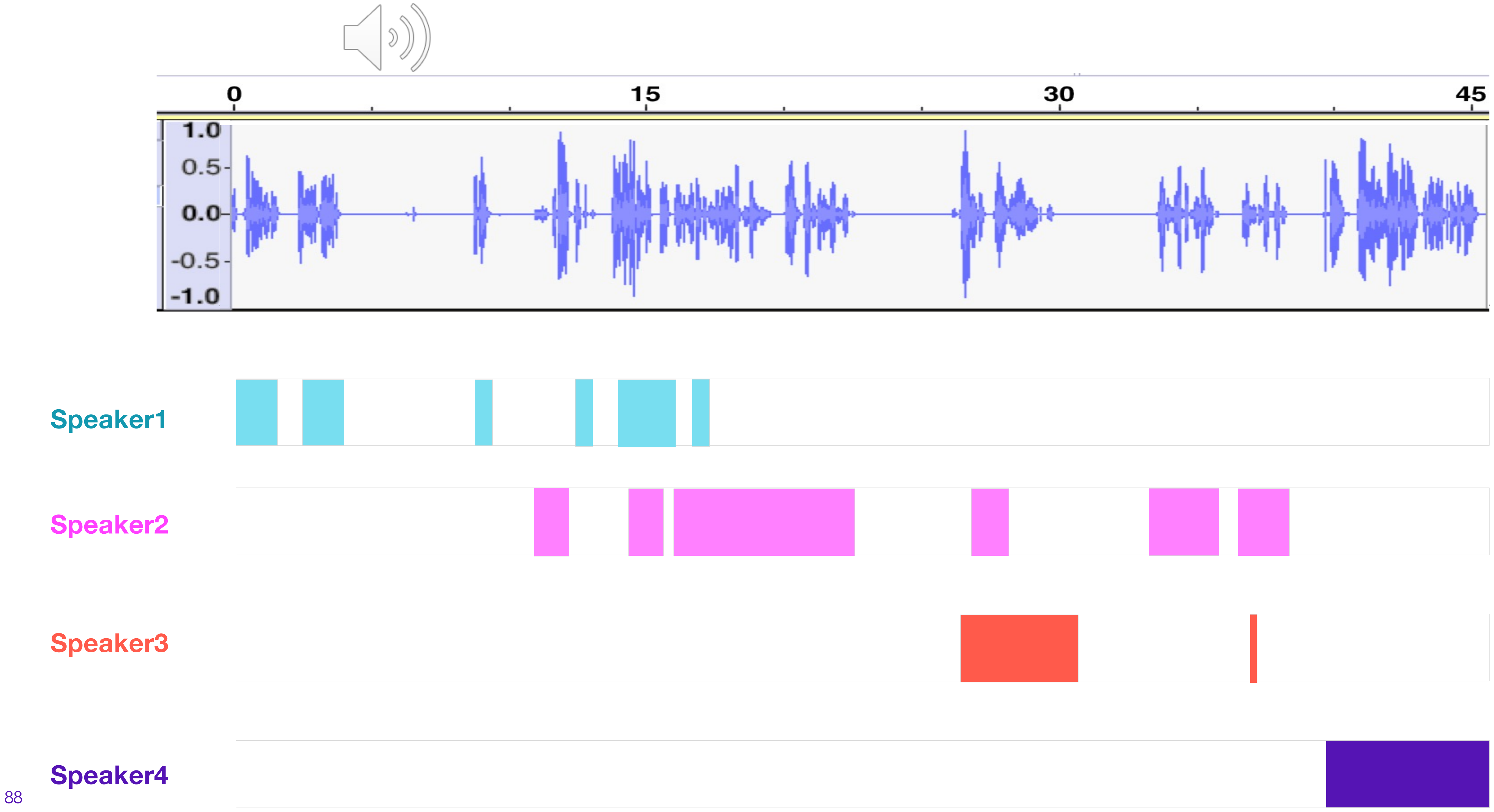
INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

Summary

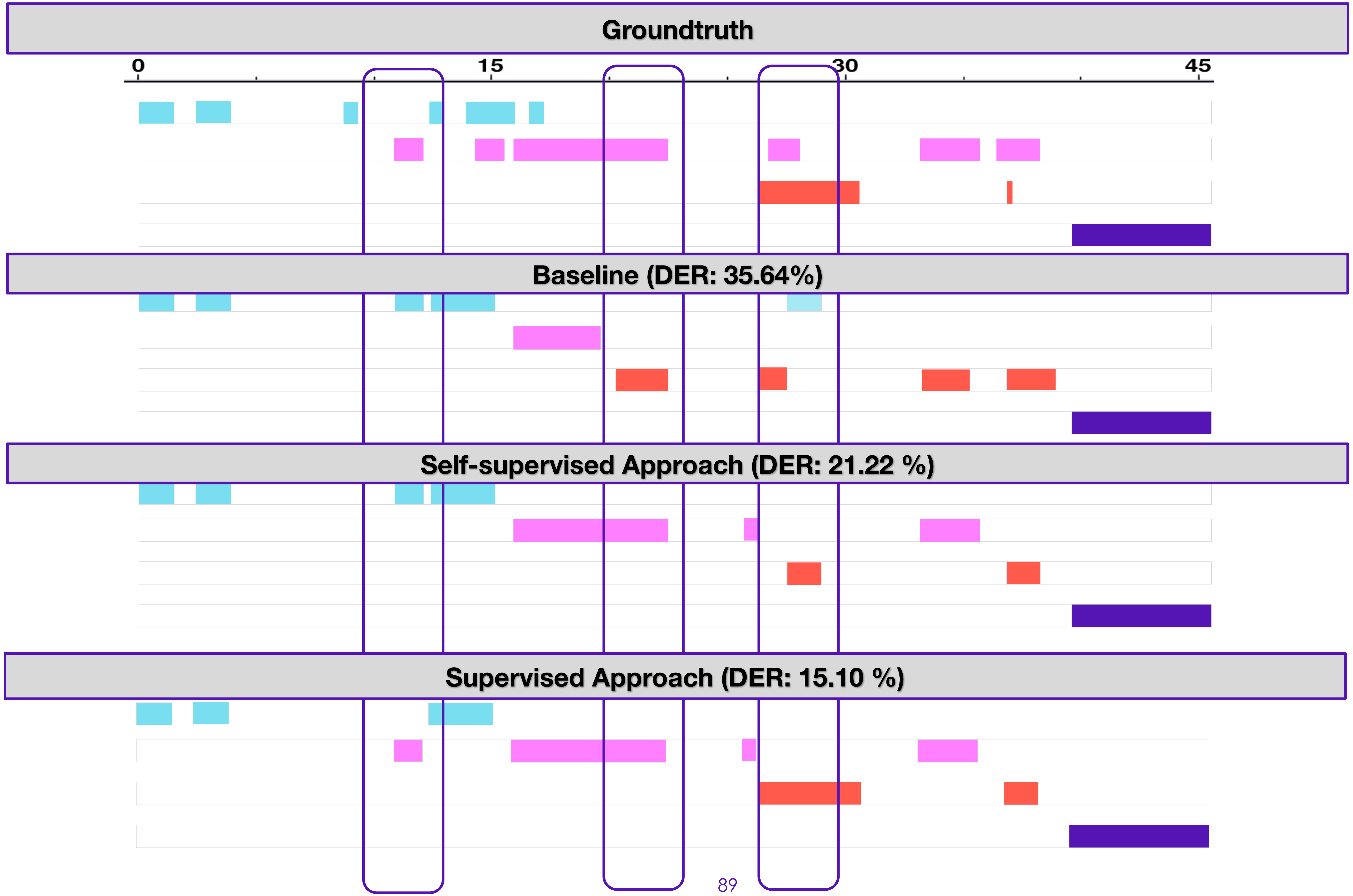
- Introduced **supervised hierarchical clustering for speaker diarization for the first time**.
- Designed an **end-to-end approach** to perform speaker diarization using **Graph Neural Networks**.
- Achieved state-of-the-art performance on two benchmark datasets.

Conclusion and Future Directions

Real world Example



Results


Groundtruth
0
15
30
45
Baseline (DER: 35.64%)
Self-supervised Approach (DER: 21.22 %)
Supervised Approach (DER: 15.10 %)

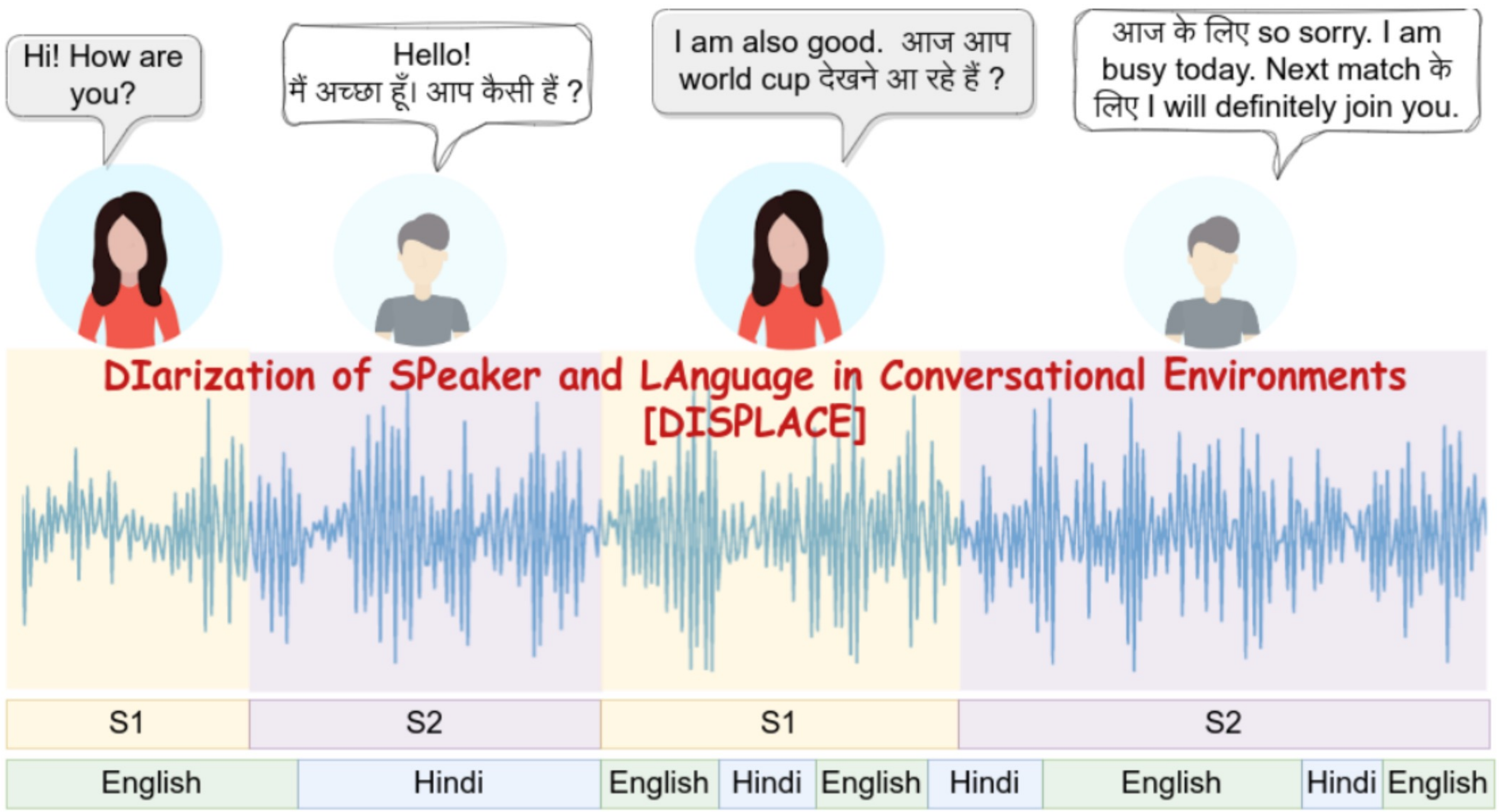
89

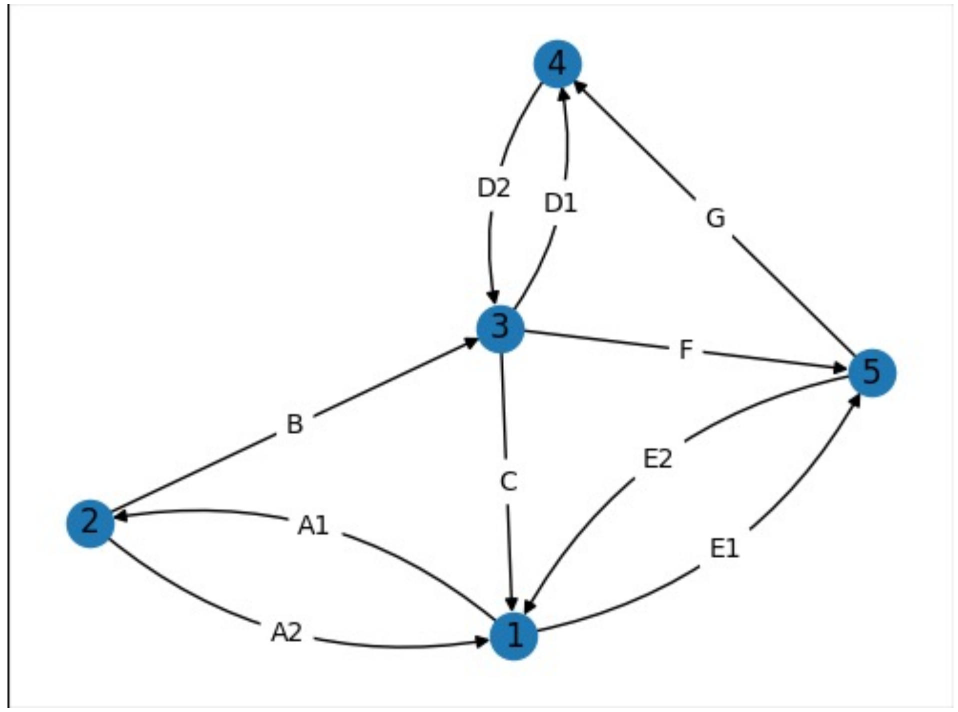
Concluding remarks

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 • Overlap detection can be added

Future Directions

Multilingual conversation Diarization



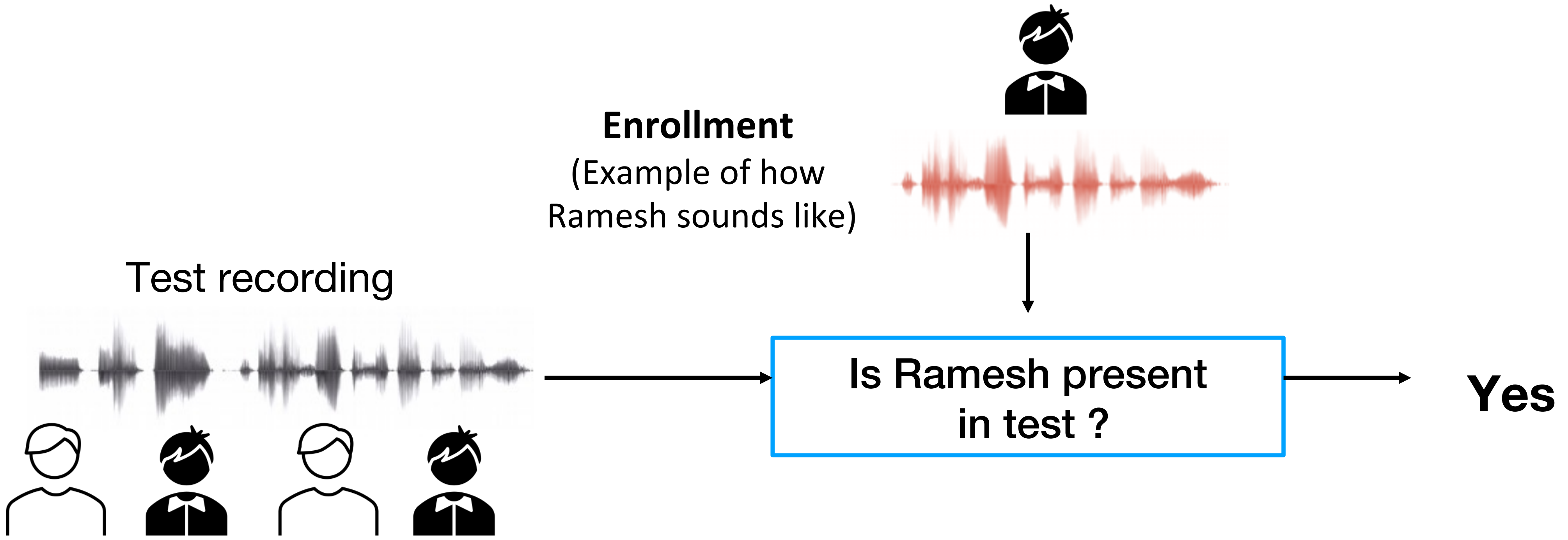


Use Multi-edge graph to perform multi-task learning

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:**
  - **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).
  - **P. Singh** and S. Ganapathy, “Speaker Diarization with Graph Based Supervised Hierarchical Clustering” (under preparation).

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