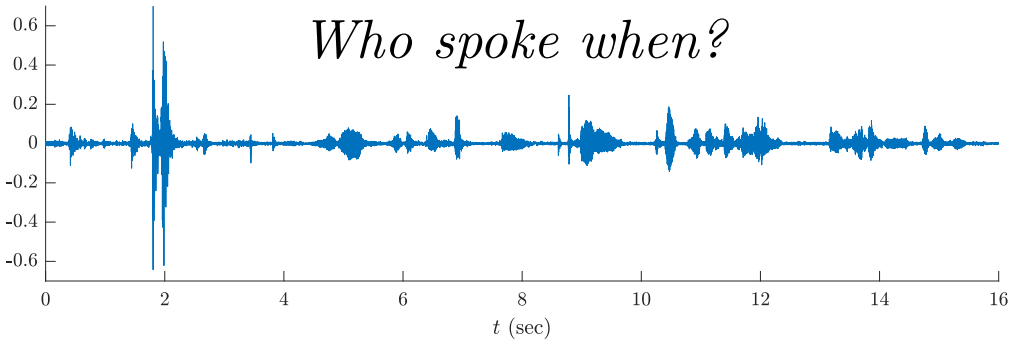
Multimodal Clustering with Role Induced Constraints for Speaker Diarization

Nikolaos Flemotomos, Shrikanth Narayanan

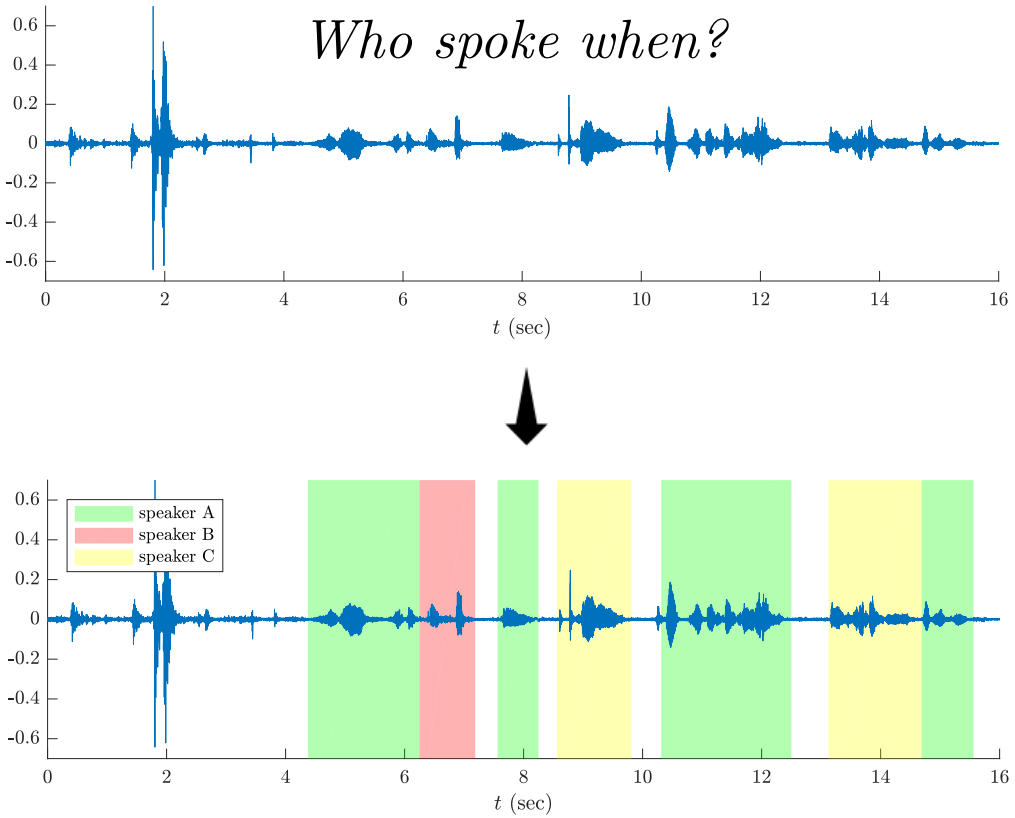
University of Southern California
Department of Electrical and Computer Engineering
Signal Analysis and Interpretation Laboratory

Interspeech 2022

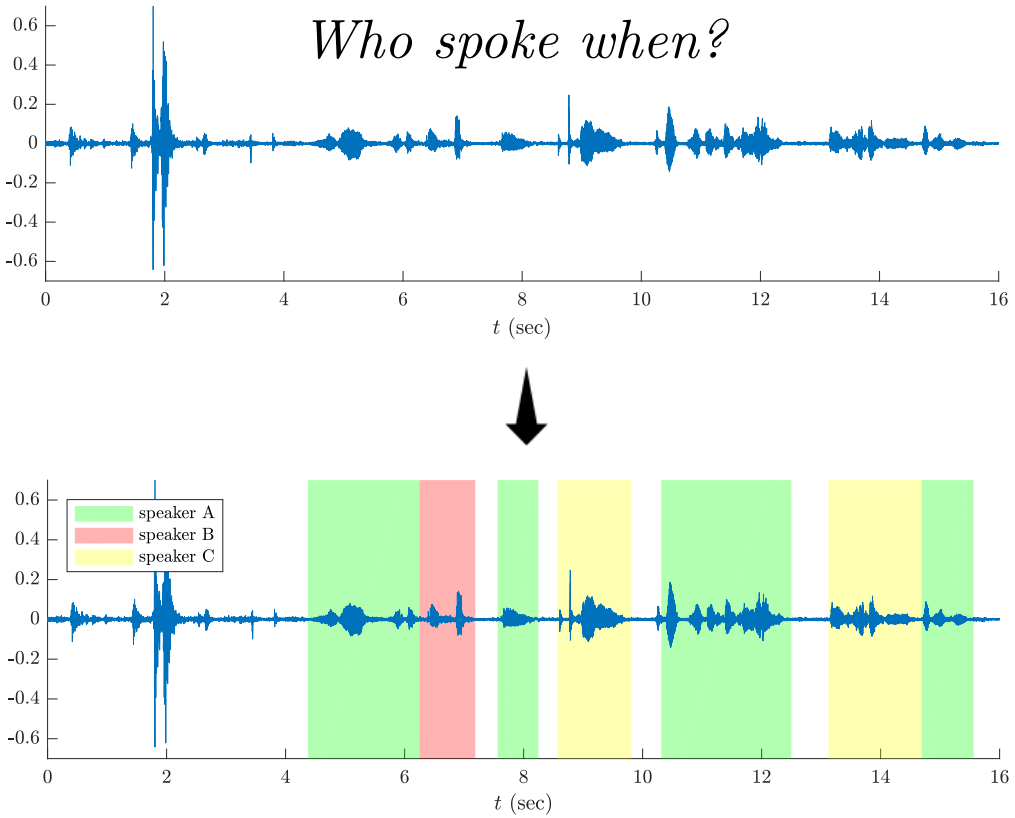
Speaker Diarization



Speaker Diarization



Speaker Diarization



Traditional approach

1. segmentation
2. clustering

Structured Scenario: speakers assume *roles*

Example scenarios:

- business meetings
- doctor-patient interactions
- broadcast news programs
- call centers
- lectures
- interviews
- ...

images from the Noun Project
creators: Nubaia Karim Barsha, Gan Khun Lay, Arafat Uddin, Llisole, ProSymbols

Structured Scenario: speakers assume *roles*

Example scenarios:

- business meetings
- doctor-patient interactions
- broadcast news programs
- call centers
- lectures
- interviews
- ...

different *roles* ⇒ distinguishable linguistic patterns

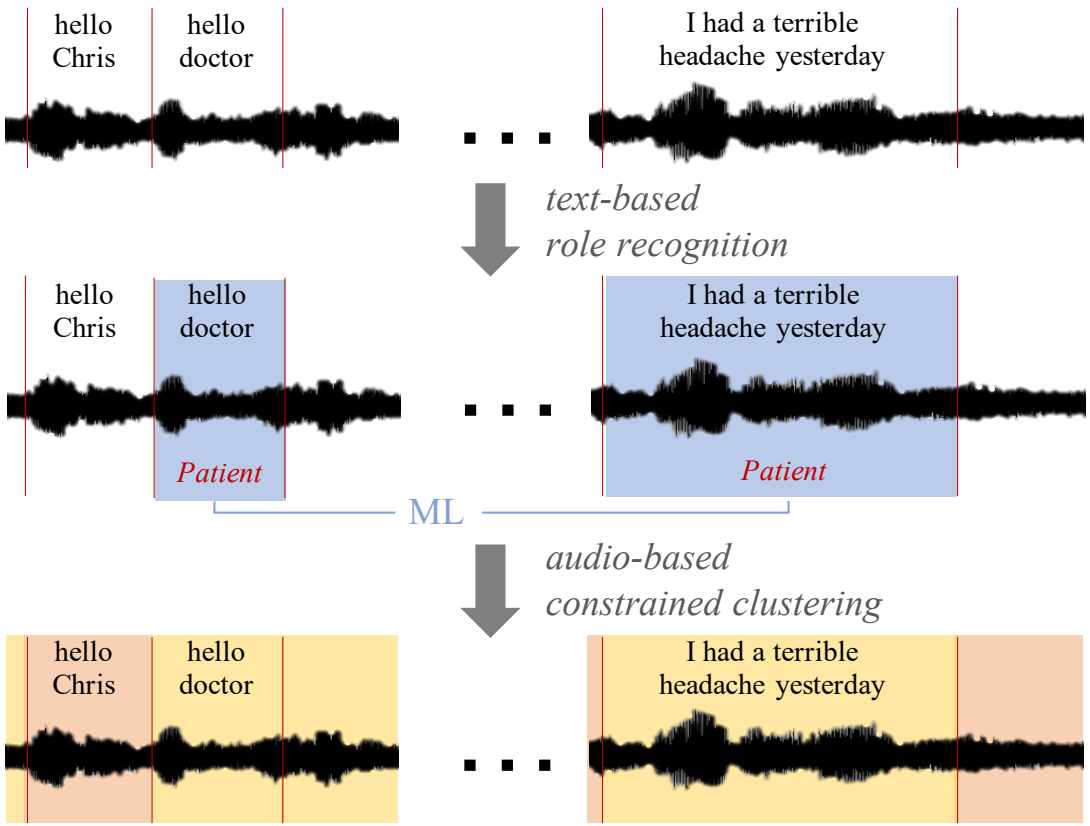
⇒ Can we use language to assist diarization?

Use Roles to Impose Constraints

- extract role information to impose constraints during audio-based clustering
- focus on segment-level pairwise constraints:
  Must-Link (ML) and Cannot-Link (CL)

Use Roles to Impose Constraints

- extract role information to impose constraints during audio-based clustering
- focus on segment-level pairwise constraints:
  Must-Link (ML) and Cannot-Link (CL)

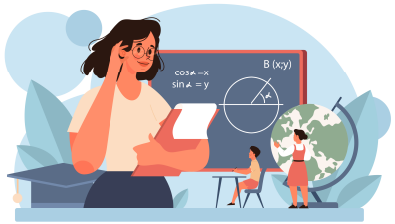


Use Roles to Impose Constraints

- extract role information to impose constraints during audio-based clustering
- focus on segment-level pairwise constraints:
  Must-Link (ML) and Cannot-Link (CL)

Some possible scenarios and strategies:

- different roles are played by different speakers
  e.g., teacher vs. students



$\Rightarrow$ CL constraints between segments with different roles

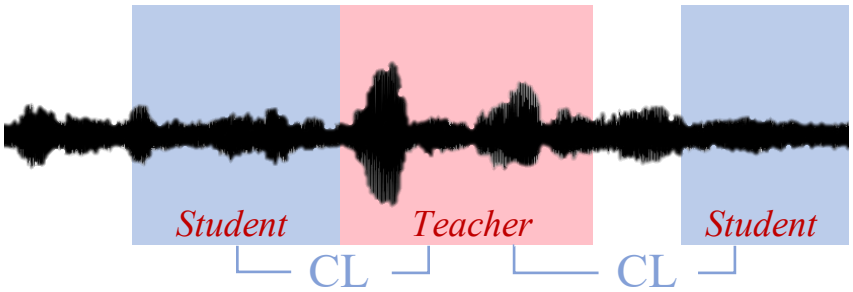


image from freepik.com by vector4stock

Use Roles to Impose Constraints

- extract role information to impose constraints during audio-based clustering
- focus on segment-level pairwise constraints:
  Must-Link (ML) and Cannot-Link (CL)

Some possible scenarios and strategies:

- different speakers play different roles
  e.g., host vs. interviewer vs. guest

$\Rightarrow$ ML constraints between segments with same roles

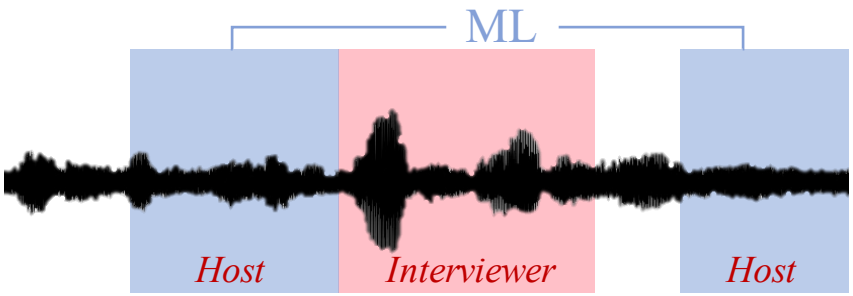


image from freepik.com by pch.vector

Use Roles to Impose Constraints

- extract role information to impose constraints during audio-based clustering
- focus on segment-level pairwise constraints:
  Must-Link (ML) and Cannot-Link (CL)

Some possible scenarios and strategies:

- every speaker mapped to a distinct role
  e.g., one doctor vs. one patient

  $\Rightarrow$ both ML and CL constraints

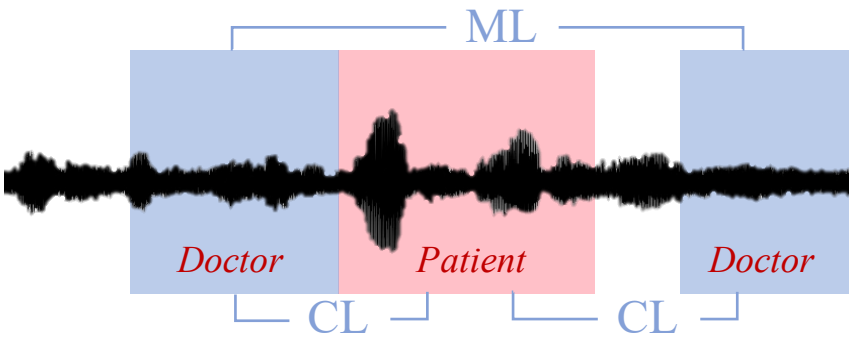


image from freepik.com by pch.vector

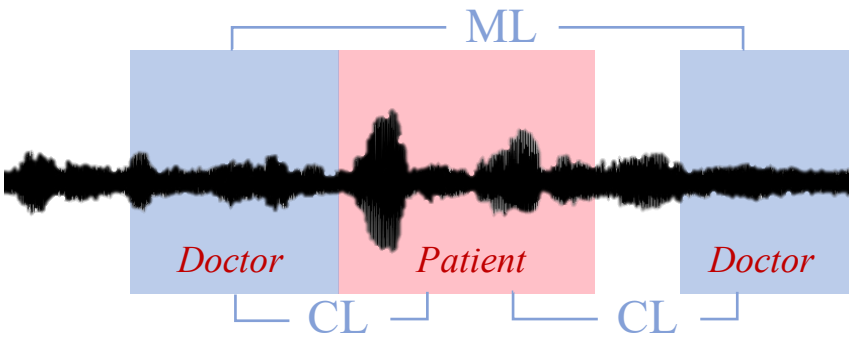
Use Roles to Impose Constraints

- extract role information to impose constraints during audio-based clustering
- focus on segment-level pairwise constraints:
  Must-Link (ML) and Cannot-Link (CL)

Some possible scenarios and strategies:

- every speaker mapped to a distinct role
  e.g., one doctor vs. one patient

  $\Rightarrow$ both ML and CL constraints



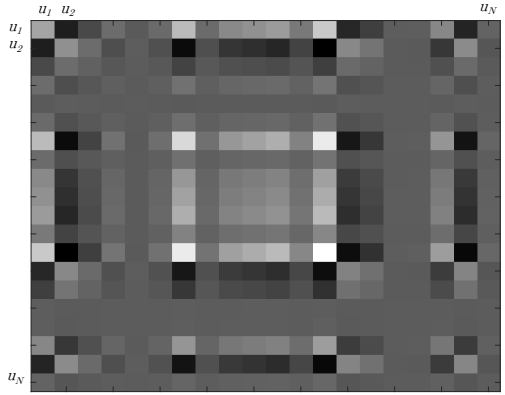
- adopt framework of constrained spectral clustering

image from freepik.com by pch.vector

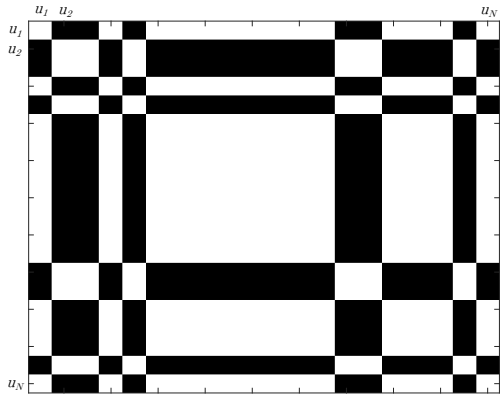
Spectral Clustering

0. speaker-homogeneous segments

1. cosine-based affinity matrix $\hat{\mathbf{W}}$



2. thresholding & symmetrization ($\mathbf{W}$)



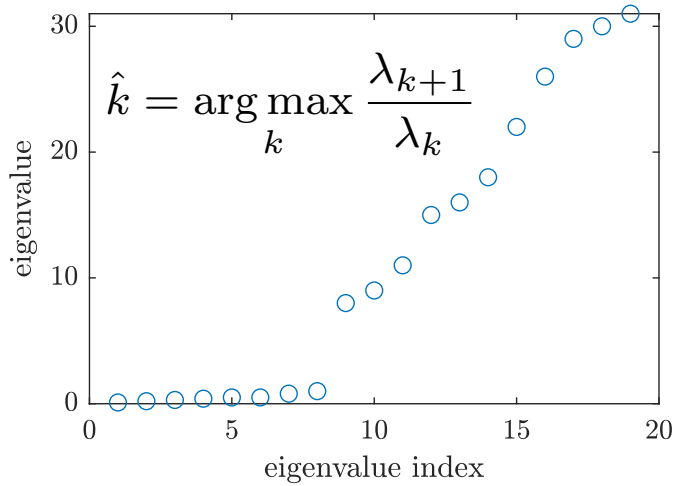
3. normalized Laplacian

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$$

$$\mathbf{D} = \text{diag}\{d_1, d_2, \cdots, d_N\}$$

$$d_i = \sum_j \mathbf{W}_{ij}$$

4. maximum eigen-gap on $\mathbf{L}$



5. $\hat{k}$-means on eigenvectors of $\mathbf{L}$

$$\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|\cdots|\mathbf{x}_{\hat{k}}]$$

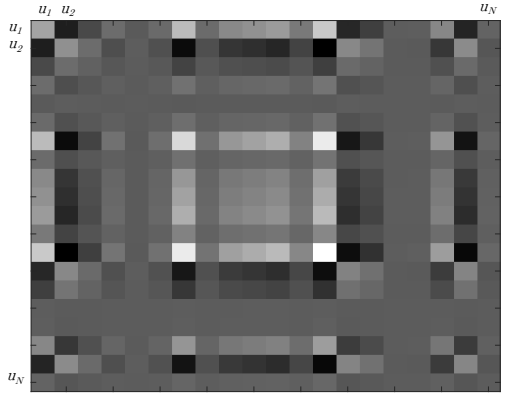
corresponding to the $\hat{k}$ smallest eigenvalues

*Eigenvalues are only given for visualization purposes; they do not correspond to $\mathbf{W}$.

Constrained Spectral Clustering

0. speaker-homogeneous segments

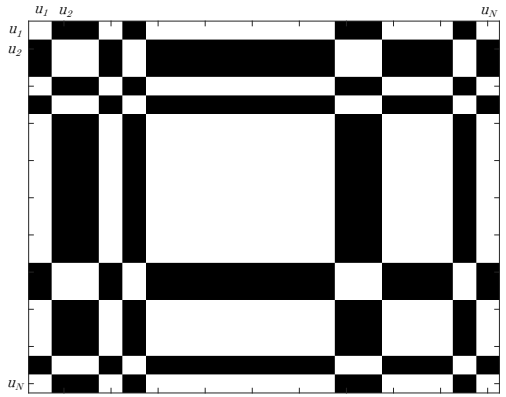
1. cosine-based affinity matrix $\hat{\mathbf{W}}$



Constrained Clustering

- increase similarity between ML-constrained pairs
- decrease similarity between CL-constrained pairs

2. thresholding & symmetrization ($\mathbf{W}$)



Integrate initial set of constraints through the Exhaustive and Efficient Constraint Propagation (E^{2}CP) algorithm:

1. construct constraint matrix $\mathbf{Z}$

$$\mathbf{Z}_{ij} = \begin{cases} +1, & \text{if } \exists \text{ ML constraint between } i \text{ and } j \\ -1, & \text{if } \exists \text{ CL constraint between } i \text{ and } j \\ 0, & \text{if } \nexists \text{ any constraint between } i \text{ and } j \end{cases}$$

2. propagate constraints to the entire session

$$\mathbf{Z}^* = (1-\alpha)^2(\mathbf{I}-\alpha\bar{\mathbf{L}})^{-1}\mathbf{Z}(\mathbf{I}-\alpha\bar{\mathbf{L}})^{-1}, \quad \bar{\mathbf{L}} = \bar{\mathbf{D}}^{-1/2}\hat{\mathbf{W}}\bar{\mathbf{D}}^{-1/2}, \quad \alpha \in [0,1]$$

α: how much to change the constraints vs. how much to change the affinity scores

$\alpha = 0 \Rightarrow \mathbf{Z}^* = \mathbf{Z} \Rightarrow$ only rely on the initial constraints
$\alpha = 1 \Rightarrow \mathbf{Z}^* = \mathbf{0} \Rightarrow$ ignore the constraints

3. update affinity scores

$$\hat{\mathbf{W}}_{ij} \leftarrow \begin{cases} 1-(1-\mathbf{Z}^*_{ij})(1-\hat{\mathbf{W}}_{ij}), & \text{if } \mathbf{Z}^*_{ij} \geq 0 \text{ (move closer to 1)} \\ (1+\mathbf{Z}^*_{ij})\hat{\mathbf{W}}_{ij}, & \text{if } \mathbf{Z}^*_{ij} < 0 \text{ (move closer to 0)} \end{cases}$$

Z. Lu & Y. Peng, "Exhaustive and efficient constraint propagation: A graph-based learning approach and its applications". International Journal of Computer Vision (2013)

Datasets

University Counseling Center (UCC) psychotherapy sessions

- dyadic conversations
- one-to-one mapping between speakers and roles
  one *therapist* vs. single *client* per session
- apply both ML and CL constraints
- total speaking time: therapist (26.7h) vs. client (46.7h)

This American Life (TAL) podcast

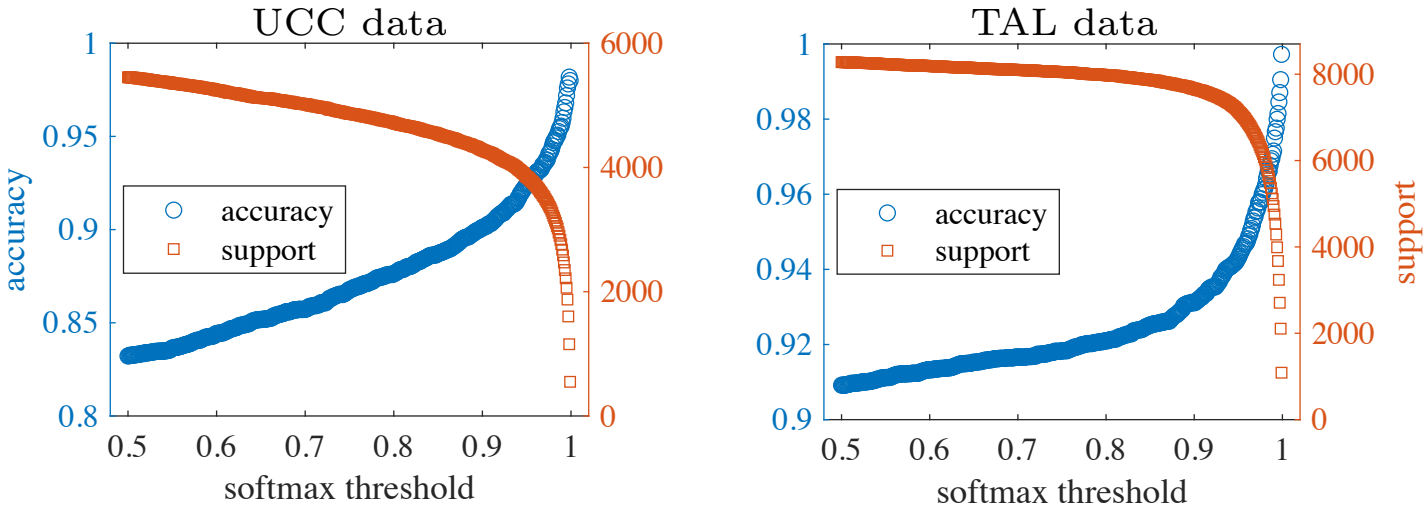
- multi-party conversations (18 speakers on average)
- partial role information
  single *host* vs. multiple *non-hosts* per episode
- apply CL constraints between segments with different roles
- total speaking time: host (118.6h) vs. non-host (519.2h)

images from freepik.com by pch.vector

Extracting Role Information

- Adapt a BERT model to classify the speaker roles

- But results are not perfect! What if we impose wrong constraints?
  - need a confidence proxy / threshold ⇒ use softmax values
  - trade-off decision: very confident or a lot of constraints??



Accuracy and support for the BERT-based classifier when only segments with softmax value above some threshold are taken into account.

- For experiments: constrain about 40% of the available segments

Diarization Results

audio-only ← | cross-modal → | language-only →

	unconstrained clustering	constrained clustering	role-based classification
UCC	1.38	**1.31**	10.34
TAL	42.22	**23.86**	63.01

Diarization Error Rate (%)—lower is better.

- experiments with manual segmentation and manual transcription
  - only evaluate clustering performance

- slight improvement for the dyadic UCC dataset
- substantial improvement for the multi-party TAL dataset
  - constraints helped estimate number of speakers (clusters) per episode

Summary

- Proposed a cross-modal framework to impose language-based role constraints during audio-based clustering.
- Improved diarization results for both dyadic and multi-party role-playing interactions.

Summary

- Proposed a cross-modal framework to impose language-based role constraints during audio-based clustering.
- Improved diarization results for both dyadic and multi-party role-playing interactions.
- What about other modalities?
  - audio- or video-based constraints
- Can we incorporate soft constraints?
  - confidence scores
  - role-based conversational dynamics