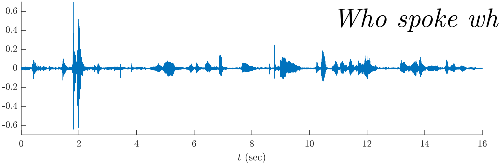
# Linguistically Aided Speaker Diarization Using Speaker Role Information

Nikolaos Flemotomos, Panayiotis Georgiou, Shrikanth Narayanan

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

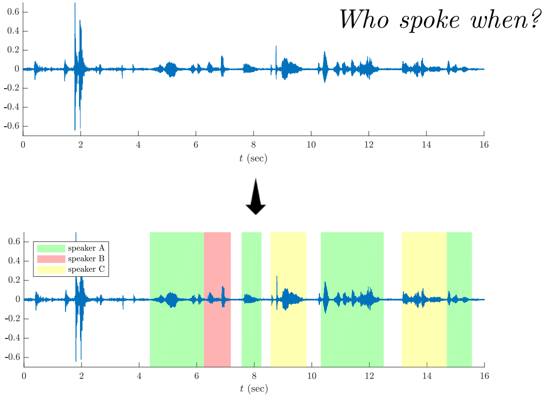
Odyssey 2020
The Speaker and Language Recognition Workshop

# Speaker Diarization

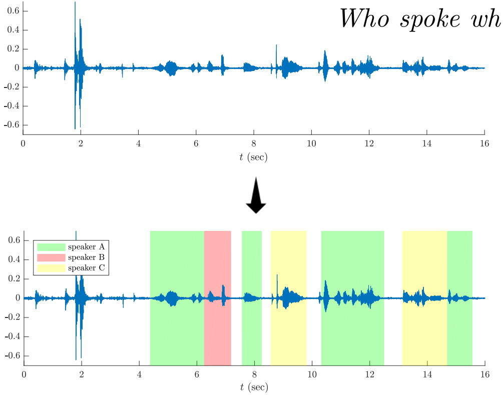


*Who spoke when?*

# Speaker Diarization


*Who spoke when?*

# Speaker Diarization

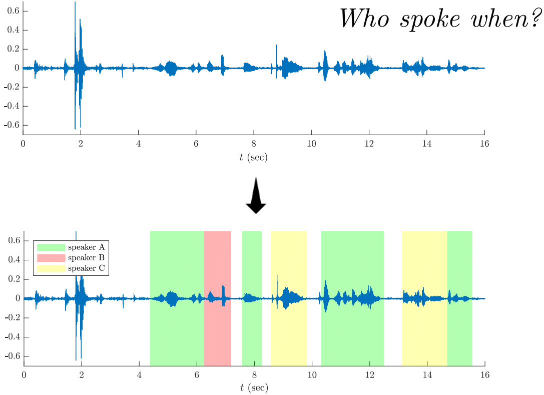


*Who spoke when?*

## Traditional approach

1. segmentation
2. clustering

# Speaker Diarization



*Who spoke when?*

## Traditional approach

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

# Structured Scenario: speakers assume *roles*

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

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

# Structured Scenario: speakers assume *roles*

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

- different *roles* ⇒ distinguishable linguistic patterns

  ⇒ Can we use language to assist diarization?

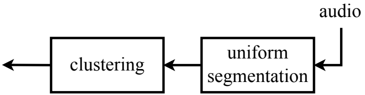
images from the Noun Project. creators: Nubaia Karim Barsha, Gan Khun Lay, Arafat Uddin, Llisole, ProSymbols

# Proposed System

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

# Proposed System

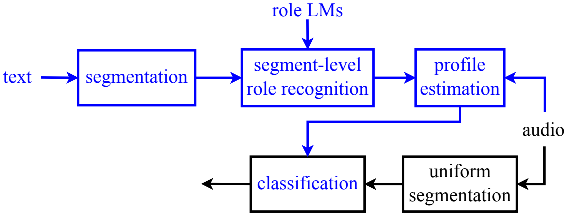
- different *roles* ⇒ distinguishable linguistic patterns
  ⇒ Can we use language to assist diarization?



traditional, audio-only system

# Proposed System

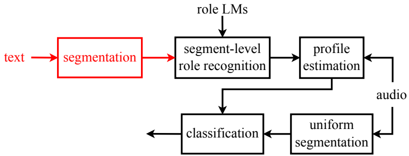
- different *roles* ⇒ distinguishable linguistic patterns
  ⇒ Can we use language to assist diarization?



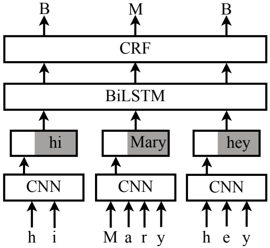
proposed, linguistically-aided system

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

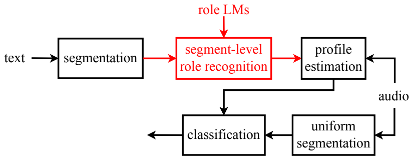
# Proposed System: Text-based segmentation



- 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



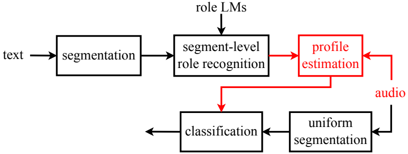
- Build a background LM $\mathcal{G}$ and $N$ role-specific LMs $\mathcal{R}_i$ ($N$ roles).
- Interpolate the LMs (n-gram):

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

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

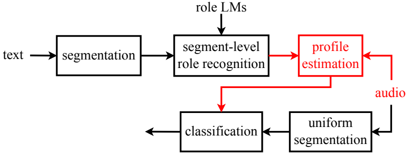
- Assign to each text segment $x$ the role $i$ that minimizes the perplexity $pp(x|\mathcal{R}_i^+)$.

# 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 $u_x : x \in R_i$.

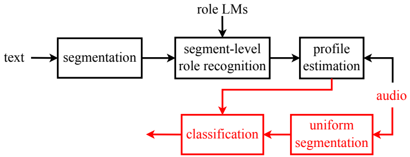
- *Are we confident about all the role assignments?*
  - Assign a confidence metric to each $x$:

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

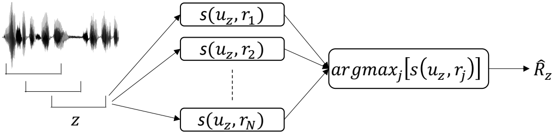
  - Take into account only the segments about which we are confident enough:

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

# Proposed System: Audio segmentation and classification



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



# 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.

# Setup and Baselines

### 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

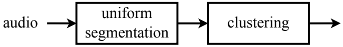
### ASR

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

### decoding & evaluation

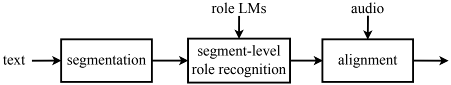
- initial oracle silence-based segmentation (1 sec threshold)
- 0.25 sec collar (metric: DER)
- ignore overlapping speech

## Baselines



audio → uniform segmentation → clustering →

audio-only baseline



text → segmentation → segment-level role recognition → alignment →

role LMs

audio

language-only baseline

# Results

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

Table: DER (%) on PSYCH-test.

# Results

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

Table: DER (%) on PSYCH-test.

- unimodal baselines:
  audio stream contains more valuable information

# Results

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

Table: DER (%) on PSYCH-test.

- tagger oversegments
  ⇒ 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 | 11.05 | 12.99 | 7.28 | **6.99** |
| | tagger | | 20.09 | 7.71 | **7.30** |
| ASR | tagger | 11.05 | 27.07 | 8.37 | **7.84** |

Table: DER (%) on PSYCH-test.

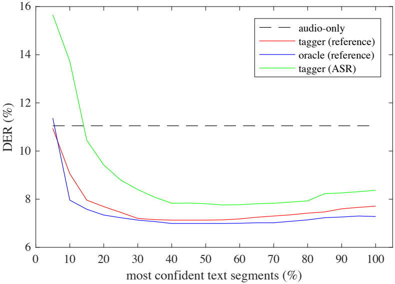
- high WER ⇒ severe degradation for language-only system
- when transcripts are only used for profile estimation (linguistically-aided) the performance gap is much smaller

# Results

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

Table: DER (%) on PSYCH-test.

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



# Conclusions

- 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.

National Institutes of Health

*Turning Discovery Into Health*