The Rich Transcription 2007 Meeting Recognition Evaluation

Jonathan G. Fiscus[1], Jerome Ajot[1,2], and John S. Garofolo[1]

[1] National Institute Of Standards and Technology, 100 Bureau Drive Stop 8940,
Gaithersburg, MD 20899
[2] Systems Plus, Inc., One Research Court – Suite 360, Rockville, MD 20850
{jfiscus,ajot,jgarofolo}@nist.gov

Abstract. We present the design and results of the Spring 2007 (RT-07) Rich Transcription Meeting Recognition Evaluation; the fifth in a series of community-wide evaluations of language technologies in the meeting domain. For 2007, we supported three evaluation tasks: Speech-To-Text (STT) transcription, "Who Spoke When" Diarization (SPKR), and Speaker Attributed Speech-To-Text (SASTT). The SASTT task, which combines STT and SPKR tasks, was a new evaluation task. The test data consisted of three test sets: Conference Meetings, Lecture Meetings, and Coffee Breaks from lecture meetings. The Coffee Break data was included as a new test set this year. Twenty-one research sites materially contributed to the evaluation by providing data or building systems. The lowest STT word error rates with up to four simultaneous speakers in the multiple distant microphone condition were 40.6 %, 49.8 %, and 48.4 % for the conference, lecture, and coffee break test sets respectively. For the SPKR task, the lowest diarization error rates for all speech in the multiple distant microphone condition were 8.5 %, 25.8 %, and 25.5 % for the conference, lecture, and coffee break test sets respectively. For the SASTT task, the lowest speaker attributed word error rates for segments with up to three simultaneous speakers in the multiple distant microphone condition were 40.3 %, 59.3 %, and 68.4 % for the conference, lecture, and coffee break test sets respectively.

1 Motivation

The National Institute of Standards and Technology (NIST) has worked with the speech recognition community since the mid 1980s to improve the state-of-the-art in speech processing technologies. [1] To facilitate progress, NIST has worked with the community to make training/development data sets 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 facilitate technical exchange and track progress trends over time. The test suites used in these benchmark tests become development resources after the formal evaluations.

R. Stiefelhagen et al. (Eds.): CLEAR 2007 and RT 2007, LNCS 4625, pp. 373–389, 2008.

In 2001, NIST began administering the Rich Transcription Evaluation series for the DARPA Effective, Affordable, Reusable, Speech-to-Text (EARS) Program in the Broadcast News (BN) and Conversation Telephone Speech (CTS) domains. The EARS community focused on building technologies to generate transcriptions of speech that are fluent, informative, readable by humans, and usable in downstream processes. To accomplish this, EARS technologies produced transcripts consisting of words and non-orthographic metadata. We refer to these metadata enriched transcripts as "rich transcriptions." While the metadata can take many forms, the EARS program worked on three main forms: which speakers spoke which words, syntactic boundaries, and dysfluent speech detection.

In 2002, the community began investigating the meeting domain as a new evaluation domain because the error rates on BN material approached 6 times that of human performance indicating the community needed a more difficult challenge problem. When error rates come close to human performance, the evaluation costs rise dramatically because transcription ambiguity in the reference becomes a disproportionately large component of the error rates. While large test sets and/or meticulously scrutinized reference transcripts can ameliorate the impact of ambiguity, they both require great expense. Instead, research in the meeting domain became a popular because it provides a unique environment to collect naturally occurring spoken interactions under controlled sensor conditions that presents several challenges to the technologies resulting in higher error rates. These include varied fora, an infinite number of topics, spontaneous highly interactive/overlapping speech, varied recording environments, varied/multiple microphones, multi-modal inputs, participant movement, and far field speech effects such as ambient noise and reverberation.

At roughly the same time in the early 2000's, a number of independent large-scale programs included the meeting domain as a component of their research and evaluation efforts. The programs included the European Union (EU) Computers in the Human Interaction Loop (CHIL), the EU Augmented Multiparty Interaction with Distant Access (AMIDA) program, and the US Video Analysis and Content Extraction (VACE) program. The programs shared many aspects of uni-modal (audio or video) and multi-modal (audio+video) research indicating a strong movement was underway in the research community to focus on building and experimenting with multi-modal technologies. However, little infrastructure was in place to support the research nor was there a public evaluation-based forum for technical interchange.

Beginning in 2005, CHIL, NIST, and VACE orchestrated a multi-year plan to bring together the disjoint speech and video processing communities through common evaluations. In 2006, the CHIL and VACE programs started the Classification of Events, Activities, and Relationships (CLEAR) evaluation [7, 14]. While the 2006 CLEAR Evaluation Workshop was held in conjunction with the 7th IEEE International Conference on Face and Gesture Recognition (FG2006), the shared use of common evaluation corpora for the RT and CLEAR evaluations in 2006 set the stage for the joint CLEAR and RT evaluations and workshops in 2007. [15]

The Rich Transcription 2007 (RT-07) Meeting Recognition evaluation, which was part of the NIST Rich Transcription (RT) series of language technology evaluations [1] [2] [6] [10], included three evaluation tasks:

- Speech-To-Text (STT) transcription – Transcribe the spoken words.
- "Who Spoke When" Diarization (SPKR) – Detect segments of speech and cluster them by speaker.
- Speaker Attributed Speech-To-Text (SASTT) – Transcribe the spoken words and associate them with a speaker.

The first two tasks, STT and SPKR, are component tasks that have always been include in the RT evaluations. The SASTT is a composite task that includes both STT and SPKR tasks. The RT-07 evaluation was the first evaluation to include the SASTT task although ICSI/SRI experiments conducted during the EARS Program [13] were very similar to the presently defined task.

The RT-07 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 with Distant Access (AMIDA) Program
- The Computers in the Human Interaction Loop (CHIL) Program
- Carnegie Mellon University (CMU)
- Edinburgh University (EDI)
- Evaluations and Language Resources Distribution Agency (ELDA)
- IBM
- International Computer Science Institute (ICSI)
- Infocomm Research Site (I2R)
- Nanyang Technological University (NTU)
- SRI International (SRI)
- 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)
- Sheffield University
- Netherlands Organisation for Applied Scientific Research (TNO)
- Universitat Politècnica de Catalunya (UPC)
- Virginia Tech (VT)

The RT-07 evaluation made use of three test sets: Conference Meetings, Lecture Meetings, and Coffee Breaks from Lecture Meetings. The multiple test sets fostered collaboration by sharing data across programmatic boundaries while accommodating the needs of individual programs and by promoting cross-disciplinary interchange via shared corpora.

2 Rich Transcription 2007 Meeting Recognition Evaluation

The RT-07 evaluation was similar to previous RT evaluations except for three changes: the addition of the Speaker Attributed Speech-To-Text task, the deletion of

Speech Activity Detection (SAD) task, and the addition of Coffee Break excerpts as a new test set.

All participating teams were required to submit a single primary system on the required task-specific evaluation condition. Developers selected their primary systems based on their efforts to build their best performing system. NIST's analysis focuses on these primary systems.

The Rich Transcription Spring 2007 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 test sets for the meeting sub-domains, the audio input conditions, the evaluation task definitions, and the evaluation corpora details.

2.1 Meeting Sub-domains and Test Sets

The meeting domain is highly variable along several dimensions. Meetings, which are verbal interactions between two or more people, range from brief informal exchanges to extremely formal proceedings with many participants following specific rules of order. However, the variability is so large that it would be impossible to build either training or testing corpora that encompasses all of these factors. Therefore, the RT evaluations have focused efforts on narrowly defined meeting sub-domains to make the problem tractable. The RT-07 evaluation material included data from two meeting sub-domains: small conference room meetings (also occasionally referred to as "board room" meetings) and interactive lectures in a small meeting room setting.

The two sub-domains represent two different participant interaction modes as well as sensor setups. The primary difference between the two sub-domains is in the group dynamics of the meetings. The first sub domain, conference meetings, consists of 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. In contrast, the second sub-domain, lecture meetings, consists of educational events where a single lecturer briefs an audience on a particular topic. While the audience occasionally participates in question and answer periods, the lecturer predominately controls the meeting.

The RT-07 evaluation included three test sets: the conference room meeting test set (*confmtg*), the lecture room meeting test set (*lectmtg*), and the coffee break (*cbreak*) test set. The *confmtg* and *lectmtg* data sets are "similar" to previous test sets because the data selection protocol did not change. The *cbreak* data consisted of excerpts selected from Lecture Meetings where the participants took a coffee break during the recording.

The recordings were sent to participants as either down-sampled, 16-bit, 16 KHz, NIST Speech Header Resources (SPHERE) files, the original 24-bit, 44.1 KHz WAV files, or headerless raw files. [12] further documents the *confmtg* data set. [11] further documents the *lectmtg* data set.

Conference Room Meetings: The *confmtg* test set consisted of nominally 190 minutes of meeting excerpts from eight different meetings. NIST selected 22.5 minutes from each meeting to include in the test set. Four sites contributed two meeting recordings for eight total meetings. The four sites were Edinburgh University (EDI), Carnegie Mellon University (CMU), the National Institute of Standards and Technology (NIST), and Virginia Tech (VT). The Linguistic Data Consortium (LDC) transcribed the test set according to the "Meeting Data Careful Transcription Specification - V1.2" guidelines [4], [12]. Table 1 gives the salient details concerning the *confmtg* evaluation corpus.

Each meeting recording met minimum sensor requirements. All meeting participants 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 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_20061115-1030	22.5	4	Discussion group
CMU_20061115-1530	22.6	4	Transcription team mtg.
EDI_20051113-1500	22.6	4	Remote control design
EDI_20051114-1500	22.7	4	Remote control design
NIST_20051104_1515	22.4	4	Planning meeting
NIST_20060216-1347	22.5	6	SWOT analysis mtg.
VT_20050408-1500	22.4	5	Problem solving scenario
VT_20050425-1000	22.6	4	Problem solving scenario
Total		35	

Lecture Room Meetings: The *lectmtg* test set consisted of 164 minutes of lecture meeting excerpts recorded at AIT, IBM, ITC-irst, UKA, and UPC. CMU selected and transcribed 32, 5-minute excerpts for the test set from 20 different meeting recordings [11]. The lectures were the CHIL "interactive lectures." The lectures involved fewer people, 3-7 participants, and contained more interactivity than the RT-06 *lectmtg* test set. The excerpts selected for the *lectmtg* test set were from the core of the meeting when either the lecturer was speaking or the lecture was answering questions.

The *lectmtg* data included more audio sensors that the confmtg data. They included four-to-six source localization arrays mounted on each of the four walls of the room, and one or two Mark III arrays mounted near the lecturer.

Coffee Break Meetings: The *cbreak* test set consisted of 41 minutes of lecture meeting excerpts recorded at AIT, IBM, ITC-irst, UKA, and UPC. CMU selected and transcribed eight, 5-minute excerpts for the test set from eight different meeting recordings [11]. The data, which came from the same meetings as the *lectmtg* data, consisted of the coffee break periods when the lecturer took a brief break from the presentation and participants stood up to walk around the room and get coffee. The

CLEAR evaluation developed the *cbreak* data set as a more challenging video processing meeting data set than the typical lecture meeting videos. While the community at large wanted to build multi-modal data sets, the RT community decided the *cbreak* data did not conform to previously used *lectmtg* test sets. Therefore, the RT community decided to make the coffee break material a separate test set rather than drastically change the makeup of the RT-07 *lectmtg* test set compared to previous years.

2.2 Microphone Conditions

The RT-07 evaluation supported seven input conditions. They were:

- Multiple distant microphones (MDM): This evaluation condition includes the audio from at least three 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 from the set of MDM microphones. Metadata provided with the meetings supplies the information to select the 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 of 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.
- Individual head microphone plus reference segmentation (IHM+REFSEG): This evaluation condition used the IHM audio and reference speech/non-speech segmentations. This evaluation condition is a laboratory control condition. The intent of the IHM condition was to provide clean, near field speech. However, the IHM recordings can at times include a significant amount of cross talk that requires significant effort to ameliorate. This condition provides developers with the opportunity to process clean, near field speech without the need to implement cross talk rejection.
- 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 [18]. The lecture domain data 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 domain data includes four or six SLAs mounted on the walls 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 but rather all microphones including the Mark III arrays and Source Localization arrays.

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 reduces 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

The RT-07 evaluation supported three evaluation tasks: the Speech-To-Text transcription task, the "Who Spoke When" Diarization Task, and the Speaker Attributed Speech-To-Text task. The following is a brief description of each of the evaluation tasks:

Speech-To-Text (STT) Transcription: STT systems output a transcript containing all of the words spoken by the meeting participants. For each word recognized by the system, the system outputs the word's orthography along with the word's start/end times and confidence score. For this task, the system outputs a single stream of words since no speaker designation is required.

The primary metric is Word Error Rate (WER). WER is the sum of 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 lower scores indicate better performance. The score for perfect performance is zero. WER scores can exceed one hundred percent since the metric includes insertion errors.

The scoring process consists of three steps: transcript normalization, segment group chunking to reduce alignment computations, and word alignment.

The process for text normalization includes many steps including spelling variant normalization, contraction expansion, optional words, etc. See the evaluation plan for a detailed enumeration of the text normalizations.

The segment group chunking splits a recording into independent units for alignment based on reference speaker segment times. Figure 1 is an example of segment group chunking which shows four segment groups. The number of active speakers in a segment group defines the "Overlap Factor" (OV) of the segment group. The overlap factor is not a measure of absolute speaker overlap (e.g., by time); rather it is a method for counting the dimensions necessary to compute a word alignment. Segment group chunking is consistent across systems; therefore, segment groups provide an effective way to bound computation and score subsets of the recordings consistently across systems. The final step in segment group chunking is to collect the system words whose time midpoints are within the span of a segment group. Each segment group along with the system output words assigned to it form an independent unit for the alignment engine.

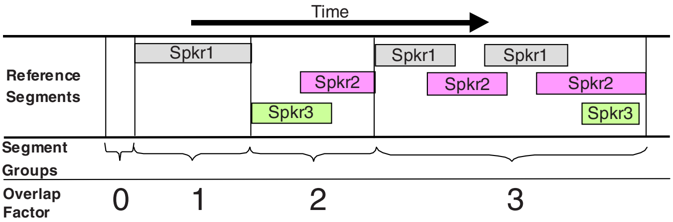


Fig. 1. Example segment group chunking analysis

The final scoring step is to align the references and the system output in order to count errors. An alignment is a one-to-one mapping between system and reference words that minimizes the edit distance to convert the system transcript into the reference transcript. NIST used the multi-dimensional, Dynamic Programming solution to sequence alignment found in the ASCLITE tool [8] of the SCTK package [5] to perform the alignment and scoring. The alignments are computationally expensive, $O(N^{\#S+\#R})$ where *N* is the number of words per speaker, *#S* is the number of system speakers, and *#R* is the number of reference speakers. The STT systems do not differentiate speakers therefore *#S* for the STT task is 1. To reduce the computational burden, several techniques discussed in [8] minimize the computational requirements.

The MDM audio input condition was the primary evaluation condition for the STT task for all test sets. The results reported for all distant microphone conditions include segment groups with up to and including overlap factor 4 ($\mathrm{WER}_{(\mathrm{OV}\leq 4)}$). Standardizing on OV≤4 was empirically determined to be a reasonable cutoff balancing evaluated material vs. the required computational resources for alignment.

Diarization "Who Spoke When" (SPKR): SPKR systems 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 speaker name or identify each speaker from a gallery – only a generic id that is unique within the processed meeting excerpt.

The Diarization Error Rate (DER) is the primary metric. DER is the ratio of incorrectly attributed speech time, (falsely detected speech, missed detections of speech, and 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.

Incorrectly clustered speech, a speaker error, occurs when a system successfully detects speech but attributes the speech to the wrong speaker. Since the system generates its own clusters and there is no a priori connection between the system and reference speaker clusters, correct speaker attribution is determined by finding a minimal cost, one-to-one mapping between the system speaker clusters and reference speaker clusters using the Hungarian solution to a bipartite graph [16]. This "speaker mapping" is the basis for determining which system speaker is correct – the mapped system/reference speakers are correct.

Preparing reference segment boundaries for the evaluation is an inherently difficult human annotation task because of the ambiguities in pinpointing speech boundaries. Ambiguities include time padding for segment-initial plosives, differentiating independent adjacent segments and single segments, and others. Instead of building arbitrary rules for annotators to follow, the evaluation infrastructure accommodates the variability with three techniques. First, the evaluation tool does not score system performance within 0.25 seconds of each reference segment boundary. This "no score" collar minimizes the amount of DER error due to segment boundary inconsistencies. Second, adjacent reference segments are merged if they are within 0.3 second of each other. Although somewhat arbitrary, 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. Finally, the process for constructing the reference segments changed for RT-07. Instead of relying on human segmentations, the reference segment times were derived from automatically computed word occurrence times. NIST used the LIMSI speech recognition tools to align the reference transcript to the speech signals thus generating the word time locations. Using these "forced word alignments," construction of the reference segments consisted of converting each word into a segment and then smoothing the segments with the 0.3 second smoothing parameter.

The MDM audio input condition was the primary evaluation condition for the SPKR task for a test sets.

Speaker Attributed Speech-To-Text (SASTT): SASTT systems output a transcript containing all of the words spoken during a meeting and attributing each word to a single speaker. The SASTT task is a joint technology development task that combines both Diarization "Who Spoke When" and Speech-To-Text technologies into a single task.

Speaker Attributed Word Error Rate (SWER) is the primary evaluation metric. SWER is the sum of transcription errors, (word substitutions, word deletions, word insertions, and speaker substitutions) divided by the number of reference words, and expressed as a percentage. WER and SWER are closely related – SWER has an additional error type, "speaker substitutions" (SpSub). Speaker substitutions are correctly recognized words attributed to the incorrect speaker. SWER is an error metric, so lowers scores indicate better performance. The score for perfect performance is zero. SWER scores can exceed one hundred percent since the metric includes insertion errors.

The SASTT scoring process, which is very similar to the STT scoring process, consists of four steps: transcript normalization, speaker mapping, segment group chunking, and word alignment. The transcript normalization and segment group chunking steps are identical to the processes used for STT scoring. The speaker mapping step is an additional step for SASTT scoring and the word alignment process is slightly different for SASTT scoring.

As stated prpeviously, SASTT systems must accurately attribute each word to a speaker. The SASTT scorer uses the SPKR evaluation tool to generate a system-to-reference speaker-mapping list that serves as the definition of a correct speaker: the correct reference speaker for a system speaker is the reference speaker mapped to it. The word alignment process uses the speaker correctness information to determine when speaker substitutions occur. We used ASCLITE's [8] Multiple System Stream-to-Multiple Reference Stream alignment capabilities to compute the word alignments.

Like the STT evaluation, the MDM audio input condition is the required condition. Unlike STT, however, the results reported for all distant microphone conditions include segment groups with up to and including overlap factor 3 (OV≤3). The additional computation burden proved too great to compute overlap factor 4 in reasonable time and with complete coverage. This is because the number of number of system speakers in a segment group can be greater than one. As an example, overlap factor 4 scoring for the AMI SASTT system would require 272 TB of memory search space to complete.

3 Results of the RT-07 Evaluation

3.1 RT-07 Evaluation Participants

The following table lists the RT-07 participants and the evaluation tasks for which they built systems.

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

Site ID	Site Name	Evaluation Task		
		SPKR	STT	SASTT
AMI	Augmented Multiparty Interaction with Distance Access	X	X	X
I2R/NTU	Infocomm Research Site and Nanyang Technological University	X		
IBM	IBM	X	X	X
ICSI	International Computer Science Institute	X		
LIA	Laboratoire Informatique d'Avignon	X		
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur	X		X
SRI/ICSI	International Computer Science Institute and SRI International		X	X
UKA	Karlsruhe University (UKA)		X	
UPC	Universitat Politècnica de Catalunya	X		

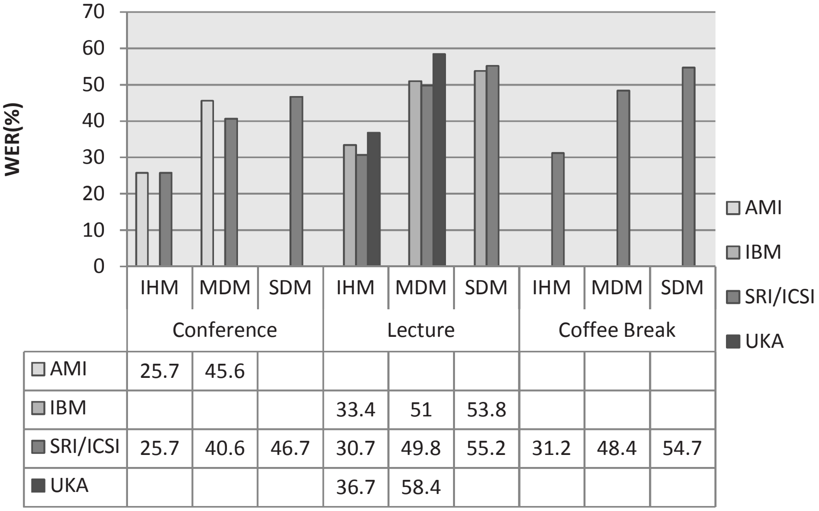
3.2 Speech-to-Text (STT) Results

Four sites participated in the STT task: AMI, IBM, SRI/ICSI, and UKA. Figure 2 contains the results of all primary systems.

The WER$_{(OV\leq 4)}$s for the MDM audio input condition for the *confmtg* data were 45.6 % and 40.6 % for AMI and SRI/ICSI respectively. The coverage of scoreable meetings for segment groups with OV≤4 was 99.3 %. The differences are significant at the 95 % confidence level using the Matched Pairs Sentence-Segment Word Error (MAPSSWE) test [17].

The WER$_{(OV\leq 4)}$s for the MDM audio input condition on the *lectmtg* data were 51.0 %, 49.8 %, and 58.4 % for IBM, SRI/ICSI, and UKA respectively. The coverage of scoreable meetings for segment groups with OV≤4 was 99.6%. All differences are significant according to the MAPSSWE test.

Only SRI/ICSI submitted outputs for the *cbreak* data. Their WER$_{(OV\leq 4)}$s for the MDM condition was 48.4, which was 2.8 % (relative) lower than their WER for the *lectmtg* data. While the error rate was lower for the *cbreak* data, it was not a significant different based a 2-Sample T-Test at the 95 % confidence level.

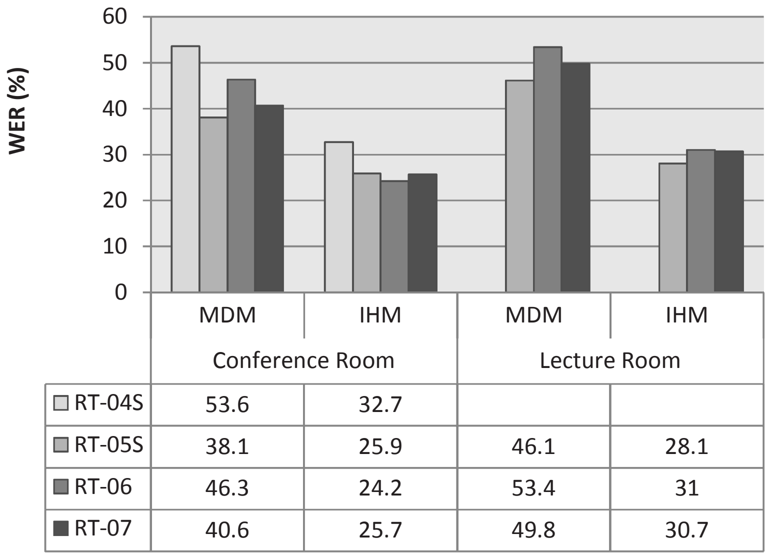


	IHM	MDM	SDM	IHM	MDM	SDM	IHM	MDM	SDM
	Conference			Lecture			Coffee Break		
□ AMI	25.7	45.6							
▨ IBM				33.4	51	53.8			
▩ SRI/ICSI	25.7	40.6	46.7	30.7	49.8	55.2	31.2	48.4	54.7
■ UKA				36.7	58.4				

Fig. 2. WERs for primary STT systems across test sets and audio input conditions. Overlap Factor 4 and less included in distant microphone conditions.[1]

Figure 3 plots the historical error rates for the MDM and IHM conditions in both domains. For the *confmtg* data, the MDM error rate was 12 % lower than the same condition for '06, but the IHM error rate was 6% higher. For the *lectmtg* data, the MDM error rates dropped 7 % relative while the IHM error rate had no change. Figure 4 sets the *confmtg* '07 results in the context of previous NIST STT evaluations. As evident from the graph, the meeting domain continues to be the most difficult actively researched domain for STT systems.

[1] All of SRI/ICSI's submissions were late accept their submissions for the IHM conditions.



	Conference Room MDM	Conference Room IHM	Lecture Room MDM	Lecture Room IHM
☐ RT-04S	53.6	32.7		
▨ RT-05S	38.1	25.9	46.1	28.1
▨ RT-06	46.3	24.2	53.4	31
■ RT-07	40.6	25.7	49.8	30.7

Fig. 3. WERs for the best STT systems from RT-04S through RT-06S. MDM results are for segment groups with OV≤4 while the IHM results include all speech.

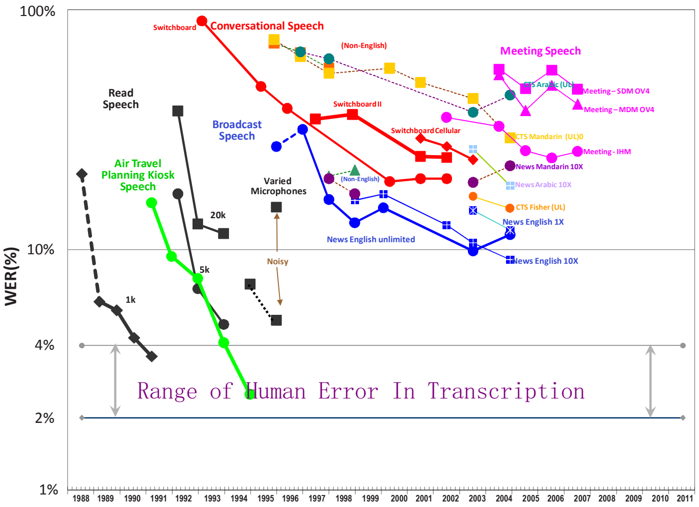
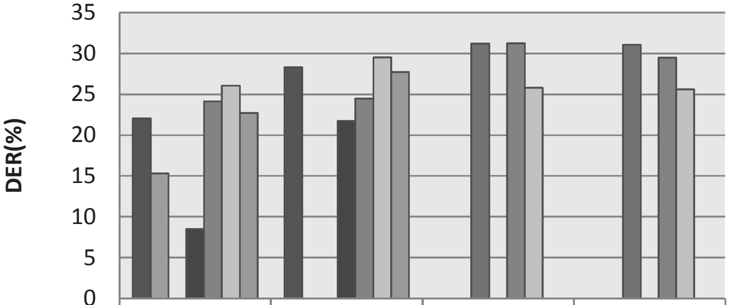


Fig. 4. STT Benchmark Test History 1988-2007

3.3 Diarization "Who Spoke When" (SPKR) Results

Eight sites participated in the SPKR task: AMDA, I2R, IBM, ICSI, LIA, LIMSI, NTU, and UPC. I2R and NTU collaborated to build a single system. All participants except IBM submitted *confmtg* systems. IBM, LIA, and LIMSI submitted *lectmtg* systems. Figure 5 contains the results of all primary systems. The lowest MDM DERs were 8.5%, and 25.8% for the *confmtg* and *lectmtg* test sets respectively.

The *lectmtg* scores for most systems at the same performance level as the previous year's, however, the 8.5% DER achieved by ICSI is very low. It is half of the closest system and a roughly a third of the rest of the systems. Further, Figure 5 the contains lowest error rate for each of the previous evaluations and shows the result was 76% relative lower than last year's best system.



	MDM	SDM	MDM	SDM
	Conference		Lec ture	
■ AMIDA	22.03	28.32		
■ I2R/NTU	15.32			
■ IBM			31.22	31.06
■ ICSI	8.51	21.74		
■ LIA	24.16	24.49	31.23	29.48
□ LIMSI	26.07	29.52	25.81	25.6
■ UPC	22.7	27.72		

Fig. 5. DERs for the primary SPKR systems across test sets and audio input conditions

Extensive discussions of the ICSI *confmtg* results occurred during the workshop. Table 3 compares the performance of the primary *confmtg* systems for differentiating speech vs. non-speech and for the system's ability to cluster speakers.

To evaluate speech/non-speech detection, we scored each SPKR submission as if it were a Speech Activity Detection (SAD) system as defined for the RT-06 SAD evaluation methodology [10]. To evaluate the system's ability cluster speakers, we computed both the average number of system speakers per meeting and the number of meetings with the correct number of speakers. The ICSI system had the second lowest SAD DER score of 3.33% with LIMSI having the lowest at 3.23%. The actual average number of speakers for the *confmtg* test set is 4.4. Six of the eight meetings had 4 speakers, one meeting had 5 speakers, and one meeting had 6 speakers. ICSI had nearly the right average and they correctly predicted the number of speakers in 7 of

Table 3. Primary SPKR system performance comparing DET to speech activity detection and speaker count prediction

Site ID	SPKR DER	SAD DER	Avg. Nbr. Sys. Speakers	Mtgs. With Correct Nbr. Speakers
ICSI	8.51	3.33	4.5	87.5%
I2R/NTU	15.32	8.65	4.4	75.0%
UPC	22.70	5.39	3.9	25.0%
LIA	24.16	3.69	4.9	12.5%
LIMSI	26.07	3.23	12.3	12.5%
AMIDA	22.03	6.73	7.1	0%

the 8 meetings. The combination of good SAD performance and accurate clustering led to ICSI's low overall SPKR DER performance. Other sites did well at one of the aspects, but not both.

The selection protocol for *confmtg* test sets will change in future evaluations. The lack of variability in the number of speakers per test excerpt is not adequately testing the SPKR systems. Next year there will be more excerpts to improve the statistical reliability of the performance estimates. Additionally, next year there will be a wider variety in the number of speakers per excerpt to test the system's ability to predict the correct number of speakers over a broader range of meeting participants.

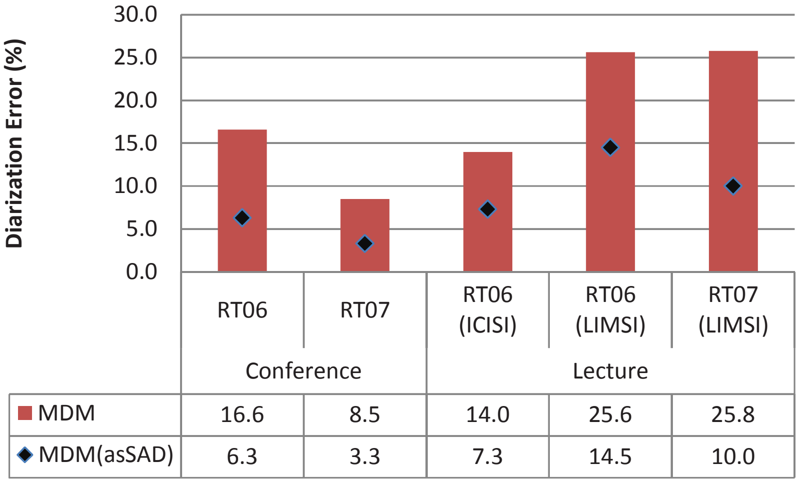


Fig. 6. DERs for the best MDM SPKR systems from RT-06 and RT-07 scored against forced alignment mediated references

Figure 5 contains the historical lowest error rates for each year when scored against forced alignment mediated references. As mentioned earlier, the SPKR error rates for the *confmtg* data dropped. The SPKR DER for the *lectmtg* data remained flat when comparing this year's LIMSI system to last year's LIMSI system. However, the LIMSI SAD DER was lower in '07.

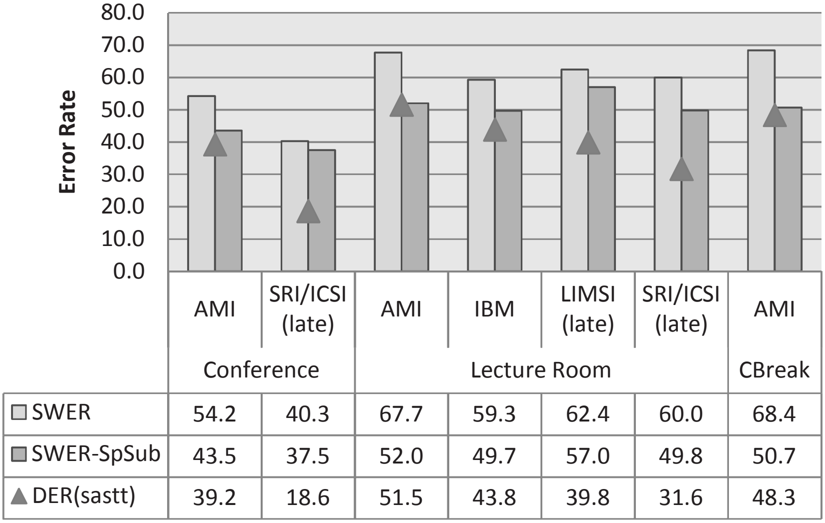
Only LIA submitted SPKR results for the Coffee Break data. The DER for their primary, MDM audio condition system on *cbreak* data was 25.5% compared to 31.2% for the same system on the *lectmtg* data. The SAD scores for the LIA system were 7.38 and 9.34 for the *cbreak* and *lectmtg* data respectively. While the error rates for the *cbreak* data are lower, the average SPKR DER by meeting excerpt are not statistically different at the 95% confidence level using a 2-sample T-Test.

3.4 Speaker Attributed Speech-to-Text (SASTT) Results

Five sites participated in the Speaker Attributed Speech-To-Text task: AMI, IBM, LIMSI, SRI/ICSI, and UKA. Figure 6 contains the results of all primary systems on the *cbreak*, *confmtg*, and *lectmtg* data for segment groups with OV≤3.

SRI/ICSI had the lowest $\text{SWER}_{(\text{OV}\leq3)}$ of 40.3 % on the *confmtg* data which was statistically different, according to the MAPSSWE test, than AMI's 54.2 %. The coverage of scoreable meetings for segment groups with OV≤3 was 84.5 %.

IBM had the lowest $\text{SWER}_{(\text{OV}\leq3)}$ of 59.3 % on the *lectmtg* data which was not statistically different, according to the MAPSSWE test, than SRI/ICSI's 60.0 %. The rest of the inter-system comparisons on the lecture data were statistically different according to the MAPSSWE test. The coverage of scoreable meetings for segment groups with OV≤3 was 97 %.



	AMI	SRI/ICSI (late)	AMI	IBM	LIMSI (late)	SRI/ICSI (late)	AMI
	Conference		Lecture Room				CBreak
□ SWER	54.2	40.3	67.7	59.3	62.4	60.0	68.4
▨ SWER-SpSub	43.5	37.5	52.0	49.7	57.0	49.8	50.7
▲ DER(sastt)	39.2	18.6	51.5	43.8	39.8	31.6	48.3

Fig. 7. RT-07 Primary SASTT system performance on the MDM condition scored without speaker substitutions and as diarization systems

AMI was the only participant to run their system on the *cbreak* data and achieved a 68.4 % $\text{SWER}_{(\text{OV}\leq3)}$.

The novel aspect of the SASTT task is to combine speaker diarization and STT systems. Figure 6 presents two data points in order to separate the errors due to diarization. The SWER-SpSub bar is the SWER minus the Speaker Substitution rate. The distance between the height of SWER and SWER-SpSub bars indicates the affect of speaker diarization errors on system performance. The second data point is DER(sastt) which is the diarization error using inferred speaker segment boundaries from the SASTT system output[2]. $\text{DER}_{(\text{sastt})}$ is not equivalent to the DER for SPKR systems, but it does correlate with Speaker Substitution errors. The notable exception is the LIMSI system where their SpSub errors were relatively low given their high $\text{DER}_{(\text{sastt})}$: this is because a large component of the LIMSI's diarization system DER is speech activity detection rather than speaker identification.

This initial evaluation of SASTT systems was a success in that developers built combined systems and the evaluation infrastructure was able to evaluate their performance. Unfortunately, none of the fielded SASTT systems for the 2007 evaluation jointly optimized their STT and SPKR systems, so one would expect future research to include joint optimization to improve error rates.

4 Conclusions and Future Evaluations

The 5th RT evaluation occurred during the 1st half of 2007. In order to promote multimodal research, the RT and CLEAR evaluations shared development/evaluation corpora and collocated their evaluation workshops.

The evaluation included three evaluation tasks: Speech-To-Text, Speaker Diarization, and Speaker Attributed Speech-To-Text.

The WERs for the STT task continue to be higher than the WERs for previous Conversational Telephone Speech evaluations by 25 % relative.

The ICSI SPKR team achieved 8.5 % DER on the *confmtg* test set which was 76 % lower than last year's best system. The ICSI system both detected speech accurately and clustered speakers accurately.

This was the first RT evaluation to include the SASTT task. Four out of five STT sites submitted systems for the SASTT tasks. The lowest speaker attributed word error rates for segment groups with OV≤4 for the MDM condition were 40.3 %, 59.3 %, and 68.4 % for the *confmtg*, *lectmtg*, and *cbreak* test sets respectively with relative increases in error of 7.7 %, 17.6 %, and 41.6 % respectively over comparable STT systems.

The Rich Transcription 2008 Evaluation will occur during the Fall of 2008. The SASTT task will likely remain as a task for the next evaluation since the implementation of the evaluation task was successful and there is enthusiasm within the community to continue to work on the task. The selection strategy for the *confmtg* test data will change for the next evaluation to include a wider variety of speaker speech durations and the number of active speakers within each excerpt.

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 and LDC for preparing reference transcriptions and annotations.

[2] The inferred boundaries were generated automatically by converting each recognized word to a speaker segment, then smoothing the segments with the 0.3 second smoothing parameter.

5 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. In: RT-2004F Evaluation Workshop Proceedings, November 7-10 (2004)
2. Garofolo, et al.: The Rich Transcription 2004 Spring Meeting Recognition Evaluation. In: ICASSP 2004 Meeting Recognition Workshop, May 17 (2004)
3. The (RT-07) Rich Transcription Meeting Recognition Evaluation Plan (2007), http://www.nist.gov/speech/tests/rt/rt2007
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. Garofolo, J.S., Fiscus, J.G., Radde, N., Le, A., Ajot, J., Laprun, C.: The Rich Transcription 2005 Spring Meeting Recognition Evaluation. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, pp. 369–389. Springer, Heidelberg (2006)
7. http://www.clear-evaluation.org/
8. Fiscus, et al.: Multiple Dimension Levenshtein Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech. In: LREC 2006: Sixth International Conference on Language Resources and Evaluation (2006)
9. http://isl.ira.uka.de/clear06/downloads/ClearEval_Protocol_v5.pdf
10. Fiscus, J., Ajot, J., Michel, M., Garofolo, J.: The Rich Transcription 2006 Spring Meeting Recognition Evaluation. In: Renals, S., Bengio, S., Fiscus, J.G. (eds.) MLMI 2006. LNCS, vol. 4299, Springer, Heidelberg (2006)
11. Burger, S.: The CHIL RT07 Evaluation Data. In: The Joint Proceedings of the 2006 CLEAR and RT Evaluations (May 2007)
12. Lammie Glenn, M., Strassel, S.: Shared Linguistic Resources for the Meeting Domain. In: The Joint Proceedings of the 2006 CLEAR and RT Evaluations (May 2007)
13. Wooters, C., Fung, J., Peskin, B., Anguera, X.: Towards Robust Speaker Segmentation: The ICSI-SRI Fall 2004 Diarization System. In: RT-2004F Workshop (November 2004)
14. Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J., Mostefa, D., Soundararajan, P.: The CLEAR 2006 Evaluation, Proceedings of the first International CLEAR Evaluation Workshop. In: Stiefelhagen, R., Garofolo, J.S. (eds.) CLEAR 2006. LNCS, vol. 4122, pp. 1–45. Springer, Heidelberg (2007)
15. StainStiefelhagen, R., Bowers, R., Rose, R.: Results of the CLEAR 2007 Evaluation. In: The Joint Proceedings of the 2006 CLEAR and RT Evaluations (May 2007)
16. http://www.nist.gov/dads/HTML/HungarianAlgorithm.html
17. http://www.nist.gov/speech/tests/sigtests/mapsswe.htm
18. 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)