Who Spoke When?

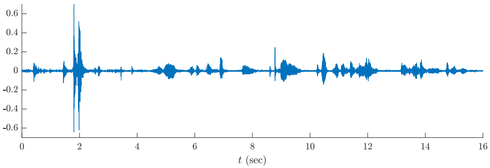
and how speaker roles can help us find the answer

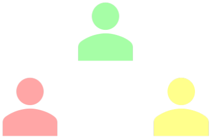
Nikolaos (Nikos) Flemotomos

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

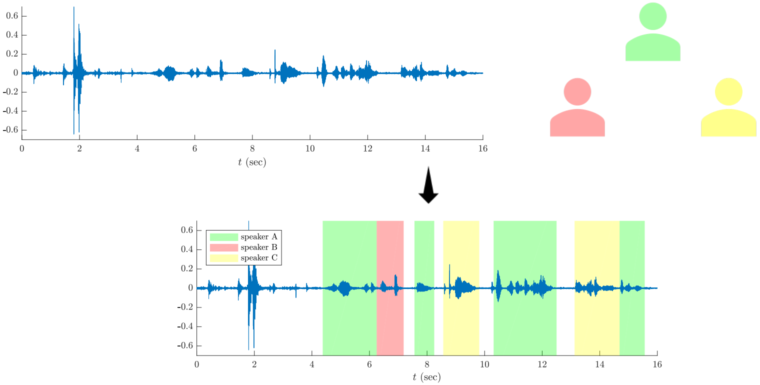
April 2, 2021

Who Spoke When: Continuous Speaker Identification

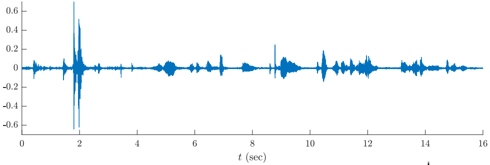


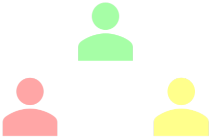


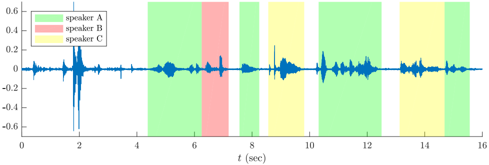
Who Spoke When: Continuous Speaker Identification



Who Spoke When: Continuous Speaker Identification



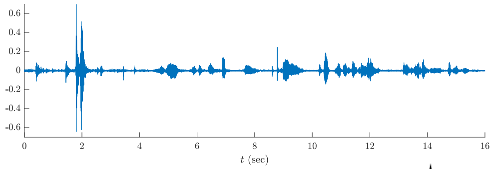


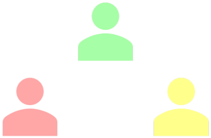


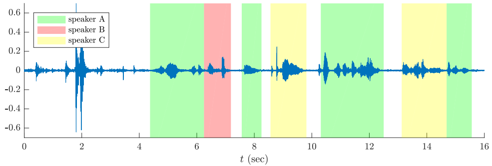
Why?

- rich transcription
- outlier detection
- speaker adaptation (ASR)
- speaker tracking

Who Spoke When: Continuous Speaker Identification



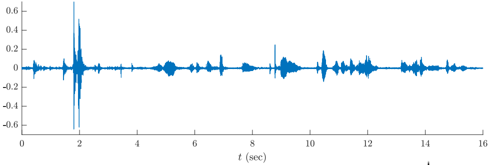


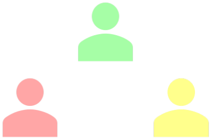


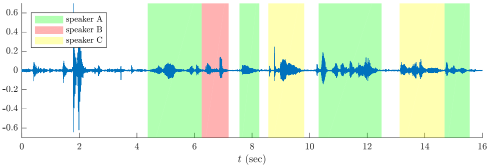
Traditional approach

1. segmentation
2. classification

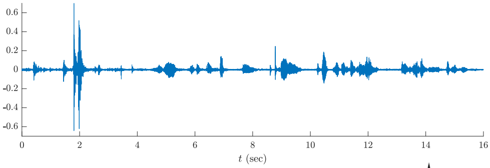
Who Spoke When: Speaker Diarization

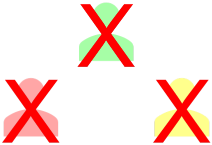


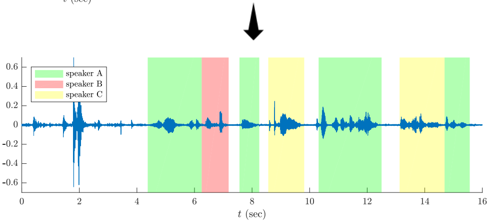




Who Spoke When: Speaker Diarization

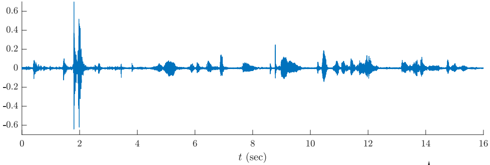


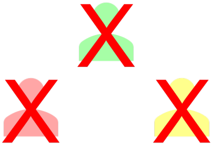


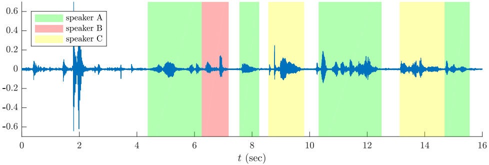


No a priori information about the speakers is given!

Who Spoke When: Speaker Diarization



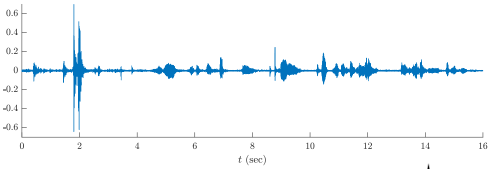


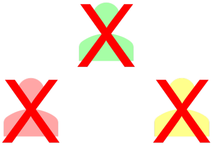


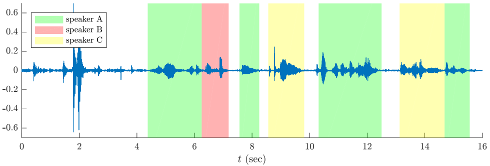
Traditional approach

1. segmentation
2. clustering

Who Spoke When: Speaker Diarization







Traditional approach

1. segmentation
2. clustering → What if...
   - very similar acoustic characteristics?
   - too much noise and/or silence?

Structured Scenario: speakers assume *roles*

- Common applications:
    - business meetings
    - doctor-patient interactions
    - broadcast news programs
    - lectures
    - interviews
    - ...

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

Structured Scenario: speakers assume *roles*

- Common applications:
  - business meetings
  - doctor-patient interactions
  - broadcast news programs
  - lectures
  - interviews
  - ...

- different *roles* $\Rightarrow$ distinguishable linguistic patterns

  $\Rightarrow$ Can we use language to assist diarization?

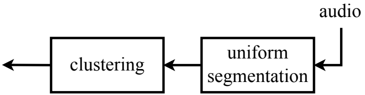
images from the Noun Project. creators: Nubaia Karim Barsha, Gan Khun Lay, Arafat Uddin, Llisole, ProSymbols

Proposed System

- different *roles* $\Rightarrow$ distinguishable linguistic patterns
  $\Rightarrow$ Can we use language to assist diarization?

Proposed System

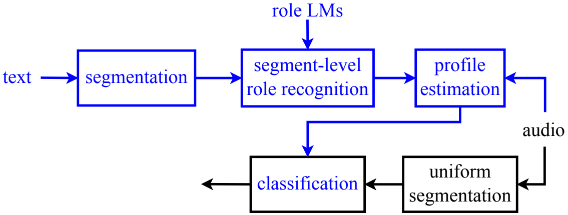
- different *roles* $\Rightarrow$ distinguishable linguistic patterns
  $\Rightarrow$ Can we use language to assist diarization?



traditional, audio-only system

Proposed System

- different *roles* $\Rightarrow$ distinguishable linguistic patterns
  $\Rightarrow$ Can we use language to assist diarization?

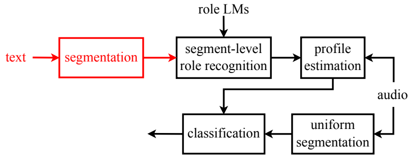


proposed, linguistically-aided system

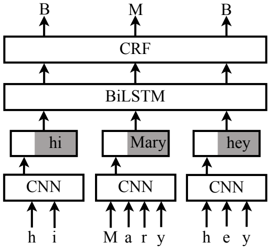
Use speaker role information to construct speaker profiles.
Turn the clustering problem into a classification one.

N. Flemotomos, P. Georgiou & S. Narayanan. “Linguistically Aided Speaker Diarization Using Speaker Role Information”, Odyssey (2020)

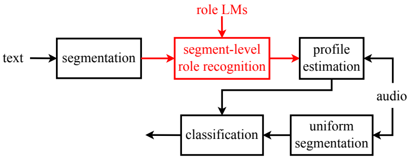
Proposed System: Text-based segmentation



- Goal: obtain speaker-homogeneous text segments
- Assumption: single speaker per sentence
  $\Rightarrow$ segment text at the sentence level
- sequence-labeling problem $\rightarrow$ CNN-BiLSTM-CRF architecture



Proposed System: Role recognition



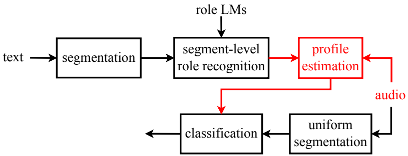
- Build a background LM $\mathcal{G}$ and N role-specific LMs $\mathcal{R}_i$ (N roles).
- Interpolate the LMs (n-gram):

$$\mathcal{R}_i^{+} = w_{g_i}\mathcal{G} \oplus w_{r_i}\mathcal{R}_i \oplus (1 - w_{g_i} - w_{r_i})\tilde{\mathcal{R}}_i$$

$$\tilde{\mathcal{R}}_i = \frac{1}{N-1}\bigoplus_{\substack{j=1\\ j\neq i}}^{N}\mathcal{R}_j$$

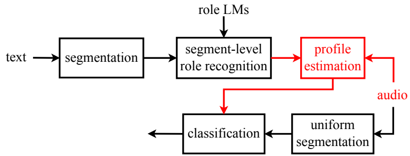
- Assign to each text segment x the role i that minimizes the perplexity $pp(x|\mathcal{R}_i^{+})$.

Proposed System: Profile Estimation



- Extract an acoustic speaker embedding (x-vector) u_x $\forall$ audio-aligned segment x assigned the role R_i.
- Define the role profile r_i as the mean of all the $u_x : x \in R_i$.

Proposed System: Profile Estimation



- Extract an acoustic speaker embedding (x-vector) u_x $\forall$ audio-aligned segment x assigned the role R_i.
- Define the role profile r_i as the mean of all the $u_x : x \in R_i$.

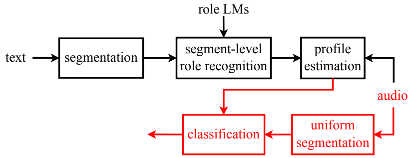
- *Are we confident about all the role assignments?*
  - Assign a confidence metric to each x:

$$c_x = \min_{j \neq i} |pp(x|\mathcal{R}_j^+) - pp(x|\mathcal{R}_i^+)|$$

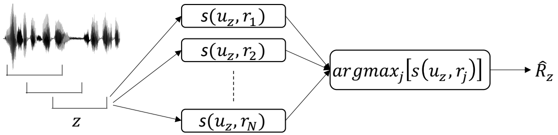
  - Take into account only the segments about which we are confident enough:

$$r_i = \frac{\sum_{x \in R_i} \mathbb{I}\{c_x > \theta\} u_x}{\sum_{x \in R_i} \mathbb{I}\{c_x > \theta\}}$$

Proposed System: Audio segmentation and classification



- Segment uniformly the speech signal (sliding window).
- Extract an acoustic speaker embedding (x-vector) u_z $\forall$ segment z
- Calculate the similarity $s(u_z, r_i)$ $\forall$ role profile r_i.
- Assign to the audio segment z the role i that maximizes $s(u_z, r_i)$.



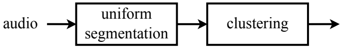
Evaluation Dataset and Baselines

- Dyadic psychotherapy interactions (Therapist vs. Patient)

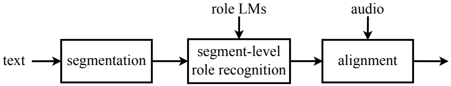
	PSYCH-train	PSYCH-dev	PSYCH-test
#sessions	74	44	25
Therapist	26.43 h	15.23 h	7.34 h
Patient	23.29 h	12.17 h	7.54 h

Table: Size of the psychotherapy dataset (PSYCH).

- Baselines



audio-only baseline



language-only baseline

Results

transcript source	text segmentation	audio only	language only	linguistically aided (all segments)	linguistically aided (best $a\%$ segments)
reference	oracle	11.05	12.99	7.28	**6.99**
	tagger		20.09	7.71	**7.30**
ASR	tagger	11.05	27.07	8.37	**7.84**

Table: DER (%) on PSYCH-test.

Results

transcript source	text segmentation	audio only	language only	linguistically aided (all segments)	linguistically aided (best $a\%$ segments)
reference	oracle	11.05	12.99	7.28	**6.99**
	tagger		20.09	7.71	**7.30**
ASR	tagger	11.05	27.07	8.37	**7.84**

Table: DER (%) on PSYCH-test.

- unimodal baselines:
  audio stream contains more valuable information

Results

transcript source	text segmentation	audio only	language only	linguistically aided (all segments)	linguistically aided (best $a\%$ segments)
reference	oracle	11.05	12.99	7.28	**6.99**
	tagger		20.09	7.71	**7.30**
ASR	tagger	11.05	27.07	8.37	**7.84**

Table: DER (%) on PSYCH-test.

- tagger oversegments
  $\Rightarrow$ short segments contain insufficient information for role recognition
  $\Rightarrow$ severe degradation for language-only system
- inaccuracies cancel out for the linguistically aided system

Results

transcript source	text segmentation	audio only	language only	linguistically aided (all segments)	linguistically aided (best $a\%$ segments)
reference	oracle	11.05	12.99	7.28	**6.99**
	tagger		20.09	7.71	**7.30**
ASR	tagger	11.05	27.07	8.37	**7.84**

Table: DER (%) on PSYCH-test.

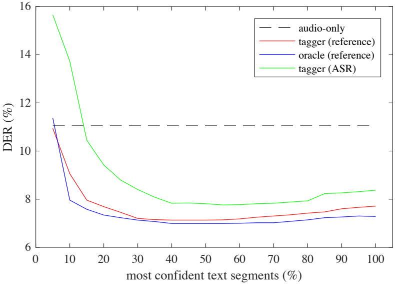
- high WER ⇒ severe degradation for language-only system
- when transcripts are only used for profile estimation (linguistically-aided) the performance gap is much smaller

Results

transcript source	text segmentation	audio only	language only	linguistically aided (all segments)	linguistically aided (best $a\%$ segments)
reference	oracle	11.05	12.99	7.28	**6.99**
	tagger		20.09	7.71	**7.30**
ASR	tagger	11.05	27.07	8.37	**7.84**

Table: DER (%) on PSYCH-test.

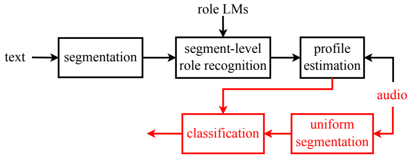
- best $a\%$ segments: use the $a\%$ of the segments we are most confident about *per session* for profile estimation
- a is optimized on dev set



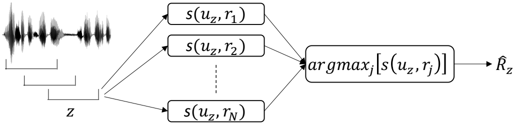
Conclusion

- Proposed a system for speaker diarization in conversational scenarios where the speakers assume specific roles.
- Used the lexical information captured within the speech signal in order to estimate the speaker profiles and follow a classification approach instead of clustering.
- Evaluated on dyadic psychotherapy interactions and demonstrated a DER relative reduction of 29.05% compared to the audio-only baseline.

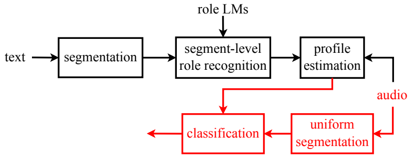
Towards Better Classification



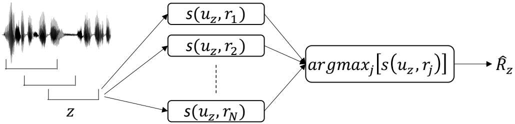
- Segment uniformly the speech signal (sliding window).
- Extract an acoustic speaker embedding (x-vector) u_z $\forall$ segment z
- Calculate the similarity $s(u_z, r_i)$ $\forall$ role profile r_i.
- Assign to the audio segment z the role i that maximizes $s(u_z, r_i)$.



Towards Better Classification



- Segment uniformly the speech signal (sliding window).
- Extract an acoustic speaker embedding (x-vector) u_z $\forall$ segment z
- Calculate the similarity $s(u_z, r_i)$ $\forall$ role profile r_i.
- Assign to the audio segment z the role i that maximizes $s(u_z, r_i)$.

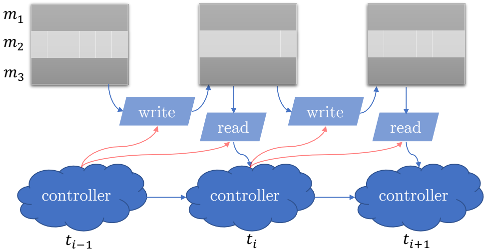


Problems

- Is the similarity metric optimal?
- Is the speaker representation appropriate for the task?
- Lack of temporal information.

Memory-Augmented Neural Networks

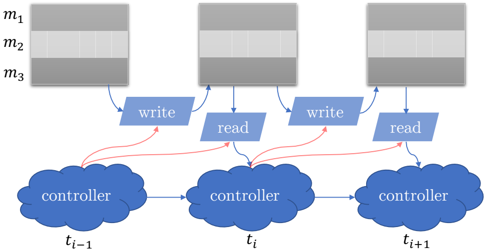
- Idea: Augment a neural architecture with a memory matrix.
- A *controller* decides how to update the memory through attention mechanisms using read and write *heads*.
- The whole system is differentiable ⇒ can learn a task-specific organization of the memory in a supervised manner through gradient descent.



Graves, Alex, *et al.* "Hybrid computing using a neural network with dynamic external memory." *Nature* (2016)
Pham, Trang, *et al.* "Relational dynamic memory networks." *arXiv preprint* (2018)
Santoro, Adam, *et al.* "Relational recurrent neural networks." *NeurIPS* (2018)

Memory-Augmented Neural Networks

- Idea: Augment a neural architecture with a memory matrix.
- A *controller* decides how to update the memory through attention mechanisms using read and write *heads*.
- The whole system is differentiable ⇒ can learn a task-specific organization of the memory in a supervised manner through gradient descent.



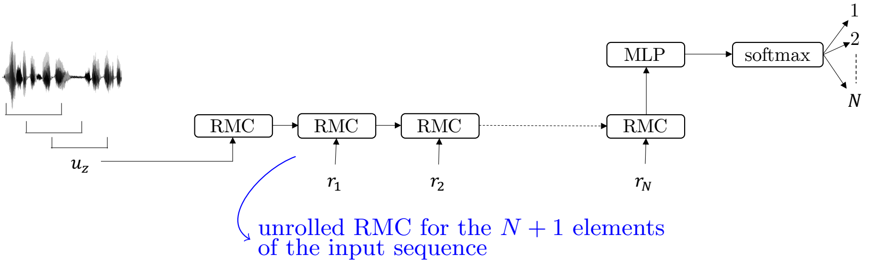
In our implementation: Relational Memory Core (RMC)

- controller [RMC] is embedded into an LSTM
- memory updates are based on a self-attention mechanism

Graves, Alex, *et al.* "Hybrid computing using a neural network with dynamic external memory." *Nature* (2016)
Pham, Trang, *et al.* "Relational dynamic memory networks." *arXiv preprint* (2018)
Santoro, Adam, *et al.* "Relational recurrent neural networks." *NeurIPS* (2018)

Proposed Architecture

- $\forall u_z$ create the sequence $\{u_z, r_1, r_2, \cdots, r_N\}$
- pass the sequence through an RMC-based network and get the label $l_z \in \{1, 2, \cdots, N\}$ corresponding to u_z; this is the one that maximizes the probability $\mathbb{P}\left[l_z = j | u_z, \mathbf{r} = \{r_j\}_{j=1}^{N}\right]$



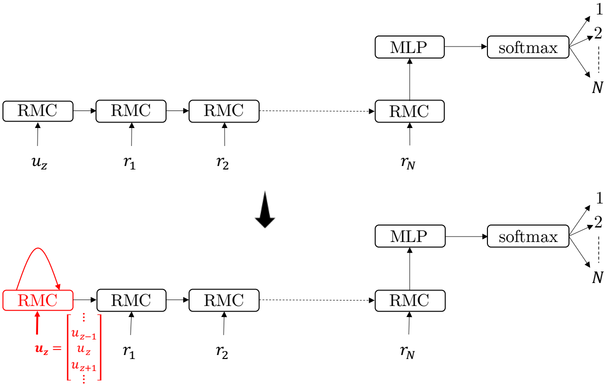
- Each element of the sequence is projected onto the "memory space".
- The RMC learns some *local* distance/similarity metric, sorts the distances and finds the r_j that minimizes the distance from u_z.

Incorporating Temporal Information

Segment length: a trade-off decision

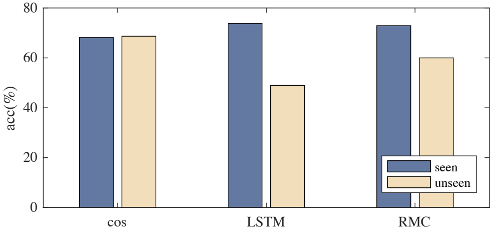
- short segments ⇒ unstable speaker representation
- long segments ⇒ multiple speakers in a single segment

Solution: reasonably short segments while keeping information from neigboring ones



Results on AMI

Simulated business meetings: 4 speakers per meeting

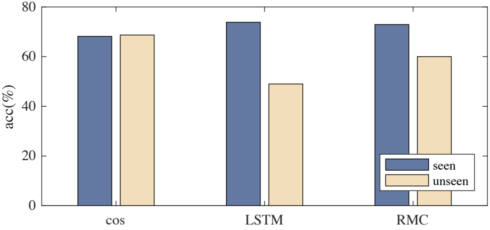


`oraclespk` segmentation, trained on AMI

- RMC captures distance information better than LSTM
- both networks fail to beat the baseline on unseen speakers (limited training speakers? ⇒ switch to VoxCeleb for training)

Results on AMI

Simulated business meetings: 4 speakers per meeting



`oraclespk` segmentation, trained on AMI

- RMC captures distance information better than LSTM
- both networks fail to beat the baseline on unseen speakers (limited training speakers? ⇒ switch to VoxCeleb for training)

system	training set	acc (%)
cos	–	68.68
RMC	AMI	60.00
	VoxCeleb clean	68.15
	VoxCeleb reverb	70.25
	VoxCeleb reverb+noise	71.90
RMC & context (±1)	VoxCeleb reverb+noise	**73.86**

`oraclespk` segmentation, evaluation on **`unseen`** AMI

Training with variable-length sequences

- Results on AMI

training seq length	4 spks	4-6 spks	2-9 spks	4-15 spks
w/o context	71.90	71.94	70.84	69.66
with context	73.86	73.77	72.67	73.42

System accuracy on `unseen` AMI set when trained with different ranges of sequence lengths. (always testing on sequences of 4 speakers)

Training with variable-length sequences

- Results on AMI

training seq length	4 spks	4-6 spks	2-9 spks	4-15 spks
w/o context	71.90	71.94	70.84	69.66
with context	73.86	73.77	72.67	73.42

System accuracy on `unseen` AMI set when trained with different ranges of sequence lengths. (always testing on sequences of 4 speakers)

- Results on real-world recorded business meetings
  9 real-world business meetings (4.6h): 4-15 speakers per meeting

	cos	RMC	RMC & context
`oraclevad` – SER (%) lower is better	20.95	18.56	**11.69**
`oraclespk` – acc (%) higher is better	70.66	72.51	**79.97**

System evaluation with different segmentation approaches on real meetings.

SAIL

Training with variable-length sequences

- Results on AMI

training seq length	4 spks	4-6 spks	2-9 spks	4-15 spks
w/o context	71.90	71.94	70.84	69.66
with context	73.86	73.77	72.67	73.42

System accuracy on **unseen** AMI set when trained with different ranges of sequence lengths. (always testing on sequences of 4 speakers)

- Results on real-world recorded business meetings

  9 real-world business meetings (4.6h): 4-15 speakers per meeting

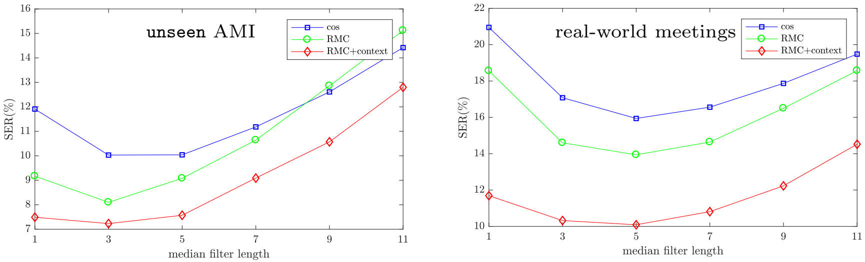
	cos	RMC	RMC & context
`oraclevad` – SER (%) lower is better	20.95	18.56	**11.69**
`oraclespk` – acc (%) higher is better	70.66	72.51	**79.97**

System evaluation with different segmentation approaches on real meetings.

Adding temporal context substantially improves the performance.
Can we do even better by incorporating temporal context at the decision level?

Smoothing at the Decision Level

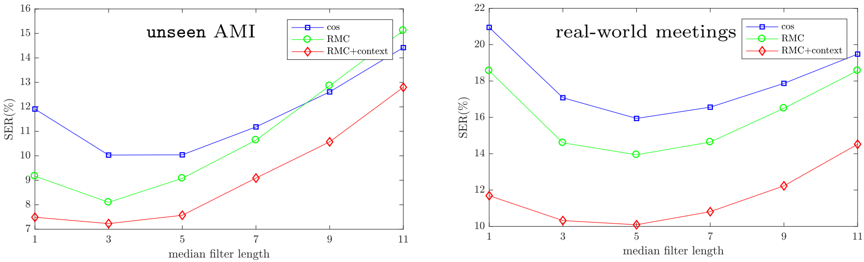
Assumption: highly improbable that isolated short segments correspond to some speaker in the middle of an utterance assigned to another speaker
$\Rightarrow$ Smooth the trajectory of the predicted speaker labels via median filtering.



System evaluation for the two datasets using different lengths of median filter for post-processing with the `oraclevad` segmentation. The RMC-based system is trained on sequences of 4-15 speakers.

Smoothing at the Decision Level

Assumption: highly improbable that isolated short segments correspond to some speaker in the middle of an utterance assigned to another speaker
⇒ Smooth the trajectory of the predicted speaker labels via median filtering.



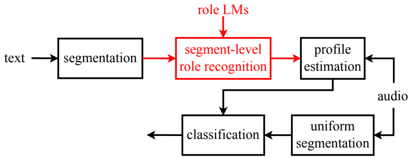
System evaluation for the two datasets using different lengths of median filter for post-processing with the **oraclevad** segmentation. The RMC-based system is trained on sequences of 4-15 speakers.

- A short median filter improves the performance for both datasets.
- Adding temporal context to the network partially acts like a data-driven smoothing filter.

Conclusion

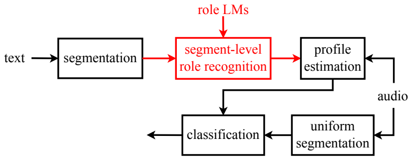
- Introduced a novel architecture for continuous speaker identification.
- Showed the importance of incorporating temporal context information both at the feature and the decision level.
- Demonstrated a SER relative reduction of 39.29% for the AMI corpus and 51.84% for the real-world business meetings, compared to the baseline when using oracle VAD information.

Towards Better Speaker Role Recognition



- Build N role-specific LMs $\mathcal{R}_i^+$ (N roles).
- Assign to each text segment x the role i that minimizes the perplexity $pp(x|\mathcal{R}_i^+)$.

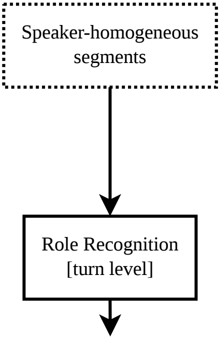
Towards Better Speaker Role Recognition



- Build N role-specific LMs $\mathcal{R}_i^+$ (N roles).
- Assign to each text segment x the role i that minimizes the perplexity $pp(x|\mathcal{R}_i^+)$.

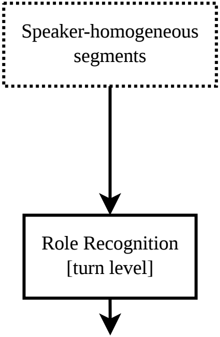
- Can we exploit the audio modality for the task of SRR?
- Do single utterances contain sufficient information for robust SRR?

Turn-level SRR



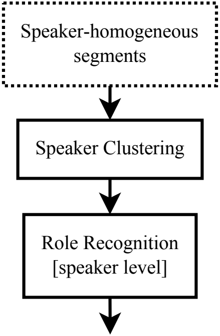
- each turn classified independently

Turn-level SRR



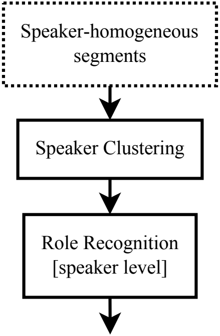
- each turn classified independently
- only role-specific information taken into account
- short segments do not contain enough information

Speaker-level SRR



- a role is assigned to each same-speaker cluster

Speaker-level SRR



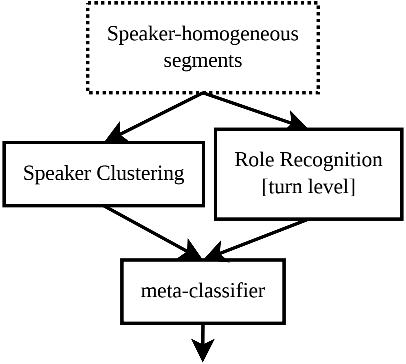
- a role is assigned to each same-speaker cluster
- error propagation between the modules

Solution?

Can we effectively combine speaker-specific and role-specific information towards better SRR performance?

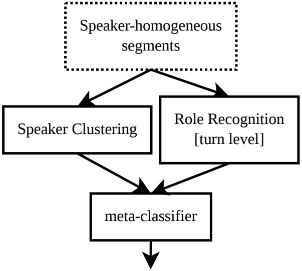
Solution?

Can we effectively combine speaker-specific and role-specific information towards better SRR performance?



N. Flemotomos, P. Papadopoulos, J. Gibson & S. Narayanan. "Combined Speaker Clustering and Role Recognition in Conversational Speech". Interspeech (2018)

Framework

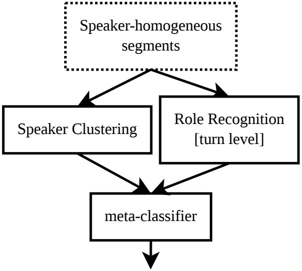


- speakers $\{S_i\}_{i=1}^N$
- roles $\{R_i\}_{i=1}^N$
- turns $x_1, x_2, \cdots, x_T$

- Speaker Clustering module:
  $(p_{1i})_{i=1}^N, (p_{2i})_{i=1}^N, \cdots, (p_{Ti})_{i=1}^N$, s.t. $x_k \leftarrow S_m$ iff $p_{km} = \max_i p_{ki}$
- Role Recognition module:
  $(q_{1i})_{i=1}^N, (q_{2i})_{i=1}^N, \cdots, (q_{Ti})_{i=1}^N$, s.t. $x_k \leftarrow R_m$ iff $q_{km} = \max_i q_{ki}$

- x_k is represented by the $2N$ scores $(p_{ki})_{i=1}^N$ and $(q_{ki})_{i=1}^N$

Framework



- speakers $\{S_i\}_{i=1}^N$
- roles $\{R_i\}_{i=1}^N$
- turns $x_1, x_2, \cdots, x_T$

- Speaker Clustering module:
  $(p_{1i})_{i=1}^N, (p_{2i})_{i=1}^N, \cdots, (p_{Ti})_{i=1}^N$, s.t. $x_k \leftarrow S_m$ iff $p_{km} = \max_i p_{ki}$
- Role Recognition module:
  $(q_{1i})_{i=1}^N, (q_{2i})_{i=1}^N, \cdots, (q_{Ti})_{i=1}^N$, s.t. $x_k \leftarrow R_m$ iff $q_{km} = \max_i q_{ki}$
- optimal mapping $M : \{S_i\}_{i=1}^N \rightarrow \{R_i\}_{i=1}^N$ defined as

role recognition module prediction

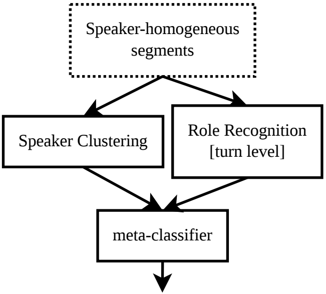
$$\hat{M} = \arg\min_M \sum_{k=1}^T \mathbb{I}(M(S'_k) \neq R'_k) d_k \quad (d_k \text{ is } x_k\text{'s duration})$$

all possible mappings

speaker clustering module prediction

- x_k is represented by the $2N$ scores $(p_{ki})_{i=1}^N$ and $(q_{ki})_{i=1}^N$

Experimental Setup and Datasets



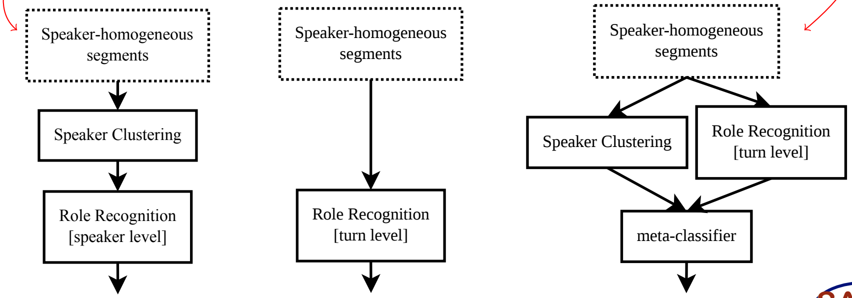
- Speaker Clustering: BIC-based hierarchical clustering, with one Gaussian modeling each cluster.
- Role Recognition: LM-based (3-gram models) and AM-based (512-component GMMs)
- meta-classifier: linear SVM

- Dyadic interactions from the psychology domain
  - *MI corpus*: Motivational Interviewing sessions between Therapist (T) and Client (Cl)
  - *ADOS corpus*: Autism Diagnostic Observation Schedule assesments between Psychologist (P) and Child (Ch)

Results

Table: Misclassification Rates (%) of the different components when used independently and when combined.
$\mathcal{R}^\dagger$: 0-error algorithm, SC: Speaker Clustering, LM & AM: Language & Acoustic Model

	$SC+\mathcal{R}^\dagger$ piped	LM only	SC+LM comb	AM only	SC+AM comb	AM+LM comb	SC+AM+LM comb
MI	3.59	9.49	2.76	35.45	3.66	9.17	**2.71**
ADOS	12.67	12.37	7.70	14.03	10.58	8.02	**5.98**

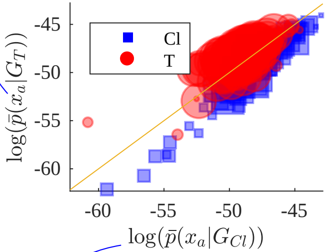


Results

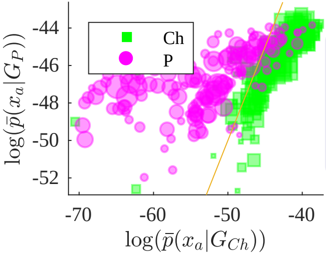
Table: Misclassification Rates (%) of the different components when used independently and when combined.
$\mathcal{R}^\dagger$: 0-error algorithm, SC: Speaker Clustering, LM & AM: Language & Acoustic Model

	SC+$\mathcal{R}^\dagger$ piped	LM only	SC+LM comb	AM only	SC+AM comb	AM+LM comb	SC+AM+LM comb
MI	3.59	9.49	2.76	35.45	3.66	9.17	**2.71**
ADOS	12.67	12.37	7.70	14.03	10.58	8.02	**5.98**

acoustic representation of a turn



(a) MI
(Therapist vs. Client)



(b) ADOS
(Psychologist vs. Child)

- 300 turns of the test set in each graph
- size proportional to duration

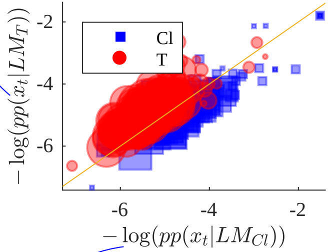
averaged log-likelihood

Results

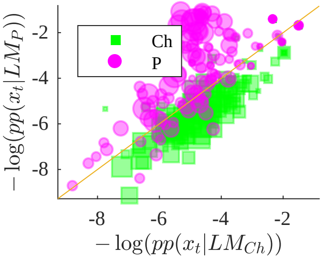
Table: Misclassification Rates (%) of the different components when used independently and when combined.
$\mathcal{R}^{\dagger}$: 0-error algorithm, SC: Speaker Clustering, LM & AM: Language & Acoustic Model

	SC+$\mathcal{R}^{\dagger}$ piped	LM only	SC+LM comb	AM only	SC+AM comb	AM+LM comb	SC+AM+LM comb
MI	3.59	9.49	2.76	35.45	3.66	9.17	**2.71**
ADOS	12.67	12.37	7.70	14.03	10.58	8.02	**5.98**

lexical representation of a turn



(a) MI
(Therapist vs. Client)



(b) ADOS
(Psychologist vs. Child)

- 300 turns of the test set in each graph
- size proportional to duration

negative log-perplexity

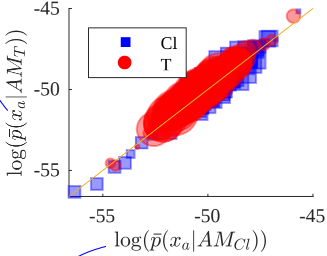
Results

Table: Misclassification Rates (%) of the different components when used independently and when combined.

$\mathcal{R}^\dagger$: 0-error algorithm, SC: Speaker Clustering, LM & AM: Language & Acoustic Model

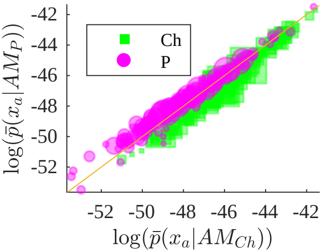
	SC+$\mathcal{R}^\dagger$ piped	LM only	SC+LM comb	AM only	SC+AM comb	AM+LM comb	SC+AM+LM comb
MI	3.59	9.49	2.76	35.45	3.66	9.17	**2.71**
ADOS	12.67	12.37	7.70	14.03	10.58	8.02	**5.98**

acoustic representation of a turn



(a) MI
(Therapist vs. Client)

averaged log-likelihood



(b) ADOS
(Psychologist vs. Child)

- 300 turns of the test set in each graph
- size proportional to duration

Conclusion

- We proposed a framework to incorporate speaker-specific and role-specific information for the SRR task, overcoming the problem of error propagation.
- We demonstrated an overall relative improvement equal to 24.5% for the MI corpus (Therapist vs. Client) and 52.8% for the ADOS corpus (Psychologist vs. Child).

Drawbacks

- more data required to train the meta-classifier
- evaluated using manually derived speaker turns and transcriptions

Thank you!

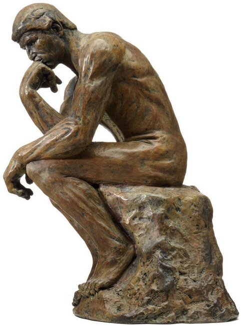
Microsoft

USC University of Southern California

THE UNIVERSITY OF UTAH®

UCSF
University of California San Francisco

W
UNIVERSITY of WASHINGTON



Questions and Discussion

NIH National Institutes of Health
Turning Discovery Into Health

SAIL