

Who Spoke When?

and how speaker roles can help us find the answer

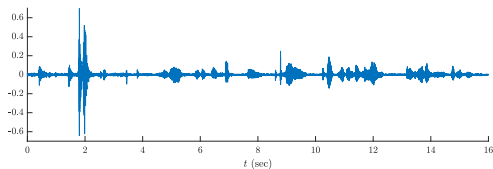
Nikolaos (Nikos) Flemotomos

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

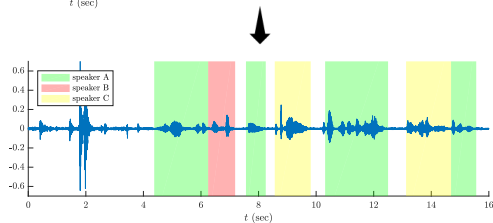
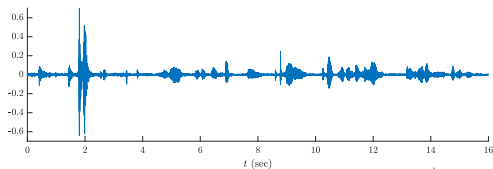
April 2, 2021



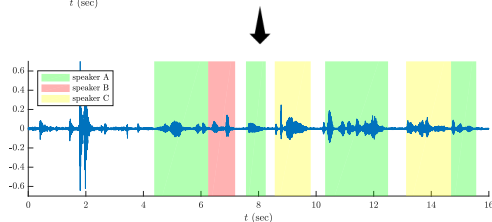
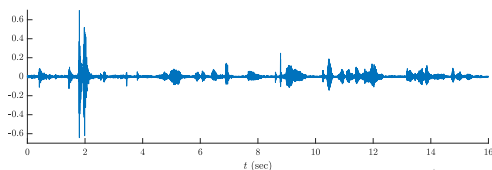
Who Spoke When: Continuous Speaker Identification



Who Spoke When: Continuous Speaker Identification



Who Spoke When: Continuous Speaker Identification

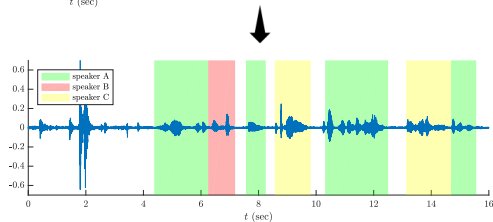
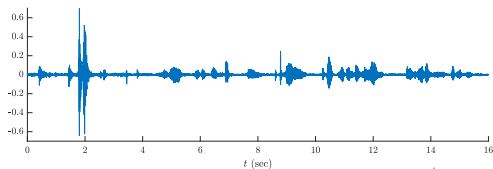


Why?

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



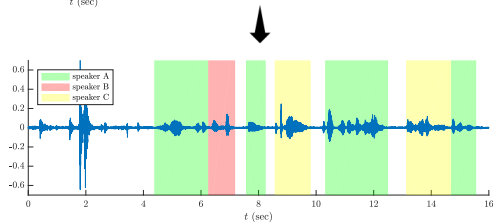
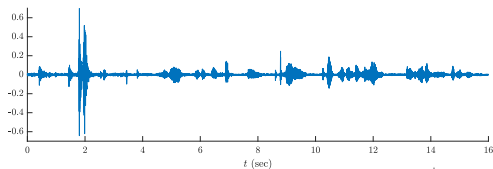
Who Spoke When: Continuous Speaker Identification



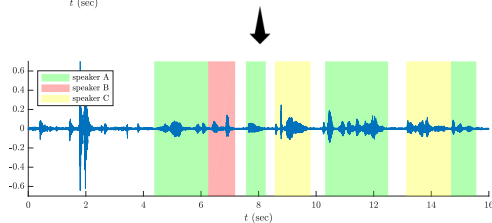
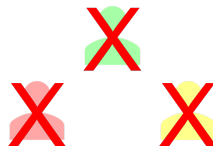
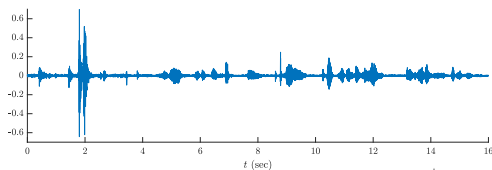
Traditional approach

- 1 segmentation
- 2 classification

Who Spoke When: Speaker Diarization

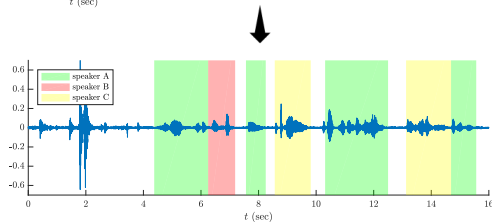
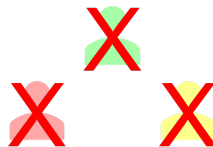
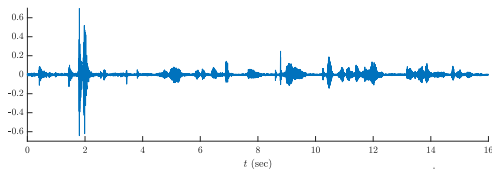


Who Spoke When: Speaker Diarization



No a priori information about the speakers is given!

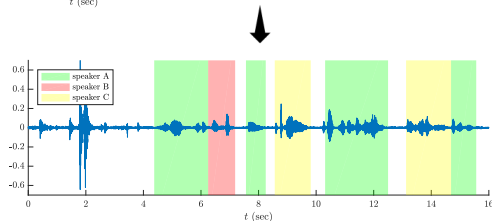
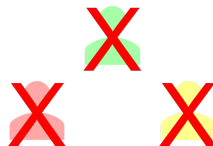
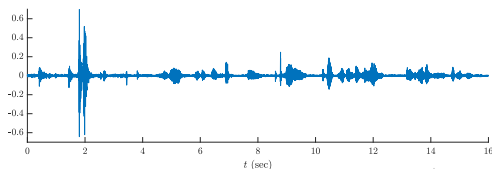
Who Spoke When: Speaker Diarization



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
 - ...



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?

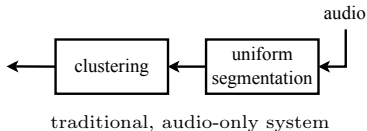
Proposed System

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



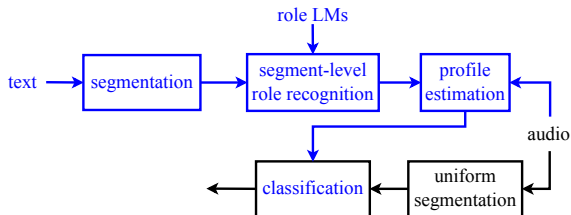
Proposed System

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



Proposed System

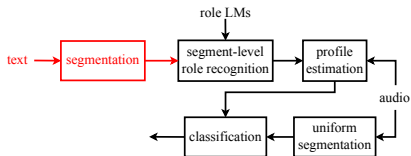
- different *roles* \Rightarrow distinguishable linguistic patterns
 \Rightarrow Can we use language to assist diarization?



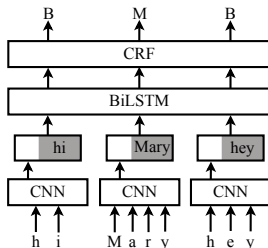
proposed, linguistically-aided system

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

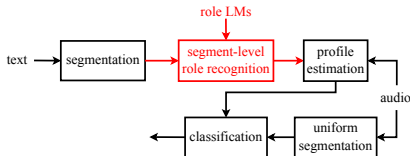
Proposed System: Text-based segmentation



- Goal: obtain speaker-homogeneous text segments
- Assumption: single speaker per sentence
⇒ segment text at the sentence level
- sequence-labeling problem → CNN-BiLSTM-CRF architecture



Proposed System: Role recognition



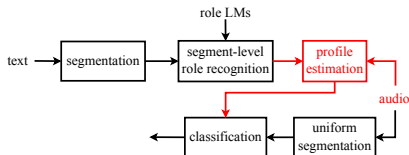
- Build a background LM \mathcal{G} and N role-specific LMs \mathcal{R}_i (N roles).
- Interpolate the LMs (n-gram):

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

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

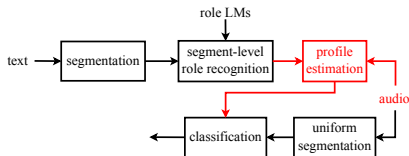
- Assign to each text segment x the role i that minimizes the perplexity $pp(x|\mathcal{R}_i^+)$.

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 :

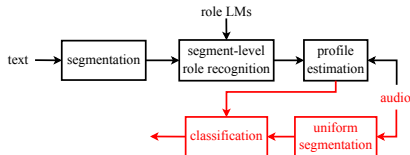
$$c_x = \min_{j \neq i} |pp(x|\mathcal{R}_j^+) - pp(x|\mathcal{R}_i^+)|$$

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

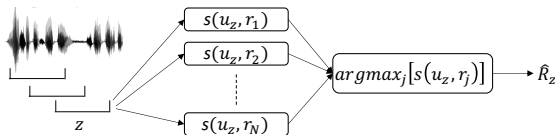
$$r_i = \frac{\sum_{x \in R_i} \mathbb{I}\{c_x > \theta\} u_x}{\sum_{x \in R_i} \mathbb{I}\{c_x > \theta\}}$$



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 $s(u_z, r_i)$.



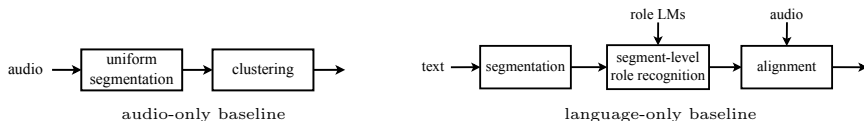
Evaluation Dataset and Baselines

- 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



Results

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

Table: DER (%) on PSYCH-test.

- tagger oversegments
⇒ short segments contain insufficient information for
role recognition
⇒ 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 tagger	11.05	12.99	7.28	6.99
			20.09	7.71	7.30
ASR	tagger	11.05	27.07	8.37	7.84

Table: DER (%) on PSYCH-test.

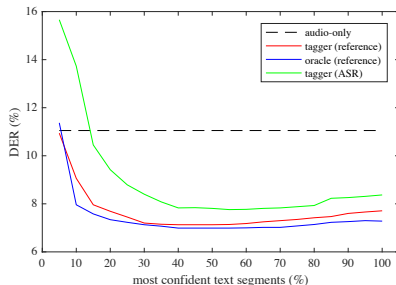
- high WER \Rightarrow severe degradation for language-only system
- when transcripts are only used for profile estimation (linguistically-aided) the performance gap is much smaller

Results

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

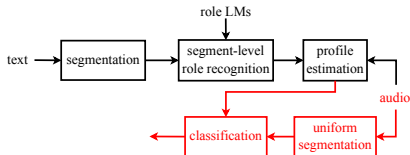
Table: DER (%) on PSYCH-test.

- 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

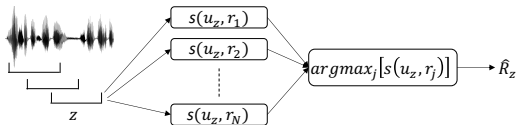


- Proposed a system for **speaker diarization** in conversational scenarios where the speakers assume specific **roles**.
- Used the **lexical information** captured within the speech signal in order to estimate the speaker profiles and follow a **classification approach instead of clustering**.
- Evaluated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

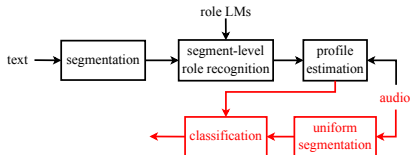
Towards Better Classification



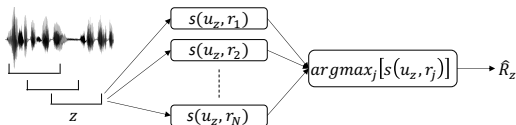
- Segment uniformly the speech signal (sliding window).
- Extract an acoustic speaker embedding (x-vector) $u_z \forall$ segment z
- Calculate the similarity $s(u_z, r_i) \forall$ role profile r_i .
- Assign to the audio segment z the role i that maximizes $s(u_z, r_i)$.



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 $s(u_z, r_i)$.



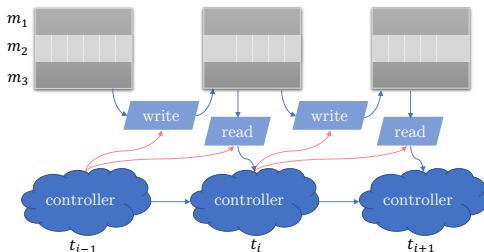
Problems

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



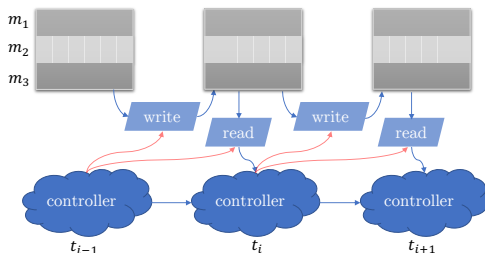
Memory-Augmented Neural Networks

- 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 \Rightarrow can learn a task-specific organization of the memory in a supervised manner through gradient descent.



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 \Rightarrow can learn a task-specific organization of the memory in a supervised manner through gradient descent.



In our implementation: Relational Memory Core (RMC)

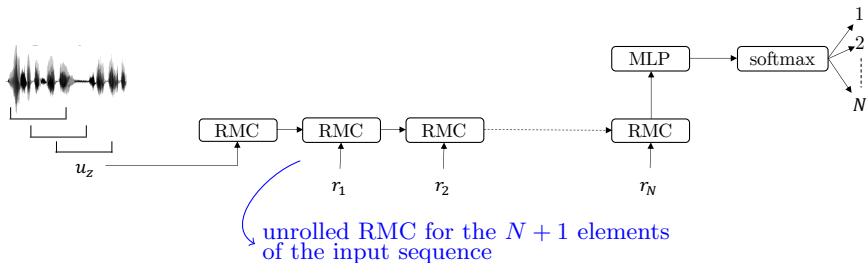
- controller [RMC] is embedded into an LSTM
- memory updates are based on a self-attention mechanism

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



- 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 .

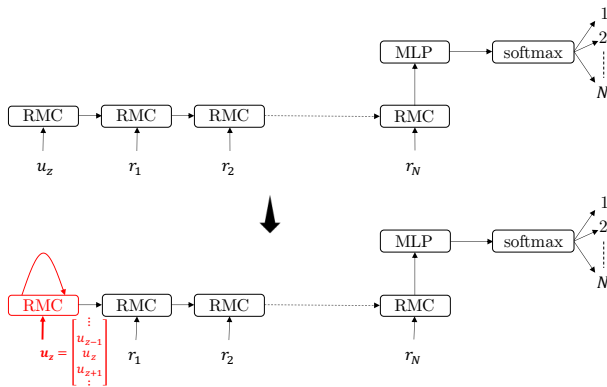


Incorporating Temporal Information

Segment length: a trade-off decision

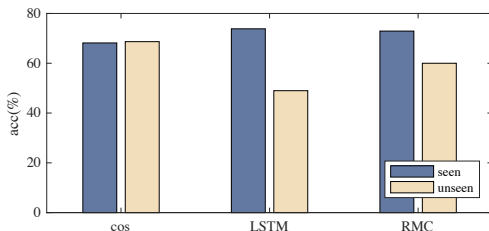
- short segments \Rightarrow unstable speaker representation
- long segments \Rightarrow multiple speakers in a single segment

Solution: reasonably short segments while keeping information from neighboring ones



Results on AMI

Simulated business meetings: 4 speakers per meeting

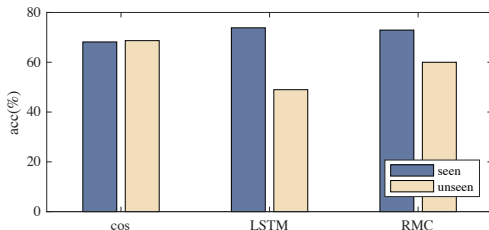


oraclespk segmentation, trained on AMI

- RMC captures distance information better than LSTM
- both networks fail to beat the baseline on unseen speakers (limited training speakers? \Rightarrow 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? \Rightarrow 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.



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.

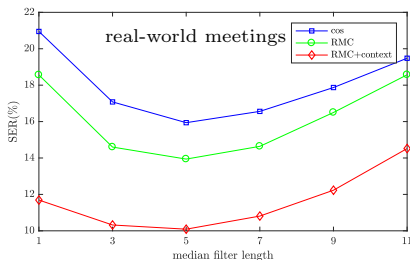
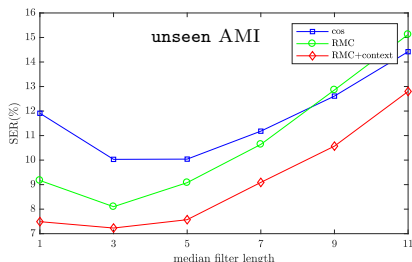
Can we do even better by incorporating temporal context at the decision level?



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

⇒ Smooth the trajectory of the predicted speaker labels via median filtering.

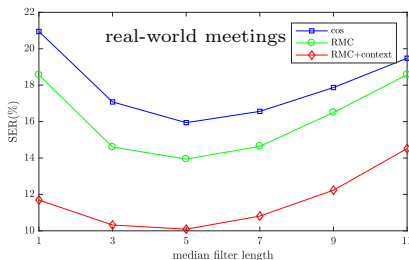
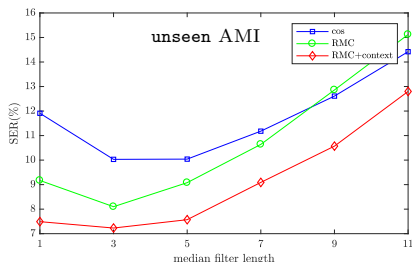


System evaluation for the two datasets using different lengths of median filter for post-processing with the `oraclevad` segmentation. The RMC-based system is trained on sequences of 4-15 speakers.

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 **oracle** vad segmentation. The RMC-based system is trained on sequences of 4-15 speakers.

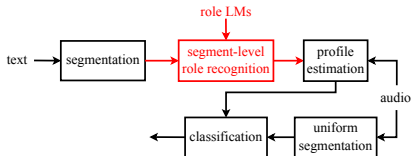
- A short median filter improves the performance for both datasets.
- Adding temporal context to the network partially acts like a data-driven smoothing filter.



- 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.

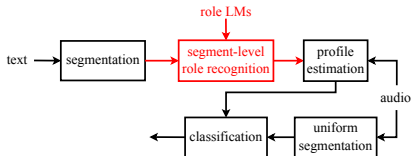


Towards Better Speaker Role Recognition

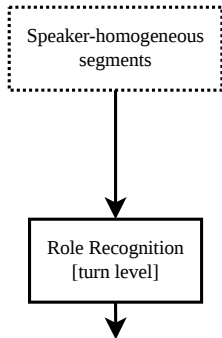


- Build N role-specific LMs \mathcal{R}_i^+ (N roles).
- Assign to each text segment x the role i that minimizes the perplexity $pp(x|\mathcal{R}_i^+)$.

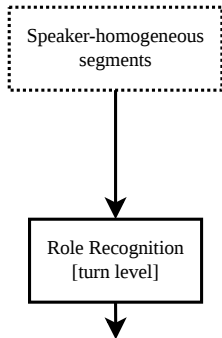
Towards Better Speaker Role Recognition



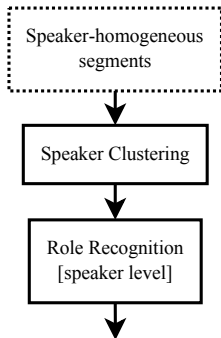
- Build N role-specific LMs \mathcal{R}_i^+ (N roles).
- Assign to each text segment x the role i that minimizes the perplexity $pp(x|\mathcal{R}_i^+)$.
- Can we exploit the audio modality for the task of SRR?
- Do single utterances contain sufficient information for robust SRR?



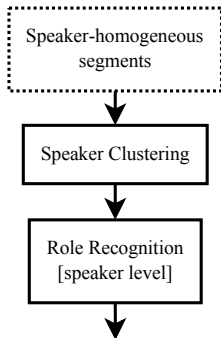
- each turn classified independently



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



- a role is assigned to each same-speaker cluster

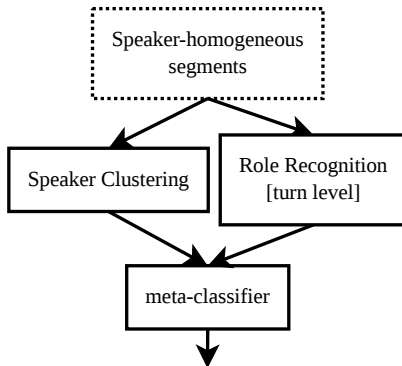


- a role is assigned to each same-speaker cluster
- error propagation between the modules

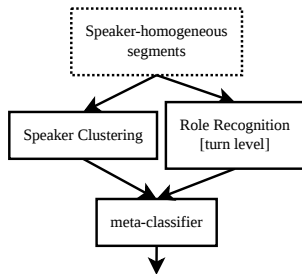
Can we effectively combine speaker-specific and role-specific information towards better SRR performance?

Solution?

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



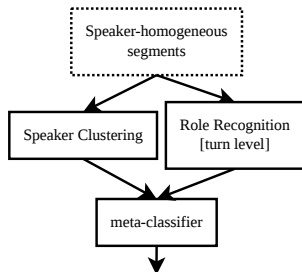
Framework



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

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



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

- Speaker Clustering module:

$(p_{1i})_{i=1}^N, (p_{2i})_{i=1}^N, \dots, (p_{Ti})_{i=1}^N$, s.t. $x_k \leftarrow S_m$ iff $p_{km} = \max_i p_{ki}$

- Role Recognition module:

$(q_{1i})_{i=1}^N, (q_{2i})_{i=1}^N, \dots, (q_{Ti})_{i=1}^N$, s.t. $x_k \leftarrow R_m$ iff $q_{km} = \max_i q_{ki}$

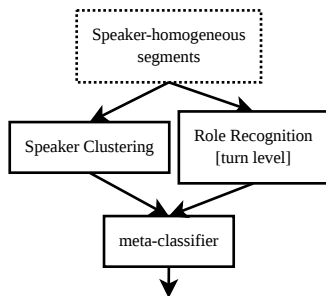
- optimal mapping $M : \{S_i\}_{i=1}^N \rightarrow \{R_i\}_{i=1}^N$ defined as

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

\swarrow all possible mappings \searrow 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



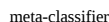
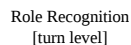
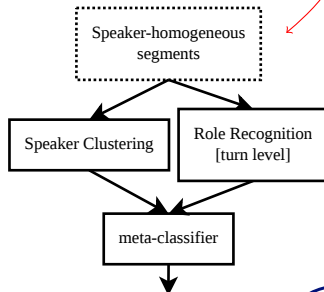
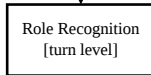
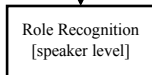
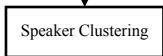
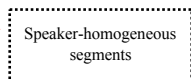
- Speaker Clustering: BIC-based hierarchical clustering, with one Gaussian modeling each cluster.
- Role Recognition: LM-based (3-gram models) and AM-based (512-component GMMs)
- meta-classifier: linear SVM

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

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 pipelined	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



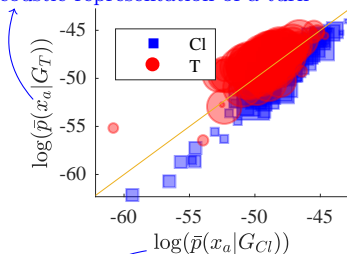
Results

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

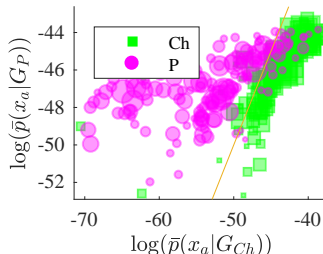
\mathcal{R}^\dagger : 0-error algorithm, SC: Speaker Clustering, LM & AM: Language & Acoustic Model

	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

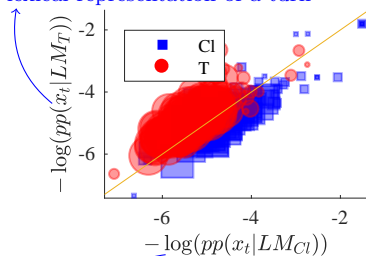
Results

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

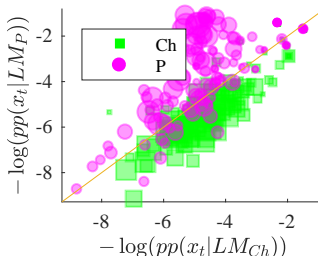
\mathcal{R}^\dagger : 0-error algorithm, SC: Speaker Clustering, LM & AM: Language & Acoustic Model

	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



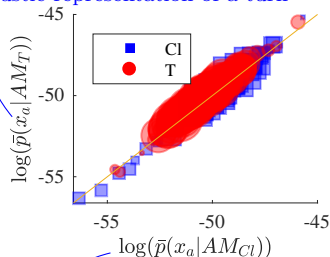
Results

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

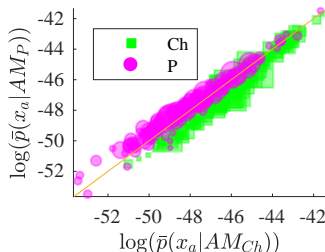
\mathcal{R}^\dagger : 0-error algorithm, SC: Speaker Clustering, LM & AM: Language & Acoustic Model

	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)



(b) ADOS
(Psychologist vs. Child)

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

averaged log-likelihood

- 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

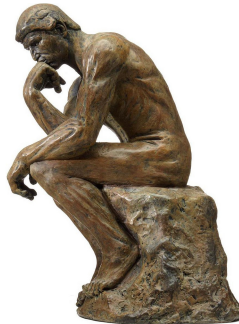
- more data required to train the meta-classifier
- evaluated using manually derived speaker turns and transcriptions



Thank you!



University of California
San Francisco



Questions and Discussion

Linguistically Aided Diarization: Datasets

- 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).

- Text-based tagger training corpus:
Fisher English transcriptions (telephone conversations)
- LM training corpora:
Fisher (background), PSYCH-train, CPTS (text-only therapy data)

	PSYCH-train	Fisher	CPTS
voc	8.17K	58.6K	35.6K
#tokens	530K	21.0M	6.52M

Table: Size of the corpora used for LM training.



Linguistically Aided Diarization: Experimental Setup

sentence tagger

- 4 CNN, 2 BiLSTM layers
- dropout ($p = 0.5$),
 l_2 regularization ($\lambda = 10^{-8}$)
- F_1 score = 0.805 (14 epochs)

uniform segmentation & embeddings

- pre-trained VoxCeleb x-vector extractor
- PLDA adapted on PSYCH
- segmentation window
length = 1.5 sec, hop = 0.25 sec

ASR

- pre-trained Kaldi ASPIRE AM
- 3-gram in-domain LM
- WER = 39.78% (PSYCH-test)

decoding & evaluation

- initial oracle silence-based segmentation (1 sec threshold)
- 0.25 sec collar (metric: DER)
- ignore overlapping speech

Appendix - Relational Recurrent MANNs: Controller

Let a memory matrix M with memory slots m_1, m_2, \dots

- Updates are based on a self-attention mechanism:

- assume no new observations

$$\tilde{M} = softmax \left(\frac{(MW^q)(MW^k)^T}{\sqrt{d_k}} \right) MW^v$$

Diagram annotations: "queries" points to MW^q , "keys" points to $(MW^k)^T$, "values" points to MW^v , and "key dimensionality" points to $\sqrt{d_k}$.

- incorporate a new observation x

$$\tilde{M} = softmax \left(\frac{(MW^q)([M; x]W^k)^T}{\sqrt{d_k}} \right) [M; x]W^v$$

Note that memory matrix dimensions do not change.

- Each memory attends to all the other memories to be updated \Rightarrow cross-memory relations are encoded

Appendix - Relational Recurrent MANNs: Recurrency

Each memory m_i is embedded into an LSTM. The resulting controller is called *Relational Memory Core (RMC)*.

$$s_{i,t} = (h_{i,t-1}, m_{i,t-1})$$

$$f_{i,t} = W^f x_t + U^f h_{i,t-1} + b^f$$

$$i_{i,t} = W^i x_t + U^i h_{i,t-1} + b^i$$

$$o_{i,t} = W^o x_t + U^o h_{i,t-1} + b^o$$

denotes element-wise multiplication

$$m_{i,t} = \sigma(f_{i,t} + \tilde{b}^f) * m_{i,t-1} + \sigma(i_{i,t}) * g(\tilde{m}_{i,t})$$

$$h_{i,t} = \sigma(o_{i,t}) * \tanh(m_{i,t})$$

$$s_{i,t+1} = (h_{i,t}, m_{i,t})$$

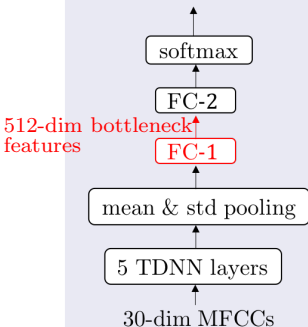
- $g(\cdot)$ is some non-linear function. In practice, modelled by a 2-layer fully connected MLP.
- all the weights and biases are shared across the memories $m_i \forall i$ more memory slots \nRightarrow more trainable parameters



Appendix - Speaker Embeddings: x-vectors

golden standard for tasks including → speaker recognition
→ diarization
→ language identification

Architecture



Training procedure

- speaker recognition on VoxCeleb 1, 2
- 2 – 4sec long speech segments
- 16kHz audio

Feature normalization

- LDA projection on a d -dimensional space ($d = 200$)
- mean- and length-normalization to $l = \sqrt{d}$



AMI Meeting Corpus

- 4 speakers per meeting
- 31 scenarios with 4 business meetings each (124 speakers):
 - 62 meetings for training (35.5*h*) → distant mic
 - 31 for evaluation [**seen**] (17.1*h*) → distant mic
 - 31 for profile extraction (8.0*h*) → close-talk mic
- 6 additional meetings with **unseen** speakers for evaluation (4.1*h*) plus 6 for profile estimation (3.8*h*)
- subsegments every 0.75*sec* both for training and evaluation (160K subsegments for training)

VoxCeleb

- 6490 speakers with > 6 utterances
- select 3 utts per speaker for profile estimation (totally 383.3h)
- subsegment the rest every 10sec (840.6K training subsegments)
- randomly create sequences of subsegments and speaker profiles for training

Internal Microsoft Meetings

- 9 business meetings (4.6h)
- subsegments every 0.75sec
- 4 – 15 speakers per meeting
- speakers already enrolled



Appendix - Experimental Setup

Training details

- Handling speaker ordering: Randomly permute the speaker profiles in every training sequence.
- When training with variable-length sequences, fix the length in each mini-batch.

Network parametrization

- $\# \text{memories} = \max \# \text{speakers} + 1$
- memory size = 2048
- MLP: 4 FC layers, 256 neurons

Evaluation metrics

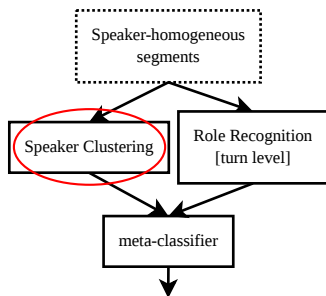
- **oraclespk**:
classification accuracy
- **oraclevad**:
Speaker Error Rate
(collar=0.25sec)

Tools

- *Kaldi* for feature and embedding extraction
- *Tensorflow* with *Sonnet* library to build the network
- *NIST md-eval.pl* for SER estimation



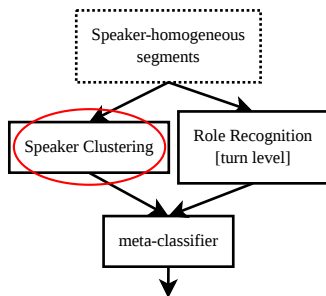
Speaker Clustering Module



Speaker Clustering module

- BIC-based algorithm, with one Gaussian modeling each cluster
- features: 13 MFCCs
- p_{ki} is the per-frame log-likelihood wrt the Gaussian corresponding to the i th speaker, averaged over the voiced frames of the turn x_k

Speaker Clustering Module

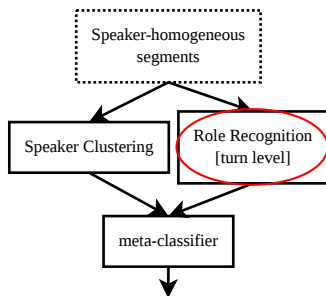


Speaker Clustering module

- BIC-based algorithm, with one Gaussian modeling each cluster
- features: 13 MFCCs
- p_{ki} is the per-frame log-likelihood wrt the Gaussian corresponding to the i th speaker, averaged over the voiced frames of the turn x_k

will be mapped to the corresponding role

Role Recognition Module



Role Recognition module – LM-based

- train one n -gram Language Model (LM) for each role
- q_{ki} is the negative log-perplexity of x_k wrt the LM corresponding to the i th role

Role Recognition module – AM-based

- train one GMM Acoustic Model (AM) for each role
- features: 13 MFCCs
- q_{ki} is the per-frame log-likelihood wrt the AM corresponding to the i th role, averaged over the voiced frames of the turn x_k

Dyadic interactions from the psychology domain

- *MI corpus*: Motivational Interviewing sessions between Therapist (T) and Client (Cl)
- *ADOS corpus*: Autism Diagnostic Observation Schedule assessments between Psychologist (P) and Child (Ch)

Table: Descriptive analysis of the corpora used.

	MI-train	MI-test	ADOS-train	ADOS-test
#sessions	242	101	141	132
mean_dur	27.24min	33.14min	3.67min	3.67min
std_dur	14.40min	17.42min	1.34min	1.65min
dur-T/P	47.30h	26.35h	2.63h	2.52h
dur-Cl/Ch	52.96h	25.87h	2.97h	2.98h
#T/P	123	53	—	—
#Cl/Ch	—	—	89	81

Therapist/Psychologist

Client/Child

- no overlapping speakers between the train/test sets