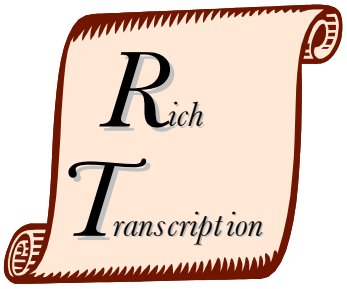
The Rich Transcription 2007 Speech-To-Text (STT) and Speaker Attributed STT (SASTT) Results

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

Jonathan Fiscus and Jerome Ajot

May 10-11, 2007

Rich Transcription 2007
Meeting Recognition Workshop



Speech-To-Text

- Task:
    - Transcribe the spoken words
- Primary input condition:
    - –Multiple Distant Mics on one or more of the sub-domains
- Participating sites:
    - –Conference Room: AMI, SRI/ICSI
    - –Lecture Room: IBM, SRIICSI, UKA
    - –Coffee Break: SRIICSI

Speech-To-Text Evaluation Protocol

- Step 1: Transcript normalization
  - Motivation: The legitimate transcription variability and ambiguity should not cause penalty
    - Differentiating /gonna/ from /going to/ is sometimes difficult
  - Text filtering rules applied to both the reference and system transcript

Speech-To-Text Evaluation Protocol

- Step 1: Transcript normalization
  - Motivation: The legitimate transcription variability and ambiguity should not cause penalty
    - Differentiating /gonna/ from /going to/ is sometimes difficult
  - Text filtering rules applied to both the reference and system transcript
- Step 2: Overlapping Speech Text Alignment
  - Motivation: Identify and classify errors by finding an optimal one-to-one mapping of reference to system words

Speech-To-Text Evaluation Protocol

- Step 1: Transcript normalization
  - Motivation: The legitimate transcription variability and ambiguity should not cause penalty
    - Differentiating /gonna/ from /going to/ is sometimes difficult
  - Text filtering rules applied to both the reference and system transcript
- Step 2: Overlapping Speech Text Alignment
  - Motivation: Identify and classify errors by finding an optimal one-to-one mapping of reference to system words
- Step 3: Error computation
  - Primary Metric: Word Error Rate (WER):

$$100 * \frac{N_{substitutions} + N_{Insertions} + N_{Deletions}}{N_{referenceWords}}$$

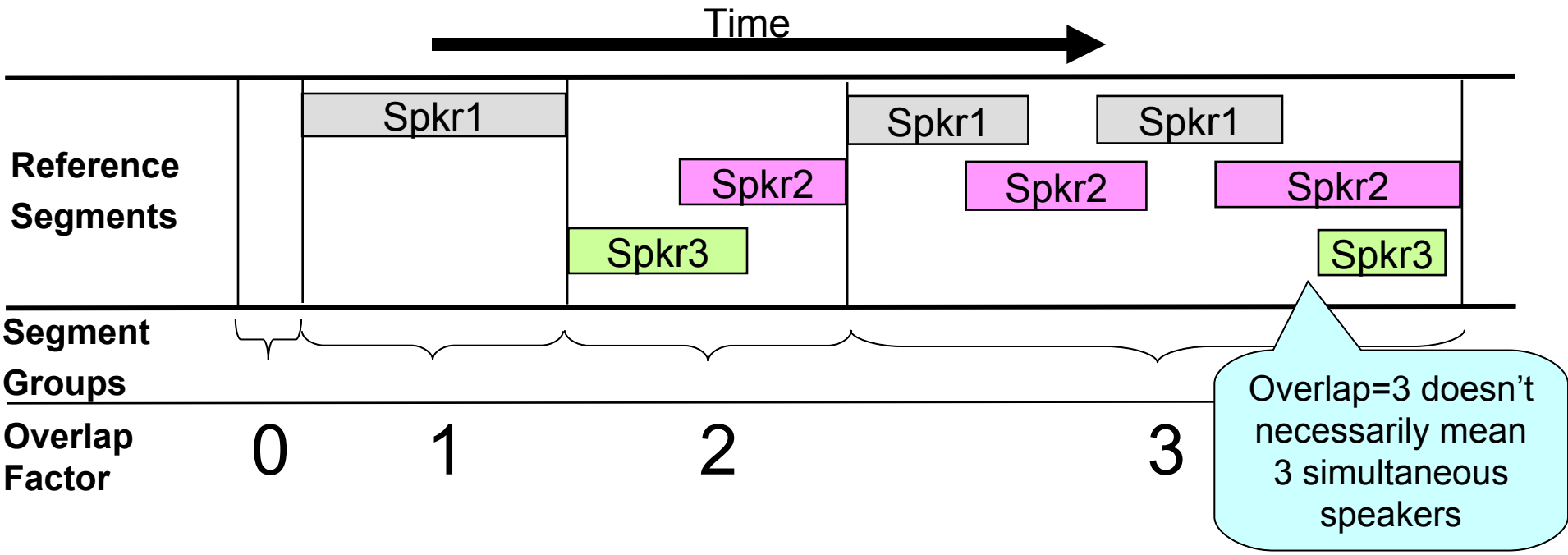
  - 0% is perfect, >100% possible

Overlapping Speech Text Alignments

- Solution: Multi-dimensional text alignments produce the 1:1 mapping
  - Each speaker (reference and system) is a dimension in a Levenshtein Edit Distance matrix
  - NIST developed the ASCLITE alignment engine
- Challenge: Computational complexity limits
  - Several techniques limit the search space
    - Pre segmenting the reference transcript into “Segment Groups”
    - Heuristic pruning, application contstraints, and memory compression
- Net Effect:
  - More evaluable data, faster scoring times, controlled conditional scoring
    - A 40GB alignment matrix can be computed in 2GB of RAM
  - However: more power is needed: can’t handle 272 TB

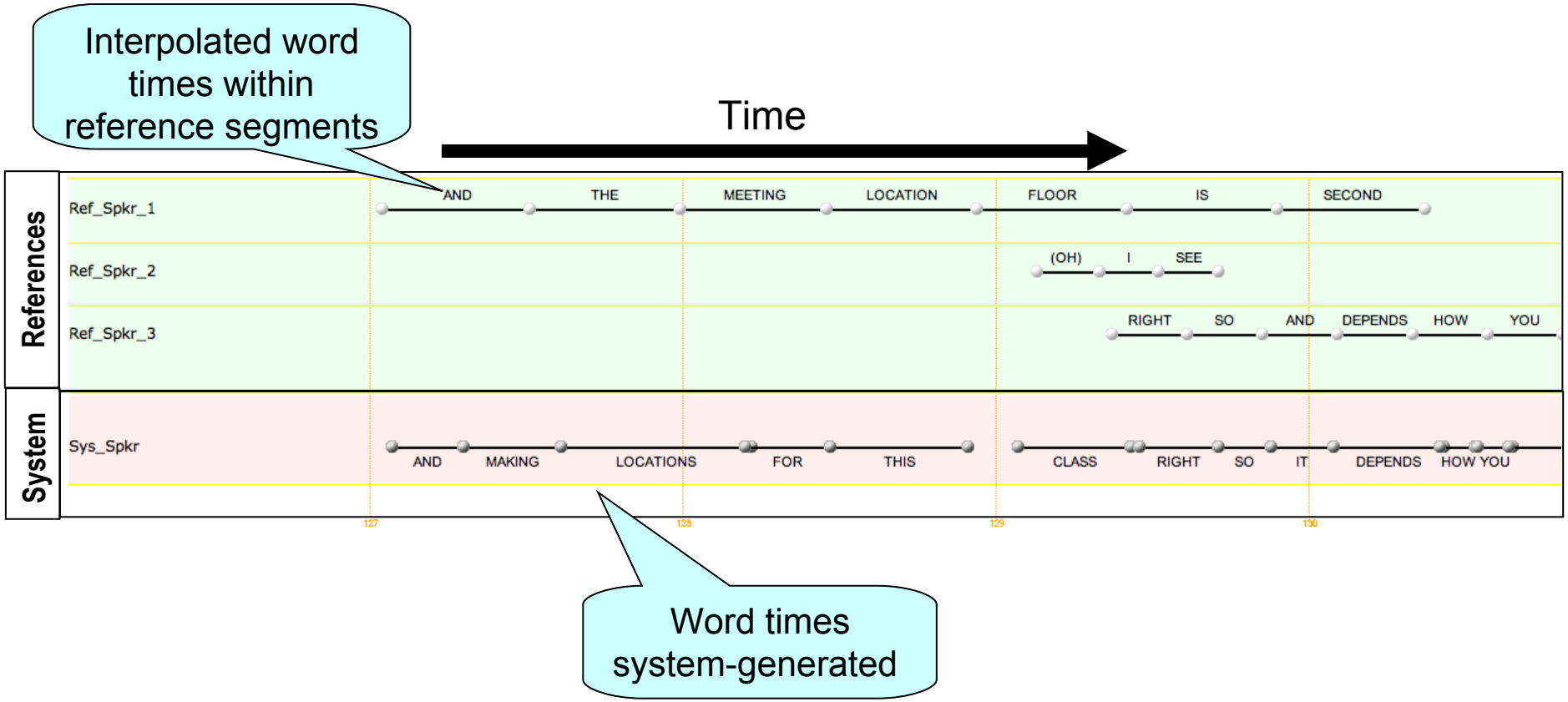
Segment Groups

- Divide the reference transcript segments into independent units based on segment times



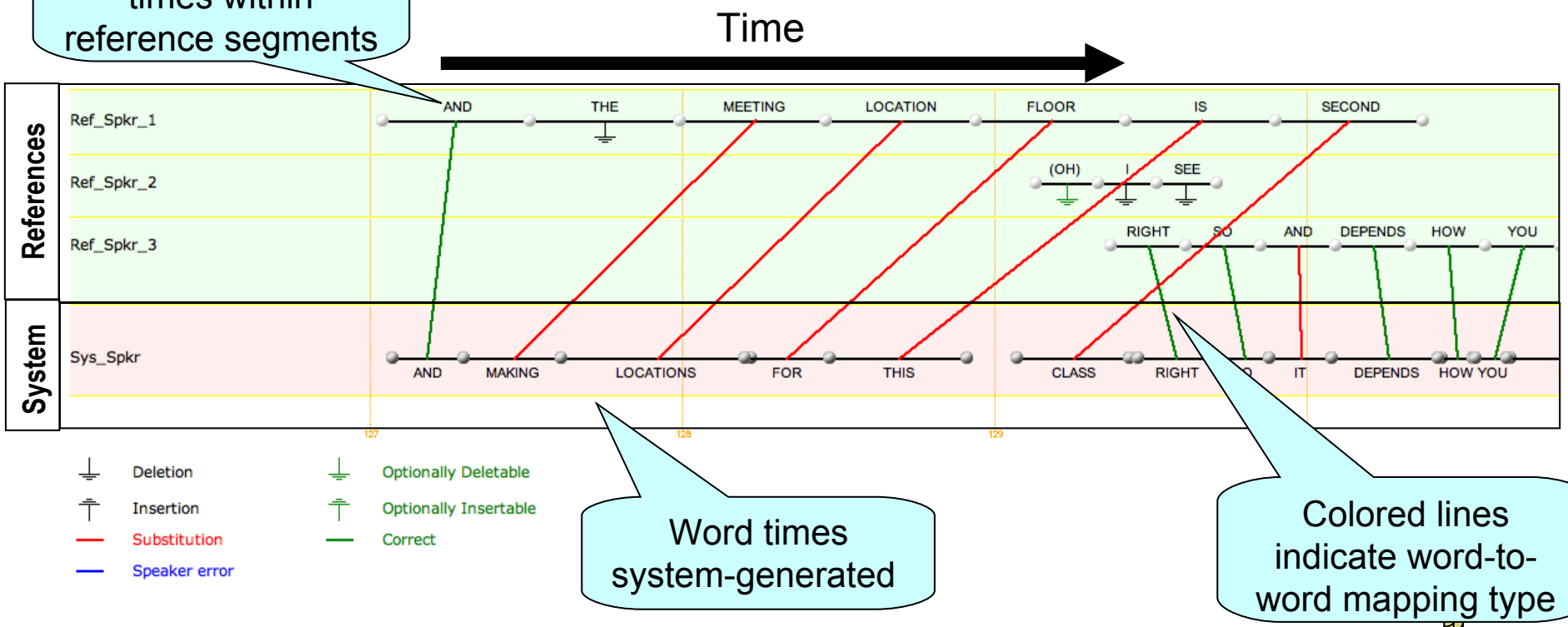
- Smaller overlap factor => faster alignment times
- Overlap factors used for conditional scoring

Multi-Dimensional Alignment Visualization for STT



Multi-Dimensional Alignment Visualization for STT

Interpolated word times within reference segments



Word times system-generated

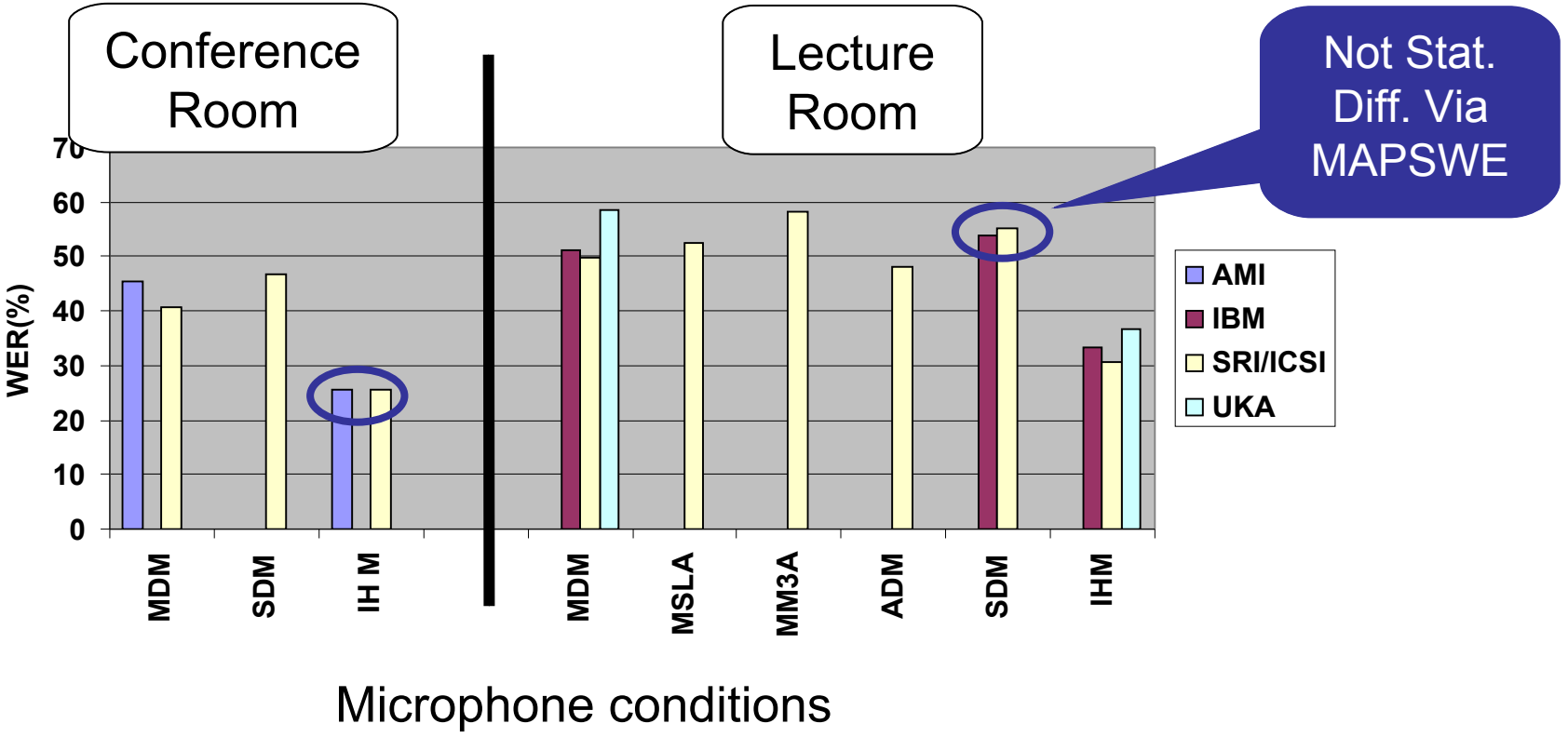
Colored lines indicate word-to-word mapping type

4 Dimensional Alignment : Labeled as Overlap == 3

0.12 MB to align

NIST National Institute of Standards and Technology

RT-07 STT Primary System Results (Overlap<=4 Results)



Distant Mic. Test Set	Percent of scored words
Conference	99.3%
Lecture	99.6%

Distant Microphone Scoring
Percentage of Evaluable Test Data

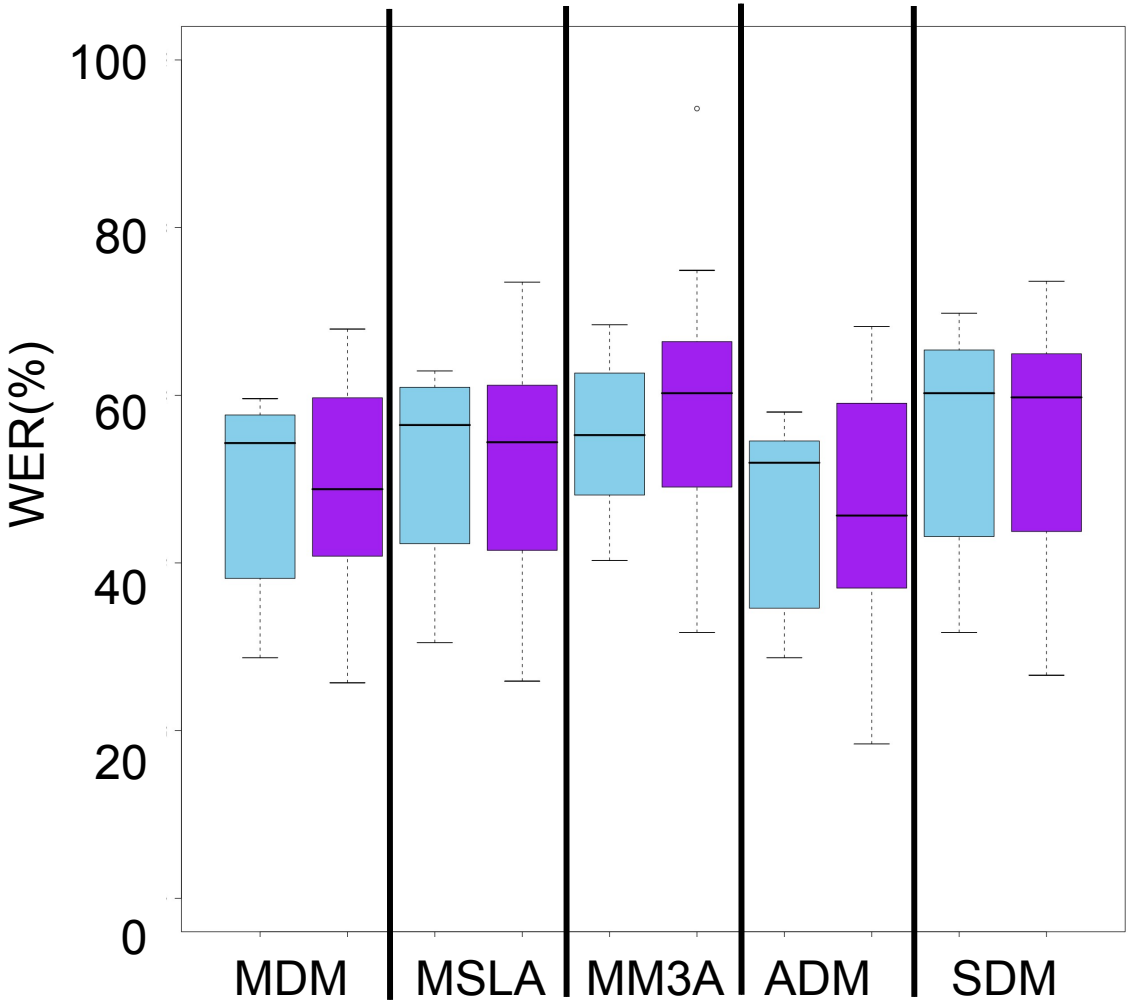
Fraction of Evaluable Words

RT-07 Distant Microphone Conditions

	Test Set	STT (Overlap <= 4)	SASTT (Overlap <= 3)
RT-07	Conference	99.3%	84.5%
	Lecture	99.6%	97.0%
	Coffee Break	100%	100%
RT-06	Conference	84.1%	
	Lecture	97.4%	

Lecture vs. Coffee Break

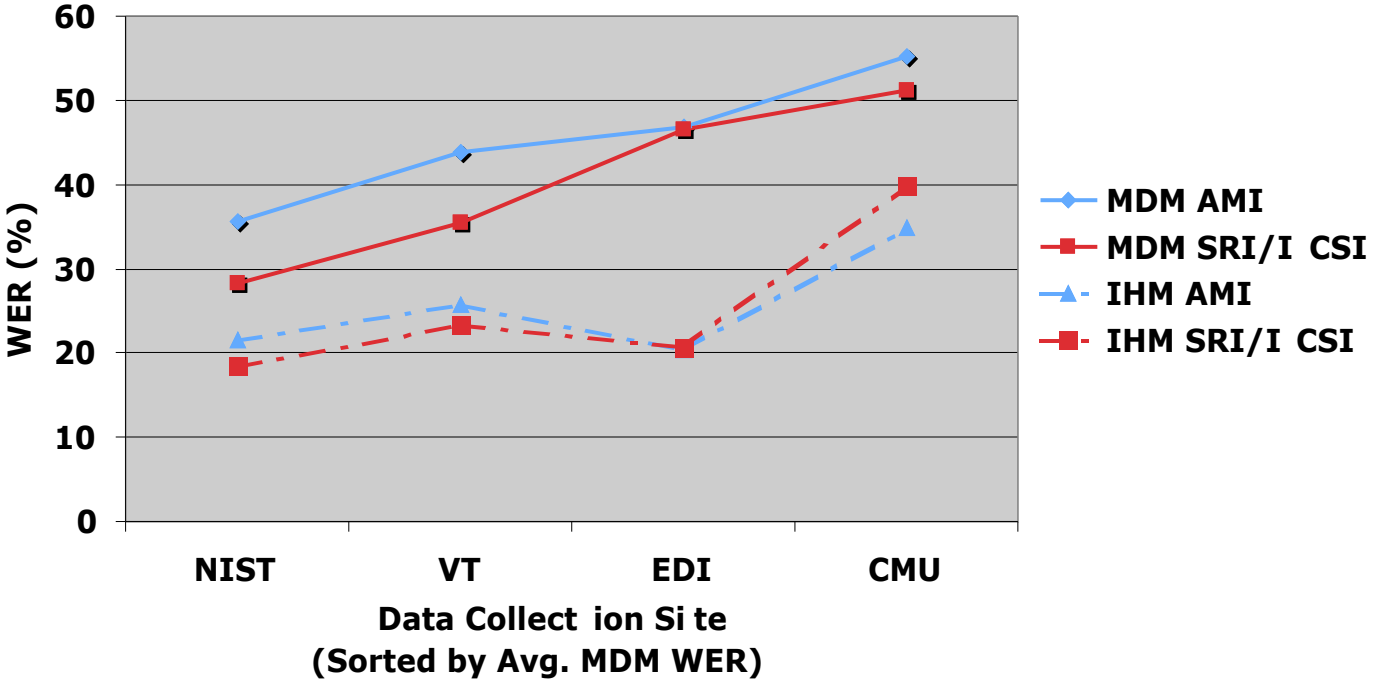
SRI/ICSI Primary Results, Excerpt WER



- Negligible difference

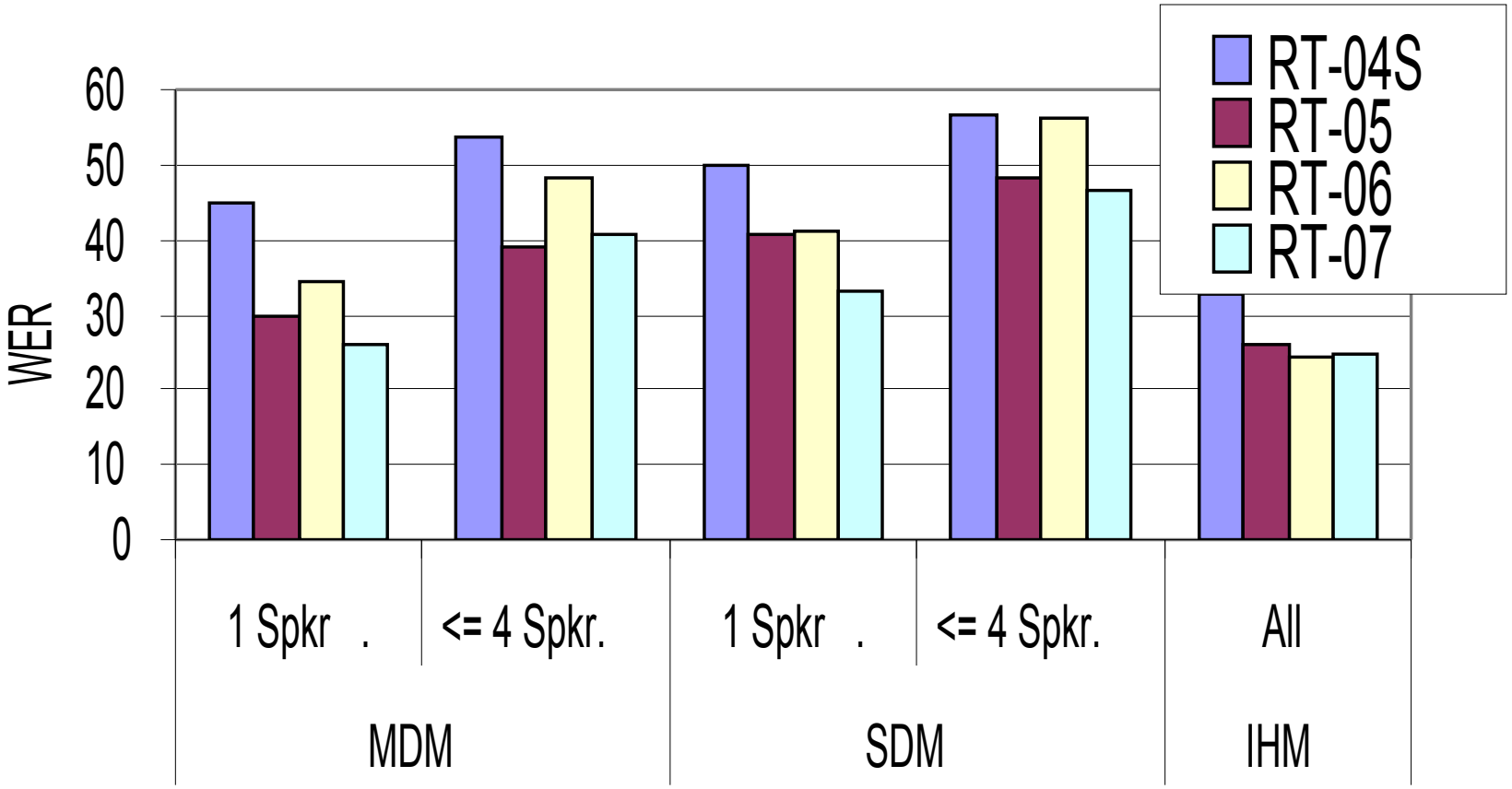
Conference Data by Collection Site

IHM and MDM Results for Primary Systems

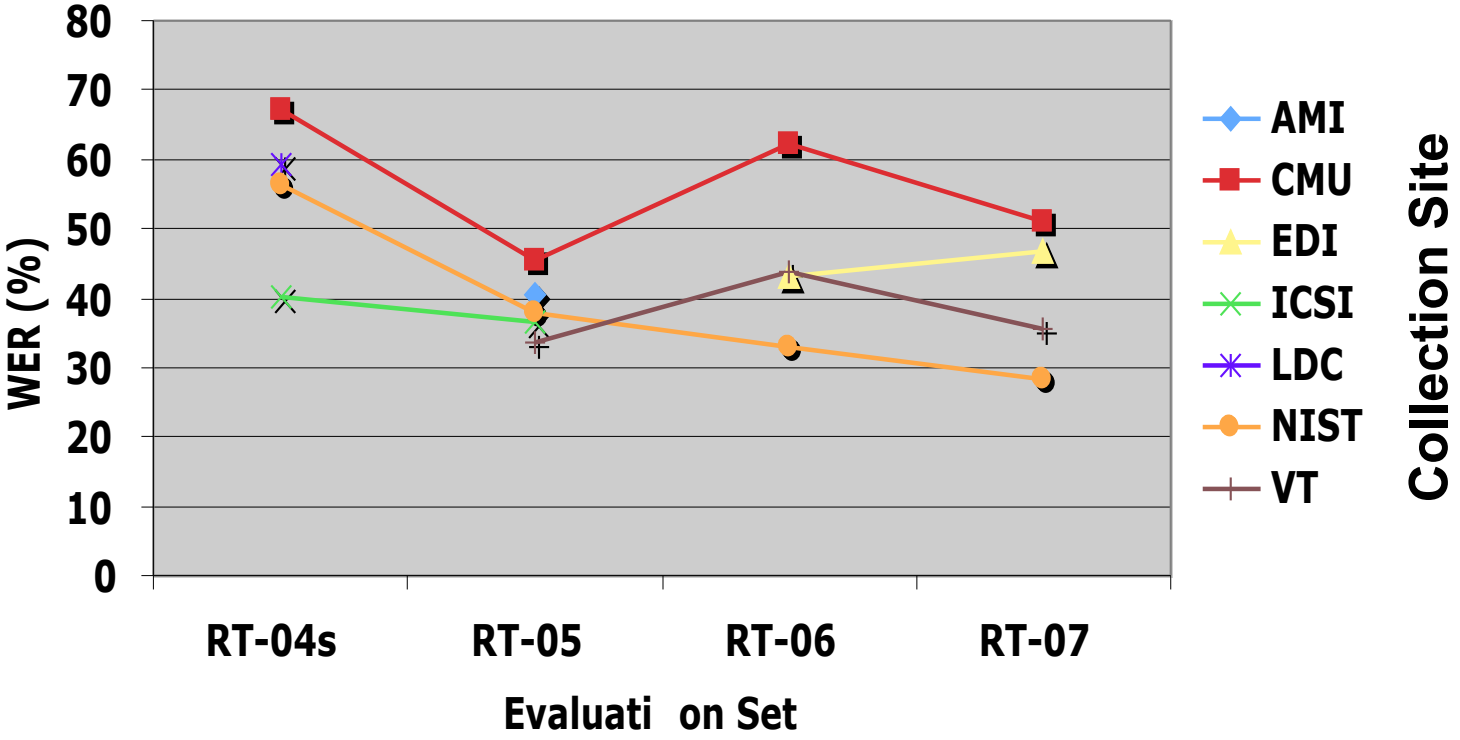


- Definite collection site effect
- CMU meetings are more difficult
- Larger difference between MDM and IHM for EDI data

Historical STT Performance in the Conference Meeting Domain

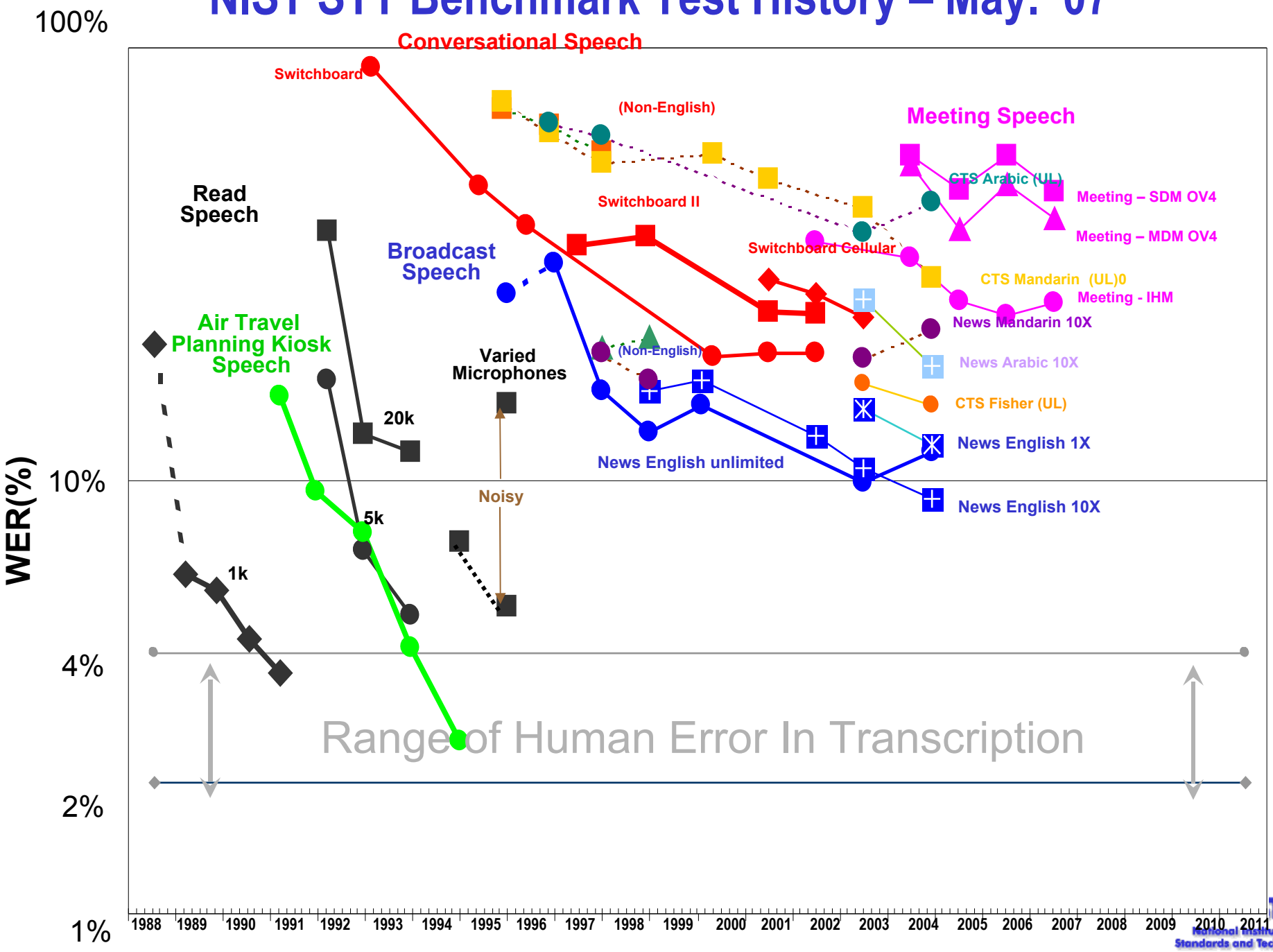


Historical Conference STT: MDM WER Split by Collection Site

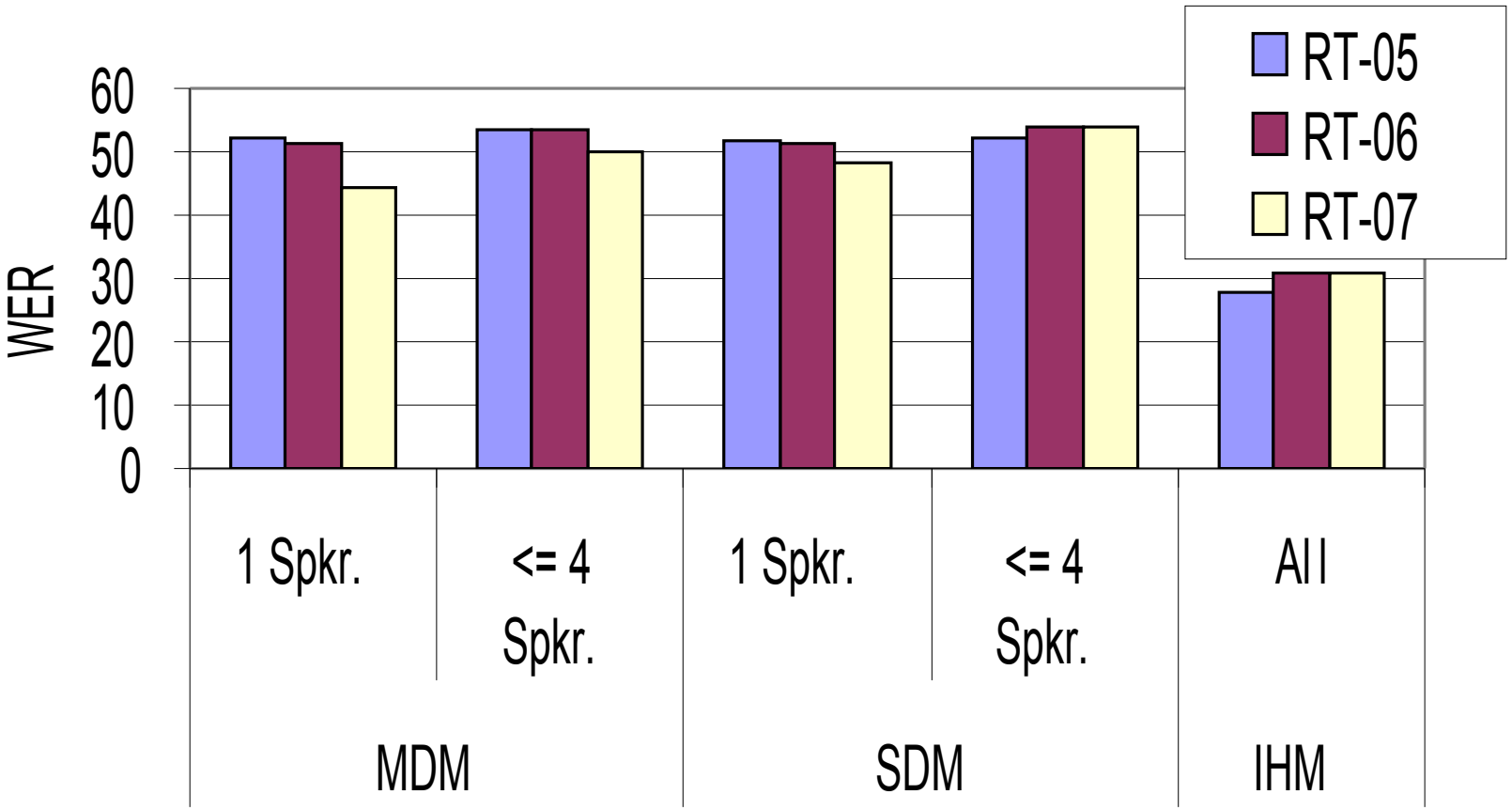


- Strong NIST trend : the “sheeps”
- Variability for other sites

NIST STT Benchmark Test History – May. '07

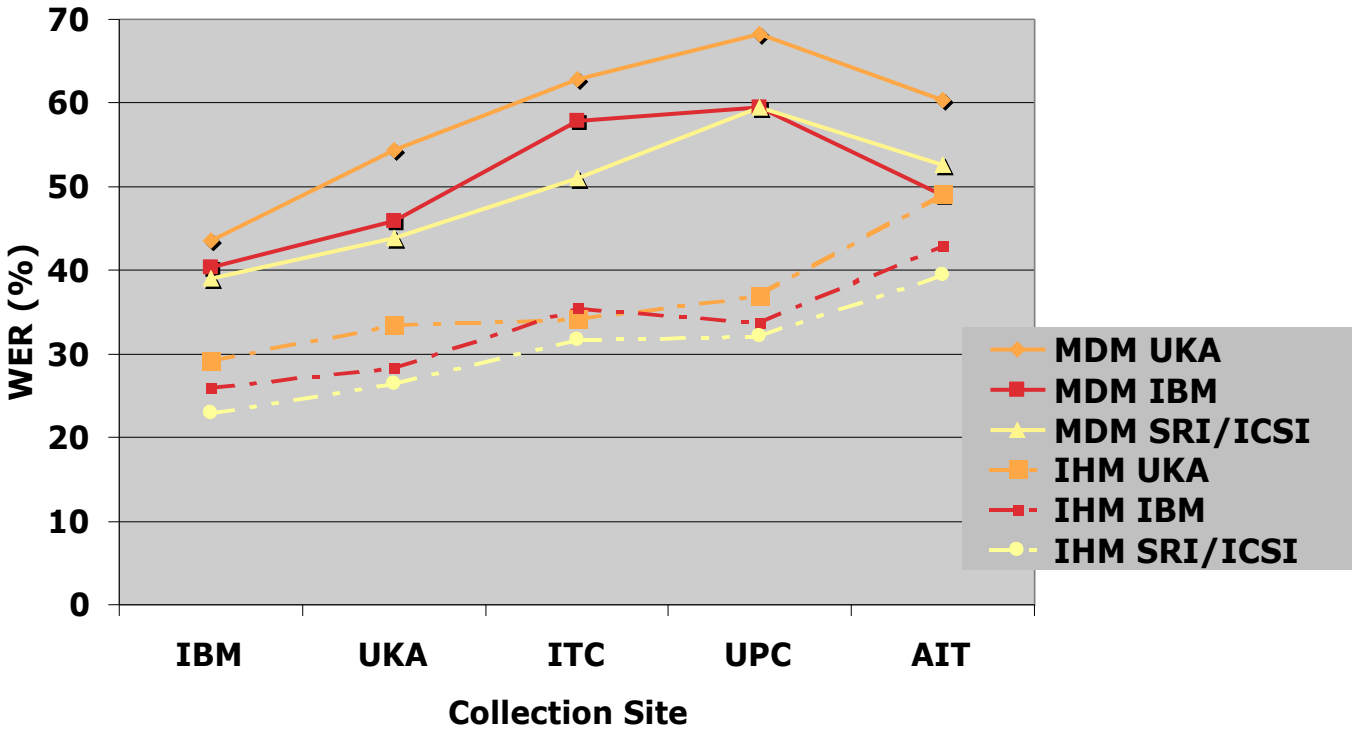


Historical STT Performance in the Lecture Meeting Domain



Lecture Data by Collection Site

IHM + MDM Results for Primary Systems



- Definite collection site effect
- AIT's acoustic challenge is less for the MDM condition

Speaker Attributed STT

- Task:
  - Transcribe the spoken words and associate them with a speaker
- New evaluation task for RT-07
  - Merge of STT and Speaker Diarization systems
  - Will require joint optimizations
- Primary input condition:
  - Multiple Distant Mics on one or more of the sub-domains
- Participating sites:
  - Conference Room: AMI, SRIICSI
  - Lecture Room: AMI, IBM, LIMSI, SRIICSI
  - Coffee Break: AMI, SRIICSI

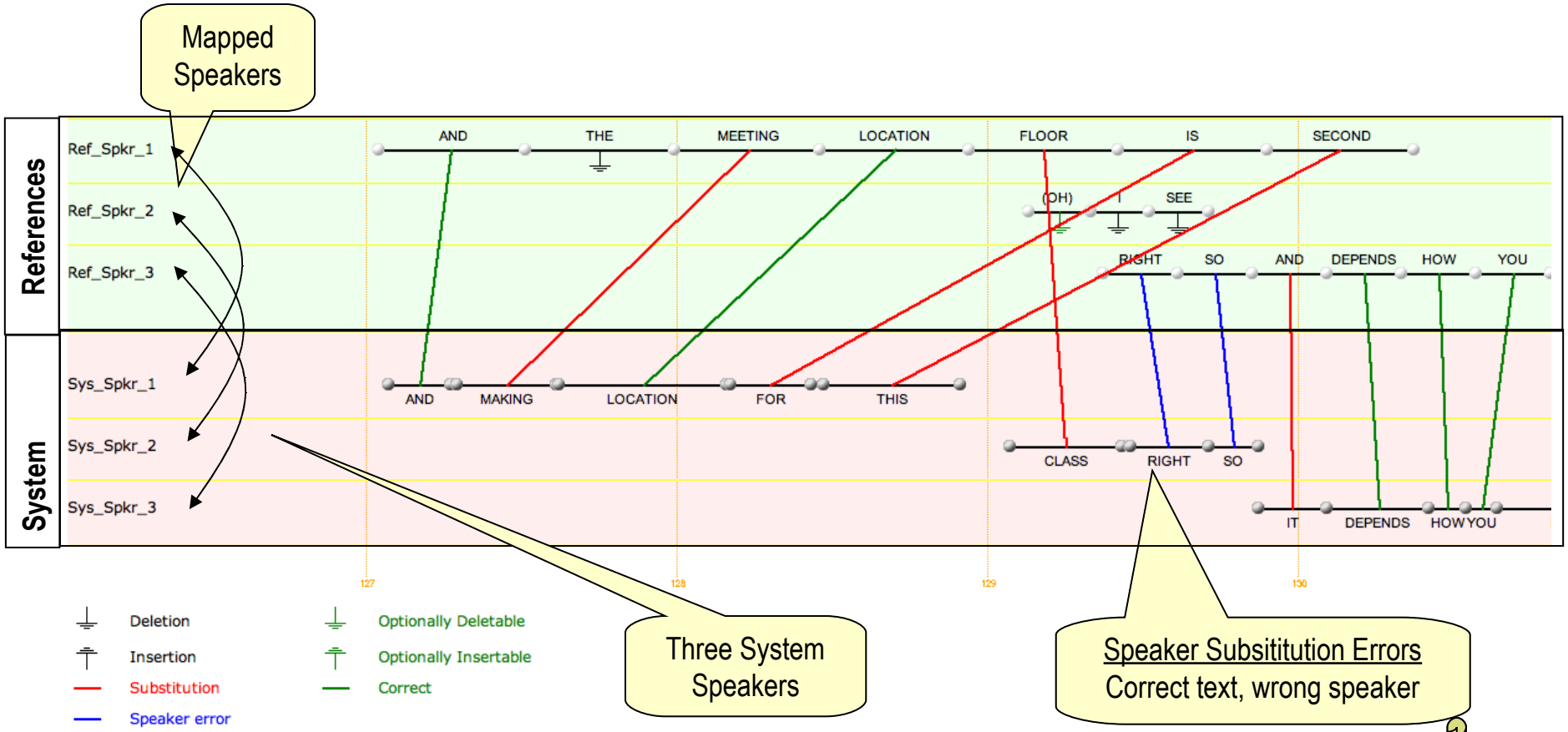
SASTT Evaluation Protocol

- Step 1: Transcript normalization
  - Identical to STT
- Step 2: Speaker Alignment
  - Define what is the “Correct” speaker
  - A one-to-one mapping between reference speakers and system speakers
  - Same time-based scoring method as used for the Speaker Diarization Task
    - Except system segments derived from recognized word locations
- Step 3: Text Alignment
  - A one-to-one mapping is found between the reference and system transcripts
  - Changes to mapping requirements:
    - Correct: matching words and mapped reference and system speaker
    - Speaker Substitution: matching words and non-mapped reference and system speakers
    - Substitutions: non-matching texts
- Step 4: Error computation
  - Primary Metric: Speaker Attributed Word Error Rate (SWER):

$$100 * \frac{N_{substitutions} + N_{Insertions} + N_{deletions} + N_{SpeakerSubstitions}}{N_{referenceWords}}$$

  - 0% is perfect, >100% possible

Multi-Dimensional Alignment Visualization for SASTT



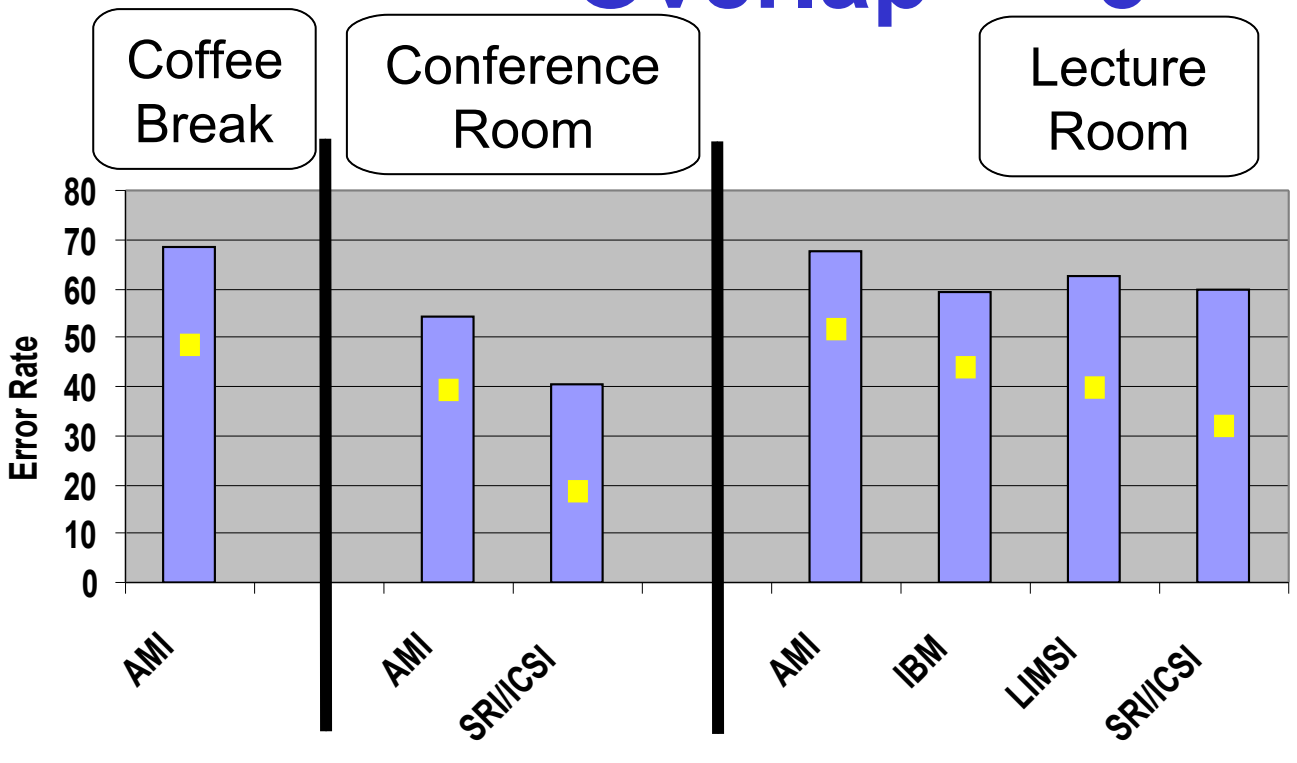
6 Dimensional Alignment : Labeled as Overlap == 3

2.12 MB to align → 18 times bigger than STT

NIST National Institute of Standards and Technology

Primary MDM SASTT Results

Overlap <= 3

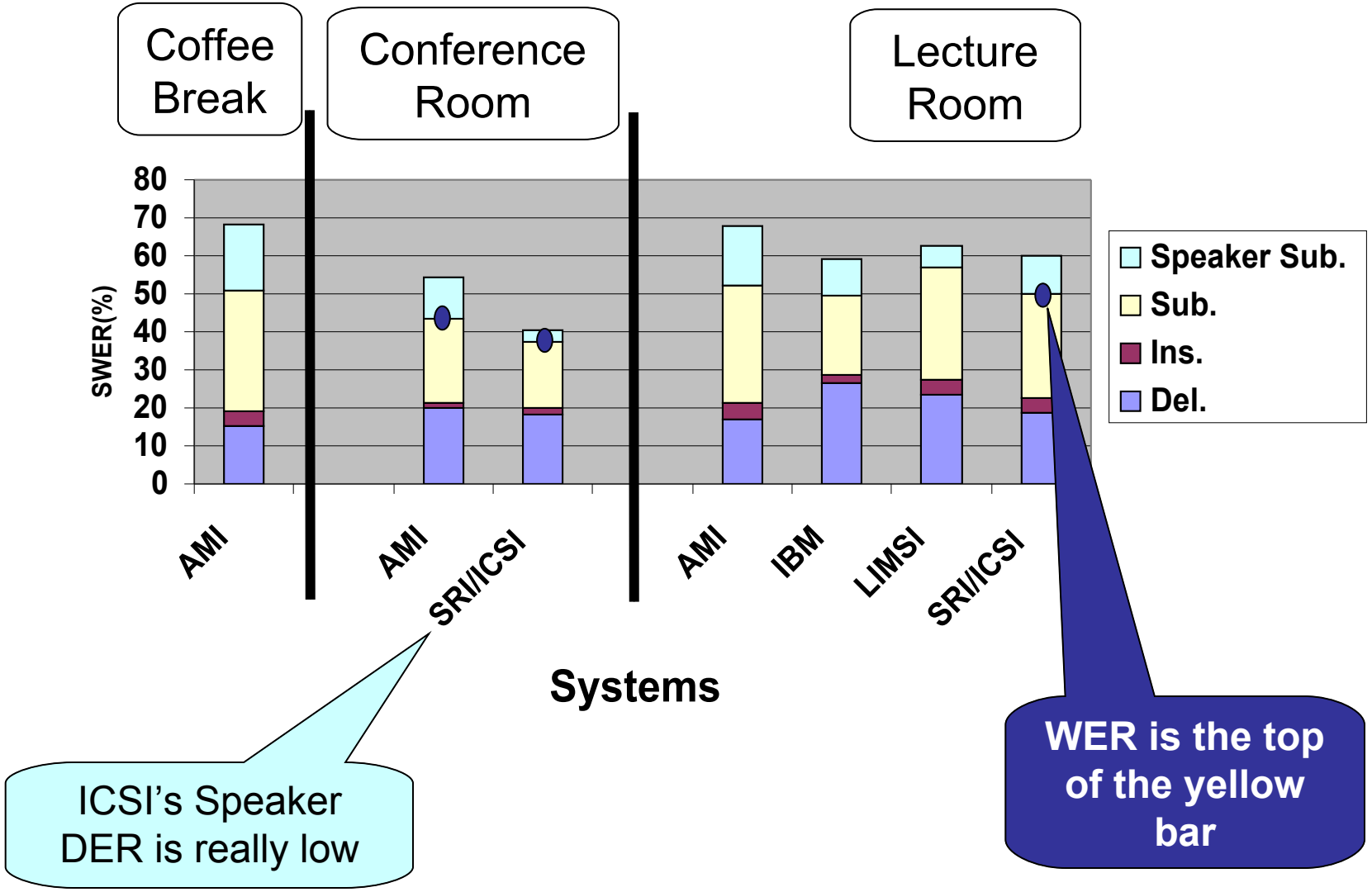


DER* is based on segments derived from recognized words

Distant Mic. Test Set	Percent of scored words
Conference	84.5%
Lecture	97.0%

Primary MDM SASTT Results

Overlap <= 3



Conclusions

- RT-07 Test Sets
  - Strong collection site effect in both test sets
- Successfully implemented the SASTT task
  - The future for Rich Transcription
  - We were able to score most Overlap<=3
  - However, the current alignment technique has hit its limits
    - Explore new techniques to handle high-dimensional alignment
- Deeper analysis
  - “Segment group”-based overlap measurements over estimate “true” simultaneous speech
    - Finer grained diagnostics for overlap speech needed
  - Word alignments have limited diagnostic ability
    - Time mediated alignments may be useful