The NIST Meeting Room Pilot Corpus

John S. Garofolo[°], Christophe D. Laprun[°†], Martial Michel[°†],
Vincent M. Stanford[°], Elham Tabassi[°]

[°] National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg MD 20899, USA
[†] Systems Plus Inc., 1370 Piccard Drive, Suite 270, Rockville, MD 20850, USA
{john.garofolo | chris.laprun | martial.michel | vince.stanford | elham.tabassi}@nist.gov

Abstract

One of the next big challenges in Automatic Speech Recognition (ASR) is the transcription of speech in meetings. This task is particularly problematic for current recognition technologies because, in most realistic meeting scenarios, the vocabularies are unconstrained, the speech is spontaneous and often overlapping, and the microphones are inconspicuously placed.

To support the development of meeting recognition technologies by both the speech recognition and video extraction research communities, NIST is providing a development and evaluation infrastructure including: a multi-media corpus of audio and video from meetings collected at NIST using a variety of microphones and video cameras, new evaluation protocols, metrics, software, rich transcription conventions, sponsoring evaluations and workshops, facilitating multi-site data pooling, and helping bring the community together to focus on the technical challenges.

To date, NIST has collected a pilot corpus of 15 hours of meetings in its specially-instrumented Meeting Data Collection Laboratory. The corpus includes digital recordings from close-talking mics, lapel mics, distantly-placed mics, 5 digitally-recorded camera views, and full speaker/word-level transcripts. This data is being used in the development and evaluation of speech technologies and by the video extraction community under the auspices of the ARDA Video Analysis and Content Exploitation (VACE) program.

Motivation

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

The development of smart meeting room core technologies that can automatically recognize and extract important information from multi-media sensor inputs will provide an invaluable resource for a variety of business, academic, and governmental applications. Such metadata will provide the basis for the development of second-tier meeting applications that can automatically process, categorize, and index meetings. Third-tier applications will provide a context-aware collaborative interface between live meeting participants, remote participants, meeting archives and vast online resources. Given that the necessary core meeting recognition technologies are in a fledgling or nonexistent state, it is essential that these first tier technologies be developed before the higher tier applications can be made useful.

The meeting domain has several important properties not found in other domains and which are not currently being focused on in other research programs:

- Multiple Forums and Vocabularies -- Meeting forums range from very informal to highly structured. Likewise, meeting vocabularies vary widely depending on both the meeting topic and degree of shared context among the participants.
- Highly-Interactive/Simultaneous Speech -- The speech found in informal meetings is spontaneous and highly interactive across multiple participants and contains frequent interruptions and overlapping speech. This poses great challenges to speech recognition technologies that are typically tailored for single-speaker speech streams.
- Multiple Distant Microphones – Meetings are typically recorded with multiple distant microphones. Speech recognition systems generally work quite poorly with distant mics. Moreover, techniques have yet to be developed which efficiently integrate input from multiple mics and take advantage of their positioning to improve recognition quality.
- Multiple Camera Views – Meetings are often recorded with multiple cameras with different and sometimes overlapping views. Much like the multi-microphone challenge above, this permits/challenges the technology to integrate data from multiple video inputs to enhance the metadata that can be extracted from the meeting and improve recognition quality.
- Multi-Media Information Integration -- It is impossible to develop a complete understanding of meetings without analyzing a number of different signal types simultaneously: audio, video and other information sources (devices/resources participants interact with).

Introduction

NIST is working to address these issues by supporting the development of audio and video recognition technologies in the context of human-human meetings. The NIST Meeting Recognition Project includes periodic technology evaluations and workshops as well as an extensive data collection effort. In addition to collecting its own multi-media meeting corpora, NIST is collaborating with several other data collection sites to provide a broad base of corpora for research, development and evaluation. The recognition technologies currently addressed by the effort include speech-to-text transcription, speaker segmentation, and video extraction in collaboration with the ARDA Video Analysis and Content Exploitation (VACE) Program.

Data collection facility

NIST has constructed a Meeting Data Collection Laboratory (MDCL) to collect corpora to support meeting domain research, development and evaluation. The lab is equipped to look and "sound" like a conventional meeting room. As such, sensors have been inconspicuously placed and noisy processors have been located outside of the room. The background noise level in the room has been measured at 42 dB A-weighted (with the video projector turned off.) The MDCL was used to collect the NIST Meeting Pilot Corpus.

Data Streaming and Synchronization

The MDCL contains a variety of microphones and several video cameras. The NIST Smart Data Flow architecture (Michel et al., 2003), developed by the NIST Smart Spaces Laboratory, streams and captures all of the sensor data from 200 mics and 5 video cameras on 9 separate data collection systems in a proprietary time-indexed "SMD" format. This architecture also ensures that all data streams are synchronized (via the Network Time Protocol and NIST atomic clock signal) to within a few milliseconds. This unique capability also permits the data collection system to be (theoretically) plugged into real-time recognition systems. We believe this approach provides a good prototype of the front-end sensor systems that might be found within future smart meeting rooms.

Meeting Room Layout

The room is approximately 22 X 22 feet (6.7 X 6.7 meters) and can be configured for a variety of meeting forums (conference, round table, classroom). However, to support initial multi-mic experiments, all of the meetings in the pilot corpus were collected using a single conference table configuration (shown in Figure 1.)

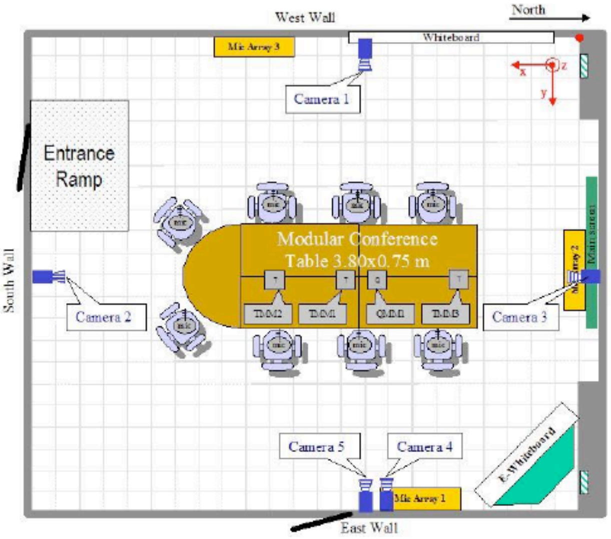


Figure 1 - NIST Meeting Data Collection Facility

Video Capture

The MDCL includes 5 Sony EVI-D30 motorized pan/tilt/zoom NTSC analog video cameras -- 4 having stationary views of the conference table (1 view from each surrounding wall), and an additional "floating" camera which is used to focus on particular participants, whiteboard, or conference table depending on the meeting forum. The video is encoded at 29.97 frames per second at 720x480 resolution. It was converted, from the SMD format, to MPEG-2 for distribution.

The MDCL also employs Camtasia Studio™ to capture the room's PC/projection screen. This capability was added late in the pilot data collection cycle, so it was used in only a few of the pilot corpus meetings. This data was recorded at 5 fps in a proprietary format and converted to MPEG-2 for distribution.

Audio Capture

Commercial Microphone Data Capture

Each meeting participant was equipped with 2 "personal" microphones (a noise-canceling Shure WCM-16 hyper-cardioid electret condenser headset mic and a Shure WL-185 cardioid electret condenser lapel mic). Both personal mics were wireless, so participants were free to move about the room. The conference table was equipped with 4 microphones: 3 Audio Technica AT841a omni-directional condenser boundary mics were positioned at the center and ends of the table and an Audio Technica AT854R 4-channel condenser boundary mic was placed at the center of the table. The AT854R is unusual in that it contains 4 separate cardioid boundary mics, together covering 360 degrees. Each channel was recorded separately. A speakerphone system with audio tap was also available, but was infrequently used and was not recorded in the pilot corpus.

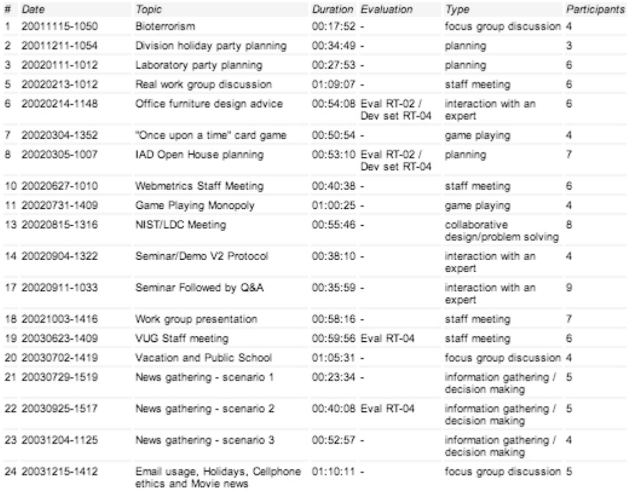
These mics were collected by a 24-channel 48-KHz/24-bit A/D system into the NIST SMD format. The data was then SPHERE-encoded, down-sampled to 16-KHz/16-bit and gain-normalized for distribution in single-channel files.

Linear Array Microphone Data Capture

To support farfield recognition experiments, three prototype NIST Smart Spaces Lab Mark-II 59-channel linear array microphones were positioned on pole mounts against the front and side walls of the room. Each array channel was collected at 22,050Hz/16-bit by a distant A/D in a PC located under the raised floor. Unfortunately, these PCs suffered from a variety of technical problems from the under-floor environment and we were unable to collect enough usable array data for distribution.

Pilot Corpus

The NIST Meeting Room Pilot Corpus consists of 19 meetings/15 hours recorded between 2001 and 2003. In total, the multi-sensor data comes to 266 hours of audio and 77 hours of video. Four meetings in the corpus have been used in the RT-02 and RT-04S evaluations. The corpus is described in Table 1.

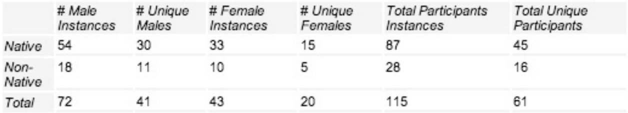


#	Date	Topic	Duration	Evaluation	Type	Participants
1	20011115-1050	Bioterrorism	00:17:52	-	focus group discussion	4
2	20011211-1054	Division holiday party planning	00:34:49	-	planning	3
3	20020111-1012	Laboratory party planning	00:27:53	-	planning	6
5	20020213-1012	Real work group discussion	01:09:07	-	staff meeting	6
6	20020214-1148	Office furniture design advice	00:54:08	Eval RT-02 / Dev set RT-04	interaction with an expert	6
7	20020304-1352	"Once upon a time" card game	00:50:54	-	game playing	4
8	20020305-1007	IAD Open House planning	00:53:10	Eval RT-02 / Dev set RT-04	planning	7
10	20020627-1010	Webmetrics Staff Meeting	00:40:38	-	staff meeting	6
11	20020731-1409	Game Playing Monopoly	01:00:25	-	game playing	4
13	20020815-1316	NIST/LDC Meeting	00:55:46	-	collaborative design/problem solving	8
14	20020904-1322	Seminar/Demo V2 Protocol	00:38:10	-	interaction with an expert	4
17	20020911-1033	Seminar Followed by Q&A	00:35:59	-	interaction with an expert	9
18	20021003-1416	Work group presentation	00:58:16	-	staff meeting	7
19	20030623-1409	VUG Staff meeting	00:59:56	Eval RT-04	staff meeting	6
20	20030702-1419	Vacation and Public School	01:05:31	-	focus group discussion	4
21	20030729-1519	News gathering - scenario 1	00:23:34	-	information gathering / decision making	5
22	20030925-1517	News gathering - scenario 2	00:40:08	Eval RT-04	information gathering / decision making	5
23	20031204-1125	News gathering - scenario 3	00:52:57	-	information gathering / decision making	4
24	20031215-1412	Email usage, Holidays, Cellphone ethics and Movie news	01:10:11	-	focus group discussion	5

Table 1: Pilot Meeting Corpus Meetings

Subjects

The participants in the pilot corpus meetings were largely NIST volunteers. Minimal effort was made to control the subject population except to maintain a balance between native/non-native speakers of American English and to minimize the occurrence of subjects with obvious speech defects. The demographics are described in Table 2.



	# Male Instances	# Unique Males	# Female Instances	# Unique Females	Total Participants Instances	Total Unique Participants
Native	54	30	33	15	87	45
Non-Native	18	11	10	5	28	16
Total	72	41	43	20	115	61

Table 2: Pilot Meeting Corpus Subject Demographics.

Pilot Corpus Meeting Types

The pilot corpus contains a mix of real meetings (which would have occurred anyway) and scenario-driven meetings (in which participants were given an artificial task to carry out) to provide a broad coverage of different meeting types and behaviors.

Meetings can range from informal, collaborative exchanges (ad hoc work group meetings) to highly structured, formal gatherings (legislative/judicial proceedings) and can be described by a set of group interaction characteristics (McGrath, 1984, Cugini et al., 1997). Accordingly, meetings can be broken down into their constituent collaborative work tasks.

For the pilot corpus, we selected a set of informal meeting types that would exhibit different types of meeting behaviors (as identified by McGrath and extended by Cugini) which would elicit natural interactions. In particular, scenarii for simulated meetings were selected to be of interest to the general population. Table 3 lists the meeting types collected in the pilot.

Meeting Type	Work Task Types	Real?
Party planning	Planning	Yes
Interactive game playing	competitive performance, dissemination of information	No
Staff meeting	Planning, negotiation, brainstorming, decision-making, dissemination of information	Yes
Interaction with expert	decision-making, dissemination of information	No
Focus group*	Dissemination of information	No
Executive summary generation	dissemination of information, decision making	No
Technical presentation	Dissemination of information	Yes

*Focus groups covered a wide topic range: bioterrorism, public schools, vacations, email, holidays, movies, cellphone ethics.

Table 3: Pilot Meeting Corpus Meeting Types

Evaluations

Rich Transcription 2002 Evaluation

NIST carried out the first community-wide evaluation of meeting domain speech-to-text transcription (STT) and speaker segmentation (SPKR) in the context of its Rich Transcription 2002 (RT-02) evaluation (NIST, 2002, eval). An 80-minute test set with 8 10-minute meeting excerpts collected at NIST, CMU (Burger, 2002), ICSI (Janin, 2003), and the LDC (Cieri, 2002) was used. Performance was measured for individual personal mics (head or lapel mics depending on data collection site), a single distant omnidirectional mic, and a personal mic mix. NIST applied its SCLITE and speaker segmentation scoring software to evaluate the tasks. As such, overlapping speech was not evaluated. However, unlike broadcast news and 4-wire telephone speech (both having little within-channel overlap), this was identified as a significant issue for recognition from distant microphones.

This exploratory evaluation included few participants (SRI and MIT-Lincoln Labs), but proved the feasibility of meeting domain evaluation and provided a performance baseline (NIST, 2002, workshop). Accordingly, the participants performed little domain-specific development. The evaluation showed that performance for the individual close-talking microphone condition was similar to that of conversational telephone speech. However, performance for the single distant microphone condition was significantly worse than for the individual personal mic condition (nearly twice as high absolute for the STT task – even excluding overlap.) A great deal of variability was also observed across meetings and data collection. Given these results, meeting research sites are now focusing on the distant mic problem as well as meeting-specific metadata extraction (Morgan, 2003).

Rich Transcription 2004 Spring Evaluation and ICASSP 2004 Meeting Recognition Workshop

NIST is conducting its second evaluation of meeting domain STT and SPKR technologies in March 2004. A new 90-minute/8-excerpt test set collected at NIST, CMU, ICSI, and the LDC is being used. Unlike in RT-02, the RT-04S personal mic condition includes only head-mic data. Of even greater importance, however, is that the primary evaluation condition is over multiple distant microphones. A single-distant mic condition is supported as a contrast. Also significant is that overlapping speech is to be scored. The results of the evaluation are to be reported at the ICASP-2004 Meeting Recognition Workshop in Montreal. The workshop will be devoted to discussion of audio and video meeting recognition technologies, evaluation, and future collaborations. Details regarding the evaluation are available on the RT-04S website (NIST, 2004).

VACE Meeting Domain Evaluation

NIST is collaborating with the ARDA VACE Program to begin evaluation of video extraction technologies in the meeting domain including text/face/hand/person detection/tracking, gesture/gaze recognition, and higher-level video event-based tasks. The first such evaluation is to occur in 2004-2005.

Lessons Learned

Quite a bit of knowledge was gained in designing, collecting, and distributing the NIST pilot and multi-site evaluation corpora.

The process of creating an infrastructure to time-synchronize all of the audio and video data was quite complicated. Issues inherent in using a non-real time operating system to generate time tags on multiple hardware systems necessitated the development of statistical techniques to correct for clock drift. After half the data collection was completed, we discovered an indexing bug that necessitated hand-synchronization of the data. To simplify future such issues, we collected the remaining meetings using a flash-equipped movie clapper.

With the complexity involved in the data collection system, a great deal of quality control was necessary to ensure that all of the channels were recorded properly. Sound levels for all the mics and views of all the cameras were checked via a continually-enhanced monitor workstation prior to each meeting and during recording.

The mic gain levels were calibrated before collection of the pilot corpus so as to avoid clipping. However, in preparing the data for distribution, we realized that our actual 16-bit gain levels were low, so we gain-adjusted all audio channels using the original 24-bit recordings.

We learned that long cable runs cause signal quality problems with analog devices – no matter how well they're shielded. We're therefore moving to mic arrays with onboard A/D and digital video cameras for the next data collection effort.

Decisions regarding the types of meetings to collect and number and types of participants is non-trivial. More complex yet is the design of scenarios for simulated meetings. We intend to employ a multi-disciplinary team in the design of new scenarios to broaden the types of meeting behaviors in our next meeting corpus.

The creation of a multi-site evaluation corpus is fraught with QC challenges. A great deal of work was spent in homogenizing the multi-site data prior to the evaluation.

Future

We plan to collect a new 20 hour meeting corpus in 2004-2005. For that collection, we will employ 4 new NIST Smart Spaces 64-element Mark-III mic arrays. An array will be positioned at each wall, suspended from the ceiling. We'll be replacing our Shure head mics with very low profile Countryman E6 directional ear-mounted mics which don't obstruct the participants' lips. We also plan to replace our analog video cameras with digital cameras. We plan to employ a linear regression model and Kalman filter for correction of the time indices to improve synchronization. We'll be employing new scenarii which facilitate gesticulation and more varied interaction and we'll be varying the room configuration and lighting – parameters which will provide useful data for video extraction research. We welcome your input.

We plan to continue our series of meeting-domain speech technology evaluations, working in concert with new meeting recognition programs both in the US and abroad. We have begun to collaborate with several US and European sites in creating a multi-site corpus of array mic data using the new NIST Mark-III arrays. We're also working with the VACE community to define and implement evaluations of video event technologies in the meeting domain.

We see meeting understanding using information integrated from multiple sensors, sensor types, and multi-tiered recognition technologies as the next frontier. We'll continue to provide resources to support research and evaluation in this important area. Our activities in this area will be documented on our website at www.nist.gov/speech/test_beds/mr_proj/.

Caveat

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

References

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

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

Cugini, J., Damianos, L., Hirschman, L., Kozierok, R., Kurtz, J., Laskowski, S., Scholtz, J. (1997). The Evaluation Working Group of the DARPA Intelligent Collaboration and Visualization Program.

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

McGrath, J. E. (1984). Groups: Interaction and Performance. Englewood Cliffs, N. J., Prentice-Hall.

Michel, M., Stanford, V., Galibert, O. (2003). Network Transfer of Control Data: An Application of the NIST Smart Data Flow. Proc CCCT 2003.

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

NIST (2002). Rich Transcription 2002 Meeting Recognition Evaluation, documentation, http://www.nist.gov/speech/tests/rt/rt2002/

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

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