Extracting and Using Speaker Role Information in Speech Processing Applications

Nikolaos (Nikos) Flemotomos

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

Ph.D. Dissertation
May 9, 2022

Advisor: Prof. Shrikanth Narayanan

Example scenarios:

- business meetings
- doctor-patient interactions
- broadcast news programs
- call centers
- lectures
- interviews
- ...

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

What is a role? Examples

Example scenarios:

- business meetings
- doctor-patient interactions
- broadcast news programs
- call centers
- lectures
- interviews
- ...

For every role we assume, we adopt specific behaviors to achieve particular goals.

According to social psychology roles are...

functions associated with a position in a group with rights and duties toward one or more other group members.

A. P. Hare. "Types of Roles in Small Groups: A Bit of History and a Current Perspective", Small Group Research (1994)

What is a role? Social perspective

According to social psychology roles are...

functions associated with a position in a group with rights and duties toward one or more other group members.

- roles are defined within the context of group interactions

A. P. Hare. "Types of Roles in Small Groups: A Bit of History and a Current Perspective", Small Group Research (1994)

According to social psychology roles are...

functions associated with a position in a group with rights and duties toward one or more other group members.

- roles are defined within the context of group interactions
- they guide our behaviors

A. P. Hare. "Types of Roles in Small Groups: A Bit of History and a Current Perspective", Small Group Research (1994)

What is a role? Social perspective

According to social psychology roles are...

functions associated with a position in a group with rights and duties toward one or more other group members.

- roles are defined within the context of group interactions
- they guide our behaviors
- they create expectations about others' behaviors

A. P. Hare. "Types of Roles in Small Groups: A Bit of History and a Current Perspective", Small Group Research (1994)

Why do we care about roles?

- Role information is useful in several multimedia applications
  - information retrieval
  - automatic summarization
  - audio indexing
  - media browser enhancement

image from the Noun Project. creator: Kamin Ginkaew

Why do we care about roles?

- Role information is useful in several multimedia applications
  - information retrieval
  - automatic summarization
  - audio indexing
  - media browser enhancement

- ... or even essential for some tasks
  - quality assessment in psychotherapy sessions
  - performance evaluation of call center employees

image from the Noun Project. creator: Kamin Ginkaew

Role Typology

- formal

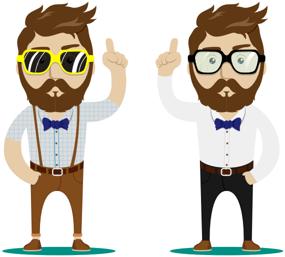
  e.g., interviewer vs. interviewee

- informal

  e.g., protagonist vs. supporter

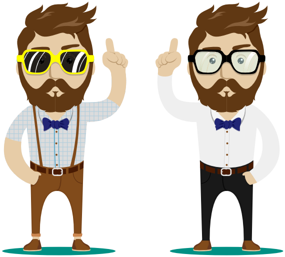
images from shutterstock. creators:Phakorn Kasikij, Lorelyn Medina

- formal
  e.g., interviewer vs. interviewee
- informal
  e.g., protagonist vs. supporter



- assigned implicitly
  e.g., lecturer vs. audience
- scripted
  e.g., roles in learning platforms
  or in psychodrama

images from shutterstock. creators:Phakorn Kasikij, Lorelyn Medina



- formal
  e.g., interviewer vs. interviewee
- informal
  e.g., protagonist vs. supporter

- assigned implicitly
  e.g., lecturer vs. audience
- scripted
  e.g., roles in learning platforms
  or in psychodrama

images from shutterstock. creators:Phakorn Kasikij, Lorelyn Medina

Conversations and Speaker Roles

- speaker roles are linked to specific communication patterns
- can be manifest through multiple modalities
- we focus on linguistic and acoustic characteristics
  - an interviewer is expected to use interrogative words
  - a teacher is expected to speak in a didactic style
  - a patient is expected to describe their symptoms

- speaker roles are linked to specific communication patterns
- can be manifest through multiple modalities
- we focus on linguistic and acoustic characteristics
  - an interviewer is expected to use interrogative words
  - a teacher is expected to speak in a didactic style
  - a patient is expected to describe their symptoms

- role recognition depends on successful speech (pre-)processing steps

speech processing

- speech activity detection
- speech recognition
- speaker segmentation
- ...

speaker role information

- speaker roles are linked to specific communication patterns
- can be manifest through multiple modalities
- we focus on linguistic and acoustic characteristics
  - an interviewer is expected to use interrogative words
  - a teacher is expected to speak in a didactic style
  - a patient is expected to describe their symptoms

- role recognition depends on successful speech (pre-)processing steps

speech processing

- speech activity detection
- speech recognition
- speaker segmentation
- ...

error propagation?

speaker role information

Conversations and Speaker Roles

- speaker roles are linked to specific communication patterns
- can be manifest through multiple modalities
- we focus on linguistic and acoustic characteristics
  - an interviewer is expected to use interrogative words
  - a teacher is expected to speak in a didactic style
  - a patient is expected to describe their symptoms

- role recognition depends on successful speech (pre-)processing steps

speech processing

- speech activity detection
- speech recognition
- speaker segmentation
- ...

speaker role information

- role information is beneficial for speech processing tasks

Thesis Statement

The behavioral patterns found within conversational interactions can help us *recognize* speaker roles and *use* them towards improved performance in speech processing tasks.

Outline

- Extracting Speaker Roles and alleviating error propagation
  - Effective speaker clustering for role recognition
  - Effective speech recognition for role recognition

- Using Speaker Roles to answer "*who spoke when*"
  - Use roles to reduce speaker clustering to speaker classification
  - Use roles to impose constraints on speaker clustering

Outline

- Extracting Speaker Roles and alleviating error propagation
  - Effective speaker clustering for role recognition
  - Effective speech recognition for role recognition

- Using Speaker Roles to answer "*who spoke when*"
  - Use roles to reduce speaker clustering to speaker classification
  - Use roles to impose constraints on speaker clustering

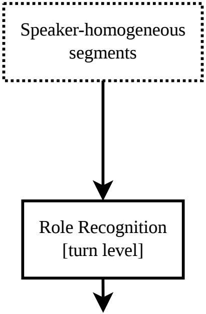
Outline

- Extracting Speaker Roles and alleviating error propagation
  - Effective speaker clustering for role recognition
  - Effective speech recognition for role recognition

- Using Speaker Roles to answer "*who spoke when*"
  - Use roles to reduce speaker clustering to speaker classification
  - Use roles to impose constraints on speaker clustering

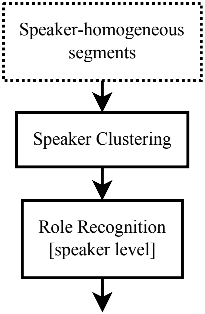
Speaker role recognition: Turn-level vs. Speaker-level

Turn-level SRR



- each turn classified independently

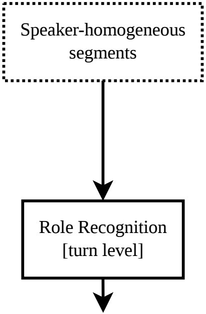
Speaker-level SRR



- a role is assigned to each same-speaker cluster

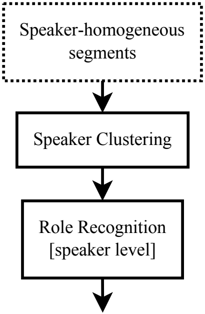
Speaker role recognition: Turn-level vs. Speaker-level

Turn-level SRR



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

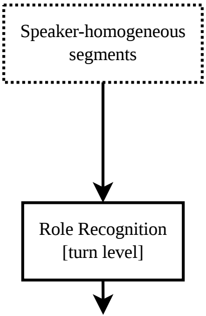
Speaker-level SRR



- a role is assigned to each same-speaker cluster

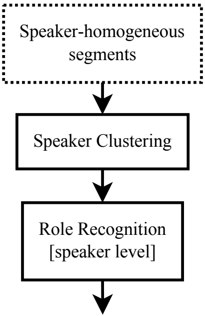
Speaker role recognition: Turn-level vs. Speaker-level

Turn-level SRR



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

Speaker-level SRR

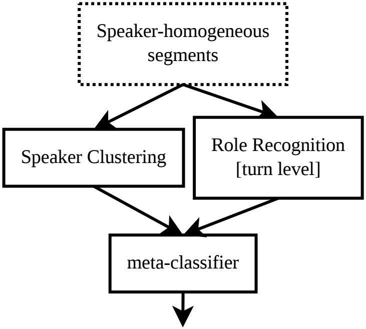


- a role is assigned to each same-speaker cluster
- error propagation between the modules

Solution?

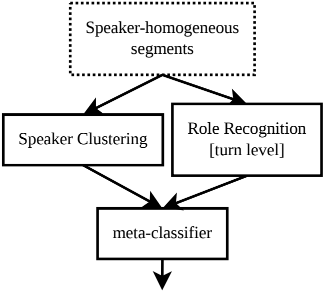
Can we effectively combine speaker-specific and role-specific information towards better SRR performance?

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

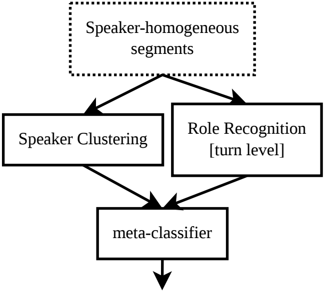


- assumption: one-to-one correspondence between speakers and roles
- each segment is represented by $2N$ scores ($N = \#\text{participants}$)
  - N scores from the speaker clustering module
  - N scores from the role recognition module
- those are fed to a meta-classifier

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



- Speaker Clustering:
  - BIC-based hierarchical clustering, with one Gaussian modeling each cluster
  - scores: log-likelihoods wrt each Gaussian
- Role Recognition:
  - LM-based (3-gram models)
    - scores: negative log perplexities wrt each LM
  - AM-based (512-component GMMs)
    - scores: log-likelihoods wrt each AM
- meta-classifier: linear SVM



- Speaker Clustering:
  - BIC-based hierarchical clustering, with one Gaussian modeling each cluster
  - scores: log-likelihoods wrt each Gaussian
- Role Recognition:
  - LM-based (3-gram models)
    - scores: negative log perplexities wrt each LM
  - AM-based (512-component GMMs)
    - scores: log-likelihoods wrt each AM
- meta-classifier: linear SVM

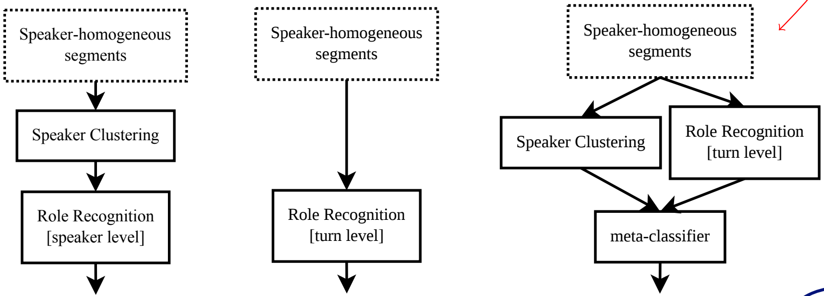
- Dyadic interactions from the psychology domain
  - *MI corpus*: Motivational Interviewing sessions between Therapist (73.7h) and Client (78.8h)
  - *ADOS corpus*: Autism Diagnostic Observation Schedule assessments between Psychologist (5.2h) and Child (5.6h)

Results: Misclassification Rates

$\mathcal{R}^\dagger$: 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**

Misclassification Rates (%)—lower is better.

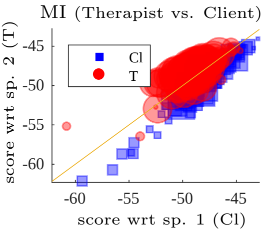


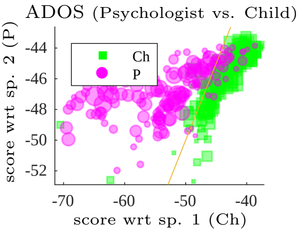
Results: Misclassification Rates

$\mathcal{R}^{\dagger}$: 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**

Misclassification Rates (%)—lower is better.



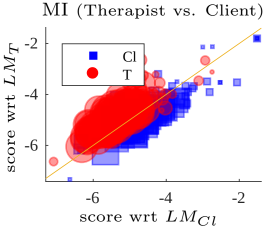


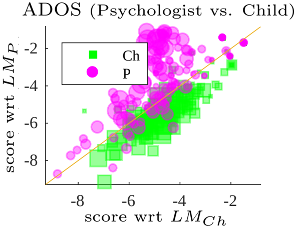
Results: Misclassification Rates

$\mathcal{R}^{\dagger}$: 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**

Misclassification Rates (%)—lower is better.



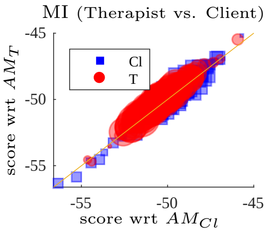


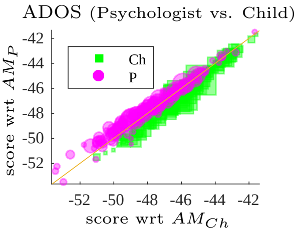
Results: Misclassification Rates

$\mathcal{R}^{\dagger}$: 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**

Misclassification Rates (%)—lower is better.





Results: Misclassification Rates

$\mathcal{R}^\dagger$: 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**

Misclassification Rates (%)—lower is better.

Final relative improvement wrt piped architecture:

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

Outline

- Extracting Speaker Roles and alleviating error propagation
  - Effective speaker clustering for role recognition
  - Effective speech recognition for role recognition
- Using Speaker Roles to answer "*who spoke when*"
  - Use roles to reduce speaker clustering to speaker classification
  - Use roles to impose constraints on speaker clustering

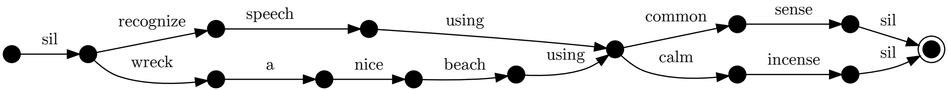
- Language patterns provide valuable cues for the task of speaker role recognition.
- But where do we find the lexical information?

Real-world text data

- Language patterns provide valuable cues for the task of speaker role recognition.
- But where do we find the lexical information?
  Automatic Speech Recognition (ASR)

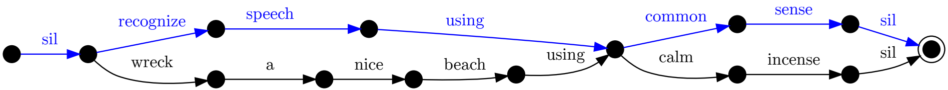
- Language patterns provide valuable cues for the task of speaker role recognition.
- But where do we find the lexical information?
  Automatic Speech Recognition (ASR)

- Given a speech utterance, ASR generates a word lattice...



- Language patterns provide valuable cues for the task of speaker role recognition.
- But where do we find the lexical information?
  Automatic Speech Recognition (ASR)

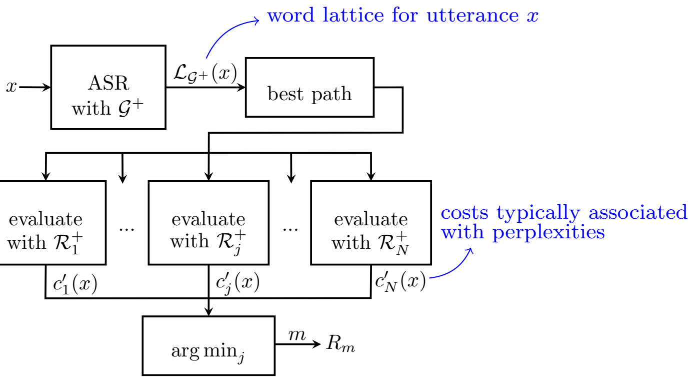
- Given a speech utterance, ASR generates a word lattice...
  ...where we find the most probable path



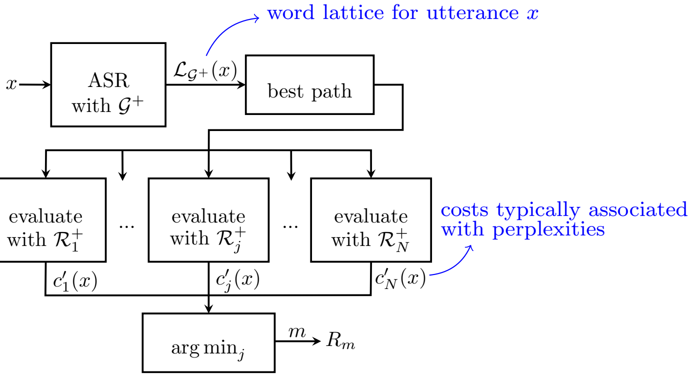
⇒ potential role-specific information loss

Text-based SRR: typical approach

- build background, generic LM $\mathcal{G}^+$
- and role-specific LMs $\mathcal{R}_1^+, \mathcal{R}_2^+, \cdots, \mathcal{R}_N^+$
- evaluate text data wrt all role-specific LMs

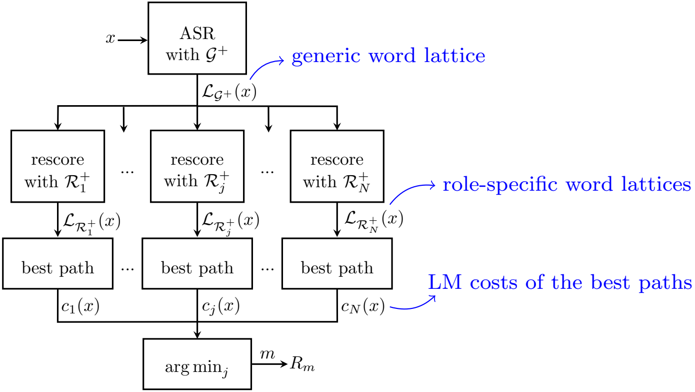


- build background, generic LM $\mathcal{G}^+$
- and role-specific LMs $\mathcal{R}_1^+, \mathcal{R}_2^+, \cdots, \mathcal{R}_N^+$
- evaluate text data wrt all role-specific LMs

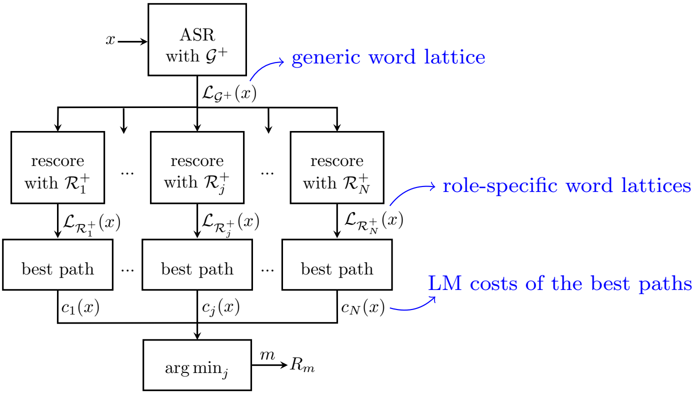


- Do we prune the lattice too early?

Text-based SRR: proposed approach



N. Flemotomos, P. Georgiou, D.C. Atkins & S. Narayanan. "Role Specific Lattice Rescoring for Speaker Role Recognition from Speech Recognition Outputs". ICASSP (2019)



Extension for speaker-level SRR

- apply speaker clustering
- define costs $c(S_i|R_j) \triangleq \sum_{x \in T_i} c_j(x)$ set of turns corresponding to speaker S_i
- assign role yielding minimum cost

N. Flemotomos, P. Georgiou, D.C. Atkins & S. Narayanan. "Role Specific Lattice Rescoring for Speaker Role Recognition from Speech Recognition Outputs". ICASSP (2019)

Results: Misclassification Rates

- PSYCH: dyadic interactions in psychotherapy
  Therapist (49.0h) vs. Client (43.0h)
- AMI: business meetings
  Project Manager (22.9h), Marketing Expert (15.3h),
  User Interface Designer (13.8h), Industrial Designer (15.2h)

Results: Misclassification Rates

- PSYCH: dyadic interactions in psychotherapy
  Therapist (49.0h) vs. Client (43.0h)
- AMI: business meetings
  Project Manager (22.9h), Marketing Expert (15.3h),
  User Interface Designer (13.8h), Industrial Designer (15.2h)

after BIC-based hierarchical clustering

	majority class	Turn-level SRR		Speaker-level SRR	
		rescoring	no rescoring	rescoring	no rescoring
PSYCH	50.67	23.58	10.75	**4.41**	5.83
AMI	62.22	64.70	63.40	**46.16**	60.94

Misclassification Rates (%)—lower is better.

- PSYCH: dyadic interactions in psychotherapy
  Therapist (49.0h) vs. Client (43.0h)
- AMI: business meetings
  Project Manager (22.9h), Marketing Expert (15.3h),
  User Interface Designer (13.8h), Industrial Designer (15.2h)

after BIC-based hierarchical clustering

	majority class	Turn-level SRR		Speaker-level SRR	
		rescoring	no rescoring	rescoring	no rescoring
PSYCH	50.67	23.58	10.75	**4.41**	5.83
AMI	62.22	64.70	63.40	**46.16**	60.94

Misclassification Rates (%)—lower is better.

- prior to speaker clustering, utterances are broken into very short speech segments
- each individual segment contains insufficient observations to infer speaker role

Results: Misclassification Rates

- PSYCH: dyadic interactions in psychotherapy
  Therapist (49.0h) vs. Client (43.0h)
- AMI: business meetings
  Project Manager (22.9h), Marketing Expert (15.3h),
  User Interface Designer (13.8h), Industrial Designer (15.2h)

after BIC-based hierarchical clustering

	majority class	Turn-level SRR		Speaker-level SRR	
		rescoring	no rescoring	rescoring	no rescoring
PSYCH	50.67	23.58	10.75	**4.41**	5.83
AMI	62.22	64.70	63.40	**46.16**	60.94

Misclassification Rates (%)—lower is better.

Relative improvement after clustering with LM rescoring:

- 24.4% for the PSYCH corpus
- 24.3% for the AMI corpus

Summary

- short speech segments contain insufficient observations to infer speaker role
  $\Rightarrow$ speaker-level SRR

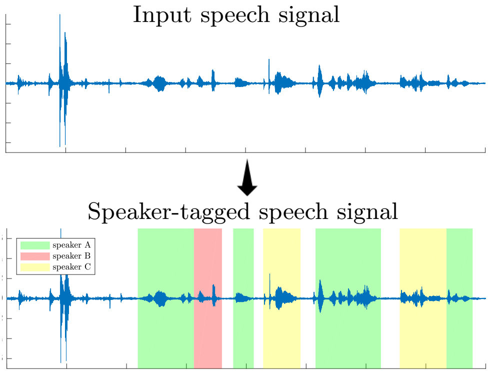
- techniques to alleviate the problem of error propagation
  - *from speaker clustering*: incorporate speaker-specific and role-specific information into a meta-classifier
  - *from ASR*: rescore the lattices with role-specific LMs

- improved SRR results for dyadic and multi-party interactions

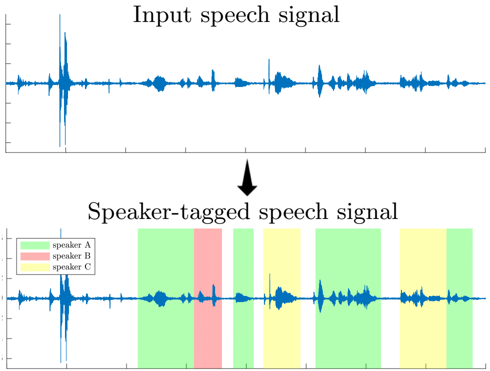
Outline

- Extracting Speaker Roles and alleviating error propagation
    - Effective speaker clustering for role recognition
    - Effective speech recognition for role recognition
- Using Speaker Roles to answer "*who spoke when*"
    - Use roles to reduce speaker clustering to speaker classification
    - Use roles to impose constraints on speaker clustering

Who Spoke When: Speaker Diarization

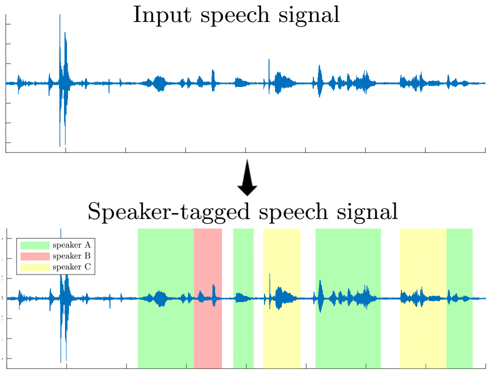


Who Spoke When: Speaker Diarization



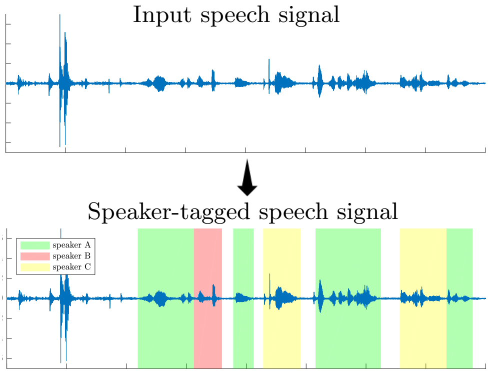
Why?

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



Traditional approach

1. segmentation
2. clustering

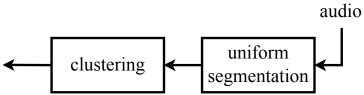


Traditional approach

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

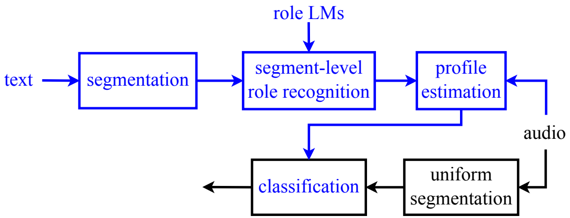
- different *roles* ⇒ distinguishable linguistic patterns
  ⇒ Can we use language to assist diarization?

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



traditional, audio-only system

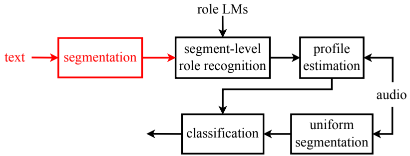
- different *roles* ⇒ distinguishable linguistic patterns
  ⇒ Can we use language to assist diarization?



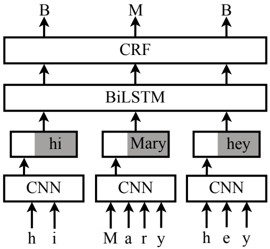
proposed, linguistically-aided system

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

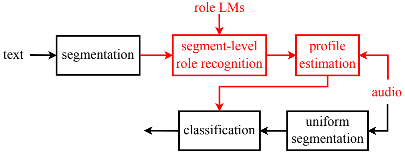
N. Flemotomos, P. Georgiou & S. Narayanan. "Linguistically Aided Speaker Diarization Using Speaker Role Information", Odyssey (2020)



- 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

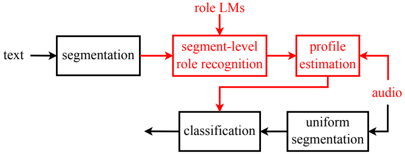


Role recognition and profile estimation



- Perform turn-level text-based SRR.
  - Assign to each text segment x the role R_i that minimizes the corresponding cost (perplexity) $pp(x|\mathcal{R}_i)$
- Extract an acoustic speaker embedding u_x $\forall$ audio-aligned segment x assigned the role R_i.
- Define the role profile r_i as the mean of all the $u_x : x \in R_i$.

Role recognition and profile estimation

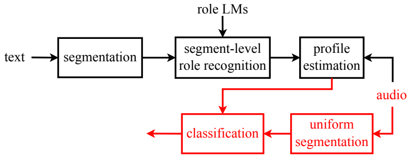


- Perform turn-level text-based SRR.
  - Assign to each text segment x the role R_i that minimizes the corresponding cost (perplexity) $pp(x|\mathcal{R}_i)$
- Extract an acoustic speaker embedding u_x $\forall$ audio-aligned segment x assigned the role R_i.
- Define the role profile r_i as the mean of all the $u_x : x \in R_i$.

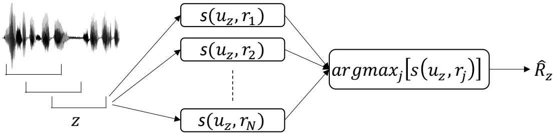
- *Are we confident about all the role assignments?*
  - Take into account only the segments about which we are confident enough:

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

Audio segmentation and classification



- Segment uniformly the speech signal (sliding window).
- Extract an acoustic speaker embedding 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 $s(u_z, r_i)$.



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**

DER (%)—lower is better—on PSYCH corpus (therapist vs. client).

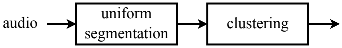
incorporates three sources of error:
missed speech, false alarm speech, speaker confusion

Results: Diarization Error Rate

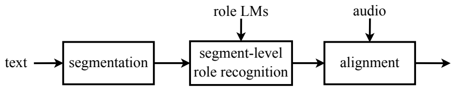
transcript source	text segmentation	audio only	language only	linguistically aided (all segments)	linguistically aided (best $a\%$ segments)
reference	oracle	11.05	12.99	7.28	**6.99**
	tagger		20.09	7.71	**7.30**
ASR	tagger	11.05	27.07	8.37	**7.84**

DER (%)—lower is better—on PSYCH corpus (therapist vs. client).

- unimodal baselines:
  audio stream contains more valuable information



audio-only baseline



language-only baseline

Results: Diarization Error Rate

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**

DER (%)—lower is better—on PSYCH corpus (therapist vs. client).

- 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: Diarization Error Rate

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**

DER (%)—lower is better—on PSYCH corpus (therapist vs. client).

- 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

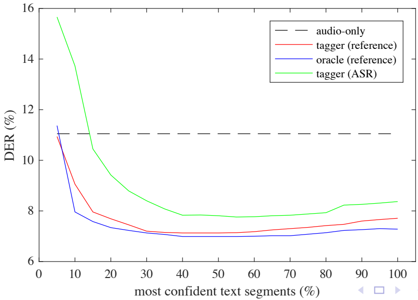
Results: Diarization Error Rate

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**

DER (%)—lower is better—on PSYCH corpus (therapist vs. client).

- 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

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



- Used lexical information to estimate acoustic speaker profiles and follow a classification approach instead of clustering for speaker diarization.

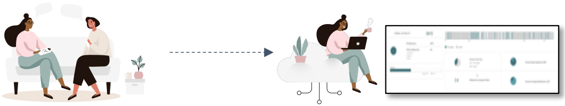
- Showed improved results in terms of DER.

Summary

- Used lexical information to estimate acoustic speaker profiles and follow a classification approach instead of clustering for speaker diarization.

- Showed improved results in terms of DER.

- Real-world downstream application: quality assessment of psychotherapy

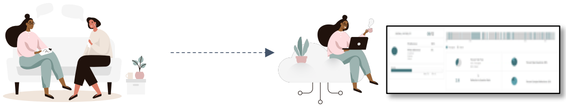


Summary

- Used lexical information to estimate acoustic speaker profiles and follow a classification approach instead of clustering for speaker diarization.

- Showed improved results in terms of DER.

- Real-world downstream application: quality assessment of psychotherapy



- Required assumption: one-to-one correspondence between speakers and roles (e.g., one therapist vs. one patient per session).

- Extracting Speaker Roles and alleviating error propagation
  - Effective speaker clustering for role recognition
  - Effective speech recognition for role recognition
- Using Speaker Roles to answer "*who spoke when*"
  - Use roles to reduce speaker clustering to speaker classification
  - Use roles to impose constraints on speaker clustering

- every speaker mapped to a distinct role
  e.g., one doctor vs. one patient

images from freepik.com by pch.vector

A Variety of Role-Playing Scenarios

- every speaker mapped to a distinct role
  e.g., one doctor vs. one patient

- many speakers assume the same role
  e.g., one judge and multiple prosecution witnesses

- many roles are played by the same speaker
  e.g., host, interviewer, and guest, where the interviewer may be the same person as the host

images from freepik.com by pch.vector and jcomp

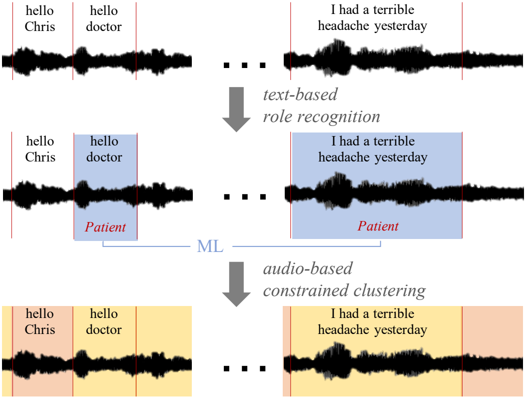
Use Roles to Impose Constraints

- extract role information to impose constraints during audio-based clustering
- focus on segment-level pairwise constraints:
  Must-Link (ML) and Cannot-Link (CL)

N. Flemotomos & S. Narayanan, "Multimodal Clustering with Role Induced Constraints for Speaker Diarization". under review (2022)

Use Roles to Impose Constraints

- extract role information to impose constraints during audio-based clustering
- focus on segment-level pairwise constraints: Must-Link (ML) and Cannot-Link (CL)



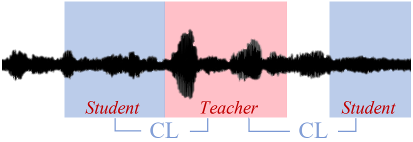
N. Flemotomos & S. Narayanan, "Multimodal Clustering with Role Induced Constraints for Speaker Diarization". under review (2022)

Use Roles to Impose Constraints

- extract role information to impose constraints during audio-based clustering
- focus on segment-level pairwise constraints:
  Must-Link (ML) and Cannot-Link (CL)

Some possible scenarios and strategies:

- different roles are played by different speakers
  e.g., teacher vs. students during lecture
  ⇒ CL constraints between segments with different roles

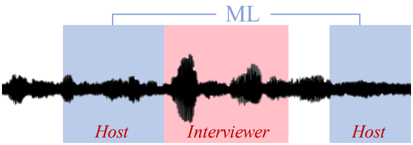


N. Flemotomos & S. Narayanan, "Multimodal Clustering with Role Induced Constraints for Speaker Diarization". under review (2022)

- extract role information to impose constraints during audio-based clustering
- focus on segment-level pairwise constraints:
  Must-Link (ML) and Cannot-Link (CL)

Some possible scenarios and strategies:

- different speakers play different roles
  e.g., host vs. interviewer vs. guest during TV show
  $\Rightarrow$ ML constraints between segments with same roles

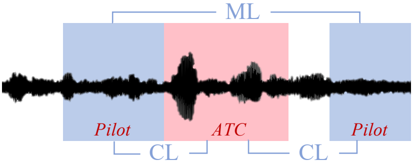


N. Flemotomos & S. Narayanan, "Multimodal Clustering with Role Induced Constraints for Speaker Diarization". under review (2022)

- extract role information to impose constraints during audio-based clustering
- focus on segment-level pairwise constraints:
  Must-Link (ML) and Cannot-Link (CL)

Some possible scenarios and strategies:

- one-to-one correspondence between speakers and roles
  e.g., pilot vs. air traffic controller during flight
  ⇒ both ML and CL constraints



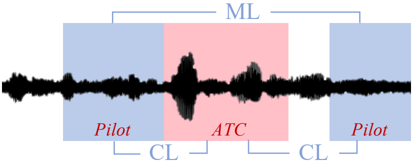
N. Flemotomos & S. Narayanan, "Multimodal Clustering with Role Induced Constraints for Speaker Diarization". under review (2022)

Use Roles to Impose Constraints

- extract role information to impose constraints during audio-based clustering
- focus on segment-level pairwise constraints:
  Must-Link (ML) and Cannot-Link (CL)

Some possible scenarios and strategies:

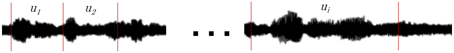
- one-to-one correspondence between speakers and roles
  e.g., pilot vs. air traffic controller during flight
  ⇒ both ML and CL constraints



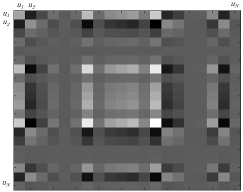
- adopt framework of constrained spectral clustering

N. Flemotomos & S. Narayanan, "Multimodal Clustering with Role Induced Constraints for Speaker Diarization". under review (2022)

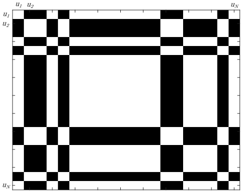
0 speaker-homogeneous segments



1 cosine-based affinity matrix $\hat{\mathbf{W}}$



2 thresholding & symmetrization ($\mathbf{W}$)



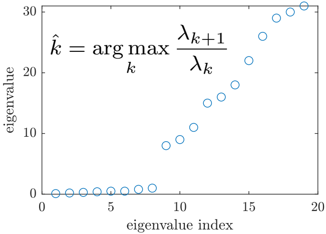
3 normalized Laplacian

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$$

$$\mathbf{D} = \mathrm{diag}\{d_1, d_2, \cdots, d_N\}$$

$$d_i = \sum_j \mathbf{W}_{ij}$$

4 maximum eigen-gap on $\mathbf{L}$



5 $\hat{k}$-means on eigenvectors of $\mathbf{L}$

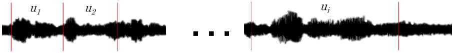
$$\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|\cdots|\mathbf{x}_{\hat{k}}]$$

corresponding to the $\hat{k}$ smallest eigenvalues

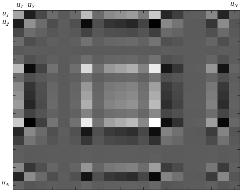
*Eigenvalues are only given for visualization purposes; they do not correspond to $\mathbf{W}$.

Constrained Spectral Clustering

0. speaker-homogeneous segments



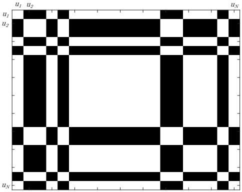
1. cosine-based affinity matrix $\hat{\mathbf{W}}$



Constrained Clustering

- increase similarity between ML-constrained pairs
- decrease similarity between CL-constrained pairs

2. thresholding & symmetrization ($\mathbf{W}$)



Integrate initial set of constraints through the Exhaustive and Efficient Constraint Propagation (E^{2}CP) algorithm:

1. construct constraint matrix $\mathbf{Z}$

$$\mathbf{Z}_{ij} = \begin{cases} +1, & \text{if } \exists \text{ ML constraint between } i \text{ and } j \\ -1, & \text{if } \exists \text{ CL constraint between } i \text{ and } j \\ 0, & \text{if } \nexists \text{ any constraint between } i \text{ and } j \end{cases}$$

2. propagate constraints to the entire session

$$\mathbf{Z}^* = (1-\alpha)^2(\mathbf{I}-\alpha\bar{\mathbf{L}})^{-1}\mathbf{Z}(\mathbf{I}-\alpha\bar{\mathbf{L}})^{-1}, \quad \bar{\mathbf{L}} = \bar{\mathbf{D}}^{-1/2}\hat{\mathbf{W}}\bar{\mathbf{D}}^{-1/2}, \quad \alpha \in [0,1]$$

α: how much to change the constraints
vs. how much to change the affinity scores

$\alpha = 0 \Rightarrow \mathbf{Z}^* = \mathbf{Z} \Rightarrow$ only rely on the initial constraints
$\alpha = 1 \Rightarrow \mathbf{Z}^* = \mathbf{0} \Rightarrow$ ignore the constraints

3. update affinity scores

$$\hat{\mathbf{W}}_{ij} \leftarrow \begin{cases} 1-(1-\mathbf{Z}^*_{ij})(1-\hat{\mathbf{W}}_{ij}), & \text{if } \mathbf{Z}^*_{ij} \geq 0 \text{ (move closer to 1)} \\ (1+\mathbf{Z}^*_{ij})\hat{\mathbf{W}}_{ij}, & \text{if } \mathbf{Z}^*_{ij} < 0 \text{ (move closer to 0)} \end{cases}$$

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

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

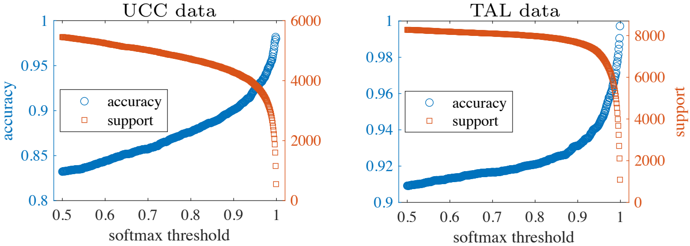
- multi-party conversations (18 speakers on average)
- partial role information
  single *host* vs. multiple *non-hosts* per episode
- apply CL constraints between segments with different roles
- total speaking time: host (118.6h) vs. non-host (519.2h)

images from freepik.com by pch.vector

Extracting Role Information

- Adapt a BERT model to classify the speaker roles

- But results are not perfect! What if we impose wrong constraints?
  - need a confidence proxy / threshold ⇒ use softmax values
  - trade-off decision: very confident or a lot of constraints??



Accuracy and support for the BERT-based classifier when only segments with softmax value above some threshold are taken into account.

- For experiments: constrain about 40% of the available segments

	unconstrained clustering	constrained clustering	role-based classification
	audio-only	cross-modal	language-only
UCC	1.38	**1.31**	10.34
TAL	42.22	**23.86**	63.01

Diarization Error Rate (%)—lower is better.

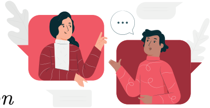
- experiments with manual segmentation and manual transcription
  - only evaluate clustering performance

- slight improvement for the dyadic UCC dataset
- substantial improvement for the multi-party TAL dataset
  - constraints helped estimate number of speakers (clusters) per episode

Summary

- Proposed a cross-modal framework to impose language-based role constraints during audio-based clustering.
    - does not need one-to-one mapping between speakers and roles
- Improved diarization results for both dyadic and multi-party role-playing interactions.

- Proposed a cross-modal framework to impose language-based role constraints during audio-based clustering.
    - does not need one-to-one mapping between speakers and roles
- Improved diarization results for both dyadic and multi-party role-playing interactions.
- What about other modalities?
    - audio- or video-based constraints
- Can we incorporate soft constraints?
    - confidence scores
    - role-based conversational dynamics



- end-to-end role-aware transcription
    - *integrated diarization, speech, and role recognition*

- analysis of informal and time-varying roles
    - *emergent roles due to social dynamics*

- intersectional analysis of speaker roles
    - *roles are just one aspect of a speaker's identity*

N. Flemotomos, Z. Chen, D.C. Atkins & Shrikanth Narayanan "Role Annotated Speech Recognition for Conversational Interactions". SLT (2018)

images from freepik.com by pch.vector and storyset

Relevant Publications

1. N. Flemotomos & S. Narayanan. “Multimodal Clustering with Role Induced Constraints for Speaker Diarization”. *under review* (2022)
2. N. Flemotomos, P. Georgiou & S. Narayanan. “Linguistically Aided Speaker Diarization Using Speaker Role Information”. *Odyssey* (2020)
3. N. Flemotomos, P. Georgiou, D.C. Atkins & S. Narayanan. “Role Specific Lattice Rescoring for Speaker Role Recognition from Speech Recognition Outputs”. *ICASSP* (2019)
4. N. Flemotomos, Z. Chen, D.C. Atkins & S. Narayanan. “Role Annotated Speech Recognition for Conversational Interactions”. *SLT* (2018)
5. N. Flemotomos, P. Papadopoulos, J. Gibson & S. Narayanan. “Combined Speaker Clustering and Role Recognition in Conversational Speech”. *Interspeech* (2018)

6. C.S. Soma, B. Wampold, N. Flemotomos, R. Peri, S. Narayanan, D.C. Atkins & Z.E. Imel. "The Silent Treatment?: Changes in patient emotional expression after silence". *Counseling and Psychotherapy Research* (2022)
7. Z. Chen, N. Flemotomos, K. Singla, T.A. Creed, D.C. Atkins & S. Narayanan. "An Automated Quality Evaluation Framework of Psychotherapy Conversations with Local Quality Estimates". *Computer Speech & Language* (2022)
8. N. Flemotomos, V.R. Martinez, Z. Chen, T.A. Creed, D.C. Atkins & S. Narayanan. "Automated Quality Assessment of Cognitive Behavioral Therapy Sessions Through Highly Contextualized Language Representations". *PLOS ONE* (2021)
9. N. Flemotomos, V.R. Martinez, Z. Chen, K. Singla, V. Ardulov, R. Peri, J. Gibson, M.J. Tanana, P. Georgiou, J. Van Epps, S.P. Lord, T. Hirsch, Z.E. Imel, D.C. Atkins & S. Narayanan. "Automated Evaluation of Psychotherapy Skills Using Speech and Language Technologies". *Behavior Research Methods* (2021)
10. Z. Chen, N. Flemotomos, V. Ardulov, T.A. Creed, Z.E. Imel, D.C. Atkins & S. Narayanan. "Feature Fusion Strategies for End-to-End Evaluation of Cognitive Behavior Therapy Sessions". *EMBC* (2021)
11. S.B. Goldberg, N. Flemotomos, V.R. Martinez, M. Tanana, P. Kuo, B.T. Pace, J.L. Villatte, P. Georgiou, J. Van Epps, Z.E. Imel, S. Narayanan & D.C. Atkins. "Machine Learning and Natural Language Processing in Psychotherapy Research: Alliance as Example Use Case". *Journal of Counseling Psychology* (2020)
12. N. Flemotomos & D. Dimitriadis. "A Memory Augmented Architecture for Continuous Speaker Identification in Meetings". *ICASSP* (2020)
13. T.J. Park, M. Kumar, N. Flemotomos, M. Pal, R. Peri, R. Lahiri, P. Georgiou & S. Narayanan. "The Second DIHARD Challenge: System Description for USC-SAIL Team". *Interspeech* (2019)
14. V.R. Martinez, N. Flemotomos, V. Ardulov, K. Somandepalli, S.B. Goldberg, Z.E. Imel, D.C. Atkins & S. Narayanan. "Identifying Therapist and Client Personae for Therapeutic Alliance Estimation". *Interspeech* (2019)
15. N. Flemotomos, V. Martinez, J. Gibson, D.C. Atkins, T. Creed & S. Narayanan. "Language Features for Automated Evaluation of Cognitive Behavior Psychotherapy Sessions". *Interspeech* (2018)
16. K. Singla, Z. Chen, N. Flemotomos, J. Gibson, D. Can, D.C. Atkins & S. Narayanan. "Using Prosodic and Lexical Information for Learning Utterance-level Behaviors in Psychotherapy". *Interspeech* (2018)

Thank you!

Questions and Discussion