

# The Rich Transcription 2006 Spring Meeting Recognition Evaluation

Jonathan G. Fiscus<sup>1</sup>, Jerome Ajot<sup>1</sup>, Martial Michel<sup>1,2</sup>, John S. Garofolo<sup>1</sup>

<sup>1</sup>National Institute Of Standards and Technology, 100 Bureau Drive Stop 8940, Gaithersburg, MD 20899

<sup>2</sup> Systems Plus, Inc., One Research Court – Suite 360, Rockville, MD 20850  
{jfiscus,ajot,martial.michel,jgarofolo}@nist.gov

**Abstract.** We present the design and results of the Spring 2006 (RT-06S) Rich Transcription Meeting Recognition Evaluation; the fourth in a series of community-wide evaluations of language technologies in the meeting domain. For 2006, we supported three evaluation tasks in two meeting sub-domains: the Speech-To-Text (STT) transcription task, and the “Who Spoke When” and “Speech Activity Detection” diarization tasks. The meetings were from the Conference Meeting, and Lecture Meeting sub-domains. The lowest STT word error rate, with up to four simultaneous speakers, in the multiple distant microphone condition was 46.3% for the conference sub-domain, and 53.4% for the lecture sub-domain. For the “Who Spoke When” task, the lowest diarization error rates for all speech were 35.8% and 24.0% for the conference and lecture sub-domains respectively. For the “Speech Activity Detection” task, the lowest diarization error rates were 4.3% and 8.0% for the conference and lecture sub-domains respectively.

## 1. Motivation

The National Institute of Standards and Technology (NIST) has been working with the speech recognition community since the mid 1980s to improve the state-of-the-art in speech processing technologies. To facilitate progress, NIST has worked with the community to make training/development data collections available for several speech domains. NIST collaborated with the research community to define performance metrics and create evaluation tools for technology developers to perform hill-climbing experiments and measure their progress. NIST also coordinates periodic community-wide benchmark tests and technology workshops to inform the research community and Government sponsors of progress, and to promote technical exchange. The test suites used in these benchmark tests are generally made available to the community as development tools after the formal evaluations.

NIST evaluations have demonstrated great progress in the state-of-the-art in speech-to-text (STT) transcription systems[1]. STT systems in the late 80s focused on read speech from artificially-constrained domains. As the technology improved, the NIST evaluations focused the research community on increasingly difficult challenges with regard to speech modality, speaker population, recording characteristics, language, vocabulary, etc.

The meeting domain presents several new challenges to the technology. These include varied fora, an infinite number of topics, spontaneous highly interactive and overlapping speech, varied recording environments, varied/multiple microphones, multi-modal inputs, participant movement, and far field speech effects such as ambient noise and reverberation. In order to properly study these challenges, laboratory-quality experiment controls must be available to enable systematic research. The meeting domain provides a unique environment to collect naturally-occurring spoken interactions under controlled sensor conditions.

The Rich Transcription Spring 2006 (RT-06S) Meeting Recognition evaluation is part of the NIST Rich Transcription (RT) series of language technology evaluations [1] [2] [7]. These evaluations have moved the technology focus from a strictly word-centric approach to an integrated approach where the focus is on creating richly annotated transcriptions of speech, of which words are only one component. The goal of the RT series is to create technologies to generate transcriptions of speech which are fluent and informative and which are readable by humans and usable in downstream processing by machines. To accomplish this, lexical symbols must be augmented with important informative non-orthographic metadata. These resulting metadata enriched transcripts are referred to as “rich transcriptions”. These metadata can take many forms (e.g., which speakers spoke which words, topic changes, syntactic boundaries, named entities, speaker location, etc.)

The RT-06S evaluation is the result of a multi-site/multi-national collaboration. In addition to NIST, the organizers and contributors included: Athens Information Technology (AIT), the Augmented Multiparty Interaction (AMI) program, the Computers in the Human Interaction Loop (CHIL) program, Carnegie Mellon University (CMU), Evaluations and Language resources Distribution Agency (ELDA), IBM, International Computer Science Institute and SRI International (ICSI/SRI), Institut National de Recherche en Informatique et Automatique (INRIA), The Center for Scientific and Technological Research (ITC-irst), Karlsruhe University (UKA), the Linguistic Data Consortium (LDC), Laboratoire Informatique d'Avignon (LIA), Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), Universitat Politècnica de Catalunya (UPC), and Virginia Tech (VT).

Two tests were built for the evaluation with different types of meeting data: the Conference Meeting meeting sub-domain and Lecture Meeting meeting sub-domain test sets. The two test sets fostered collaboration between the many research programs by providing backward compatible test sets (previous evaluations used Conference Meeting data) and sharing data across programmatic boundaries while accommodating individual programmatic needs.

## **1.2 Rich Transcription Relation to Multi-Modal Technology Evaluations**

Beginning in the early 2000's, a number of independent meeting-domain focused research and evaluation efforts were started: the European Union (EU) Computers in the Human Interaction Loop (CHIL), the EU Augmented Multiparty Interaction (AMI) program, the US Video Analysis and Content Extraction (VACE) program, and the NIST Rich Transcription Evaluation series which shared many aspects of unimodal and multi-modal research. Since the recognition of human-human communi-

cations in meetings is multi-modal by nature, NIST decided to expand the evaluations it supports in this area to facilitate the development of a multi-modal research community.

NIST decided to take several steps to create a collaborative international evaluation effort that would share knowledge and resources across research programs both in the US and abroad, leverage efforts, standardize data, metrics, and interchange formats across efforts, and help increase the critical mass of research in multi-modal meeting understanding technologies. Advisory committees were established to develop plans for upcoming evaluations and workshops that selected cross-program evaluation tasks to support. As a result, the RT evaluation became a program-independent evaluation forum for language technologies with a focus on the meeting domain and the extraction of language content from both audio and video source channels. The second result was to create the Classification of Events, Activities, and Relationships (CLEAR) evaluation and workshop which focuses on spatial analysis problems.

During 2006, RT remained co-located with the 3rd Joint Workshop on Multi-modal Interaction and Related Machine Learning Algorithms (MLMI-06) and the CLEAR workshop occurred earlier as a separate event. For 2007, both the RT and CLEAR workshops will be co-located so that initial discussions regarding fusion technologies can be discussed and future evaluations can be planned accordingly.

## **2. Rich Transcription Spring 2005 Meeting Recognition Evaluation**

The RT-06S evaluation was similar to the RT-05S evaluation. Two major changes were made to the evaluation: first, the Source Localization evaluation task was moved to the Classification of Events, Activities, and Relationships (CLEAR) [8] evaluation and second, overlapping speech was evaluated in both the STT and Diarization “Who Spoke When” (SPKR) tasks instead of restricting the scoring to only non-overlapping speech.

All participating teams were required to submit a single primary system on the required task-specific evaluation condition. The primary systems are expected, by the developers, to be their best performing systems. NIST’s analysis focuses on these primary systems.

The Rich Transcription Spring 2006 Evaluation plan [3] describes in detail the evaluation tasks, data sources, microphone conditions, system input and output formats, and evaluation metrics employed in the evaluation. This section summarizes the evaluation plan by discussing the meeting sub-domains in the test set, the audio input conditions, the evaluation task definitions, and the evaluation corpora details.

### **2.1 Meeting Sub-Domains: Conference Room vs. Lecture Room**

The meeting domain is highly variable along several dimensions. In the broad sense, any interaction between 2 more people may be considered to be a meeting. As such, meetings can range from brief informal exchanges to extremely formal proceedings

with many participants following specific rules of order. It is well known that the type, number, and placement of sensors have a significant impact on the performance of recognition tasks. The variability is so large that it would be impossible to build either a training or testing corpus that encompasses all of these factors. To make the problem tractable, the RT evaluations have attempted to focus efforts on two specific sub-domains: small conference room meetings (also occasionally referred to as “board room” meetings) and classroom-style lectures in a small meeting room setting. The two sub-domains are used to differentiate between two very different participant interaction modes as well as two different sensor setups. The RT-06S evaluation includes a separate test set for each of these two sub-domains, labeled “*confmtg*” and “*lectmtg*.”

In addition to differences in room and sensor configuration, the primary difference between the two sub-domains is in the group dynamics of the meetings. The RT conference meetings are primarily goal-oriented, decision-making exercises and can vary from moderated meetings to group consensus building meetings. As such, these meetings are highly-interactive and multiple participants contribute to the information flow and decisions made. In contrast, lecture meetings are educational events where a single lecturer briefs an the audience on a particular topic. While the audience occasionally participates in question and answer periods, it rarely controls the direction of the interchange or the outcome.

Section 2.4 describes the corpora used for both the *lectmtg* and *confmtg* domains in the RT-06S evaluation.

## 2.2 Microphone Conditions

Seven input conditions were supported for RT-06S. They were:

- Multiple distant microphones (MDM): This evaluation condition includes the audio from at least 3 omni-directional microphones placed (generally on a table) between the meeting participants.
- Single distant microphone (SDM): This evaluation condition includes the audio of a single, centrally located omni-directional microphone for each meeting. This microphone channel is selected from the microphones used for the MDM condition. Based on metadata provided with the recordings, it is selected so as to be the most centrally-located omni-directional microphone.
- Individual head microphone (IHM): This evaluation condition includes the audio recordings collected from a head mounted microphone positioned very closely to each participant’s mouth. The microphones are typically cardioid or super cardioid microphones and therefore the best quality signal for each speaker. Since the IHM condition is a contrastive condition, systems can also use any of the microphones used for the MDM condition.
- Multiple Mark III microphone arrays (MM3A): This evaluation condition includes audio from all the collected Mark III microphone arrays. A Mark III microphone arrays is a 64-channel, linear topology, digital microphone array[11]. The *lectmtg* dataset contains the data from each channel of one or two Mark-III microphone array per meeting.

- **Multiple source localization microphone arrays (MSLA):** This evaluation condition includes the audio from all the CHIL source localization arrays (SLA). An SLA is a 4-element digital microphone array arranged in an upside down ‘T’ topology. The lecture room meeting recordings include four or six SLAs, one mounted on each wall of the room.
- **All distant microphones (ADM):** This evaluation conditions permits the use of all distant microphones for each meeting. This condition differs from the MDM condition in that the microphones are not restricted to the centrally located microphones and the Mark III arrays and Source Localization arrays can be used. This condition was new for RT-06S.
- **Multiple beamformed signals (MBF):** This evaluation condition permits the use of the just the blind source separation-derived signals from the Mark-III arrays. This condition was new for RT-06S.

The troika of MDM, SDM, and IHM audio input conditions makes a very powerful set of experimental controls for black box evaluations. The MDM condition provides a venue for the demonstration of multi-microphone input processing techniques. It lends itself to experimenting with beamforming and noise abatement techniques to address room acoustic issues. The SDM input condition provides a control condition for testing the effectiveness of multi-microphone techniques. The IHM condition provides two important contrasts: first, it effectively eliminates the effects of room acoustics, background noise, and most simultaneous speech, and second it is most similar to the Conversational Telephone Speech (CTS) domain [1] and may be compared to results in comparable CTS evaluations.

### 2.3 Evaluation tasks

Three evaluation tasks were supported for the RT-05S evaluation: a speech-to-text transcription task and two diarization tasks: “Who Spoke When” and “Speech Activity Detection”. The following is a brief description of each of the evaluation tasks:

**Speech-To-Text (STT) Transcription:** STT systems are required to output a transcript of the words spoken by the meeting participants along with the start and end times for each recognized word. For this task, no speaker designation is required. Therefore, the speech from all participants is to be transcribed as a single word output stream.

Systems were evaluated using the Word Error Rate (WER) metric. WER is defined to be the sum of system transcription errors, (word substitutions, deletions, and insertions) divided by the number of reference words and expressed as a percentage. It is an error metric, so lowers scores indicate better performance. The score for perfect performance is zero. Since insertion errors are counted, it is possible for WER scores to exceed one hundred percent.

WER is calculated by first harmonizing the system and reference transcript through a series of normalization steps. Then the system and reference words are aligned using a Dynamic Programming solution. Once the alignment mapping between the system and reference words is determined, the mapped words are compared to classify

them as either correct matches, inserted system words, deleted reference words, or substituted system words. The errors are counted and statistics are generated.

The MDM audio input condition was the primary evaluation condition for the STT task for both meeting sub-domains. The *confmtg* data set also supported the SDM and IHM conditions. The *lectmtg* data supported the SDM, IHM, MSLA, MM3A, and MBF conditions. Participants could submit systems for the *confmtg* domain, the *lectmtg* domain, or both sub-domains.

For the RT-06S evaluation, the distant microphone systems were evaluated on speech including up to 4 simultaneous speakers. Previous evaluations ignored overlapping speech for these conditions. To compute these scores, the ASCLITE [9] module of the NIST Scoring Toolkit (SCTK) [5] was used.

**Diarization “Who Spoke When” (SPKR):** SPKR systems are required to annotate a meeting with regions of time indicating when each meeting participant is speaking and clustering the regions by speaker. It is a clustering task as opposed to an identification task since the system is not required to output a name for the speakers – only a generic id which is unique within the meeting excerpt being processed.

The Diarization Error Rate (DER) metric is used to assess SPKR system performance. DER is the ratio of incorrectly attributed speech time, (either falsely detected speech, missed detections of speech, or incorrectly clustered speech) to the total amount of speech time, expressed as a percentage. As with WER, a score of zero indicates perfect performance and higher scores indicate poorer performance.

In order to determine incorrectly clustered speech, the Hungarian solution to a bipartite graph<sup>1</sup> is used to find a one-to-one mapping between the system-generated speaker segment clusters and the reference speaker segment clusters. Once the mapping is found, speech time within a system speaker cluster not matching speech time in the mapped reference speaker cluster is classified as the incorrectly clustered speech.

New for 2006, the primary measure of DER was calculated for all speech including overlapping speech. This harmonizes the scores with the STT task which now includes the evaluation of overlapping speech.

Inherent ambiguities in pinpointing speech boundaries in time and annotator variability result in a small degree of inconsistency in the time annotations in the reference transcript. To address this, a 0.25 second “no score” collar is created around each reference segment. This collar effectively minimizes the amount of DER error due to reference annotation inconsistencies.

Another challenge is in determining how large a pause in speech must be to cause a segment break. Although somewhat arbitrary, a cutoff value of 0.3 seconds was empirically determined to be a good approximation of the minimum duration for a pause in speech resulting in an utterance boundary. As such, segments that are closer than 0.3 seconds apart are merged in both the reference and system output transcripts.

The MDM audio input condition was the primary evaluation condition for the SPKR task for both meeting sub-domains. The *confmtg* data supported one contrastive condition, SDM, and the *lectmtg* data supported four contrastive conditions:

---

<sup>1</sup> <http://www.nist.gov/dads/HTML/HungarianAlgorithm.html>

SDM, MSLA, MM3A, and MBF. Participants could submit systems for the *confmtg* domain, the *lectmtg* domain, or both the sub-domains.

**Diarization “Speech Activity Detection” (SAD):** SAD systems are required to annotate a meeting with regions of time indicating when at least one person is talking. The SAD task is therefore a simplified version of the SPKR task (because no speaker clustering is performed by the system). The task was introduced as an entry point for new participants in the RT evaluation series and to gauge the contribution of SAD errors to the SPKR and STT tasks.

The task was a dry run for the RT-05S evaluation but was considered a full evaluation task for RT-06S.

Since SAD is viewed as a simplification of the SPKR task, the SPKR DER scoring metric is also used to score the SAD task. The same no-score collar, 0.25 seconds, was applied during scoring and the same smoothing parameter, 0.3 seconds, was applied to the reference files. The reference files were derived from the SPKR reference files by simply merging the reference speaker clusters into a single cluster and then merging segments that either overlap or were within the 0.3 second smoothing parameter.

The MDM audio input condition was the primary evaluation condition for the SAD task for both meeting sub-domains. The *confmtg* data supported two contrastive conditions: SDM and IHM, and the *lectmtg* data supported five contrastive conditions: SDM, IHM, MSLA, MM3A, and MBF. Participants could submit system outputs for the *confmtg* domain, the *lectmtg* domain, or both sub-domains

The SAD task using IHM data is not directly comparable to SAD on distant microphone data, (i.e., MDM, SDM, MSLA, or MM3A data). An IHM channel includes both the wearer’s speech and cross-talk from other meeting participants. For the purposes of this evaluation, this cross talk is not considered detectable speech even though it was human generated. Therefore, an IHM SAD system has the challenging task of detecting the primary speaker’s speech and differentiating it from the cross-talk.

## 2.4 RT-06S Evaluation Corpora Details

As indicated previously, the RT-06S evaluation data consisted of two test sets: a conference room meeting test set and a lecture room meeting test set. The recordings were sent to participants as either down-sampled, 16-bit, 16Khz NIST Speech Header Resources (SPHERE) files or in the original 24-bit, 44.1 Khz WAV sample format as well as headerless raw files. The recordings of the meetings in the *confmtg* data set were distributed in their entirety while only the selected excerpts from the *lectmtg* data were distributed. All meeting recordings included video recordings. The video recordings were not distributed to the RT participants but they were distributed to the CLEAR evaluation participants.

**Conference Room Meetings:** The *confmtg* test set consisted of nominally 162 minutes of meeting excerpts from nine different meetings. NIST selected 18 minutes from each meeting to include in the test set. For two of the nine meetings, the ex-

cerpts were not contiguous. Five sites contributed data: the Augmented Multi-party Interaction (AMI) Project provided two meetings collected at Edinburgh University (EDI) and one meeting collected at TNO. Carnegie Mellon University (CMU), the National Institute of Standards and Technology (NIST), and Virginia Tech (VT) each contributed two meetings. The Linguistic Data Consortium (LDC) transcribed the test set according to the “Meeting Data Careful Transcription Specification - V1.2” guidelines [4]. Table 1 gives the salient details concerning the *confmtg* evaluation corpus.

Each meeting recording evaluation excerpt met minimum sensor requirements. Each meeting participant wore a head-mounted close talking microphone and there were at least three table-top microphones placed between the meeting participants. The dialects were predominately American English with the exception of the EDI meetings. In addition to these sensors, the EDI and TNO meetings included an eight-channel circular microphone array placed on the table between the meeting participants.

**Table 1** Summary of Conference Room Meeting evaluation corpus

Meeting ID	Duration (minutes)	Number of Participants	Notes
CMU_20050912-0900	17.8	4	Transcription team mtg.
CMU_20050914-0900	18.0	4	Transcription team mtg.
EDI_20050216-1051	18.0	4	Remote control design
EDI_20050218-0900	18.2	4	Remote control design
NIST_20051024-0930	18.1	9	Project planning mtg.
NIST_20051102-1323	18.2	8	Data resource planning
TNO_20041103-1130	18.0	4	Remote control design
VT_20050623-1400	18.0	5	Problem solving scenario
VT_20051027-1400	17.7	4	Candidate selection
Total		46	
Unique speakers		43	

**Lecture Room Meetings** The *lectmtg* test set consisted of 190 minutes of lecture meeting excerpts recorded at AIT, IBM, ITC-irst, and UKA. There were 38, 5-minute excerpts included in the evaluation from 26 recordings. Two types of lectures were recorded for the evaluation: “lectures” and “interactive lectures”. Both lecture types were technical language technology talks given by invited lecturers. The lectures involved a single lecturer and a large group of audience members: only a few of which wore head microphones. In contrast, the interactive lectures were smaller groups and included not only the recording of the lecture but also people entering the room and coffee breaks. All participants in the interactive lectures wore head microphones. While the coffee breaks and person movements were useful for CLEAR evaluation, the data was unlike the data used for previous RT evaluations.

The excerpts were selected and transcribed by ELDA. After the evaluation, CMU revised the transcripts to include speech only recorded on the table-top microphones.

During the revision, twelve of the excerpts were deemed to be of insufficient audio quality for the evaluation and removed from test set. This resulted in a 130-minute test set from 24 meetings.

The audio sensors used in the *lectmtg* data were configured differently than the *confmtg* data. Only the lecturer and two-to-four audience members, of potentially several, wore head-mounted, close-talking microphones. The rest of the audience was audible on the distant microphones. Three-to-four microphones were placed on the table in front of the lecturer and a fifth table-top microphone was placed in the corner of the room. Four-to-six source localization arrays were mounted on each of the four walls of the room. Finally, one or two Mark III arrays were placed directly in front of the lecturer.

### 3. Results of the RT-06S Evaluation

#### 3.1 RT-06S Evaluation Participants

The following table lists the RT-06S participants and the evaluation tasks each site took part in. In total, there were ten sites submitting system outputs.

**Table 2** Summary of evaluation participants and the tasks for which systems were submitted.

Site ID	Site Name	Evaluation Task		
		STT	SPKR	SAD
AIT	Athens Information Technology		X	X
AMI	Augmented Multiparty Interaction Program	X	X	X
IBM	IBM	X		X
ICSI/SRI	International Computer Science Institute and SRI International	X	X	X
INRIA	Institut National de Recherche en Informatique et en Automatic			X
ITC-irst	Center for Scientific and Technological Research			X
LIA	Laboratoire Informatique d'Avignon		X	X
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur	X	X	X
UKA	Karlsruhe University (UKA)	X		
UPC	Universitat Politècnica de Catalunya			X

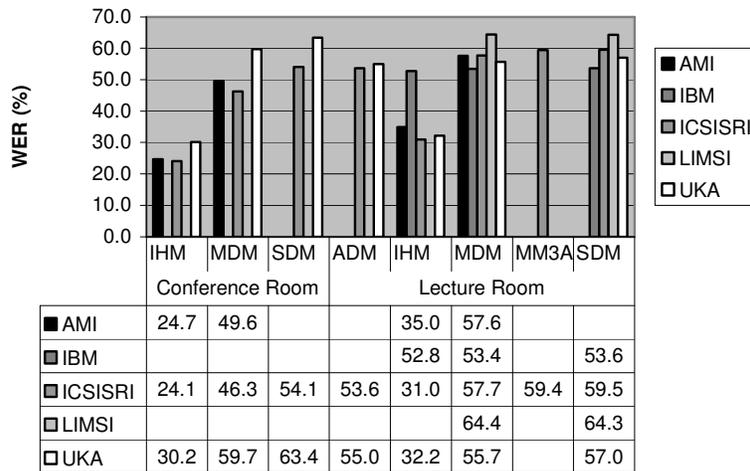
### 3.2 Speech-To-Text (STT) Results

Five sites participated in the STT task: AMI, IBM, ICSI/SRI, LIMSI, and UKA. This was the first year for IBM, LIMSI, and UKA. AMI, ICSI/SRI, and UKA submitted system outputs for the *confmtg* data while all sites submitted system outputs for the *lectmtg* data.

Figure 1 contains the results of all primary systems. The MDM WERs for *confmtg* data were 49.6, 46.3, and 59.7 for AMI, ICSI/SRI, and UKA respectively. The MDM WERs for the *lectmtg* data were 57.6, 53.4, 57.7, 64.4, and 55.7 for AMI, IBM, ICSI/SRI, LIMSI, and UKA. The *lectmtg* WERs for AMI and ICSI/SRI were 16% and 25% (relative) higher than *confmtg* data. However, UKA did 6% (relative) better on the *lectmtg* data. Last year, AMI and ICSI/SRI had higher error rates for the *lectmtg* data.

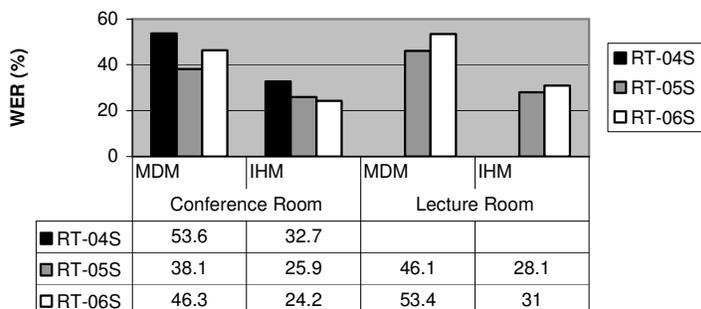
Unlike last year, the IHM error rates are higher for the *lectmtg* data: 41%, 28%, and 6% relative for AMI, ICSI/SRI and UKA respectively. One possible explanation for the increase is the dialect of the speakers. Many *lectmtg* speakers speak with strong, non-English accents, e.g., German- and Spanish-accented English.

A notable result from the *lectmtg* data was ICSI/SRI's 59.4% MM3A score. This result used the beamformed signal from built by UKA's Probabilistic Data Association Filters [10]. This was the first time in an RT evaluation that an automatic, blind source separation algorithm was applied to the output of Mark III arrays for use in an STT system.



**Figure 1. WERs for primary STT systems across test sets and audio input conditions. Up to 4 simultaneous speakers included in distant mic. conditions.**

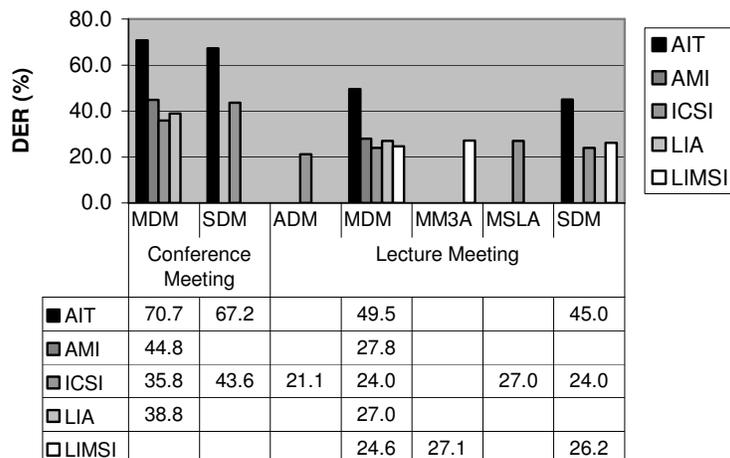
Figure 2 plots the historical error rates for the MDM and IHM conditions in both domains. There was a slight reduction in IHM WERs for the *confmtg* data. However, both MDM and IHM error rates were higher for the RT-06 *lectmtg* data.



**Figure 2. WERs for the best STT systems from RT-04S through RT-06S. MDM results are for segments with  $\leq 4$  active speakers while the IHM results include all speech.**

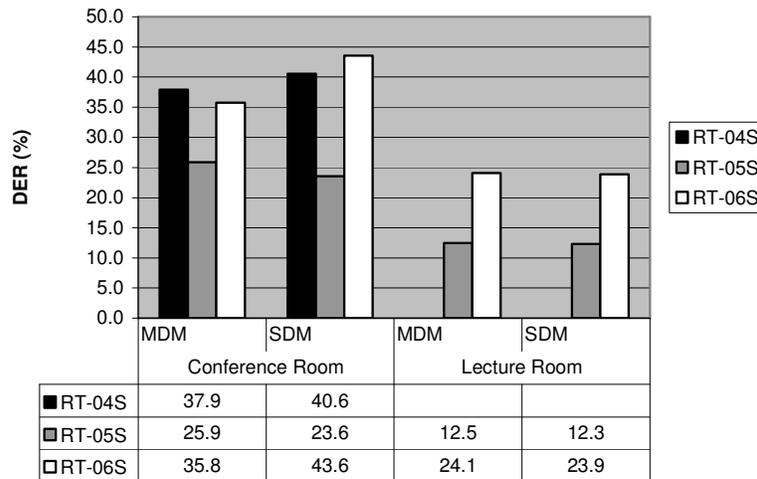
### 3.3 Diarization “Who Spoke When” (SPKR) Results

Five sites participated in the SPKR task: AIT, AMI, ICSI, LIA, and LIMSI. Of the five, only ICSI participated in the RT-05S SPKR evaluation. Figure 3 contains the results of all primary systems. The MDM DERs for *confmtg* data were 70.7, 44.8, 35.8, and 38.8 for AIT, AMI, ICSI, and LIA respectively. The MDM DERs for the *lectmtg* data were 49.5, 27.8, 24.0, 27.0, and 24.6 for AIT, AMI, ICSI, LIA, and LIMSI.



**Figure 3. DERs for the all speech for the primary SPKR systems across test sets and audio input conditions.**

The scores were appreciably higher than last year. Figure 4 contains the historical lowest error rates for each year. There are a couple of factors that may have attributed to the increase. First, these are all new systems. Second, the reference file generation continued to be problematic. Using human segmentation annotations are problematic in that consistency is hard to achieve. Future evaluations will use reference files derived from word-level forced alignments of reference transcription.

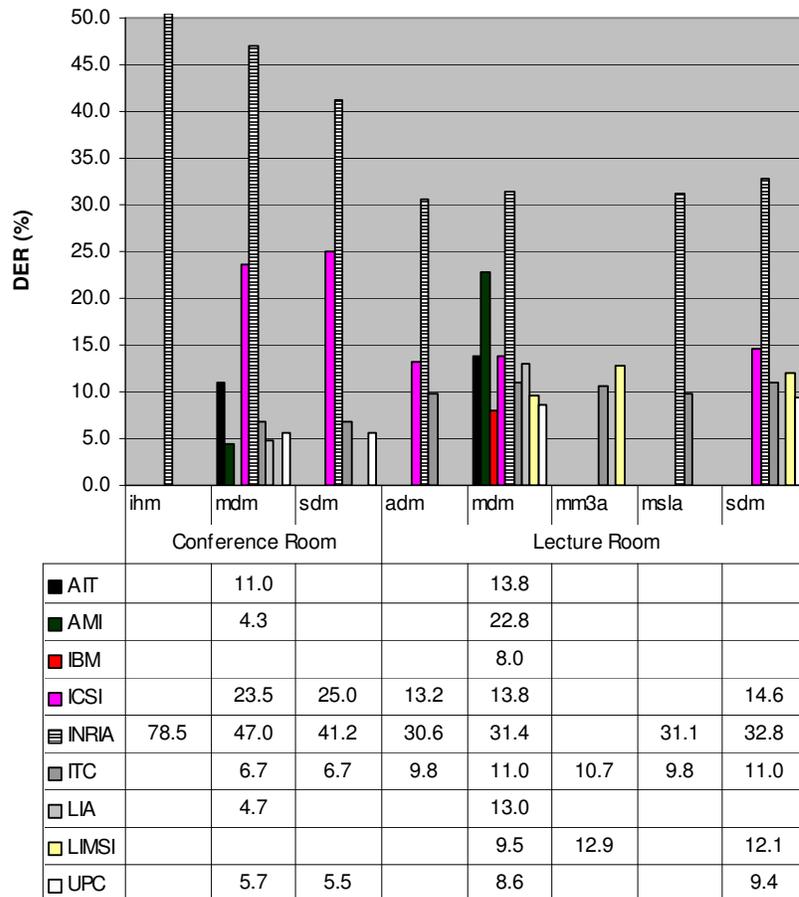


**Figure 4. DERs for the best MDM and SDM SPKR systems from RT-04S through RT-06S**

### 3.4 Diarization “Speech Activity Detection” (SAD) Results

Nine sites participated in this first formal evaluation SAD task during the RT evaluation: AIT, AMI, IBM, ICSI, INRIA, ITC-irst, LIA LIMSI, and UPC. Figure 5 shows the lowest MDM error rate for the *confmtg* data was achieved by AMI with a DER of 4.3%. For the *lectmtg* data, IBM had the lowest MDM DER of 8.0%. As with the other tasks, the *lectmtg* data had higher error rates than the *confmtg* data.

The SAD task will be continued in future evaluations since both the Dry Run in 2005 and the evaluation in 2006 were successful.



**Figure 5. DERs for primary SAD systems across test sets and audio input conditions**

#### 4.0 Conclusions and Future Evaluations

The collaboration between RT and the CLEAR evaluation as well as the AMI, CHIL and VACE programs has boosted the RT community on many levels. For the first time, the RT evaluation corpora has been annotated and used for the evaluation of both language and video processing/extraction tasks. The collaboration has also led to expanded task participation: almost twice the number of systems were built by the

participants even though the number of participating sites remained constant. We look forward to continued progress and evaluations in the meeting domain.

The RT-07 evaluation is being planned for the Spring of 2007. As with 2006, the same evaluation corpora will be used for both RT and CLEAR. In addition, the RT and CLEAR evaluation workshops will be co-located.

Applying blind source separation techniques RT is an exciting new direction for RT systems. We anticipate further sensor fusion will be possible as the CLEAR and RT communities are merged.

## 5.0 Acknowledgements

NIST would like to thank AIT, EDI, CMU, IBM, ITC, VT, UKA, TNO, and UPC for donating meeting recordings to the evaluation. Special thanks go to CMU, ELDA, and LDC for preparing reference transcriptions and annotations. The authors thank and appreciate the edits provided by Vince Stanford.

## 6.0 Disclaimer

These tests are designed for local implementation by each participant. The reported results are not to be construed, or represented, as endorsements of any participant's system, or as official findings on the part of NIST or the U. S. Government. Certain commercial products may be identified in order to adequately specify or describe the subject matter of this work. In no case does such identification imply recommendation or endorsement by NIST, nor does it imply that the products identified are necessarily the best available for the purpose.

## References

1. Fiscus et. al., "Results of the Fall 2004 STT and MDE Evaluation", RT-04F Evaluation Workshop Proceedings, November 7-10, 2004.
2. Garofolo et. al., "The Rich Transcription 2004 Spring Meeting Recognition Evaluation", ICASSP 2004 Meeting Recognition Workshop, May 17, 2004
3. Spring 2006 (RT-06S) Rich Transcription Meeting Recognition Evaluation Plan, <http://www.nist.gov/speech/tests/rt/rt2006/spring/>
4. LDC Meeting Recording Transcription, <http://www ldc.upenn.edu/Projects/Transcription/NISTMeet>
5. SCTK toolkit, <http://www.nist.gov/speech/tools/index.htm>
6. Michel et. al., "The NIST Meeting Room Phase II Corpus", 3<sup>rd</sup> Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI-06), 1-3, May 2006.
7. Fiscus et. al., "The Rich Transcription 2005 Spring Meeting Recognition Evaluation", The joint proceedings Rich Transcription Workshop and the 2nd Joint Work-

shop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI), 11-13 July 2005

8. <http://www.clear-evaluation.org/>
9. Fiscus et. al., "Multiple Dimension Levenshtein Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech", LREC 2006: Sixth International Conference on Language Resources and Evaluation
10. Gehrig and McDonough, "Tracking Multiple Simultaneous Speakers with Probabilistic Data Association Filters", 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI-06)
11. Stanford, V.: The NIST Mark-III microphone array - infrastructure, reference data, and metrics. In: Proceedings International Workshop on Microphone Array Systems - Theory and Practice, Pommersfelden, Germany (2003)
12. [http://isl.ira.uka.de/clear06/downloads/ClearEval\\_Protocol\\_v5.pdf](http://isl.ira.uka.de/clear06/downloads/ClearEval_Protocol_v5.pdf)