THE ICSI MEETING CORPUS

Adam Janin[1,4], *Don Baron*[1,4], *Jane Edwards*[1,4], *Dan Ellis*[1,2], *David Gelbart*[1,4],*Nelson Morgan*[1,4]
Barbara Peskin[1], *Thilo Pfau*[1], *Elizabeth Shriberg*[1,3], *Andreas Stolcke*[1,3], *Chuck Wooters*[1]

[1]International Computer Science Institute, Berkeley, CA
[2]Columbia University, New York, NY
[3]SRI International, Menlo Park, CA
[4]University of California, Berkeley, CA

ABSTRACT

We have collected a corpus of data from natural meetings that occurred at the International Computer Science Institute (ICSI) in Berkeley, California over the last three years. The corpus contains audio recorded simultaneously from head-worn and table-top microphones, word-level transcripts of meetings, and various metadata on participants, meetings, and hardware. Such a corpus supports work in automatic speech recognition, noise robustness, dialog modeling, prosody, rich transcription, information retrieval, and more. In this paper, we present details on the contents of the corpus, as well as rationales for the decisions that led to its configuration. The corpus will be delivered to the Linguistic Data Consortium (LDC) [1] by December, 2002, and we expect it to be available through the LDC by the summer of 2003.

1. INTRODUCTION

The process of setting up and collecting the ICSI Meeting Corpus involved many, many decisions. The purpose of this paper is to describe not only the data actually in the corpus, but also the reasons we chose a particular approach.

In this section, we will discuss some of the high-level decisions regarding meeting types, participants, equipment, etc. In the following sections, we will describe the audio (section 2), meeting-specific information (section 3), speaker information (section 4), and the transcripts (section 5). We finish up with conclusions (section 6), related work (section 7), and acknowledgments (section 8).

One of the decisions we made early in the process was to record only "natural" meetings, meaning meetings that would have taken place anyway. Another option would have been scenario-based meetings, in which people are asked to discuss a particular topic, solve a problem, play a game, etc. As a result of this decision, it became clear that the bulk of the meetings would be with people at the International Computer Science Institute (ICSI), since we had set up the recording room there. Although it is possible to convince people to move a regular meeting to a new location, it is far easier to simply record meetings that would have occurred in that room (or nearby). Most of the recordings are of regular group meetings of research groups at ICSI. Consequently, some of the speakers appear in many meetings. Also, there are many examples of a given topic. See sections 3 and 4 for more information on the meeting types and speaker demographics.

Another key decision was to simultaneously record head-worn microphones and several desktop microphones. Although this introduces an unnatural component to the meeting (wearing a microphone), we felt it was crucial. First, the near-field signals allow us to separate far-field acoustic effects from language and dialog effects. Second, it provides a high-quality baseline for human transcription. Speech activity detection is also much easier, since the signal from the mic on the participant's head will likely be quite strong compared to signals from neighboring mics. Third, it provides a baseline for analyzing techniques designed to compensate for far-field effects (noise, reverberation, crosstalk, etc). Finally, it allows non-acoustic research (e.g. dialog act analysis) to proceed without the penalty of a poor acoustic signal.

At the beginning or end of each meeting, we also asked participants to read digit strings similar to those found in TIDIGITS [2]. The full task of performing automatic speech recognition from the far-field signal on the unconstrained meeting task is quite daunting, and current accuracy is poor. Providing the digits data allows research into far-field acoustic issues without the additional complexity of large vocabulary, spontaneity, and human-to-human interactions.

2. AUDIO

The meetings all took place in a conference room at ICSI. The room seats approximately 12 people along a long and thin table. A projection screen is located at the end of the room. Although the projector was seldom used, its fan was active during the recordings.

For each meeting, we simultaneously recorded up to 10 close-talking head-worn microphones, 4 desktop omni-directional PZM microphones, and a "dummy" PDA containing two inexpensive microphones. The PZM microphones were arranged in a staggered line along the center of the conference table, and the PDA was placed roughly in the center of the table.

A few of the earlier meetings also used a single lapel-style microphone instead of one of the head-worn microphones. Because of problems with background noise and crosstalk (hearing a neighboring voice on the lapel's channel), we stopped using the lapel mic early on in the data collection. We also moved to an all-wireless system for the head-worn microphones. Although this was a more expensive solution, it allowed participants to move about the room, and eliminated one of the most common hardware faults: broken connectors.

The waveform for each channel was stored in a separate file. The data were down-sampled on the fly from 48 kHz to 16 kHz, and encoded using 16 bit linear NIST SPHERE format. A software gain setting controlled which 16 of the 24 available bits per sample

were used.

We chose 16 kHz and 16 bits both to reduce the data storage requirements, and because higher quality settings do not appear to be necessary for automatic speech recognition systems.

Each file was then compressed using a lossless algorithm [3]. We obtained very good compression because the near-field signals contain a large amount of silence. As an example, a 55 minute meeting with 9 participants takes about 1.5 gigabytes of disk space uncompressed, while the actual compressed meeting takes 0.55 gigabytes of disk space.

Sections of a meeting that participants want excluded from public release (see section 4.2) were replaced with a pure tone on all channels. This is necessary since a speaker's voice is often audible on other channels. The corresponding text was also removed from the transcript.

The corpus as it will be released contains 75 meetings, for a total of about 72 hours. The meetings average about 6 participants per meeting, and each meeting also includes the audio from the 6 table-top microphones.

3. MEETING-SPECIFIC INFORMATION

Most of the meetings in the corpus are regularly scheduled weekly group meetings held at the ICSI in Berkeley, California. Some of the meetings (for example "Meeting Recorder" and "Robustness") have a significant number of speakers in common. Others are mostly speaker disjoint. The following meeting types have been recorded:

Name	Code	Count
Even Deeper Understanding	Bed	15
Meeting Recorder	Bmr	29
Robustness	Bro	23
Network Services & Applications	Bns	3
Other one-time only meetings	varies	5

In the *Even Deeper Understanding* (Bed) meetings, the participants discuss natural language processing and neural theories of language. The *Meeting Recorder* (Bmr) meetings are concerned with the ICSI Meeting Corpus. *Robustness* (Bro) involves methods to compensate for noise, reverberations, and other environmental issues in speech recognition. The *Network Services* (Bns) group researches internet architectures, standards, and related networking issues. The remaining recordings include meetings among the transcriptionists of the corpus, site visits from collaborators, and miscellaneous other meetings.

For each meeting, we store a small XML file describing some meeting-specific information:

Date-time stamp The date and time of the meeting. The duration can be inferred from the size of the audio files.

Unique tag Each meeting gets a unique tag consisting of the location of the meeting, the meeting type, and a number. For example, the meeting tag "Bro003" indicates the meeting took place in the meeting room at ICSI in Berkeley, the topic was "Robustness", and it was the third such meeting. A separate file provides a translation from the letter codes to the full description. The short, fixed-width tag allows for easy sorting of files that use the tag in their name. Note that the corpus as released only contains meetings recorded at ICSI in Berkeley.

Participant information Each speaker was assigned a channel (see next entry), a unique ID (see section 4 below), and a seat. The seats were labeled numerically clockwise around the table. The seat position provides an approximate location of the participant for speaker localization work, as well as providing adjacency information. Unfortunately, we did not start recording seat location until about 30 meetings were already recorded.

Channel The channel contains information related to the audio from a microphone. It includes a code for the microphone type (e.g. "s1" is a Sony ECM-310BMP headset mic), a code for the transmission method (e.g. "j1" for a wired jack, "w1" for a particular arrangement of Sony wireless transmitters and receivers), the software gain setting, and some statistical information about the audio on the channel (currently just the standard deviation of the amplitude). Providing a short code for microphone and transmission type allows us to easily sort by these factors. We include the statistical information simply because it can be expensive to compute.

Notes A free-form area for notes. Typically, we use it to note technical problems with a meeting (such as a dead battery in a wireless microphone), acoustic problems (e.g. "Lots of breath noise on channel 3"), people entering late or leaving early, etc.

4. SPEAKER INFORMATION

Each speaker was asked to fill out a speaker form prior to their first recorded session. The actual form is available on the web [4]. The data were then entered into an XML database.

Because all the meetings would have occurred regardless of the recording, we felt that the forms should be short and easy for the participants to fill out. If we interfered too much with the meeting process, not only could we bias the data, but groups might become reluctant to participate.

We request the speaker's name and contact information on the form, although, because of privacy concerns, this information is excised in the released corpus. Gender is also recorded.

A unique tag for each speaker is generated by concatenating **m** for male or **f** for female, followed by **e** for a native speaker of English[1] or **n** for a non-native speaker of English, followed by a unique three digit number. For example, a female non-native English speaker is **fn002**. The short tags allow easy sorting by the major categories of speakers.

Because of interest in meeting dynamics, and because most of the meetings were among people in a research setting, we ask for the education level of each participant. The possible choices are "Undergrad", "Grad", "PhD", "Professor", and "Other". We provide a write-in box for the "Other" category. For our group of participants, this provides a fairly good indication of position in the hierarchy.

We also request each speaker's age, although it is marked as optional on the form. We felt that making this information non-optional might compromise its accuracy. Of the 53 total unique participants, 45 provided their age. The youngest reported was 20 years, and the oldest reported was 62 years.

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

Information on the participant's language is split into two sections. The first section is filled out by all participants, and is intended to elicit information about the type of English used in the meeting. We ask, "What variety of English do you speak?" The possible answers are "Other", "American", "British", or "Indian". We provide a write-in box for the "Other" category. We also provide a write-in box for the region, although few participants filled in the region box.

Although it would have been possible to provide an exhaustive list of regions and varieties, we felt that only a trained linguist could really assign the categories. Since we wanted it to be very easy for the participant's to fill out the forms, we chose to provide only the write-in box for the region, and a very small number of choices for the variety.

The second part of the language information is for non-native speakers of English. We ask for their native language and the region, although again, few participants provided region information. We also wanted some indication of the speaker's proficiency in English. Rather than ask them to self-evaluate their English skills, we instead ask for the number of years spent in an English speaking country, and which country that was.

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.

4.1. Speaker Demographic Data

Figure 1 is a summary of the demographic data for the speakers. Note that for almost all of the questions, there were some participants who did not provide an answer. In these cases, they are grouped into the category "Unspecified". The exception is "Native Language", which we assume is English if the speaker fails to fill out the "For non-native English speakers" part of the speaker form.

Because region information was so variable, it is not included in figure 1. Also, we did not change any entries to conform to standard spelling, but rather kept the spelling as provided by the participants.

53	Unique speakers
13	Female
40	Male
#	Age
18	20–29
18	30–39
4	40–49
4	50–59
1	60+
8	Unspecified
#	Native Language
28	English
12	German
5	Spanish
1	Chinese
1	Czeck
1	Dutch
1	French
1	Hebrew
1	Malayalam
1	Norwegian
1	Turkish

#	Education Level
21	Grad
20	PhD
7	Professor
4	Undergrad
1	Postdoc
#	Variety of English
36	American
6	British
2	German
2	Indian
1	Czeck
1	Norwegian
5	Unspecified
#	Time Spent in English Speaking Country
9	$<$ 1 Year
3	1–2 Years
4	2–5 Years
6	$>$ 5 Years
3	Unspecified

Fig. 1. Speaker Demographics

4.2. Participant Approval

In addition to the speaker form, we also ask each participant to sign an approval form. The form fulfills several functions. First, it briefly describes the project to the participants. Second, it satisfies the University of California's Human Subjects committee requirements. Finally, it explicitly states that participants are responsible for monitoring their own speech, and must inform us of deletion requests. We did not censor any data except as specifically requested by participants. This includes instances of people's names being spoken during the meeting — we removed them only if requested to do so.

Because of privacy concerns, the participant approval forms are not part of the released corpus.

To aid the participants in checking the data, we provide a web interface to the transcripts and the audio. The interface allows the participants to listen to the meetings and review the transcripts, to mark sections as incorrectly transcribed, and to mark sections that the participant wants removed from the corpus. The participants then have 3 weeks to request deletions before the data is marked as finalized.

In the entire corpus, there were 19 requests for data removal in 7 meetings totaling 2.6 minutes. Of that, one passage accounts for more than one minute.

5. TRANSCRIPTS

For each meeting, the corpus contains an XML file with a word-level transcription. In addition to the full words, other information is also provided, such as word fragments, restarts, filled pauses, back-channels, contextual comments (e.g. "while whispering"), and non-lexical events such as cough, laugh, breath, lip smack, door slam, microphone clicks, etc.

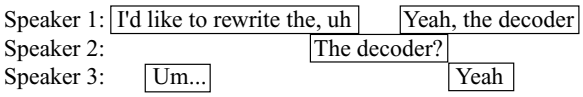

Fig. 2. Example of overlap.

Overlap between participant's speech is *extremely* common in our meetings [5]. Therefore, it is important to capture the details. Figure 2 shows a section of a meeting with overlap.

In the transcript, we mark the speaker, the start time, and the end time of each of the utterances. So for figure 2, there would be 5 entries.

The process of transcribing the data was quite complex. Full details are beyond the scope of this paper, but a brief description, applicable to most of the transcripts, follows.

Each of the near-field signals was transcribed separately, and went through several passes of transcription, correction, and quality assurance. For the first pass, we linearized the data by taking each chunk of speech as derived from a speech-activity detector, and sequentially pasted them together. Figure 3 shows an example

of linearizing the speech chunks from figure 2. The linearized audio was sent to a commercial transcription service. Upon return, we divided the audio back into a separate channel for each speaker. We then corrected the first pass transcription using a version of the Transcriber [6] tool modified for multiple channels. Finally, a senior transcriptionist verified the data.

I'd like to rewrite the, uh	Um...	The decoder?	Yeah, the decoder	···

Fig. 3. Linearization of Figure 2.

The XML transcription format was designed specifically for this collection. A complete DTD and description of the format will be distributed with the corpus. We will also provide software for translating from our format to other common formats.

When a participant requests that a portion of the meeting be removed, we "bleep" the audio portion (see section 2), and remove from the transcript any words which overlap the requested section.

6. CONCLUSIONS

We presented a description of the ICSI Meeting Corpus, which contains audio and transcripts of natural meetings recorded simultaneously with head-worn and table-top microphones. The corpus contains 75 meetings of 4 main types and 53 unique speakers. We will deliver the corpus to the LDC [1] by December, 2002, and expect it to be available through the LDC by the summer of 2003.

We have published many papers related to the corpus, including research on automatic transcription, speech activity detection for segmentation, overlap analysis, applications, prosody, automatic punctuation, noise robustness, and reverberation. For an overview, please see [7], which appears in the special session on Smart Meeting Rooms in these Proceedings. For a complete listing of publications from ICSI on the Meeting Corpus, see our web page [4].

In addition to the data released with the corpus, we also continue to annotate the corpus with additional information, including dialog act labeling and prosodic features. We hope that others will also contribute to the corpus, either with additional meeting data, or with more annotations of the existing data.

7. RELATED WORK

Several other groups have collected and analyzed meeting data similar to the ICSI Corpus, often with the addition of video and collaboration tools. These include Carnegie Mellon University [8], LDC [9], NIST [10], and the IM2 [11] and M4 [12] projects. Many commercial companies have investigated meeting capture, which is similar in many ways to our work. Some recent examples include Xerox [13] and Microsoft [14].

8. ACKNOWLEDGMENTS

We would like to thank the many people who kindly consented to be recorded for this corpus. Thanks are also due to Jim Beck for help with setting up the hardware, and to Jennifer Alexander, Helen Boucher, Robert Bowen, Jennifer Brabec, Mercedes Carter, Hannah Carvey, Leah Hitchcock, Joel Hobson, Adam Isenberg, Julie Newcomb, Cindy Ota, Karen Pezzetti, Marisa Sizemore, and Stephanie Thompson, the ICSI transcribers. We would also like to thank our collaborators at other sites, most notably: Mari Ostendorf, Jeff Bilmes, and Katrin Kirchhoff from University of Washington; Hynek Hermansky from the Oregon Graduate Institute; and Brian Kingsbury of IBM.

This work was funded in part under the DARPA Communicator project (in a subcontract from the University of Washington), in part by a ROAR "seedling" from DARPA, and supplemented by an award from IBM.

9. REFERENCES

[1] "Linguistic data consortium (LDC) web page," http://www.ldc.upenn.edu/.

[2] 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.

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

[4] "ICSI meeting corpus web page," http://www.icsi.berkeley.edu/speech/mr.

[5] 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.

[6] "Transcriber web page," http://www.etca.fr/CTA/gip/Projets/Transcriber/.

[7] Nelson Morgan, et. al, "Meetings about meetings: research at ICSI on speech in multiparty meetings," in *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-2003)*, Hong Kong, April 2003.

[8] Alex Waibel, Michael Bett, Florian Metze, Klaus Ries, Thomas Schaaf, Tanja Schultz, Hagen Soltau, Hua Yu, and Klaus Zechner, "Advances in automatic meeting record creation and access," in *Proceedings of ICASSP 2001*, May 2001.

[9] Christopher Cieri, David Miller, and Kevin Walker, "Research methodologies, observations and outcomes in (conversational) speech data collection," in *Proceedings of HLT 2002 The Human Language Technologies Conference*, San Diego, March 2002.

[10] "NIST automatic meeting transcription project," http://www.nist.gov/speech/test_beds/mr_proj/.

[11] "IM2 web page," http://www.im2.ch.

[12] "M4 web page," http://www.dcs.shef.ac.uk/spandh/projects/m4.

[13] Patrick Chiu, Ashutosh Kapuskar, Lynn Wilcox, and Sarah Reitmeier, "Meeting capture in a media enriched conference room," in *Cooperative Buildings*, 1999, pp. 79–88.

[14] Ross Cutler, Yong Rui, Anoop Gupta, JJ Cadiz, Ivan Tashev, Li wei He, Alex Colburn, Zhengyou Zhang, Zicheng Liu, and Steve Silverberg, "Distributed meetings: A meeting capture and broadcasting system," in *ACM Multimedia*, Juan Les Pins, France, December 2002.