MULTIMODAL SYSTEMS, RESOURCES AND EVALUATION

Mark T. Maybury
Information Technology Division
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA
maybury@mitre.org
www.mitre.org/resources/centers/it

Abstract

This paper considers multimodal systems, resources, and evaluation. We first motivate the value of multimodal information access with a vision of multimodal question answering and an example of content based access to broadcast news video. We next describe intelligent multimodal interfaces, define terminology, and summarize a range of applications, required corpora, and associated media. We then introduce a jointly created roadmap for multimodality and show an example of an open source multimodal spoken dialogue toolkit. We next describe requirements for and an abstract architecture of multimodal systems. We conclude discussing multimodal collaboration, multimodal instrumentation, and multilevel evaluation.

1. Multimodal Question Answering

A long range vision of ours is to create software that will support natural, multimodal information access. As implied by Figure 1, this suggests transforming the conventional information retrieval strategy of keyword-based document/web page retrieval into one in which multimodal questions spawn multimodal information discovery, multimodal extraction, and personalized multimodal presentation planning. In Figure 1 the user of the future is able to naturally employ a combination of spoken language, gesture, and perhaps even drawing or humming to articulate their information need which is satisfied using an appropriate coordinated integration of media and modalities, extracted from source media.

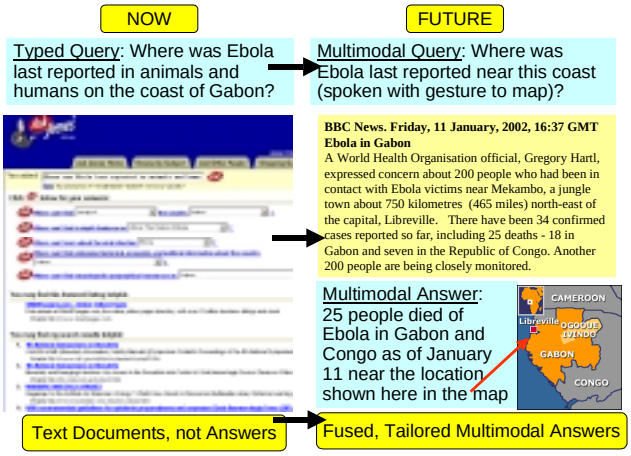

Figure 1. Ask Multimodal Questions, Get Multimodal Answers

The inadequacy of the current document retrieval strategy most closely associated with web search engines is underscored by Figure 2. Figure 2 illustrates that while (normalized) computing power doubles every 18 months and storage capacity doubles every 12 months, the fastest changing area of infrastructure is optical networking, where network speed is doubling every 8 months. Coupled with the rapid deployment of wireless devices and infrastructure, the ability to support mobile, multimodal access is becoming reality.

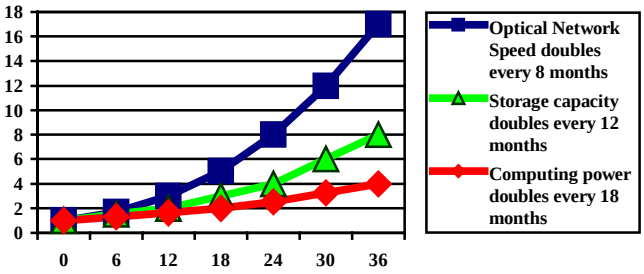

Figure 2. Acceleration of Infrastructure Growth

2. Broadcast News Access

As a step toward multimodal question answering, we have been exploring tools to help individuals access vast quantities of non-text multimedia (e.g., imagery, audio, video). Applications that promises on-demand access to multimedia information such as radio and broadcast news on a broad range of computing platforms (e.g. kiosk, mobile phone, PDA) offer new engineering challenges. Synergistic processing of speech, language and image/gesture promise both enhanced interaction at the interface and enhanced understanding of artifacts such as web, radio, and television sources (Maybury 2000). Coupled with user and discourse modeling, new services such as delivery of intelligent instruction and individually tailored personalcasts become possible.

Figure 3 illustrates one such system, the Broadcast News Navigator (BNN) (Merlino et al. 1997). The web-based BNN gives the user the ability to browse, query (using free text or named entities), and view stories or their multimedia summaries. For example,

Figure 3 displays all stories about the Russian nuclear submarine disaster from multiple North American broadcasts from 14-18 August 2000. This format is called a Story Skim. For each story, the user can view story details, including a closed caption text transcription, extracted named entities (i.e., people, places, organizations, time, and money), a generated multimedia summary, or the full original video.

Figure 3. Tailored Multimedia News

In empirical studies, Merlino and Maybury (1999) demonstrated (see Figure 4) that users enhanced their retrieval performance (a weighted combination of precision and recall) when utilizing BNN's Story Skim and Story Details presentations instead of mono-media presentations (e.g., text, key frames, video). In addition to performance enhancement, users reported increased satisfaction (8.2 on a scale of 1 (dislike) to 10 (like)) for mixed media display (e.g., story skim, story details).

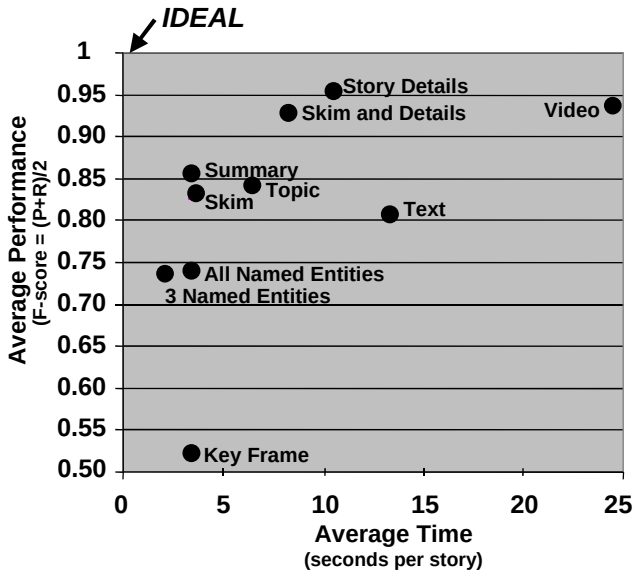

Figure 4. Relevancy Judgement Performance with Different Multimedia Displays

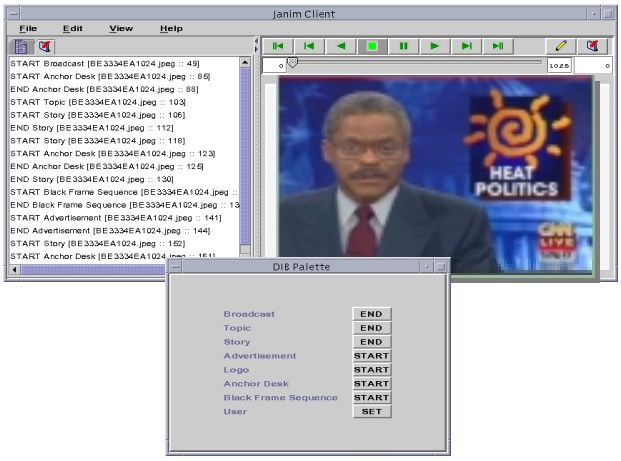

Figure 5. Video Annotation

As illustrated in Figure 5, during system development we utilized annotation tools to markup a corpus of video for features such as program start/stop as well as commercial and story segments. Using this gold standard, we can apply hidden Markov models to automatically learn a cross modal statistical model for video segmentation and transition detection. Learned models can then detect such video elements as the start of commercial or the transition from a desk anchor to a reporter in the field (Boykin and Merlino 2000). Rapid creation of this multimodal corpora is essential.

3. Multimodal Interfaces

Another vision is of intelligent multimodal interfaces[1] that support more sophisticated and natural input and output, enable users to perform complex tasks more quickly, with greater accuracy, and improve user satisfaction. Intelligent multimodal interfaces are becoming more important as users face increasing information overload, system complexity, and mobility as well as an increasing need for systems that are locally adaptive and tailorable to heterogeneous user populations. Intelligent multimodal interfaces are typically characterized by one or more of the following three functions (Maybury and Wahlster 1998, Maybury 1999):

Multimodal input – they process potentially ambiguous, impartial, or imprecise combinations of mixed input such as written text, spoken language, gestures (e.g., mouse, pen, dataglove) and gaze.

Multimodal output – they design coordinated presentations of, e.g., text, speech, graphics, and gestures, which may be presented via conventional displays or animated, life-like agents.

Interaction management – they support mixed initiative interactions that are context-dependent based on system models of the discourse, user, task and media.

[1] See www.mitre.org/resources/centers/it/maybury/iui99 for an on-line tutorial on intelligent interfaces.

This new class of interfaces promises *knowledge or agent-based multimodal dialogue*, in which the interface gracefully handles errors and interruptions, and dynamically adapts to the current context and situation. Exploiting explicit monitoring of user attention, intention, and task progress, an interface can explain why an action failed, predict a user's next action, warn a user of undesirable consequences of actions, or suggest possible alternative actions.

4. Media, Mode, and Code

In the above visions, it is useful to distinguish among media, mode, and code. A *mode* or *modality* is a human sense employed to process incoming information, e.g., vision, audition, olfaction, haptic (touch), and taste. In contrast, a *medium* is the material object (e.g., the physical carrier of information such as paper or CD-ROM) used for presenting or saving information and, particularly in the context of human computer interaction, computer input/output devices (e.g., microphone, speaker, screen, pointer). Finally, a *code* is a system of symbols (e.g., natural language, pictorial language, gestural language) used to represent and reason about media and modality.

5. Applications, Corpora, and Media

Table 1 illustrates a range of multimodal applications and associated corpora and media. What's different about these corpora from traditional linguistic corpora? Notably, the applications and associated multimodal corpora incorporate temporal and/or spatial dimensions Consider the following examples:

Multimodal question answering. The ability of users to articulate queries by typing, speaking, drawing, or singing and the ability to receive results in a range of integrated but heterogeneous media.

Broadcast Media. Automatically indexing television news or radio requires among other skills such as multimedia query, segmentation, extraction, and summarization.

Car-Driver Interactions (exemplifying intelligent multimodal interfaces). Recordings of the interactions of a driver in a car-driver scenario must also include instrumentation of the environment (e.g., car speed, location, time of day, temperature, other cars), as well as instrumenting the user as well as possibly other users.

Mobile Users. The interactions of a biking or walking individual, e.g., someone walking through a mall or supermarket and making purchases or someone enjoying a personalized museum tour.

Meeting transcription. Video tapings of human behavior that include not only (written or spoken) language discourse and visual events, but also capture the physical location of participants (in space but also in the video frames), changes in their properties over time (e.g., position to one another, attention, emotional state), and so on.

Multimodal authentication in which multiple biometric signatures of users (e.g., voice, face, eyes, gestures) are utilized to determine the identity of an individual in order to provide access control and behavior monitoring.

Each of these situations might imply audio, visual, and/or tactile modalities. Associated media have temporal extent and implied sequencing. They frequently contain information with spatial extent, coming in the form of user input, information accessed, or properties of the environment. For the user, spatial information can come from gaze or gestures (facial, hand, body) articulated by the user or system, the location (absolute or relative) of the user or the retrieved information or object (e.g. GPS coordinates of a car on a road) or simply a characteristic or property of the information retrieved (e.g., a map, blueprint, CAD/CAM diagram).

APPLICATION AREA	CORPORA (and models)	MEDIA
Multimodal question answering	Question and answer corpora	Text, speech, graphics, video
Intelligent multimodal interfaces	Human-machine interaction corpora	Text, speech, non-speech audio (e.g., sounds, music), gaze, video, gesture
Lifelike interface agents and/or Robotic interfaces	Interaction corpora (human physiology models)	Speech, gaze, gestures (facial, hand, body)
Meeting transcription (and human behavior analysis)	Human human communication corpora, meeting corpora	Video analysis of speech, gaze, gesture, drawings
Authentication	Multimodal biometric corpora	Text, speech, face, iris, gesture

Table 1. Applications, Corpora, and Media

Collection and annotation of multimedia corpora is challenging. Application requirements differ in needs, such as fidelity (e.g., degree of geoposition specificity), accuracy/error rate, and timeliness. There are no standard mark up languages much less common ontologies for such phenomena as time and location, although there are several ongoing international initiatives (Cunningham et al. 2000). Evaluation of these applications is also challenging for a number of reasons, not the least of which is they are often interactive and thus it is almost impossible to replicate exact human behavior across sessions.

6. Multimodal Roadmap

In an attempt to get a better handle on the future of this area, a recent Dagstuhl Workshop (Bunt, Maybury, and Wahlster 2001) drafted a baseline technology and capability roadmap for multimodality. As shown in Figure 6, the roadmap articulates three lanes to distinguish among developments in corpora analysis, advanced methods, and toolkits. These all lead to a medium term objective of creating mobile, human-centered and intelligent multimodal interfaces. For example, in the left lane labeled "empirical and data-driven models of multimodality", the group identified important steps toward multimodal corpora to include capturing examples of the value of multimodality, XML-encoding of human-human and human-machine multimodal corpora, analysis of frequency and complexity of phenomena, task specific corpora, and standards for multimodal annotation.

The center lane in the road labeled "advanced methods for multimodal communication" includes key milestones such as the creation of a common knowledge representation of multimodal content, task/situation/user aware multimodal interaction, models for effective multimodal human computer interaction, multiparty multimodal interaction, and multimodal barge in. The right lane focuses on activities to create "toolkits for multimodal systems" including markup languages for multimodal dialogue, reuseable components and a plug-and-play architecture, hybrid modules for input fusion, models of mutual disambiguation, and tools for universal access and mobile interaction. While the exact temporal location of each of these capabilities may be disputed, what is clear is the extensive research and development required to advance toward the vision.

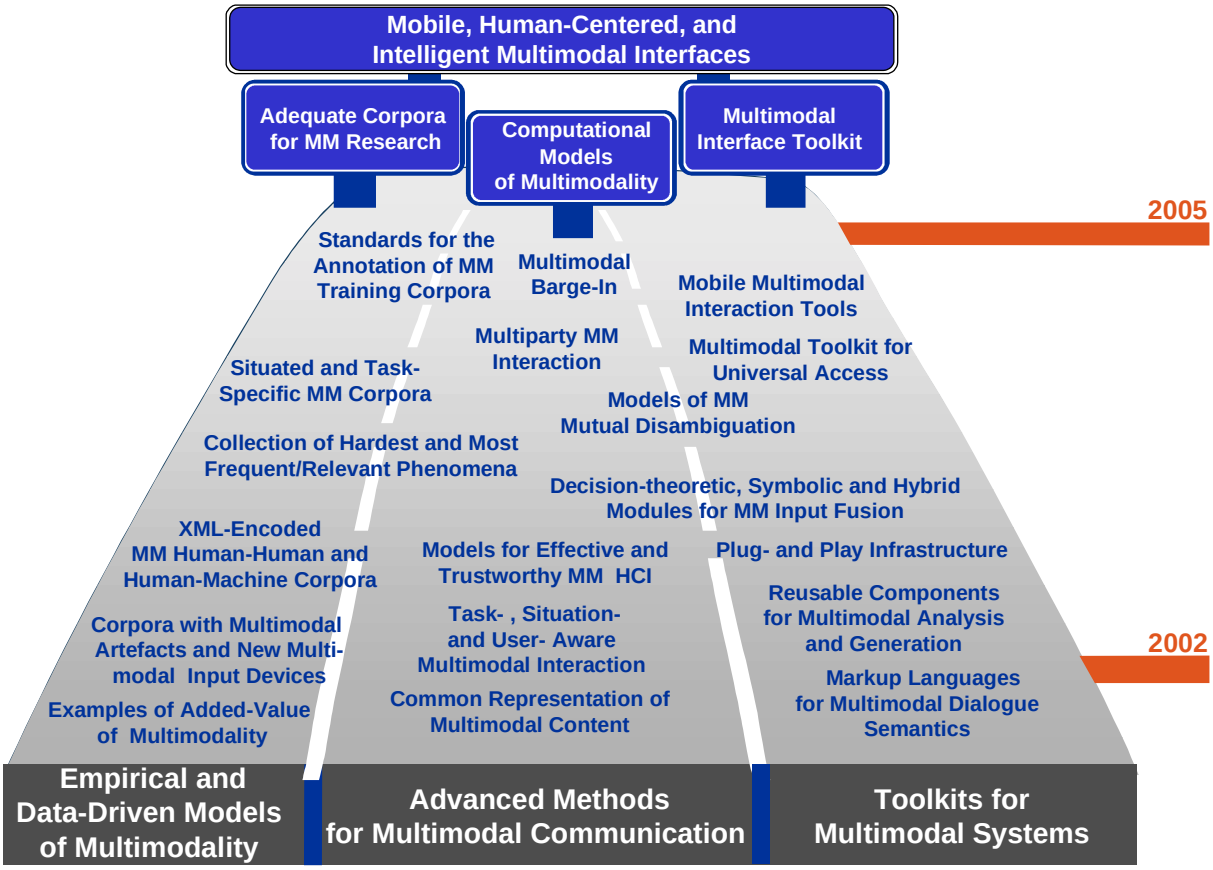

Figure 6. Multimodal Roadmap

7. Communicator

An example of an initiative to advance toolkits that might support multimodal interaction, the DARPA Communicator initiative aims to support natural conversational interaction to distributed on-line resources. This includes spoken access to web content, navigation, and summarization. Research foci of this initiative include dialogue management, multimodal input (speech, gesture), and output (synthesis, generation). One of the objectives of Communicator is to create a market and facilitate development via a component based, distributed architecture (see Figure 7) that is available via an open source repository (communicator.sourceforge.net).

The Communicator initiative is leveraging standards for plug and play and portability to new domains with the intent of lowering entry barrier to system development through componentware. For example, the JUPITER demonstration system created at the MIT LCS provides user with mobile access to weather information via a speaker independent phone interface (www.sls.lcs.mit.edu/sls/whatwedo/applications/jupiter.html.)

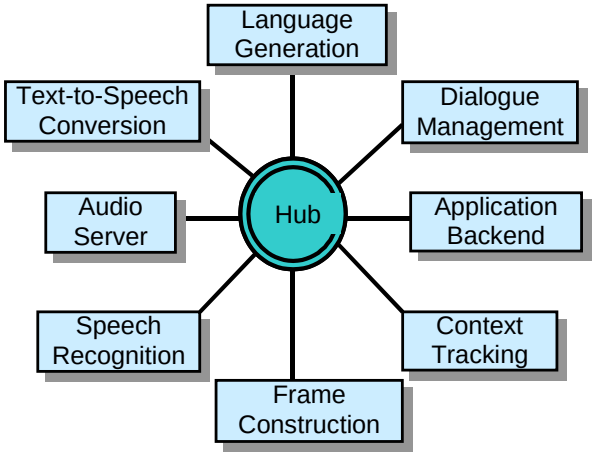

Figure 7. Galaxy Communicator Architecture

8. Multimodal Systems Requirements

In addition to the technology roadmap of Figure 6, participants at Dagstuhl analyzed approximately a dozen implemented multimodal interface systems. This included both historical and contemporary systems (e.g., "Put that there", CUBRICON, EMBASSI, SmartKom, DARPA Galaxy Communicator). Detailed systems analysis identified a number of essential functional and technical requirements. For example, SmartKom's (Wahlster 2001) MultiModal Markup Language (M3L) enables intermodule communication. Its word and gesture lattices support mutual disambiguation across partial processing results. Collectively, the identified requirements include the need to:

- support modality integration (both fusion of input and design of coordinated output)
- provide situation (user, task, application) appropriate real-time sensing/response (e.g., supporting barge-in, perceptual sensing/feedback)
- represent (modules and data structures) at varying levels of granularity
- manage feedback, both locally and globally
- support incremental processing
- support incremental development
- be scaleable

In addition these functional requirements, there are a number of important system/technical requirements these systems should exhibit, including:

- technical means for processing/fusing multimodal input (e.g., parallel processing)
- modular, composable elements and algorithms (possibly distributed processing)
- efficient algorithms and efficient implementations of those
- support for varying time scales, and temporal and spatial resolutions (as well as of course temporal resolution)
- shared (even after partial processing) data structures
- open and extensible protocols for interprocess and intermodule communication

While no single architecture has, or perhaps can, satisfy all of these requirements, we next describe a framework that captures the critical aspects of these systems but also serves as a more general description of this class of systems.

9. Abstract Multimodal Architecture

An abstract architecture can provide a common, community framework from which to understand, compare and contrast burgeoning multimodal systems. Motivated by the above requirements analysis, Dagstuhl participants formulated an abstract architecture for multimodal systems motivated by the architecture articulated in Maybury and Wahlster (1998). The result is the extended and refined architecture shown in Figure 8. As defined in the beginning of this article, the architecture utilizes the definitions of "media" as a material-centered notion including interactive devices (e.g., keyboard, mouse, microphone) and artifacts (audio, video, text, graphics), "mode" as human-centered perceptual processes (e.g., visual, auditory, tactile), and "code" as the formal languages that specific the elements, syntax, semantics, pragmatics and so on that govern the use of media and modes.

As can be seen in the figure, this abstract architecture includes functionality for media input processing and media output rendering as well as deeper media/mode analysis and synthesis, which would draw upon at least underlying models of media and modes (language, graphics, gesture). Following analysis, multimodal input is fused and then interpreted within the current state of the discourse, context (time, space, task, domain and so on) and user model including such functions as cross-modal mutual disambiguation. Once the intention of the user (in an interactive setting) is recognized, the system might interact with the backend application (possibly initiating or terminating sessions, requesting and integrating information or responding to application requests). Finally the system might plan a response to the user, which in turn might require the design of a multimodal presentation (including content selection, media design, allocation, coordination, layout) which would then need to be synthesized and rendered on specific media for the user.

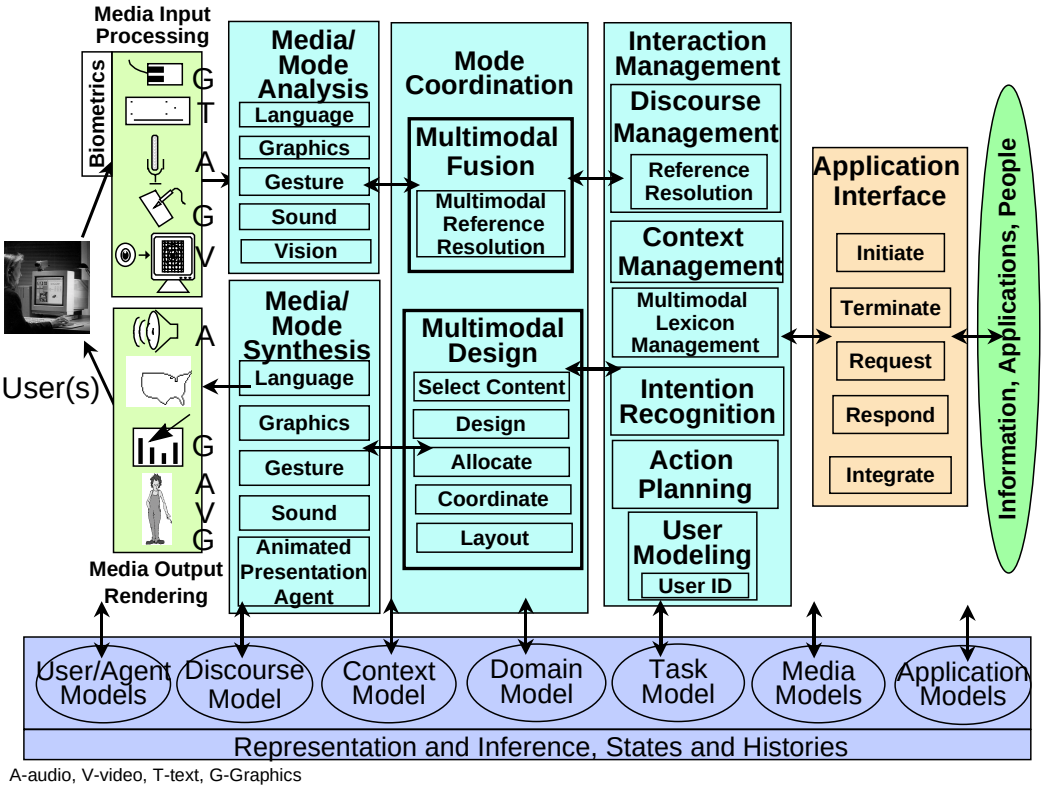

Figure 8. Abstract Multimodal Architecture

10. Multimodal Human-Human Interaction

Just as it is important to provide mechanisms for multimodal human machine interaction, so too it is important to enable multimodal human human interaction, augmenting current face-to-face interactions. Figure 9 graphically depicts the importance of team efforts and attempts to relate several levels of human collaboration, which build upon one another. Levels range from awareness of individuals, groups and activities, to sharing information with one another, to coordinating individual activities, to working jointly together, ultimately leading up to shared intent.

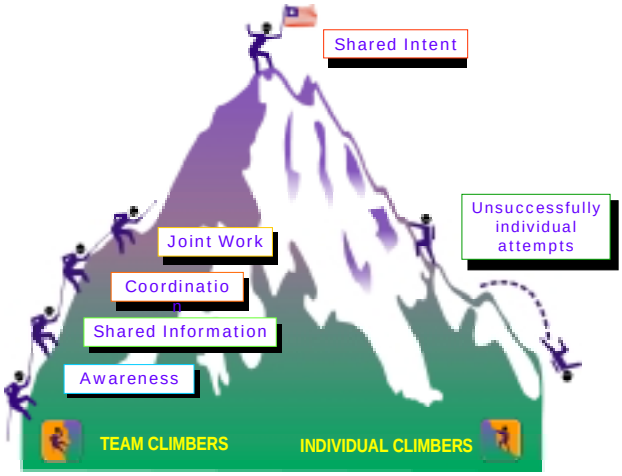

Figure 9. Levels of Collaboration

As detailed in Table 2, each of these levels of interaction implies different activities, classes of tools and associated media and modalities. For example, basic awareness of others, their communication capabilities (e.g., text, audio, video), availability, and perhaps even their activities is a fundamental prerequisite to collaboration. Tools such as electronic calenders, publish/subscribe mechanisms, presence information, and expertise finding tools can facilitate this awareness. Communication of awareness information typically occurs using text, graphics, and audio or visual alerts.

At the next level users can share information with one another at conferences, workshops, tutorials or just using personal communication in electronic mail, chat or video teleconference. Users can go beyond information sharing to coordination, the next level, which might involve creating shared assessments or shared plans in group brainstorming or decision meetings, possibly supported by decision support tools. Coordination might rely upon many media and modalities.

Joint work can occur face-to-face but can also be mediated by tools such as shared whiteboards or shared applications which can capture user preferences and application interactions. Workflow tools can facilitate sequencing and controlling interdependent efforts. Finally, building upon all of the underlying levels, the establishment of shared intent in a relationship typically grows over many, often face-to-face, interactions.

LEVEL	Activities	Tools	Media
Shared Intent	• Shared Purpose • Co-dependent	• Strategic Alliances	• Face-to-face
Joint Work	• Shared goals • Joint goal creation • Cross-organizational teams	• Workflow • Whiteboard • Shared applications	• Application actions • Gesture • Text • Audio • Video
Co-ordination	• Shared plans • Group meetings	• Decision Support • Brainstorming tools	• Text • Audio • Video
Shared Information	• Meetings • Conferences • Briefings and presentations • Training	• E-mail, chat, VTC • Web pages, Portals • Publications	• Text • Messages • Audio • Video
Awareness	• Shared calendars • Shared presence	• Electronic calendars • Publish/subscribe • Alerts • Presence • Expert finding	• Text • Graphics • Audio • Video

Table 2. Collaboration Levels, Tools, and Media

For a number of years we have been exploring human human group collaborations within distributed, virtual environments. Our work has resulted in the open source software (cvw.sourceforge.net), Collaborative Virtual Workplace (CVW), a screenshot of which is shown in Figure 10. CVW incorporates a comprehensive suite of tools that support many of the tasks outlined in Table 2, including shared whiteboarding, audio/video/text conferencing, user presence awareness, access control, and persistent virtual spaces (i.e., virtual rooms which contain applications, documents, and users).

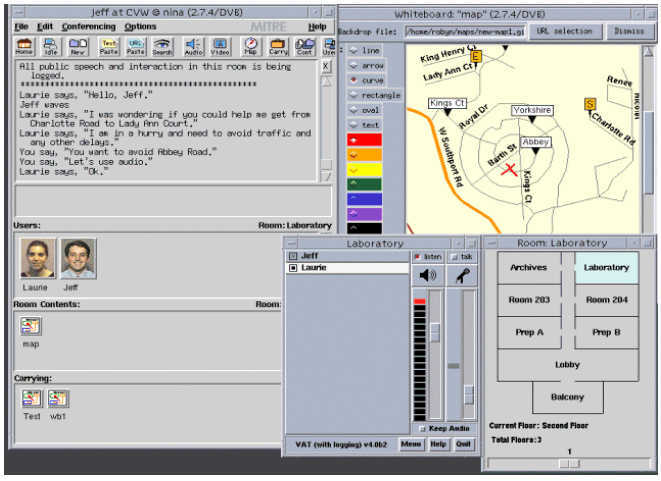

Figure 10. Collaborative Virtual Workplace

Maybury (2001) describes the functionality and operational use of this place-based environment by hundreds and thousands of users in two major organizational settings for analysis and planning. In order to understanding the operational impact and evaluate the effectiveness of these tools, as well as to understand technical infrastructure issues, we have found it essential to instrument user activities within these virtual environments. We have used MITRE's multimodal logger to accomplish this, which we describe next.

11. Multimodal Logging and Evaluation

MITRE's Multimodal logger (Bayer et al. 1999) supports the recording, retrieval, annotation and visualization of data collected in human-computer and human-human interactions. The Multimodal logger incorporates a database structure which groups datapoints by application (e.g., audio utterance, text chat, whiteboard use, video conference) and applications by session. It supports the typing of data points via MIME types, provides an easy-to-use API for instrumenting existing applications and tools for reviewing and annotating data collected via instrumentation.

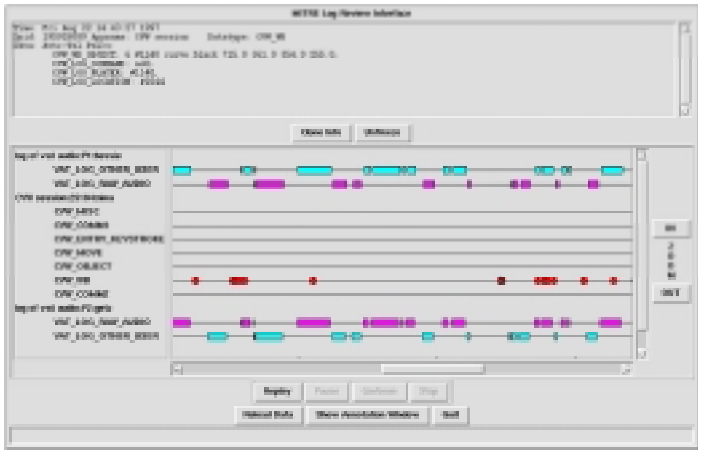

Figure 11. Multimodal Logging and Annotation

Figure 11 illustrates the visualization of multimedia events across a range of applications such as whiteboarding (CVW_WB), start, end and duration of events in audio conferencing (VAT), movements among virtual rooms (CVW_MOVE) and object manipulation (CVW_OBJECT). The user can zoom in or out to inspect specific events as well as add further annotations to this automatically constructed event log. This supports analyses, for example, of multiparty communication to look at properties such as frequency of user communications and actions, discourse events such as interruptions, and cross modal events such as co-occurring speech and gestures.

DARPA's Intelligent Collaboration and Visualization initiative (zing.ncsl.nist.gov/nist-icv) utilized MITRE's multimodal logger in support of collaboration system evaluation. Working initially with NIST, NIMA and CMU, MITRE developed an assessment methodology for collaboration systems (Damianos et al. 2000) that includes a framework of four levels of abstraction as illustrated in Figure 12. The requirements level captures the work and transition tasks to be performed, and the social protocols and characteristics of the group performing the tasks; the next level specifies the

capabilities (e.g., shared workspace, communications, etc.) required to perform the work; the services level describes specific services (e.g., text chat, whiteboard) that could be used to deliver the capabilities, and the technology level describes specific implementations of services. Associated with each level are appropriate assessment metrics. Assessments can be made at multiple levels of this framework, depending on the intended needs of the evaluators, whether they are users, researchers, or systems designers. Community defined multimodal evaluations are essential for progress, and that the key to such progress is a shared infrastructure of benchmark tasks, evaluation tools, and training and test sets to support cross-site performance comparisons.

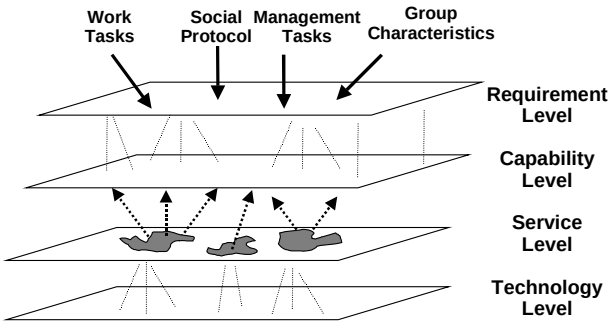

Figure 12. Multilevel Multimodal Evaluation

12. Conclusion

This article has considered multimedia and multimodal applications, resources, and annotation tools. We envisioned multimodal research areas and presented a multimodality roadmap and abstract architecture. We described two open source contributions: a multimodal spoken dialogue toolkit and a multimodal logger for instrumenting applications and users. Finally, we sketched a multimodal evaluation methodology. Important areas for future work include annotation standards, richer corpora, and community evaluation.

13. References

Bayer, S. 2001. DARPA Communicator. Open source and documentation available at fofoca.mitre.org and www.sourceforge.net/projects/communicator.

Bayer, S., Damianos, L., Kozierok, R., Mokwa, L. 1999. The MITRE Multi-Modal Logger: Its Use in Evaluation of Collaborative Systems. ACM Computing Surveys, March 1999. www.mitre.org/technology/logger.

Boykin, S. and Merlino, M. Feb. 2000. Machine Learning of Event Segmentation for News on Demand. *Communications of the ACM*. Vol 43(2): 35-41.

Bunt, H., Maybury, M. and Wahlster, W. Dagstuhl Seminar on Coordination and Fusion in Multimodal Interaction. Oct. 28-Nov. 2, 2001. www.dfki.de/~wahlster/Dagstuhl_Multi_Modality

Cunningham, H., Roy, D. and Wittenburg, P. 2000. First EAGLES/ISLE Workshop on Meta-Descriptions and Annotation Schemes for Multimodal/Multimedia Language Resources and Data Architectures and Software Support for Large Corpora. LREC 2000, Athens, Greece, 29/30 May.

Damianos, L., Drury, J., Fanderclai, T., Hirschman, L., Kurtz J., and Oshika, B. 2000. Evaluating Multi-party Multi-modal Systems. In Proceedings of LREC 2000, vol. III, pages 1361-1368.

Drury, J., Damianos, L, Fanderclai, T., Kurtz, J., Hirschman, L. and Linton, F. 1999. Methodology for Evaluation of Collaboration Systems. zing.ncsl.nist.gov/nist-cv/documents/methodv4.htm

Merlino, A., Morey, D. and Maybury, M. 1997. "Broadcast News Navigation using Story Segments", ACM International Multimedia Conference, Seattle, WA, November 8-14, 381-391. www.acm.org/sigmm/MM97/papers/morey

Merlino, A. and Maybury, M. 1999. An Empirical Study of the Optimal Presentation of Multimedia Summaries of Broadcast News. Mani, I. and Maybury, M. (eds.) *Automated Text Summarization*, MIT Press.

Maybury, M. T. 1999. Multimedia Interaction for the New Millennium. Proceedings of 6th European Conference on Speech Communication and Technology (Eurospeech '99). Budapest, Hungary. September 6-9, 1999. vol 1. p. KN 15-19.

Maybury, M. February 2000. News on demand: Introduction. *Communications of the ACM*. Vol 43(2): 32-34.

Maybury, M. December 2001. Collaborative Virtual Environments for Analysis and Decision Support. *Communications of the ACM* 14(12): 51-54. www.acm.org/cacm/1201/1201toc.html.

Maybury, M. T. and Wahlster, W. editors. 1998. *Readings in Intelligent User Interfaces*. Morgan Kaufmann Press.

Wahlster, W. 2001. Smartkom: Dialog-based Human-Computer Interaction by Coordinated Analysis and Generation of Multiple Modalities. In International Status Conference on HCI October 26, 2001, Saarbrücken, Germany. www.smartkom.de

14. Acknowledgements

I thank Jean-Claude Martin, Wolfgang Wahlster, Harry Bunt, and the Dagstuhl workshop participants for their contributions to the creation of the roadmap and abstract architecture. Appreciation for leadership of MITRE activities described herein goes to Laurie Damianos, Bea Oshika, Sam Bayer, Lynette Hirschman, Stanley Boykin, Warren Greif, Anita King, Rod Holland, Jay Carlson and Michael Krutsch.