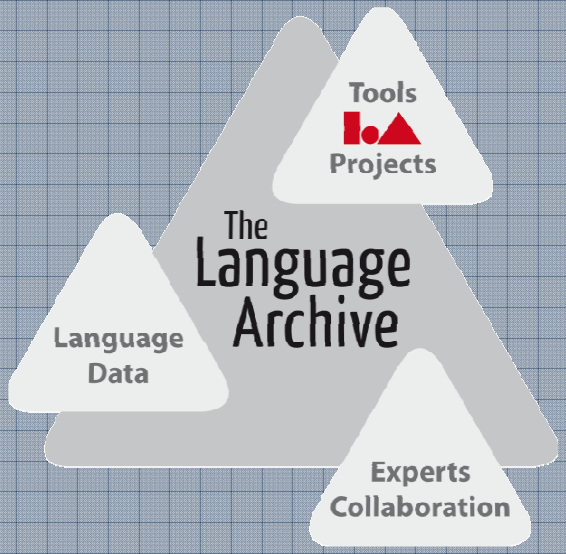


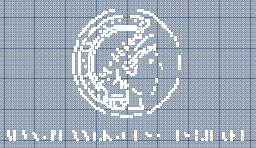
Speaker Diarization using Gesture

Binyam Gebrekidan Gebre
The Language Archive – Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

Speaker diarization: Definition

- Is the task of determining **who spoke when?** in an audio/video recording

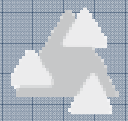
Speaker Diarization: Applications



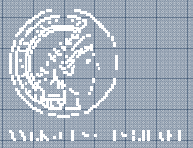
- Speech and speaker indexing
  (used for video navigation and retrieval)

- Speaker attributed speech-to-text transcription
  (used for speech translation and message summarization)

- Speaker model adaptation
  (used for enhancing speaker recognition)
- And so on

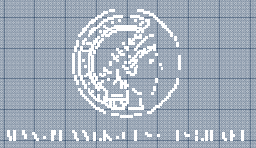


Speaker Diarization: Hypothesis



The gesturer is the speaker

Speaker Diarization: Evidence



- Gestures occur mainly during speech
- Delayed auditory feedback
- The inborn blind do gesture
- Fluency affects gesturing

Speaker Diarization: Algorithm

- Determine in the video

a) the number of speakers

b) their location and

c) whether or not they gestured

Speaker Diarization: Experiment

Table 1. Features of experiment videos

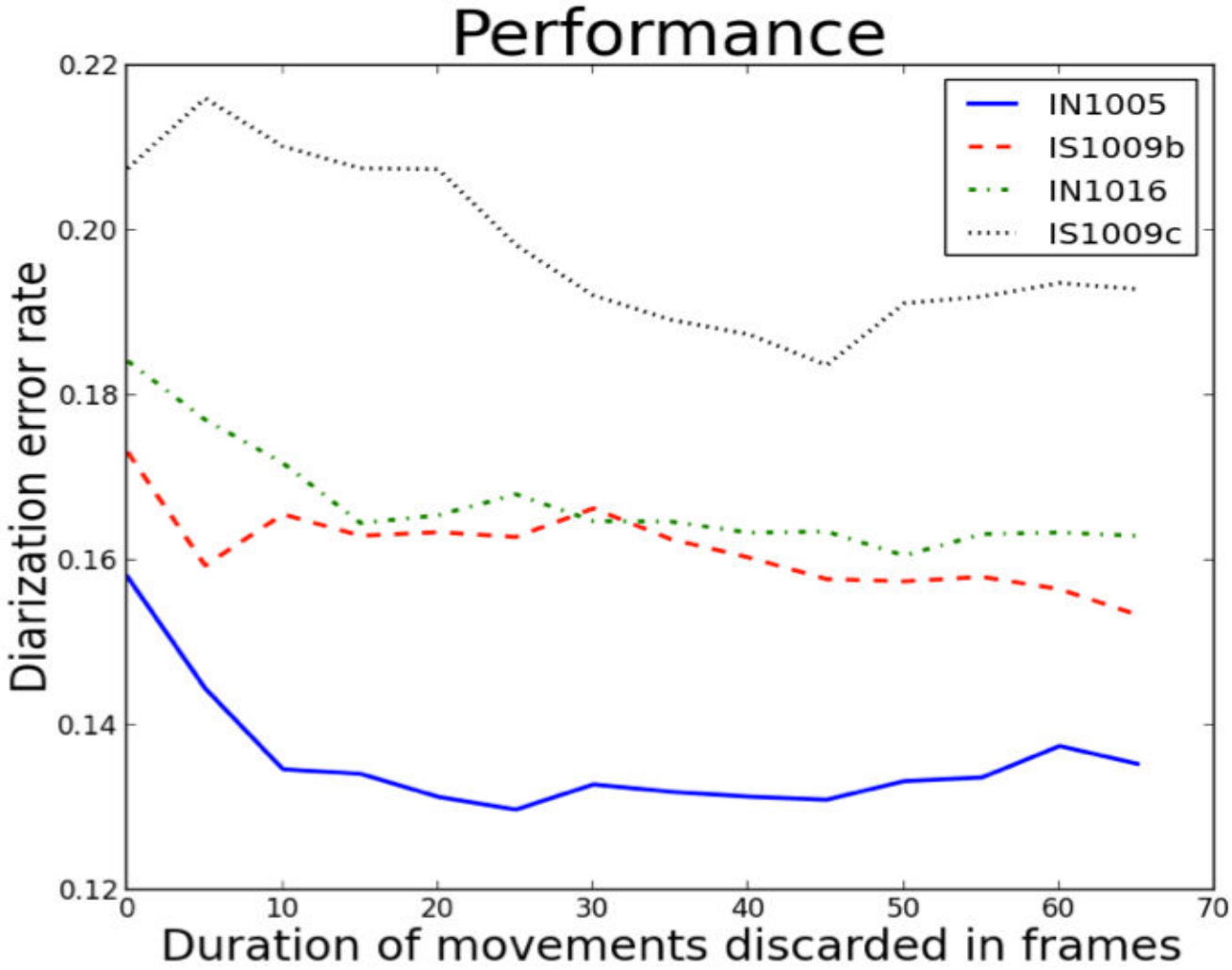
Name	Video length (min)	Speech time (%)	Speech overlap (%)	Turn switches (per min)
IN1005	46	94.90	9.53	7.35
IS1009b	34	87.88	8.97	6.48
IN1016	59	96.95	18.27	12.30
IS1009c	30	84.16	4.23	4.85
IN1002*	41	93.15	14.31	10.03
IN1009	20	89.67	12.61	4.57
IS1009a*	13	75.15	10.27	3.25
IS1009d*	32	80.83	8.58	8.45
IN1007*	40	96.46	22.57	9.43
IN1012	51	96.89	28.44	12.82
IN1014*	61	90.49	12.21	10.00
IN1008*	56	90.81	9.27	12.40
IN1013	51	96.04	26.64	12.88

Speaker Diarization: Results

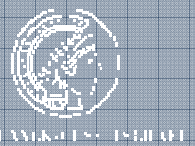
Table 2. Diarization Error Rates (DER)

Name	Gesture time (%)	Gesture overlap (%)	Turn switches (per min)	DER (%)
IN1005	62.54	0.03	1.07	13.53
IS1009b	72.23	0.00	0.78	15.33
IN1016	72.45	0.00	1.58	16.29
IS1009c	66.40	0.00	0.70	19.28
IN1002*	63.65	0.00	0.95	22.27
IN1009	59.50	0.00	0.67	22.47
IS1009a*	60.84	0.00	0.28	22.57
IS1009d*	68.82	0.00	0.58	23.49
IN1007*	67.06	0.04	1.37	23.86
IN1012	64.00	0.00	1.67	23.90
IN1014*	71.60	0.00	1.15	26.33
IN1008*	57.80	0.00	1.88	27.47
IN1013	69.47	0.01	1.42	28.04

Speaker Diarization: Results



Possible Extensions



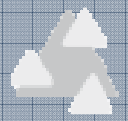
- Speaker Diarization using Gesture ***and Speech***

Conclusions

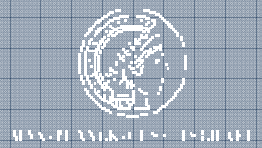
- Simple idea: the gesturer is the speaker
- Simple algorithm: use movement to approximate gesture
- Good performance: comparable with previous state-of-the-art performance

  The official NIST 2009:

  – For batch audio [17.24 -- 31.30%].
  – For online audio [39.27% -- 44.61%].
  – For audiovisual [16%, 32.56%]



Acknowledgments



- Idea
  - Peter Wittenburg
  - Tom Heskes
  - Przemyslaw Lenkiewicz
  - Anna Lenkiewicz
  - Han Sloetjes

Discussion

- Thank you for your attention
  - Questions?
  - Comments?