UNIVERSITY OF CRETE
COMPUTER SCIENCE DEPARTMENT

COURSE CS-564 (OPTIONAL)
ADVANCED TOPICS
IN HUMAN – COMPUTER INTERACTION

Course Convenor: Constantine Stephanidis

Multimodal Interaction

What are multimodal systems?

- Multimodal systems process
  - ⇨ two or more **combined user input modes**
    - e.g., speech, pen, touch, manual gestures, gaze, head and body movements
  - ☞ in a coordinated manner with **multimedia** system **output**

Aims and Objectives of multimodal systems

- Two unique features of multimodal architectures and processing are:
  - the fusion of different types of data
  - real-time processing and temporal constraints imposed on information processing
- Multimodal systems represent a new class of user-machine interfaces, different from standard WIMP interfaces
  - They emphasize the use of richer and more natural ways of communication, such as speech or gestures, and more generally all the five senses
- The objective of multimodal interfaces is twofold:
  - to support and accommodate users' perceptual and communicative capabilities; and
  - to integrate computational skills of computers in the real world, by offering more natural ways of interaction to humans

Advantages of multimodal systems

- Multimodal interfaces were first seen as more efficient than unimodal interfaces
  - however, evaluations showed that multimodal interfaces only speed up task completion by 10%
  - Hence, efficiency should not be considered the main advantage of multimodal interfaces

Advantages

- Improvement of error handling and reliability
  - users made 36% fewer errors with a multimodal interface than with a unimodal interface
- Expressive power and precision in visual-spatial tasks
- Support for users' preferred interaction style
  - since 95%-100% of users prefer multimodal interaction over unimodal interaction

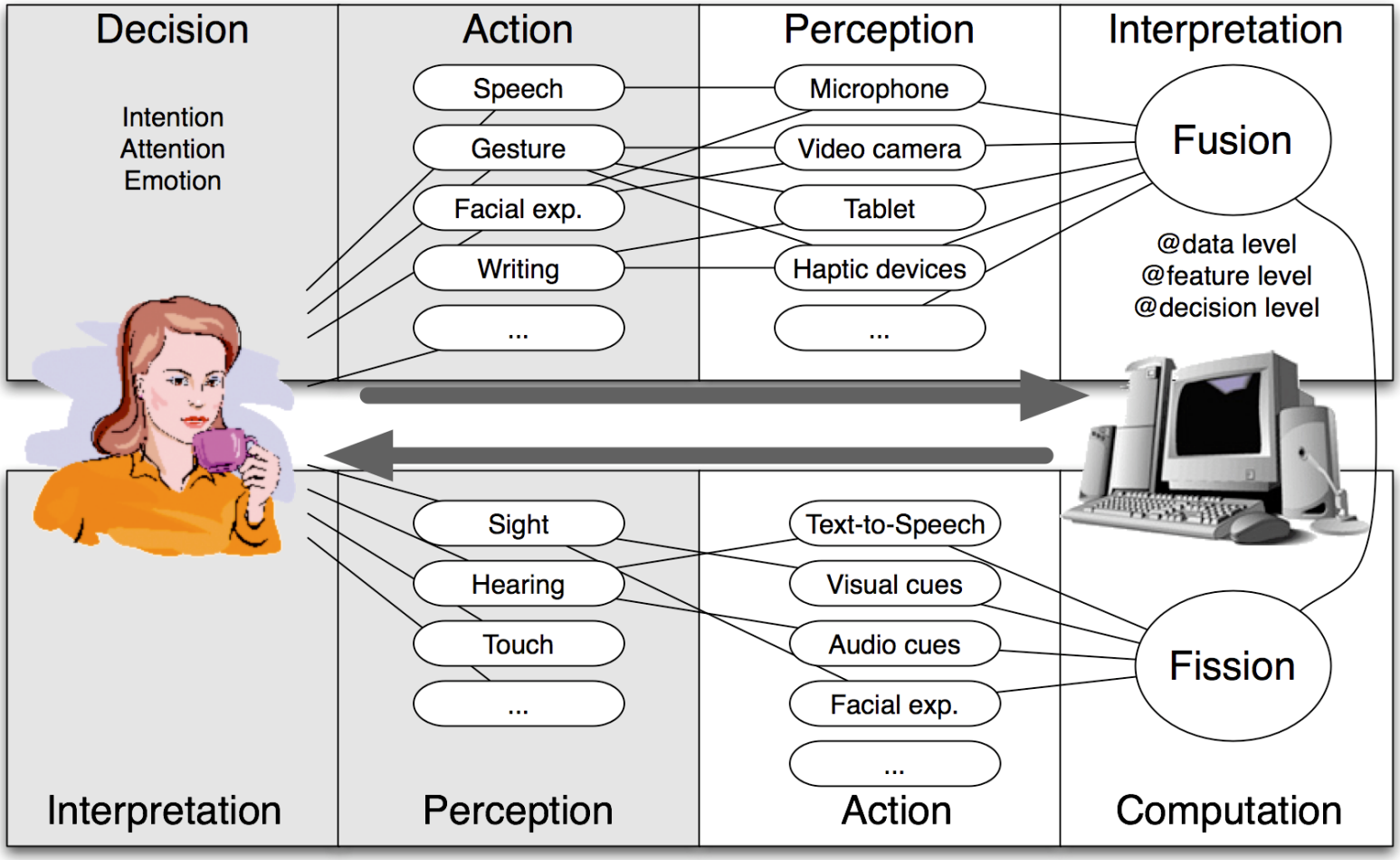
Multimodal vs. graphical user interfaces

GUI	Multimodal
Single input stream: a singular physical input device is used to control the position of a cursor and present information organized in windows and represented with icons	**Multiple input streams:** various modalities can be used as input streams
Atomic, deterministic: either mouse position or characters typed on a keyboard are used to control the computer	**Continuous, probabilistic:** input streams have to be first interpreted by probabilistic recognizers; events are not always clearly temporally delimited and thus require a continuous interpretation
Sequential processing	**Parallel processing:** due to the multiple recognizers necessary to interpret multimodal input and the continuous property of input streams
Centralized architectures	**Distributed & time-sensitive architectures**

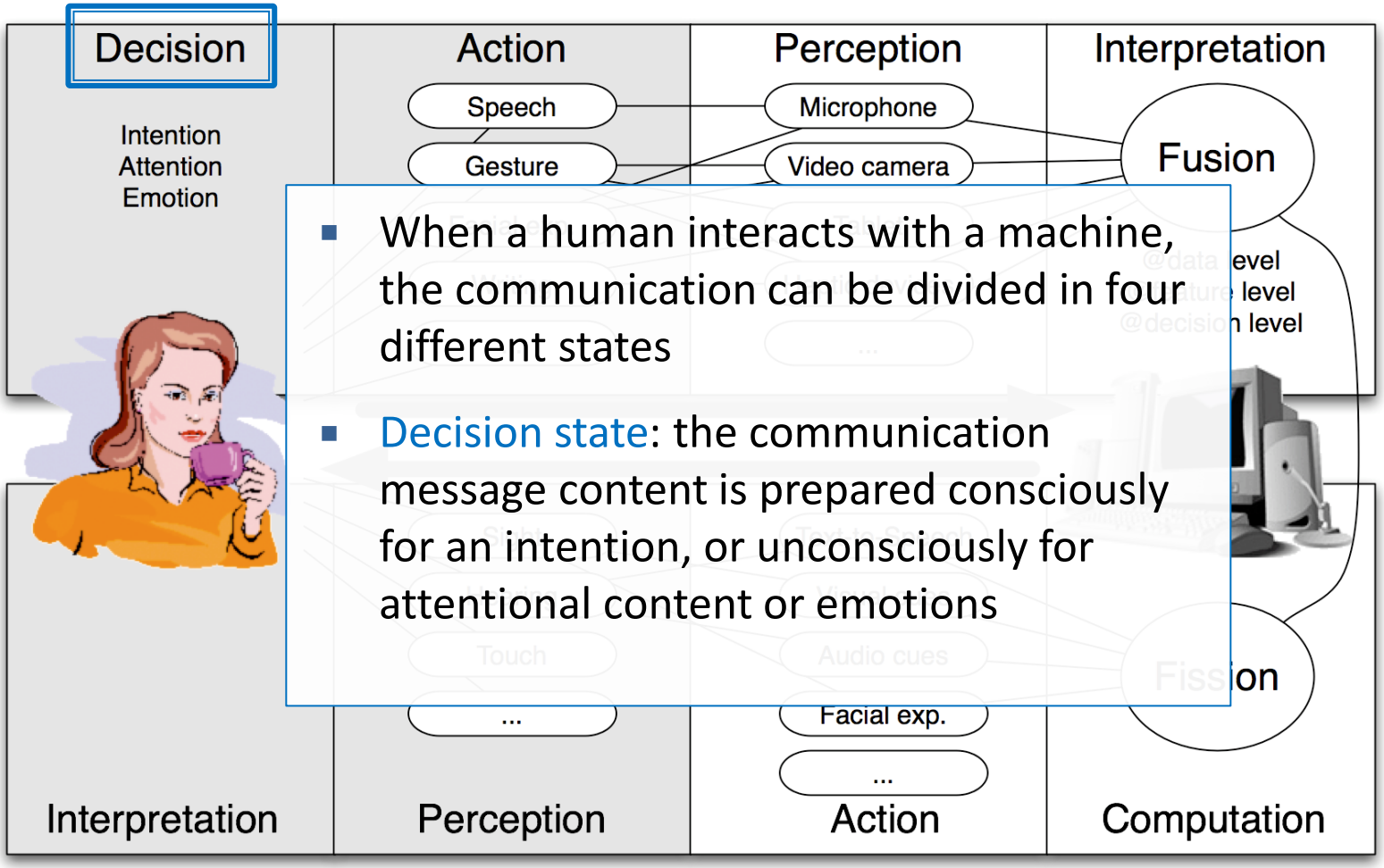
Cognitive foundations

- Findings in cognitive psychology reveal that:
  - humans tend to reproduce interpersonal interaction patterns during multimodal interaction with a system
  - humans are able to process modalities partially independently and, thus, presenting information with multiple modalities increases human working memory
    - effective working memory increased by presenting information in a dual-mode form rather than a purely visual one, could expand processing capabilities (Mousavi et al., 1995)
  - human performance is improved when interacting multimodally due to the way human perception, communication, and memory function
    - For example, when processing both auditory and visual information during speech, a listener is able to extract a higher rate of lexical intelligibility (Grant & Greenberg, 2001)

The multimodal man-machine interaction loop (1/6)

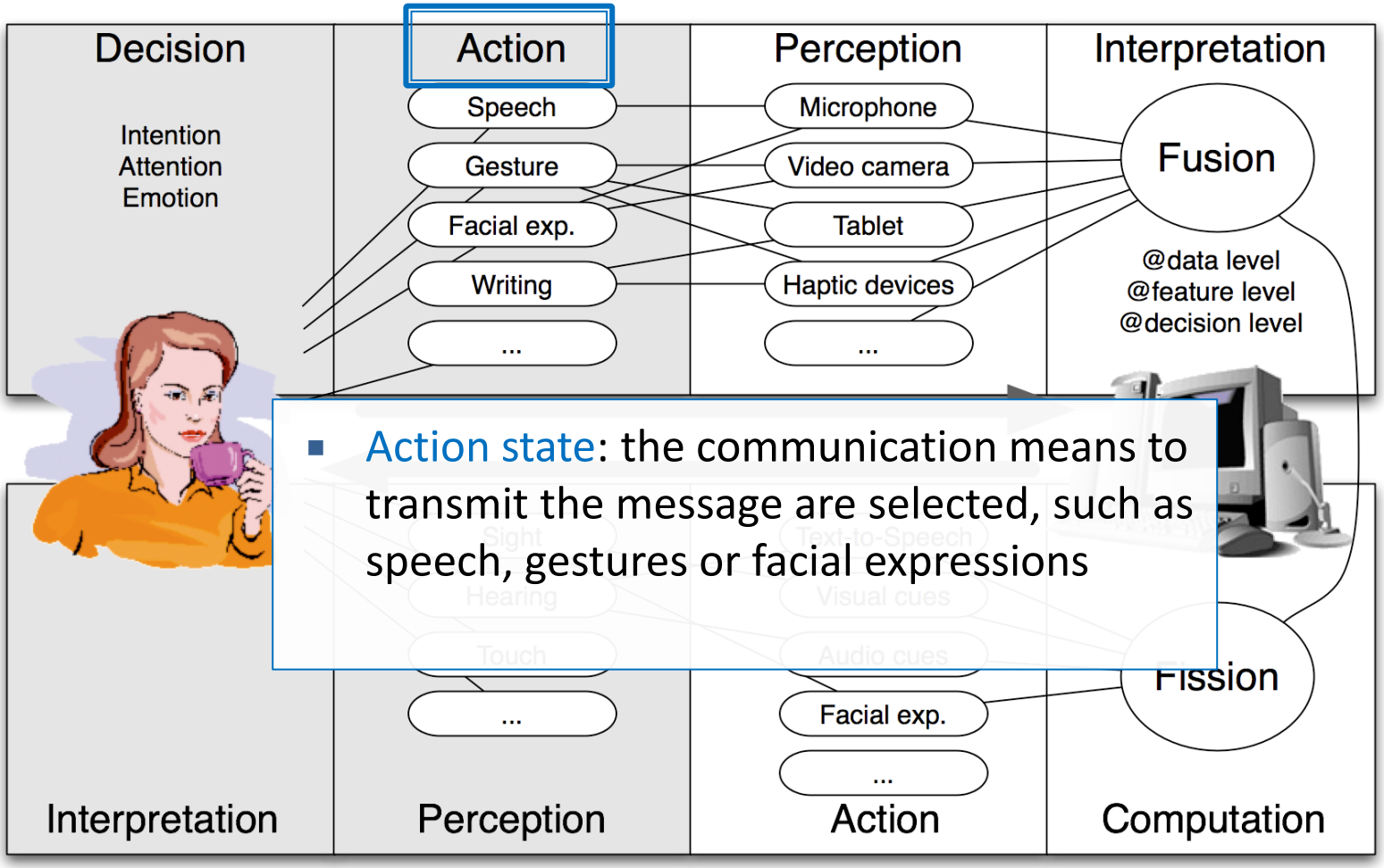


The multimodal man-machine interaction loop (2/6)



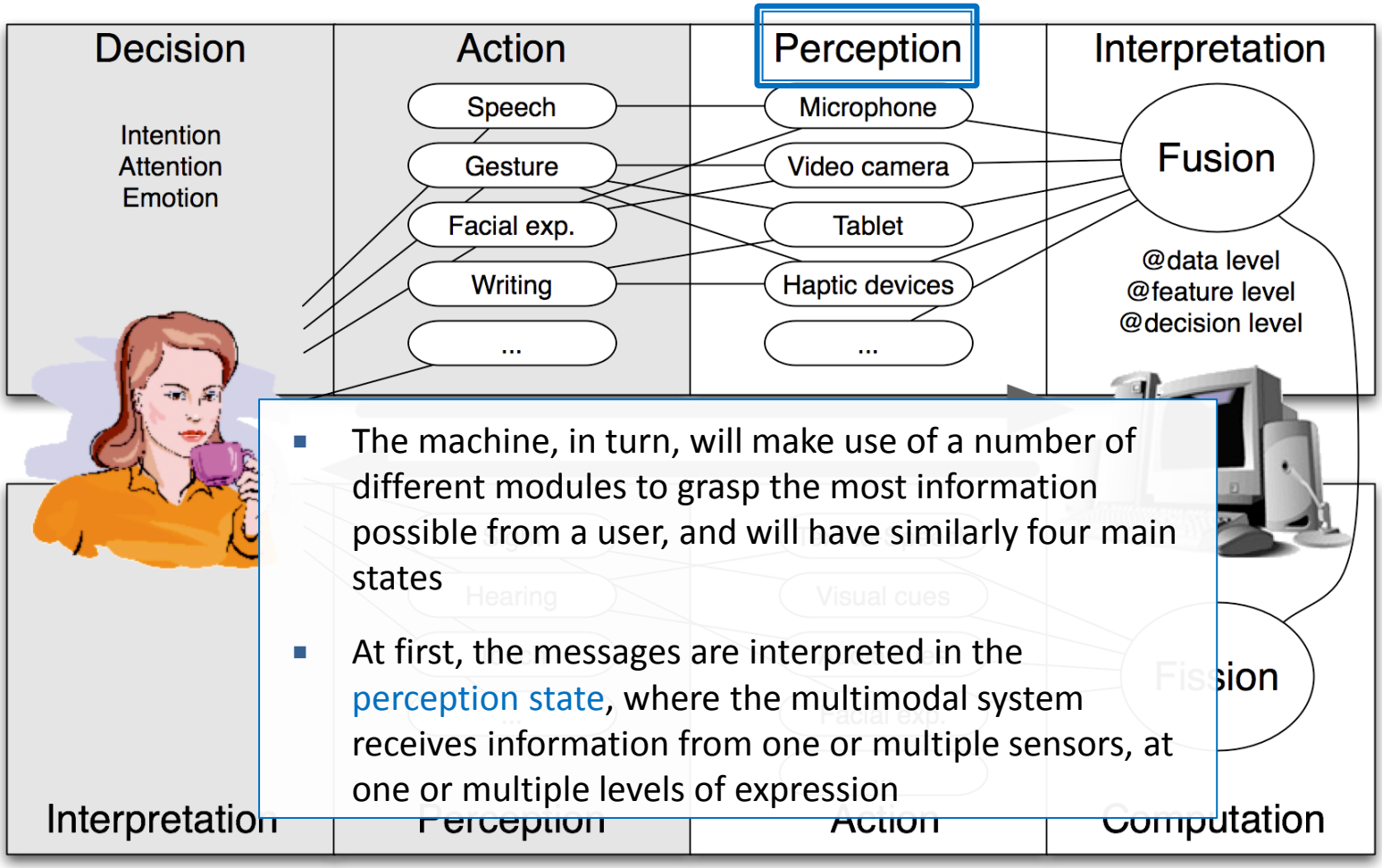
- When a human interacts with a machine, the communication can be divided in four different states
- Decision state: the communication message content is prepared consciously for an intention, or unconsciously for attentional content or emotions

The multimodal man-machine interaction loop (3/6)



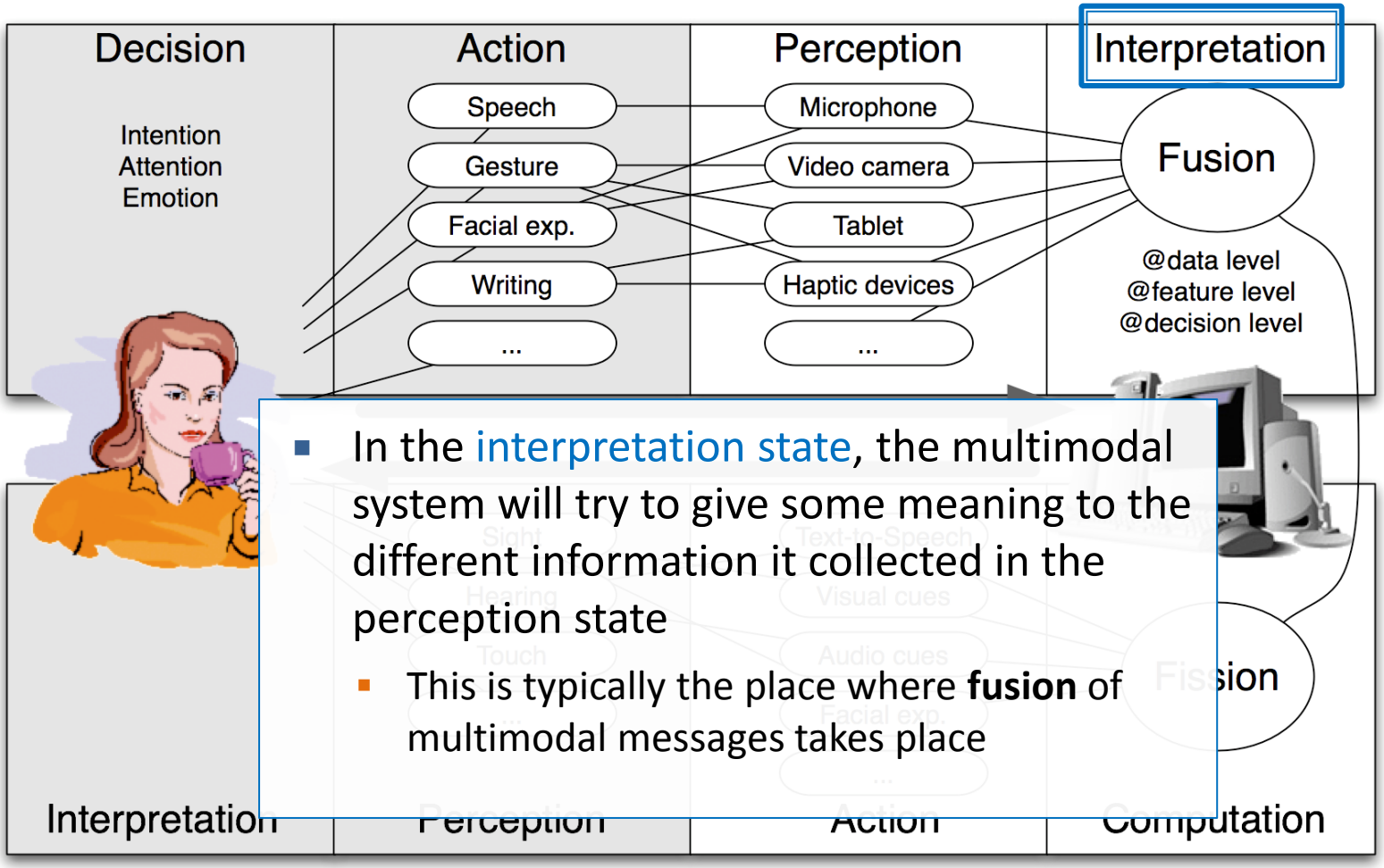
- Action state: the communication means to transmit the message are selected, such as speech, gestures or facial expressions

The multimodal man-machine interaction loop (4/6)



- The machine, in turn, will make use of a number of different modules to grasp the most information possible from a user, and will have similarly four main states
- At first, the messages are interpreted in the perception state, where the multimodal system receives information from one or multiple sensors, at one or multiple levels of expression

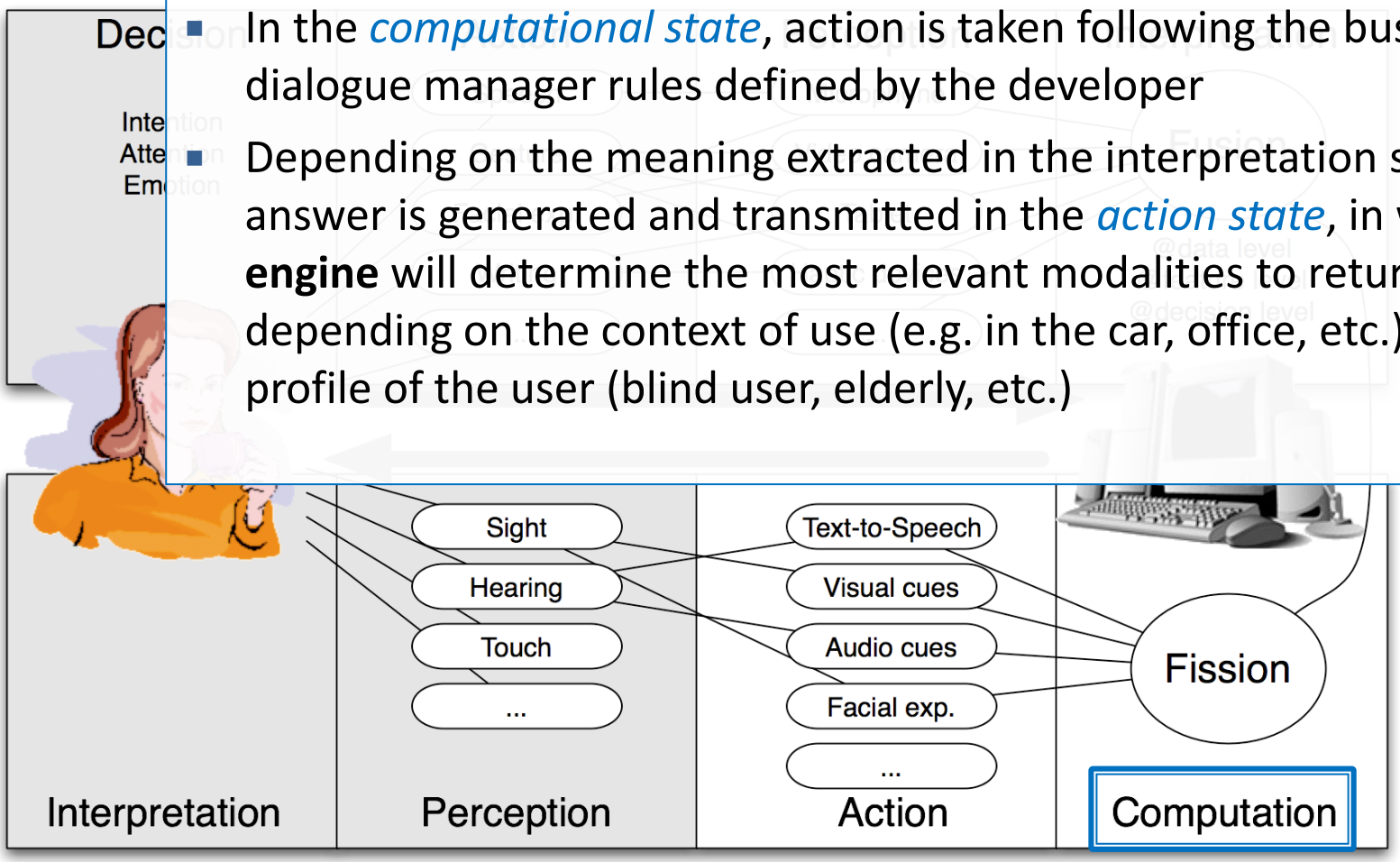
The multimodal man-machine interaction loop (5/6)



- In the interpretation state, the multimodal system will try to give some meaning to the different information it collected in the perception state
  - This is typically the place where **fusion** of multimodal messages takes place

The multimodal man-machine interaction loop (6/6)

- In the *computational state*, action is taken following the business logic and dialogue manager rules defined by the developer
- Depending on the meaning extracted in the interpretation state, an answer is generated and transmitted in the *action state*, in which a **fission engine** will determine the most relevant modalities to return the message, depending on the context of use (e.g. in the car, office, etc.) and the profile of the user (blind user, elderly, etc.)



Input modalities

Modality combinations

- Historically, the main trend has focused on pointing and speech, combined using:
  - Speech and (mouse or touchpad) pointing
  - Speech and pen input
  - Speech and manual gesturing
  - Speech and lip movements
  - Speech and gaze tracking
- Later multimodal interfaces evolved beyond pointing into richer interaction, allowing users to produce symbolic gestures such as arrows and encircling
- Another direction in multimodal research has been speech/lip movement integration

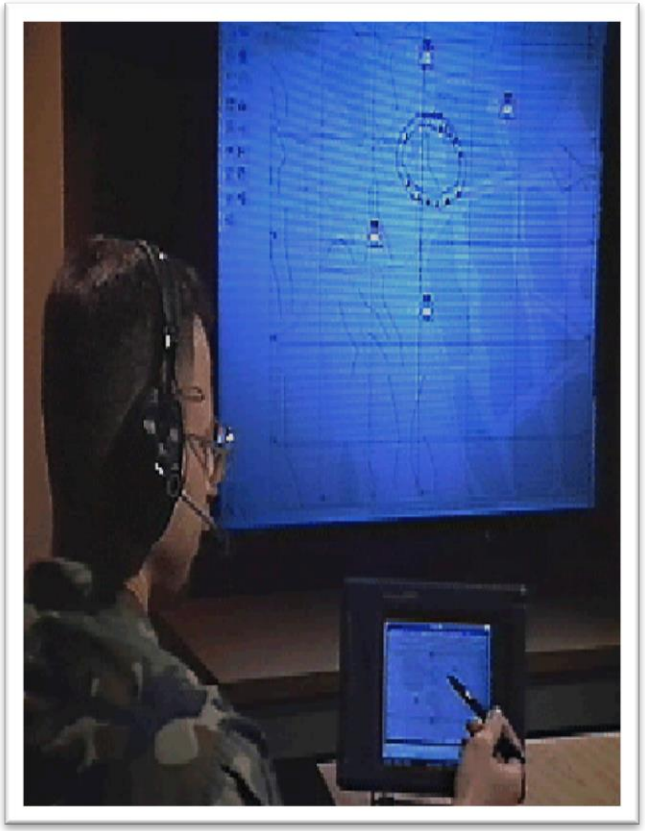
Speech and pointing

Put-That-There Demo:

http://www.youtube.com/watch?v=0Pr2KIPQOKE

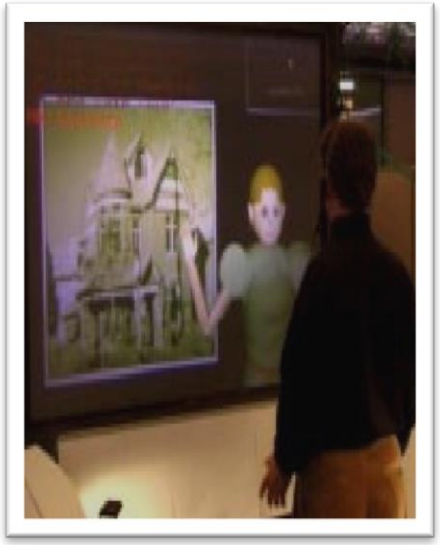
- Put That There Interface (Bolt, 1980)
  - Users sat in front of a projection of "Dataland" in "the Media Room"
  - Using the Put That There interface they could use speech and pointing to create and move objects on a 2-D large-screen display

Speech and pen input



- Spoken language is sometimes processed along with complex pen-based gestural input
  - Involving hundreds of different symbolic interpretations beyond pointing
- QuickSet System (Cohen et al, 1997)
  - Runs on a hand-held PC
  - Multimodal command to place the correct number, length and orientation of aircraft landing strips on a map
    - "Airstrips… facing this way <draws arrow> and facing this way <draws arrow>

Speech and manual gesturing

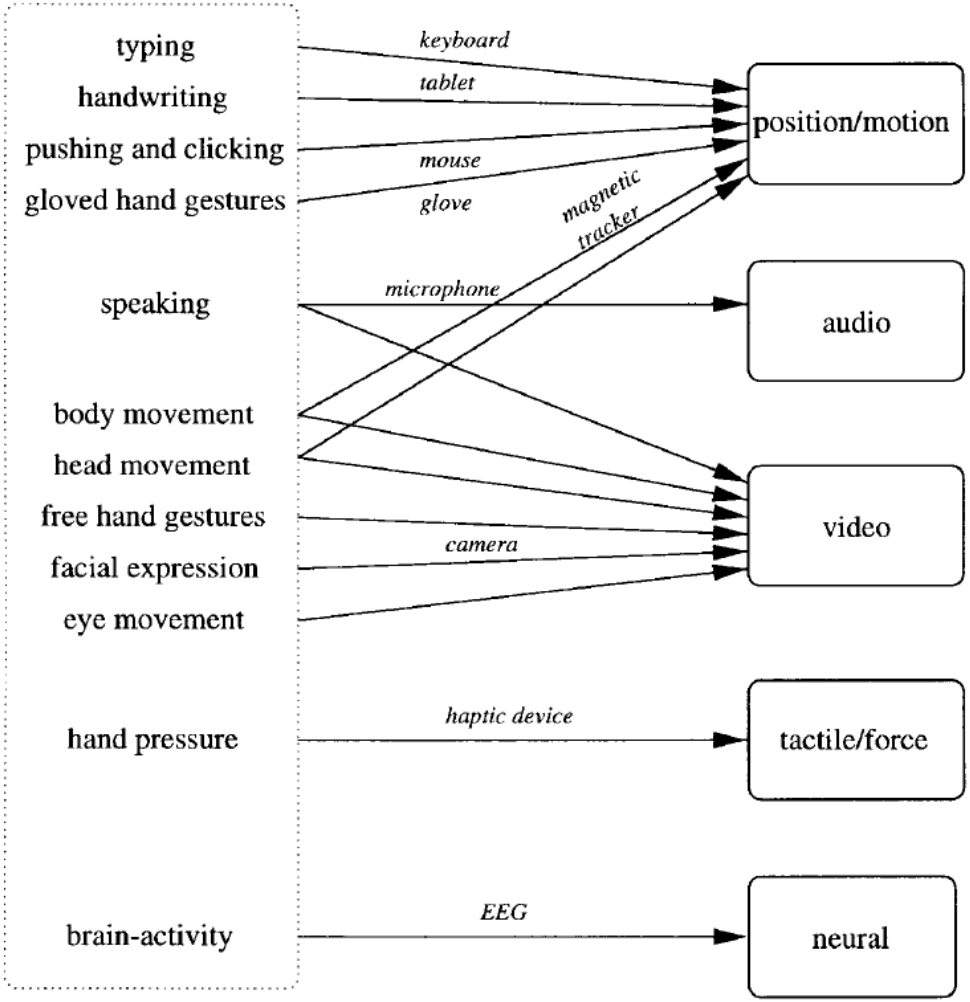


- At present, 3D hand and arm movements are most often considered for pointing at large screens, especially in immersive virtual reality environments
  - e.g., reality centers or caves
- Cameras are currently able to track and interpret pointing movements with fair accuracy
- Sensors or specific pointing equipments (e.g., data gloves, magic wands) can also be used for this purpose

Speech and lip movements

- Spoken language is processed along with corresponding human lip movements
- Research has been driven by cognitive science interest in intersensory audiovisual perception and the coordination of speech output with lip and facial movements
  - This kind of work has included classification of human lip movement (visemes) and the viseme-phoneme mappings that occur during articulated speech
- Recently, researchers in this area have begun to explore adaptive techniques for improving system robustness in noisy environmental contexts

Mapping of modalities



- Mapping of different human-action modalities to computer-sensing modalities for HCI
  - Multiple human actions, such as facial expressions and hand or eye movement, can be sensed through the same "devices" and used to infer different information

Source: Sharma et al. (1998)

Modality properties

Property	Example
Required interaction channels	•spoken language is conveyed auditively, written text visually
Salience	•auditive signals are more attention catching than visual ones
Local selectivity	•visual data are perceived only if they are paid attention to
Degree of user control	•static modalities like written text allow more user control over pacing than dynamic modalities such as videos or spoken text
Learning requirements	•arbitrary modalities such as newly defined symbols require more learning efforts than those building upon existing conventions
Expressiveness	•analogous modalities such as graphics are preferred for conveying spatial relationships whereas linguistic modalities like text convey conceptual information such as detailed descriptions better

Input modes integration

- The ability to develop multimodal systems depends on knowledge of the natural integration patterns that typify people's combined use of different input modes
- The design of new multimodal systems
  - depends on knowledge of
    - the properties of different modes and the information content they carry
    - the unique characteristics of multimodal language and its processability
    - the integration and synchronization characteristics of users' multimodal interaction
  - It also relies on accurate prediction of
    - when users are likely to interact multimodally
    - how alike different users are in their specific integration patterns

Modality Selection According to Task Properties

- The first step in designing multimodal interactive systems is to elicit interaction modalities that are appropriate for the current task, based on the properties of each modality

Input							Output					
	Speech Input	Typing	Handwriting	Pointing	Free Gestures	Eye movements		Auditive Output	Visual Output	Haptic Output	Motor Input	Speech Input
Sketching	–	–	–	+	?	?	Urgent Information	+	?	?		
Graphic manipulation	?	?	?	+	?	?	Highly current information	+	?	?		
Selecting (small sets)	+	?	?	+	+	?	Status information	–	+	?		
Selecting (large sets)	+	+	?	?	?	?	Private information	–	+	+	+	–
Text input	?	+	+	?	?	?	Security relevant visual primary task	+	–		–	+

Exemplary task-modality matrix

Interaction Constraints Based on Context of Use (1/3)

- After selecting (several alternative) task appropriate modalities, the designer has to check further interaction constraints imposed by
  - user characteristics,
  - device characteristics
  - and the environment
- The designer first checks which interaction modalities are most appropriate for the tasks to be supported by the system
- Then he checks whether for each individual candidate modality, the factors listed in the columns outweigh the factors listed in the rows and contrasts these results for each interaction modality

Interaction Constraints Based on Context of Use (2/3)

			Speech output preferred						Audio tones preferred					Visual text preferred			Graphics preferred		
			Illiterate people	Visually impaired people	Car user interfaces	Phone user interfaces	Smartphones, PDAs	Bad lighting conditions	Visually impaired people	Car user interfaces	Phone user interfaces	Smartphones, PDAs	Bad lighting conditions	Hearing impaired people	Noisy environments	Public, crowded environments	Hearing impaired people	Noisy environments	Public, crowded environments
Speech output	discou-raged	Hearing impaired people	X	X	X	X	X	X											
		Noisy environments	X	X	X	X	X	X											
		Public / crowded environments	X	X	X	X	X	X											
audio tones	discou-raged	Hearing impaired people							X	X	X	X	X						
		Noisy environments							X	X	X	X	X						
		Public / crowded environments							X	X	X	X	X						
Visual text	impos-sible	Illiterate People												X	X	X			
		Car user interfaces												X	X	X			
	discou-raged	Visually impaired people												X	X	X			
		Bad lighting conditions												X	X	X			
Graphics	discou-raged	Visually impaired people															X	X	X
		Car user interfaces															X	X	X
		Bad lighting conditions															X	X	X

Exemplary output constraints

- crossed out squares indicate cases of conflicting usability goals

Interaction Constraints Based on Context of Use (3/3)

			Speech input preferred							Typing preferred			Hand-writing preferred		Pointing preferred					
			Illiterate people	Visually impaired people	Motor impaired people	Car user interfaces	Phone user interfaces	Smartphones, PDAs	Bad lighting conditions	People with speech disorders	Noisy environments	Public and crowded environments	People with speech disorders	Smartphones, PDA, Tablets	People with speech disorders	Public information Kiosk	Phone user interfaces	Smartphones, PDAs	Noisy Environments	Public environments
Speech input	discouraged	Hearing impaired people	X	X	X	X	X	X	X											
		People with speech disorders	X	X	X	X	X	X	X											
		Noisy environments	X	X	X	X	X	X	X											
		Public / crowded environments	X	X	X	X	X	X	X											
Typing	impossible	Illiterate people								X	X	X								
		Car user interfaces								X	X	X								
	discouraged	Visually impaired								X	X	X								
		Motor impaired								X	X	X								
		Smartphones, PDAs								X	X	X								
		Glaring environments								X	X	X								
Handwriting	impossible	Illiterate People											X	X						
		Motor impaired People											X	X						
	discouraged	Visually impaired people											X	X						
		Car user interfaces											X	X						
Pointing	discouraged	Visually impaired people													X	X	X	X	X	X
		Motor impaired people													X	X	X	X	X	X
		Car user interfaces													X	X	X	X	X	X
		Bad lighting conditions													X	X	X	X	X	X

Exemplary input constraints

- crossed out squares indicate cases of conflicting usability goals

Design Guidelines

Guidelines for multimodal UI design (1/2)

- Multimodal systems should be designed for the broadest range of users and contexts of use, since the availability of multiple modalities supports flexibility
- Designers should take care to address privacy and security issues when creating multimodal systems
- Modalities should be integrated in a manner compatible with user preferences and capabilities

Guidelines for multimodal UI design (2/2)

- Multimodal systems should be designed to adapt easily to different contexts, user profiles and application needs
- Error prevention and handling is a major advantage of multimodal interface design, for both user- and system-centered reasons
  - Specific guidelines include integrating complementary modalities to improve system robustness, and giving users better control over modality selection so they can avoid errors

Ten myths about multimodal interaction (1/2)

1. If you build a multimodal system, users will interact multimodally
2. Speech & pointing is the dominant multimodal integration pattern
3. Multimodal input involves simultaneous signals
4. Speech is the primary input mode in any multimodal system that includes it
5. Multimodal language does not differ linguistically from unimodal language

Ten myths about multimodal interaction (2/2)

6. Multimodal integration involves redundancy of content between modes
7. Individual error-prone recognition technologies combine multimodally to produce even greater unreliability
8. All users' multimodal commands are integrated in a uniform way
9. Different input modes are capable of transmitting comparable content
10. Enhanced efficiency is the main advantage of multimodal systems

Applications

Application areas

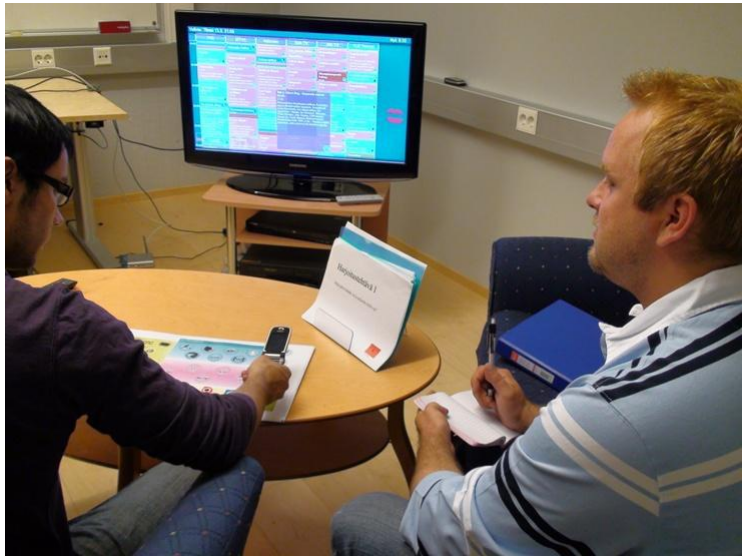
- The literature on applications that use MMHCI is vast
- Some of the major areas are:
  - Ambient spaces
  - Mobile / wearable
  - Virtual environments
  - Art
  - Applications for users with disabilities
- This section introduces recent multimodal applications

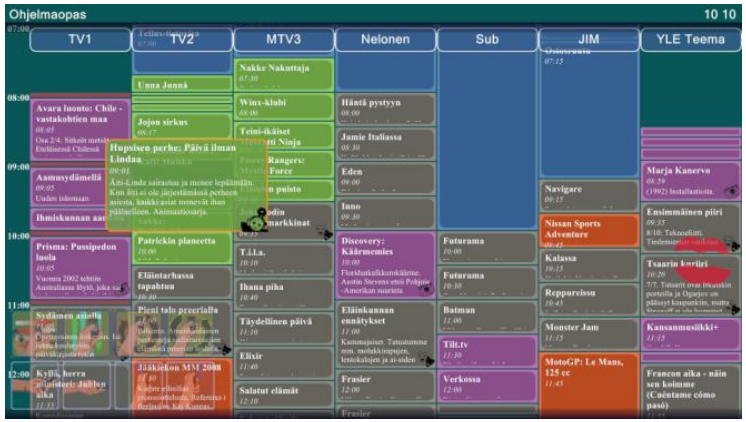
Manual deskterity | Pen & Touch

Manual Deskterity: http://www.youtube.com/watch?v=9sTgLYH8qWs

- A prototype digital drafting table that supports both pen and touch input
- The pen writes, touch manipulates, and the combination of pen+touch yields new tools
- The richness of the gestures arises from how the primitives are combined
  - For the pen the primitives we use are tap, drag-off, crossing, or drawing a stroke
  - For multi-touch, single-finger tap, single-finger hold (i.e. a tap with a long duration), holding with thumb and forefinger, and crossing have been employed

Multimodal media center | Speech & Touch



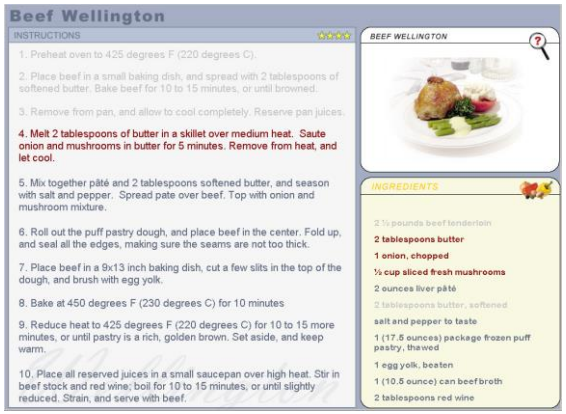


- The application is based on a combination of a large high-definition display and a mobile phone
  - Users can interact with the system using speech input (speech recognition), physical touch (touching physical icons with the mobile phone), and gestures
  - Gestures are used together with the mobile phone keypad. Here, the different orientations of the mobile phone alter how the keypad works
    - In the vertical/down orientation keys are used to move selection
    - in the horizontal orientation keys perform zooming functions
    - a third position is used as a replacement to push-to-talk
  - The media center user interface consists of several screens for different media applications (e.g., photographs and music). The interfaces consists of a grid, where columns represent television channels, while rows represent time slots. Cells are individual television programs.

Multimodal Multiplayer Tabletop Gaming | Speech & Gestures

- A multimodal multiplayer gaming environment has been created that allows players to interact directly atop a digital table via speech and rich whole-hand gestures
- Multiplayer multimodal wrappers were built atop of two commercial single-player games: Warcraft III (a command-and-control strategy game) and The Sims (a simulation game)
- A gesture is not only recognized as input, but is easily understood as a communicative act providing explicit and consequential information of one's actions to the other players
- Likewise, speech is intended not only for the control of the system, but also for the benefit of one's collaborators
- Combination of speech and gestures: all gestures indicate locations not provided by the speech act

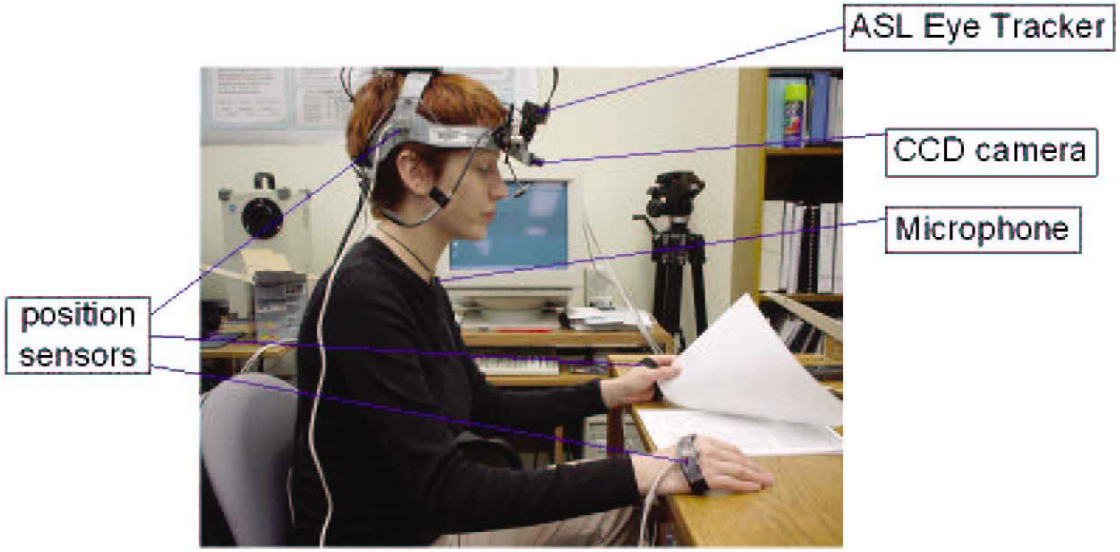
eyeCOOK | eye gaze & speech



- eyeCOOK is a multimodal attentive cookbook to help a non-expert computer user cook a meal
- The user communicates using eye-gaze and speech commands, and eyeCOOK responds visually and/or verbally, promoting communication through natural human input channels without physically encumbering the user
- Once the user begins the actual process of cooking, the system automatically sets timers and reminders
- The relationship between current, previous and future cooking steps and the ingredients they involve is implicit via a simple color-coding scheme
- This dynamic colouring provides a cue situating the user's current cooking task within the recipe as a whole

A Multimodal Learning Interface | Speech & gaze, head direction, hand movement

- The learning system is trained in an unsupervised mode in which users perform everyday tasks while providing natural language descriptions of their behaviors
- The system collects acoustic signals in concert with user-centric multisensory information from non-speech modalities, such as user's perspective video, gaze positions, head directions, and hand movements
- A multimodal learning algorithm uses this data to first spot words from continuous speech and then associate action verbs and object names with their perceptually grounded meanings
- The central ideas are to make use of nonspeech contextual information to facilitate word spotting, and utilize body movements as deictic references to associate temporally cooccurring data from different modalities and build hypothesized lexical items



Other applications (1/2)

	Gesture	Head/ Body	Speech /Voice	Pen	Gaze	Touch	Pointing
XTRA (Kobsa et al., 1986)	■		■				
Visual environment for use in CAD and teleoperation (Weimer & Ganapathy, 1989)	■		■				
Multi-user virtual world (Codella et al., 1992)	■		■				
Charade (Baudel et al., 1993)	■						■
Finger-Pointer (Fukumoto et al. 1994)	■		■				
Jeanie (Vo & Wood, 1996)			■	■			
GALAXY (Seneff et al., 1996)			■				■
Hands-free computer control (Malkewitz, 1998)		■	■				
VizSpace (Lucente et al., 1998)	■	■	■				
MVIEWS (Cheyer & Julia, 1998)			■	■			
MSVT multimodal scientific visualization tool (Laviola 2000)	■		■				

Other applications (2/2)

	Gesture	Head/ Body	Speech /Voice	Pen	Gaze	Touch	Pointing
Interactive audience participation (Maynes-Aminzade et al., 2002)		■					■
MUST tourist guide (Almeida et al., 2002)			■	■			
XISM & DAVE_G (Sharma et al., 2003)	■		■				
XWand (Wilson & Shafer, 2003)	■						■
3D Multimodal Interaction in AR and VR (Kaiser et al, 2003)	■		■				
Eyes-Free' interaction techniques for wearable devices (Brewster et al., 2003)	■	■					
Multiparty conversation analysis (Otsuka et al., 2007)		■	■		■		
Speech and Gaze Control for Realistic Desktop Environments (Castellina et al., 2008)			■		■		

Multimodal Interfaces & Universal Access

- Increased range of available modalities and styles of multimodal interaction compensates for the diversity of **physical disabilities**
- **Flexible** use of input modes
- Improved **efficiency**, especially when manipulating graphical information
- Improved **error handling**, both in terms of error avoidance and graceful recovery from errors

Multimodality and affective HCI (1/2)

- Most current MMHCI systems do not account for the fact that human–human communication is always socially situated and that we use emotion to enhance our communication
- HCI systems that can sense the affective states of the human (e.g., stress, inattention, anger, boredom, etc.) and are capable of adapting and responding to these affective states are likely to be perceived as more natural, efficacious, and trustworthy
- On the other hand, emotion is intricately linked to other functions such as attention, perception, memory, decision-making, and learning. This suggests that it may be beneficial for computers to recognize the user's emotions and other related cognitive states and expressions

Multimodality and affective HCI (2/2)

- Few research efforts towards implementing **a multimodal affective analyzer**
- Studies in psychology suggest that the **combined face and body approaches** are the most informative
- Facial expression recognition
  - Most works in this area use still photographs or video sequences without speech
- Vocal emotion recognition
  - Works on this area often use only audio information
- How much information does the face, as compared to speech, and body movement, contribute to natural interaction?
  - Most experimenters suggest that the face is more accurately judged, produces higher agreement, or correlates better with judgments based on full audiovisual input than on voice input

Mulltimodality and Ambient Intelligence

In AmI environments

- A **great variety** of input and output modalities, or **combinations** of modalities, is necessary
- Interaction with interfaces has to be **flexible**
  - as it takes place in various physical settings and while users are engaged in some daily routine activity (e.g., driving or cooking)
- The new contexts of use:
  - **restrict** the availability of the user's **perceptual and motor** resources
  - induce **constraints** upon input and output media and modality choices
- The volume and content of information exchanges between user and system have to be tailored to fit the specific throughput and expressive power of available interaction devices and modalities

References

Dumas, B., Lalanne, D., & Oviatt, S. (2009). Multimodal interfaces: A survey of principles, models and frameworks. In *Human Machine Interaction* (pp. 3-26). Springer Berlin Heidelberg. *(Available through the course website)*

Jaimes, A., & Sebe, N. (2007). Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, *108*(1), 116-134. *(Available through the course website)*

Additional Reading

- Bradbury, J. S., Shell, J. S., & Knowles, C. B. (2003, April). Hands on cooking: towards an attentive kitchen. In *CHI'03 extended abstracts on Human factors in computing systems* (pp. 996-997). ACM.
- Carbonell, N. (2006). Ambient multimodality: towards advancing computer accessibility and assisted living. Universal Access in the Information Society, 5(1), 96-104. (Available through the course website)
- Hinckley, K., Yatani, K., Pahud, M., Coddington, N., Rodenhouse, J., Wilson, A., Benko, H. and Buxton, B. (2010). Manual deskterity: an exploration of simultaneous pen + touch direct input. In Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems (CHI EA '10). ACM, New York, NY, USA, pp. 2793-2802
- Oviatt, S. (2007). Multimodal Interfaces. In J. Jacko and A. Sears (eds.) The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, Taylor & Francis Group, pp. 413 – 432
- Ratzka, A. (2008). Steps in Identifying Interaction Design Patterns for Multimodal Systems. In *Engineering Interactive Systems* (pp. 58-71). Springer Berlin Heidelberg.
- Sharma, R., Pavlovic, V. I., & Huang, T. S. (1998). Toward multimodal human-computer interface. Proceedings of the IEEE, 86(5), 853-869
- Tse, E., Greenberg, S., Shen, C. and Forlines, C. (2007). Multimodal Multiplayer Tabletop Gaming. In ACM Computers in Entertainment, 5 (2), Article 12
- Turunen, M., Kallinen, A., Sànchez, I., Riekki, J., Hella, J., Olsson, T., Melto, A., Rajaniemi, J.-P., Hakulinen, J., Mäkinen, E., Valkama, P., Miettinen, T., Pyykkönen, M., Saloranta, T., Gilman, E., and Raisamo, R. (2009). Multimodal interaction with speech and physical touch interface in a media center application. In Proceedings of the International Conference on Advances in Computer Enterntainment Technology (ACE '09). ACM, New York, NY, US
- Yu, C., & Ballard, D. H. (2004). A multimodal learning interface for grounding spoken language in sensory perceptions. ACM Transactions on Applied Perception (TAP), 1(1), 57-80

The End

Questions?