THE ICSI MEETING PROJECT: RESOURCES AND RESEARCH

Adam Janin[1,2], *Jeremy Ang*[1], *Sonali Bhagat*[1], *Rajdip Dhillon*[1]
Jane Edwards[1,2], *Javier Macías-Guarasa*[1,4], *Nelson Morgan*[1,2], *Barbara Peskin*[1]
Elizabeth Shriberg[1,3], *Andreas Stolcke*[1,3], *Chuck Wooters*[1], *Britta Wrede*[1,5]

[1]International Computer Science Institute, Berkeley CA, USA
[2]University of California, Berkeley CA, USA
[3]SRI International, Menlo Park CA, USA
[4]Speech Technology Group, Dept. of Electronic Engineering, Univ. Politécnica de Madrid, Spain
[5] Applied Computer Science Group, Bielefeld University, Germany

ABSTRACT

This paper provides a progress report on ICSI's Meeting Project, including both the data collected and annotated as part of the project, as well as the research lines such materials support. We include a general description of the official "ICSI Meeting Corpus", as currently available through the Linguistic Data Consortium, discuss some of the existing and planned annotations which augment the basic transcripts provided there, and describe several research efforts that make use of these materials. The corpus supports wide-ranging efforts, from low-level processing of the audio signal (including automatic speech transcription, speaker tracking, and work on far-field acoustics) to higher-level analyses of meeting structure, content, and interactions (such as topic and sentence segmentation, and automatic detection of dialogue acts and meeting "hot spots").

1. INTRODUCTION

The speech recognition community has continually accepted new challenges in processing spoken language, progressing from early work in low-noise isolated word and digit recognition, to continuous read speech, to spontaneous speech, and to more difficult tasks such as the recognition of speech from natural conversations. Natural multi-party interaction has become a new frontier, with data collected from meetings being an ideal testbed. This domain presents significant challenges not just to speech recognition, but also to speaker technologies, to discourse modeling, to spoken language understanding, and to audio retrieval, to name just a few of the disciplines challenged by this material. There is now an international effort in Meetings recognition and understanding (primarily funded in Europe at the moment), which has already produced promising research results, and initiated new collaborations. See [1], [2], [3], and [4] for further information on key European projects in this area.

Four years ago, ICSI inaugurated its Meeting Recorder project to address this important new research area. Since that time, we have collected nearly 100 hours of meeting room data, have transcribed and publicly released a 75-meeting subset, and have continued to enrich the collection with additional levels of annotation. We have also launched a number of research projects exploring many of the challenges presented by this material. Efforts range from low-level processing of the speech signal, including core automatic speech recognition (ASR), work on far-field acoustics, and speaker segmentation, to higher-level analyses of meeting structure, content, and interactions. We have reported on the early stages of the project in [5] with a later research update in [6]. Our Meeting Recorder website [7] provides additional visibility into these efforts.

This paper provides a progress report on the Meeting Project at ICSI, including both the data collected and annotated as part of the project, as well as the research lines that such materials support. We first provide a review of the official "ICSI Meeting Corpus", as currently available through the Linguistic Data Consortium, and then discuss some of the existing and planned annotations which augment the basic corpus transcripts. We then describe several research efforts that make use of these resources, focusing primarily on higher-level analyses of meeting structure. Core ASR is covered lightly, as a detailed description of our meeting transcription system is provided in the companion paper [8] also found in these proceedings.

2. THE ICSI MEETING CORPUS

In this section, we present a brief summary of the contents of the corpus. For a more complete description, including many of the design decisions made along the way, see [9].

The ICSI Meeting Corpus consists of 75 meetings recorded in a conference room at the International Computer Science Institute in Berkeley between the years 2000–2002. The meetings were "natural", in the sense that they would have occurred regardless of the recording process. Most of the meetings were regularly scheduled weekly group meetings.

Each meeting participant wore a head-mounted microphone (a few early meetings contain an occasional lapel microphone instead of the head-worn one). Additionally, 6 tabletop microphones simultaneously recorded the audio. Meetings averaged slightly under an hour and involved 3-10 participants, with an average of 6. This resulted in 72 hours of meeting audio, approximately 85 hours of recorded speech[1], and about 900 channel-hours of total audio.

The waveform from each microphone was stored in a separate file. The data were downsampled on the fly from 48 kHz to

[1] There is more recorded speech than meeting audio because of speaker overlap.

16 kHz, encoded using 16-bit linear NIST SPHERE format, and compressed using a lossless algorithm [10].

For each meeting, we stored a small XML file describing some meeting-specific information, including the date and time of the meeting, the primary topic, a unique identifier, participant information (see below), microphone and channel types, and some free-form notes. The notes typically recorded technical or acoustic problems, information about late arrivals and early exits, and other meeting idiosyncrasies.

Each speaker was asked to fill out a speaker form prior to their first recorded session. Information requested included name, sex, education level, age, and language information. For language information, we asked if the participant is a native speaker of English[2], and what variety of English they speak (e.g. “American”, “British”, “Indian”). For non-native speakers of English, we asked for their native language and region, as well as the number of years spent in an English-speaking country.

It is important to note that information on the speaker form is self-reported. This is especially relevant to native language and dialect information, since people are often unable to identify the particular region of their dialect. In addition, most fields were optional, so that some information (e.g. age) may not always be available.

For each meeting, the corpus also contains an XML file with a complete word-level transcription. In addition to standard lexical entries, these include word fragments, filled pauses, and nonspeech vocalizations such as laughs, gasps, and lip smacks, as well as nonvocal acoustic events such as door slams, microphone clicks, etc. The transcripts are also heavily commented, with notes qualifying the speech (e.g. mangled pronunciations or “while laughing”) and providing contextual information (e.g. about who is being addressed or other activities such as writing on the whiteboard).

Overlap between participants’ speech is *extremely* common in our meetings [11]. In the transcript, we mark the speaker, the start time, and the end time of each of the utterances. Overlaps can therefore be determined by overlapping utterance times for speakers, but each speaker’s speech is individually transcribed.

Transcription was performed primarily with the close-talking channels. Occasionally, the far-field microphone channels were used to clarify events not well captured on the close-talking channels, such as nonspeech events and off-mic remarks. Use of the close-talking channel permitted careful transcription during overlapped speech, as well as the capture of soft-spoken back-channels and self-vocalizations.

In addition to the meetings themselves, we also asked participants to read digit strings similar to those found in TIDIGITS [12] at the beginning or end of each meeting. The transcripts of the digits task are included in the meeting transcripts. This simpler task provides an opportunity for research on far-field acoustics without the additional complexity of dealing with large-vocabulary spontaneous multi-party speech, but still involving the same speakers, microphones, and room acoustics as in the main meetings.

The XML formats of the transcripts and metadata were designed specifically for this collection. A complete DTD and description of the format are distributed with the corpus. We also provide software for translating from our format to other common formats.

The ICSI Meeting Corpus is now available from the Linguistic Data Consortium [13] as publications LDC2004S02 (Speech) and LDC2004T04 (Transcripts). Further information can be found in the extensive documentation available with the corpus.

3. ADDITIONAL ANNOTATIONS

Only word-level transcriptions have been released with the corpus, but we are currently engaged in a number of efforts to augment these transcripts with additional levels of hand annotation.

3.1. Dialog acts and adjacency pairs: the ICSI MRDA corpus

Understanding meetings requires more than just the words. An obvious level up from words is the annotation of larger units (on the order of a sentence) according to their function in the conversation. We have annotated the full 75-meeting corpus for dialog acts (DAs), such as whether the utterance is a statement, question, backchannel, and so on. This auxiliary corpus, called the ICSI Meeting Recorder Dialog Act (MRDA) Corpus, is described in [14] and is now freely available to the community for research purposes [15]. We plan to make the corpus available in the future through the Linguistic Data Consortium.

The MRDA corpus consists of over 180,000 hand-annotated dialog act tags, using an annotation system adapted from SWBD-DAMSL [16]. SWBD-DAMSL provided a good initial match to the types of phenomena in our meetings, but it was nevertheless necessary to modify and adapt the system in a number of ways to fit the meeting data. The annotation system and numerous real examples from our data are provided in a detailed manual [17].

The system marks each DA with one of 11 “general” tags (for example, “statement”) and a variable number of 39 possible further descriptive tags (for example, “suggestion”, “disagreement”, “joke”). Interlabeler agreement on a random subset of the data showed excellent agreement for basic class groupings of tags as measured by κ, which adjusts for chance agreement (hence values for absolute agreement are always much higher). Our κ value was 0.80 for a six-way classification; further agreement statistics are provided in [14].

The annotation involved not only labeling DAs, but also segmenting the annotations into DA units (which differ from sentence and segment units in the original corpus; alignment information is provided). In addition, the annotations include marking of adjacency pairs, or utterances that refer to content of other utterances (e.g., an answer is coindexed with the question it applies to). Such information is quite complex in meetings, since the multiple speakers and overlaps often mean that adjacency pairs are not consecutive, and the same DA can also lead to responses from multiple participants.

3.2. Involvement and “Hot Spots”

Another level of annotation that we predict will be useful for summarizing and browsing meetings is to mark “hot spots”, or locations of high participant involvement. Although raised involvement at times does occur for only one participant, more typically it is a feature of the *interaction* among two or more participants. For example, speakers may become involved in a heated disagreement, or they may strongly agree on a particular proposed solution. Note that it is not the type of situation that determines whether a re-

[2] In retrospect, we probably should have distinguished between native and non-native speakers of *American* English.

gion of speech or exchange is "hot" but rather the level of affective involvement on the part of the participants.

Labeling of involvement for isolated utterances.

While the idea of annotating involvement level may sound dangerously subjective, we found that agreement on such annotation by human judges on isolated utterances is significantly above chance — in fact better than we expected.

We removed the utterances from their surrounding context in order to determine how much information was in the utterance itself. This is useful for modeling purposes, and allowed more experimenter design control over the listening task. Human listeners who were native speakers of English and knew the speakers, but had not been in any of the recorded meetings, were asked to simply mark involvement level after hearing each utterance. They were not given information on how to make their judgments, and often reported that they felt they were assigning random labels. Yet they agreed with each other with $\kappa = .63$ [18].

Labeling of the "hot spot anatomy".

In current work, we are extending our annotation to label "hot spot anatomy" or common features we have found to be identifiable in marking a stretch of utterances that we perceive at a hot spot, as judged by 3 human labelers. Note that unlike the work in isolated utterances, this annotation involves listening to the interactions among participants and marking off regions in time that span multiple participant turns. Each hot spot begins with an utterance identified as a "trigger" (which is often not "hot" itself), and ends with a "closure" that is based on the semantic and pragmatic resolution of the hot spot. Inside this region, utterances that are said with high involvement are marked as "peaks". The "hotness" of the overall hot spot is given a rating for level. In addition, we mark the hot spot "type". For example, we distinguish between disagreements and amusement; we also distinguish hot spots due only to intense interaction from those due to interaction plus semantic/pragmatic content. We have recently iterated on interlabeler agreement and have begun labeling of the 75 released meetings by two annotators. We will continue to monitor interlabeler reliability using randomly selected meeting excerpts. Labeling appears to proceed at approximately four times real time, but can vary considerably by meeting.

3.3. Other annotations

Now that the ICSI Meeting Corpus is becoming a shared resource across numerous research sites, other labs are also beginning to add their own annotations.

For example, the Speech, Signal and Language Interpretation (SSLI) Lab at the University of Washington is in the process of annotating the corpus with anaphoric (pronouns and pro-verbs) and deictic (e.g. "these", "this", "that") words, marking these words and their referents in the textual transcriptions of the meetings. Ultimately, the goal is to build automatic anaphora resolution algorithms that feed into topic detection and annotation tools. Annotations have been performed on an initial set of meeting transcripts[3], and about 30 meetings have been annotated at this point.

In addition, Columbia University annotated 25 of the meetings for topic segmentation [19], resulting in an average of 7.5 segments per meeting. At least 3 annotators labeled locations of topic shift in each meeting, and a substantial level of agreement was obtained amongst team members. They have also explored other annotation to assist with meeting summarization work, in particular the utterance-level scoring of significance and salience, but the interlabeler agreement was much weaker. They are currently extending this work by augmenting the corpus with hand-written summaries (that is, "minutes" of meetings) to evaluate and possibly train a meeting summarization system. The annotated data are available from their website [20].

Further annotations are planned by our collaborators as part of the EU Integrated Project, AMI. We hope more people working with the corpus will wish to add new levels of annotation to this collection, and we would welcome information about such efforts. We also hope that such resources, like those described above, can be made widely available to the Speech and Language community.

4. A RESEARCH SAMPLING

The resources described above support a wide range of research activities, from low-level processing of the speech signal to higher-level analyses of meeting structure and content. In this section, we provide a sampling of several different research efforts to illustrate the richness of the meetings domain.

We begin with a very brief summary of work in far-field acoustics, speaker segmentation, and core automatic speech recognition — all key areas of research in processing meeting data. However, since this work is largely reported elsewhere, including in our companion paper in these proceedings [8], we here focus on other efforts, including work on recognition of accented speech, but concentrating primarily on work addressing higher-level analysis of meeting structure.

4.1. Far-field acoustics, speaker segmentation, and core ASR

The digit sequences recorded as part of the ICSI Meeting Corpus provide a good starting point for exploring the acoustics of the meeting recordings. Early work included a study on noise reduction and deconvolution processing for single-microphone far-field speech recognition using the recording from a single high-quality tabletop microphone [21]. We also investigated the simultaneous use of the two inexpensive electret microphones that were part of the personal digital assistant mockup placed on the meeting room table [22]. Michael Seltzer, then at Carnegie Mellon University, used the recordings from all four high-quality tabletop microphones in his work on microphone array speech recognition [23].

A key challenge in moving from digit sequences to conversational interactions is segmenting the speech into speaker turns. While this is clearly an issue for the far-field microphones, where all speakers co-exist on the same channel, it remains a surprisingly significant problem for the close-talking channels as well, especially because of cross-talk from neighboring speakers. In [24], we presented one approach to this problem. We continue to explore new techniques for handling speaker segmentation.

Our basic Meetings ASR system has improved dramatically over the course of the Meetings project, starting from a simplified form of SRI's recognizer designed for conversational telephone speech. Improvements are due both to evolution in the basic SRI system as well as to improved modeling to specifically address the meeting domain. Word error rates for native English speakers on close-talking channels are now comparable to those we see for conversational telephone tasks such as Switchboard. Our early

[3] Unfortunately, this was an early release of the data, differing slightly from the transcripts released to the LDC.

work in this area was reported in [5] and [6]. Details of the current ASR system can be found in the companion paper [8].

4.2. Speech recognition for non-native speakers

Given the international nature of ICSI and the multi-dialectal characteristic of the corpus, we have begun work on the problem of recognizing non-native speech in the corpus. This was done despite the fact that the amount of recorded speech and the number of different speakers is insufficient for adequately modeling and testing most accent groups. The database contains 15 variants of accented English, although only three of them (American, German and Spanish) have more than two different speakers (23 American, 12 German and 5 Spanish). We therefore concentrated attention on only the German and Spanish speakers.

The recognition experiments were carried out on a carefully selected database partition to ensure, whenever possible, an adequate balance between training and testing material, including gender, number of speakers and amount of recorded speech.

The baseline experiments used a slightly simplified version of the SRI recognizer trained for recognition of conversational telephone speech [25], but with dictionary and language model adapted to the meetings domain. Initial results with this baseline system are shown in the first row of Table 1. The error rates for non-native speech are markedly higher than for American speakers, and higher than other results reported in the literature [26]. This is probably caused by the lower English proficiency of some of the speakers in the ICSI corpus. For native speech, the results are very close to those obtained in the standard conversational telephone speech tasks, indicating that meeting data recorded with close-talking microphones is an accessible task using current technology.

	American	German	Spanish
Baseline SI	34.1%	52.3%	104.2%
TaskMAP	32.5%	46.4%	95.2%
AccMAP	32.5%	46.0%	96.8%
TaskAccMAP	32.5%	45.8%	95.0%
+ phoneloop	30.4%	42.3%	93.2%
Rel. Δ Error	[-10.9%]	[-19.1%]	[-10.6%]

Table 1. Word error rates for accent-dependent speech

In these initial experiments, only acoustic adaptation was performed, using maximum a posteriori (MAP) techniques[4]. Three strategies were studied:

- Task adaptation (TaskMAP row in Table 1), in which all the speech in the training subset was used for adaptation.
- Accent adaptation (AccMap row in Table 1), in which accent dependent models were adapted using the corresponding subset of the adaptation data.
- Task adaptation followed by Accent adaptation (TaskAccMAP row in Table 1).

The results in Table 1 demonstrate that the combined strategy performed the best. Applying a final phone-loop adaptation stage at test-time, the improvements for non-native speech are on the average higher than the ones obtained for native speakers, with the German speakers showing substantial improvement although the more limited Spanish speaker pool still lags behind.

Clearly, there is still considerable room for improvement. Future work will incorporate additional methods for adaptation to non-native speech, such as modifications to pronunciation modeling [27], as it is likely that the Spanish speakers' performance is limited by dictionary pronunciations inadequate for their speech.

Some preliminary work is also being carried out in accent identification, applying classification algorithms traditionally used in language identification and combining their results using neural networks. Such "accent ID" would be an important pre-processing step in order to incorporate accent-sensitive models during speech transcription.

4.3. Discourse markers' role in inferring topic structure and social structure

A main goal of the our meeting research is to develop automatic methods for identifying and representing the content structure and social structure of meetings. An important part of this involves detecting important utterances within topics and linkages between topics. One promising route for accomplishing this is via analysis of discourse markers (DMs), such as "now," "well," and "so." These items serve to bracket a unit of talk and explicitly mark the relationship between that unit and what precedes it [28].

Of this class of items, "so" seems uniquely well suited for meeting structuring for several reasons. First, several of its uses involve topic management in some form, e.g., introducing a new topic, resuming a previous topic, expressing a conclusion. Second, it also has some social uses, such as when it is used turn-finally to signal a speaker's willingness to relinquish the floor.

Previous work on some DMs ("now" and "oh") has found different prosodic characteristics in their different uses (e.g., [29], [30]). Such work has not yet been done for "so."

The work here represents a feasibility study extending that approach to "so" and using the information to infer meeting structure. In addition, it seeks to determine whether the uses of "so" correlate with speaker role or speaker style; for example, whether highly dominant speakers use topic-initiating "so" more often than less dominant speakers, and whether a particular speaker uses them more often when leading a discussion than otherwise.

Six meetings were examined, representing two each of three different meeting types. The types varied in their number of participants, the degree to which they were agenda-driven, and the degree to which power was centralized or shared equally. In addition, the leader in one meeting type was a non-leader participant in another.

Analyses have only started, and so far only two of the meetings have been carefully studied, but the results are in the expected direction.

Regarding content structure: it was found that the meeting leaders produced more topic-relevant, turn-initiating uses of "so" than the other participants. In addition, the person who switched status from leading in one meeting to not leading in another produced more "topic-relevant" uses in the former than the latter.

Regarding social structure: two findings warrant discussion.

The utterance-final "so" was described by Schiffrin [28] as a method speakers use to signal willingness to relinquish the turn. Consistent with her findings, utterance-final "so" in one of the examined meetings correlated with turn shift 90% of the time.

[4] Preliminary experiments showed that MAP was better than MLLR given the amount of adaptation data available.

Utterance-initial “so” was found to be used quite often as a floor grabber in this meeting. This use was not found in Schiffrin’s data, perhaps because her data were obtained via interviews rather than meetings. Turn-initial “so” may be an effective way to gain the floor for both phonetic and semantic reasons. The leading /s/ is easily heard above overlapping speech and the semantics of “so” give the expectation that the speaker’s utterance will be collaborative (in contrast to an adversarial “but”), and important or definitive in some way (in contrast to “and”). It’s possible that it may be more frequent in more competitive meetings, or that it may be more characteristic of some speakers than others. This will be examined in future analyses.

Future results will include analysis of the other meetings and will also incorporate prosodic profiles. Prosodic characterizations for automatic recognition of the different uses of “so” will be based on the parameters such as those developed for [31]. These include duration of the word, position in the utterance, fundamental frequency, and other intonational characteristics.

4.4. Automatic detection of dialog acts and hot spots

Given the hand annotation of dialog acts and hot spots reported above, it is now possible to explore whether such labeling can be performed automatically and to examine the interrelationships among the various types. Here we describe some initial research along these lines.

Classification of four frequent DAs.

In initial research on dialog acts, we looked at automatic classification of DAs in a focused study of four types of single-word and frequently-occurring DAs: *Acceptance/Agreement*, *Acknowledgment*, *Backchannel*, and *Floor-grabber*. Focusing on these 4 frequent DAs made it feasible to look at automatic prediction using limited annotated data (only 20 meetings had been annotated for DAs at the start of our work). The interesting thing about these DAs, aside from their frequency of occurrence, is that in English they share lexical forms. For example, each of the words “yeah”, “okay”, “uhhuh”, “right” can function as any of the four.

We were interested in whether automatically extracted prosodic features could help to distinguish these different types of DAs. As described in [32], we extracted and automatically normalized a variety of prosodic features based on the output of forced alignment recognition. The feature types included duration features, pause features (both within and across speakers, thus capturing turn pauses in the latter), pitch features, and energy features. We found that decision tree classifiers were able to distinguish among the DAs using only prosodic features at rates significantly above chance. What was interesting was the specific patterning of features associated with the different distinctions. As one case in point, we found that while floor-grabbers are dramatically higher in energy and pitch than are backchannels, the two forms differ little in duration. This goes against the typical finding that duration, pitch and energy pattern together in conventional prosody. Our interpretation is that speakers in grabbing the floor want to make sure they are heard, but are “feigning” politeness by jumping in and out quickly while another speaker is still talking. Further interpretations of different classifier results are described in [32].

Classification of all DAs.

Upon our recent completion of DA annotations for all 75 meetings, we have begun to construct a database of both language model scores and prosodic features for all dialog acts, using truth transcripts. Our plans are to first look at automatic DA classification performance using both DA-level language and DA-level prosody. A number of different experiments are underway in this area, examining different ways of grouping DAs as well as different modeling techniques. Our future plans include constructing a database based on automatically recognized words, adding a DA-sequence grammar modeling component, and modeling adjacency pair information.

Human and machine labeling of involvement.

In research on our hot spot labels, we asked whether human judgments of involvement on isolated utterances could be predicted by automatically extracted prosodic cues. Acoustic analyses of these utterances showed that involvement is correlated with higher mean and maximum F0 values (after normalizing by speaker pitch range), whereas energy values are only moderately affected [18]. This suggests that these prosodic cues are reliable features to automatically detect involvement in meeting data. It also agrees with general findings in the linguistics and prosody literatures — that although the perception of raised voice is often attributed to increased *loudness* (energy), it is often actually *pitch range* that is more significantly modified.

Analysis of the relationship between dialog acts and involvement.

In an effort to better understand the relationship between discourse phenomena and affective phenomena in our meetings, we also examined the relationship between the human dialog act annotations and the human hot spot annotations on common meetings [33]. Importantly, these two different types of annotations were performed completely independently, and by different teams of annotators, making it a fair study. Results showed that involved utterances contain significantly more evaluative and subjective statements such as jokes, suggestions, and extremely positive or negative answers. Noninvolved utterances are characterized by more backchannels and floor-holders. Quite interestingly, involved utterances do not appear to contain more information, as measured by statistical perplexity. (The one exception here is jokes, which often contain out-of-context information or even out of vocabulary words). The overall similarity in perplexity of involved and noninvolved suggests that involvement reflects not the content itself, but a speaker’s *attitude toward that content*. This makes it all the more important, we think, to include affective information in automatic summarization, since words alone (as used in classical summarization approaches) may fail to capture these highly charged regions.

4.5. Other research

The above projects represent just a few of the many lines of research supported by the meeting data. Additional efforts on automatically detecting meeting structures, which make use of the resources reported here but are more fully documented elsewhere, include:

Automatic detection of punctuation and disfluencies.

In early work on pre-released meeting data [31], we examined cues for automatic labeling of “hidden events” such as sentence boundaries and disfluencies. Such labeling is important for downstream natural language processing from ASR output, since the processing techniques typically assume fluent, punctuated text. We found that combining lexical cues with automatically extracted prosodic information produced better performance than either knowledge source alone. Prosodic information was more robust than lexical information if using ASR output. In addition, prosody degraded much less than did language information for the task of on-line (no lookahead) prediction of sentence boundaries.

Unsupervised learning for detecting agreement and disagreement.

In joint work with the University of Washington [34], we investigated whether unsupervised training techniques could help limit the need for hand-labeled data on a dialog-related task. We looked at automatic classification of whether utterances corresponded to an agreement, disagreement, or other type of utterance. Results showed that while hand-labeled data is best, adding more data via automatic labeling significantly improved both lexical and prosodic modeling for the recovery of agreements and disagreements.

Visualization of topic and speaker structures.

In [35], Renals and Ellis report on a number of preliminary investigations, exploring the ICSI Meeting data and presenting novel ways of analyzing and accessing meeting structure. Their studies include spoken document retrieval for meetings and visualizing topic structure, automatic structuring of meetings based on self-similarity matrices of speaker turn patterns, and a simple model of speaker activity in terms of "talkativity".

5. FUTURE DIRECTIONS

The efforts reported above only begin to suggest the richness of the meetings domain and of the resources we are compiling to explore it. We and our collaborators continue to expand the resources available to the research community, as well as continuing our efforts to understand and automatically mark meeting structures.

6. ACKNOWLEDGMENTS

Many, many people were involved in the collection, annotation, and exploration of the ICSI Meeting Corpus. We would like to thank the participants who donated their voices to science; the ICSI transcription team; ICSI students, postdocs, and staff, including Don Baron (meeting processing, prosodic database scripts), Jim Beck (hardware assistance), Hannah Carvey (DA annotation), Ashley Krupski (DA annotation system development), and Chris Oei (annotation interfaces). Finally, we would like to thank our collaborators at other sites, most notably: Dan Ellis, Michel Galley, and Kathy McKeown from Columbia University; Steve Renals from the University of Edinburgh; and Mari Ostendorf, Jeff Bilmes, and Katrin Kirchhoff from University of Washington.

The corpus collection and preparation was funded in part by DARPA (through the Communicator project and a ROAR seedling), by the Swiss National Science Foundation (through NCCR's IM2 project), and by a supplemental award from IBM. ICSI's ongoing work on Meetings annotation and research is supported in part by the Swiss Research Network IM2, by an NSF-ITR on "Mapping Meetings", by the European Union 5th FWP M4, and by the European Union 6th FWP IST Integrated Project, AMI (Augmented Multi-party Interaction, FP6-506811, publication AMI-6).

In addition, Britta Wrede's work on hot spots was supported by the German Research Foundation (DFG), and Javier Macías's work on non-native speakers was supported by the Spanish Ministry of Science and Technology (MCYT).

7. REFERENCES

[1] Augmented Multi-party Interaction (AMI), "Web page," http://www.amiproject.org.

[2] Computers In the Human Interaction Loop (CHIL), "Web page," http://chil.server.de.

[3] Multi-Modal Meeting Manager (M4), "Web page," http://www.dcs.shef.ac.uk/spandh/projects/m4.

[4] Interactive Multimodel Information Management (IM2), "Web page," http://www.im2.ch.

[5] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Human Language Technologies Conference*, San Diego, March 2001.

[6] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "Meetings about meetings: research at ICSI on speech in multiparty conversations," in *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-2003)*, Hong Kong, April 2003.

[7] International Computer Science Institute, "ICSI meeting corpus web page," http://www.icsi.berkeley.edu/speech/mr.

[8] Andreas Stolcke et al., "The ICSI-SRI-UW 2004 meeting recognition system," in *NIST Meeting Recognition Workshop*, Montreal, May 2004.

[9] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-2003)*, Hong Kong, April 2003.

[10] Tony Robinson, "Shorten: Simple lossless and near-lossless waveform compression," Tech. Rep., Cambridge University Engineering Department, 1994, CUED/F-INFENG/TR.156.

[11] Elizabeth Shriberg, Andreas Stolcke, and Don Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *8th European Conference on Speech Communication and Technology (Eurospeech-2001)*, Aalborg, September 2001.

[12] R.G. Leonard, "A database for speaker independent digit recognition," in *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-84)*, San Diego, CA, 1984.

[13] The Linguistic Data Consortium (LDC), "Web page," http://www.ldc.upenn.edu/.

[14] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Special Interest Group on Discourse and Dialogue (SIGdial)*, Boston, April–May 2004.

[15] International Computer Science Institute, "ICSI MRDA corpus web site," http://www.icsi.berkeley.edu/~ees/dadb.

[16] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL shallow discourse function annotation coders manual, draft 13," Tech. Rep. 97-02, Institute of Cognitive Science Technical, University of Colorado, Boulder, 1997.

[17] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg, "Meeting recorder project: Dialog act labeling guide," Tech. Rep. TR-04-002, ICSI, 2004.

[18] B. Wrede and E. Shriberg, "Spotting "hotspots" in meetings: Human judgments and prosodic cues," in *European Conference on Speech Communication and Technology (Eurospeech-2003)*, Geneva, September 2003.

[19] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, Sapporo, Japan, 2003.

[20] M. Galley, "Columbia University tools and corpus web page," http://www1.cs.columbia.edu/~galley/research.html.

[21] D. Gelbart, "Mean subtraction for automatic speech recognition in reverberation," M.S. thesis, Univ. of California Berkeley, 2004.

[22] L. Docio-Fernandez, D. Gelbart, and N. Morgan, "Far-field ASR on inexpensive microphones," in *European Conference on Speech Communication and Technology (Eurospeech-2003)*, Geneva, September 2003.

[23] M. L. Seltzer, *Microphone Array Processing for Robust Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, 2003.

[24] T. Pfau, D.P.W. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI meeting recorder," in *2001 IEEE Workshop on Automatic Speech Recognition and Understanding*, Madonna di Campiglio, Italy, December 2001.

[25] A. Stolcke, H. Franco, R. Gadde, M. Graciarena, K. Precoda, A. Venkataraman, D. Vergyri, W. Wang, and J. Zheng, "Speech-to-text research at SRI-ICSI-UW," Presentation at RT-03S NIST Workshop, http://nist.gov/speech/tests/rt/rt2003/spring/presentations /sri+-rt03-stt.pdf.

[26] Laura Mayfield Tomokiyo, "Handling non-native speech in lvcsr: A preliminary study," in *Proceedings of the EUROCALL/CALICO/ISCA workshop on Integrating Speech Technology in (Language) Learning (InSTIL)*, August 2000.

[27] Silke Goronzy, Stefan Rapp, and Ralf Kompe, "Generating non-native pronunciation variants for lexicon adaptation," *Speech Communication*, vol. 42, no. 1, pp. 109–123, January 2004.

[28] D. Schiffrin, *Discourse markers*, Studies in interactional sociolinguistics. Cambridge University Press, 1987.

[29] J. Hirschberg and D. Litman, "Empirical studies on the disambiguation of cue phrases," in *Computational Linguistics*, 1992, vol. 19, pp. 501–530.

[30] J. Local, "Conversational phonetics: Some aspects of news receipts in everyday talk," *York Papers in Linguistics*, vol. 15, pp. 37–80, 1992.

[31] D. Baron, E. Shriberg, and A. Stolcke, "Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues," in *Proceedings International Conference on Spoken Language Processing*, Denver, September 2002, pp. 949–952.

[32] S. Bhagat, H. Carvey, and E. Shriberg, "Automatically generated prosodic cues to lexically ambiguous dialog acts in multiparty meetings," in *Proceedings International Congress of Phonetic Sciences*, Barcelona, August 2003.

[33] B. Wrede and E. Shriberg, "The relationship between dialogue acts and hot spots in meetings," in *Proceedings IEEE Speech Recognition and Understanding Workshop*, St. Thomas, U.S. Virgin Islands, 2003.

[34] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proc. HLT-NAACL Conference*, Edmonton, Canada, May 2003.

[35] S. Renals and D. Ellis, "Audio information access from meeting rooms," in *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-2003)*, Hong Kong, April 2003.