RT-07 Speaker Diarization Results

http://www.nist.gov/speech/tests/rt/rt2007

Jonathan Fiscus and Jerome Ajot

May 8-9, 2007

Rich Transcription 2007
Meeting Recognition Workshop

RT-07 Evaluation Participants

Site ID	Site Name	Evaluation Task		
		SPKR	STT	SASTT
AMIDA	Augmented Multi-party Interaction with Distance Access	6	4(*1)	4
I2R/NTU	Infocomm Research Site and Nanyang Technological University	4		
IBM	IBM	4	4(*1)	8
ICSI	International Computer Science Institute	2		
LIA	Laboratoire Informatique d'Avignon	16		
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur	4		1(1)
SRI/ICSI	SRI International and International Computer Science Institute		18(*12)	7(*7)
UKA	Karlsruhe University		2	
UPC	Universitat Politècnica de Catalunya	3		

* Number of late submissions

Diarization "Who Spoke When" (SPKR)

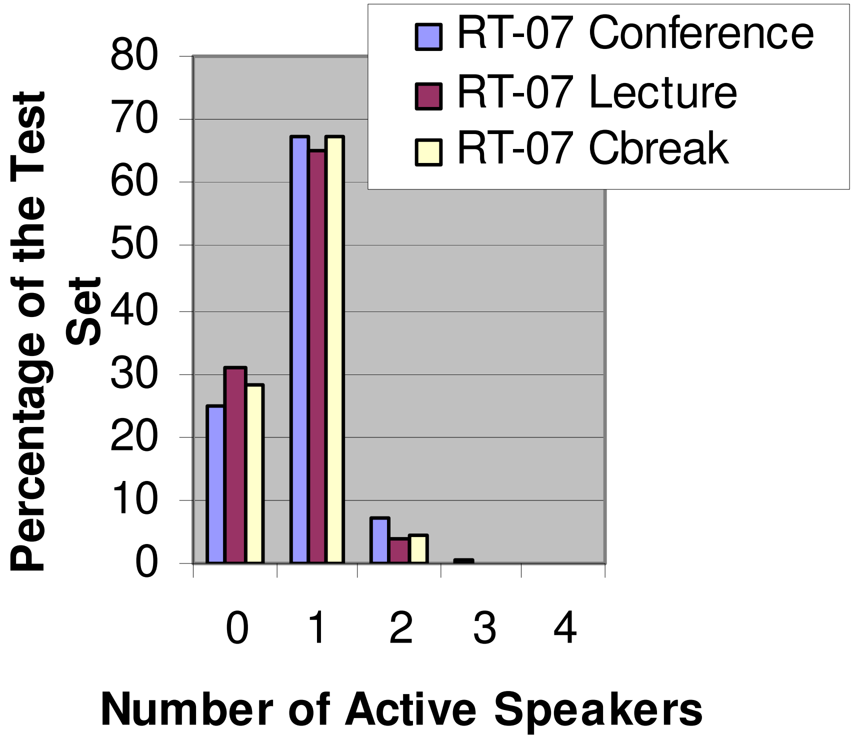
- Task:
  - –Detect segments of speech an cluster them by speaker
- Primary input condition:
  - –Multiple Distant Mics on one or more of the sub-domains
- Participating sites:
  - –Conference Room: AMIDA, I2R/NTU, ICSI, LIA, LIMSI, UPC
  - –Lecture Room: IBM, LIA, LIMSI
  - –Coffee Break: AMI
- Changes for RT-07
  - –Reference segments determined from forced word alignments generated with LIMSI tools

SPKR System Evaluation Method

- Step 1: Speaker alignment
  - A one-to-one mapping between reference speaker segment clusters and system determined speaker clusters
  - The mdeval tool was used with a +/- 250ms no-score collar around reference segment boundaries
- Step 2: Error metric computation
  - Diarization Error Rate (DER) – the ratio of incorrectly detected speaker time to total speaker time
  - Error Types:
    - Speaker assignment errors (i.e., detected speech but not assigned to the right speaker)
    - False alarm detections
    - Missed detections
  - Three scorings performed
    - All speech (Primary metric)
    - Non-overlapping speech (for backward compatibility)
    - Scoring as a Speech Activity Detection system

Test Set Measurements:

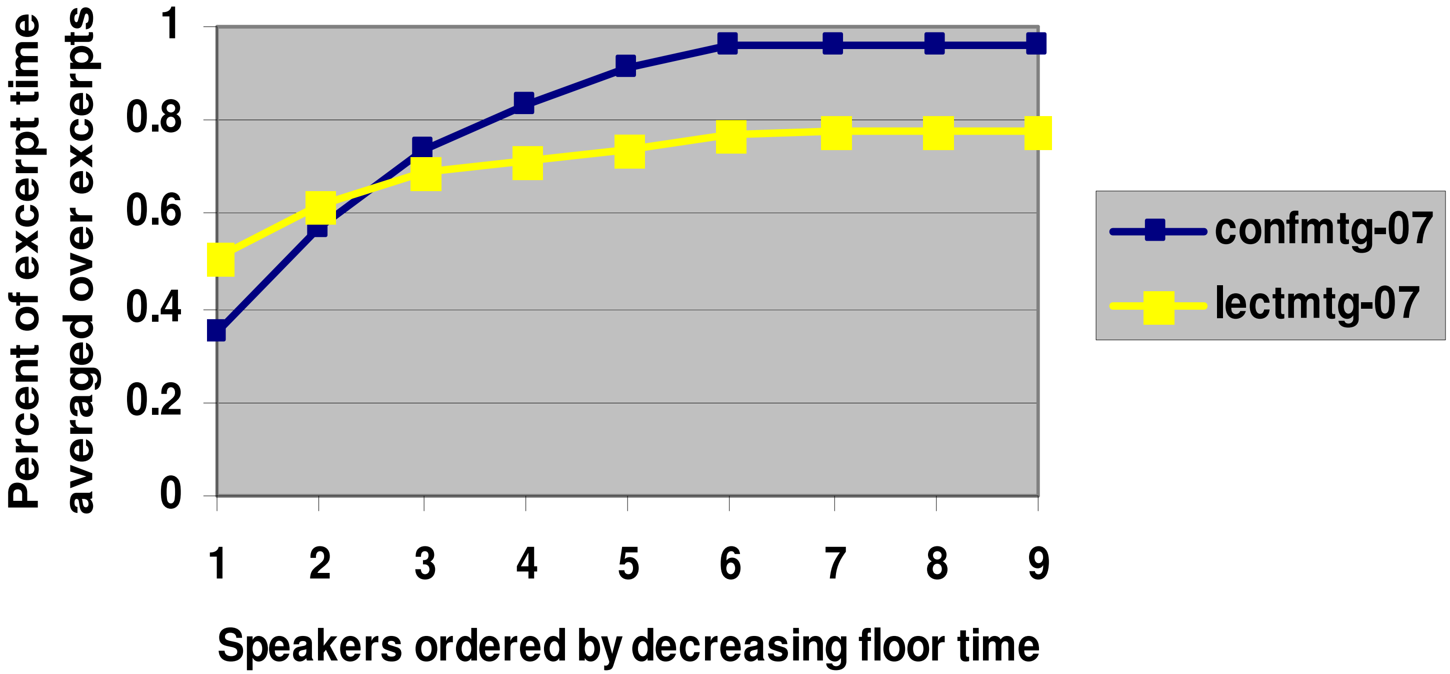
Amount of Overlapping Speech

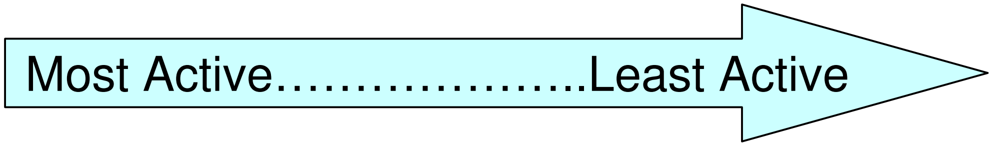


- Speaker activity measured every 0.1 second
- Conference data has more interactivity

Test Set Measurements:

“Floor” Time Averaged Over Excerpts

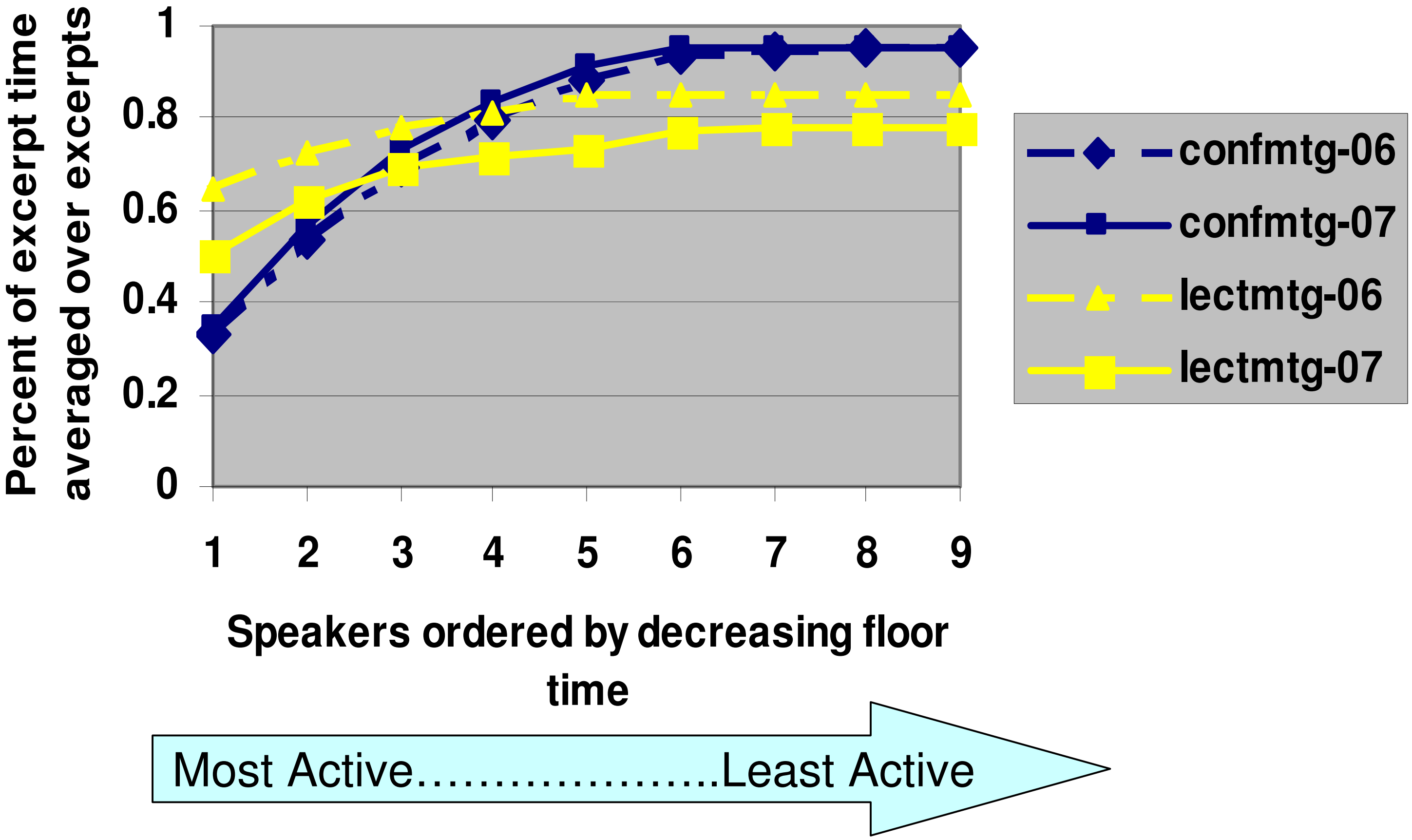



Most Active……………….Least Active

- Conference and Lecture have different distributions

Test Set Measurements:

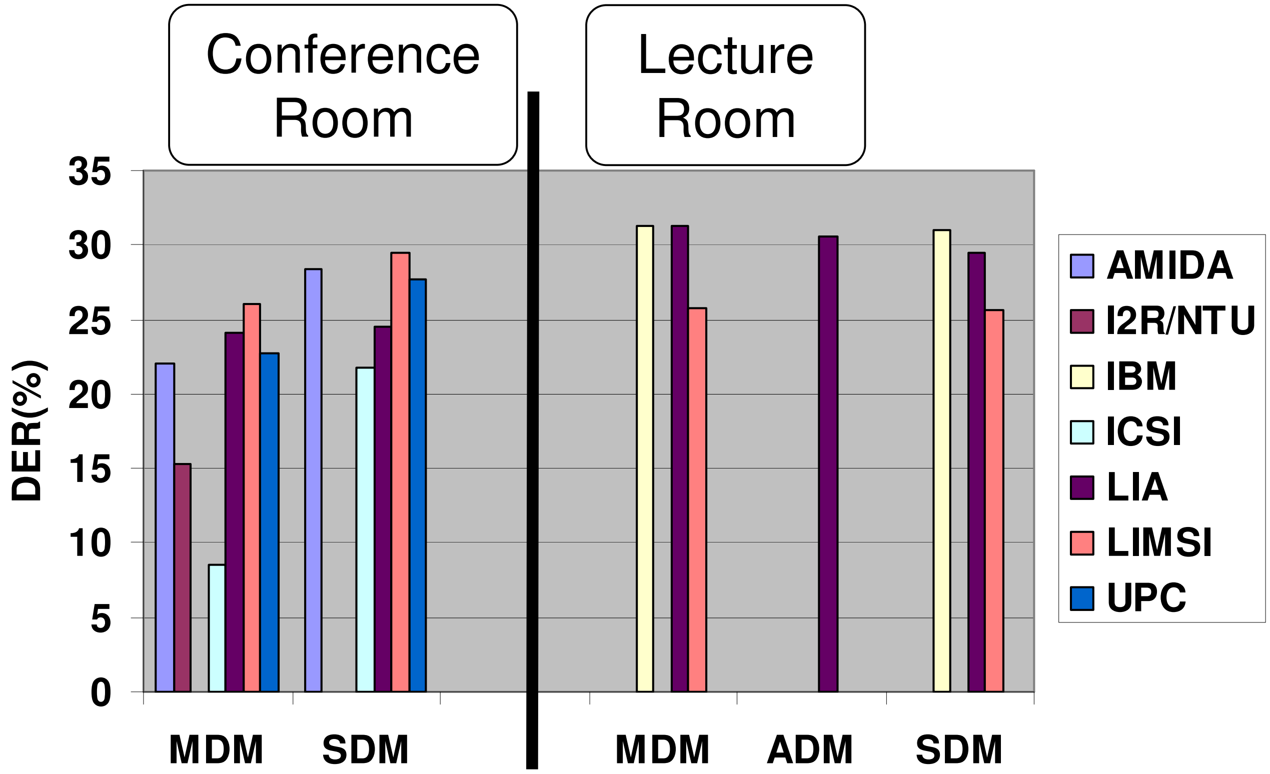
“Floor” Time Averaged Over Excerpts



- Conference and Lecture have different distributions
- ’06 Lecture data has a more dominant main speaker

RT-07 SPKR Results

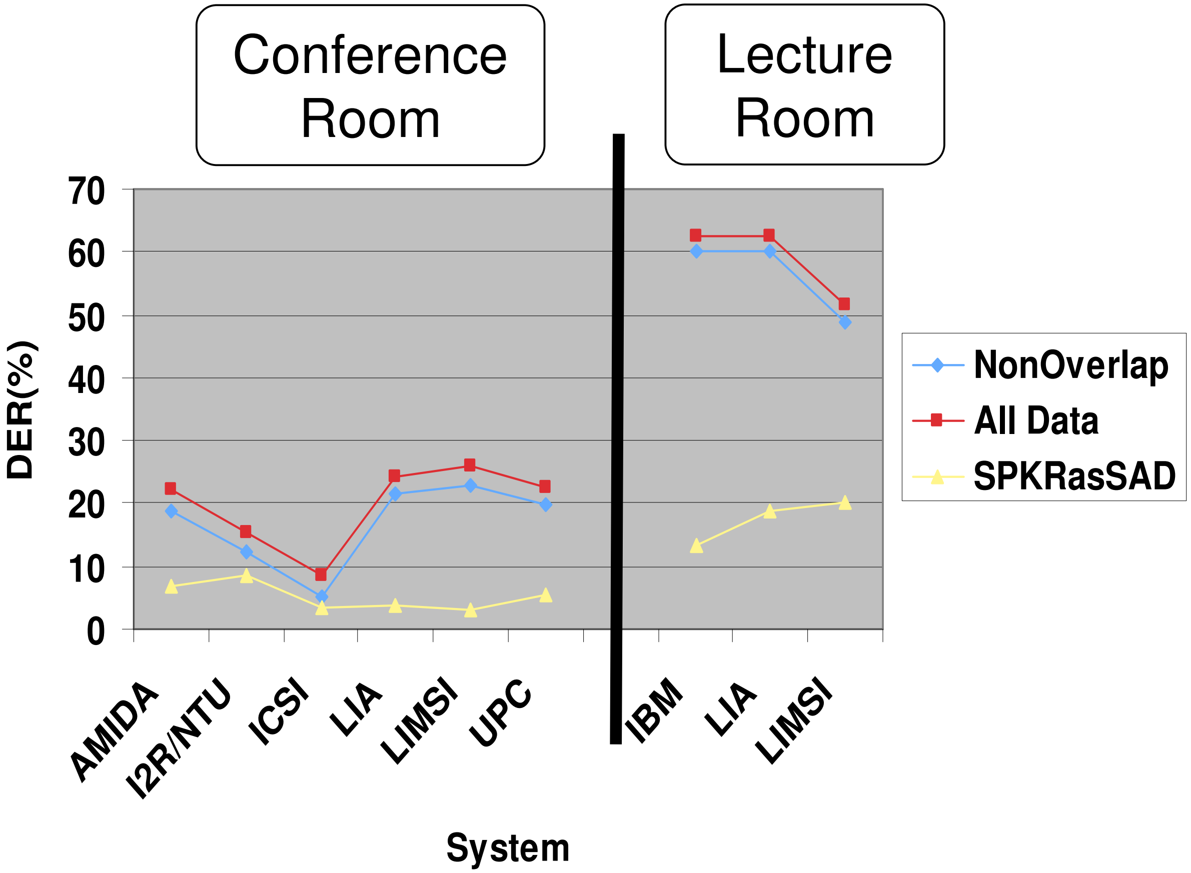
Primary Systems, All Speech



- Lecture DERs are higher that Conference
- Improvement with MDM (from SDM) is mixed
- WOW … ICSI has < 10%DER

RT-07 SPKR Results

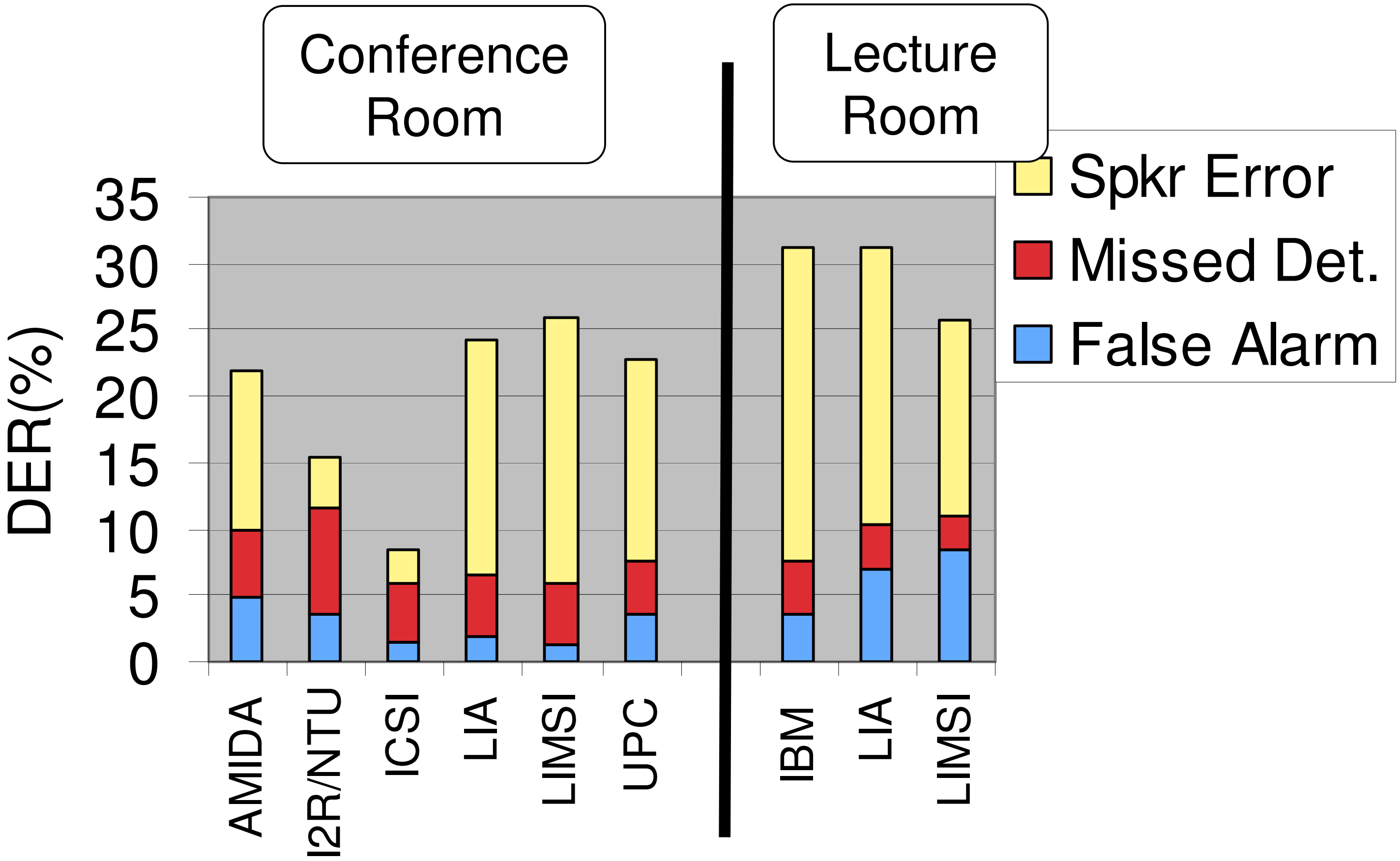
Primary Systems, All Speech



- High correlation between with/without overlap
- SAD scores are commensurate within domain

RT-07 Primary SPKR MDM Systems

DER Split by Error Type



- Speaker Errors dominate the scores

Predicting the Right Number of Speakers For Conference Data

Site	Speaker DER	Speech Activity Detection DER	Average Number System Speakers	Meetings with Correct #speakers (out of 8)	Average Incorrect Number of Speakers (Nsys-Nref)
ICSI	8.51	3.33	4.5	7 (87.5%)	0.1
I2R/NTU	15.32	8.65	4.4	6 (75%)	0
UPC	22.7	5.39	3.9	2 (25%)	-0.5
LIA	24.16	3.69	4.9	1 (12.5%)	0.5
LIMSI	26.07	3.23	12.3	1 (12.5%)	4.8
AMIDA	22.03	6.73	7.1	0 (0%)	2.8

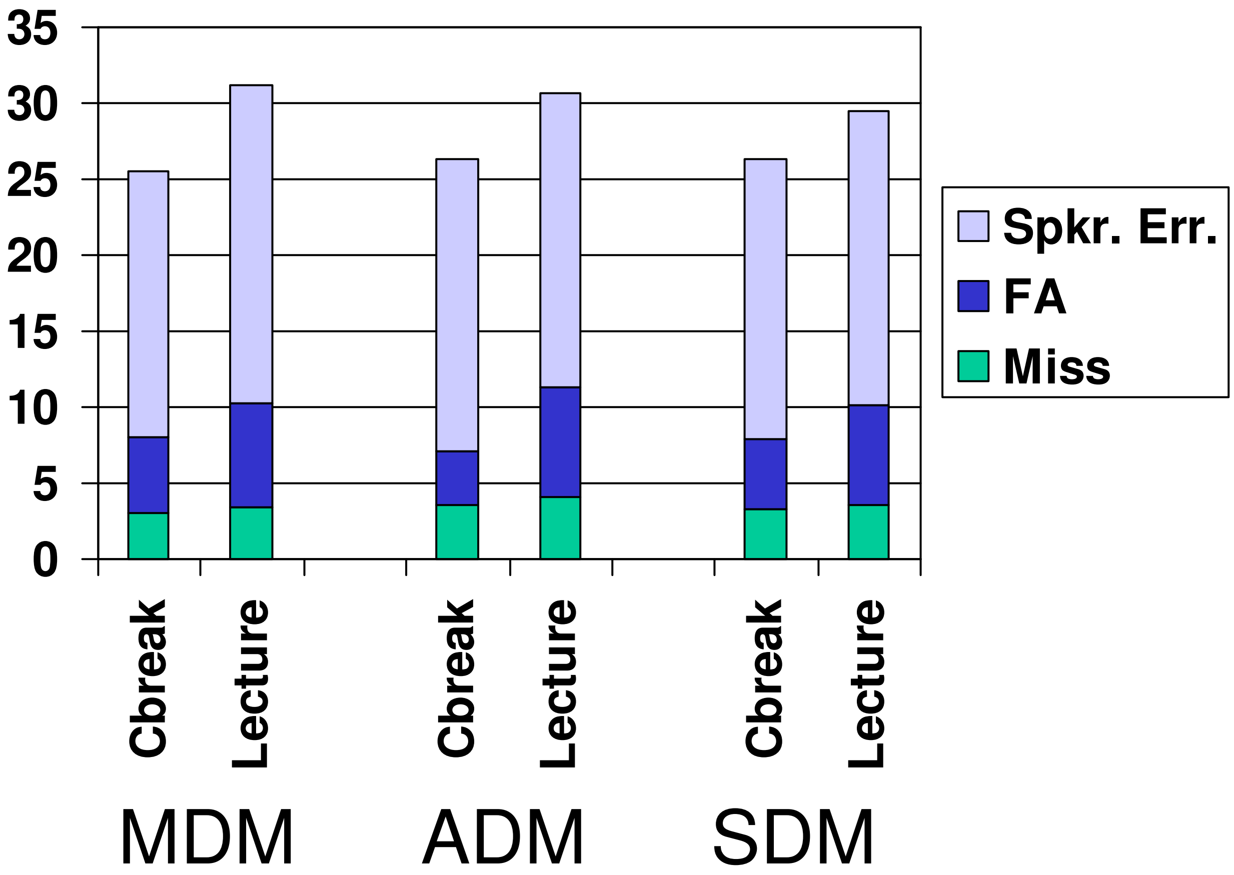
- Predicting the right number of speakers is key
- Lecture data exhibits the same pattern – incorrect speaker count

Questions to Ponder

- What is the challenging part of this task?
  - Predicting the right number of speakers
  - Handling overlap/non-overlapping speech
  - SAD
- Is the test set construction appropriate for this task?
  - 8 trials (one per excerpt), isn’t enough
  - Should the number of meetings be expanded?
  - Should the excerpts be split apart?

Lecture vs. Coffee break (LIA only)

- Large difference mostly occurring as false and speaker errors

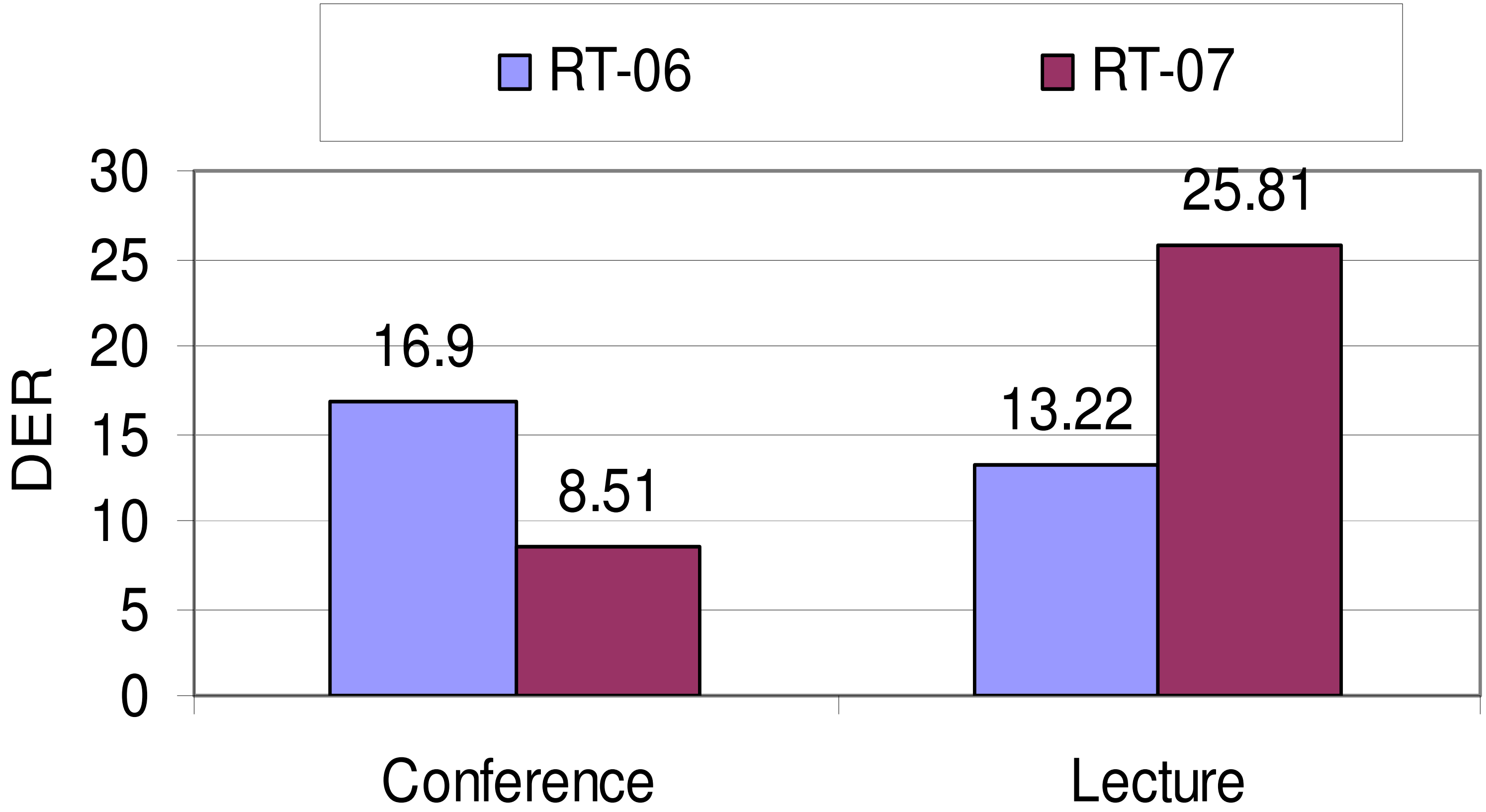


Predicting the Right Number of Speakers For Lecture Data

Site	Speaker DER	Speech Activity Detection DER	Average Number System Speakers	Meetings with Correct #speakers (out of 32)	Average Incorrect Number of Speakers (Nsys-Nref)
IBM	31.22	6.59	3	6 (18.7%)	-1.2
LIA	31.23	9.34	1.25	0 (0%)	-3.1
LIMSI	25.81	10.07	7.8	5 (15.6%)	3

Historical Best System MDM SPKR Performance

(Forced Alignment Mediated)



- drf

Conclusions

- The evaluation ran smoothly
  - Forced alignment mediated reference segmentations were used for this year's test set.
  - SAD scoring as a diagnostic is valuable
- '07 Lecture data is more similar to Conference data
  - SPKR on interactive lectures is now a harder problem
- ICSI's low DER for Conference data is impressive
  - But, this is not a solved problem
  - Is this an indication we need a larger test set?