Imperial College London

8 November 2022

Studying Human-Based Speaker Diarization and Comparing to State-of-the-Art Systems

Simon W. McKnight
Speech and Audio Processing Lab
Communications and Signal Processing Group
Electrical and Electronic Engineering

Imperial College London

GT = ground truth
GT-SAD = GT speech activity detection
BUT = Brno University of Technology

Introduction

- Human-based speaker diarization experiments:
  - Experiment 1: no prior information – *13 reviewers* – baselines pyannote.audio V2 and V1
  - Experiment 2: start from ground truth speech activity detection (GT-SAD) – *10 reviewers* – baselines pyannote.audio V1, BUT BDII and BUT ResNet101
  - Experiment 3: start from ground truth blank labels (GT-labels) – *10 reviewers* – no baselines
- 5-minute extract of AMI 2008a meeting headset recordings
  - 4 speakers, 3 female and 1 male
  - significant overlapping speech (around 4.45 to 8.52% from GT)
  - reviewers used Audacity to segment (if relevant) and label
  - instructions for consistent application (e.g. 300 ms pauses)
- Effect of GT differences and forgiveness collars in scoring

Imperial College London

Speaker Diarization

- Distinguishing speakers and specifying times they speak in a speech recording or live player
- Often referred to as “who spoke when”
  - … but most diarization systems distinguish speakers but do not identify them
  - diarization challenges expect systems not to have heard speakers before
  - nonetheless, current top performing systems train on labelled data (e.g. VoxCeleb 1 and 2) for a discriminative model, then make generative
- Inaccurate and inconsistent labelling of speaker and speech boundaries is a big problem for both training and scoring
  - subjectivity in human ground truth labelling
  - splitting speech on pauses: AMI general v NIST 300 ms v DIHARD 200ms
  - scoring moving away from forgiveness collars and excluding overlapping speakers
  - use of validation/development sets helps to a degree

M = miss
FA = false alarm
SE = speaker error
UEM = unpartitioned evaluation map

Evaluating speaker diarization performance

- Standard time-based diarization error rate (DER) measure

$$DER = \frac{\tau_M + \tau_{FA} + \tau_{SE}}{\tau_{TOTAL}} = M_\tau + FA_\tau + SE_\tau$$

- Overlapping speakers included
- Generally exclude laughter/coughing etc, but some subjectivity
- Examples imprecise v precise GT labelling:

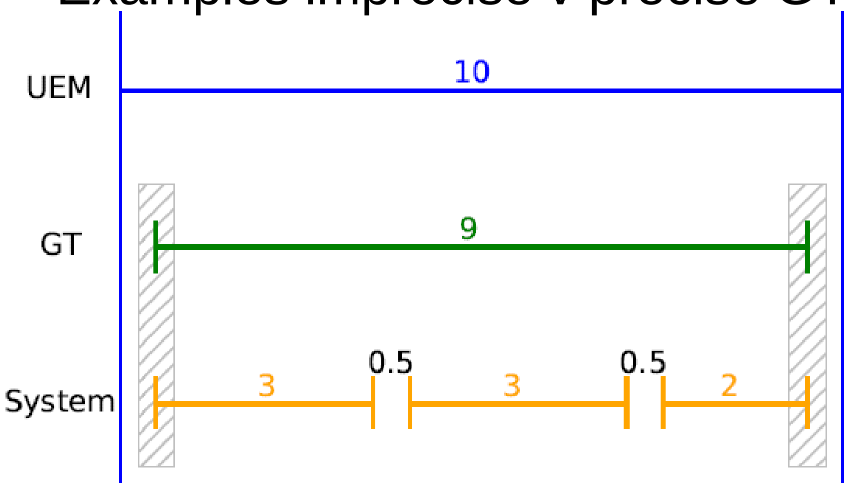


Fig. 1 – imprecise GT labelling (11.1% DER collar, 11.8% no collar)

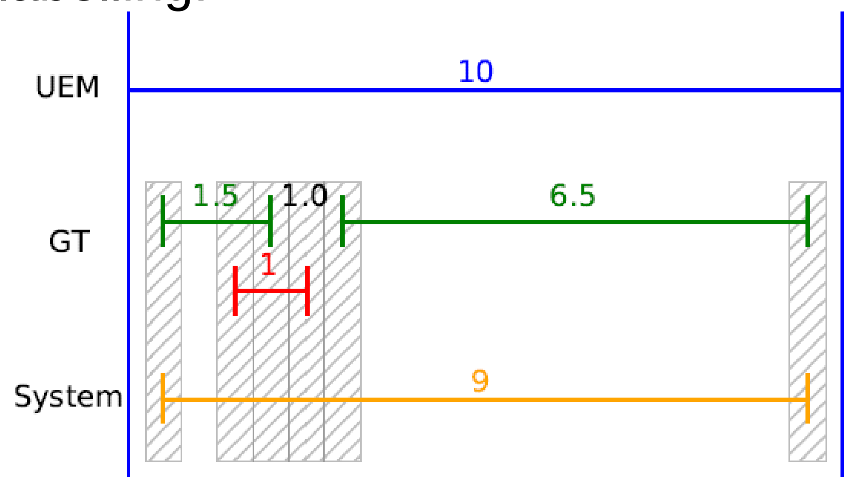


Fig. 2 – precise GT labelling with overlaps (0% DER collar, 16.7% no collar)

Imperial College London

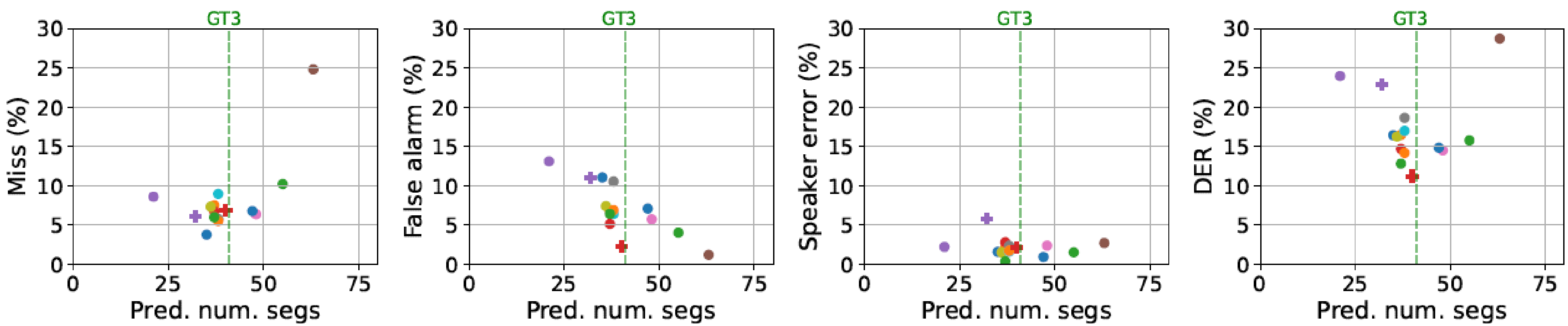
STD = standard deviation
exc. = excludes laughter/coughing
inc. = includes laughter/coughing

Experiment 1 Results

- Scores for human reviews very considerably with 2 outliers
- Sensitive to ground truth chosen – Table II DERs in %

	GT	250 ms Means	250 ms STDs	0 ms Means	0 ms STDs	
AMI	GT1	11.93	1.51	18.94	1.43	exc.
	GT2	11.02	1.46	17.20	1.45	inc.
BUT	GT3	8.95	1.60	15.60	1.53	exc.
	GT4	10.27	1.66	17.62	1.44	inc.

- Predicting same number of segments as ground truth used is biggest driver of good performance



Experiment 1 Results

- Forgiveness collars reduce DER means, but increase STDs
  - means down from 15.60% to 8.95% (±250 ms to 0 ms)
  - but STDs up from 1.53% to 1.60%
  - this would not be expected if differences were primarily due to insignificant timing differences around speaker boundaries
- pyannote.audio V2 (but not V1) outperforms humans on segmentation/ timings
  - was it just because it got closer to the right number of segments than all the human reviewers?
  - had been trained on AMI generally

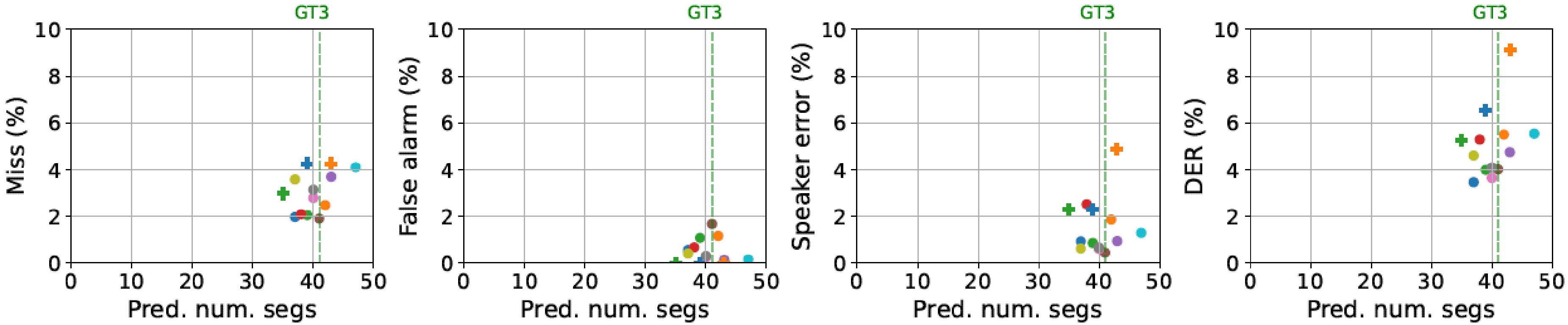
Imperial College London

Experiment 2 Results

- Much better results than for Experiment 1

GT	250 ms Means	250 ms STDs	0 ms Means	0 ms STDs
GT3	2.03	0.64	4.49	0.73

  - mean DERs improved 11.11% without collar, 6.92% with
- Misses reflect missed overlapping speakers
- 7 of 10 human reviews outperformed best baseline system
  - 9 of 10 in the speaker error component



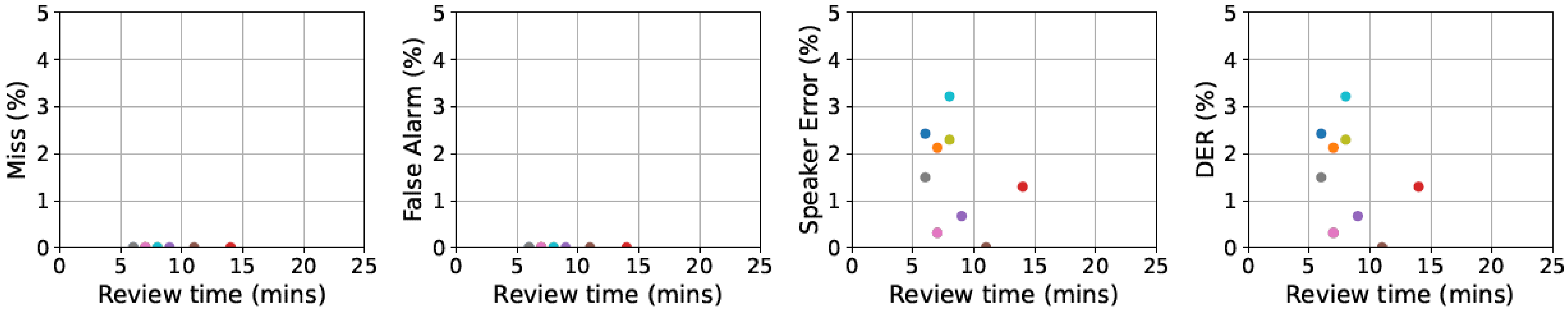
Imperial College London

Experiment 3 Results

- Scores dramatically better

GT	250 ms Means	250 ms STDs	0 ms Means	0 ms STDs
GT3	0.68	0.69	1.41	1.03

  - misses and false alarms naturally fall to zero
  - speaker errors improve, but still non-zero due to multiple overlapping speaker difficulties and inconsistent speaker pitch



Imperial College London

Reviewer Observations

- Recordings generally clear, but heavy breathing annoying
  - old style microphones in front of mouths
- Several reviewers noted the female speakers had similar pitch
  - used semantic information to distinguish them at times rather than vocal pitch or timbre
  - 2 reviewers who were non-native English speakers felt they were at a disadvantage compared to native English speakers
  - times when an existing female speaker interjected in a higher-pitched voice or showing more emotion were often incorrectly thought to have been a different speaker altogether
- All reviewers coped well with 2 overlapping speakers, but not 3
  - difficult because overlaps tended to be short
  - not all vocal sounds easy to classify as speech or not

Imperial College London

ASR = automatic speech recognition

Conclusions and Further Information

- Use of forgiveness collars not recommended in scoring
- Scoring sensitivity to ground truth means probably better off combining ASR with diarization and assigning word error rates scores based on correct speaker allocation
  - … though only an option if ASR involved, there are other uses of speaker diarization
- Humans struggle with timings, but still better at distinguishing speakers
- Instructions to reviewers and results at
  - https://github.com/swm1718/HumanReviews
  - https://tinyurl.com/4ys4ba7t