The Meeting Project at ICSI

Nelson Morgan[1,4] Don Baron[1,4] Jane Edwards[1,4] Dan Ellis[1,2] David Gelbart[1,4]
Adam Janin[1,4] Thilo Pfau[1] Elizabeth Shriberg[1,3] Andreas Stolcke[1,3]
[1]International Computer Science Institute, Berkeley, CA
[2]Columbia University, New York, NY
[3]SRI International, Menlo Park, CA
[4]University of California at Berkeley, Berkeley, CA

{morgan,dbaron,edwards,dpwe,gelbart,janin,tpfau,ees,stolcke}@icsi.berkeley.edu

ABSTRACT

In collaboration with colleagues at UW, OGI, IBM, and SRI, we are developing technology to process spoken language from informal meetings. The work includes a substantial data collection and transcription effort, and has required a nontrivial degree of infrastructure development. We are undertaking this because the new task area provides a significant challenge to current HLT capabilities, while offering the promise of a wide range of potential applications. In this paper, we give our vision of the task, the challenges it represents, and the current state of our development, with particular attention to automatic transcription.

1. THE TASK

We are primarily interested in the processing (transcription, query, search, and structural representation) of audio recorded from informal, natural, and even impromptu meetings. By "informal" we mean conversations between friends and acquaintances that do not have a strict protocol for the exchanges. By "natural" we mean meetings that would have taken place regardless of the recording process, and in acoustic circumstances that are typical for such meetings. By "impromptu" we mean that the conversation may take place without any preparation, so that we cannot require special instrumentation to facilitate later speech processing (such as close-talking or array microphones). A plausible image for such situations is a handheld device (PDA, cell phone, digital recorder) that is used when conversational partners agree that their discussion should be recorded for later reference.

Given these interests, we have been recording and transcribing a series of meetings at ICSI. The recording room is one of ICSI's standard meeting rooms, and is instrumented with both close-talking and distant microphones. Close-mic'd recordings will support research on acoustic modeling, language modeling, dialog modeling, etc., without having to immediately solve the difficulties of far-field microphone speech recognition. The distant microphones are included to facilitate the study of these deep acoustic problems, and to provide a closer match to the operating conditions ultimately envisaged. These ambient signals are collected by 4 omnidirectional PZM table-mount microphones, plus a "dummy" PDA that has two inexpensive microphone elements. In addition to these 6 distant microphones, the audio setup permits a maximum of 9 close-talking microphones to be simultaneously recorded. A meeting recording infrastructure is also being put in place at Columbia University, at SRI International, and by our colleagues at the University of Washington. Recordings from all sites will be transcribed using standards evolved in discussions that also involved IBM (who also have committed to assist in the transcription task). Colleagues at NIST have been in contact with us to further standardize these choices, since they intend to conduct related collection efforts.

A segment from a typical discussion recorded at ICSI is included below in order to give the reader a more concrete sense of the task. Utterances on the same line separated by a slash indicate some degree of overlapped speech.

> **A:** Ok. So that means that for each utterance, .. we'll need the time marks.
>
> **E:** Right. / **A:** the start and end of each utterance.
>
> [a few turns omitted]
>
> **E:** So we - maybe we should look at the um .. the tools that Mississippi State has.
>
> **D:** Yeah.
>
> **E:** Because, I - I - I know that they published .. um .. annotation tools.
>
> **A:** Well, X-waves have some as well, .. but they're pretty low level .. They're designed for uh - / **D:** phoneme / **A:** for phoneme-level / **D:** transcriptions. Yeah.
>
> **J:** I should -
>
> **A:** Although, they also have a nice tool for - .. that could be used for speaker change marking.
>
> **D:** There's a - there are - there's a whole bunch of tools
>
> **J:** Yes. / **D:** web page, where they have a listing. **D:** like 10 of them or something.
>
> **J:** Are you speaking about Mississippi State per se? or
>
> **D:** No no no, there's some .. I mean, there just - there are - there are a lot of / **J:** Yeah.
>
> **J:** Actually, I wanted to mention - / **D:** (??)
>
> **J:** There are two projects, which are .. international .. huge projects focused on this kind of thing, actually .. one of them's MATE, one of them's EAGLES .. and um.
>
> **D:** Oh, EAGLES.
>
> **D:** (??) / **J:** And both of them have
>
> **J:** You know, I shou-, I know you know about the big book.
>
> **E:** Yeah.
>
> **J:** I think you got it as a prize or something.
>
> **E:** Yeah. / **D:** Mhm.
>
> **J:** Got a surprise. {laugh} {J. thought "as a prize" sounded like "surprise"}

Note that interruptions are quite frequent; this is, in our experience, quite common in informal meetings, as is acoustic overlap

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **2001**	2. REPORT TYPE	3. DATES COVERED **00-00-2001 to 00-00-2001**

4. TITLE AND SUBTITLE **The Meeting Project at ICSI**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Computer Science Department,Columbia University,New York City,NY,10027**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**7**	

between speakers (see the section on error rates in overlap regions).

2. THE CHALLENGES

While having a searchable, annotatable record of impromptu meetings would open a wide range of applications, there are significant technical challenges to be met; it would not be far from the truth to say that the problem of generating a full representation of a meeting is "AI complete", as well as "ASR complete". We believe, however, that our community can make useful progress on a range of associated problems, including:

- ASR for very informal conversational speech, including the common overlap problem.
- ASR from far-field microphones - handling the reverberation and background noise that typically bedevil distant mics, as well as the acoustic overlap that is more of a problem for microphones that pick up several speakers at approximately the same level.
- Segmentation and turn detection - recovering the different speakers and turns, which also is more difficult with overlaps and with distant microphones (although inter-microphone timing cues can help here).
- Extracting nonlexical information such as speaker identification and characterization, voice quality variation, prosody, laughter, etc.
- Dialog abstraction - making high-level models of meeting 'state'; identifying roles among participants, classifying meeting types, etc. [2].
- Dialog analysis - identification and characterization of fine-scale linguistic and discourse phenomena [3][10].
- Information retrieval from errorful meeting transcriptions - topic change detection, topic classification, and query matching.
- Summarization of meeting content [14] - representation of the meeting structure from various perspectives and at various scales, and issues of navigation in thes representations.
- Energy and memory resource limitation issues that arise in the robust processing of speech using portable devices [7].

Clearly we and others working in this area (e.g., [15]) are at an early stage in this research. However, the remainder of this paper will show that even a preliminary effort in recording, manually transcribing, and recognizing data from natural meetings has provided some insight into at least a few of these problems.

3. DATA COLLECTION AND HUMAN TRANSCRIPTION

Using the data collection setup described previously, we have been recording technical meetings at ICSI. As of this writing we have recorded 38 meetings for a total of 39 hours. Note that there are separate microphones for each participant in addition to the 6 far-field microphones, and there can be as many as 15 open channels. Consequently the sound files comprise hundreds of hours of recorded audio. The total number of participants in all meetings is 237, and there were 49 unique speakers. The majority of the meetings recorded so far have either had a focus on "Meeting Recorder" (that is, meetings by the group working on this technology) or "Robustness" (primarily concerned with ASR robustness to acoustic effects such as additive noise). A smaller number of other meeting types at ICSI were also included.

In addition to the spontaneous recordings, we asked meeting participants to read digit strings taken from a TI digits test set. This was done to facilitate research in far-field microphone ASR, since we expect this to be quite challenging for the more unconstrained case. At the start or end of each meeting, each participant read 20 digit strings.

Once the data collection was in progress, we developed a set of procedures for our initial transcription. The transcripts are word-level transcripts, with speaker identifier, and some additional information: overlaps, interrupted words, restarts, vocalized pauses, backchannels, and contextual comments, and nonverbal events (which are further subdivided into vocal types such as cough and laugh, and nonvocal types such as door slams and clicks). Each event is tied to the time line through use of a modified version of the "Transcriber" interface (described below). This Transcriber window provides an editing space at the top of the screen (for adding utterances, etc), and the wave form at the bottom, with mechanisms for flexibly navigating through the audio recording, and listening and re-listening to chunks of virtually any size the user wishes.

The typical process involves listening to a stretch of speech until a natural break is found (e.g., a long pause when no one is speaking). The transcriber separates that chunk from what precedes and follows it by pressing the Return key. Then he or she enters the speaker identifier and utterance in the top section of the screen. The interface is efficient and easy to use, and results in an XML representation of utterances (and other events) tied to time tags for further processing.

The "Transcriber" interface [13] is a well-known tool for transcription, which enables the user to link acoustic events to the wave form. However, the official version is designed only for single-channel audio. As noted previously, our application records up to 15 parallel sound tracks generated by as many as 9 speakers, and we wanted to capture the start and end times of events on each channel as precisely as possible and independently of one another across channels. The need to switch between multiple audio channels to clarify overlaps, and the need to display the time course of events on independent channels required extending the "Transcriber" interface in two ways. First, we added a menu that allows the user to switch the playback between a number of audio files (which are all assumed to be time synchronized). Secondly, we split the time-linked display band into as many independent display bands as there are channels (and/or independent layers of time-synchronized annotation). Speech and other events on each of the bands can now be time-linked to the wave form with complete freedom and totally independently of the other bands. This enables much more precise start and end times for acoustic events.

See [8] for links to screenshots of these extensions to Transcriber (as well as to other updates about our project).

In the interests of maximal speed, accuracy and consistency, the transcription conventions were chosen so as to be: quick to type, related to standard literary conventions where possible (e.g., - for interrupted word or thought, .. for pause, using standard orthography rather than IPA), and minimalist (requiring no more decisions by transcribers than absolutely necessary).

After practice with the conventions and the interface, transcribers achieved a 12:1 ratio of transcription time to speech time. The amount of time required for transcription of spoken language is known to vary widely as a function of properties of the discourse (amount of overlap, etc.), and amount of detailed encoding (prosodics, etc.), with estimates ranging from 10:1 for word-level with minimal added information to 20:1, for highly detailed discourse transcriptions (see [4] for details).

In our case, transcribers encoded minimal added detail, but had two additional demands: marking boundaries of time bins, and switching between audio channels to clarify the many instances of overlapping speech in our data. We speeded the marking of time bins by providing them with an automatically segmented version (described below) in which the segmenter provided a preliminary set of speech/nonspeech labels. Transcribers indicated that the pre-segmentation was correct sufficiently often that it saved them time.

After the transcribers finished, their work was edited for consistency and completeness by a senior researcher. Editing involved checking exhaustive listings of forms in the data, spell checking, and use of scripts to identify and automatically encode certain distinctions (e.g., the distinction between vocalized nonverbal events, such as cough, and nonvocalized nonverbal events, like door slams). This step requires on average about 1:1 - one minute of editing for each minute of speech.

Using these methods and tools, we have currently transcribed about 12 hours out of our 39 hours of data. Other data have been sent to IBM for a rough transcription using commercial transcribers, to be followed by a more detailed process at ICSI. Once this becomes a routine component of our process, we expect it to significantly reduce the time requirements for transcription at ICSI.

4. AUTOMATIC TRANSCRIPTION

As a preliminary report on automatic word transcription, we present results for six example meetings, totalling nearly 7 hours of speech, 36 total speakers, and 15 unique speakers (since many speakers participated in multiple meetings). Note that these results are preliminary only; we have not yet had a chance to address the many obvious approaches that could improve performance. In particular, in order to facilitate efforts in alignment, pronunciation modeling, language modeling, etc., we worked only with the close-mic'd data. In most common applications of meeting transcription (including those that are our chief targets in this research) such a microphone arrangement may not be practical. Nevertheless we hope the results using the close microphone data will illustrate some basic observations we have made about meeting data and its automatic transcription.

4.1 Recognition system

The recognizer was a stripped-down version of the large-vocabulary conversational speech recognition system fielded by SRI in the March 2000 Hub-5 evaluation [11]. The system performs vocal-tract length normalization, feature normalization, and speaker adaptation using all the speech collected on each channel (i.e., from one speaker, modulo cross-talk). The acoustic model consisted of gender-dependent, bottom-up clustered (genonic) Gaussian mixtures. The Gaussian means are adapted by a linear transform so as to maximize the likelihood of a phone-loop model, an approach that is fast and does not require recognition prior to adaptation. The adapted models are combined with a bigram language model for decoding. We omitted more elaborate adaptation, cross-word triphone modeling, and higher-order language and duration models from the full SRI recognition system as an expedient in our initial recognition experiments (the omitted steps yield about a 20% relative error rate reduction on Hub-5 data).

It should be noted that both the acoustic models and the language model of the recognizer were identical to those used in the Hub-5 domain. In particular, the acoustic front-end assumes a telephone channel, requiring us to downsample the wide-band signals of the meeting recordings. The language model contained about 30,000 words and was trained on a combination of Switchboard, CallHome English and Broadcast News data, but was not tuned for or augmented by meeting data.

4.2 Speech segmentation

As noted above, we are initially focusing on recognition of the individual channel data. Such data provide an upper bound on recognition accuracy if speaker segmentation were perfect, and constitute a logical first step for obtaining high quality forced alignments against which to evaluate performance for both near- and far-field microphones. Individual channel recordings were partitioned into "segments" of speech, based on a "mixed" signal (addition of the individual channel data, after an overall energy equalization factor per channel). Segment boundary times were determined either by an automatic segmentation of the mixed signal followed by hand-correction, or by hand-correction alone. For the automatic case, the data was segmented with a speech/nonspeech detector consisting of an extension of an approach using an ergodic hidden Markov model (HMM) [1]. In this approach, the HMM consists of two main states, one representing "speech" and one representing "nonspeech" and a number of intermediate states that are used to model the time constraints of the transitions between the two main states. In our extension, we are incorporating mixture densities rather than single Gaussians. This appears to be useful for the separation of foreground from background speech, which is a serious problem in these data.

The algorithm described above was trained on the speech/nonspeech segmentation provided manually for the first meeting that was transcribed. It was used to provide segments of speech for the manual transcribers, and later for the recognition experiments. Currently, for simplicity and to debug the various processing steps, these segments are synchronous across channels. However, we plan to move to segments based on separate speech/nonspeech detection in each individual channel. The latter approach should provide better recognition performance, since it will eliminate cross-talk in segments in which one speaker may say only a backchannel (e.g. "uhhuh") while another speaker is talking continuously.

Performance was scored for the spontaneous conversational portions of the meetings only (i.e., the read digit strings referred to earlier were excluded). Also, for this study we ran recognition only on those segments during which a transcription was produced for the particular speaker. This overestimates the accuracy of word recognition, since any speech recognized in the "empty" segments would constitute an error not counted here. However, adding the empty regions would increase data load by a factor of about ten—which was impractical for us at this stage. Note that the current NIST Hub-5 (Switchboard) task is similar in this respect: data are recorded on separated channels and only the speech regions of a speaker are run, not the regions in which they are essentially silent. We plan to run all speech (including these "empty" segments) in future experiments, to better assess actual performance in a real meeting task.

4.3 Recognition results and discussion

Overall error rates. Table 1 lists word error rates for the six meetings, by speaker. The data are organized into two groups: native speakers and nonnative speakers. Since our recognition system is not trained on nonnative speakers, we provide results only for the native speakers; however the word counts are listed for all partici-

Table 1: Recognition performance by speaker and meeting (MRM = "Meeting Recorder meeting"; ROB = "Robustness meeting"). Speaker gender is indicated by "M" or "F" in the speaker labels. "* . . . *" marks speakers using a lapel microphone; all other cases used close-talking head-mounted microphones. "—" indicates speakers with severely degraded or missing signals due to incorrect microphone usage. Word error rates are in boldface, total number of words in Roman, and out-of-vocabulary (OOV) rates in *italics*. **OOV rate is by token, relative to a Hub-5 language model. WER is for conversational speech sections of meetings only, and are not reported for nonnative speakers.**

Meeting	MRM002	MRM003	MRM004	MRM005	ROB005	ROB004
Duration (minutes)	45	78	60	68	81	70
Native speakers						
M_004		**42.4** 4550 *2.07*	**48.1** 3087 *2.75*	**44.3** 3432 *1.60*	**48.4** 4912 *2.12*	**45.1** 5512 *1.61*
M_001	**42.4** 2311 *1.82*	**50.6** 2488 *2.09*	**37.6** 1904 *2.78*	**38.6** 3400 *1.56*		
F_001	**45.2** 3008 *2.59*	**43.2** 3360 *3.18*	**42.9** 2714 *4.05*	**41.9** 2705 *2.14*		
M_009		***100.1*** 1122 *1.59*	***115.8*** 367 *2.45*	**38.2** 1066 *1.88*		***68.7*** 696 *2.01*
F_002		**45.2** 1549 *2.26*	**43.7** 1481 *2.64*	***46.0*** 2480 *1.63*		
M_002	***55.6*** 990 *2.12*					
Speakers with low word counts						
M_007					**55.6** 198 *2.97*	— 69 *2.90*
M_008					**72.7** 55 *5.45*	**59.5** 121 *5.79*
M_015						— 59 *6.56*
Non-native speakers (total words only)						
M_003 (British)	2189					
M_011 (Spanish)		2653	1239	663		
F_003 (Spanish)					620	220
M_010 (German)			28			
M_012 (German)			639			
M_006 (French)					3524	2648

pants for completeness.[1]

The main result to note from Table 1 is that overall word error rates are not dramatically worse than for Switchboard-style data. This is particularly impressive since, as described earlier, no meeting data were used in training, and no modifications of the acoustic or language models were made. The overall WER for native speakers was 46.5%, or only about a 7% relative increase over a comparable recognition system on Hub-5 telephone conversations. This suggests that from the point of view of pronunciation and language (as opposed to acoustic robustness, e.g., for distant microphones), Switchboard may also be "ASR-complete". That is, talkers may not really speak in a more "sloppy" manner in meetings than they do in casual phone conversation. We further investigate this claim in the next section, by breaking down results by overlap versus nonoverlap regions, by microphone type and by speaker.

Note that in some cases there were very few contributions from a speaker (e.g., speakers M_007, M_008, and M_015), and such speakers also tended to have higher word error rates. We initially suspected the problem was a lack of sufficient data for speaker adaptation; indeed the improvement from adaptation was less than for other speakers. Thus for such speakers it would make sense to pool data across meetings for repeat participants. However, in looking at their word transcripts we noted that their utterances, while few, tended to be dense with information content. That is, these were not the speakers uttering "uhhuh" or short common phrases (which are generally well modeled in the Switchboard recognizer) but rather high-perplexity utterances that are generally harder to recognize. Such speakers also tend to have a generally higher overall OOV rate than other speakers.

Error rates in overlap versus nonoverlap regions. As noted in the previous section, the overall word error rate in our sample meetings was slightly higher than in Switchboard. An obvious question to ask here is: what is the effect on recognition of overlapping speech? To address this question, we defined a crude measure of overlap. Since segments were channel-synchronous in these meetings, a segment was either non-overlapping (only one speaker was talking during that time segment), or overlapping (two or more speakers were talking during the segment). Note that this does not measure amount of overlap or number of overlapping speakers; more sophisticated measures based on the phone backtrace from forced alignment would provide a better measure for more detailed analyses. Nevertheless, the crude measure provides a clear first answer to our question. Since we were also interested in the interaction if any between overlap and microphone type, we computed results separately for the head-mounted and lapel microphones. Results were also computed by speaker, since as shown earlier in Table 1, speakers varied in word error rates, total words, and words by microphone type. Note that speakers M_009 and F_002 have data from both conditions.

As shown, our measure of overlap (albeit crude), clearly shows that overlapping speech is a major problem for the recognition of speech from meetings. If overlap regions are removed, the recognition accuracy overall is actually better than that for Switchboard. It is premature to make absolute comparisons here, but the fact that the same pattern is observed for all speakers and across microphone conditions suggests that it is not the inherent speech properties of participants that makes meetings difficult to recognize, but rather the presence of overlapping speech.

Table 2: Word error rates broken down by whether or not segment is in a region of overlapping speech.

Speaker	No overlap		With overlap	
	Headset	Lapel	Headset	Lapel
M_004	41.0	-	50.3	-
M_001	34.2	-	47.6	-
F_001	40.5	-	45.8	-
M_009	30.7	41.0	40.7	117.8
F_002	37.7	29.8	50.5	56.3
M_002	-	48.6	-	71.3
M_007	52.2	-	81.3	-
M_008	50.9	-	69.9	
Overall	39.9	38.5	48.7	85.2

Furthermore, one can note from Table 2 that there is a large interaction between microphone type and the effect of overlap. Overlap is certainly a problem even for the close-talking head-mounted microphones. However, the degradation due to overlap is far greater for the lapel microphone, which picks up a greater degree of background speech. As demonstrated by speaker F_002, it is possible to have a comparatively good word error rate (29.8%) on the lapel microphone in regions of no overlap (in this case 964/2480 words were in nonoverlapping segments). Nevertheless, since the rate of overlaps is so high in the data overall, we are avoiding the use of the lapel microphone where possible in the future, preferring head-mounted microphones for obtaining ground truth for research purposes. We further note that for tests of acoustic robustness for distant microphones, we tend to prefer microphones mounted on the meeting table (or on a mock PDA frame), since they provide a more realistic representation of the ultimate target application that is a central interest to us - recognition via portable devices. In other words, we are finding lapel mics to be too "bad" for near-field microphone tests, and too "good" for far-field tests.

Error rates by error type. The effect of overlapping speech on error rates is due almost entirely to insertion errors, as shown in Figure 1. Rates of other error types are nearly identical to those observed for Switchboard (modulo a a slight increase in substitutions associated with the lapel condition). This result is not surprising, since background speech obviously adds false words in the hypothesis. However, it is interesting that there is little increase in the other error types, suggesting that a closer segmentation based on individual channel data (as noted earlier) could greatly improve recognition accuracy (by removing the surrounding background speech).

Error rates by meeting type. Different types of meetings should give rise to differences in speaking style and social interaction, and we may be interested in whether such effects are realized as differences in word error rates. The best way to measure such effects is within speaker. The collection of regular, ongoing meetings at ICSI offers the possibility of such within-speaker comparisons, since multiple speakers participate in more than one type of regular meeting. Of the speakers shown in the data set used for this study, speaker M_004 is a good case in point, since he has data from three "Meeting Recorder" meetings and two "Robustness" meetings. These two meeting types differ in social interaction; in the first, there is a fairly open exchange between many of the partici-

[1] Given the limitations of these pilot experiments (e.g., no on-task training material and general pronunciation models), recognition on nonnative speakers is essentially not working at present. In the case of one nonnative speaker, we achieved a 200% word error rate, surpassing a previous ICSI record. Word error results presented here are based on meeting transcripts as of March 7, 2000, and are subject to small changes as a result of ongoing transcription error checking.

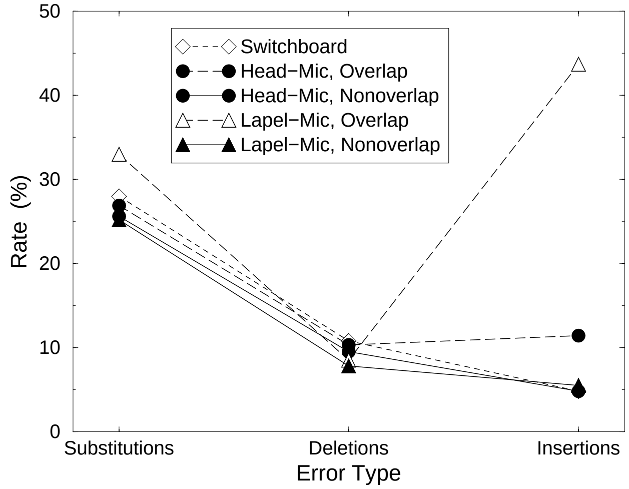

Figure 1: Word error rates by error type and microphone/overlap condition. Switchboard scores refer to an internal SRI development testset that is a representative subset of the development data for the 2001 hub-5 evals. It contains 41 speakers (5-minute conversation sides), from Switchboard-1, Switchboard-2 and Cellular Switchboard in roughly equal proportions, and is also balanced for gender and ASR difficulty. The other scores are evaluated for the data described in the text.

pants, while in the second, speaker M_004 directs the flow of the meeting. It can also be seen from the table that speaker M_004 contributes a much higher rate of words relative to overall words in the latter meeting type. Interestingly however, his recognition rate and OOV rates are quite similar across the meeting types. Study of additional speakers across meetings will allow us to further examine this issue.

5. FUTURE WORK

The areas mentioned in the earlier section on "Challenges" will require much more work in the future. We and our colleagues at collaborating institutions will be working in all of these. Here, we briefly mention some of the work in our current plans for the study of speech from meetings.

Far-field microphone ASR. Starting with the read digits and proceeding to spontaneous speech, we will have a major focus on improving recognition on the far-field channels. In earlier work we have had some success in recognizing artificially degraded speech [6][5], and will be adapting and more fully developing these approaches for the new data and task. Our current focus in these methods is on the designing of multiple acoustic representations and the combination of the resulting probability streams, but we will also compare these to methods that are more standard (but impractical for the general case) such as echo cancellation using both the close and distant microphones.

Overlap type modeling. One of the distinctive characteristics of naturalistic conversation (in contrast to monolog situations) is the presence of overlapping speech. Overlapping speech may be of several types, and affects the flow of discourse in various ways. An overlap may help to usurp the floor from another speaker (e.g., interruptions), or to encourage a speaker to continue (e.g., back channels). Also, some overlaps may be accidental, or a part of joint action (as when a group tries to help a speaker to recall a person's name when he is in mid-sentence). In addition, different speakers may differ in the amount and kinds of overlap in which they engage (speaker style). In future work we will explore types of overlaps and their physical parameters, including prosodic aspects.

Language modeling. Meetings are also especially challenging for the language model, since they tend to comprise a diverse range of topics and styles, and matched training data is hard to come by (at least in this initial phase of the project). Therefore, we expect meeting recognition to necessitate investigation into novel language model adaptation and robustness techniques.

Prosodic modeling. Finally, we plan to study the potential contribution of prosodic (temporal and intonational) features to automatic processing of meeting data. A project just underway is constructing a database of prosodic features for meeting data, extending earlier work [10, 9]. Goals include using prosody combined with language model information to help segment speech into coherent semantic units, to classify dialog acts [12], and to aid speaker segmentation.

6. ACKNOWLEDGMENTS

The current work has been funded under the DARPA Communicator project (in a subcontract from the University of Washington), supplemented by an award from IBM. In addition to the authors of this abstract, the project involves colleagues at a number of other institutions, most notably: Mari Ostendorf, Jeff Bilmes, and Katrin Kirchhoff from the University of Washington; and Hynek Hermansky from the Oregon Graduate Institute.

7. REFERENCES

[1] M. Beham and G. Ruske, Adaptiver stochastischer Sprache/Pause-Detektor. *Proc. DAGM Symposium Mustererkennung*, pp. 60–67, Bielefeld, May 1995, Springer.

[2] D. Biber, *Variation across speech and writing.* 1st pbk. ed. Cambridge [England]; New York: Cambridge University Press, 1991.

[3] W. Chafe, Cognitive constraints on information flow. In R. S. Tomlin (ed.) *Coherence and grounding in discourse.* Philadelphia: John Benjamins, pp. 21–51, 1987.

[4] J. Edwards, The transcription of Discourse. In D. Tannen, D. Schiffrin, and H. Hamilton (eds). *The Handbook of Discourse Analysis.* NY: Blackwell (in press).

[5] H. Hermansky, D. Ellis, and S. Sharma, Tandem connectionist feature stream extraction for conventional HMM systems, Proc. ICASSP, pp. III-1635–1638, Istanbul, 2000.

[6] H. Hermansky and N. Morgan, RASTA Processing of Speech, *IEEE Trans. Speech and Audio Processing* 2(4), 578–589, 1994.

[7] A. Janin and N. Morgan, SpeechCorder, the Portable Meeting Recorder, *Workshop on hands-free speech communication*, Kyoto, April 9-11, 2001.

[8] http://www.icsi.berkeley.edu/speech/mtgrcdr.html

[9] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3-4):439–487, 1998.

[10] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154, 2000.

[11] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng. The SRI March 2000 Hub-5 conversational speech transcription system. *Proc. NIST Speech Transcription Workshop*, College Park, MD, May 2000.

[12] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech, *Computational Linguistics* 26(3), 339–373, 2000.

[13] http://www.etca.fr/CTA/gip/Projets/Transcriber/

[14] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen, Meeting Browser: Tracking and Summarizing Meetings, *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, 1998.

[15] H. Yu, C. Clark, R. Malkin, and A. Waibel, Experiments in Automatic Meeting Transcription Using JRTK, *Proc. ICASSP*, pp. 921–924, Seattle, 1998.