THE RICH TRANSCRIPTION 2004 SPRING MEETING RECOGNITION EVALUATION

John S. Garofolo[†], Christophe D. Laprun[†], Jonathan G. Fiscus[†]*

[†] National Institute of Standards and Technology
[*] Systems Plus Inc.

ABSTRACT

This paper presents the design and results of the Rich Transcription 2004 Spring Meeting Recognition Evaluation. The evaluation included both Speaker Segmentation (SPKR) and Speech-to-Text Transcription (STT) tasks. Three microphone type conditions were supported:

Multiple Distant Microphones (the primary condition of interest),

Single Distant Microphone (SDM), and

Individual Head Microphones (IHM) (for the STT task only).

The 3 microphone type conditions permitted the examination of performance for distant vs. close-talking microphones and single vs. multiple distant microphones.

Multi-site training and development corpora were provided to the evaluation participants. The evaluation test set consisted of 8 11-minute meeting excerpts collected at Carnegie Mellon University (CMU), the International Computer Science Institute (ICSI), the Linguistic Data Consortium (LDC), and the National Institute of Standards and Technology (NIST). The development and evaluation corpora were transcribed by the LDC. Because meetings contain a great deal of overlapping/spontaneous speech, the evaluation featured a new experimental scoring of overlapping speech for the STT task.

1. MOTIVATION

Huge efforts are being expended in mining information in newswire, news broadcasts, and conversational speech and in developing interfaces to metadata extracted in these domains. However, little has been done to address such applications in the more challenging and equally important meeting domain.

The meeting domain has several important properties not found in other domains and which are not currently being focused on in other research programs including: highly diverse forums, participant hierarchies, relationships, and vocabularies, highly-interactive and simultaneous speech from multiple speakers, multiple distant microphones, one or more camera views, and a variety of multi-sensor and integration issues.

The development of smart meeting room core technologies that can automatically recognize and extract important information from multi-media sensor inputs will provide an invaluable resource for a variety of business, academic, and governmental applications.

To facilitate progress in fledgling core meeting recognition technologies, NIST is sponsoring an evaluation and workshop series focusing on the challenge of recognizing speech in meetings.

While this evaluation series has, thus far, not incorporated video source signal input or multi-media integration, these factors are important issues to be addressed in the future.

2. RT-02 MEETING RECOGNITION

NIST carried out the first community-wide evaluation of meeting domain Speech-to-Text Transcription (STT) and Speaker Segmentation (SPKR) in the context of its Rich Transcription 2002 evaluation[6][7]. An 80-minute test set with 8 10-minute meeting excerpts collected at NIST[3], CMU[1], ICSI[4], and the LDC[2] was used. Performance was measured for individual personal mics (head or lapel mics depending on data collection site), a single distant omni-directional mic, and a personal mic mix. NIST applied its speech recognition and speaker segmentation scoring software to evaluate the tasks. The *SCLITE* STT scoring software did not support evaluation of overlapping speech, so segments containing overlapping speech were not evaluated. However, unlike broadcast news and 4-wire telephone speech (both having little within-channel overlap), this was identified as a significant issue for recognition from distant microphones.

This exploratory evaluation included only 2 participants (SRI and MIT-Lincoln Labs), but proved the feasibility of evaluating STT and SPKR technologies in the meeting domain and provided a performance baseline[7]. Accordingly, the participants performed little domain-specific development. The evaluation showed that performance for the individual close-talking microphone condition was similar to that of conversational telephone speech. However, performance for the single distant microphone condition was significantly worse than for the individual personal mic condition (nearly twice as high absolute for the STT task – even excluding overlapping speech.) A great deal of variability was also observed

across meetings and data collection. Given these results, meeting research sites are now focusing on the distant mic problem as well as meeting-specific metadata extraction[5].

3. RT-04 SPRING MEETING RECOGNITION EVALUATION

While the Rich Transcription 2004 Spring Meeting Recognition Evaluation shares several attributes with the earlier 2002 evaluation, a number of improvements have been made to the test methodology and source data set.

Like the 2002 evaluation, the 2004 evaluation included both Speaker Segmentation and Speech-to-Text Transcription tasks. In 2002, systems were permitted to use the reference segmentation to perform the STT task. This was done away with in 2004 – all processing was required to be fully automatic. The 2002 evaluation included a head-mic mix condition. This was deemed to be artificial by the participants and it was done away discontinued. Instead, the participants were interested in the challenge of optimizing recognition performance over multiple farfield mic inputs. So, the focus of the 2004 evaluation shifted to a multiple distant microphone condition[1].

Likewise, a great deal of speech was not evaluated in the 2002 STT task because it contained overlapping simultaneous speech. Traditional transcription practices had largely excluded such speech and the NIST *SCLITE* scoring tool could not process multiple parallel transcriptions – a feature necessary to score overlapping speech. In the past, this was not a significant impediment to evaluation since overlapping speech is not prominent in broadcast news or 4-wire conversational telephone speech. However, informal meeting domain speech generally contains a great deal of overlapping speech. In particular, 78 % of the words in the RT04S meeting test set occur in segments[2] containing overlapping speech. Since *SCLITE* evaluates recognition performance on a segment basis, all of these words are excluded from the *SCLITE* scoring procedure. To address this challenge, NIST developed a prototype STT scoring tool to evaluate segments containing overlapping speech. The new tool and the results of an experimental evaluation of the RT-04S STT results using it are discussed in Section 7.

Information regarding the evaluation was provided on the RT-04S website[8]. The specifications for the evaluation were given in a detailed evaluation plan [9] on the website.

3.1. SPEAKER SEGMENTATION TASKS

A "who spoke when" speaker segmentation task was included in RT04S and was designed to be similar to the RT03 Spring broadcast news diarization task. For this task, systems were required to identify temporal regions of speech from different talkers with an arbitrary unique ID. The following test conditions were supported for this task:

Multiple Distant Microphones (MDM) – systems were presented with recordings from multiple distantly placed microphones (depending on the data collection site)[3] and were required to produce a single speaker speech index for all the speakers in each meeting. This condition was considered primary for the evaluation.

Single Distant Microphone (SDM) – systems were presented with a recording of a single "centrally placed" distant microphone (as determined by the data collection site). This contrast condition was designed to demonstrate the effectiveness of using multiple distant microphone inputs vs. a single distant microphone input.

To evaluate the performance, the system output segments were mapped to human-annotated reference segments and scored using the NIST *spkrsegeval* speaker segmentation scoring software. For the purposes of this evaluation, within-speaker speech pauses of less than 300 milliseconds were not to cause breaks in segments. Since the resources did not exist to provide extremely accurate segment boundaries, a 250-millisecond time "collar" was applied when determining the accuracy of segment boundaries – thus forgiving small boundary errors due to inaccuracies in the reference. The primary metric was **speaker segmentation diarization error** which is a ratio of incorrectly-attributed speech time to the total speech time in the test set. Scores were computed for both segments without overlapping speech and segments where

[1] The evaluation was of the speech recognition systems with certain broad types of microphone inputs, not of the mics themselves. The mics were chosen to be representative of two broad classes: commercial close-talking/noise-cancelling head-boom mics and commercial distant conference room mics. The particular models used were not considered to be of particular consequence. The evaluation therefore emphasized comparison of performance on the classes of mics, not the particular models. The microphone models used varied from data collection site to data collection site (sometimes even from meeting-to-meeting within a data collection site).

[2] *SCLITE* segments are pause-separated sequences of words from a single speaker. Segment times cannot overlap and word times within a segment cannot overlap.

[3] CMU recorded only a single distant mic in the two meetings it contributed for the evaluation

overlapping speech could be properly attributed in the reference.

3.2. SPEECH-TO-TEXT (STT) TASKS

The RT04S evaluation included a Speech-To-Text Transcription (STT) task. Systems were required to output the words spoken by the meeting participants without any speaker designation (except for the IHM task in which the recordings of each speaker's head mic was presented explicitly in separate files) along with the start and end time of each recognized word.

The 2004 evaluation contained the following test conditions for the STT task:

Multiple Distant Microphones (MDM)[4] – systems were presented with recordings of multiple distantly placed microphones (depending on the data collection site)[3] and were required to output a single transcript containing all of the words spoken by all the meeting participants in each meeting. This condition was considered primary for the evaluation.

Single Distant Microphone (SDM)[4] – systems were presented with a single "centrally placed" distant microphone (as determined by the data collection site). This contrast condition was designed to demonstrate the effectiveness of using multiple distant microphone inputs vs. a single distant microphone input.

Individual Head Microphones (IHM) – systems were presented with a separate head-mic recording for each meeting participant and were required to output a separate transcript for each participant in each meeting.

The STT tasks were evaluated using the standard NIST *SCLITE* speech recognition scoring software which mapped the system output words to the words in the human reference. The primary metric was Word Error Rate (WER) which is a ratio of the sum of the number of erroneous words inserted, missed, and substituted by the recognition system to the total words in the test set. The *SCLITE* scoring included only words in non-overlapping speech segments – segments bounded by silence in which only a single participant spoke.

4. CORPORA

Unlike the RT-02 meeting recognition evaluation in which only a partially transcribed development test set and no training data was provided, a fully transcribed set of training data, development test data, as well as the evaluation test set was provided for the RT-04S evaluation. To make the challenge as broad and realistic as possible, data was employed from 4 data collection sites: CMU, ICSI, the LDC, and NIST. No effort was made to control the corpus parameters across sites. As such vastly different data collection systems, subject populations, microphones, scenarios, and transcription conventions were employed at the different data collection sites.

4.1. TRAINING DATA

A 95-hour multi-site training corpus was made available by 3 of the contributing data collection sites for this evaluation. The data is described in Table 1.

Collection Site	Hours	Meetings
CMU	10	18
ICSI	72	75
NIST	13	17

Table 1 – RT-04S Multi-Site Training Data

4.2. DEVELOPMENT DATA

The RT-02 evaluation test set was deemed to be the development test data for this evaluation. It consisted of 8 10-minute excerpts from meetings collected at CMU, ICSI, LDC, and NIST data (2 meeting excerpts per data collection site) – the same data collection sites represented in the evaluation data. This dataset was similar in structure to the RT-04S evaluation test set with one significant difference – only lapel mics were collected as personal mics at CMU and the LDC for the development data. So, it did not contain representative head-mic data from all the data collection sites. The development test set was transcribed by the LDC in 2002.

4.3. EVALUATION DATA

The evaluation test set consisted of 8 11-minute excerpts from 8 meetings collected at each of the 4 data collection sites (2 excerpts per data collection site). All meetings were recorded with a noise-canceling head-mounted mic on each subject to support the IHM condition and 1 or more distantly-placed mics to support the SDM and MDM conditions. If multiple distant mics were collected, the data collection sites specified a central distant mic for the SDM condition. The meetings from ICSI, the LDC, and NIST contained recordings from multiple distant microphones. The CMU data contained only a single distant microphone.

[4] The same datasets were used for the MDM and SDM conditions for both the SPKR and STT tasks.

The test set contained 31 unique participants with 40 participant instances. Therefore, some participants took part in multiple meetings in the test set. Of the 31 unique participants, 4 were non-native speakers of American English. The meetings ranged from 3-person gatherings to 10-person group meetings. The meeting forums included 2 moderated focus group discussions from the LDC, 2 un-moderated discussion group meetings from CMU, 2 technical working group meetings from ICSI, and a technical working group meeting and a scenario-task-driven meeting from NIST. The evaluation test set meetings are described more extensively in Appendix G.

The LDC transcribed the evaluation test set using a convention adapted from the existing Rich Transcription convention for conversational telephone speech. The transcription convention was designed to make the SPKR and STT tasks and lexical forms as similar to the other RT tasks as possible[13].

The test set contained a total of 20,697 word tokens. Of those, only 4,496 occurred in non-overlapping speech segments. Therefore, only 22 % of the word tokens in the test set could be evaluated in the traditional STT Word Error Rate scoring by *SCLITE*.

5. SPEAKER SEGMENTATION RESULTS

Three sites participated in the Speaker Segmentation task evaluation: (1) a team consisting of LIA and CLIPS, (2) a team consisting of CMU and the University of Karlsruhe, and (3) Macquarie University.

The best performance for the MDM condition was achieved by the LIA-CLIPS2 system (running in 20xRT) with a Diarization Error Rate (DER) of 23.54 % for non-overlapping speech and 37.53 % for overlapping speech. It is worth noting, though, that LIA/CLIPS' contrastive system achieved the best overall performance with a 22.79 % DER for non-overlapping speech and 37.22 % DER for overlapping speech.

For the SDM condition, the best performance was achieved by the LIA-CLIPS0 system with a 26.46 % DER on non-overlapping speech and 39.46 % DER on overlapping speech.

The results clearly show that performance for overlapping speech is significantly worse than for non-overlapping speech. It is interesting to note that the SDM results for the LIA-CLIPS0 system are close to the results for that system on the MDM data with a difference of only 2 % absolute. This system was one of the best overall performers (for both overlapping and non-overlapping speech, see Figure 13), but did not seem to benefit from the multiple mic inputs. The complete results of the Speaker Segmentation evaluation are shown in Appendix A.

6. SPEECH-TO-TEXT TRANSCRIPTION RESULTS

Four sites participated in the Speech-to-Text Transcription task evaluation: (1) a team consisting of CMU/University of Karlsruhe (CMU/KU is represented as "ISL" in the appendices), (2) a team consisting of ICSI/SRI/University of Washington (ICSI/SRI/UW), (3) Panasonic, and (4) Virage. Table 2 shows which sites participated in which tasks. It is important to note that it is difficult to conclusively compare the different systems' results since they were run at different processing speeds.

	MDM	SMD	IHM
CMU/KU	X	X	X
ICSI/SRI/UW	X	X	X
Panasonic	X		X
Virage		X	X

Table 2 - Site/Task Participation Matrix

The primary STT task was the MDM condition. Please refer to Appendix B for the results of the STT MDM evaluation. The best performance was achieved by CMU/KU with an overall Word Error Rate (WER) of 44.9 %. The WER for the ICSI/SRI/UW system was 47 % but ran at only 10xRT while the CMU/KU system ran in unlimited time. Since the system speeds were different, it is inappropriate to compare these systems statistically. ICSI/SRI/UW also submitted results for a revised bug-fixed system. That system, which ran in unlimited time, equaled the performance of the CMU/KU system and is not statistically different using the MAPSSWE test. However, it should be noted that the ability of the statistical significance test to distinguish performance differences is severely weakened by the small test set size and small number (4496) of non-overlapping evaluable word tokens.

For the SDM condition (results in Appendix C), the best performance was achieved by the CMU/KU system (running in unlimited time) with a WER of 49.8 %. The ICSI/SRI/UW system (running in 10xRT) achieved comparable results with a WER of 51.3 % overall. The MAPSSWE test found a significant difference between these two systems.

Interestingly, for the distant mic conditions, systems performed worse for male speakers than female speakers. This was particularly obvious for the SDM condition where the WER was 4.3 % higher for males than females for the ICSI/SRI/UW system and 2.9 % higher for CMU/KU system. However, Virage's SDM results were 1.6 % better for the male speakers than for the female speakers.

It's interesting to note that one of the NIST meetings (*NIST_20030925-1517*) was particularly difficult for the distant mic conditions. This meeting featured a scenario-driven task. We reviewed the video corresponding to the excerpt that was chosen for the evaluation and saw that during the excerpt, one participant spoke from the whiteboard at the side of the room during the beginning of the excerpt and was, therefore, several feet from the nearest microphone. However, this also caused the two participants sitting at the table to turn in his direction – away from the microphones. The fourth participant was facing the main view screen on the North Wall (see [10]) and typing on the keyboard as she spoke. This example dramatically shows that far-field recognition is particularly sensitive to microphone placement and position of participants relative to the mics.

A comparison of the MDM and SDM system performance (see Figure 26 and Figure 27) clearly shows that systems performed better overall when presented with multiple distant mic inputs than a single centrally-placed distant mic input. This encouraging finding demonstrates that the systems were making effective use of the multiple signal inputs. Interestingly, however, one of the meeting excerpts (*LDC_20011207-1800*) proved to be more challenging for the MDM condition than for the SDM condition. Also of interest is that the CMU/KU system achieved a different WER for the SDM and MDM conditions for the CMU meetings. This is surprising since CMU provided only a single distant mic for its meetings. So, in theory, the results should have been identical for the MDM and SDM conditions for those meetings.

For the IHM condition (results in Appendix D), the best performance was achieved by the ICSI/SRI/UW 10xRT system with a WER of 34.8 %. Their bug-fixed system improved on this result with a WER of 32.7 %. However, this system fell into the unlimited time category. The CMU/KU unlimited time system achieved comparable results with a 35.7 % WER. The Virage and Panasonic systems performed quite poorly on this data with WER that were significantly higher than their SDM/MDM systems with very high insertion error rates. This is most likely because the Virage and Panasonic systems did not employ an echo cancellation algorithm in their IHM systems to ignore cross-talk in the close-talking mics.

The comparison between the IHM and MDM results provide also somewhat surprising results (with sites performing worse on some meetings on the IHM condition than on the MDM condition as seen on Figure 28). As of this writing, the reason for this anomaly is unclear and will require further investigation. It may be significant that the IHM and SDM/MDM tests were scored using different token sets since all overlapping speech was ignored in the SDM/MDM evaluation and not in the IHM evaluation (since the overlapping speech was separated on different channels). This warrants an additional IHM condition scoring using the same token set as the SDM/MDM conditions.

7. STT TASK OVERLAPPING SPEECH EVALUATION EXPERIMENT

Meeting participants frequently interrupt each other and/or talk at the same time – especially in informal meetings. In the distant microphone recordings of the RT-04S test set, 78 % of the word tokens occurred in overlapping speaker turns. Previously, using the standard NIST *SCLITE* speech recognition scoring tool, only the non-overlapping 22 % of the data could have been evaluated. This is because *SCLITE* requires a one-to-one mapping of the reference transcription stream to the system output stream. To address this problem, a new multi-dimensional network alignment program was developed at NIST which enabled multiple transcription streams to be aligned and scored.

So as not to redefine the existing STT task, NIST implemented a scoring protocol that did not require speaker attribution for each system-generated token. Instead, the scoring protocol relies on the reference transcript for speaker identification and freely aligns system tokens to any reference token. The algorithm aligns the single stream system output tokens to the multi-stream reference tokens so as to minimize the overall Word Error Rate – thus providing a maximally forgiving error measurement.

With the exception of the multi-stream alignment procedure, the steps taken to evaluate overlapping speech are essentially the same as has been traditionally used for all NIST-sponsored STT evaluations: normalize the reference and hypothesis transcriptions, segment the reference transcript into silence-bounded regions, align and score the system output transcription against the reference. As such, the token alignment step was the most significant change from *SCLITE*.

7.1. REFERENCE TRANSCRIPT SEGMENTATION

The segmentation step amounts to analyzing the reference transcript turn times to find time points that clearly divide the recording into turn- and silence-bounded speech regions. This is necessary to bound the alignment computation. The result is that the recording timeline is divided into segmented speech regions containing 1 or more talkers and inter-segment gap regions where no speech exists so that all of the timeline is accounted for. The mid-point of the system output token times are then used to assign the tokens to putative regions in the reference. The region-mapped system output tokens are then aligned to the reference tokens using the multidimensional alignment algorithm described below.

7.2. TOKEN ALIGNMENT

The token alignment process employs a new, stream-based multi-dimensional network alignment algorithm which is an extension of the *SCLITE* two-dimensional network dynamic programming (DP) string alignment procedure (based on Kruskal and Sankoff's "Directed Network Alignment"[12]). In the *SCLITE* algorithm, the system transcript and the reference transcript for a single speaker turn are formed into two word transition networks that are aligned. In the new algorithm, the reference transcript for each speaker in the speech region is formed into individual networks that constitute additional alignment dimensions – one for each speaker. Thus the number of networks to align is the number of reference speakers plus one.

The computational complexity of the DP alignment is proportional to $O(N^d)$ where d is the number of dimensions in the alignment and where N is the number of tokens to align in each dimension. This is therefore a very computationally expensive algorithm and search constraints must be employed to make the computation tractable. The following is a list of the constraints currently employed:

> Reference tokens cannot align to each other.
>
> System tokens cannot align to each other.
>
> Aligned tokens must overlap in time.[5]
>
> The time synchrony of aligned token pairs is monotonic. Thus, an aligned token pair that occurs at times (T1:T2) cannot be followed by a pair of aligned tokens whose times precede T1.

Even with these four constraints, the alignments become intractable when the reference transcript contains several overlapping speakers. We implemented several experiments on the RT-04S evaluation data to determine feasibility. Our experiments showed that the algorithm sometimes becomes too memory-intensive when more than 3 speakers talk in the same region (Appendix F shows the degree of overlapping speech in each of the RT-04S test set meetings.) However, limiting the scoring to 3-person overlap still rendered an additional 49 % of the words in the reference as scorable. Therefore, only 29 % of the words in the reference remain un-scored.

The results of the scoring are shown in Appendix F. The scores clearly indicate increasing error rates for 2- and 3-party overlapping speech regions. The lowest WER for speech containing 3-party or less overlap for the MDM condition was achieved by the CMU/KU system with a WER of 56 %. This result is 11 % higher absolute than the 45 % WER obtained from the same scoring applied to only single-party speech. Note that this number does not correspond exactly to the *SCLITE*-generated result of 44.9 %. We are researching this difference, but it is most likely due to subtle differences in how the reference is generated and how the alignment is generated by *SCLITE* and the new algorithm. We have found, however, that the numbers are almost identical when we use the reference generated for *SCLITE* non-overlapping scoring.

These initial results show empirically that multi-party overlapping speech poses great challenges for recognition technology and that the error rates increase as more speakers are introduced. We intend to continue work on the multi-stream scoring algorithm and we hope this new tool enables the community to better focus on this difficult problem.

7.3. RICH TRANSCRIPTION 2004 SPRING MEETING RECOGNITION WORKSHOP

The results of the evaluation were presented at a NIST-sponsored ICASSP-2004 satellite workshop in Montreal on May 17, 2004. The workshop was broadly attended by the evaluation participants, data collectors, other researchers working in related meeting domain technologies, and by representatives of large governmental programs funding research and development in meeting domain technologies.

8. CONCLUSION AND FUTURE DIRECTIONS

In implementing this evaluation and preparing for this workshop, it has become clear than many research sites are now working on the problem of speech recognition in the meeting domain. Anecdotally, it also seems that performance has improved since the initial such evaluation in 2002.

There are several issues we must address with regard to the test set design prior to conducting the next such evaluation. Because the small test set and high degree of overlapping speech greatly reduced the number of evaluable word tokens (and statistical inferences we could make) for traditional *SCLITE* scoring, we must either abandon such scoring as the primary form of evaluation in this domain or greatly increase the test set size. We might also consider some constraints on the data collection parameters for the test set so as to make it somewhat less diffuse. However, with that said, we must be careful to avoid making the task artificial. We will continue discussion of approaches for the evaluation of overlapping speech and will continue to explore overlapping speech scoring algorithms.

[5] Since accurate reference word times were not available for this evaluation, system output tokens were only required to overlap in time with the reference turn times to be aligned.

We believe that the core component tasks (STT/SPKR) included in the RT-04S evaluation were the most important at this time for this domain. However, it may be desirable to begin to explore fusion tasks involving a combination of technologies such as a speaker-identified STT task. If additional resources become available, we may also wish to explore additional metadata extraction tasks such as the sentence unit and disfluency detection tasks currently being addressed in the DARPA EARS program. Finally, with increasing interest from the European Community, we may wish to explore these tasks using multi-lingual meeting recordings. While we believe that the RT-04S test conditions (MDM/SDM/IHM) are the most important ones to probe the dimensions of interest, resources permitting, we might also consider adding an array microphone contrast to support research efforts in that area.

It will also be important to eventually begin exploring multi-modal fusion tasks using video as well as audio inputs. We did find that the video from the NIST meetings was invaluable in our analysis of the results of the evaluation and we wished we had had such resources from the other data collection sites.

The increasingly broad participation of data collectors, evaluation participants, workshop participants, and international research programs in multi-media recognition in the meeting domain is a strong indicator that the research community is now focusing on building meeting understanding technologies. We plan to continue this evaluation and workshop series on possibly an annual basis. Input from the community on the design of the evaluation and the workshop is most welcome.

9. ADDENDA – ERRATA

We realized, while preparing the reference data for public distribution that, contrary to what was specified in the evaluation specifications (in Section 4 of that document), we did not perform any data smoothing for the diarization part of the evaluation. In the context of this evaluation, small pauses of less than 0.3 seconds in the speaker speech were not to be considered as segmentation breaks. Bridging (smoothing) such segments into a single continuous segment should have been performed on the reference transcripts (as well as by the systems).

We have now performed the smoothing of the reference transcripts in the version that will be publicly available. We include in this paper updated diarization results based on the corrected version of the reference transcripts.

The following table compares diarization results with and without smoothing of the reference data. Additionally, Appendix A has also been augmented to include the complete updated results.

	Non-Overlapping		Overlapping	
MDM	**Original**	**Smoothed**	**Original**	**Smoothed**
LIA-CLIPS0_1	24.61	25.13	37.63	38.36
LIA-CLIPS1_1	22.79	23.26	37.22	37.93
LIA-CLIPS2_1	23.54	24.1	37.53	38.29
ISL	28.17	27.66	40.19	40.21
SDM	**Original**	**Smoothed**	**Original**	**Smoothed**
LIA-CLIPS0_1	26.46	27.46	39.46	40.6
LIA-CLIPS1_1	28.99	29.98	41.37	42.41
LIA-CLIPS2_1	30.02	31.11	42.74	43.86
MQU	62.02	62.27	69.09	69.19

Table 3 - Comparison between original and corrected results.

10. CAVEAT

Certain commercial products are mentioned to explain the processes used. NIST does not recommend particular commercial products nor does it believe that the products used were necessarily the best for the tasks described.

11. REFERENCES

[1] Burger, S., MacLaran, V., Yu, H. (2002). The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style, Proc. ICSLP 2002, Denver.

[2] Cieri, C., Miller, D., Walker, K. (2002) Research Methodologies, Observations and Outcomes in Conversational Speech Data Collection, Proc. HLT 2002, San Diego, CA

[3] Garofolo, J.S., Laprun, C.D., Michel, M., Stanford, V.M., Tabassi, E., The NIST Meeting Room Pilot Corpus, Proc. LREC 2004, Lisbon, Portugal.

[4] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C. (2003). The ICSI Meeting Corpus, Proc. ICASSP 2003, Hong Kong.

[5] Morgan, N., Baron, D., Bhagat, S., Carvey, H., Dhillon, R., Edwards, J., Gelbart, D., Janin, A., Krupski, A., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C. (2003). Meetings about Meetings: Research at ICSI on Speech in Multiparty Conversations, Proc. ICASSP 2003, Hong Kong.

[6] NIST (2002). Rich Transcription 2002 Meeting Recognition Evaluation,

documentation, http://www.nist.gov/speech/tests/rt/rt2002/

[7] NIST (2002). Rich Transcription 2002 STT and Metadata Extraction results, presentations, RT-02 Workshop, http://www.nist.gov/speech/tests/rt/rt2002/presentations/index.htm

[8] NIST (2004), Rich Transcription 2004 Spring Meeting Recognition Evaluation, documentation, http://nist.gov/speech/tests/rt/rt2004/spring/

[9] NIST (2004), Rich Transcription Spring 2004 Meeting Recognition Evaluation Plan, http://nist.gov/speech/tests/rt/rt2004/spring/documents/rt04s-meeting-eval-plan-v1.pdf

[10] NIST (2004), NIST Data Collection Facility Layout, http://nist.gov/speech/test_beds/mr_proj/meeting_corpus_1/room.html#layout

[11] NIST (2004) NIST Speech Recognition Scoring Toolkit, http://www.nist.gov/speech/tools/index.htm

[12] Sankoff, D., Kruskal, J., "Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison", Chapter 10, Section 1, 1983

[13] Strassel, S., Glenn, M., Shared Linguistic Resources for Human Language Technology in the Meeting Domain, Proc. Rich Transcription 2004 Spring Meeting Recognition Workshop, Montreal.

APPENDIX A. DIARIZATION RESULTS

I MDM RESULTS (ORIGINALLY REPORTED RESULTS)

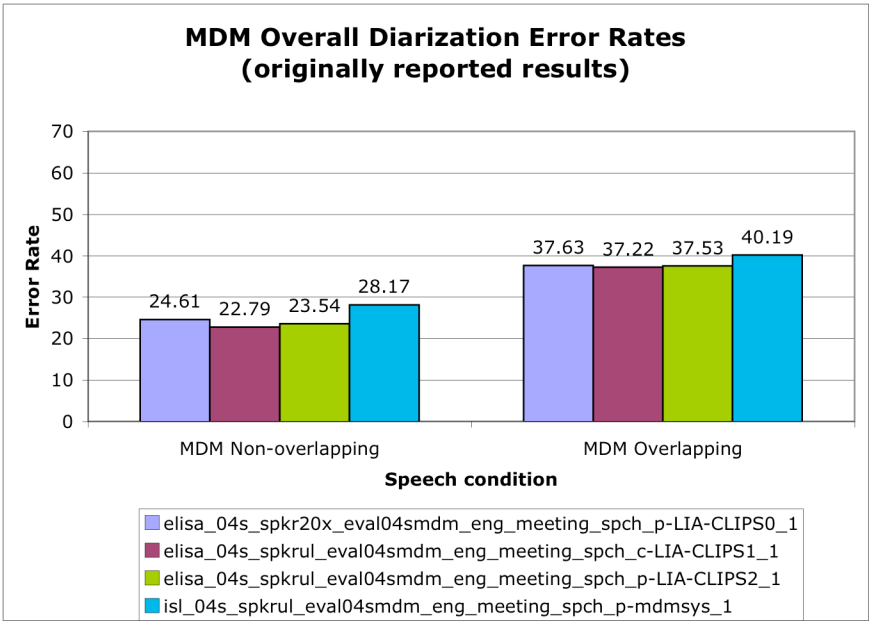
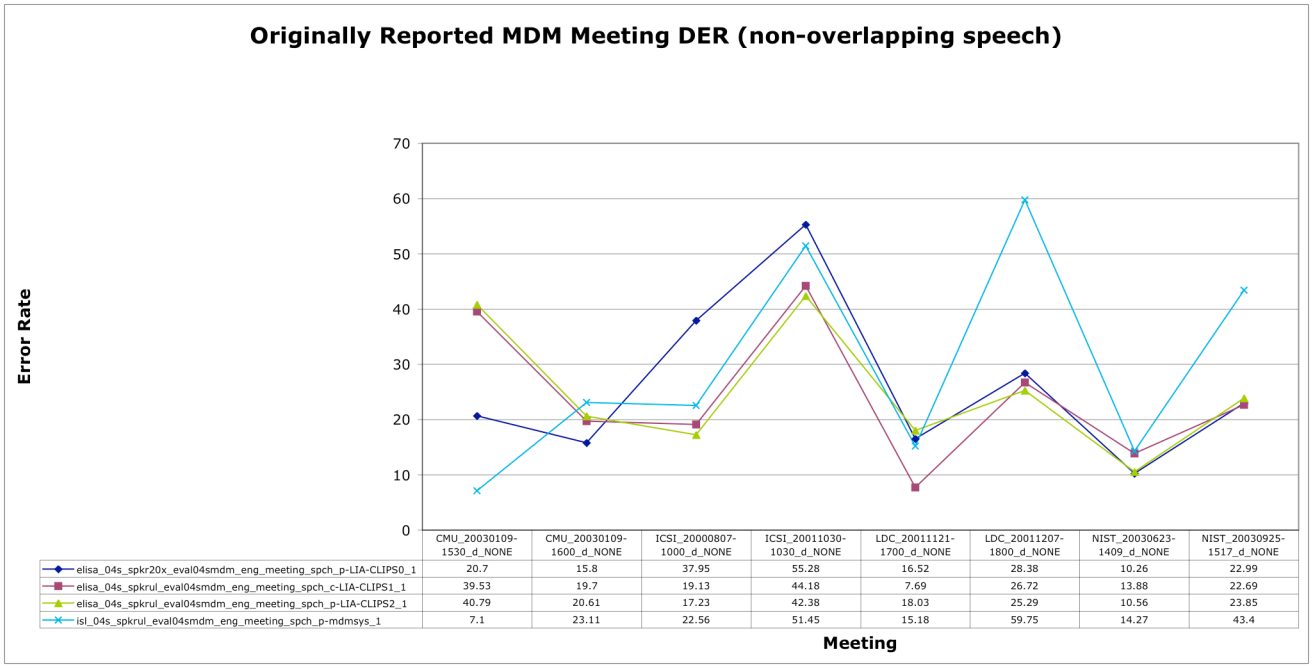


Figure 1 - MDM Overall Diarization Error Rates



	CMU_20030109-1530_d_NONE	CMU_20030109-1600_d_NONE	ICSI_20000807-1000_d_NONE	ICSI_20011030-1030_d_NONE	LDC_20011121-1700_d_NONE	LDC_20011207-1800_d_NONE	NIST_20030623-1409_d_NONE	NIST_20030925-1517_d_NONE
elisa_04s_spkr20x_eval04smdm_eng_meeting_spch_p-LIA-CLIPS0_1	20.7	15.8	37.95	55.28	16.52	28.38	10.26	22.99
elisa_04s_spkrul_eval04smdm_eng_meeting_spch_c-LIA-CLIPS1_1	39.53	19.7	19.13	44.18	7.69	26.72	13.88	22.69
elisa_04s_spkrul_eval04smdm_eng_meeting_spch_p-LIA-CLIPS2_1	40.79	20.61	17.23	42.38	18.03	25.29	10.56	23.85
isl_04s_spkrul_eval04smdm_eng_meeting_spch_p-mdmsys_1	7.1	23.11	22.56	51.45	15.18	59.75	14.27	43.4

Figure 2 - MDM Meeting DER (non-overlapping speech)

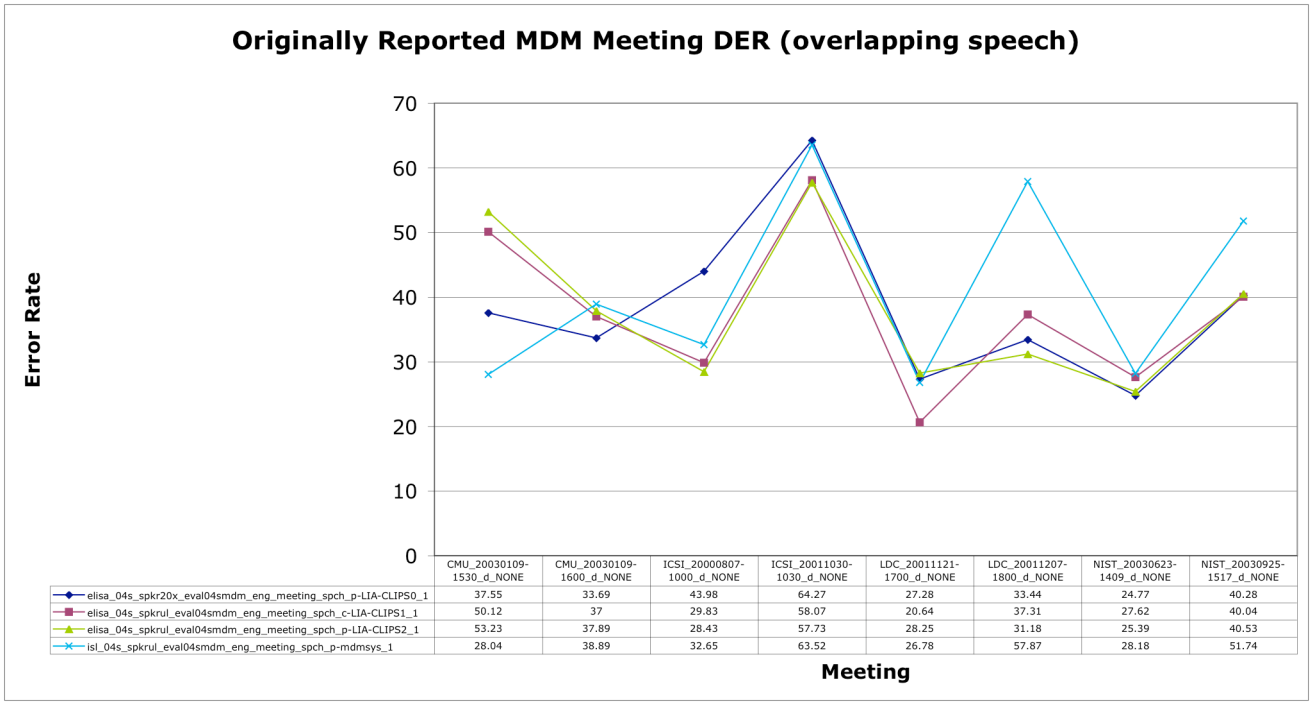


Figure 3 - MDM Meeting DER (overlapping speech)

II MDM RESULTS (INCLUDING SMOOTHING)

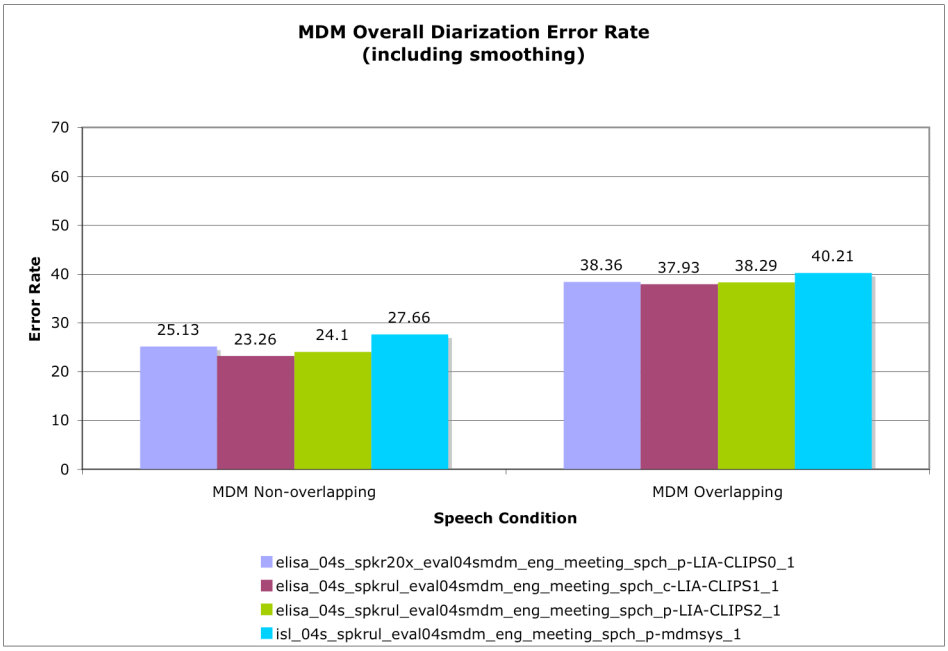
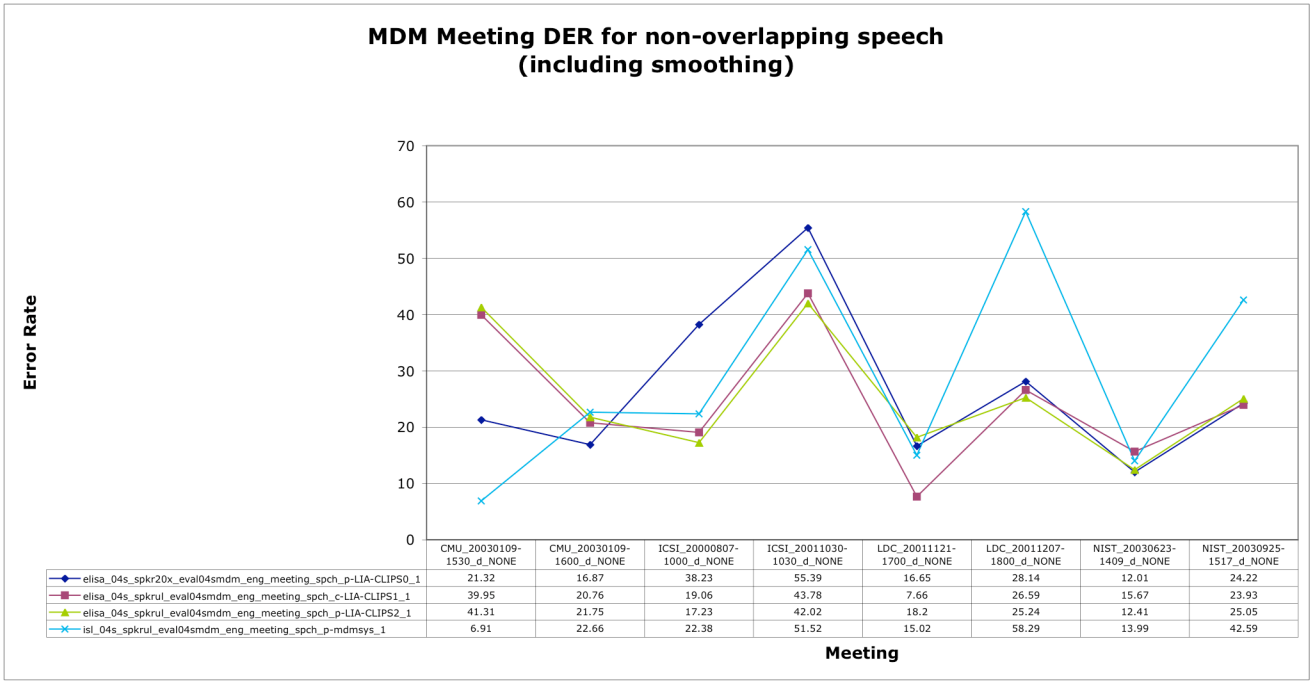
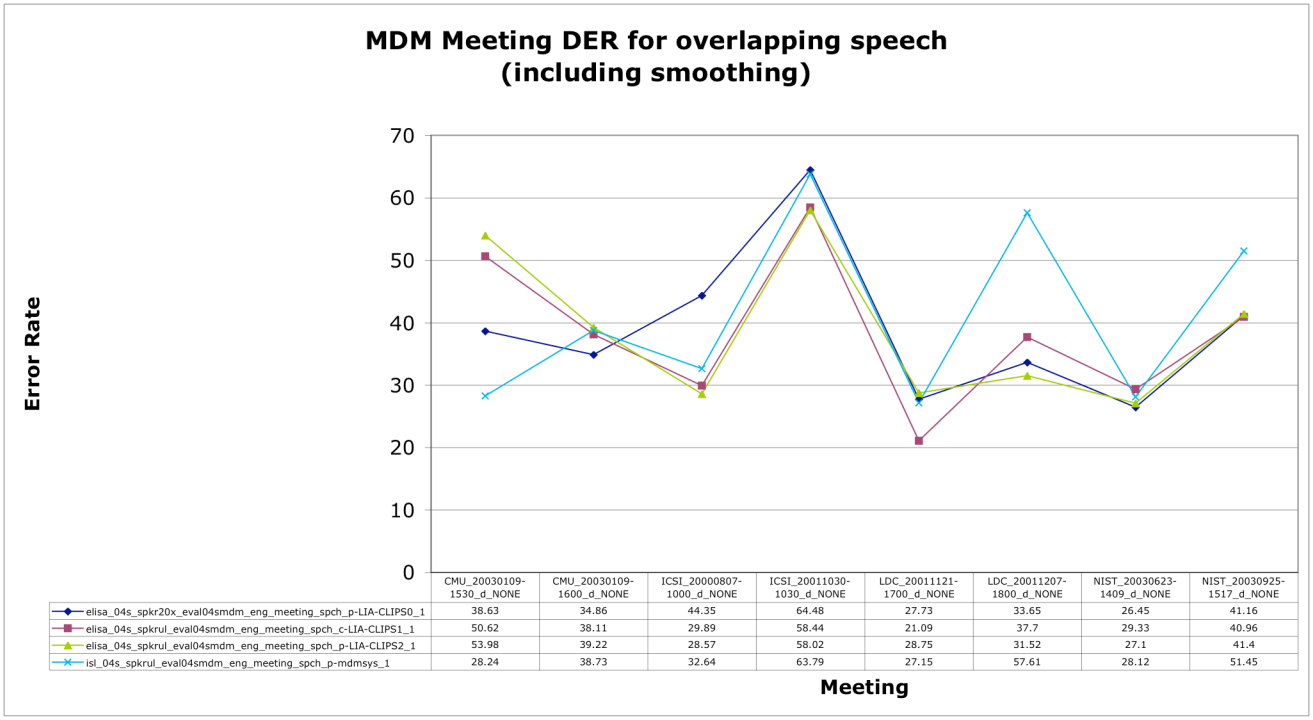


Figure 4 - MDM Overall Diarization Error Rates (including smoothing)



	CMU_20030109-1530_d_NONE	CMU_20030109-1600_d_NONE	ICSI_20000807-1000_d_NONE	ICSI_20011030-1030_d_NONE	LDC_20011121-1700_d_NONE	LDC_20011207-1800_d_NONE	NIST_20030623-1409_d_NONE	NIST_20030925-1517_d_NONE
elisa_04s_spkr20x_eval04smdm_eng_meeting_spch_p-LIA-CLIPS0_1	21.32	16.87	38.23	55.39	16.65	28.14	12.01	24.22
elisa_04s_spkrul_eval04smdm_eng_meeting_spch_c-LIA-CLIPS1_1	39.95	20.76	19.06	43.78	7.66	26.59	15.67	23.93
elisa_04s_spkrul_eval04smdm_eng_meeting_spch_p-LIA-CLIPS2_1	41.31	21.75	17.23	42.02	18.2	25.24	12.41	25.05
isl_04s_spkrul_eval04smdm_eng_meeting_spch_p-mdmsys_1	6.91	22.66	22.38	51.52	15.02	58.29	13.99	42.59

Figure 5 - MDM Meeting DER for non-overlapping speech (taking smoothing into account)



	CMU_20030109-1530_d_NONE	CMU_20030109-1600_d_NONE	ICSI_20000807-1000_d_NONE	ICSI_20011030-1030_d_NONE	LDC_20011121-1700_d_NONE	LDC_20011207-1800_d_NONE	NIST_20030623-1409_d_NONE	NIST_20030925-1517_d_NONE
elisa_04s_spkr20x_eval04smdm_eng_meeting_spch_p-LIA-CLIPS0_1	38.63	34.86	44.35	64.48	27.73	33.65	26.45	41.16
elisa_04s_spkrul_eval04smdm_eng_meeting_spch_c-LIA-CLIPS1_1	50.62	38.11	29.89	58.44	21.09	37.7	29.33	40.96
elisa_04s_spkrul_eval04smdm_eng_meeting_spch_p-LIA-CLIPS2_1	53.98	39.22	28.57	58.02	28.75	31.52	27.1	41.4
isl_04s_spkrul_eval04smdm_eng_meeting_spch_p-mdmsys_1	28.24	38.73	32.64	63.79	27.15	57.61	28.12	51.45

Figure 6 - MDM Meeting DER for overlapping speech (taking smoothing into account)

III SDM RESULTS (ORIGINALLY REPORTED RESULTS)

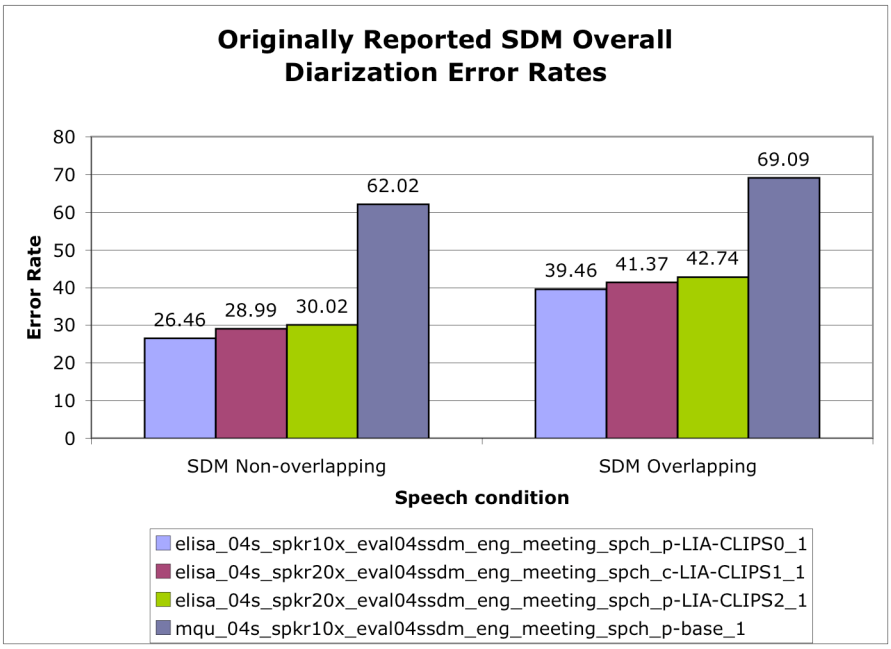
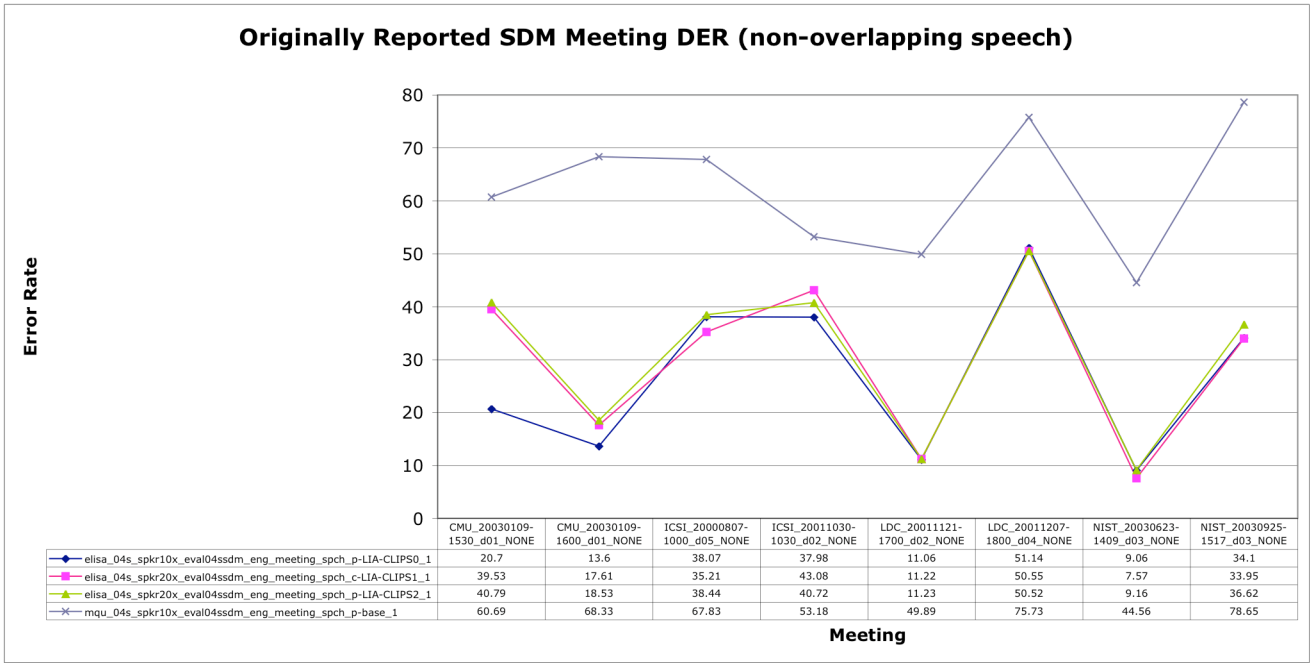


Figure 7 - SDM Overall Diarization Error Rates



	CMU_20030109-1530_d01_NONE	CMU_20030109-1600_d01_NONE	ICSI_20000807-1000_d05_NONE	ICSI_20011030-1030_d02_NONE	LDC_20011121-1700_d02_NONE	LDC_20011207-1800_d04_NONE	NIST_20030623-1409_d03_NONE	NIST_20030925-1517_d03_NONE
elisa_04s_spkr10x_eval04ssdm_eng_meeting_spch_p-LIA-CLIPS0_1	20.7	13.6	38.07	37.98	11.06	51.14	9.06	34.1
elisa_04s_spkr20x_eval04ssdm_eng_meeting_spch_c-LIA-CLIPS1_1	39.53	17.61	35.21	43.08	11.22	50.55	7.57	33.95
elisa_04s_spkr20x_eval04ssdm_eng_meeting_spch_p-LIA-CLIPS2_1	40.79	18.53	38.44	40.72	11.23	50.52	9.16	36.62
mqu_04s_spkr10x_eval04ssdm_eng_meeting_spch_p-base_1	60.69	68.33	67.83	53.18	49.89	75.73	44.56	78.65

Figure 8 - SDM Meeting DER (non-overlapping speech)

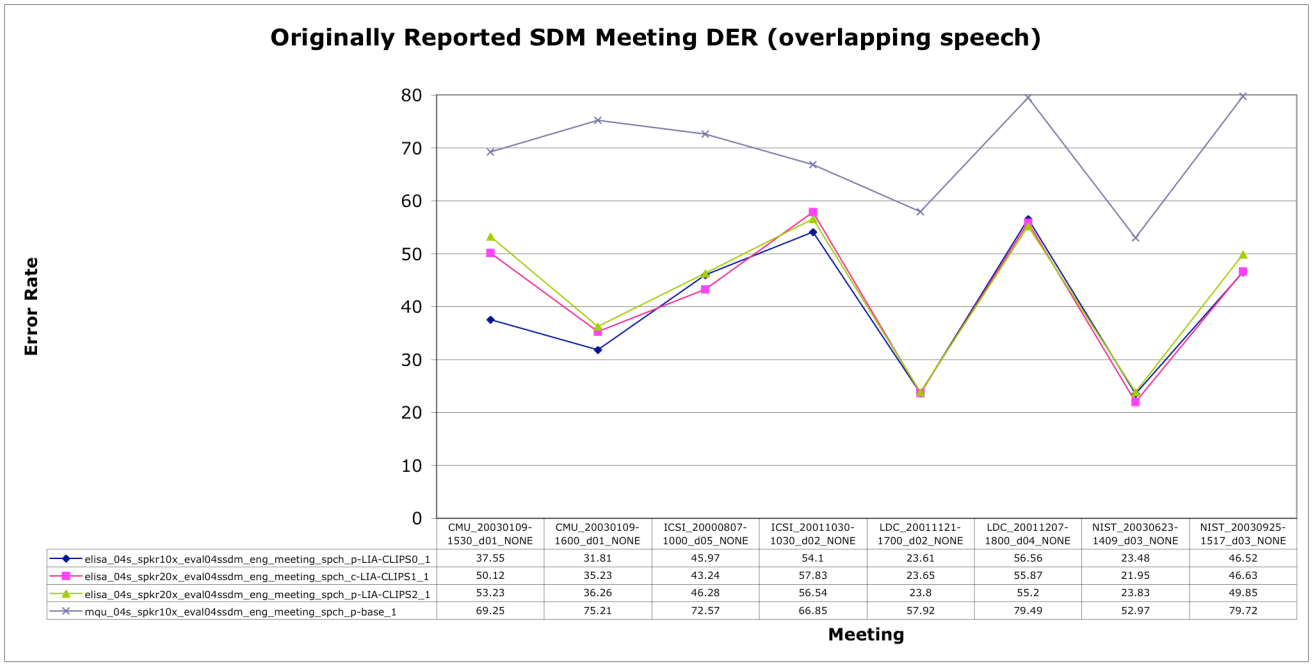


Figure 9 - SDM Meeting DER (overlapping speech)

IV SDM RESULTS (INCLUDING SMOOTHING)

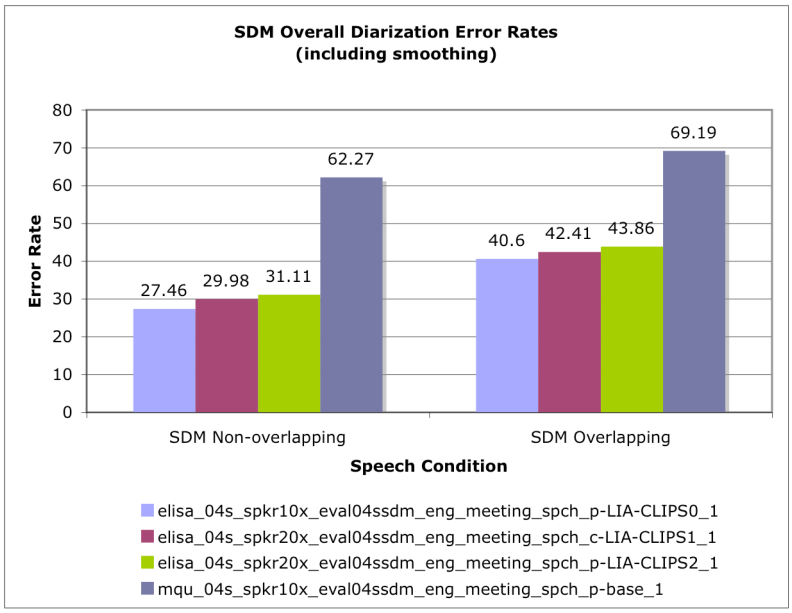
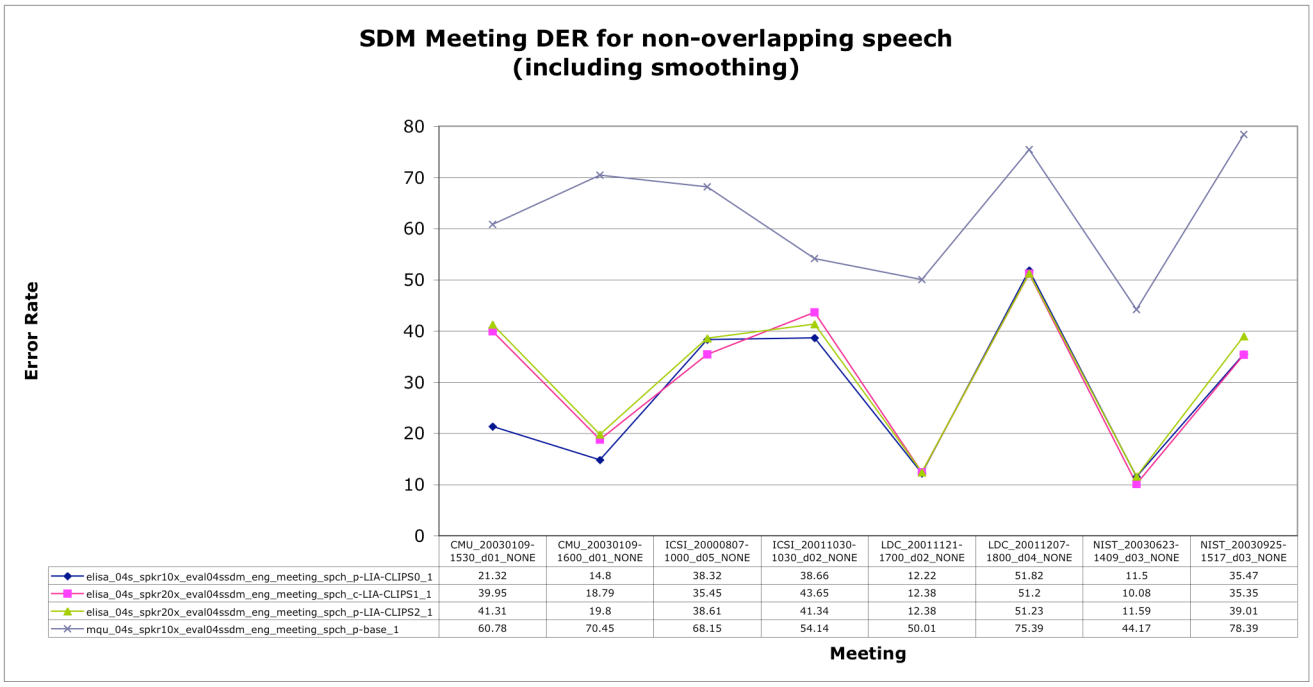
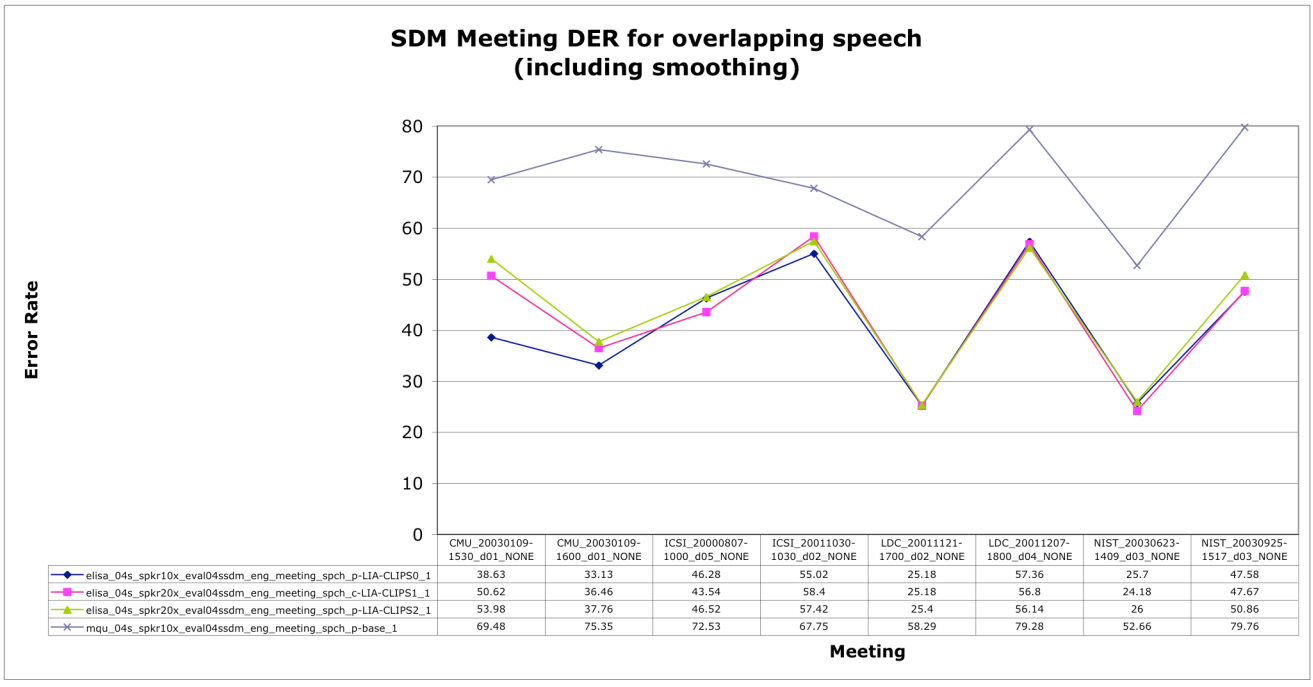


Figure 10 - SDM Overall Diarization Error Rates (including smoothing)



	CMU_20030109-1530_d01_NONE	CMU_20030109-1600_d01_NONE	ICSI_20000807-1000_d05_NONE	ICSI_20011030-1030_d02_NONE	LDC_20011121-1700_d02_NONE	LDC_20011207-1800_d04_NONE	NIST_20030623-1409_d03_NONE	NIST_20030925-1517_d03_NONE
elisa_04s_spkr10x_eval04ssdm_eng_meeting_sph_p-LIA-CLIPS0_1	21.32	14.8	38.32	38.66	12.22	51.82	11.5	35.47
elisa_04s_spkr20x_eval04ssdm_eng_meeting_sph_c-LIA-CLIPS1_1	39.95	18.79	35.45	43.65	12.38	51.2	10.08	35.35
elisa_04s_spkr20x_eval04ssdm_eng_meeting_sph_p-LIA-CLIPS2_1	41.31	19.8	38.61	41.34	12.38	51.23	11.59	39.01
mqu_04s_spkr10x_eval04ssdm_eng_meeting_sph_p-base_1	60.78	70.45	68.15	54.14	50.01	75.39	44.17	78.39

Figure 11 - SDM Meeting DER for non-overlapping speech (including smoothing)



	CMU_20030109-1530_d01_NONE	CMU_20030109-1600_d01_NONE	ICSI_20000807-1000_d05_NONE	ICSI_20011030-1030_d02_NONE	LDC_20011121-1700_d02_NONE	LDC_20011207-1800_d04_NONE	NIST_20030623-1409_d03_NONE	NIST_20030925-1517_d03_NONE
elisa_04s_spkr10x_eval04ssdm_eng_meeting_sph_p-LIA-CLIPS0_1	38.63	33.13	46.28	55.02	25.18	57.36	25.7	47.58
elisa_04s_spkr20x_eval04ssdm_eng_meeting_sph_c-LIA-CLIPS1_1	50.62	36.46	43.54	58.4	25.18	56.8	24.18	47.67
elisa_04s_spkr20x_eval04ssdm_eng_meeting_sph_p-LIA-CLIPS2_1	53.98	37.76	46.52	57.42	25.4	56.14	26	50.86
mqu_04s_spkr10x_eval04ssdm_eng_meeting_sph_p-base_1	69.48	75.35	72.53	67.75	58.29	79.28	52.66	79.76

Figure 12 - SDM Meeting DER for overlapping speech (including smoothing)

V COMPARISON BETWEEN MDM AND SDM CONDITIONS (ORIGINALLY REPORTED RESULTS)

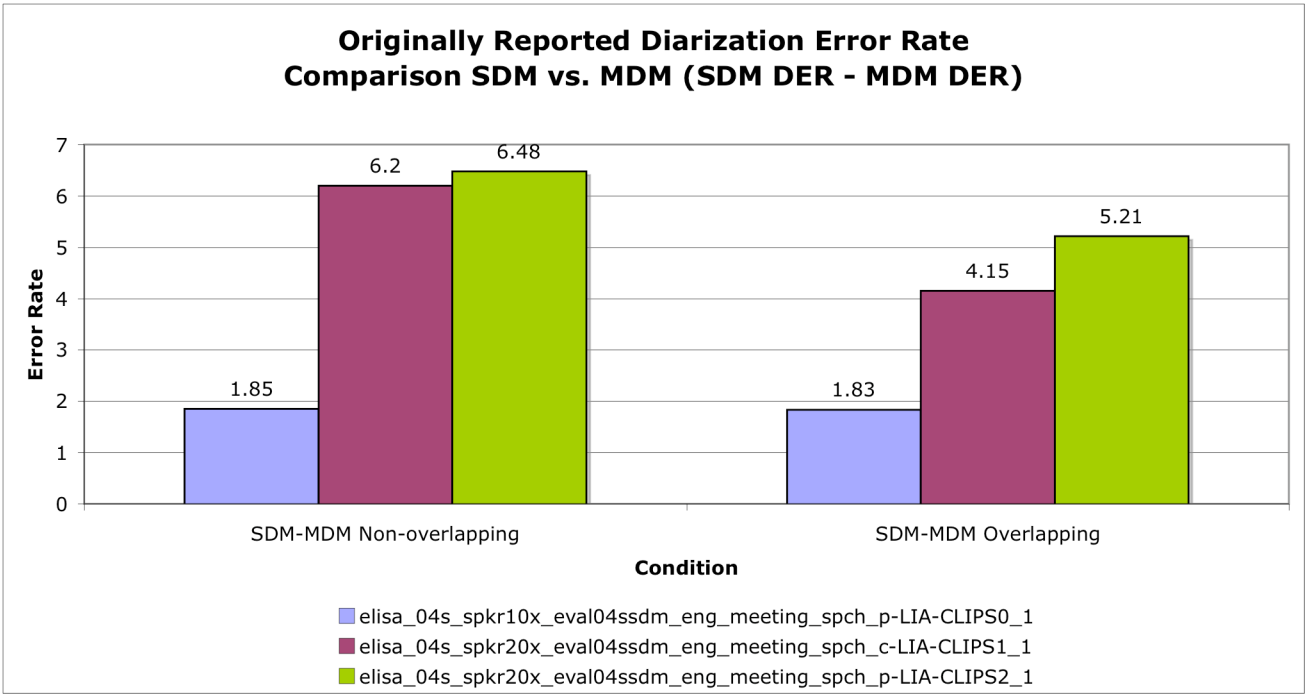


Figure 13 – Originally Reported Diarization Error Rate Comparison SDM vs. MDM

VI COMPARISON BETWEEN MDM AND SDM CONDITIONS (INCLUDING SMOOTHING)

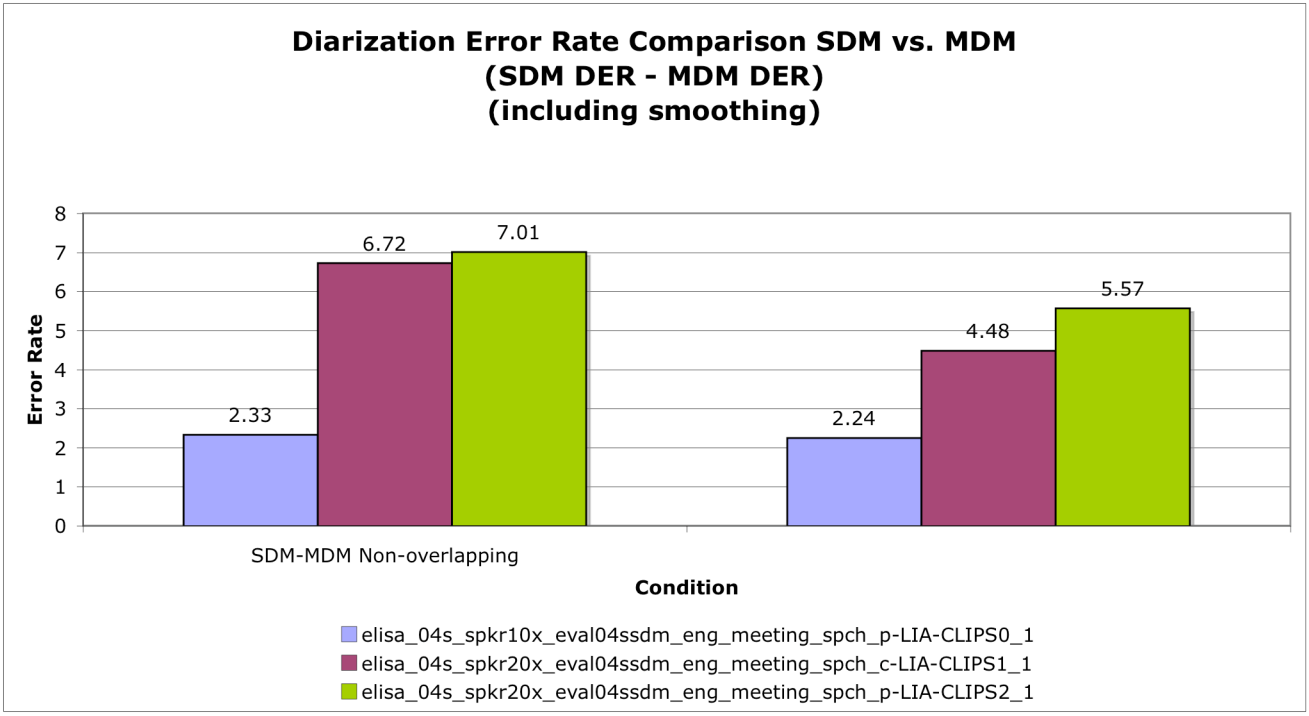


Figure 14 - Diarization Error Rate Comparison SDM vs. MDM (including smoothing)

APPENDIX B. STT MDM RESULTS – SCLITE NON-OVERLAPPING

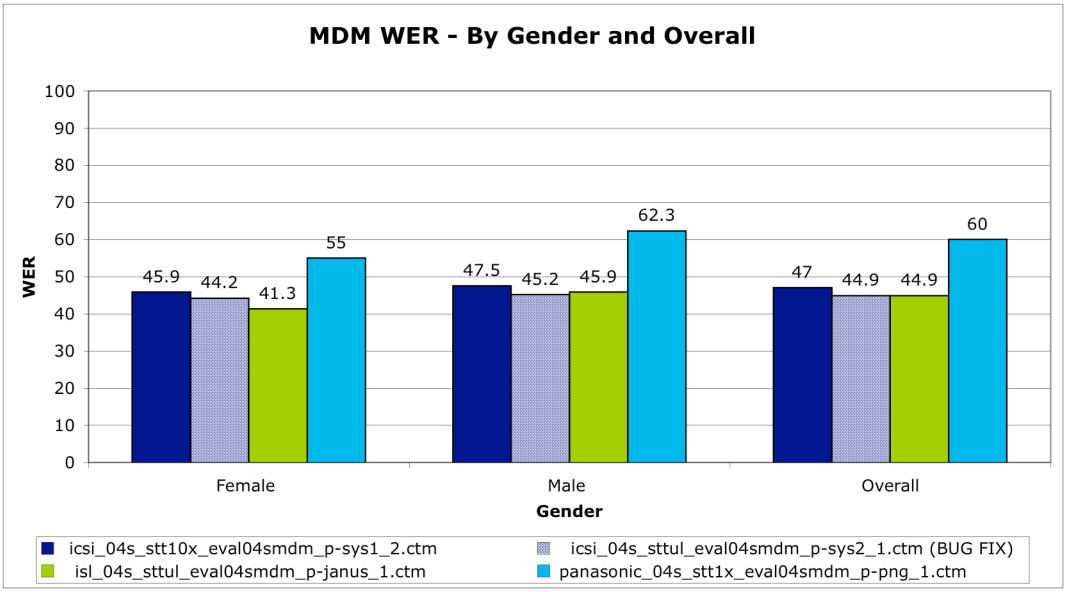
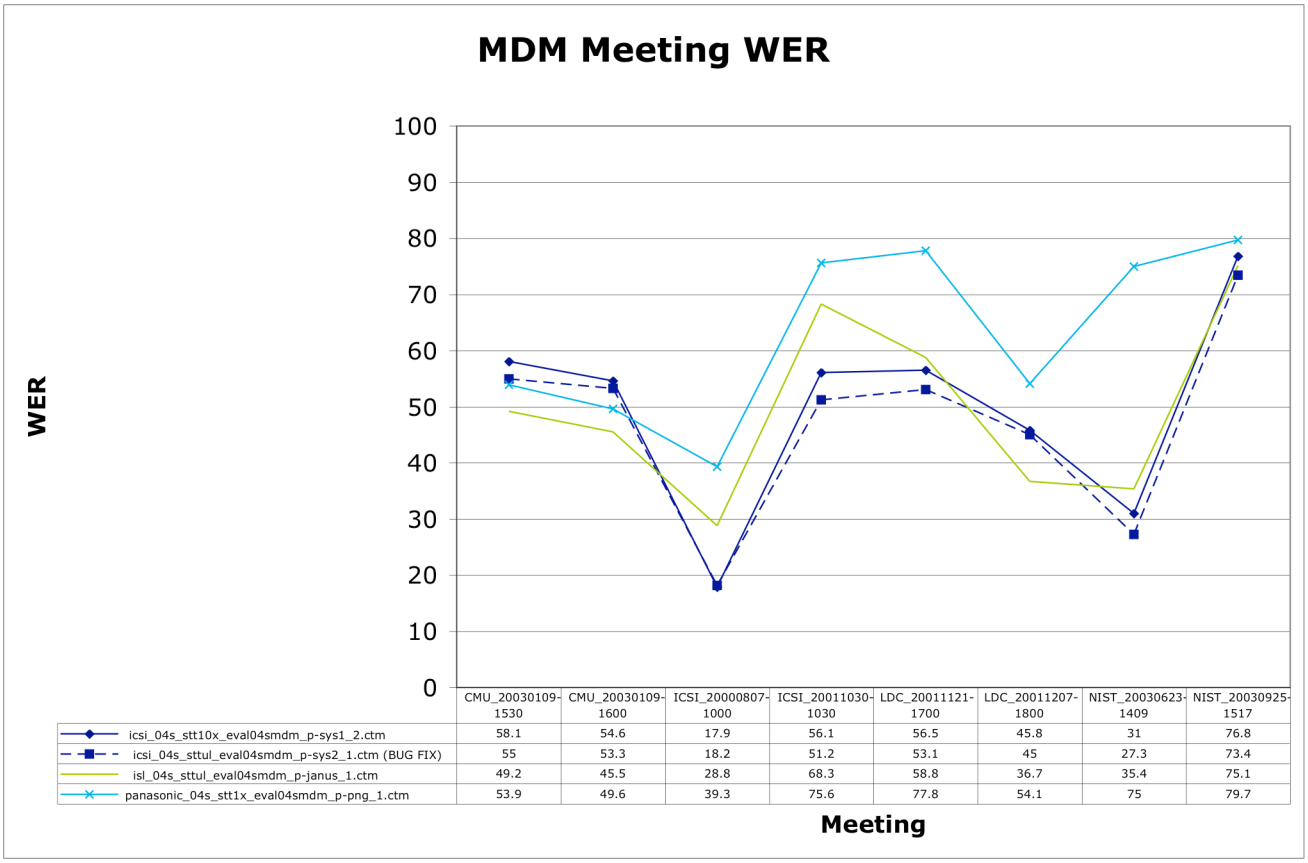


Figure 15 - MDM WER – By Gender and Overall



	CMU_20030109-1530	CMU_20030109-1600	ICSI_20000807-1000	ICSI_20011030-1030	LDC_20011121-1700	LDC_20011207-1800	NIST_20030623-1409	NIST_20030925-1517
icsi_04s_stt10x_eval04smdm_p-sys1_2.ctm	58.1	54.6	17.9	56.1	56.5	45.8	31	76.8
icsi_04s_sttul_eval04smdm_p-sys2_1.ctm (BUG FIX)	55	53.3	18.2	51.2	53.1	45	27.3	73.4
isl_04s_sttul_eval04smdm_p-janus_1.ctm	49.2	45.5	28.8	68.3	58.8	36.7	35.4	75.1
panasonic_04s_stt1x_eval04smdm_p-png_1.ctm	53.9	49.6	39.3	75.6	77.8	54.1	75	79.7

Figure 16 - MDM Meeting WER

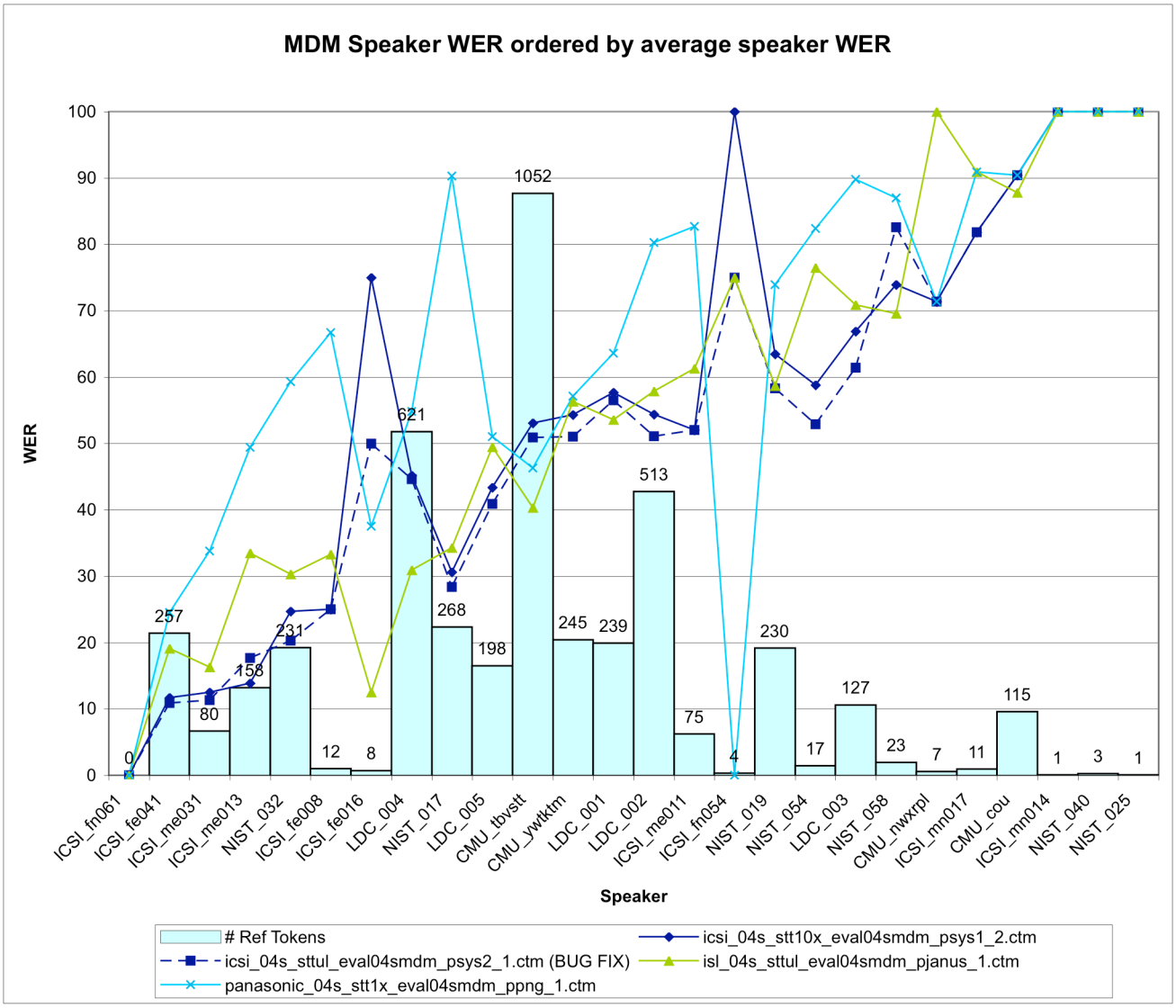


Figure 17 - MDM Speaker WER ordered by average speaker WER

APPENDIX C. STT SDM RESULTS – SCLITE NON-OVERLAPPING

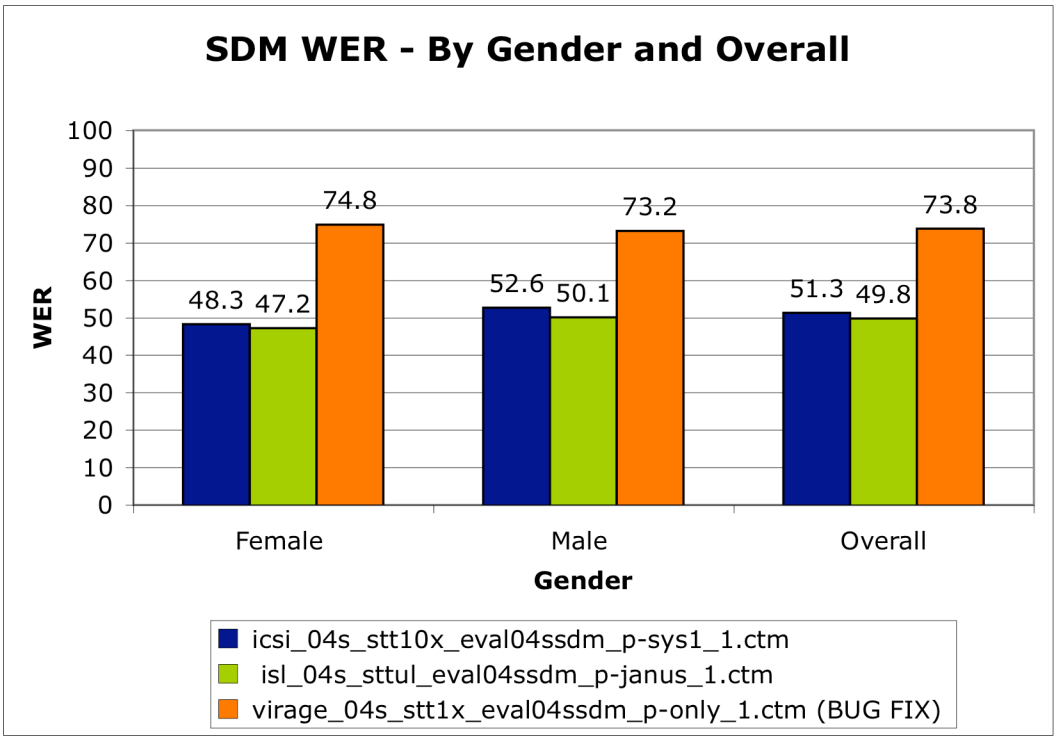
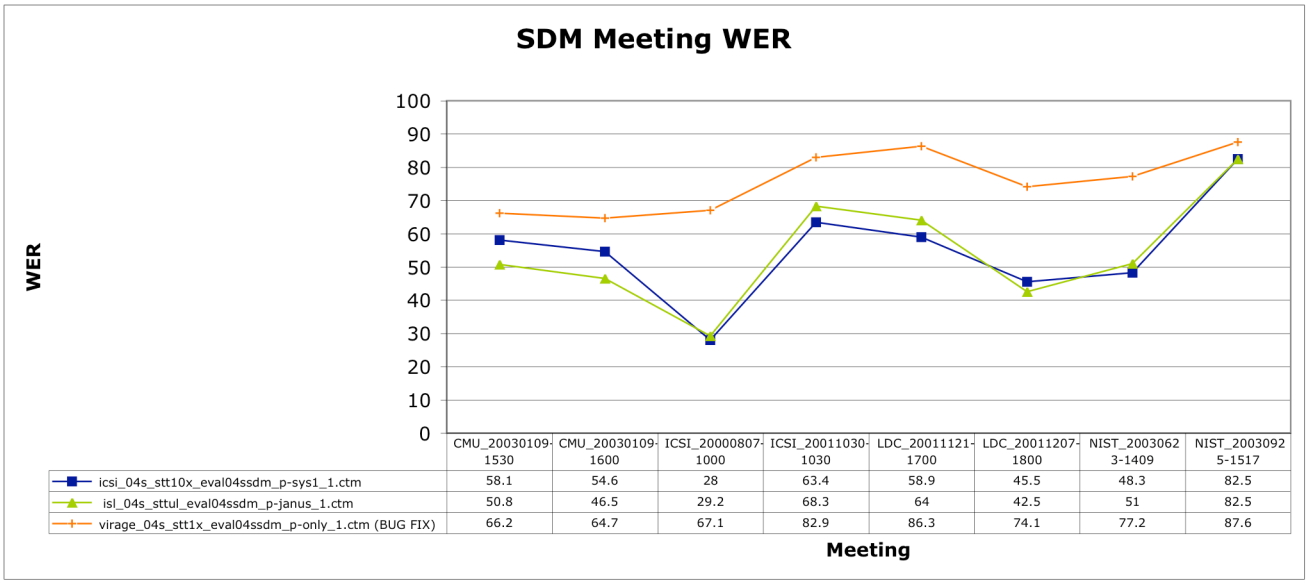


Figure 18 - SDM WER - By Gender and Overall



	CMU_20030109-1530	CMU_20030109-1600	ICSI_20000807-1000	ICSI_20011030-1030	LDC_20011121-1700	LDC_20011207-1800	NIST_20030623-1409	NIST_20030925-1517
icsi_04s_stt10x_eval04ssdm_p-sys1_1.ctm	58.1	54.6	28	63.4	58.9	45.5	48.3	82.5
isl_04s_sttul_eval04ssdm_p-janus_1.ctm	50.8	46.5	29.2	68.3	64	42.5	51	82.5
virage_04s_stt1x_eval04ssdm_p-only_1.ctm (BUG FIX)	66.2	64.7	67.1	82.9	86.3	74.1	77.2	87.6

Figure 19 - SDM Meeting WER

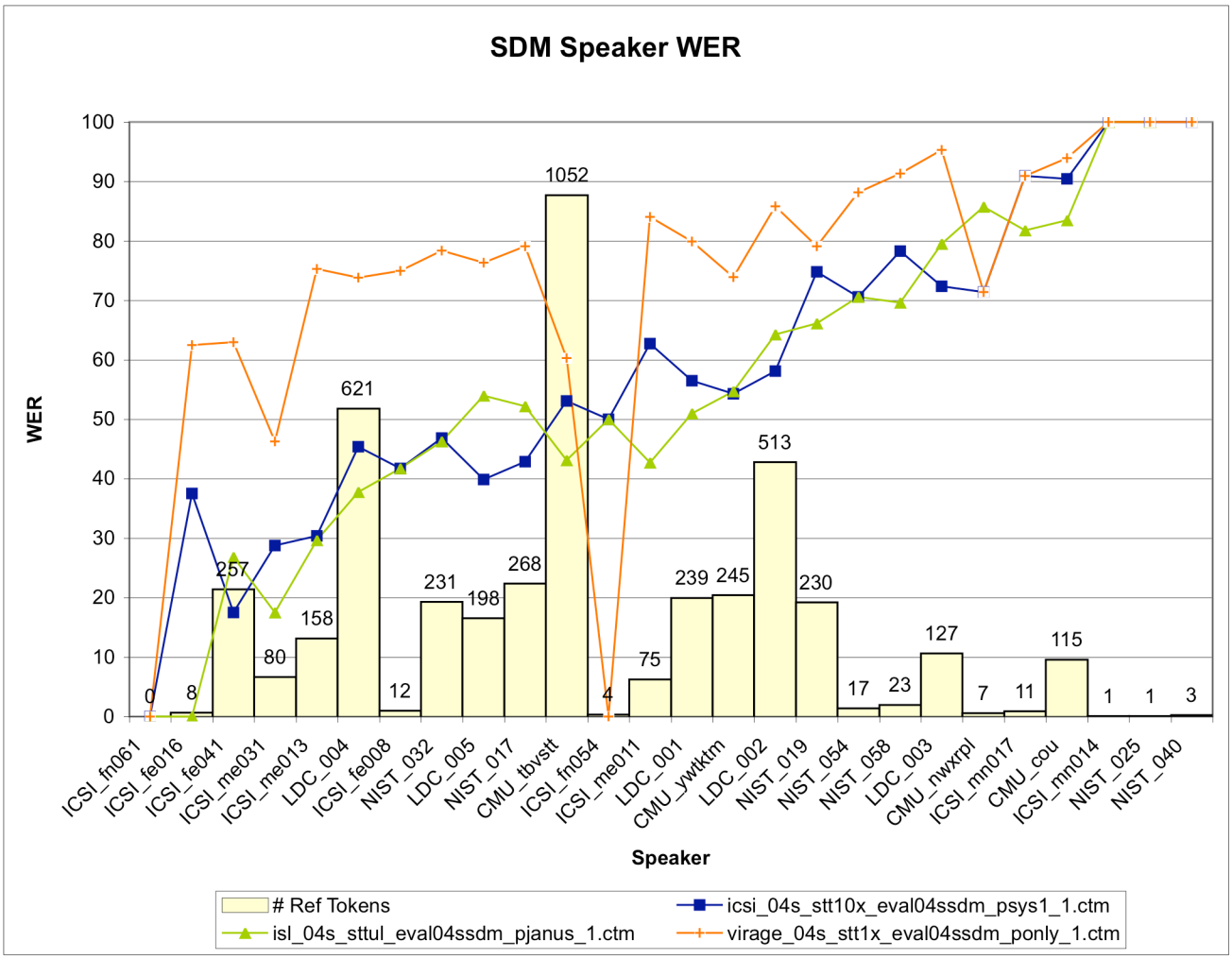


Figure 20 - SDM Speaker WER

APPENDIX D. STT IHM RESULTS – SCLITE NON-OVERLAPPING

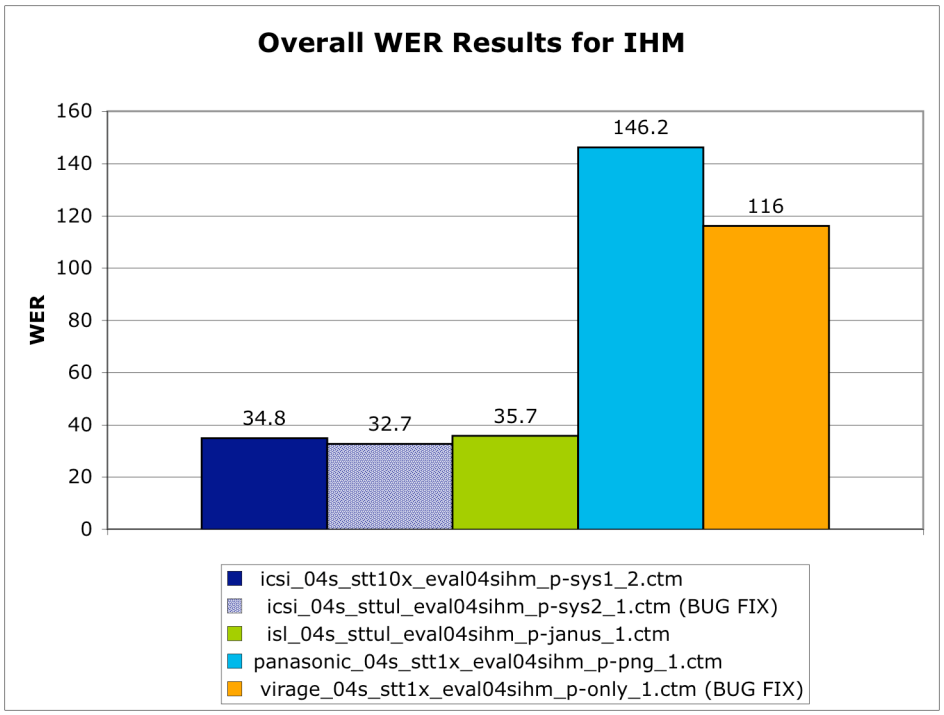


Figure 21 - Overall WER Results for IHM

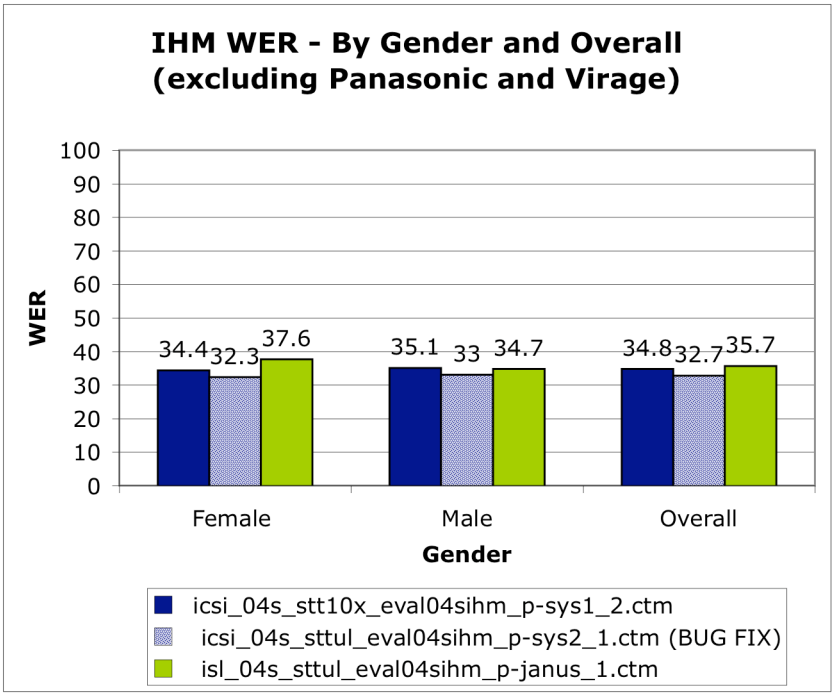
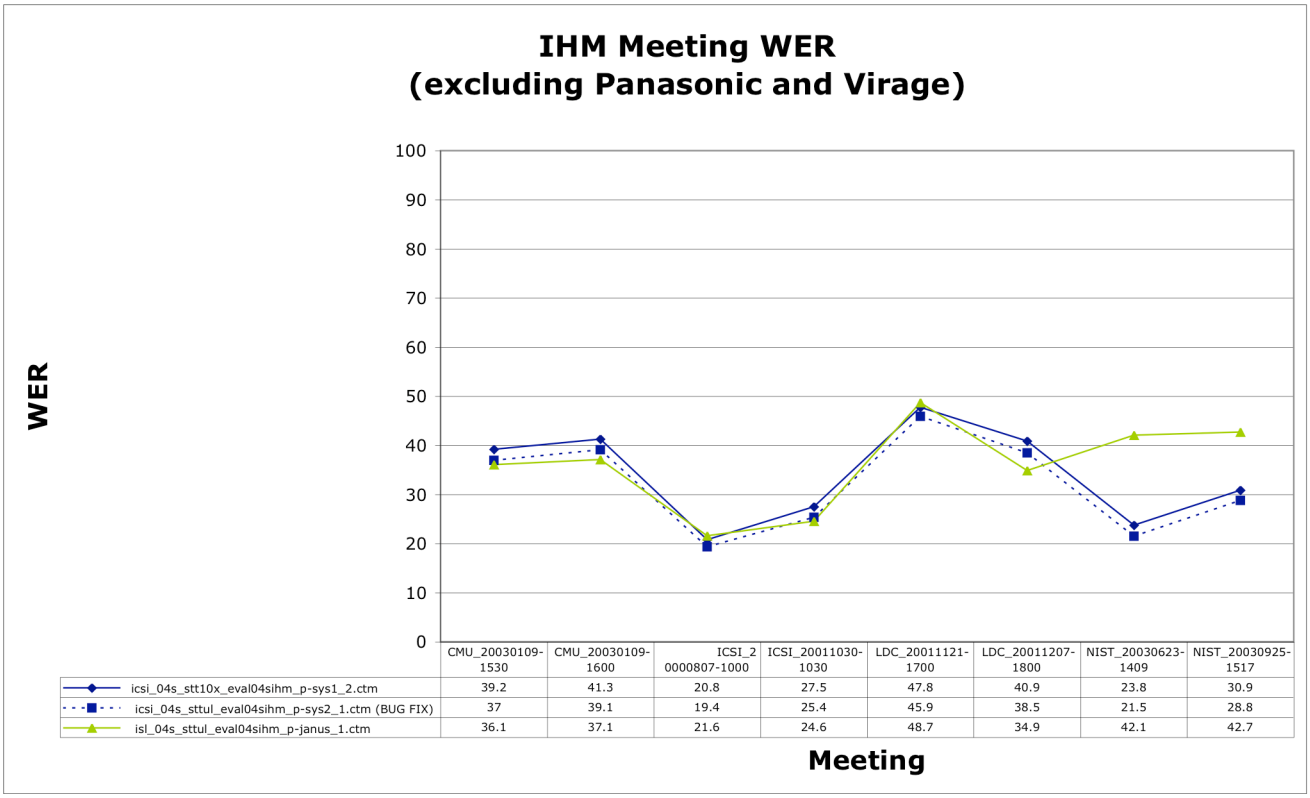


Figure 22 - IHM WER - By Gender and Overall (excluding Panasonic and Virage)



	CMU_20030109-1530	CMU_20030109-1600	ICSI_20000807-1000	ICSI_20011030-1030	LDC_20011121-1700	LDC_20011207-1800	NIST_20030623-1409	NIST_20030925-1517
icsi_04s_stt10x_eval04sihm_p-sys1_2.ctm	39.2	41.3	20.8	27.5	47.8	40.9	23.8	30.9
icsi_04s_sttul_eval04sihm_p-sys2_1.ctm (BUG FIX)	37	39.1	19.4	25.4	45.9	38.5	21.5	28.8
isl_04s_sttul_eval04sihm_p-janus_1.ctm	36.1	37.1	21.6	24.6	48.7	34.9	42.1	42.7

Figure 23 - IHM Meeting WER (excluding Panasonic and Virage)

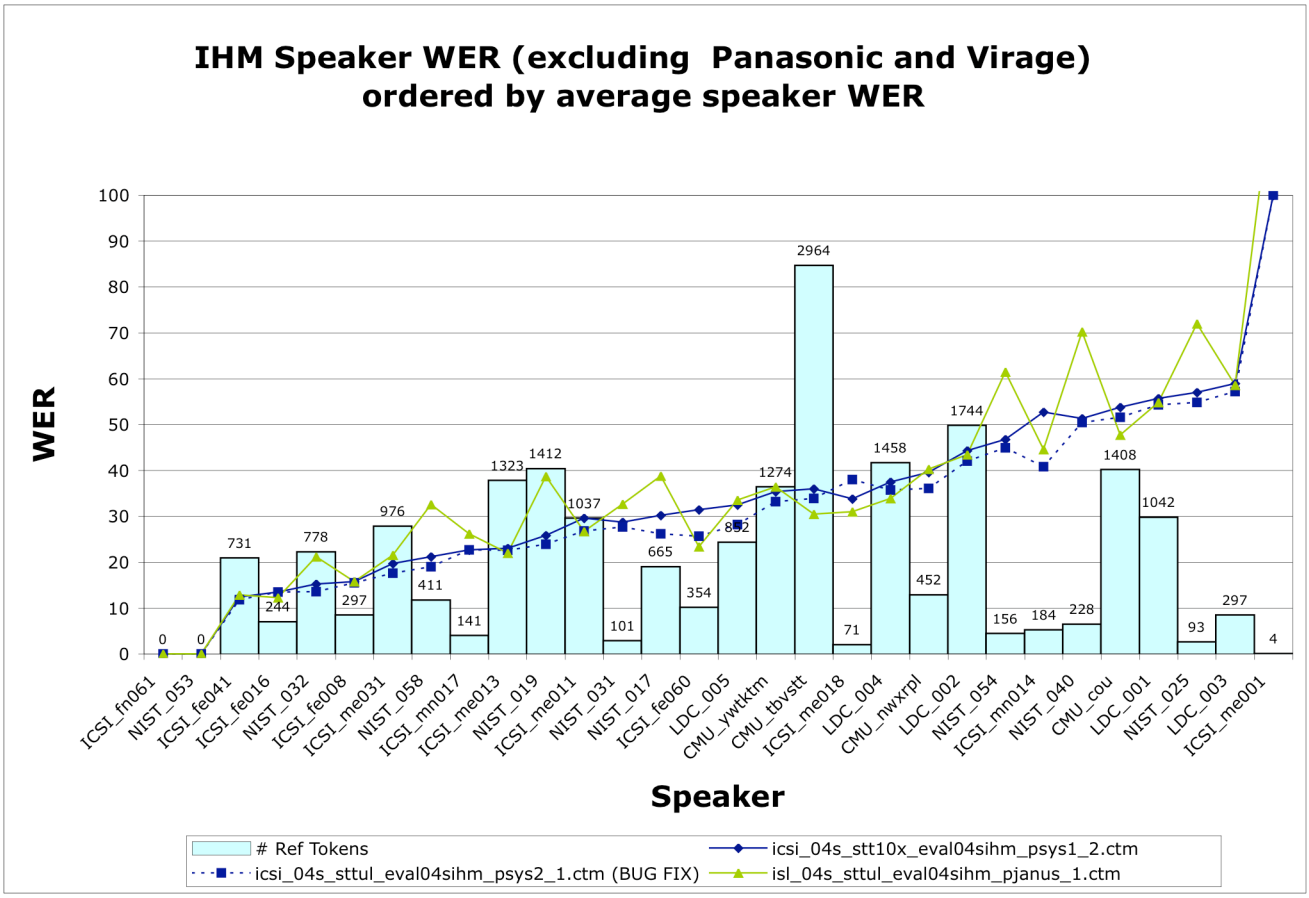


Figure 24 - IHM Speaker WER (excluding Panasonic and Virage)

APPENDIX E. CROSS-CONDITION COMPARISON – SCLITE NON-OVERLAPPING

I COMPARISON OF OVERALL WER FOR THE THREE MIC CONDITIONS

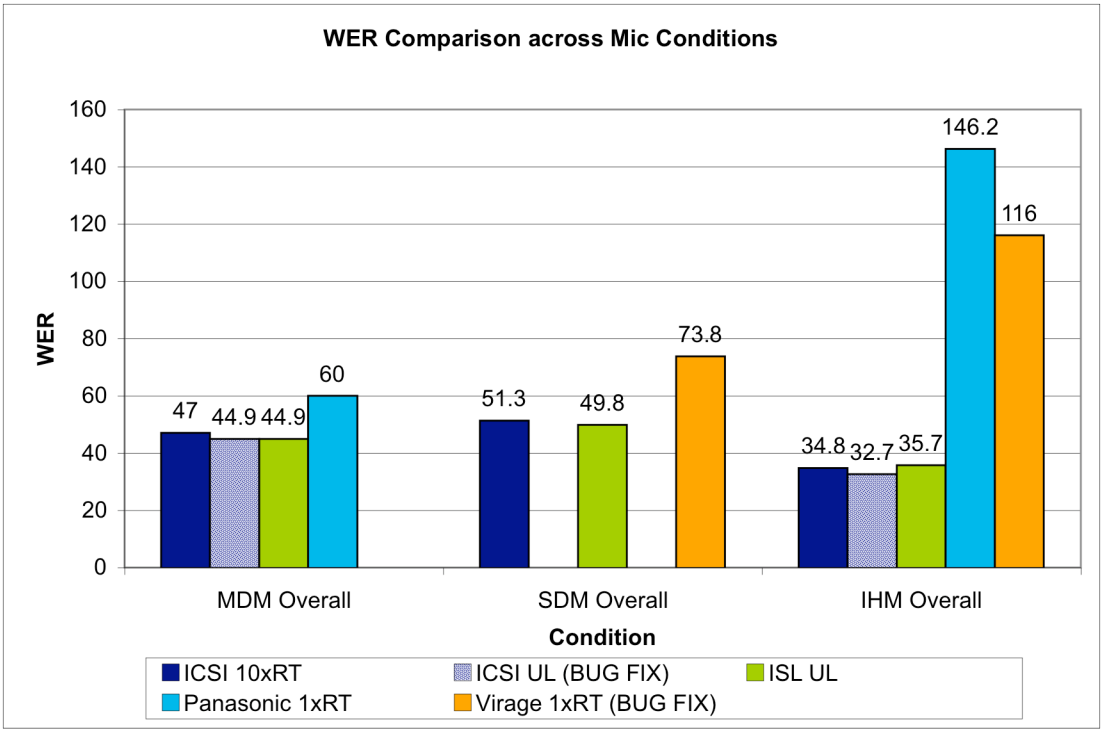


Figure 25 - WER Comparison across Mic Conditions

II SDM VS. MDM

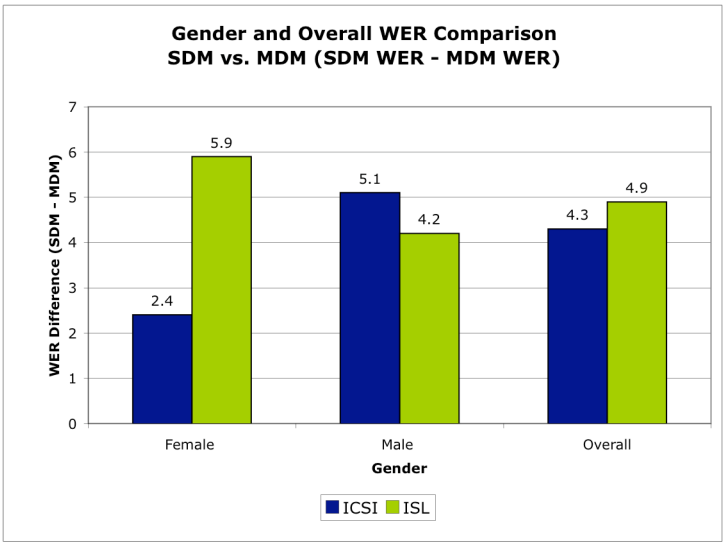
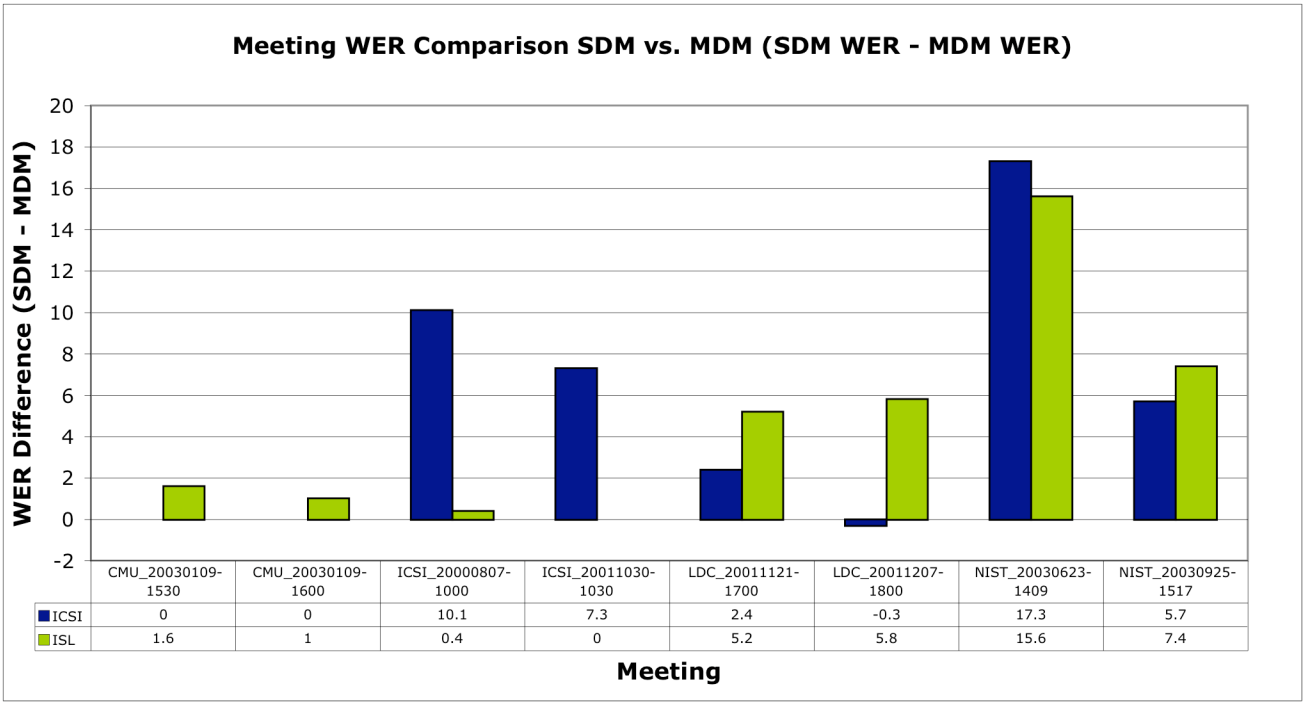


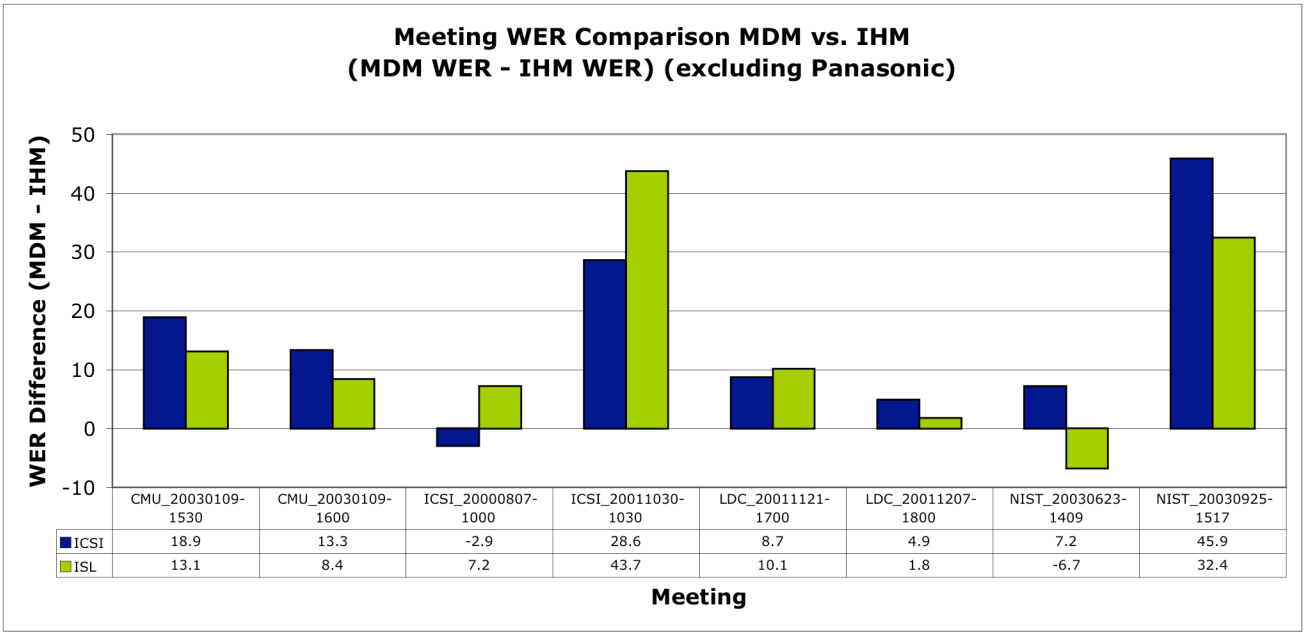
Figure 26 –Gender and Overall WER Comparison SDM vs. MDM



	CMU_20030109-1530	CMU_20030109-1600	ICSI_20000807-1000	ICSI_20011030-1030	LDC_20011121-1700	LDC_20011207-1800	NIST_20030623-1409	NIST_20030925-1517
ICSI	0	0	10.1	7.3	2.4	-0.3	17.3	5.7
ISL	1.6	1	0.4	0	5.2	5.8	15.6	7.4

Figure 27 – Meeting WER Comparison SDM vs. MDM

III MDM VS. IHM



	CMU_20030109-1530	CMU_20030109-1600	ICSI_20000807-1000	ICSI_20011030-1030	LDC_20011121-1700	LDC_20011207-1800	NIST_20030623-1409	NIST_20030925-1517
ICSI	18.9	13.3	-2.9	28.6	8.7	4.9	7.2	45.9
ISL	13.1	8.4	7.2	43.7	10.1	1.8	-6.7	32.4

Figure 28 – Meeting WER Comparison MDM vs. IHM (excluding Panasonic)

APPENDIX F. EXPERIMENTAL STT OVERLAPPING SPEECH SCORING RESULTS FOR THE MDM CONDITION

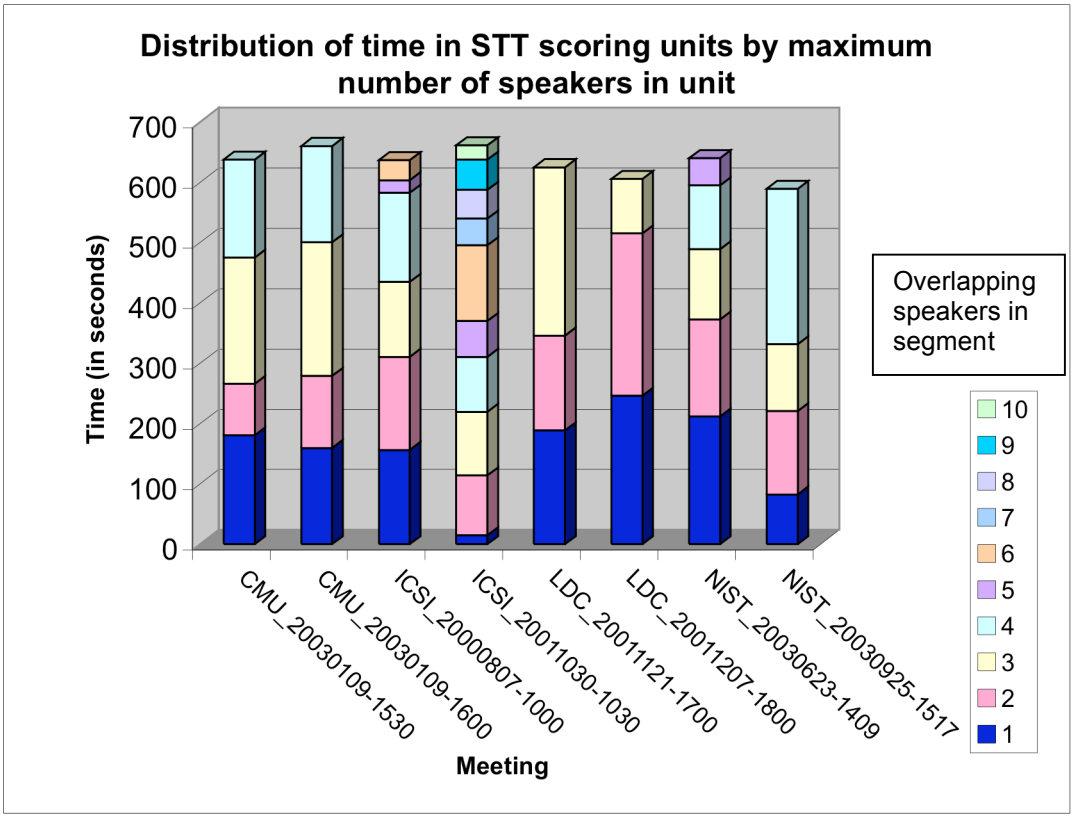


Figure 29 - Distribution of time in STT scoring units by maximum number of speakers in unit

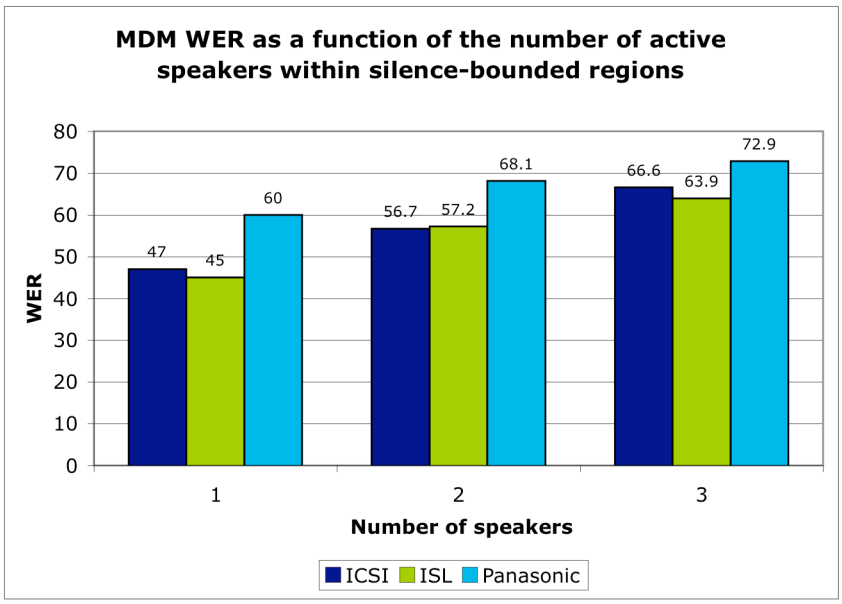


Figure 30 – WER as a function of the number of active speakers within silence-bounded regions

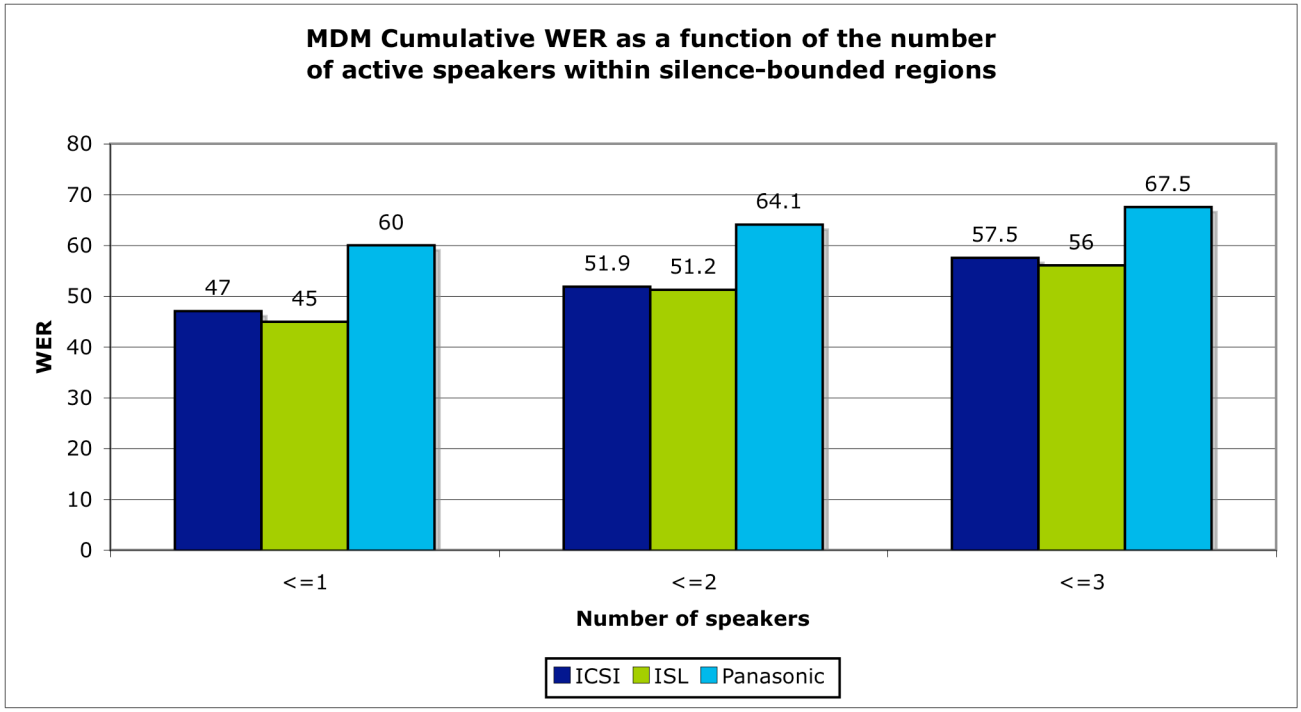
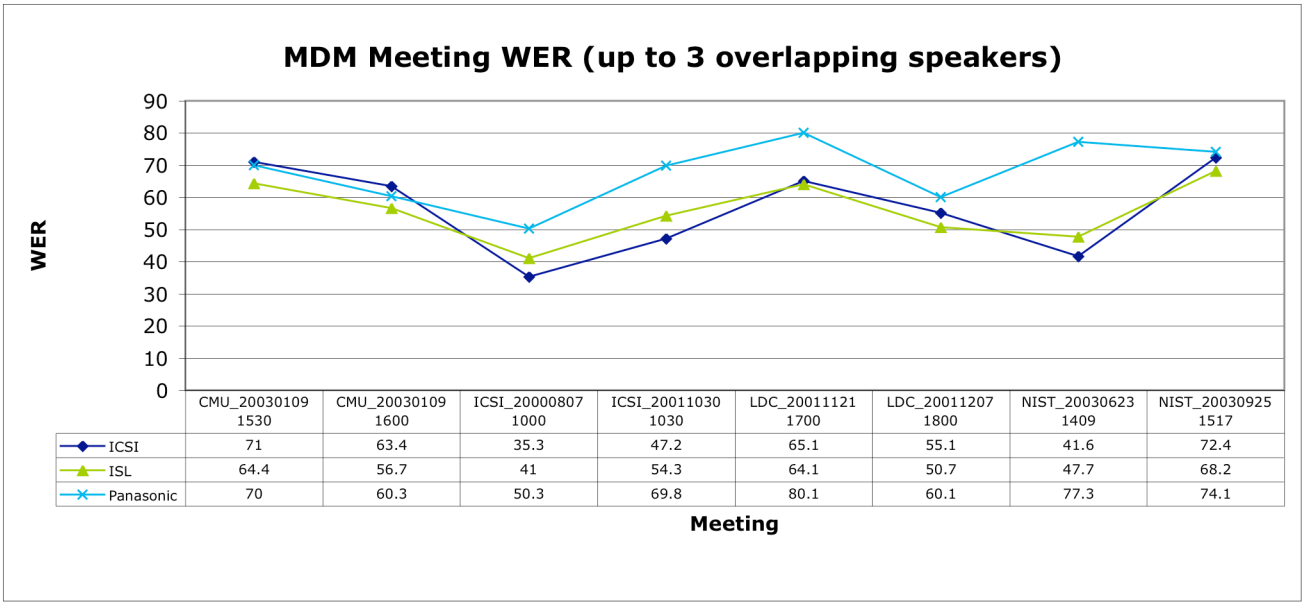


Figure 31 - Cumulative WER as a function of the number of active speakers within silence-bounded regions



	CMU_20030109 1530	CMU_20030109 1600	ICSI_20000807 1000	ICSI_20011030 1030	LDC_20011121 1700	LDC_20011207 1800	NIST_20030623 1409	NIST_20030925 1517
ICSI	71	63.4	35.3	47.2	65.1	55.1	41.6	72.4
ISL	64.4	56.7	41	54.3	64.1	50.7	47.7	68.2
Panasonic	70	60.3	50.3	69.8	80.1	60.1	77.3	74.1

Figure 32 - Meeting WER (up to 3 overlapping speakers)

APPENDIX G. EVALUATION TEST SET MEETINGS

Meeting Excerpt	Duration	Participants	Male	Female	Native	Non-native	Distant Mics	Notes
CMU								
20030109-1530	11.02	4	3	1	4	0	1	unmoderated discussion of EU, planes - college students
20030109-1600	11.10	4	3	1	4	0	1	unmoderated discussion of SUVs, bikes - college students
	Unique	4	3	1	4	0		
ICSI								
20000807-1000	11.37	5	3	2	5	0	6	technical meeting on meeting room setup between ICSI / UW
20011030-1030	11.50	10	6	4	7	3	6	technical meeting on transcription/annotation
	Unique	12	7	5	9	3		
LDC								
20011121_1700	11.03	3	3	0	3	0	10	moderated discussions of memories of two war veterans
20011207_1800	11.62	3	1	2	3	0	4	moderated discussion of political issues surrounding war on Iraq
	Unique	5	3	2	5	0		
NIST								
20030623-1409	11.23	6	3	3	5	1	7	NIST visualization group staff meeting
20030925-1517	11.03	5	2	3	5	0	7	news team scenario
	Unique	10	5	5	9	1		
Totals	89.90	40	24	16	36	4		
	Unique	31	18	13	27	4		