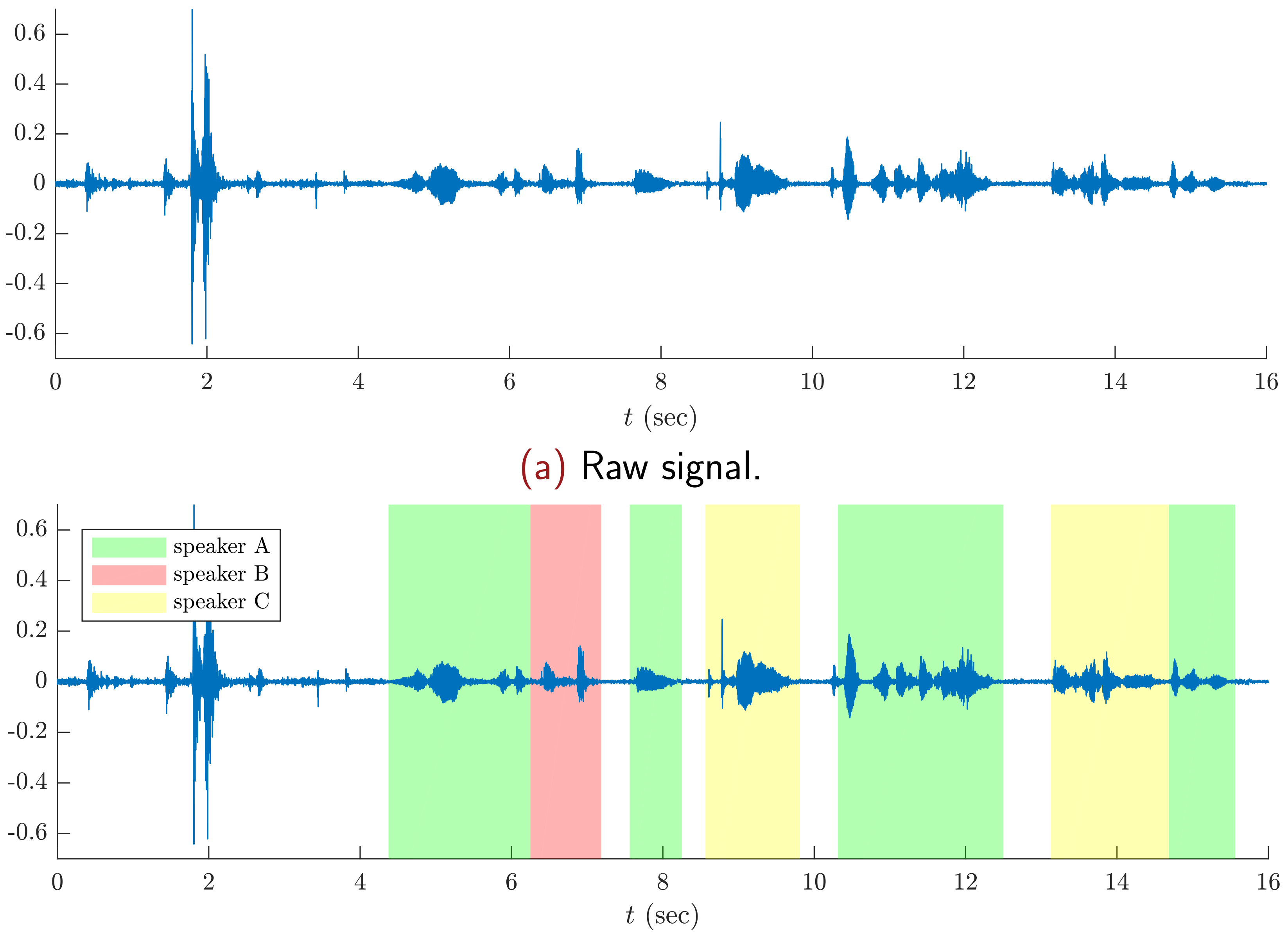
Multimodal Clustering with Role Induced Constraints for Speaker Diarization

Nikolaos Flemotomos, Shrikanth Narayanan

Signal Analysis and Interpretation Lab (SAIL), 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



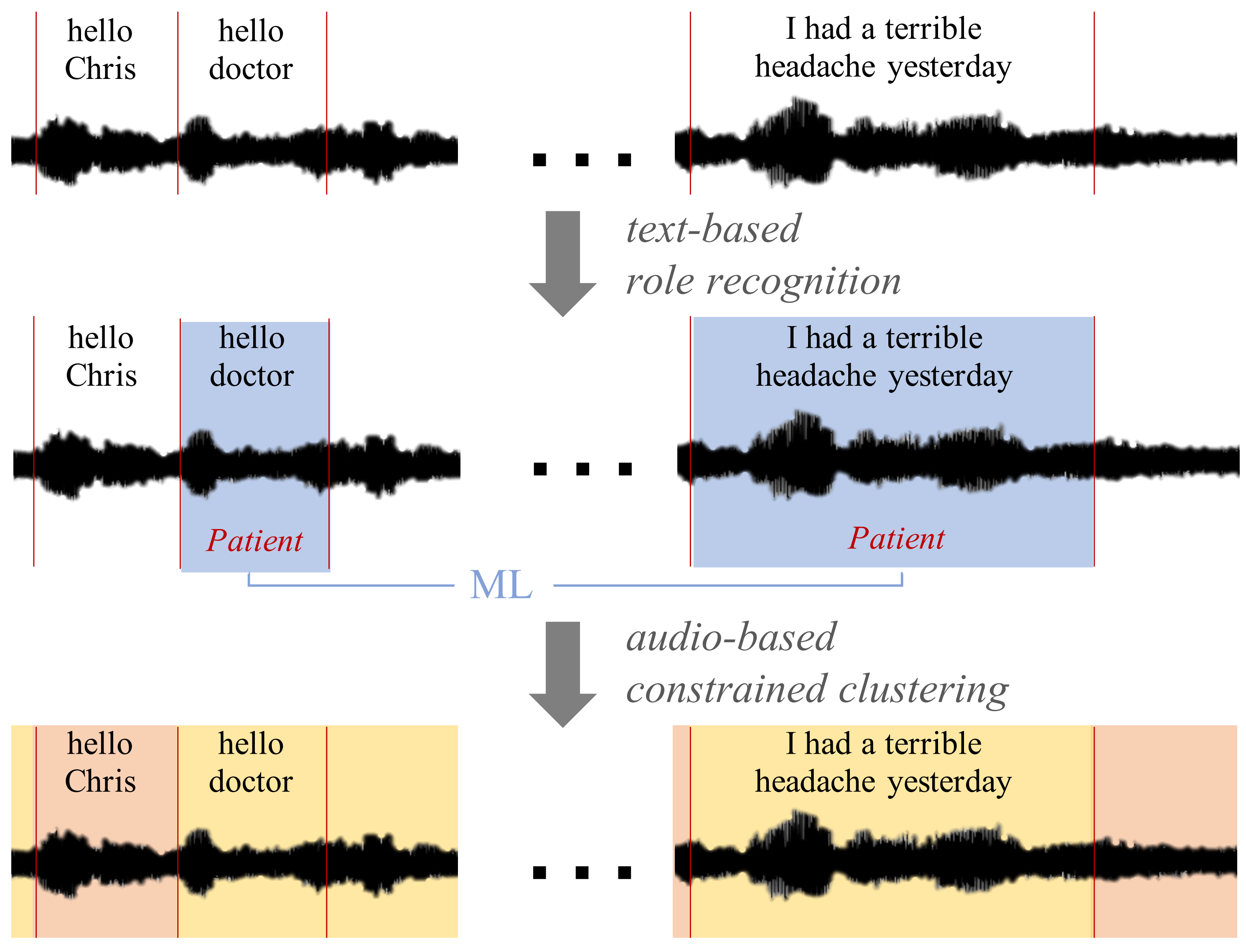
(a) Raw signal.

(b) Diarization output.

- focus on scenarios where speakers assume *roles*
  - examples: interviews, lectures, TV shows, etc.
- roles are associated with distinguishable linguistic patterns
  - interviewer uses more interrogative words
  - teacher speaks in a more didactic style
- 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


hello Chris
hello doctor
I had a terrible headache yesterday
. . .
text-based role recognition
hello Chris
hello doctor
Patient
I had a terrible headache yesterday
Patient
ML
audio-based constrained clustering
hello Chris
hello doctor
I had a terrible headache yesterday

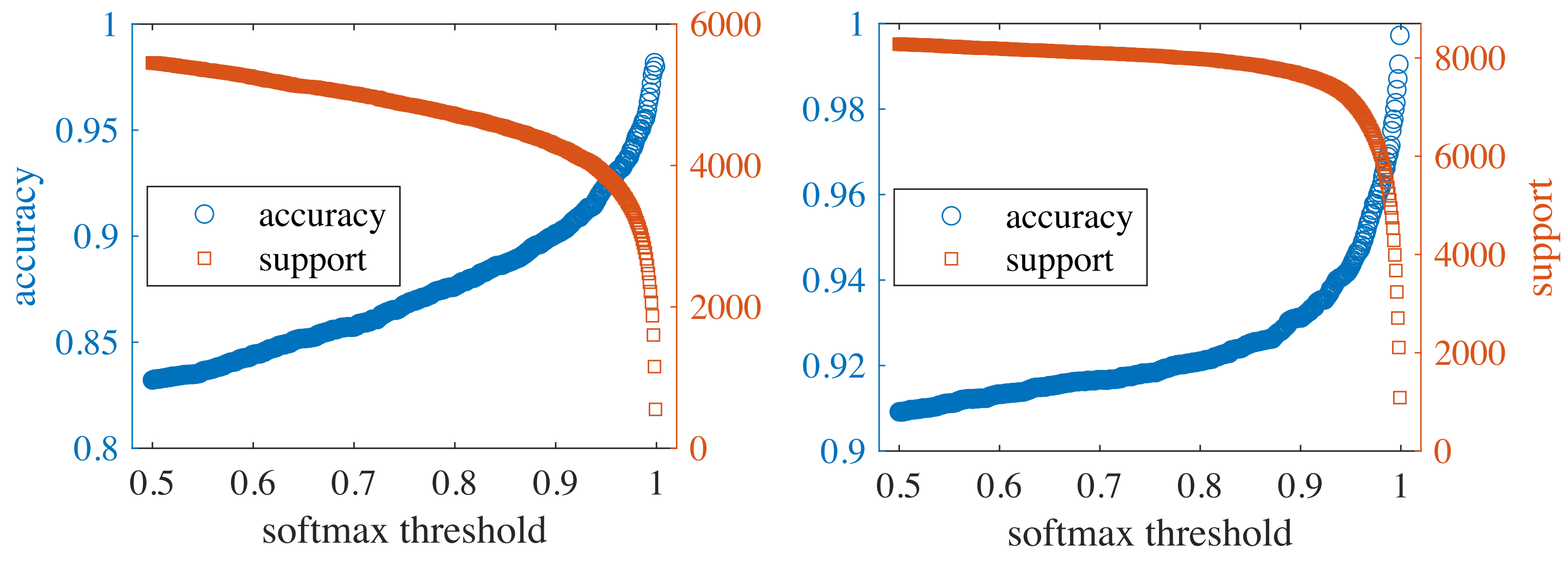
- possible scenarios
  - different roles played by different speakers (e.g., teacher vs. students) ⇒ CL constraints between segments with different roles
  - different speakers play different roles (e.g., host vs. interviewer vs. host) ⇒ ML constraints between segments with same roles
  - every speaker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

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?



(a) UCC data (b) TAL data

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
- apply initial ML/CL constraints on ~ 40% of the segments
- propagate constraints via *Exhaustive and Efficient Constraint Propagation* (E^{2}CP) algorithm[a]
- apply spectral clustering

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)

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

Conclusion

- cross-modal framework: impose language-based role constraints during audio-based clustering
- 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?

flemotom@usc.edu

September 22, 2022