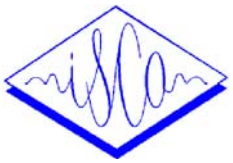

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

MULTIMEDIA INTERACTION FOR THE NEW MILLENNIUM

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

ABSTRACT

Spoken language processing has created value in multiple application areas such as document transcription, data base entry, and command and control. Recently scientists have been focusing on a new class of application that promises on-demand access to multimedia information such as radio and broadcast news. In separate research, augmenting traditional graphical interfaces with additional modalities of interaction, such as spoken language, gesture, or eye tracking, promises to enhance human computer interaction. In this address I discuss the synergy of speech, language and image processing, introduce a new idea for corpus based multimedia interfaces, and identify some remaining challenging research areas.

MULTIMEDIA INFORMATION ON DEMAND

Information on demand, the ability to provide information tailored to specific user needs, promises new capabilities for research, education and training, and electronic commerce (e.g., on-line information access, question answering, and customer service). Whereas significant commercial activity has focused on providing access to documents, web pages, and structured data sources, less attention has been given to multimedia information on demand. To achieve effective multimedia information on demand, however, requires a confluence of capabilities from several fields including image, speech and language processing, information retrieval, information extraction, translation, summarization, and presentation design.

Over the past several years scientists have been exploring a range of systems to provide tailored, content-based access to multimedia including text, imagery, audio, and video [19]. For example, by synergistically combining techniques for processing speech, language, and imagery, MITRE developed a sophisticated news understanding system, the Broadcast News Navigator (BNN) [21]. 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 (Figure 1 displays all stories about Diana on CNN Prime News during August-September 1997). For each story, the user is given the ability to view its closed caption text, named entities (i.e., people, places, organizations, time, money), a generated multimedia summary, or view the full original video of a story. The user can also graph trends of named entities in the news for given sources and time periods. For example, Figure 2 graphs the onset and abatement of stories on Princess Diana and Mother Teresa, 8-15 September 1997.

Figure 1. Tailored Multimedia News

Analyzing audio, video, and text streams from digitized video, BNN segments stories, extracts named entities, summarizes stories, and designs presentations to provide the end user with content-based, personalized web access to the news [21]. For example, within the video stream color histograms are used to classify frames and detect scene changes. In the audio stream, algorithms detect silence, speaker changes, and transcribe the spoken language. Finally, the closed caption stream and/or speech transcription is processed to extract named entities. This fully automated broadcast news system stands in contrast to the current method of manual transcription and summarization of broadcast news (e.g., via closed captioning services) which is expensive, error prone, and can result in dissemination delays. BNN has been integrated into a larger system called GeoNODE [12], which correlates named entities across stories to create story clusters and then partitions

these clusters from a constructed hypergraph to automatically identify topics.

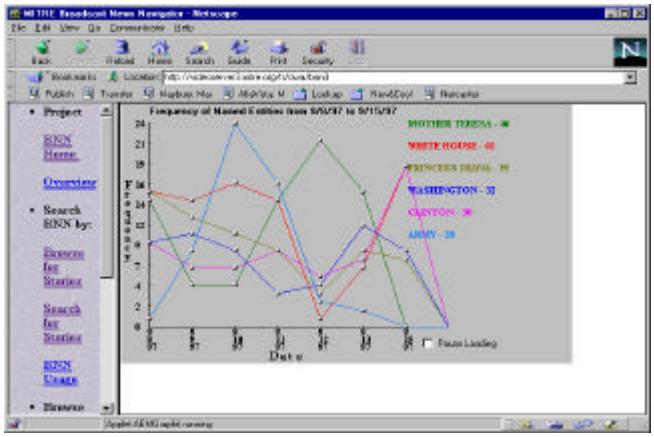

Figure 2. Temporal Visualization of Named Entities

Our analyses show that in key tasks such as segmenting stories, audio and imagery processing can enhance algorithms which are based only on linguistic cues (e.g., explicitly stated anchor welcomes, anchor to reporter handoffs, story introductions). For example, silence detection, speaker change detection, and key frame detection (e.g., black frames, logos) can improve the performance of text-only story segmentors. By modeling story transitions using hidden Markov models and learning the most effective combination of cross media cues [4], successive versions of the system have incrementally increased performance. As illustrated in Figure 3, the system's version 2.0 performance over a range of broadcast sources (e.g., CNN, MS-NBC, and ABC) averaging across all cues is 38% precision and 42% recall. In contrast, performance for the best combination of multimodal cues rises to 53% precision and 78% recall. When visual anchor booth recognition cues are specialized to a specific source (e.g., ITN broadcasts that have more regular visual story change indicators) the performance rises to 96% precision and recall.

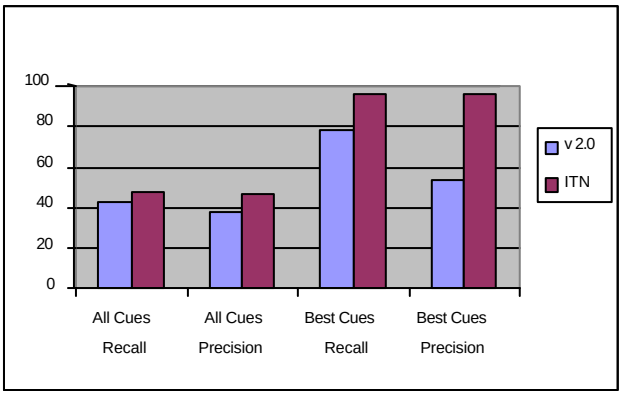

Figure 3. BNN Segmentation Performance by Version

Given properly delineated story boundaries, BNN is able to summarize the story in a variety of fashions. This includes extracting [1] the most significant named entities, extracting visual elements (e.g., keyframes) and/or summarizing the story text and creating from these elements a multimedia summary. We have integrated Carnegie Mellon University's SPHINX-II speech system into BNN and have begun experiments in extracting named entities from transcribed stories. For example, Figure 4 shows a manual transcript of a segment of a story with human markup of locations (bold italic) and organizations (bold underlined). In contrast, Figure 5 shows an automated transcription of a news segment followed by automated markup of locations (bold italic) and organizations (bold underlined). Notice the errors in the automated transcript of omission of punctuation and content (e.g., "A", "AND") and substitutions (e.g., "EVELYN" for "CONSIDERING", "TRIAL" for "PANEL"). Also note errors in the named entity identification in Figure 5 ("BLACK" is identified as a location, "UNITED STATES" and "CONGRESSIONAL BLACK CAUCUS" are missed). In the DARPA HUB-4 information extraction evaluation, the version of the system described in [26] achieved 71-81% accuracy on broadcast news speech data with a wide range of word error rates (WERs) ranging from 13% to 28%. The model achieved 88% accuracy on reference transcriptions with WER 0% (like Figure 4).

WHITE HOUSE OFFICIALS SAY THE PRESIDENT IS CONSIDERING TWO STEPS: A NEW PRESIDENTIAL PANEL MODELED ON THE 1968 **KERNER COMMISSION**, WHICH CONCLUDED THERE WERE TWO SOCIETIES IN THE ***UNITED STATES***, ONE BLACK, ONE WHITE SEPARATE AND UNEQUAL.

A **WHITE HOUSE** SPONSORED CONFERENCE THAT WOULD INCLUDE CIVIL RIGHTS ACTIVISTS, EXPERTS, POLITICAL LEADERS AND OTHERS THE NEW HEAD OF THE **CONGRESSIONAL BLACK CAUCUS** SAYS

Figure 4. Manual Transcription & Extraction

WHITE HOUSE OFFICIALS SAY THE PRESIDENT IS EVELYN TWO STEPS WHAT NEW PRESIDENTIAL TRAIL OF ALL ON THE NINETEEN SIXTY EIGHT **KERNER COMMISSION** WHICH CONCLUDED THE OUR TWO SOCIETIES IN THE UNITED STATES ONE ***BLACK*** ONE WHITE SEPARATE AND UNEQUAL

WHITE HOUSE SPONSORED CONFERENCE THAT WILL INCLUDE CIVIL RIGHTS ACTIVISTS EXPERTS POLITICAL LEADERS AND OTHERS AND THE NEW HEAD OF THE CONGRESSIONAL BLACK CAUCUS OF

Figure 5. Automated Transcription & Extraction

Even when dealing with closed-captioned text, we face a 10-15% word error rate because of errors introduced during manual transcription. The word error rates for the best automated speech transcription systems (depending upon processing speed) range widely from 13-28% on

studio quality speech (e.g., anchor segments) to 40% or higher on shots with degraded audio (e.g., reporters in the field, speakers calling in over the phone, music in the background). Furthermore, neither closed captions nor speech transcripts have case information to use, for example, to recognize proper nouns. In addition, speech transcripts contain no punctuation and can contain disfluencies (e.g., hesitations, false starts), which further deteriorates performance. However, it is important to point out that the type of errors introduced during human transcription are distinct from those made by automated transcription, which can have different consequences for subsequent named entity processing. In the HUB-4 evaluations [8], named entity extraction on clean, caseless, and punctuationless manual transcripts was approximately 90% in contrast to the best extraction performance on newswire with case, which was approximately 94%.

Significantly, MITRE's named entity extraction system, called Alembic[1] [1], consists of rule sequences that are automatically induced from an annotated corpus using error-based transformational learning. This significantly enhances cross domain portability and system build time (from years or months to weeks) and also results in more perspicuous rule sets. The Alembic group is presently pursing the induction of relational structure with the intent of automated event template filling.

USER STUDIES

We have performed a number of user studies [23] with BNN and discovered users can perform their information seeking tasks both faster and more accurately by using optimal mixes of news story summaries. For example, using the BNN display shown in Figure 1 (a key video frame plus the top three entities identified per segmented story which we term a "skim" in Figure 6), users can find relevant stories with the same rate of precision but in one sixth the time required to retrieve (unsegmented) indexed video using a digital VCR. If we add in a little more information to the display (e.g., an automatically generated one-line summary of each story and a list of extracted topics), the user's retrieval rate slows to only one-third the time of the indexed video (called "Full" in Figure 6). However, recall and precision are as good as if the user watched the full video. Results for this identification task are summarized in Figure 6 with the best performance for both precision and recall resulting from "full" or "story details", displays integrating multiple media elements (e.g., key frame, summary, top named entities, access to video and closed caption).

[1] The Alembic Workbench, which enables graphical corpus annotation, can be downloaded from http://www.mitre.org/ resources/centers/it/g063/nl-index.html.

In addition to this story retrieval task, we measured performance in question answering (called the comprehension task). As expected, on average answering comprehension questions took 50% longer to perform (~4 seconds per question/story) than story identification (~2 normalized seconds per story). The performance of users on comprehension tasks using presentations such as that shown in Figure 1 were not as beneficial. We attribute this to multiple factors, including the fact that summaries may be less valuable for comprehension tasks but also because of limits in the state of the art in information extraction. Nevertheless, on a satisfaction scale of 1 to10 (1 dislike, 10 like), for both story retrieval and question answering tasks, users prefer mixed media displays like that in Figure 1 about twice as much (7.8 average rating for retrieval, 8.2 for comprehension) over other displays including text transcripts, video source, summaries, named entities or topic lists (average ratings of the 5.2 and 4.0 for retrieval and comprehension tasks, respectively).

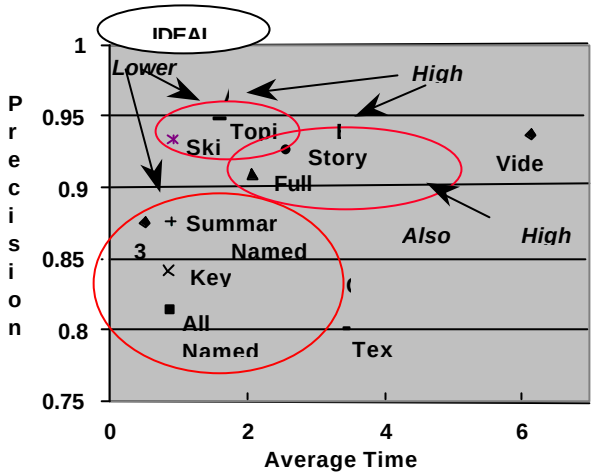

Figure 6. Avg. Precision vs. Time for Story Identification

Interestingly, giving users less information (e.g., only the three most frequently occurring named entities in a story) reduces task time without significantly affecting performance. This broadcast news understanding experimentation illustrates how synergistic processing of multimodal streams to segment, extract, summarize, and tailor content can enhance information seeking task performance and user satisfaction. It is also the case that the user's task performance can be enhanced by increasing intelligence in the interaction between user and system, which we consider next.

CORPUS-BASED MULTIMEDIA INTERFACES

In the future, it would be desirable to have interfaces and evaluation methodologies that would both facilitate the kind of user study just described and learn from feedback from usage to improve performance. It is noteworthy that corpus-based methods have resulted in regular and

predictable improvements in individual processing components of intelligent interfaces such as speech, language, and image processors. For example, community wide evaluations in speech and language processing have achieved considerable success by using standard training and test sets, establishing common tasks, and enabling multi-site comparison of performance. This methodology has been successful in several application areas including information retrieval [10], information extraction [11], and topic tracking and detection [2]. We have hypothesized that such success can be transferred more generally to human computer interactions [20]. That is, just as we can learn models that can label spoken language elements (e.g., phones, vocabulary, pronunciation, and phrases) and learn models to label language elements (e.g., part-of-speech, phrase structure, named entities, and events), so too we should be able to induce models of elements of interaction (e.g., reference, topic transitions, and turn-taking). In addition, we should be able to learn models of the most effective interface elements such as defaults or layouts for widgets (e.g., menu order or window layout [25]), hypertext structure [5] or preferred multimedia presentation designs [16, 18].

Some architectural efforts aim to create open, standard interface services that can enable plug and play of components and facilitate interface customization. For example, the DARPA Communicator program (http://fofoca.mitre.org/) aims to provide the next generation of intelligent conversational interfaces [29] to distributed information. The goal is to support the creation of speech-enabled interfaces that scale gracefully across modalities, from speech-only to interfaces that include graphics, maps, pointing and gesture. Its primary objectives include the desire (1) to increase modularity to allow plug-and-play interoperability of components to enhance system conversational capabilities, (2) to provide intelligent interaction and appropriate easy-to-understand output to facilitate creation of mobile, multi-modal conversational interfaces, and (3) to allow application developers to build interfaces to new applications. These kinds of shared architectures, together with community wide evaluation methods and metrics, can play a central role in the advancement of spoken dialogue science and technology. The challenge is to expand corpus-based processing to this class of systems.

The success of a corpus-based strategy is predicated upon successful corpora creation (which can be expensive and so should be shared), standard annotations together with standard evaluation measures, metrics, and methods, and an application and user-driven focus on evolutionary fielding. With respect to human computer interaction, instrumentation is key to effective evaluation. Evaluation can serve a range of purposes including benchmarking, comparison (of system to own, modified, or human performance), hypothesis testing, confirmation of experimental results, and discovery of strengths and weaknesses. Traditionally, the primary means for such evaluation are usability methods such as inspection (e.g., [14]), task analysis to predict performance (e.g., [6]), or user studies (e.g., Wizard-of-Oz, usability lab evaluation).

A corpus or data-driven approach may require instrumentation of interface widgets and new evaluation methods. The instrumentation of interface and application elements itself can have tremendous value. Logs of application usage are rich sources for developers for refining engineered solutions and making scientific and technical discoveries. Cuomo [7] points out that there is useful information to be discovered in interaction data logs using techniques such as hidden Markov models or sequence analysis to analyze error states, discover bottlenecks, or detect repeating sequences which could be automated. Linton et al. [15] tracked 16 Microsoft Word users and found only 45% of the application's features were ever used. Moreover, of the 642 commands available, just 20 commands accounted for 90% of use. The average person used only 57 commands in six months, and taken together, all of the users used only 152 commands in 18 months. This knowledge can help identify what users don't know about or don't need functionally in modern software and can guide decisions regarding on-line training or evolution of software. In addition, collaborative filtering of user activity can help identify and connect up users that have unique or complementary patterns of usage that may reflect competencies or development needs to address.

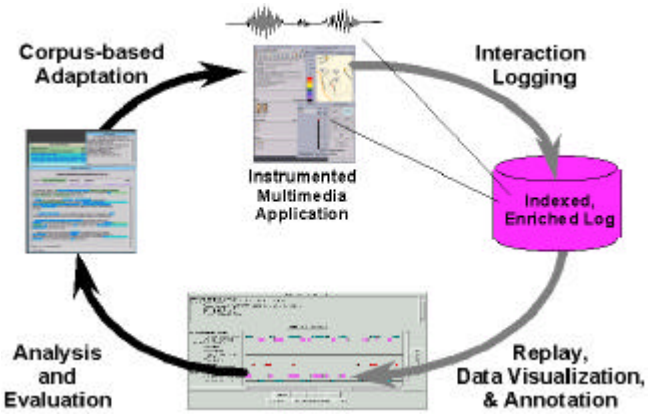

Figure 7. Evaluation Methodology

Figure 7 illustrates the steps in a corpus-based approach to multimodal interfaces that MITRE researchers are exploring. These researchers have developed logging tools for capturing and visualizing multimodal interactions (http://www.mitre.org/research/logger), evaluation methods [3], and are beginning to experiment testing hypotheses using such an environment. This enables rich possibilities for experimental design. For

example, one could readily establish an instrumented, multisite, multimodal collaboration environment and then perform ablation studies and evaluate task and user success using the instrumented environment as an experimental platform. An additional challenge is integrating multimodal (e.g., speech, graphical, gesture) interactions to detect error states using clues such as command repetition, help invocation, and clarification subdialogues in one media and repetitions and disfluencies in another such as speech.

INTELLIGENT MULTIMEDIA INTERFACES

Empirical methods, such as those outlined above, promise a new approach for moving toward a vision of intelligent mulitmodal interaction. Intelligent multimedia interfaces[2] support more sophisticated and natural input and output and enable users to perform potentially complex tasks more quickly, with greater accuracy, and with improved satisfaction. These systems are typically characterized by one or more of the following properties [15, 16]:

1. *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.
2. *Multimodal output* – they design coordinated presentations of, e.g., text, speech, graphics, and gestures.
3. *Knowledge or agent-based dialogue* – mixed initiative interactions that are context-dependent based on system models of the discourse, user, and task.

In addition to supporting a much richer range of interaction styles, these interfaces enable the user to do things they perhaps could not otherwise. For example, by explicitly monitoring user attention, intention, and task progress, an interface can explain why an action failed, predict a users next action, warn a user of undesirable consequences of actions, or suggest possible alternative actions. These systems, however, tend to be heavily knowledge based which is both a systems engineering challenge as well as a machine learning opportunity.

[2] An on-line tutorial of intelligent user interfaces is available at http://www.mitre.org/resources/centers/it/maybury/iui99/index.htm

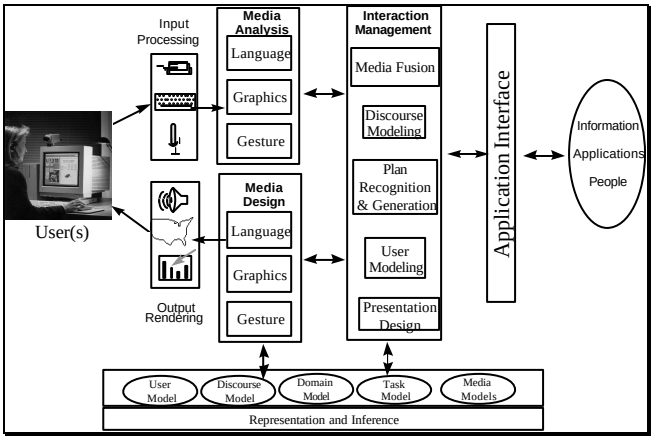

Figure 8. Intelligent Multimedia Interfaces

Figure 8 illustrates a high level architecture of intelligent multimedia user interfaces [22]. We first distinguish between *mode* or *modality* which refers primarily to the human senses employed to process incoming information, e.g., vision, audition, olfaction, haptic (touch), and taste. In contrast, *medium* refers to 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). We use the term *code* to refer to a system of symbols (e.g., natural language, pictorial language, gestural language). For example, a natural language code might use typed/written text or speech, which in turn would rely upon visual or auditory modalities and associated media (e.g., keyboard, microphone). While these distinctions are important, we frequently use the term multimedia to refer to this general area of processing.

These interfaces perform a number of intelligent functions including analyzing and interpreting input, designing and rendering output, managing the interaction, and representing and reasoning about models that support intelligent interaction. An example of a model is a user model, more generally, an agent model (e.g., which could represent the user, system, intermediary, addressee, etc.) The "intelligence" in these systems that distinguishes them from traditional interfaces is indicated in bold in Figure 8 and includes mechanisms that perform automated media analysis, design, and interaction management. For example, these systems can use gesture or gaze input to resolve ambiguous linguistic input and vice versa [13]. For example, systems have been implemented in several domains (e.g., direction giving [18], electronic commerce [28]) to manage interactions by reasoning about communicative acts such as (spoken or typed) linguistic, graphical, display control, dialogue and physical actions. Some of these actions are cross media, such as attentional acts, which can be realized as gesture, speech, or written language.

The architecture in Figure 8 has been adopted and extended in the context of the SmartKom initiative [28], an academic and industrial consortium which plans to create intuitive multimodal interactive devices for public kiosks, home, and mobile applications. Extensions include supporting biometrics for security for input and animated presentation agents for output.

In the context of the class of systems shown in Figure 8, spoken language solutions will need to operate with multiple interactive devices, work under a range of operating conditions (e.g., office, outside, on the move), operate in a broad set of information domains, and support real-time, tailored access to information, tools, and people. Accessing information will increasingly need to take into account multimedia (e.g., text, radio, video) and multilingual content and solutions will need to scale to massive data sets which will be heterogeneous, noisy and incomplete. Finally, testing needs to be comprehensive and include multidimensional measures such as speed, accuracy, fault tolerance/graceful degradation, and user enjoyment.

RESEARCH OPPORTUNITIES

There are many exciting areas for new science with the advent of intelligent multimedia systems, a few of which we outline here.

Multimedia Input Analysis. Many research challenges remain in areas such as individual and inter-media segmentation, ill-formed and partial input parsing and interpretation, and ambiguous and partial multimedia reference resolution. New interactive devices (e.g., force, olfactory, and facial expression detectors/generators) will need to be developed and tested and will provide new possibilities, such as emotional state detection and tracking. Techniques for media integration and aggregation need to be further refined to ensure synergistic coupling among multiple media, managing input that is impartial, asynchronous or varies in level of abstraction. Algorithms developed for multimedia input analysis have proven beneficial for multimedia information access [19]. The crossover of algorithms from input analysis to artifact processing and vice versa will continue to be an area of further research opportunity.

Multimedia Output Generation. Important questions remain regarding methods for effective content selection (choosing what to say), media allocation (choosing which media to say what in, such as choosing among language, non-speech audio or gesture to direct attention), and modality selection (e.g., realizing language as visual text or aural speech). In addition, further investigation remains to be done in media realization (i.e., choosing how to say items in a particular media), media coordination (cross modal references, synchronicity), and media layout (size and position of information) [15].

Agent Interfaces. Anthropomorphic interface agents are found increasingly in learning environments, games, and customer service applications. They can mitigate complex tasks, bring expertise to the user, and provide more natural (anthropomorphic) interaction. For example, they might be able to adapt sessions to a user, deal with dialogue interruptions or follow-up questions, and help manage focus of attention. Agents raise important technical and social questions but equally provide opportunities for research in representing, reasoning about and realizing agent belief and attitudes (including emotions). Creating believable behaviors and supporting listening, speaking and gesturing agent displays [24] are important user interface requirements. Research issues include what can and should an agent do, how and when should they do it (e.g., implicit versus explicit tasking, activity, and reporting) and by what means should they carry out communications (e.g., text, audio, video). Other important issues include how do we instruct agents to change their future behavior and who is responsible when things go wrong.

Multimedia Collaboration. Discovering more effective means of human-human computer-mediated interaction is increasingly important as our world becomes more wired. The multimedia application in Figure 7 is in fact a multimodal collaboration environment including text, audio, and video conferencing, shared applications (documents and whiteboards), and persistent stores of information and tools using a room-based paradigm. In these virtual place-based environments, many questions remain including: How do people find one another? How does an individual discover meetings/collaborations? What are the most effective multimedia interfaces in these environments for different purposes, individuals, and groups? Multimodal processing has many potential roles in these environments ranging from transcribing and summarizing meetings to correlating voices, names, and faces, to tracking individual and group attention and intention across media. For example, just as we are beginning to model richer models of communicative acts (e.g., speech acts, discourse acts, media acts) [18], so too we should aim to identify collaboration acts in which multiparty behaviors occur such as introducing participants, motivating participation, clarifying positions, or, in general, facilitating interaction. Careful and clever instrumentation and evaluation of collaboration environments [3, 9] will be key to learning more about just how people collaborate.

Machine Learning. Machine learning of algorithms using multimedia corpora promises portability across users, domains, and environments. There remain many research opportunities in machine learning applied to multimedia interaction such as on-line learning from one

medium to benefit processing in another (e.g., learning new words that appear in newswires to enhance spoken language models for transcription of radio broadcasts). A central challenge will be the rapid learning of explainable and robust systems from noisy, partial, and small amounts of learning material. Community defined evaluations will be essential for progress; the key to this progress will be a shared infrastructure of benchmark tasks with training and test sets to support cross-site performance comparisons.

Neuroscience-inspired Processing and Architectures. Neuroimaging of our existence proof of multimedia processing, the human mind, is yielding functional architectures of the brain and mind that are highly parallel and interconnected but yet have observable specialization. Interestingly, computer processing architectures for multimedia are simultaneously maturing from sequential to parallel and highly connected (note many two-way arrows in Figure 8). Neuroscience is also helping cognitive psychologists refine their models of cognition [16], such as discovering that humans directly access brain areas associated with semantic knowledge following visual and auditory perception of numbers, contradicting long held beliefs that humans re-code auditory stimuli in visual short term memory prior to semantic access.

Observations of child learning and neuroscience suggest that exploiting information from multiple modalities (i.e., audio, imagery, haptic) reduces processing complexity. For example, researchers [27] have begun to explore early word acquisition from natural acoustic descriptions and visual images (e.g., shape, color) of everyday objects in which mutual information appears to dramatically reduce computational complexity. This work, which exploits results from speech processing, computer vision, and machine learning [27], is being validated by observing mothers in play with their pre-linguistic infants performing the same task.

Neuroscientists and cognitive psychologists are only beginning to discover and, in some cases, validate abstract functional architectures of the human mind. However, even the relatively abstract models available from today's measurement techniques (e.g., low fidelity measures of gross neuroanatomy via indirect measurement of neural activity such as cortical blood flow) promise to provide us with new insight and inspire innovative processing architectures and machine learning strategies.

CONCLUSION

The use of multimedia information and interfaces continues to rise and our ability to harness multimedia for user benefit will remain a key challenge as we move into the next millennium. Methodologies such as corpus-based systems and biologically inspired processing give us reason to be optimistic that we will improve our understanding of and, if successful, achieve regular and predictable progress with sophisticated multimedia interfaces to information and people.

ACKNOWLEDGEMENTS

I would like to thank Lynette Hirschman, David Palmer, John Burger, and Andy Merlino for providing the spoken language processing examples for BNN. Sam Bayer and Laurie Damianos are responsible for the multimodal logger. I also thank Stanley Boykin and Andy Merlino for providing BNN story segmentation performance results.

REFERENCES

[1] Aberdeen, A., Burger, J., Day, D., Hirschman, L., Robinson, P. & Vilain, M. 1995. MITRE: Description of the Alembic System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference* (MUC-6), Columbia, MD, 6-8 November 1995, 141-155.

[2] Allan, J., Carbonell, J., Doddington, G., Yamron, J. & Yang, Y. 1998. Topic Detection and Tracking Pilot Study: Final Report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. February 1998. p. 194-218.

[3] Bayer, S., Damianos, L., Kozierok, R., Mokwa, J. 1999. The MITRE Multi-Modal Logger: Its Use in Evaluation of Collaborative Systems. ACM Computing Surveys. March.

[4] Boykin, S. and Merlino, A. forthcoming. Improving Broadcast News Segmentation Processing. IEEE International Conference on Multimedia and Computing Systems. Florence, Italy. 7-11 June 1999.

[5] Boyle, C. and Encarnacion, A. O. 1994 An Adaptive Hypertext Reading System. *UMUAI* 4(1): 1-19.

[6] Card, S. K.; Moran, T. P. & Newell, A. 1983. *The Psychology of Human-Computer Interaction*. Hillsdale, N.J.: Erlbaum

[7] Cuomo, D. 1994. Understanding the Applicability of Sequential data Analysis Techniques for Analysing Usability Data. *Behavior and Information Technology*: 13 (1,2): 171-182.

[8] DARPA Broadcast News Workshop, 28 February – 3 March 1999, Herndon, VA.

[9] Goodman, B., Soller, A., Linton, F. and Gaimri, R. 1998. Encouraging Student Reflection and Articulation Using a Learning Companion. *International Journal of Artificial Intelligence in Education*, 9:237-255.

[10] Harman, D. 1998. The Text Retrieval Conferences (TRECs) and the Cross-Language Track. In Rubio, A. Gallardo, N., Castro, R. & Tejadaeck, A. (eds.) *Proceedings of the First International Conference on Language Resources and Evaluation*, 517-522. European Language Resources Association, Granada, Spain.

[11] Hirschman, L. 1998. The Evolution of Evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language*. 12: 281-305.

[12] Hyland, R., Clifton, C., and Holland, R. 1999. Geonode: Visualizing News in Geospatial Context. AFCEA Federal Data Mining Symposium. Washington, D.C.

[13] Koons, D. B., Sparrell, C. J., and Thorisson, K. R. 1993. Integrating Simultaneous Output from Speech, Gaze, and Hand Gestures. In Maybury, M. (ed.) *Intelligent Multimedia Interfaces*, 243-261. Menlo Park: AAAI/MIT Press.

[14] Lewis, C.; Polson, P.; Wharton, C. & Rieman, J. 1990. Testing a Walkthrough Methodology for Theory-based Design of walk-up-and-use Interfaces. In Proceedings of CHI, 235-242. Seattle, WA 1-5 April. New York: ACM.

[15] Linton, F., Charrron, A. and Joy, D. 1998. OWL: A Recommender System for Organization-wide Learning. AAAI Workshop on Recommender Systems. Madison, WI.

[16] Kosslyn, S. 1994. *Image and Brain*. Cambridge, MA: MIT Press.

[17] Maybury, M. T. (ed.) 1993. *Intelligent Multimedia Interfaces*. Menlo Park: AAAI/MIT Press. (http://www.aaai.org:80/Press/Books/Maybury1)

[18] Maybury, M. T. 1993. Planning Multimedia Explanations using Communicative Acts. In Maybury, M. (ed.) *Intelligent Multimedia Interfaces*, Cambridge, MA: AAAI/MIT Press, 60-74.

[19] Maybury, M. T. (ed.) 1997. *Intelligent Multimedia Information Retrieval*. Menlo Park: AAAI/MIT Press. (http://www.aaai.org/Press/Books/Maybury-2)

[20] Maybury, M. T., Bayer, S. and Linton, F. 1999. Corpus-based User Interfaces. International Conference on Human Computer Interaction, Munich, 26 August 1999.

[21] Maybury, M., Merlino, A., and Morey, D. 1997. Broadcast News Navigation using Story Segments, ACM International Multimedia Conference, Seattle, WA, November 8-14, 381-391.

[22] Maybury, M. T. and Wahlster, W. (eds.) 1998. *Readings in Intelligent User Interfacees*. Morgan Kaufmann: Menlo Park, CA.

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

[24] Nagao, K. and Takeuchi, A. 1994. Speech Dialogue with Facial Displays: Multimodal Human-Computer Conversation, ACL-94, 102-109.

[25] Neal, J.G. & Shapiro, S.C. 1991. Intelligent Multi-Media Interface Technology. In Sullivan, J. W., & Tyler, S. W. (eds.) *Intelligent User Interfaces*. Frontier Series. New York: ACM Press, 11-43.

[26] Palmer, D., Burger, J. and Ostendorf, M. 1999. Information Extraction from Broadcast News Speech Data. DARPA Broadcast News Workshop, 28 February – 3 March 1999, Herndon, VA.

[27] Roy, D. and Pentland, A. 1998. Learning Words from Audio-Visual Input. International Conference on Spoken Language Processing, Sydney, Australia, Dec, 1998. Vol. 4, p. 1279.

[28] Wahlster, W. Keynote. Agent-based Multimedia Interaction for Virtual Web Pages. ACM International Conference on Intelligent User Interfaces. Los Angeles, CA, 6 January 1999.

[29] Zue, V. 1997. Conversational Interfaces: Advances and Challenges. ESCA Eurospeech97. Rhodes, Greece, Page KN-9-16.