

Multimodal Clustering with Role Induced Constraints for Speaker Diarization

Nikolaos Flemotomos[†], Shrikanth Narayanan

Signal Analysis and Interpretation Lab (SAIL), University of Southern California

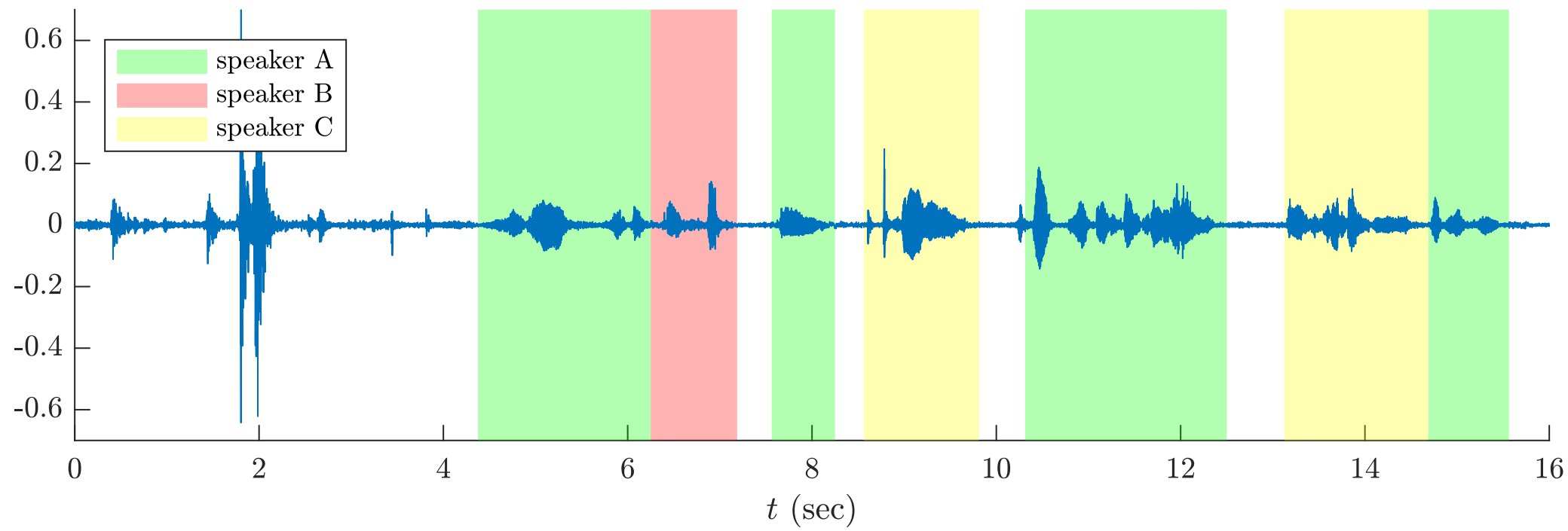
[†]currently with Apple



USC University of Southern California

Speaker Diarization & Speaker Roles

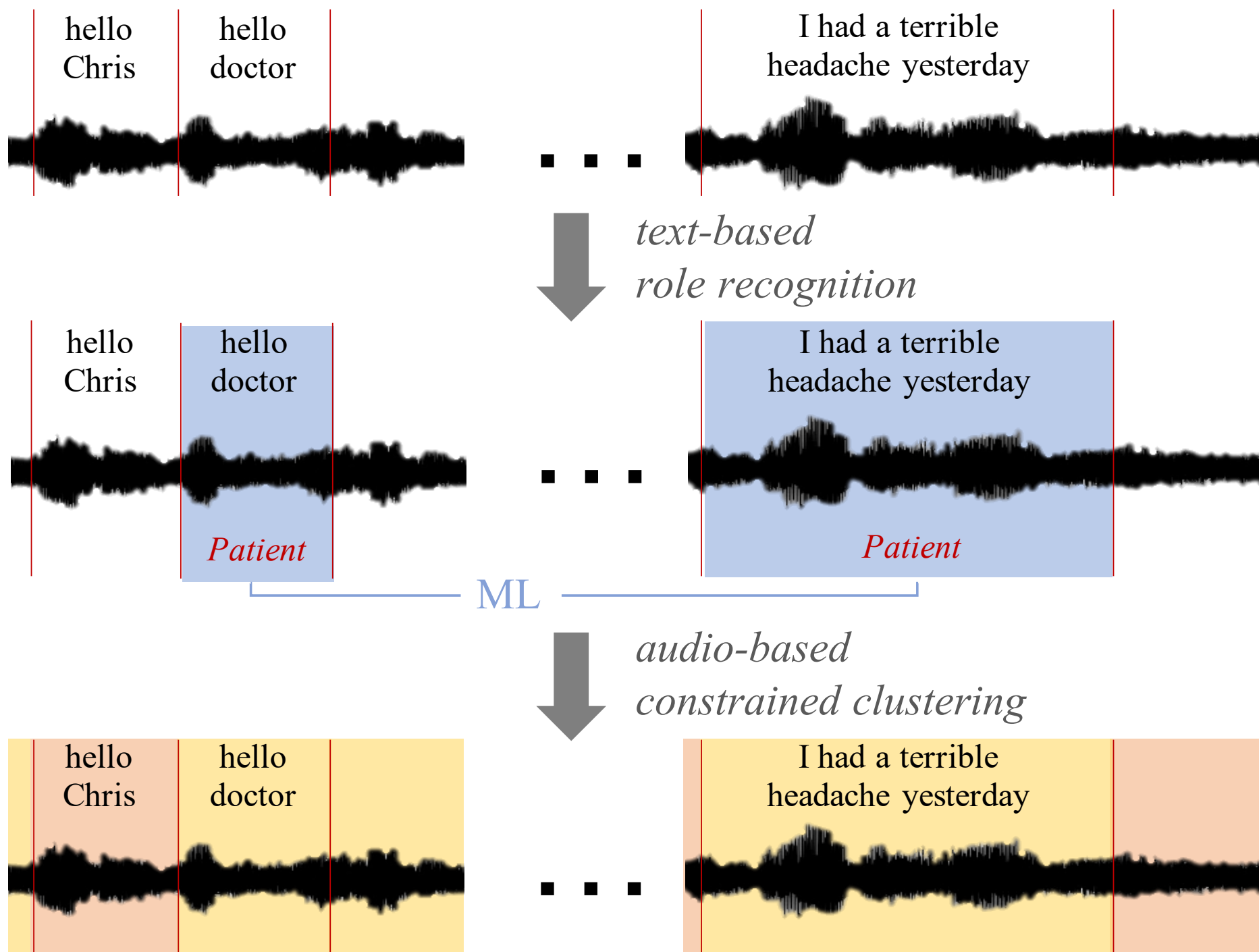
- ▶ diarization answers the question “who spoke when?”
- ▶ conventional approach:
 - ▶ speaker segmentation: find speaker change points
 - ▶ *speaker clustering*: cluster speaker-homogeneous segments



- ▶ focus on scenarios where speakers assume *roles*
 - ▶ examples: interviews, lectures, TV shows, etc.
- ▶ roles are associated with distinguishable linguistic patterns
- ▶ can we use role-specific language to assist diarization?

Role-Induced Constrained Clustering

- ▶ extract *language-based* role information to impose constraints during *audio-based* clustering
- ▶ focus on segment-level pairwise constraints
 - ▶ must-link (ML): 2 segments *should* be in the same cluster
 - ▶ cannot-link (CL): 2 segments *should not* be in the same cluster



- ▶ possible scenarios
 - ▶ different roles played by different speakers (e.g., teacher vs. students)
 - ⇒ CL constraints between segments with different roles
 - ▶ different speakers play different roles (e.g., host vs. interviewer vs. host)
 - ⇒ ML constraints between segments with same roles
 - ▶ every speaker mapped to a distinct role (e.g., doctor vs. patient)
 - ⇒ both ML and CL constraints

Constrained Spectral Clustering

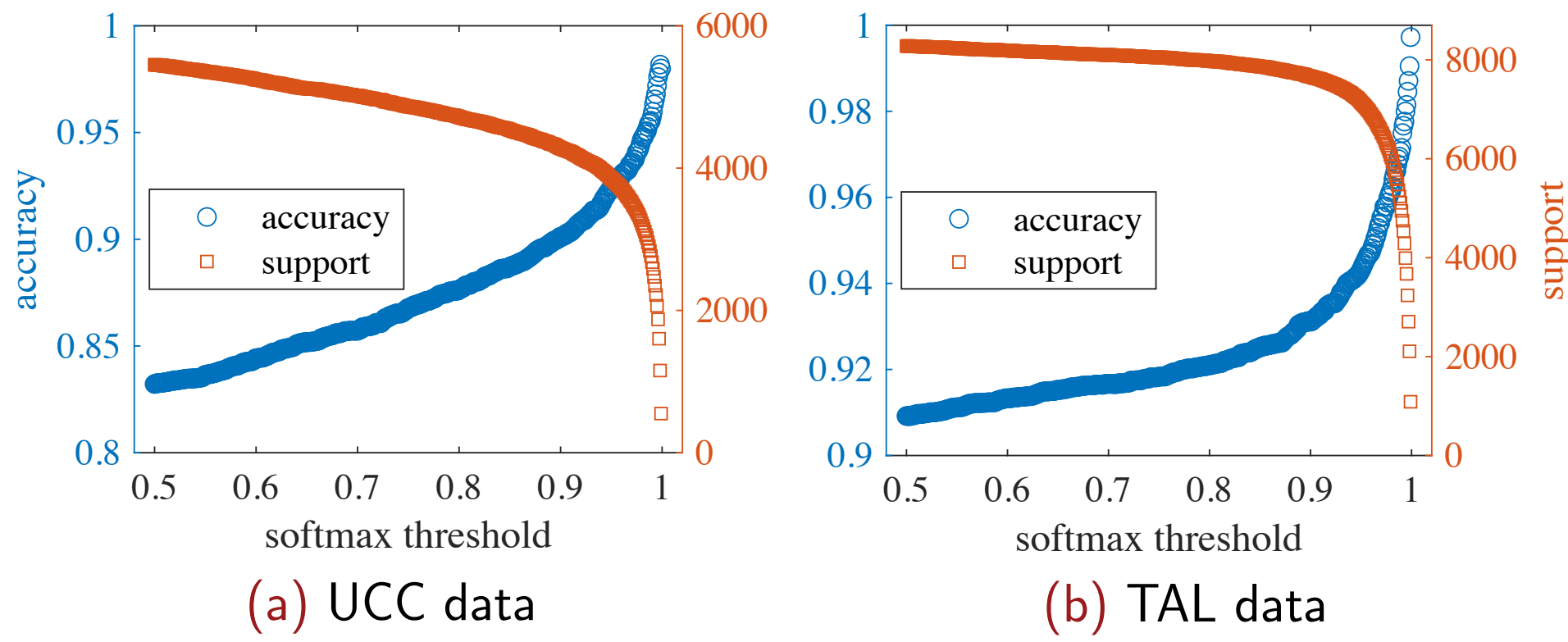
- ▶ construct pairwise similarity matrix \mathbf{W}
- ▶ construct role-based constraint matrix \mathbf{Z} for a high-confidence subset of segments
$$\mathbf{Z}_{ij} = \begin{cases} +1, & \text{if } \exists \text{ ML constraint between } i \text{ and } j \\ -1, & \text{if } \exists \text{ CL constraint between } i \text{ and } j \\ 0, & \text{if } \nexists \text{ any constraint between } i \text{ and } j \end{cases}$$
- ▶ propagate constraints via *Exhaustive and Efficient Constraint Propagation* (E²CP) algorithm [1] and update \mathbf{W}
- ▶ apply spectral clustering

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)

Extracting Role Information

- ▶ adapt a BERT model to classify the speaker roles
- ▶ make sure we don't impose wrong constraints
 - ▶ need for confidence proxy ⇒ use softmax values of classifier
 - ▶ 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

Experiments & Results

- ▶ use oracle segmentation + oracle transcriptions ⇒ only evaluate clustering performance
- ▶ speaker representation: x-vectors
- ▶ apply initial ML/CL constraints on ~ 40% of the segments and integrate constraints via E²CP

diarization error rate (%) – lower is better

	unconstrained clustering (audio-only)	constrained clustering (multimodal)	role-based classification (language-only)
UCC	1.38	1.31	10.34
TAL	42.22	23.86	63.01*

*results contain 2 speakers (due to the binary classification)

Conclusion

- ▶ improved diarization results for both dyadic and multi-party role-playing interactions
 - ▶ improved estimation of the number of speakers in the multi-party scenario
- ▶ future work
 - ▶ focused on language-based constraints – what about other modalities?
 - ▶ can we incorporate soft constraints?

References

- [1] Z. Lu, Y. Peng, “Exhaustive and efficient constraint propagation: A graph-based learning approach and its applications”. Int J Comput Vis (2013)
- [2] A. Tripathi, et. al., “Turn-to-diarize: Online speaker diarization constrained by transformer transducer speaker turn detection”. ICASSP (2022)
- [3] N. Flemotomos, P. Georgiou, S. Narayanan, “Linguistically aided speaker diarization using speaker role information”. Odyssey (2020)