Language Technology – a Survey of the State of the Art

Language Resources – Multimodal Language Resources

Mark T. Maybury
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730
maybury@mitre.org

Abstract

This article provides an overview of research in multimodal language processing and associated resources. It defines multimodal processing, describes key challenges, identifies potential benefits, and outlines the major tasks, including multimodal input interpretation, multimodal output generation, and multimodal information access. The article exemplifies the state of the art in multimedia and multimodal processing, describes multimodal language resources and annotation, and finally presents a 2003-2006 roadmap that points the way to the future.

Keywords: multimedia, multimodal, resources, language, speech, gesture, gaze, audio, video, text, graphics
Terms for Glossary term on language technology: multimedia, multimodal, multicodal

Introduction

Human-human communication by its nature exploits multiple input and output modalities. Humans draw upon a rich mixture of communicative mechanisms such as speech, gesture, gaze, facial expression, and body posture for face-to-face communication as well as communication via computers and via artifacts such as video. Equipping computers with human-inspired multimodal facilities should enhance the naturalness of human computer interaction. Designed wisely, we might also improve communicative speed and accuracy. And we might even enhance human-human computer-mediated interaction by increasing the bandwidth of communication (which occurs today primarily via keyboard and mouse), increasing signal-to-noise ratio, increasing the learnability of interfaces, and increasing the joy of the interactive experience.

To achieve these benefits, several national and international programs focus on multimodal resources, including the United States DARPA Human Language Technology (HLT) program, the European Union HLT program under FP5-IST, the German Mensch-Technik-Interaktion (MTI) Program[1], the Francophone AUF program, and others. The European 6th Framework program (FP6)[2], planned for a start in 2003, includes multilingual and multisensorial communication as major research and development issues. Multimodal

[1] http://www.dlr.de/IT/IV/MTI

[2] http://www.cordis.lu/rtd2002/fp-debate/fp.htm

resources are necessary to enable technology development, evaluation, and application maturation.

Definitions

Following Maybury and Wahlster (1988), we distinguish media, modes, and codes. By *medium* we mean the material on which or through which information is captured, conveyed or interacted with (i.e., text, audio, video). In contrast, we use *mode* to refer to the human perceptual systems that enable sensing (e.g., visual, auditory, tactile modalities). Both media and modes may be formalized in a variety of syntactic, semantic, and pragmatic languages, so we also define the notion of *code* which includes representations for and interrelations among language, graphics, gesture, and so on. By multimedia, multimodal, and multicodal, we imply the synergistic combination of two or more of these.

Grand Challenges

A number of visionary capabilities could be enabled by multimedia and multimodal processing. Two examples of these capabilities and associated grand challenge problems include:

Intelligent Multimodal Interfaces: The interpretation and generation of cross media input and output, tailored to the specific needs and desires of the user. This requires multimedia input interpretation, including the ability to understand ambiguous, impartial or inconsistent cross modal input. It also implies tailored multimedia presentation generation, wherein both the selection of content and its form of presentation (media and modalities) are dynamically adapted to the situation and needs of the user. Finally, cross media interaction management means the system is actively monitoring the content, computing platform, environment, and user's choice of and reaction to media and modalities to modify its behavior to optimize the likelihood of communication success.

Intelligent Multimodal Presentation Planning: The automated selection of relevant content, structuring and ordering of material, allocation of content to media, design, realization, and coordination of media and modalities, and generation of effective layout. This is aimed at providing tailored information presentation sensitive to the user, domain, task, available media and modalities, and application environment. This is necessary to ensure a coherent, cohesive, and effective presentation.

Multimedia Content Understanding: The processing of multimedia artifacts (e.g., captioned images, broadcast news video, surveillance or meeting video) to interpret the simultaneous speech, non-speech audio, (still and motion) imagery, and any associated text streams (e.g., camera meta data, closed captions) to including retrieval, extraction, summarization, visualization and so on. This would enable a range of applications such as advanced video analysis, personalized news, meeting information access, and automated behavior interpretation. These applications may rely upon the ability to transcribe, retrieve, translate, extract, summarize, visualize, or in general analyze content from possibly massive, heterogeneous, multilingual, and multimedia archives.

Benefits

There are multiple potential benefits from having a computer system support multimodal interaction. These include:

Flexibility – With multiple methods of interaction, users have a choice of input and/or output media. This might be necessary, as in the case of a user who is unable to communicative via a particular modality such as speech, vision, or gesture (e.g., blind, deaf, paraplegic).

Efficiency – Certain tasks can be performed more quickly with appropriate input or output devices. For example, selection of geospatially attributed (e.g. designating a preferred region on a map when searching for houses or apartments) might be more efficiently accomplished with a hand/pen gesture. In contrast, selecting a subset of objects based on an abstract property (e.g., the price of the house) can be accomplished more rapidly using language or selecting from a menu rather than selecting individual objects.

Task Effectiveness – certain tasks are performed more accurately, with fewer errors, in the appropriate modality. Speech recognizers have high word error rates when transcribing conversational speech or in noisy environments whereas simple menu selections may yield very few errors, although be more constraining with regard to the range of input.

Usability – Humans enjoy certain kinds of interfaces over others. For example, André et al. (1999) showed empirically that while animated interface agents don't necessarily improve the efficiency or effectiveness of users in information seeking tasks, they do enjoy the experience more. This could yield decreases in stress and increases in user motivation. Or simply social/environmental factors may be critical for usability. For example, using speech input for a passcode at an ATM in a public space is undesireable because of privacy and security.

Cross-modal synergy. In addition, errors in one mode (e.g., imprecise, ambiguous, or incorrect) gestural input can often be corrected by processing and integrating synchronous input from another modality (e.g., speech).

All of the benefits, however, rely upon careful design and implementation of interaction.

Tasks and Abstract Architecture

Multimodal language processing can be divided into several key task areas. These include multimodal input processing, multimodal output generation, and multimodal content processing. Figure 1 (cf. Maybury and Wahlster 1988) illustrates the relationship of many these areas within an abstract systems architecture. As is illustrated in the figure, in a multimedia interface, users might interact via a variety of input devices such as mouse, keyboard, microphone, and possibly even body, face, and eye trackers. This media input is analyzed by processors that interpret language, graphics, gesture, and so on. This analysis might include biometrics (e.g., voice, gesture and/or retina/iris analysis or physiological analyses such as breath and heart rate or skin conductivity) for identification, authentication, and/or status monitoring of users. If input is cross modal, a mode

coordinator may need to fuse media as well as mutually disambiguate media inputs during a process of cross modal reference resolution. An interaction manager will then perform such tasks as recognizing user identity, goals, and intentions, and populate models of the user, the unfolding discourse, task, and environmental context. A multimodal system may be used to access information, applications, or other users. Multimodal information access might include the need for algorithms that process multimedia and/or multimodal artifacts such as audio or video archives.

Having retrieved relevant information, the interaction manager then needs to package possibly heterogeneous elements into a coherent and cohesive presentation. This could include both media design (e.g., content selection, media allocation, structure, order, layout of language, graphics, and gesture) and synthesis and rendering of output onto a variety of presentation mechanisms such as maps, spoken language, gesture, and display devices (e.g., monitors, speakers, animated agents). This entire process generates and relies upon detailed models and histories of the user/agents, discourse, context, domain, task, media, and applications.

Having described this abstract architecture, we next consider some key developments that have occurred in the areas of multimodal input processing, multimodal output generation, and multimodal content processing.

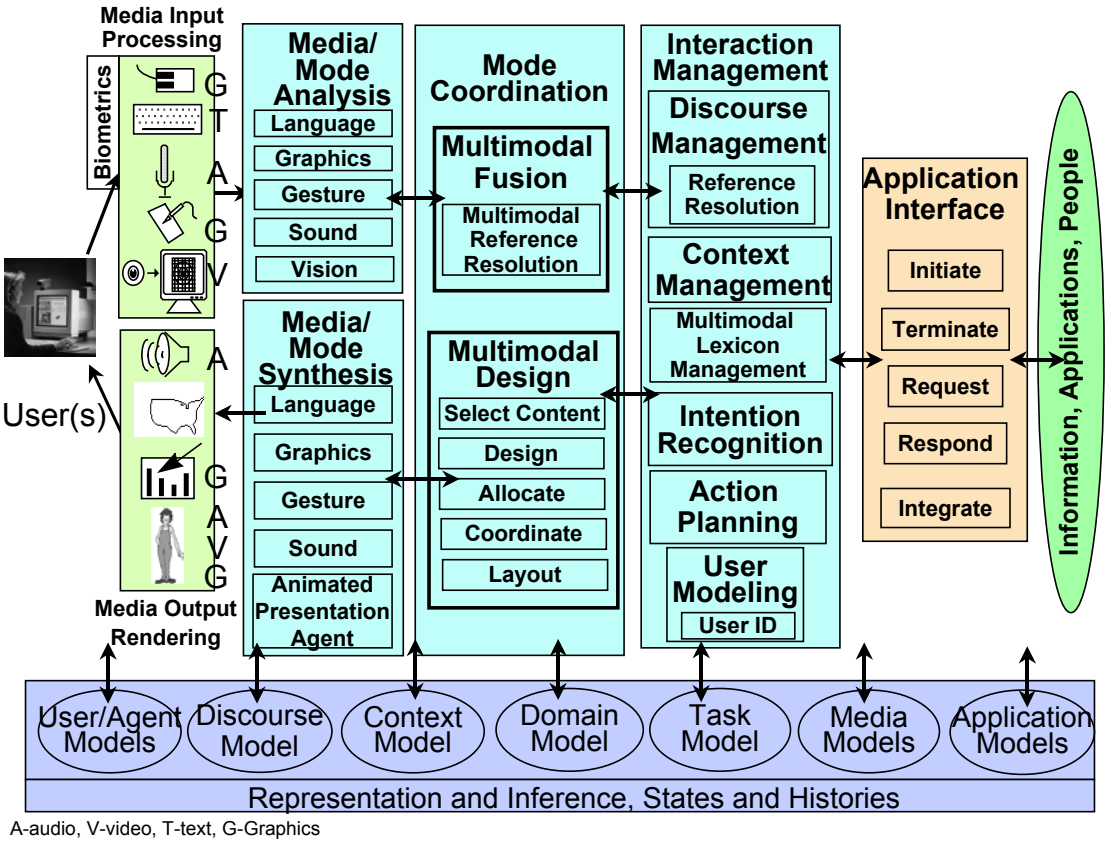


Figure 1. Multimedia and Multimodal Interfaces: Abstract Architecture

Multimodal Interpretation

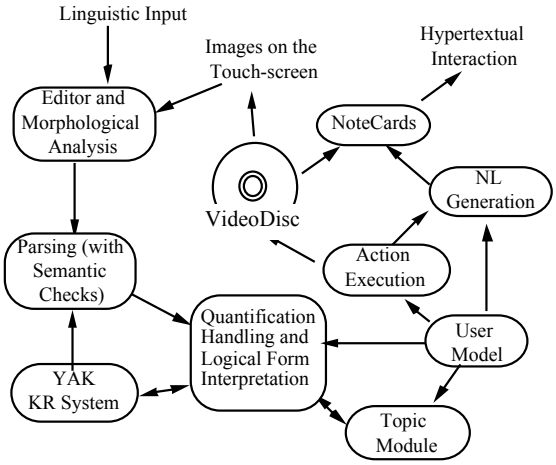
There have been many pioneering efforts in interpreting mixed and asynchronous multimedia user input, such as spoken language input with gesture. These include the classic "Put that there" system by Bolt (1980), that enabled users to combine spoken language commands with hand gestures to manipulate blocks-world shapes around a graphical display. The user could create, modify, move, delete, and even name objects, and in the process use pronouns such as "that" or "there" to refer to objects or their locations, which are resolved by corresponding gestural input.

Similarly, the XTRA (eXpert TRAnslator) interface for filling out tax forms (Wahlster 1991), had a subcomponent named TACTILUS (Kobsa et al. 1986) that enabled the interpretation of mixed language and gesture. The user could choose from a menu of deictic gestures of varying "granularity" and function (e.g., a pencil, index finger, hand, or region encircler). There were no pre-defined screen areas, hence no one-to-one correspondence between a location on the screen and a domain object. The system would resolve inexact and pars-pro-toto (part for whole) pointing by first computing a "plausibility value" (the portion demonstratum covered by the pointer). Next, it would prune candidates using the semantics of associated language and dialogue. Thus, if a user points to a portion of the form between the name and date box and then types in a name, it is pretty clear what he or she intended by the pointing gesture. Whereas in TACTILUS the

language and pointing occur sequentially, in subsequent work this constraint is relaxed (Koons et al. 1993, 1994).

Other research has investigated the role of direct manipulation and natural language. In the ALFRESCO art system for exploration (Stock 1993), users can navigate art masterpieces and gesture while asking natural language questions. The system (see Figure 2) explores the mutual advantages of hypermedia and natural language processing. Hypermedia organizes heterogeneous and unstructured information, enabling direct manipulation integrated with language, thus facilitating exploration. Natural language parsing provides direct query (of nodes or links, of subnetworks), rapid navigation, helping to overcome disorientation and cognitive overhead caused by too many links. Both a gesture and a linguistic expression may be ambiguous and yet yield a unique referent through mutual constraint. Simple natural language generation can be combined with more complex canned text (e.g., art critiques) and images. As Figure 2 illustrates, users can interact in a combination of language and gesture and the system can similarly respond.

SYSTEM **TRANSCRIPT**



USER: Vorrei sapere se Ambrogio Lorenzetti ha dipinto un'opera che raffigura una scena sacra con un angelo. (I would like to know if Ambrogio Lorenzetti ever painted a work that represents a sacred scene with an angel.)
ALFRESCO: Sì. Per esempio: A. Lorenzetti, l'Annunciazione. L'angelo é l'Arcangelo Gabriele. (Yes, for example, A Lorenzetti, the annunciation. The angel is the Arcangel Gabriel.)
USER: Chi è questa persona? [pointing at Mary on the touch screen] (Who is this person?)
ALFRESCO: a Madonna. (Mary.)
USER: Puoi mostrarmi un ingrandimento che la contiene? (Can you show me an enlargement containing her?)
ALFRESCO: [The system shows an enlargement]

Figure 2. ALFRESCO: Language and Gesture Input

Similarly, CUBRICON (Neal and Shapiro 1992) explored both the interpretation and generation of spoken natural language with coordinated gestures in the context of a mission planner. An augmented transition network (ATN) for natural language interpretation was extended with mouse gestures support for both noun phrase and locative adverbials. This enabled mixed modality input including:

- Interrogative: "Is this <point> a surface to air missile?"
- Imperative: "Enter this <user-points-to-map-icon> here <user-points-to-slot-in-form>."
- Declarative: Units from this <point-1> airbase will strike these targets <point-2> <point-3> <point-4>."

An even more sophisticated multimodal analysis system is required to address continuous, overlapping, and ambiguous input. For example, Koons et al. (1993, 1994) explored the integration of simultaneous speech, gesture, and eye movement for reference resolution for map and blocks world interaction. These researchers developed a model of sequences of gesture features (hand posture, orientation, and motion) to classify hand movements as a complex set of actions such as pointing, which consists of an attack, sweep, and end reference. Semantic features from spoken language, gestures, and gaze (fixations, saccades, blinks) are interpreted in parallel in order to mutually constrain ambiguous expressions such as “put that blue square below the red triangle” as well as use speech and “depictive" gestures in a three-dimensional blocks world to describe some action on the objects in the displayed scene.

Oviatt (1999) found that multimodal input can not only support preferred interaction styles by providing more choice, but also enhance robustness. For example, via user studies, she found that users reduce errors via their natural mode selection. In some applications she found 80% error avoidance via methods such as mutual disambiguation across media. In the context of the Quickset (Cohen et al. 1997) interface for map-based planning, Oviatt explored the use of 100 Quickset commands and 200 military symbols using a lexicon of approximately 500 words and 9 gestures. For example, in an analysis of 2600 within subject commands in a mobile noisy environment, Oviatt found a 41% reduction in speech error rate and a 19% error reduction in mobile environments during the use of the multimodal “PAN” command. The Quickset system exploits the N-best results of speech and gesture input and “pulls-up” lower ranking interpretations if there is a consistent cross modal interpretation. In only about 2% of the cases is there a failure in both speech and gesture modes. Oviatt notes that diverse user groups (e.g., children or accented speakers) as well as field environments drive requirements for improved error handling and robustness.

In a related finding, Oviatt et al. (1994) found that when users spoke phrases and sentences to fill a slot in a visual form as opposed to speaking to an open workspace they exhibited a three-fold reduction in bigram perplexity, syntactic complexity, semantic integration, and spoken disfluencies. This demonstrated how user interfaces could positively impact the processibility of utterances, overcoming some of the weaknesses of communication via language alone, again demonstrating the complementarity of direct manipulation and natural language (Cohen 1992). Other research has demonstrated how visual information (e.g., lip movements, body posture) can also help resolve ambiguous speech as well as convey additional information (e.g., about focus of attention, communication success, and participant attitudes and opinions).

Multimodal Presentation

In addition to general investigations of information visualization techniques (Gershon and Eick 1995), researchers have developed a range of methods automatically generate multimodal presentations. The first of these systems explicitly represented graphical knowledge (Card et al. 1991) and made decisions among graphical encoding mechanisms by reasoning about the expressiveness and effectiveness of underlying representations and resulting presentations. In addition, others have investigated presenting information in sound or sonification (Kramer 1994) as well as presenting information in spoken and

written language and knowledge-based graphics (Feiner et al. 1993). More generally, the concept of multimedia interfaces (Maybury 1993) incorporates a range of media and modalities during interpretation and presentation.

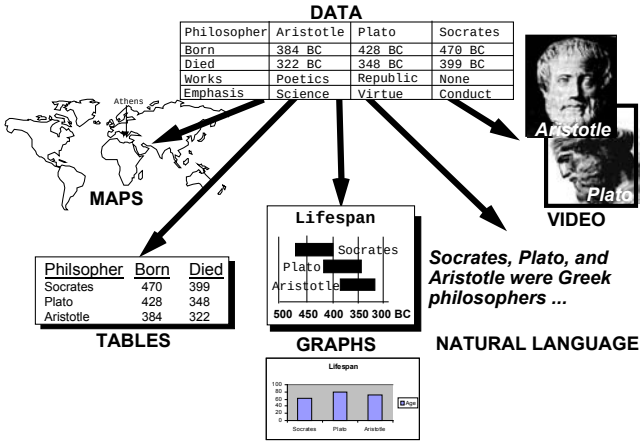


Figure 3. Multimedia/Multimodal Presentation

As Figure 3 illustrates using an example about philosophers, the same information can be presented in a variety of media artifacts such as maps, tables, graphs, spoken or typed natural language, and even video. Challenges include how to decide what content about philosophers should be chosen to satisfy a given user information need, what information should be allocated to what media, how should a media artifact be generated and realized in a manner tailored to the user, how can it be realized as text, graphics or combinations of media, which then need to be realized and coordinated. These challenges are related in Figure 4, which illustrates the key tasks in presentation design which are shown as a set of cascaded, co-constraining processes. By the latter, we mean the kind of content that will influence the layout and the available presentation media or modalities and will constrain the range of content that can be conveyed. Following an analysis of presentation design tasks, a standard reference model (SRM) for intelligent multimodal presentation systems (IMMPS) presentation was created (Bordegoni et al. 1997).

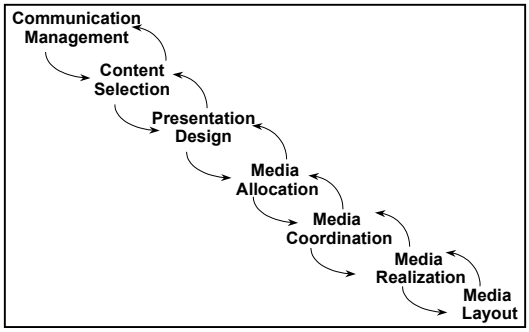


Figure 4. Presentation Tasks

Figure 5 shows the phases of processing required in the Personalized Plan-based Presenter (PPP) (André et al. 1996, 1999) which reasons about and plans the communicative actions and interactions of a life-like agent who narrates animated mixed media presentations. Multimodal communicative actions are driven from the presentation task and include reasoning about presentation acts, scheduling them, and then realizing them in a mix of animated agent and graphical actions.

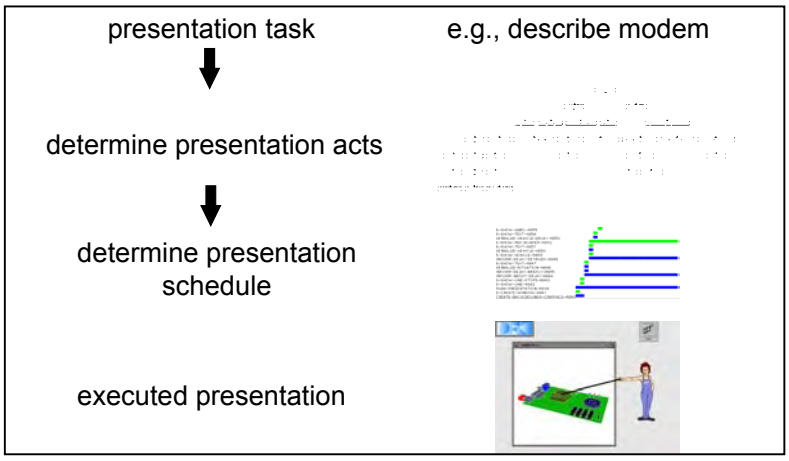


Figure 5. Personalized Plan-based Presenter (PPP)

Multimodal Dialogue

When interpretation and generation are integrated together with components to support discourse analysis, error recovery, and interaction management, we have a multimodal dialogue. Figure 6 illustrates several steps in a multimodal conversation between a human and a virtual agent (named Smartakus) in the SmartKom system (www.smartkom.org) (Wahlster 2001). The user's spoken language, gesture, and facial expressions serve as primary input. Smartakus can then select speech, graphics, and its own facial expressions to convey information back to the user. Because of lack of pre-existing resources, the SmartKom project is creating its own gesture and facial expression database.

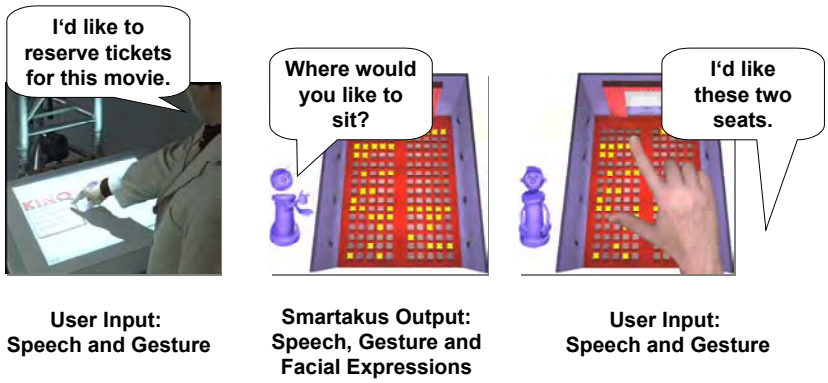


Figure 6. Multimodal Dialogue in SmartKom

Researchers have also explored the integration of spoken dialogue interaction with facial expressions (Nagao and Takeuchi 1994). Figure 7 illustrates examples of a system in which a life-like human agent converses about computer products with a human in Japanese. Facial muscles were modeled to capture emotional displays (see examples in left of Figure 7) and phonemes and visemes were temporally coordinated. Facial expressions are synthesized using wire-frame models of key facial muscles and varying over time, lip synchronized (Waters and Levergood 1994). This enabled mapping from emotional state to expressions using a range of facial displays.

Figure 7. Sony Lifelike Spoken Language Dialogue

What is evident from interactive multimedia is the need for models of human bodies and faces, physical and communicative behavior, and interaction scenarios.

Full-bodied and environmentally situated conversational agents have been subsequently explored by a number of researchers, as exemplified by Jack (Badler, Phillips, and Webber 1993), Steve (Johnson and Rickel 1998), and Rea (Cassell et al. 1994). For example, the Steve project (Johnson and Rickel 1998) at USC/ISI has developed a pedagogical agent named Steve that provides training in virtual environments, both in individual and team settings. Working together with a human in a virtual environment, Steve can demonstrate physical tasks, perform actions together with human or virtual partners, detect and correct human task errors, and use combinations of synthesized speech and eye, head, hand, and body gesture, as well as plan recognition and generation to teach and/or accomplish tasks. Figure 8 (left) shows Steve describing the operation of a machine.

In contrast, The MIT Media Lab's Rea has a fully articulated body, interpreting user speech and sensing user gestures and head movement passively through cameras (Cassell et al. 1994). The agent, named Rea (for Real Estate Agent) (see Figure 8 right), plays the role of a real estate salesperson who interacts with users to determine their needs, shows them around virtual properties, and attempts to sell them a house. Real estate sales was chosen as an application area because of opportunities for both task-oriented and socially-oriented conversation. Coordinated speech, hand gestures, body movements, and facial expressions are synthesized based on a grammar, lexicon, and communicative context. This has inspired the researchers to create animation toolkits to express behaviors (Cassell et al 2001).

Figure 8 Embodied Agents:
Steve points out a power light to a student (left); REA shows off a house (right)

Multimodal Content Analysis

Multimedia and multimodal information occurs not only in human computer interactive contexts, but also in artifacts such a text-captioned images and video. Researchers have begun to explore functions such as topic detection and tracking (TDT) in news (e.g., Maybury 2000) and meeting analysis. However, much research has focused on individual media such as text, audio or image analysis, the state of the art for which are briefly outlined in Table 1. We first highlight a couple of less investigated monomedia (specifically, graphics and sound analysis), and then exemplify research in cross modal analysis.

One less investigated area is the analysis of graphical media. For example, Chuah et al.'s (1997) SageBook enables search and customization of stored data graphics, including data-graphic query, representation (i.e., content description), indexing, search, and adaptation. Queries are formulated via a graphical direct-manipulation interface (called SageBrush) by selecting and arranging spaces (e.g., charts, tables), objects contained within those spaces (e.g., marks, bars), and object properties (e.g., color, size, shape, position). Retrieved data-graphics can be manually adapted. SageBook maintains an internal representation of the syntax and semantics of data-graphics, which includes spatial relationships between

objects, relationships between data-domains (e.g., interval, 2D coordinate), and the various graphic and data attributes.

In related research (Wold et al 1996, Blum et al. 1997), researchers have explored the acoustic and perceptual processing of sound. For example, in their SoundFisher system (www.soundfisher.com) users can access audio using simile (e.g., sounds like "the sound of a herd of elephants" or like a class of sounds, e.g., "applause "), acoustic features (e.g., brightness, pitch, and loudness), subjective features (e.g., "scratchiness," "shimmering"), and onomatopoeia (making a sound similar in some quality to the sound you are looking for, e.g., user makes a buzzing sound to find bees or electrical hum.). In an evaluation the performance of the SoundFisher system has been illustrated on a database of 400 widely ranging sound files (e.g., captured from nature, animals, instruments, speech).

These researchers aim to provide sound-effects designers, computer animators, and presentation designers with the ability to access sounds by browsing or querying by value (e.g., specifying a pitch and duration), using a weighted query by value (e.g., foreground and transition with >.8 metallic and >.7 plucked aural properties and 2000 Hz < average pitch < 300 Hz and duration of 3 seconds), or querying by example and searching for similar sounds. This enables complex queries such as "Find all AIFF encoded files with animal or human vocal sounds that are similar to goose sounds without regard to duration or amplitude."

In other research, multiple modalities of artifacts are processed. For example, consider captioned images. While general image recognition remains unsolved, if a caption names and locates individuals in a picture, face detection of the image can be associated with language processing of the caption.

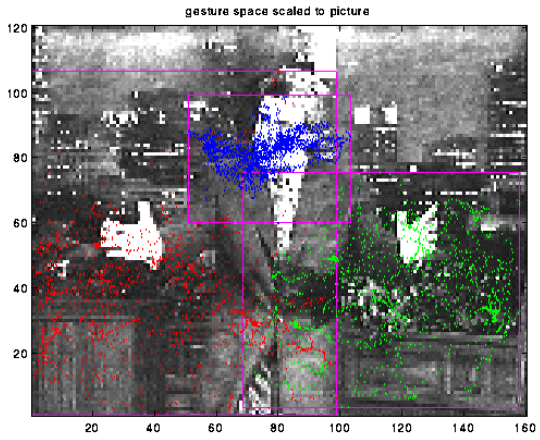


Figure 9. Joke Detector:
Synergistic Image and Audio Processing

Some researchers have explored the area of multimodal video analysis. For example, Casey et al (1995) and Wachman and Picard (2001) analyzed the audio pitch together with head and hand motion in video from "The Tonight Show" comedy monologue and were

able to accurately predict the completion of a joke by stand-up comic Jay Leno. In the example in Figure 9, the boxes bound the hand and head motion throughout the monologue; the dots indicate the centroid location of the head or hands. By measuring hand positions and velocity together with the distribution of pauses and voice pitch in the audio channel, the authors were able to automatically find portions of the comic's monologue where he makes large gestures at long pauses in his speech. This typically corresponds to a point of emphasis at the conclusion of a joke, thus this combination of visual and auditory features results in the so-called Joke Detector.

In multimodal news on demand research (Maybury, 2000), several groups also take advantage of cross modal analysis, integrating the results of image, speech, and text processing of a digital video to detect story segments, extract named entities, represent key frame and sentence summaries, and present this to the user. Figure 10 exemplifies the results of one such system, MITRE's Broadcast News Navigator (BNN), responding to a user query requesting all news stories regarding "Iraq" between Monday to Thursday, November 11 to 14, 2002. One hundred forty-six Iraq stories were found. Figure 10 displays 12 keyframe and key entity summaries as well as the beginning details of the first story. For each story matching the query, the system presents a key frame, the three most frequent named entities within the story, and the source and date of the story. Compared with sequential digital video access, this kind of presentation was shown empirically to more than double the speed of analysts accessing information (Light and Maybury 2002) with no loss in retrieval accuracy.

Figure 10. Broadcast News Navigator: Story Skim (left) and Story Details (right)

Enabling Technologies: State of the Art and Research

Multimedia and multimodal processing requires the ability to process component media such as text, speech, graphics, and imagery, both in isolation and in coordination. It requires the ability to represent and reason about (human and machine) communicative agents and necessary subtasks, such as the ability to interpret and generate cross modal referring expressions. Table 1 indicates component areas of multimodal language processing, the state of the art, near-term research and long-term research.

Area	State of the Art	Near-Term Research	Long-Term Research
Text processing	Commercial named entity extraction (SRA, BBN) at 95% precision and recall (P&R); relation extraction at approximately 80% P&R, and many hand-crafted, domain-specific systems for relation and event extraction at about 60% P&R; large cost to port to new domains; incremental sentence generation, limited document generation	Demonstrate portability of TIPSTER technology to support multilingual information extraction and spoken language retrieval; incremental text generation; text summarization; topic detection and tracking	Scaleable, trainable, portable algorithms; document-length text generation
Speech processing	Commercial small-vocabulary recognizers (Corona, HARK); large-vocabulary (60,000+ words) recognizers exist in research labs (BBN, SRI, Cambridge University).	Robust speaker, language, and topic identification; prosodic analysis; natural-sounding synthesis	Large-vocabulary, speaker-independent systems for speech-enabled interfaces; large-vocabulary systems for multilingual video and radio transcription, noisy environments
Image processing	Color, shape, texture-based indexing and query of imagery. Primitive object (e.g., human, car) detection and tracking in video.	Motion-based indexing of imagery and video; video segmentation. Simple behavior detection (e.g., person interaction, object transfer)	Visual information indexing and extraction, more complex human behavior recognition (e.g., suspicious behavior)
Graphics processing	Graphical User Interface Toolkits (e.g., object-oriented, reusable window elements such as menus, dialogue boxes	Tools for automated creation of graphical user interface elements; limited research prototypes of automated graphics design.	Automated, model-based creation and tailoring of graphical user interfaces
Gesture Processing	Two-dimensional mice; eyetrackers; tethered body-motion tracking	Tetherless, three-dimensional gesture, including hand, head, eye, and body-motion tracking	Intentional understanding of gesture; cross-media correlation (with text and speech processing); facial and body gesture recognition
Multimodal analysis	Limited research prototypes exploring ambiguous, imprecise, and incomplete input	Cross modal referring expression interpretation, speech prosodic and emotion expression recognition	Unrestricted multimedia and multimodal interpretation, advanced modality interpreters (e.g., olfactory), robust human identification.

Multimodal generation	Genre specific presentation generation (e.g., multimodal how-to instructions). Highly complex systems with typically knowledge rich methods of presentation planning and realization.	Content selection, media selection and allocation, media coordination, media realization for multimedia generation. Mixed media (e.g., text, graphics, video, speech and non-speech audio) and mode (e.g., linguistic, visual, auditory) displays tailored to the user and context	Automated generation of coordinated speech, natural language, gesture, animation, non-speech audio, generation, possibly delivered via interactive, animated life-like agents. Cross modal referring expression generation; multimedia and multimodal generation; investigation of less-examined senses (e.g., taction, olfaction).
Animated agents	Many prototypes and preliminary applications of lifelike agents. First life-like, fully articulated anthropomorphic agents capable of engaging in human-like conversation including verbal and nonverbal behaviors.	Agents engaging/motivating users. Agents interpreting and responding to cross modal user input (e.g., speech, gesture, facial movements), and responding with same	Agents capable of engaging in socially, culturally, and individually appropriate conversational behavior. Agents that build relationships with users over time. Social and user implications of conversational virtual humans
Discourse modeling	Limited prototypes in research and government	Error handling (ill-formed and incomplete input/output), two-party conversational model, discourse annotation schemes, discourse data collection and annotation, conversation tracking	Context tracking/dialogue management; multiuser conversation tracking, annotation standards; model-based conversational interaction
User modeling	Fragile research prototypes available from academia; one-user modeling shell (BGP-MS).	Track user focus and skill level to interact at appropriate level; empirical studies in broad range of tasks in multiple media	Hybrid stereotypical/personalized and symbolic/statistical user models, rich modeling of cognition and emotion

Table 1: Component Capabilities key to Multimodal Language Processing (modified from: http://www.nap.edu/readingroom/books/screen/tab1.html)

Multimodal Corpora

Analyzing human multimodal behavior and training algorithms for multimodal processing requires data. *Multimodal corpora* contain primary data (text, audio, and video files) and encodings, possibly on different modality tracks or layers and at different levels of granularity. These encodings can be descriptive or interpretative. For spoken language, standard encodings for human language often include word transcription, part-of-speech, syntactic structures (e.g., noun phrase, sentence), named entities (e.g., person, organization, location), relations among entities (e.g., employee-of, a-part-of, a-kind-of), co-reference, rhetorical relations, dialogue acts, and so on, possibly conforming to a standard like MPEG-7. Non-speech audio such as music, laughing, clapping, noise, and so on may also be annotated, although there are no standard schemes. In closed captioning, standard

conventions exist (e.g., ">>" annotates a speaker shift and ">>>"annotates a topic shift), however, one study found error rates as high as 20% or more on even these simple tags because of human annotation error.

For the visual modality, current research is focused on the description of nonverbal communication through the human body, typically hand or arm gestures (and some posture) and facial expressions. The latter are most often annotated using the Facial Action Coding System, or FACS (Ekman and Friesen 1978). Some researchers focus on annotation of individuals, objects, and their relations in scenes for detection and tracking purposes, e.g., to detect the exchange of a package between two people in a parking lot (cf VACE). While annotation tools are typically based on time, spatio-temporal encodings are becoming increasingly important with applications such as embodied conversational agents or robotic tour guides situated in real and virtual environments. Similarly, annotation of haptics, including pressure and texture on hands, feet, or torso is important for design, gaming, and analysis applications (cf. PHANTOM[3]).

Standardizing encoding schemes enables data sharing and reuse among researchers across a range of applications, including annotation, visualization, query, and analysis. Standard coding schemes exist for part of speech, syntax trees, dialogue acts, and even temporal expressions (Ingria and Pustejovsky 2002). The ISO/TC 37/SC 4 committee (Ide and Romary 2001) is developing a unified coding scheme specification language, enabling data interoperability and reuse across applications such as speech recognizers, language parsers, generators, and so on. General standards such as the Extensible Markup Language (XML) and Synchronized Multimedia Integration Language (SMIL) are often used for data markup. The Hamburg Sign Language Notation System (HamNoSys)[4] is a "phonetic" transcription system often used for gesture markup.

Important annotation tool capabilities include complex search, statistical analysis, visual access to coding schemes, and semi-automatic documentation facilities. Also important are bootstrapping techniques to increase efficiency, especially where standard taxonomies are used (POS, syntax, etc.). Multi-coder annotation can provide update/merge functions (versioning), concurrent coding, and reliability checks.

Several groups have created annotation tools and collected corpora. General data collection and standardization initiatives in the United States include the National Institute of Standards and Technology (NIST)[5] and Linguistic Data Consortium (LDC)[6] and include such collections as broadcast news. In Europe there is the European Language Resources Association (ELRA) with its operational arm, the European Language Resources Distribution Agency (ELDA)[7]. In the United Kingdom there is the Arts Humanities Data

[3] http://www.sensable.com/haptics/haptics.html

[4] http://www.sign-lang.uni-hamburg.de/Projects/HamNoSys.html

[5] http://www.nist.gov

[6] http://www.ldc.upenn.edu

[7] http://www.elda.fr

Service (AHDS) and in Japan the International Committee for the Coordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)[8]. Initiatives to build standard tools include the Architecture and Tools for Linguistic Analysis Systems (ATLAS)[9] in the United States and the Natural Interactivity Tools Engineering (NITE)[10] in Europe (successor of MATE). ELRA fosters the founding of central national agencies for the collection of native language corpora, and organizes the International Conference on Linguistic Resources and Evaluation[11] (LREC).
With respect specifically to multimodality, there is the International Standards for Language Engineering (ISLE)[12] project (formerly EAGLES), and in particular the Natural Interactivity and Multimodality (NIMM)[13] subgroup (Knudsen et al. 2002ab), as well as the TalkBank[14] project and a project at MITRE (Bigbee et al. 2001).

Corpus metadata that specify file formats and annotation schemes are essential for understanding and reuse. This led to the founding of the Open Language Archives Community (OLAC)[15] based on the Dublin Core Metadata Set (DCMS)[16], a standard resource description model. The ISLE MetaData Initiative (IMDI)[17] is also working on meta-data, specifically for multimedia/multimodal corpora. IMDI includes freely available, integrated, Java-based tools including an editor, browser, search and efficiency tools for linguists and software and language engineers.

Toward the Future: A Multimodal Roadmap

While many exciting developments have occurred in the last few years, it is clear that research into multimodal language processing yields more questions than answers, ensuring an active area of research in the near future. Figure 11 illustrates a roadmap created at an international workshop (Maybury and Martin 2002) that depicts three "lanes" of multimodal developments leading up to natural multimodal systems: resources; theories, methods and algorithms; and systems. The roadmap distinguishes between planned activities (in italics font in Figure 11) and desirable ones (in regular font). For example, in terms of resources, there exists the Berlin gesture database, the NITE D2.1 gesture concordance (to be released at nite.nis.sdu.dk/deliverables), an example of a digitally recorded meeting from NIST-MITRE-LDC, the FORM annotation specification and data,

[8] http://www.slt.atr.co.jp/cocosda

[9] http://www.nist.gov/speech/atlas

[10] http://nite.nis.sdu.dk

[11] http://www.lrec-conf.org

[12] http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm

[13] http://isle.nis.sdu.dk

[14] http://www.talkbank.org

[15] http://www.language-archives.org

[16] http://dublincore.org

[17] http://www.mpi.nl/ISLE/

and the SmartKom corpus, the latter two of which are being prepared for distribution in the near future. Subsequent to this in the resources lane are a number of desired developments including a multimodal database of PDA, in-vehicle, or human-human interaction, an evaluated corpora, benchmarks for multimodal systems, cross-cultural databases (e.g., of language, gesture and/or facial expressions), and annotation standards and corpora for social protocol encoding. As shown diagonally along the left hand side of the roadmap, over the next several years we require the development and standardization of metadata, coding standards, mappings, and tools (e.g., ISO TC 37/SC4/WG1) as well as better human factors knowledge of multimodal interaction. Also, requiring data and models as well as new algorithms are emotional and/or user state modeling and synthesis (e.g., to drive life-like characters), then the same for user emotion/state recognition, and longer term (past 2006) robust recognition and management of user states.

A second lane in the road addresses multimodal theories, methods, and algorithms. Following the first multimodal prototypes in the late 1970s and 1980s, we see a current emphasis on universal access, enabling users increased flexibility with choices of modality input and output. To that end, near-term effort is being expended on multimodal evaluation and ISO usability guidelines for multimodal systems. In the near term there is a need for methods to support multimodal dialogue and deep interaction across modalities. In the next few years researchers see progress on 3-D task-specific gesture recognition leading eventually to large vocabulary iconic gesture identification and in the very long term (beyond 2006), general multimodal gesture recognition. Also in the mid term are projected standards specifications for multimodal systems and more fully-featured animated agents, likely including groups of conversational animated agents, beyond the task specific interface agents of today. A little further out we will begin to see exquisite “virtuoso” interfaces for experts that exploit multiple sensory modalities and media devices but require special training and expertise to use. In the longer term, researchers see progress on human-human multimodal interaction models leading eventually to models of the neurocognitive mechanisms supporting such interaction. Ultimately this will lead up to multiparty, multimodal dialogue systems.

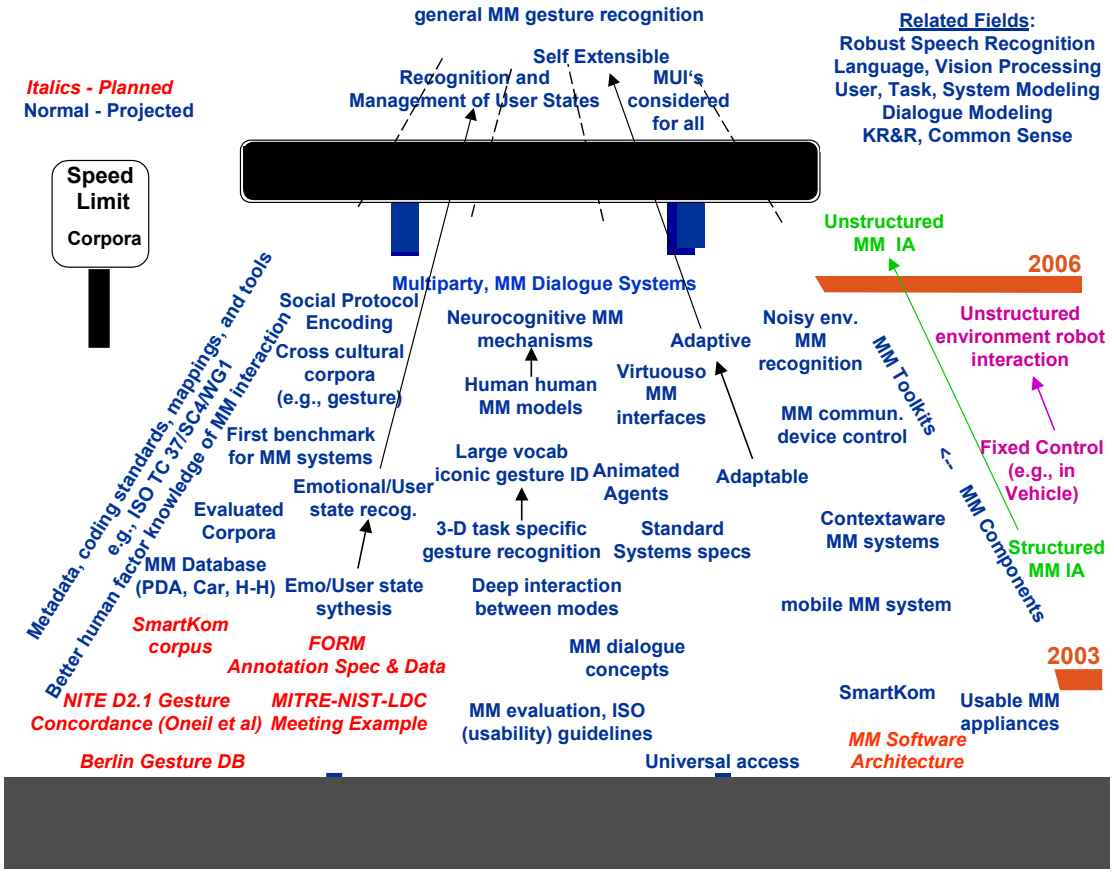


Figure 11. Multimodal Roadmap from 2003-2006

A third lane in the road considers multimodal systems. Based on early efforts, researchers created a common reference model for intelligent multimodal presentation systems (IMMPS) (Bordegoni et al. 1997), and a reference model for multimodal interfaces (Maybury and Wahlster, 1998), evolved to that shown in Figure 1. There are near term multimodal systems (e.g., SmartKom) and there is marketplace desire for "usable" multimodal appliances (e.g., the famous difficulty of using a VCR that might yield to combined speech and direct manipulation interaction). Very shortly we expect mobile, then context-aware multimodal systems leading to multimodal control of devices and robust interaction in noisy environments. Systems in the mid- to long-term and beyond will evolve from providing adaptable (by the user), then adaptive (by the system), and ultimately self-extensible capabilities and behaviors. Ultimately we will begin to see Multimodal User Interfaces (MUIs) replacing the well-known Graphical User Interfaces (GUIs) predominant today. As shown diagonally along the right hand side of the systems lane, multimodal system components and ultimately toolkits will be required to enable rapid system progress. As illustrated to the right of the roadmap, we will also see a progression over the upcoming years from monomodal access to primarily structured multimodal information to increasing levels of sophistication in multimodal access to unstructured multimodal information (e.g., text, audio, video). We will see a concomitant

progression in multimodal control of fixed systems (e.g., in vehicle) to multimodal control in unstructured environments (e.g., multimodal robot control in an open environment).

A number of related fields enable or are closely related to natural multimodal systems, including robust speech recognition, language, vision processing, user, task, and system modeling, dialogue modeling, knowledge representation and reasoning, and common sense. Their success or failure will ultimately pace the progress toward the vision of natural multimodal systems.

Acknowledgements

I thank Michael Kipp for his insights into multimodal standards, encoding schemes, and associated tools. I thank Nancy Ide and Laurent Romary for discussions regarding international language standards and Justine Cassell for her insight on animated agents. I thank all of the participants in the Dagstuhl and LREC multimodal workshops for their contributions to the multimodal systems roadmap.

References

André, E., Müller, J., and Rist, T. 1996. The PPP Persona: A Multipurpose Animated Presentation Agent. In *Advanced Visual Interfaces*, pages 245-247. ACM Press.

André, E., Rist, T., and Müller, J. 1999. Employing AI methods to control the behavior of animated interface agents. *Applied Artificial Intelligence*. 13:415-448.

Badler, N. I., Phillips, C. B., and Webber, B. L. 1993. *Simulating Humans*. New York: Oxford University Press.

Badler. N. 2001. "Virtual Beings," *Communications of the ACM* 44(2): 33-35. March 2001.

Badler, N., Erignac, C., and Liu, Y. 2002. "Virtual humans for validating maintenance procedures," *Communications of the ACM*, 45(7): 56-63, July 2002.

Bigbee, T. and Loehr, D. and Harper, L. 2001. "Emerging Requirements for Multi-Modal Annotation and Analysis Tools." In: *Proceedings of Eurospeech*, pages 1533-1536.

Blum, T., Keislar, D., Wheaton, J. and Wold, E. 1997. Audio Databases with Content-based Retrieval. In Maybury, M. T. (ed.). *Intelligent Multimedia Information Retrieval*. pages 113-135, AAAI/MIT Press.

Bolt, R. A. 1980. Put-That-There: Voice and gesture at the graphics interface. ACM SIGGRAPH *Computer Graphics*, 14(3): 262-270.

Bordegoni M., Faconti G., Feiner S., Maybury M.T., Rist T. Ruggieri S., Trahanias P., and Wilson M. 1997. A Standard Reference Model for Intelligent Multimedia Presentation Systems. In: *Computer Standards & Interfaces* 18(6,7): 477-496. December 1997, North-Holland. (http://www.dfki.uni-sb.de/~rist/csi.html)

Card, S. K., Robertson, G. G., and Mackinlay, J. D. 1991. The Information Visualizer, An Information Workspace. In Proceedings of the Computer Human Interactions: Human Factors in Computing Systems, pp. 181-188. New Orleans, La., April 28-May 2.

Casey, M. Wachman, J, and Wexelblat, A. 1995. "Unsupervised, Cross-modal Characterization of Discourse in 'The Tonight Show' monologues or An Automatic Joke Detector" Final Class Project. "Machine Understanding of Video" Profs. Pentland, S., Picard, R., and Cassell, J. MIT Media Lab.

Cassell, J., Pelechaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone M. 1994. Animated conversation: rule-based generation of facial expression, gesture, and spoken intonation for multiple conversational agents. In Proceedings of ACM SIGGRAPH '94. Computer Graphics Annual Conference Series, 413-420.

Cassell, J., Vilhjalmsson, H., and Bickmore, T. 2001. "BEAT: the Behavior Expression Animation Toolkit" Proceedings of ACM SIGGRAPH 2001, Los Angeles, August 12-17, p.477-486

Chuah, M., Roth, S., Kerpedjiev, S. Sketching, Searching and Customizing Visualizations: A content based approach to Design Retrieval. 1997. In Maybury, M. T. (ed.). *Intelligent Multimedia Information Retrieval*. AAAI/MIT Press, p. 83-111

Cohen, P. R. 1992. The role of natural language in a multimodal interface, Proceedings of the User Interface Software Technology Conference, ACM Press, 1992, 143-149.

Cohen, P.R.; Johnston, M.; McGee, D.; Oviatt, S.,; Pittman, J.; Smith, I.; Chen, L. and Clow, J. 1997. Quickset: Multimodal Interaction for Distributed Applications. ACM International Multimedia Conference, pages 31-40, 1997.

Ekman, P. and Friesen, W.V. (1978). *Facial Action Coding System*. Palo Alto, CA: Consulting Psychologists Press.

Feiner, S., Litman, D., McKeown, K., and Passonneau, R. 1993. Towards Coordinated Temporal Multimedia Presentations. *Intelligent Multimedia Interfaces*, M. Maybury (Ed.), pp. 139-147. AAAI/MIT Press, Menlo Park, Calif.

Gershon, N., and Eick, S. (Eds.). 1995. Proc. Information Visualization '95. IEEE Computer Society Press, Los Alamitos, Calif.

Ingria, R. and J. Pustejovsky 2002. TimeML Specification, time2002.org.

Ide, N. and Romary, L. 2001. "Standards for Language Resources", In: *Proceedings of the IRCS Workshop on Linguistic Databases*, pages 141-149.

Johnson, W. L. and Rickel, J. W. 1998. Steve: An animated pedagogical agents: for procedural training in virtual environments. *SIGART Bulletin* 8: 16-21.

Kendon, A. 1990. *Conducting Interaction*. Cambridge: Cambridge University Press.

Knudsen, M. W., Martin, J. C., Berman, S., Bernsen, N. O., Choukri, K., Dybkjær, L., Heid, U., Mapelli, V., Pelachaud, C., and Poggi, I. 2002a. *Survey of NIMM Data Resources, Current and Future User Profiles, Markets and User Needs for NIMM Resources*, ISLE Deliverable D8.1, http://isle.nis.sdu.dk.

Knudsen, M. W., Martin, J. C., Bernsen, N. O., Dybkjær, L., Heid, U., Pelachaud, C., Poggi, I., Reithinger, van ElsWijk, G., Wittenburg, P., Llisterri, J. and Ayuso, M. J. M, N., Carletta, J. 2002b. *Survey of Multimodal Annotation Schemes and Best Practice*, ISLE Deliverable D9.1, http://isle.nis.sdu.dk.

Kobsa, A., Allgayer, J. Reddig, C. Reithinger, N. Schmauks, D. Harbush, K. and Wahlster, W. 1986. Combining Deictic Gestures and Natural Language for Referent Identification. *Proceedings of the 11th COLING*, Bonn, West Germany, 356-361

Koons, D. B., Sparrell, C. J., and Thorisson, K. R. 1993. Integrating Simultaneous Output from Speech, Gaze, and Hand Gestures. In *Intelligent Multimedia Interfaces*, ed. M. Maybury, 257-276. Menlo Park: AAAI/MIT Press

Koons, D. B. and Sparrell, C. J., 1994. ICONIC: Speech and Depictive Gestures at the Human-Machine Interface. In Proceedings of CHI '94 (SIGRAPH/SIGCHI video review). Boston, Mass. 1994

Kramer, G. (Ed.). 1994. *Auditory Display: Sonification, Audification, and Auditory Interfaces*. Addison-Wesley, Reading, Mass.

Light, M. and Maybury, M. May 2002. Personalized Multimedia Information Access: Ask Questions, Get Personalized Answers. *Communications of the ACM* 45(5): 54-59. (www.acm.org/cacm/0502/0502toc.html). In Brusilovsky, P. and Maybury, M. (eds). Special Section on The Adaptive Web

Mariani, J. 1996. Multimedia. *Survey of the State of the Art in Human Language Technology*. Cole, R. (ed). http://cslu.cse.ogi.edu/HLTsurvey/ch9node2.html#Chapter9

Maybury, M. (Ed). 1993. *Intelligent Multimedia Interfaces*. AAAI/MIT Press, Menlo Park, CA.

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

Maybury, M. and Martin, J. 2002. Workshop on Multimodal Resources and Multimodal Systems Evaluation. Third International Conference On Language Resources And Evaluation (LREC'2002), Las Palmas, Canary Islands, Spain. Saturday, June 1, 2002. www.lrec-conf.org/lrec2002/lrec/wksh/Multimodality.html

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

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

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

Oviatt, S. L., Cohen, P. R., and Wang, M. Q. 1994. "Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity," *Speech Communication*, 15: 283-300.

Oviatt, S. L. 1999. "Mutual Disambiguation of Recognition Errors in a Multimodal Architecture." ACM Conference on Human Factors in Computing Systems (CHI'99), Pittsburgh, PA, May 15-20, pp. 576-583.

SMIL. Synchronized Multimedia Integration Language (SMIL) 1.0 Specification . W3C Recommendation REC-smil-19980615, World Wide Web Consortium, 1998.

Stock, O. and the ALFRESCO Project Team. 1993. ALFRESCO: Enjoying the Combination of Natural Language Processing and Hypermedia for Information Exploration. In *Intelligent Multimedia Interfaces*, ed. M. Maybury, 197-224. Menlo Park: AAAI/MIT Press

Taylor, M., and Bouwhuis, D. G. (eds) 1989. *The Structure of Multimodal Dialogue*. B. U.: Elsevier Science Publishers

Video Analysis and Content Extraction (VACE) Program. http://www.ic-arda.org/InfoExploit/vace.

Wachman, J. and Picard, R. 2001. Tools for Browsing a TV Situation Comedy Based on Content Specific Attributes. *Multimedia Tools and Applications*. 13(3):255-284.

Wahlster, W. 1991. User and Discourse Models for Multimodal Communication. In Sullivan, J. W., and Tyler, S. W. (eds.) *Intelligent User Interfaces*. Frontier Series. New York: ACM Press, 45-67.

Wahlster, W. 2001. Dialog-based human-computer interation by coordinated analysis and generation of multiple modalities. International HCI Status Conference. 26/27 October, 2001. Saarbrueken, Germany. http://www.dlr.de/IT/IV/Tagungsberichte/MTI_Tagung/lectures/smartkom_vortrag.pdf

Wold, E., Blum, T., Keislar, D., and Wheaton, J. 1996. Content-Based Classification, Search, and Retrieval of Audio. *IEEE Multimedia* 3,(3): 27-36. Fall 1996.

Waters, K. and Levergood, T. 1994. An automatic lip-synchronization algorithm for synthetic faces. Proceedings of ACM Multimedia, San Francisco, 15-20 October, 149-156.

XML. Extensible Markup Language (XML) 1.0. W3C Recommendation REC-xml-19980210, World Wide Web Consortium, 1998.