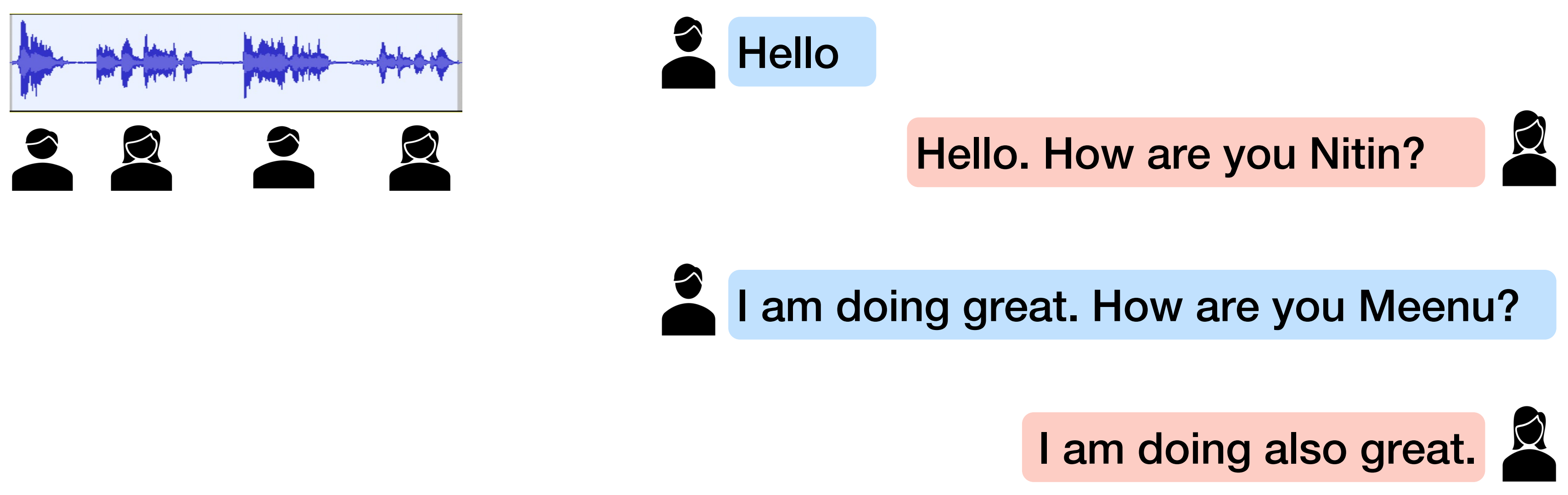
ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

Supervised Hierarchical Clustering using Graph Neural Networks for Speaker Diarization

Prachi Singh, Amrit Kaul, Sriram Ganapathy,
LEAP Lab, Dept. of Electrical Engineering,
Indian Institute of Science, Bangalore

Introduction

- **Conversational audio** contains multiple speakers engaged in a conversation.
- **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.

Speaker diarization: the task of partitioning and labelling an input audio file into segments based on speaker identity.

Prior work- Drawbacks

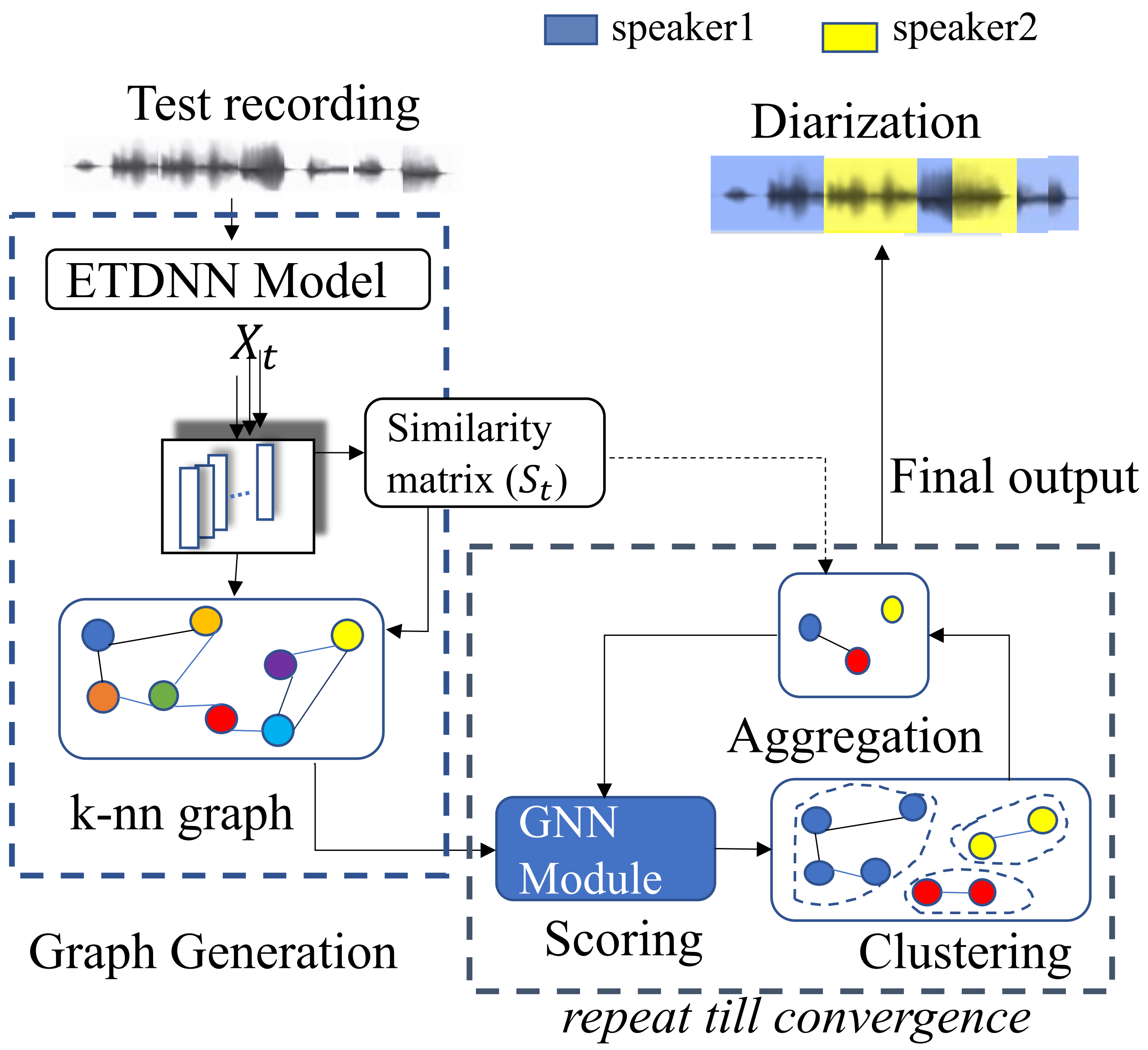
- Conventional X-vector clustering is a multi-step approach where each component is optimized independently.
- The end goal is to minimize the clustering errors to improve performance so why not to train a model with same objective.

Proposed Approach

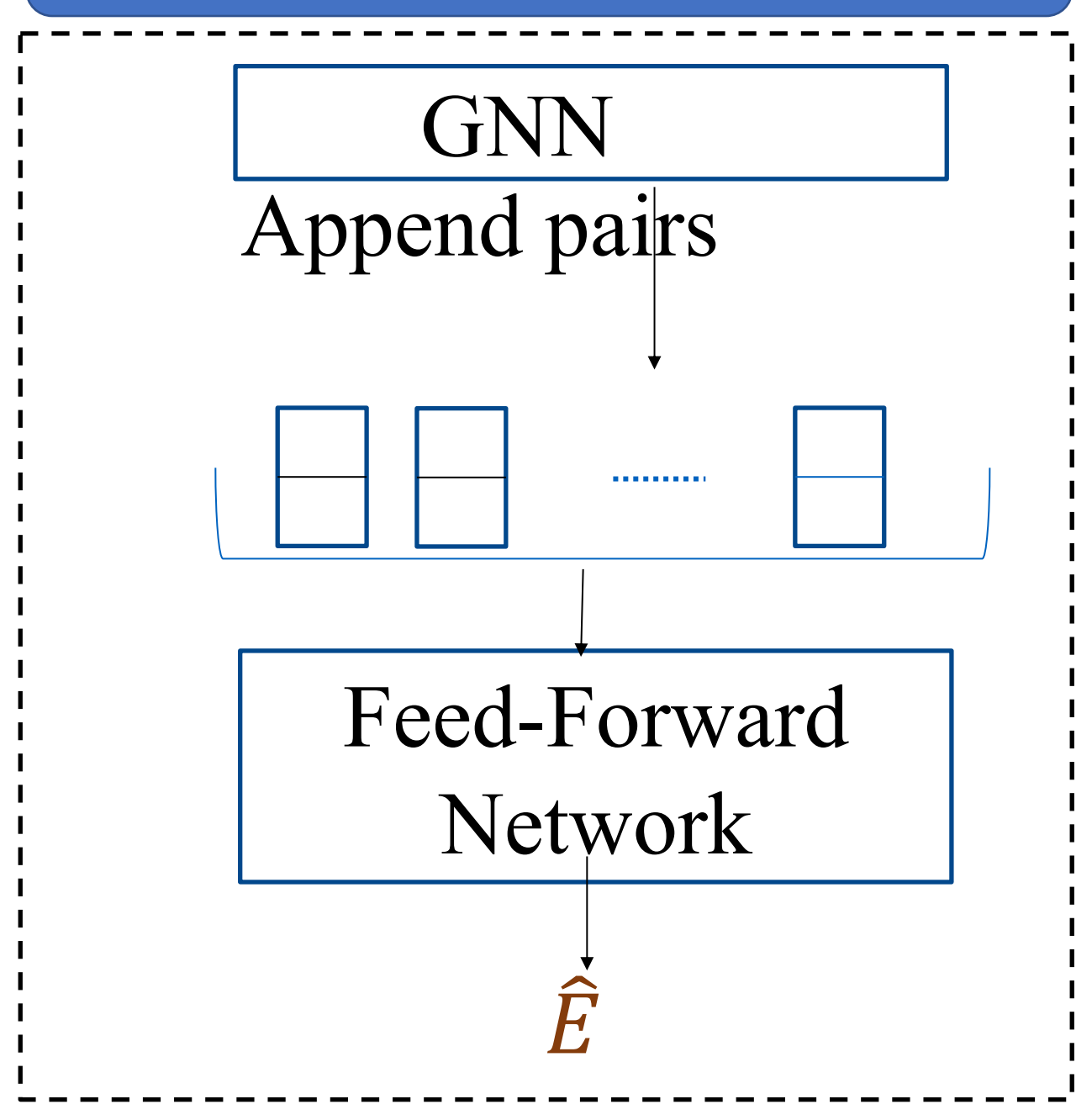
Supervised HierArchical gRaph Clustering algorithm (SHARC)

- 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.

SHARC Inference



GNN Module



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

Clustering

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

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

$$\varepsilon(i) = \{j | (v_i, v_j) \in E_m, \hat{d}_i \le \hat{d}_j \;\; p_{ij} \ge p_\tau\}$$

SHARC Training

Loss: $L = L_{conn} + L_{den}$

$L_{conn} = BCE\left(p_{ij}, q_{ij}\right) ; L_{den} = MSE\ loss\left(d_i, \hat{d}_i\right)$

q_{ij}- Ground truth edge labels,

p_{ij} - predicted edge labels
$\forall i, j \in V$

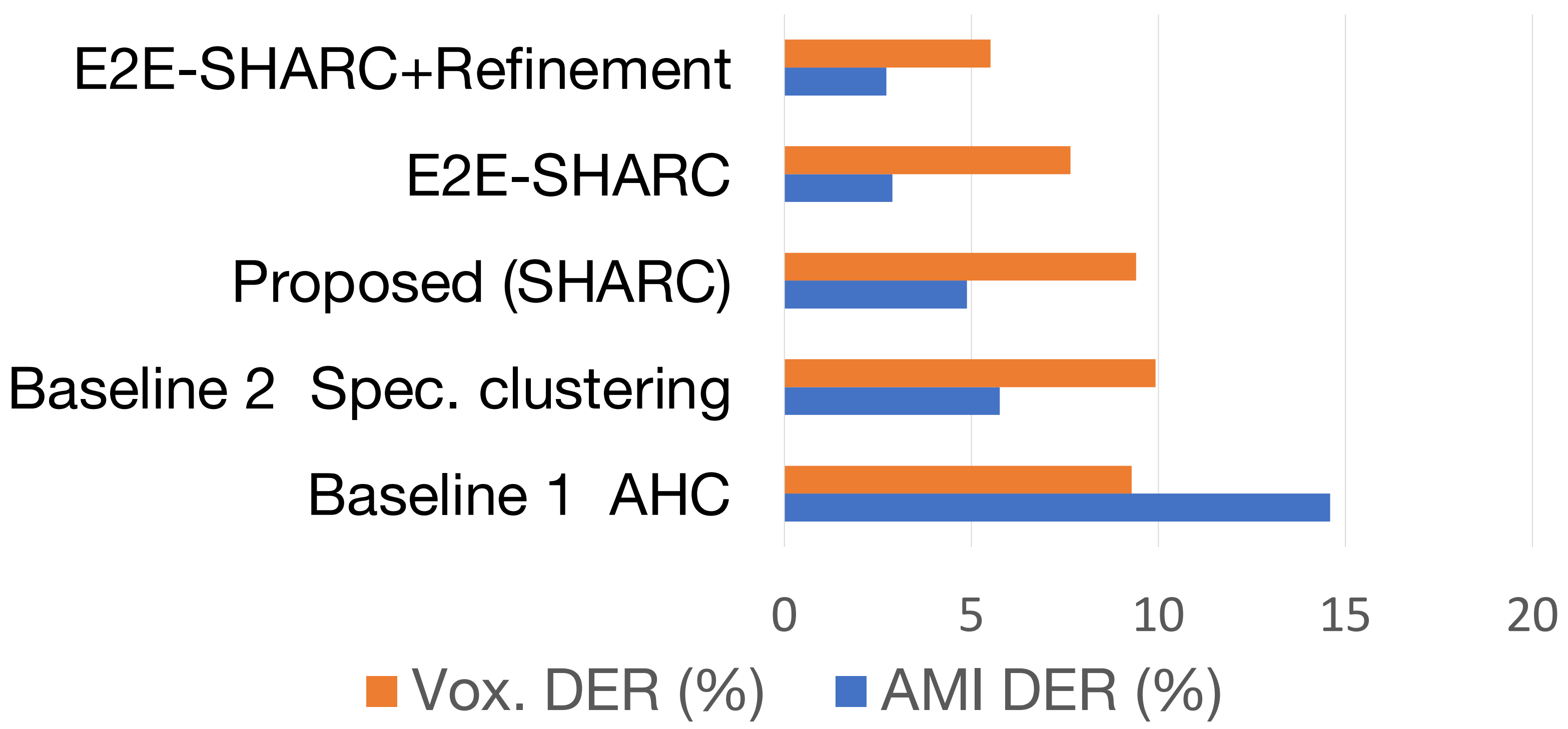
Database

- **AMI:** Meeting dataset containing 20-60 mins audio files with 3-5 speakers, dev set: 18 and eval set:16.
- **Voxconverse**: Conversations extracted from YouTube videos containing 22s - 20mins audio files with 1-21 speakers. dev: 216 and eval: 232.

Results

Diarization Error Rate (DER): Miss + False alarm + Speaker confusion errors
Hyper parameters: k-nn: 60 (AMI), 30 (Vox.)
p_τ**:** 0 (AMI), 0.8 (Vox.)

DER Performance comparision



AMI MDM	Eval DER (%)
ECAPA-TDNN	3.01
SHARC + VBx	2.11
Voxconverse	**Eval DER (%)**
Wang et. al.	5.82
E2E SHARC + VBx	5.51

Conclusions

- Introduced supervised hierarchical clustering for speaker diarization.
- Designed an end-to-end approach to perform speaker diarization using Graph Neural Networks.
- 53% and 41% rel. improvement on AMI and Voxconverse datasets

References: Yifan Xing et. al, “Learning hierarchical graph neural networks for image clustering,” in Proc. IEEE ICCV, 2021,
Hossein Zeinali et. al., “BUT system description to vox- celeb speaker recognition challenge 2019,” *arXiv preprint arXiv:1910.12592*, 2019.