Combined Speaker Clustering and Role Recognition in Conversational Speech

Nikolaos Flemotomos, Pavlos Papadopoulos,
James Gibson, Shrikanth Narayanan

University of Southern California
Signal Analysis and Interpretation Laboratory

Interspeech 2018
September 4

Speaker Role Recognition

- Goal: assign a specific *role* to each speaker turn
  - role: characterized by the task a speaker performs and the objectives related to it

Speaker Role Recognition

- Goal: assign a specific *role* to each speaker turn
  - role: characterized by the task a speaker performs and the objectives related to it
- Examples:
  - broadcast news programs
  - business meetings
  - psychotherapy sessions
  - ...


https://www.alternet.org/sites/default/files/story_images/news.jpg
https://img.clipartxtras.com/6f8c4841fb7c7f646a2a5d3fe2290eb6_business-meeting-clipart-free-clip-art-library-business-meeting-clipart-images_2560-1024.jpeg
https://cdn-images-1.medium.com/max/1600/1*BHBHQcA30AjIpkEJFoVJxg.jpeg

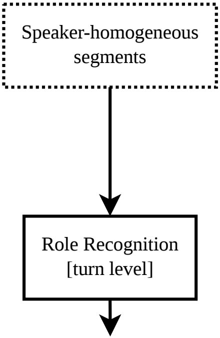
Speaker Role Recognition

- Goal: assign a specific *role* to each speaker turn
  - role: characterized by the task a speaker performs and the objectives related to it
- Examples:
  - broadcast news programs
  - business meetings
  - psychotherapy sessions
  - ...


- Turn-level vs. Speaker-level SRR

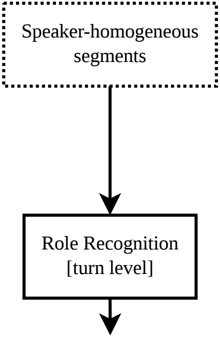
https://www.alternet.org/sites/default/files/story_images/news.jpg
https://img.clipartxtras.com/6f8c4841fb7c7f646a2a5d3fe2290eb6_business-meeting-clipart-free-clip-art-library-business-meeting-clipart-images_2560-1024.jpeg
https://cdn-images-1.medium.com/max/1600/1*BHBHQcA30AjIpkEJFoVJxg.jpeg

Turn-level SRR



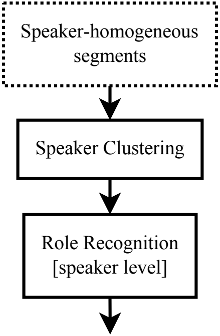
- each turn classified independently

Turn-level SRR



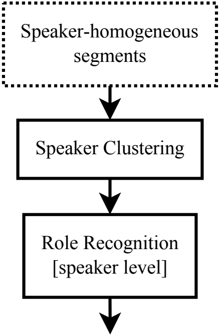
- each turn classified independently
- only role-specific information taken into account

Speaker-level SRR



- a role is assigned to each same-speaker cluster

Speaker-level SRR



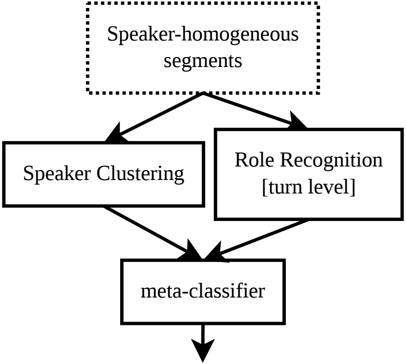
- a role is assigned to each same-speaker cluster
- error propagation between the modules

Solution?

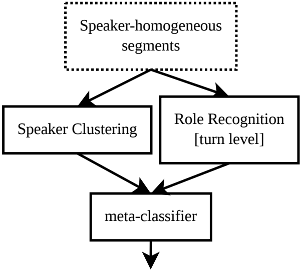
Can we effectively combine speaker-specific and role-specific information towards better SRR performance?

Solution?

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



Framework

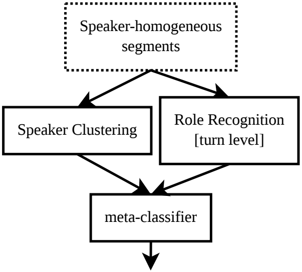


- speakers $\{S_i\}_{i=1}^N$
- roles $\{R_i\}_{i=1}^N$
- turns $x_1, x_2, \cdots, x_T$

- 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

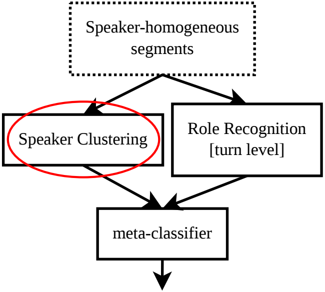
$$\hat{M} = \arg\min_{M} \sum_{k=1}^{T} \mathbb{I}(M(S'_k) \neq R'_k) d_k \quad (d_k \text{ is } x_k\text{'s duration})$$

all possible mappings

speaker clustering module prediction

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

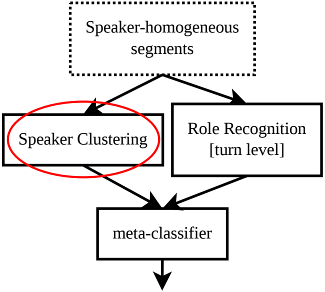
Speaker Clustering Module



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 ith speaker, averaged over the voiced frames of the turn x_k

Speaker Clustering Module

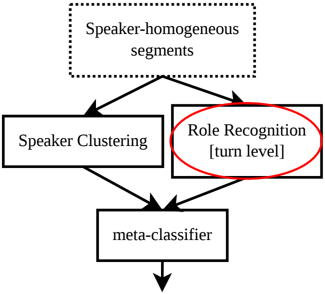


Speaker Clustering module

- 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 ith 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 ith 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 ith role, averaged over the voiced frames of the turn x_k

Datasets

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)

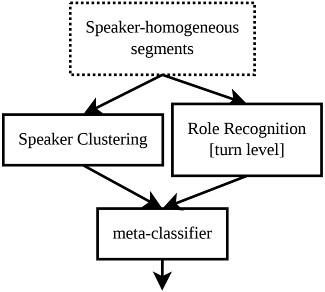
Table: Descriptive analysis of the corpora used.

	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



- train the LMs (3-gram models) and AMs (512-component GMMs) for all the roles on the training set
- linear support vector machine as meta-classifier

- 5-fold cross-validation on the test set
- evaluation metric: Misclassification Rate (MR)

true role

duration

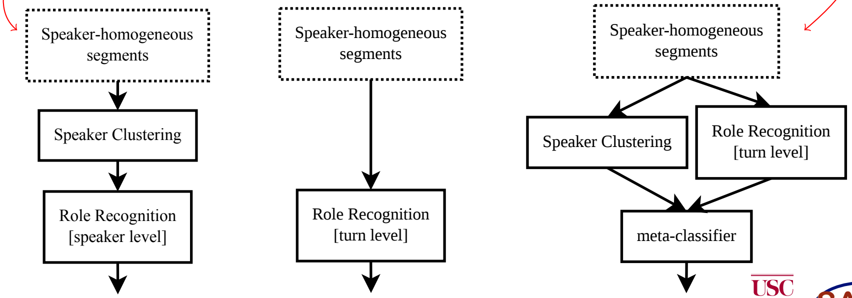
$$\mathrm{MR} = \frac{\#\text{misclassified frames}}{\text{total } \#\text{frames}} = \frac{\sum_k \mathbb{I}(R_k \neq \hat{R}_k) d_k}{\sum_k d_k}$$

predicted role

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**

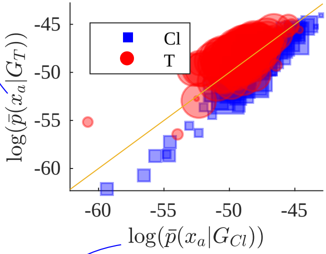


Results

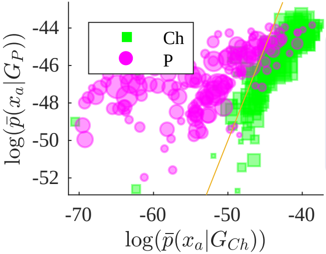
Table: Misclassification Rates (%) of the different components when used independently and when combined.
$\mathcal{R}^\dagger$: 0-error algorithm, SC: Speaker Clustering, LM & AM: Language & Acoustic Model

	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

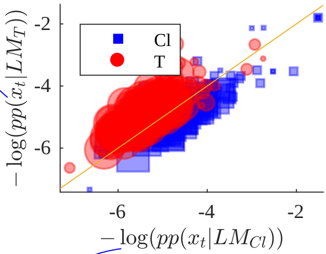
averaged log-likelihood

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

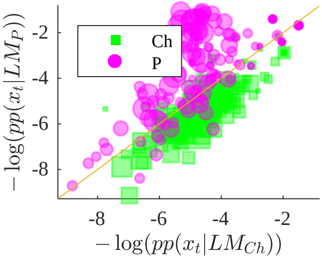
	SC+$\mathcal{R}^{\dagger}$ piped	LM only	SC+LM comb	AM only	SC+AM comb	AM+LM comb	SC+AM+LM comb
MI	3.59	9.49	2.76	35.45	3.66	9.17	**2.71**
ADOS	12.67	12.37	7.70	14.03	10.58	8.02	**5.98**

lexical representation of a turn



(a) MI
(Therapist vs. Client)

negative log-perplexity



(b) ADOS
(Psychologist vs. Child)

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

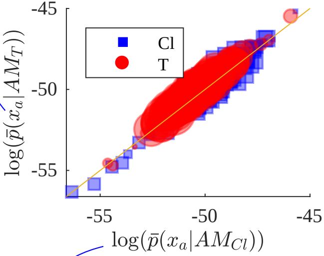
Results

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

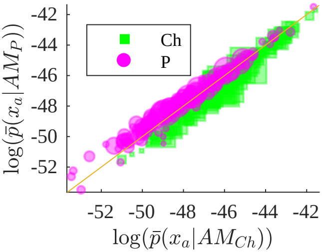
	SC+$\mathcal{R}^{\dagger}$ piped	LM only	SC+LM comb	AM only	SC+AM comb	AM+LM comb	SC+AM+LM comb
MI	3.59	9.49	2.76	35.45	3.66	9.17	**2.71**
ADOS	12.67	12.37	7.70	14.03	10.58	8.02	**5.98**

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

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**

Final relative improvement wrt piped architecture:

- 24.5% for the MI corpus (Therapist vs. Client)
- 52.8% for the ADOS corpus (Psychologist vs. Child)

Conclusions

We proposed a framework to incorporate *speaker-specific* and *role-specific* information for the SRR task, *overcoming the problem of error propagation.*

Drawbacks

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

Future Work

- apply the method to multi-role databases
- formulate the framework to accomodate more than one speaker clustering and/or role recognition modules