THE NIST SMART SPACE AND MEETING ROOM PROJECTS: SIGNALS, ACQUISITION, ANNOTATION, AND METRICS

Vincent Stanford, John Garofolo, Olivier Galibert, Martial Michel, and Christophe Laprun

National Institute of Standards and Technology, Information Technology Laboratory,
Information Access Division, 225 Technology Building, Gaithersburg, Maryland 20899

ABSTRACT

Pervasive Computing devices, sensors, and networks, provide infrastructure for context aware smart meeting rooms that sense ongoing human activities and respond to them. This requires advances in areas including networking, distributed computing, sensor data acquisition, signal processing, speech recognition, human identification, and natural language processing. Open interoperability and metrology standards for the sensor and recognition technologies can aid R&D programs in making these advances. The NIST Smart Space and Meeting Room projects are developing tools for data formats, transport, distributed processing, and metadata. We are using them to create annotated multi modal research corpora and measurement algorithms for smart meeting rooms, which we are making available to the research and development community.

1. OVERVIEW: SMARTFLOW DATA AND ATLAS METADATA

The vision of Smart meeting rooms, our theme for this ICASSP Special Session, must include multi-sensor, multi-modal, user interfaces which are able to react to the desires of the groups meeting within them. Our Smart Space, which can be found at (www.nist.gov/smartspace), and Meeting Room, (www.nist.gov/speech/test_beds/mr_proj) projects at NIST address issues in data and metadata standardization. Data in the Meeting Room are digitized signals from microphones, and video cameras, i.e. direct sensor data, and meeting metadata are semantically meaningful tags like spoken words and others described below.

We address broad issues of data using the NIST SmartFlow system, which is a set of tools that allow components from various developers to interoperate in a data rich environment containing flows from many sensors, comprising a reference implementation for laboratory use. The NIST Meeting Room has over two hundred microphones, five cameras, a smart whiteboard, and will soon have a locator system for the meeting attendees. These generate over a gigabyte per minute of sensor data, which are time tagged to millisecond resolution and stored for research uses. Significant engineering challenges emerge in developing data acquisition and distributed processing systems to acquire and track this profusion of raw data. We manage the sensor streams using the NIST SmartFlow system, a data flow middleware layer that provides a data transport abstraction, and offers consistent formats for the data streams.

We address broad issues of metadata, or annotations of the data streams, with semantic descriptions using the Architecture and Tools for Linguistic Analysis Systems (ATLAS). Standardization of metadata derived directly from the sensor data streams, and subsequent higher-level annotations of meeting context, which may allow indexing, transcription, and possibly even summarization, are one of the major design features of ATLAS. Some significant meeting metadata under investigation include:

- Spoken words
- Speaker identity
- Sentence-like units and disfluencies
- Speaker locations
- Time tags

Beyond these low-level metadata, smart meeting rooms will eventually have to make higher-level inferences about tasks the users are undertaking to become context aware. But this will only be possible using a hierarchy of metadata, each level building upon those below it. Also, numerous problems in sensor fusion, and cross-channel registration, must be solved before activities can be automatically recognized in Smart meeting rooms. Realizing the goal of a context aware meeting room system will take time, and will stimulate many research projects along the way.

Fortunately, though, interesting possibilities for collaborative interfaces open, well before complex activities can be recognized. For example a meeting chairman could dictate minutes, or command presentation resources, retrieve information using spoken language, while discourse by other meeting attendees can be ignored.

In 2002, we ran our first common evaluation of meeting speech recognition collected using the SmartFlow data transport system and annotated using ATLAS. We found that special challenges posed by meeting speech, e.g.: distant microphones, room acoustics, and overlapping simultaneous speech, make recognition in this domain more difficult than any previously attempted.

0-7803-7663-3/03/$17.00 ©2003 IEEE

2. COMPONENTS AND DIRECTIONS IN SMART MEETING ENVIRONMENTS

A discussion of the technologies needed by smart meeting rooms will help clarify the engineering, metrology, and standards issues the R&D community will need to address. Flexible smart meeting rooms require integration of technologies including:

Devices: Wireless PDAs, smart locator badges, and wearable computers. These will transport authentication and biometric characterizations for things like speaker, speech, and face recognition. Most current recognition technologies, e.g. large vocabulary speech recognition systems, require speaker specific training, to achieve satisfactory recognition performance.

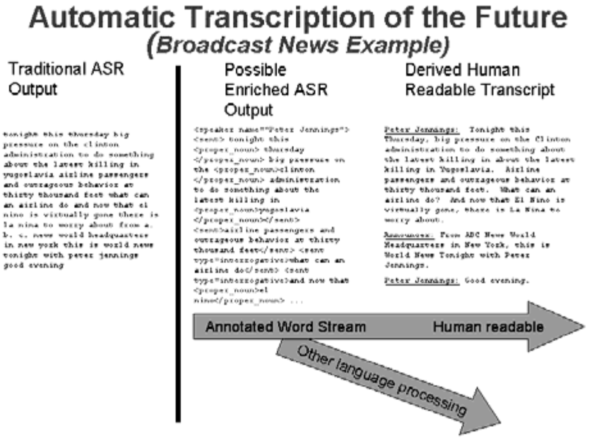
Networking: Dynamic discovery, authentication, and incorporation of pervasive devices people wear and carry into smart meeting rooms. Some fairly well developed standards for this include Java/Jini, UPnP, Rendezvous, ZeroConf, and Bluetooth.

Interfaces – At the core of smart meeting rooms are multi-modal user interfaces employing numerous devices and sensors that create interfaces centered on people rather than devices. Clearly, smart meeting room interfaces must make significant developments over conventional computer interfaces in order to function in this way. Some will include transitions from:

- Bimodal to Multi-modal
- Passive to Perceptual/Reactive
- Nominal to Personal
- Individual to Collaborative

The dominant computer window interface is passive. It waits for keys to be pressed, and changes its display in response. Video and acoustic sensors, combined with appropriate recognition technologies, will provide data for Perceptual/Reactive interfaces that see and hear users as they work. Biometrics will allow the interface to identify users as they speak and move in the room, which will make new levels of selective responsiveness possible. Applicable recognition technologies exist, although at varying levels of maturity. Some include: speech recognition, speaker identification, face recognition, gesture, pen input, and even gaze tracking.

These personalized, user aware, interfaces will use metadata from the perceptual interface systems, including, say, who is asking the system to “Show me my appointments for the day.” Or user aware smart meeting interfaces could offer accessibility by the disabled. For example, someone with a mobility limitation could have his pervasive devices negotiate with a smart meeting room to launch hands free services such as microphone arrays, speech recognition and head tracking to dictate to the interface in combination with the head driven pointer. We are also working with the INCITS (www.ncits.org) V2 technical committee on Information Technology Access Interfaces to define XML transaction structures to make such user preferences known to smart meeting spaces. Another example is that meeting groups attempting to collaborate often ask a member to act as the secretary and capture, as best he can, the proceedings in notes. The perceptual interfaces will allow the smart room to be the secretary, say taking meeting minutes from the chairperson by in response to a command. The NIST Rich Transcription Evaluation series, which began in 2002 seeks to support the development of these technologies. The columns of figure 2 show a raw machine generated transcript, XML metadata enrichments, and human-readable form.



Automatic Transcription of the Future

(Broadcast News Example)

Traditional ASR Output

```
tonight this thursday big
pressure on the clinton
administration to do something
about the latest killing in
yugoslavia airline passengers
and outrageous behavior at
thirty thousand feet what can
an airline do and now that el
nino is virtually gone there is
la nina to worry about from a.
b. c. news world headquarters
in new york this is world news
tonight with peter jennings
good evening
```

Possible Enriched ASR Output

```
<speaker name""Peter Jennings">
<sent> tonight this
<proper_noun> thursday
</proper_noun> big pressure on
the <proper_noun>clinton
</proper_noun> administration
to do something about the
latest killing in
<proper_noun>yugoslavia
</proper_noun></sent>
<sent>airline passengers and
outrageous behavior at thirty
thousand feet</sent> <sent
type"interrogative>what can an
airline do</sent> <sent
type"interrogative>and now that
<proper_noun>el
nino</proper_noun> ...
```

Derived Human Readable Transcript

Peter Jennings: Tonight this Thursday, big pressure on the Clinton administration to do something about the latest killing in about the latest killing in Yugoslavia. Airline passengers and outrageous behavior at thirty thousand feet. What can an airline do? And now that El Nino is virtually gone, there is La Nina to worry about.

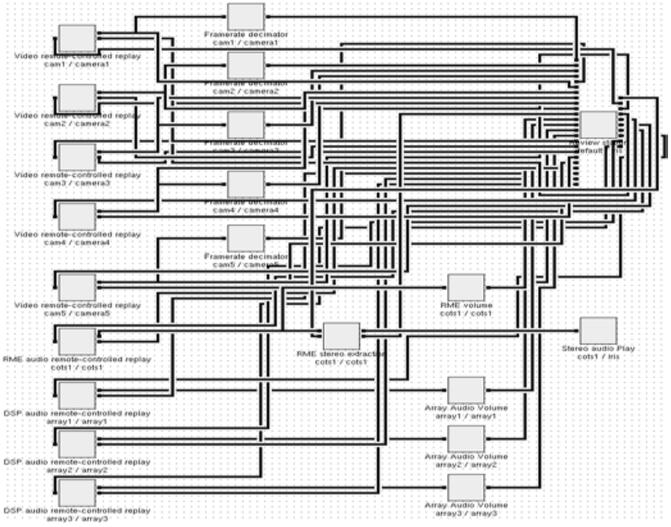
Announcer: From ABC News World Headquarters in New York, this is World News Tonight with Peter Jennings.

Peter Jennings: Good evening.

Figure 2 – Rich transcription concept: raw transcript, XML metadata enrichments, and human readable version.

3. SMARTFLOW: THE NIST DATA TRANSPORT AND INTEROPERABILITY TOOLKIT

The NIST SmartFlow system was developed in response to the need to provide connectivity to the large number of sensors and devices that will be needed to construct the smart meeting rooms [1].



Framerate decimator cam1 / camera1
Video remote-controlled replay cam1 / camera1
Framerate decimator cam2 / camera2
Video remote-controlled replay cam2 / camera2
Framerate decimator cam3 / camera3
Video remote-controlled replay cam3 / camera3
Framerate decimator cam4 / camera4
Video remote-controlled replay cam4 / camera4
Framerate decimator cam5 / camera5
Video remote-controlled replay cam5 / camera5
Review default
RME volume cots1 / cots1
RME audio remote-controlled replay cots1 / cots1
RME stereo extraction cots1 / cots1
Stereo audio Play cots1 / iris
DSP audio remote-controlled replay array1 / array1
Array Audio Volume array1 / array1
DSP audio remote-controlled replay array2 / array2
Array Audio Volume array2 / array2
DSP audio remote-controlled replay array3 / array3
Array Audio Volume array3 / array3

Figure 1- An executable SmartFlow graph for Meeting Room data processing and review.

Figure 1 shows an example data flow graph for data review in the NIST Meeting Room. The SmartFlow system generates the connections and transports the data among the clients.
Hand crafting the inter-process communication was found to be very labor intensive, and brittle with respect to changing requirements for new sensors and configuration changes to accommodate equipment faults. Dragging and dropping the components and naming of the flows can be used to reconfigure the application flow graphs. The system consists of a defined middleware API for real-time data transport, and a connection server for sensor data source, processing, display, and storage graph nodes.

An example application in our Meeting Room consists of a microphone array, which acquires a sixty-four channel audio input flow and offers it for subscription. A beamformer client subscribes to the multi channel flow and reduces it to a single channel, and offers it as an audio flow. A speaker identification system subscribes to this flow, and a speaker-dependent speech recognition client also subscribes to the audio and the speaker identity flows. Speech segments that match the correct speaker can then be recognized by a speaker dependent dictation system. The SmartFlow layer makes it possible to integrate components from multiple sources, such as speaker identification and speech recognition systems. With many sensor and recognition technologies under separate development in industry, issues of interoperability and integration are crucial in new generation smart working environments. The SmartFlow system is being used to integrate technologies including:

- Speech recognition
- Speaker identification
- Face localization and recognition
- Source localization/separation
- Channel normalization
- Video and acoustic displays
- Integration of wireless PDAs
- Standardized formats for multimedia data streams
- Archiving, retrieval, and review tools

The SmartFlow toolkit has components for graphical configuration of flow graphs, and allocation of the graph nodes to distributed systems, and connection by TCP/IP. The data transport code is provided in the SmartFlow libraries. It also has a code generator for a simplified construction of clients that operate as nodes of SmartFlow graphs. This can very substantially reduce the systems programming burden in research and development laboratories working on multi modal sensor based interfaces, or advanced classification systems. Our goal is to work with industrial and academic users to refine and develop this system to better to meet their needs as they develop future systems. We hope that it forms the basis of a standards working group for smart environments and pervasive computing.

3. METADATA ARCHITECTURE AND TOOLS FOR LINGUISTIC ANALYSIS SYSTEMS (ATLAS)

ATLAS is designed to support the integrated annotation of a variety of phenomena from multi-modal sensor signals in meeting rooms. It is a signal-independent, linguistic annotation framework, which provides infrastructure for creation and management of necessary metadata. It does not prescribe a single approach or tag set, but rather allows for defining and managing most signal-based annotation schemes. It is based on several simple but powerful ideas as follows.

First, being designed to represent a wide variety of possible annotations, it is based a relatively simple, but generic, annotation ontology; so it is expressive enough to represent annotations for diverse phenomena, some linguistic and some not. Its framework allows annotations, or content definitions, to be associated with regions identified in a signal. Using its content ontology as a starting point, a data model has been developed, providing a set of core constructs and associated relations. These constructs can be understood as basic building blocks that ATLAS can use and combine hierarchically to create more complex annotations, and new annotation schemes. Note also that this data model is signal-independent, meaning that the same concepts can be applied to text, audio, video or more complex data sources as appropriate. Moreover, data model structures can be represented using an XML file format for easy interchange.

Second, ATLAS provides a software infrastructure to implement the data model and provides useful services, e.g.: serialization of annotations, and automatic validation. It makes these accessible from an Application Programming Interface (API). The API includes access to a generic and extensible set of object classes to handle low-level annotation management details. Developers can leverage this framework and focus on application-level issues while creating multi modal sensor based applications. Generic GUI components are built on top of the data model components of the ATLAS API. These GUI components are organized in an object framework, making it easy for developers to create applications that incorporate them.

Third, a Meta-Annotation Infrastructure for ATLAS (MAIA) provides integrated type-definition support. MAIA has increased ATLAS usability by providing an easy, declarative way for users to create new annotation schemes. MAIA provides a simple, XML-based language allowing users to describe how to combine ATLAS core constructs to represent the complex annotations they wish to make, as well as how these annotations relate to one another.

4. WHAT METADATA SHOULD BE DERIVED FROM MEETING SENSORS?

From 1988 to the present, the NIST Speech Group developed a variety of standard data sets and metrics for speech recognition system development and test [2]. These data sets progressed from a small structured speech task, to read speech, to broadcast speech, to spontaneous telephone speech, and most recently to meeting room speech. Most of this work focused on the Word Error Rate (WER) metric and dealt with ASCII transcripts of the spoken words. While simple speech transcription can be viewed as first level of metadata, or direct annotation of acoustic data with words describing segments of the acoustic signal, they did not provide enough information for readily human readable transcripts, and will not provide enough for context aware recognition systems. Consultation with communities interested in recognition and metadata extraction suggested some linguistic and non-linguistic elements described in table 1 below, not all of which have been addressed to date.

Non-linguistic Metadata	Linguistic metadata
Speaker	Punctuation, capitals, formatting
Speaker Gender	Named entity and type
Multiple speakers	Utterance boundary and type
Background Music	Disruption points
Background Noise	Verbal edit interval
Time Tags	Filled pause
	Quotation
	Parenthetical/aside

Table 1 – Example metadata elements to be derived from acoustic sensor streams.

The Rich Transcription 2002 pilot evaluations were conducted using material from three domains: Broadcast News, Telephone Conversations, and Meeting Room data [3]. Since no publicly available training corpus for meeting recognition existed at the time, a small data set of similar size and properties to the evaluation test set was provided for training. For the pilot evaluation, we settled on word transcription and speaker segmentation. For RT-03 we are adding sentence-like units, and disfluency markups. This gives us the capability to produce rudimentary human-readable transcripts.

We are currently working toward creating a generic evaluation engine based on the ATLAS architecture that can be used for a variety of recognition, detection, and classification tasks relevant to spoken language domains, including meeting speech.

5. THE NIST MEETING ROOM

In addition to the SmartFlow and ATLAS software toolkits described above, we have constructed a data collection laboratory to instrument meetings. Figure 3 shows a schematic plan view of the layout and sensor arrangements in the room. It is a sensor rich environment, which provides many views of the meetings using twenty-four random placement microphones, three linear microphone arrays, five camera views, and an electronic white board. We are currently developing enhancements to this facility. One includes a Mk-III version microphone array, which will offer improved signal to noise ratio, and onboard conversion of the data streams to UDP/IP packets for direct transport across fast Ethernet. Also, an additional system of sensors for meeting participant locations using smart badges is being added.

7. THE NIST MEETING ROOM CORPUS

We have recorded a meeting room data set for use by industry and academic research and development groups. This consists of twenty hours of meeting data, at more than sixty gigabytes per hour aggregate data rate for all of the sensors. These meetings had various subjects including focus groups, game playing, expert interviews, and planning meetings. They varied in length from fifteen minutes to one hour, and had from three to eight participants. This data will be made available through the Linguistic Data Consortium.

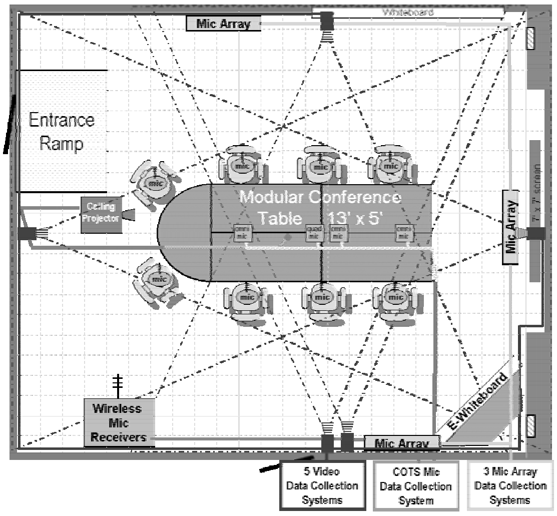


Figure 3 – Floor plan, sensor arrangements and data path for the NIST Meeting Room facility.

6. CONCLUSION

In order to aid our existing and prospective partners' industrial R&D programs in advanced multi modal computer interface and recognition technologies we have developed software and hardware tools necessary to create a meeting data corpus, and make this data available. The SmartFlow and ATLAS toolkits are in the public domain, and source code is available.

7. REFERENCES

[1] Proceedings of the 1998 DARPA/NIST Smart Spaces Workshop, July 1998, National Institute of Standards and Technology, pp. 3-1 to 3-14, available by request.

[2] Measurements in Support of Research Accomplishments, Pallett, D.S., Garofolo, J.S., Fiscus, J.G., Communications of the ACM, pp. 75-79, Vol. 43, No. 2, February 2000.

[3] Laprun, C., Fiscus, J., Garofolo, J., Pajot, S., A Practical Introduction to ATLAS, Proc. Language Resources and Evaluation Conference, Gran Canaria, Spain, May 29-31, 2002.

[4] Garofolo, J., Fiscus, J., Martin, A., Pallett, D., Przybocki, M., NIST Rich Transcription 2002 Evaluation: A Preview Proc. Language Resources and Evaluation Conference, Gran Canaria, Spain, May 29-31, 2002