Modalities and Multimodalities Introduction

Domain

Definitions

Challenges

1

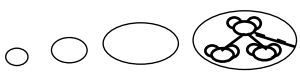
Introduction

Man-Machine Interface
Handling multimodal interaction

A modality

A multimodal system

Software architecture model for multimodal systems



Fusion of different objects from various modelling techniques:
How ?

At which level of abstraction?

2

Domain and definitions

- Beyond the traditional User Interface (UI)
  - Windows: scroll, resize, move
  - Icons: representations, drag/drop
  - Menus: pop-up, pull-down
  - Pointers: mouse, digitizer, trackball, etc.
- Multimodal systems
  - Multi-modal refers to interfaces that support non-GUI interaction
  - Speech and pen input are two common examples - and are complementary

3

Domain and definitions

- Multimodal systems
  - Multi-Sensori-Motor Systems
  - extend the sensori-motor capabilities of computer systems

4

Domain and definitions

"New Interfaces" extend the sensori-motor capabilities of computer systems

Multimodal ≠ Multimedia

Multimodal ≠Speech interface

New interaction capabilities appear

5

Media - Modality

- Media
  - material (signal on a channel)
  - the support of communication
- Modality
  - a channel or path of communication between the human and the computer
    - sensorial (audition, vision, etc.)
    - of communicating (voice, gestures, facial expressions, etc.)
  - A modality is a process of receiving and producing chunks of information

6

Multimedia - Multimodality

- Multimedia system
  - transport signals of different kinds
    - For ex.: a sound clip attached to a presentation
- Multimodal system
  - interpret signs belonging to various sensory and communication modalities
    - For ex.: the combined input of speech and typing in a word processor

7

Multimodality

- Multimodality is the use of two or more of the five senses for the exchange of information
- A multimodal system represents and manipulates information from different human communication channels at multiple levels of abstraction

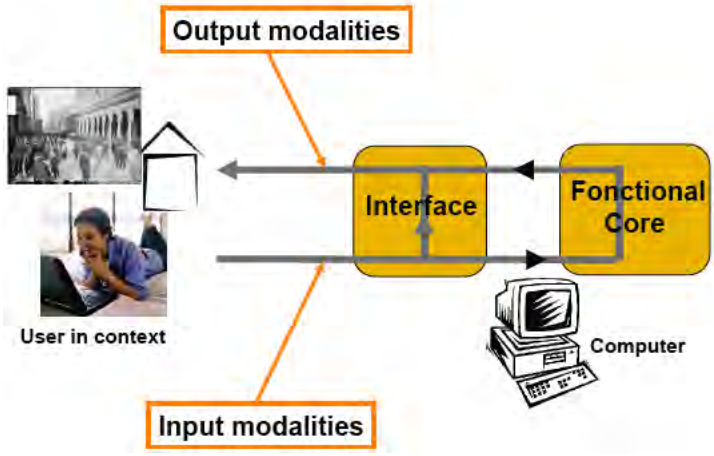
8

Multimodal and crossmodal

- Multimodal interaction makes use of several input and/or feedback modalities in interacting with a computer system.
  - Examples of modalities: manual gestures, gaze, touch, speech, head & body movements
  - Modality: human sensory channel, different representation modality, or different input method
- Crossmodal interaction makes use of a different human sensory modality to present information typically presented through another modality.

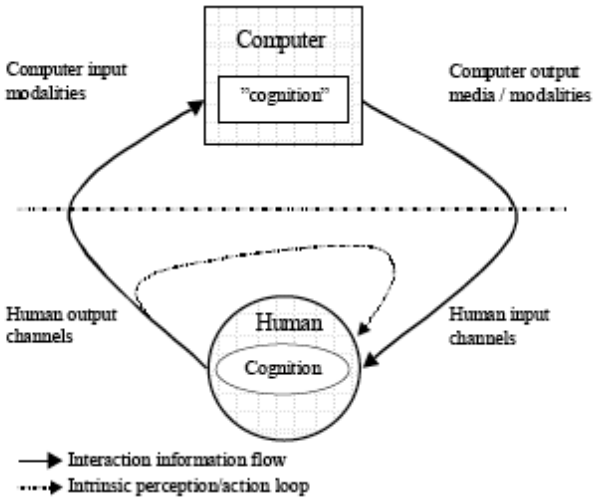
9

Input/Output modality



10

Multimodal interaction



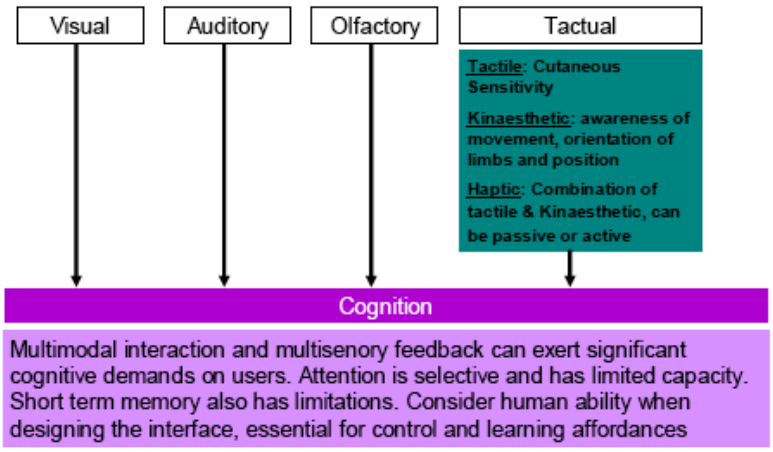
11

Human senses and modalities

Sensory perception	Sense organ	Modality
Sense of sight	Eyes	Visual
Sense of hearing	Ears	Auditive
Sense of touch	Skin	Haptic
Sense of smell	Nose	Olfactory
Sense of taste	Tongue	Gustatory
Sense of balance	Organ of equilibrium	Vestibular

12

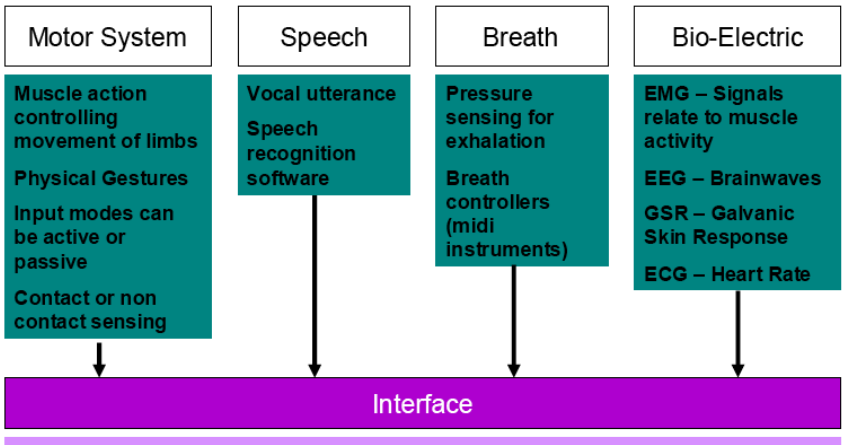
Human perceptual modalities (Multimodal output interfaces)



Multimodal interaction and multisenory feedback can exert significant cognitive demands on users. Attention is selective and has limited capacity. Short term memory also has limitations. Consider human ability when designing the interface, essential for control and learning affordances

13

Human output modalities (Multimodal input interfaces)



Passive (inherent) System Feedback – e.g. A 'click' from pressing a button, not used by the system

Active System Feedback – Those sounds & images etc, produced by the system as a result of human control (active or passive input modes)

14

Multimodal (MM) versus GUI

- GUI interfaces often restrict input to single non-overlapping events, while MM interfaces handle all inputs at once
- GUI events are unambiguous, MM inputs can be based on recognition

15

Why multimodal?

- Most technologies are mature
- Seek to optimize the distribution of information over different modalities
- For adaptive, cooperative and flexible interaction among people

16

Why multimodal?

- Naturalness
  - provide more "natural" interfaces Usability
- Usability / flexibility
  - improve ease-of-use
- Robustness/Efficiency/Accuraccy
  - decrease error rates (Mutual disambiguation of recognition errors)
- Perception
- Relieve burden on the visual channel
- Support users with disabilities

17

Natural interaction and multimodality

- Natural interaction is the long-term goal of being able to communicate with machines in the same ways in which humans communicate with one another
  - Input/output audiovisual speech, facial expression, gesture, gaze, body posture, physical action, touch, etc.
- **Natural interaction is multimodal by nature**

18

Why multimodal?

- Flexibility for Robutness
  - Advantages for error recovery
    - Users intuitively pick the modality that is less error-prone
    - Language is often simplified
    - Users intuitively switch modality after an error, so that the same problem is not repeated
- Flexibility for
  - Users with disability (permanent or temporary)
  - Variable usage context (**mobile support, ubiquitous computing**)
- The flexibility of a multimodal interface can accommodate **a wide range of users, tasks, and environments** for which any given single mode may not suffice

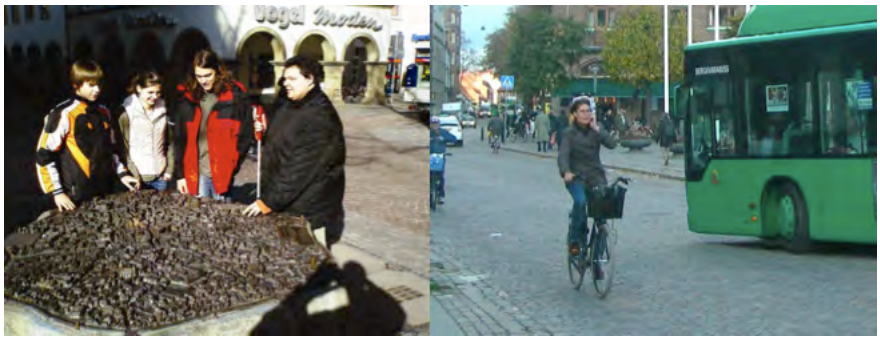
19

Input Multimodality

- Because of the user's circumstances – including her task, her background, her training, her knowledge, and the physical and interactive behaviour of the computer interface – the user may well have preferences as to how she communicates with the computer.
  - A familiar example is that if the user is engaged in a task which occupies her hands, she may prefer to use speech.
  - Another example: Suppose that the user wishes to book a flight from somewhere in Europe to Las Vegas. She may not know what is the nearest international airport, so she would prefer to indicate her destination by pointing on a map – or at the very least, by choosing from an appropriately filtered list of airports.

Why multimodal?

- What do these persons have in common?



21

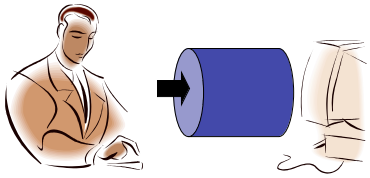
Why multimodal?

- Enabling the user
- New multimodal technologies enable the user to be better engaged in the interaction to receive more information through several modalities
- Multimodal interaction makes using of information technology possible for people with special needs, e.g., for blind and visually impaired people

22

Why multimodal?

- The combination of human output channels effectively (multimodal input interaction) increases the bandwidth of the human machine channel.
  - *This has been discovered in many empirical studies of multimodal human computer interaction*



23

Why multimodal? Nevertheless…

- **Input multimodality**
- Adding extra input modality requires more neurocomputational resources and will lead to deteriorated output quality resulting in reduced effective bandwidth.
- Two types of effects are usually observed:
  - a slowdown of all output processes, and
  - interference errors due to the fact that attention cannot be divided between the number of channels.
- Two examples of this: writing when speaking, and speaking when driving a car.

24

Why multimodal? Nevertheless…

- **Input multimodality**
- Example: Two handed-interaction
  - Psychological Theory -Kinematic chain – Y. Guiard
    - Right-to-left reference: The right hand performs its motion relative to the frame of reference set by the left hand
    - Asymmetric scales: Different temporal-spatial scales of motion
    - Left hand precedence: The left hand precedes the right: for example, the left hand first positions the paper, then the right hand begins to write
    - Right hand preference: Is the one finishing the action, touching the world

25

Why multimodal? Nevertheless…

- **Output multimodality**
- The designers of computer interfaces exploit the power of vision, in making maximum use of visual displays
- With more thought, there might be better ways of presenting information, ways that will not increase the visual complexity and the user's task load
- Goal of output multimodal interaction
  - It is to be expected that designers are aware of the possibility of using different output channels when appropriate

  => Human perception

26

Human multisensory perception

- Humans have several different senses through which information about the environment is obtained
- Information from different senses interact with each other to form the integrated representation of objects and events

From Jukka Raisamo and Roope Raisamo, Tampere Univ.

Human multisensory perception

- Sensory modalities
  - No information processing system is powerful enough to perceive and act accurately under all conditions
  - If a single modality is not enough to come up with a robust estimate, information from several modalities can be combined

From Jukka Raisamo and Roope Raisamo, Tampere Univ.

Human multisensory perception: sensory combination

- The human brain reconstructs the environment from the incoming streams of – often ambiguous – sensory information and generates unambiguous interpretations of the world
- To do so many different sources of sensory information are constantly processed, analyzed and **combined**
  - Moving train illusion:
    - Is it your train or the other train that is moving?
    - The brain collects more and more information about the perceptual event and finally resolves the ambiguity

From Jukka Raisamo and Roope Raisamo, Tampere Univ.

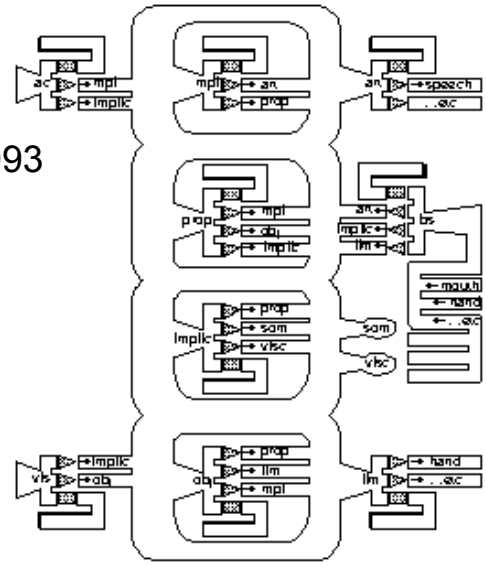
Human multisensory perception: sensory integration

- Visual dominance in sensory integration
  - Tactile information can be altered by visual information
    - For example, if the visual shape of an object differs considerably from its tactual shape (Rock & Victor experience)
  - The spatial location of a sound source can be drastically influenced by visual stimulation
    - For example, in television the voices are perceived to originate from the actors on the screen
  - Vision may alter speech perception
    - *McGurk effect explained using ICS*

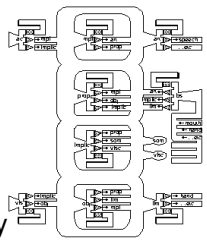
From Jukka Raisamo and Roope Raisamo, Tampere Univ.

Interactive cognitive sub-systems

- Theory ICS
  - APU Cambridge
  - Barnard & May, 1993
- ICS as predicting cognitive resources involved in using and choosing modalities

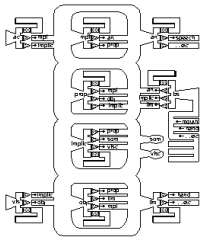


ICS



- Sensory subsystems:
  - Acoustic: sound frequency (pitch), timbre, intensity
    what we hear in the world
  - Visual: light wavelength (hue), brightness, saturation
    what we see in the world
- Perceptual subsystems:
  - Morphonolexical: Abstract structure of sounds, especially speech
    what we hear in our head, our mental voice
  - Object: Abstract structure of visual objects, their position and motion
    what we see in our head, our mind's eye

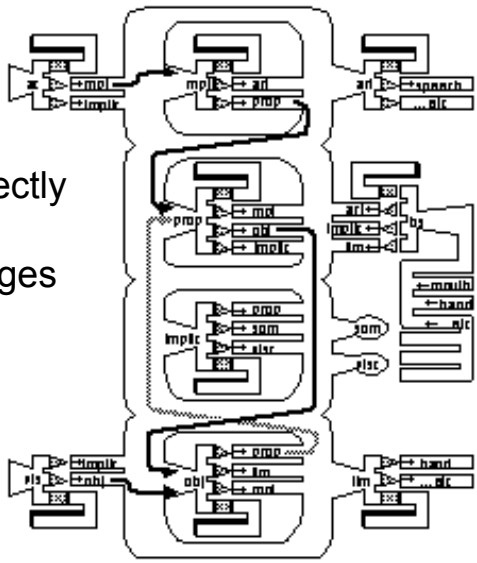
ICS



- Central subsystems:
  - Propositional: the identities of objects, their relationships, and facts about them
    what we know as facts about the world
  - Implicational: ideas about the real 'meanings' of events, situations and emotions
    what we know as 'feelings' or 'impressions'

ICS: Blending sight and sound

- Sight and sound
  - Sound cannot directly produce mental visual images



ICS: Blending sight and sound

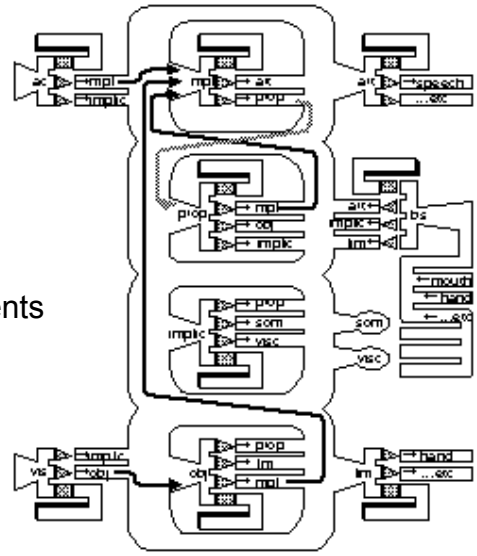
- Sight and sound
  - Sound cannot directly produce mental visual images
  - The consequence of this is that the object subsystem does not receive direct inputs of multimodal origin, and that our perception of the visual world is not directly affected by the sounds we hear

ICS: Blending sight and sound

- Sight and sound
  - our perception of the visual world is not directly affected by the sounds we hear
- however
  - there can be effects of our visual perception upon the way we interpret sound

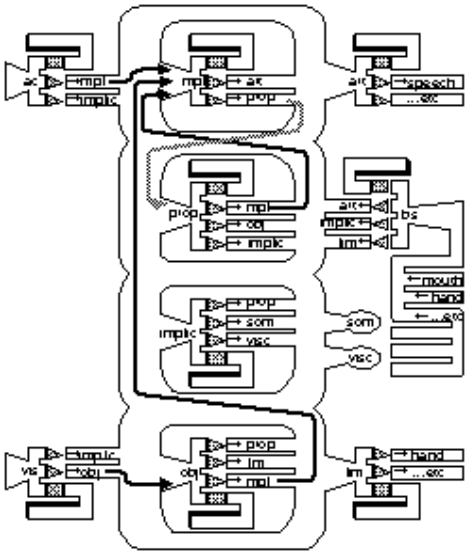
ICS: Blending sight and sound

- Sight and sound
  - there can be effects of our visual perception upon the way we interpret sound
  - a speaker's lip movements are usually blended with their speech - which is why 'out of synch' films are so difficult to watch



ICS: Blending sight and sound

- Sight and sound
  - there can be effects of our visual perception upon the way we interpret sound
  - McGurk effect
    video

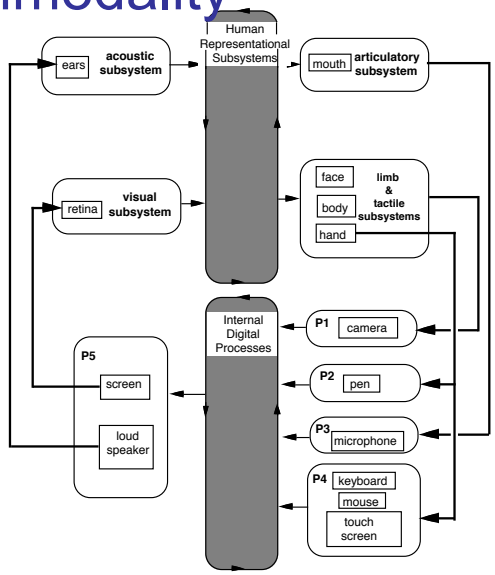


ICS: Blending sight and sound

- Sight and sound
  - there can be effects of our visual perception upon the way we interpret sound
  - McGurk effect
    A consequence of this blending of sight and sound occurs in the 'McGurk effect', where the sound of a speaker saying "ba ba ba" is dubbed onto the lip movements of them saying "ga ga ga". Most people actually report hearing the sound "da da da".

ICS and Input/Output Multimodality

- Theory ICS
  - APU Cambridge
- ICS as predicting cognitive resources involved in using and choosing modalities



Three paradigms for multimodality

- **Computer as tool**
- Multiple input modalities are used to enhance direct manipulation behavior of the system
  - the computer is a passive tool and tries to understand the user through all the different input modalities that the system recognizes
  - the user is responsible for initiating the actions
  - follows the principles of direct manipulation [Shneiderman, 1982]

41

Three paradigms for multimodality

- **Computer as partner**
- The multiple modalities are used to increase the anthropomorphism of the user interface
  - agent based conversational user interfaces
  - multimodal output is important: talking heads and other humanlike presentation modalities
  - speech recognition is a common input modality in these systems, and speech synthesis is used as an output modality

42

Three paradigms for multimodality

- **Proactive computing (ubicomp, PUI, …)**
- The multiple modalities are used to sense the user and the environment
  - multimodal (multisensory) input is important
  - the functionality of the system depends on the level of deduction (AI) the system is capable of
  - proactive functionality is often in the background and only indirectly visible for the user, predicting his/her actions and needs

43

Multimodality: challenges

- Parallel recognition
  - processing multiple input streams
- Joint interpretation
  - was interaction multimodal?
  - should streams be combined?
  - to what utterance does a particular gesture correspond?
- Interpretation
  - compensation if there is an error in any modality

44

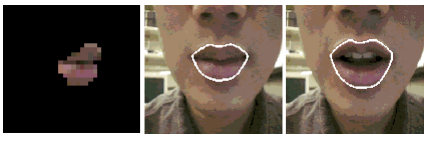
Multimodality: challenges

- Using multimodal input generally requires advanced recognition methods:
  - For each modality
  - For combining redundant information
  - For combining non-redundant information: “open this file (pointing)”
- Information is combined at two levels:
  - Feature level (early fusion)
  - Semantic level (late fusion)

45

Multimodality: challenges

- Early fusion
  - applies to combinations like speech+lip movement
    - Speech Recognition degrades in noisy environments
    - Use of Image based modeling of the lips can improve accuracy



  - Difficult because:
    - Of the need for MM training data
    - data need to be closely synchronized
    - Computational and training costs

46

Multimodality: challenges

- Late fusion
- for combinations of complementary information, like pen+speech.
  - Recognizers are trained and used separately
  - Unimodal recognizers can be available off-the-shelf
  - It is still important to accurately time-stamp all inputs: typical delays are known between e.g. gesture and speech

47

System or interface challenges?

- What we need are generic interaction technologies for converging towards full multimodal interaction:
- Including increasingly powerful:
  - enabling technologies, such as signal processing for speech, vision, touch, etc.
  - early (signal level) and late (semantic) input fusion
  - semantic-level input and output processing
  - generic interaction management engines
  - output presentation: graphics, speech, touch, movement, etc.

48

Systems or interfaces?

- The systems created may be content-”heavy” or not
- But the generic nature of solutions implies that specific contents are less important to meeting the challenges
- Contents are basically application-specific
- What matters is the interaction machinery, which can be
  - far ”more” than a traditional interface but still ”less” than a complete application

49

Multimodality: Path to evolution

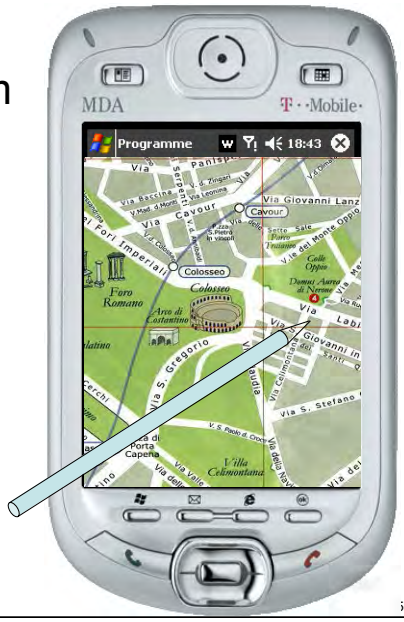
- Since 1980 “Put that there” paradigm
  R. Bolt MIT

50

Multimodality: Path to evolution

- “Put that there” paradigm
  R. Bolt MIT

„Zoom in here”



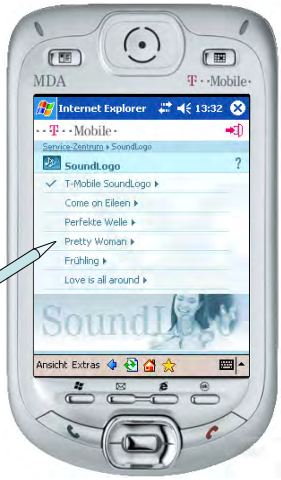
User selects a point of interest clicking with a stylus and speaking in order to focus it.

51

Multimodality: Path to evolution

- “Put that there” paradigm
  R. Bolt MIT

„Play this sound logo”

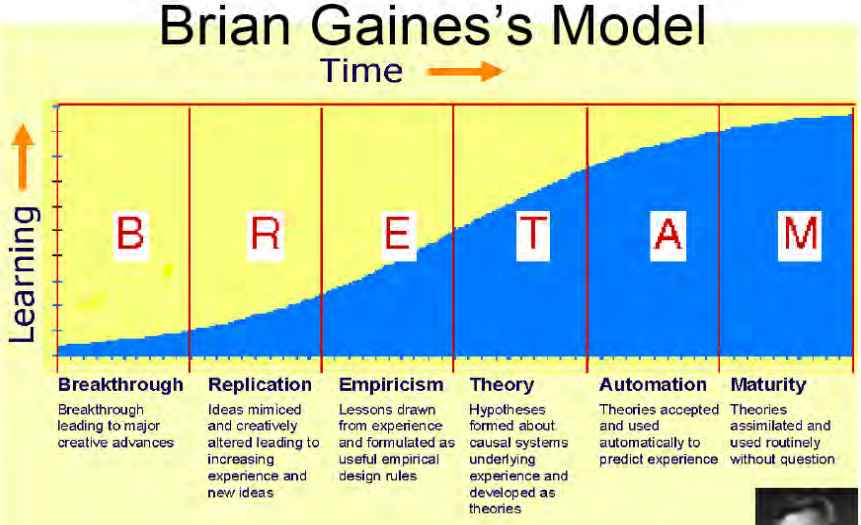


User selects a sound logo by clicking on the title with a stylus and speaking in order to hear it

52

Multimodality: Path to evolution

Brian Gaines's Model



In the 80's, Brian Gaines introduced a model on how science technology develops over time

Brian Gaines

53

Readings

- Bolt, R. A. "Put-that-there": Voice and gesture at the graphics interface. Proceedings of SIGGRAPH'80, 14, 3 (1980), 262–270
- Martin, J. C. TYCOON: Theoretical Framework and Software Tools for Multimodal Interfaces. Intelligence and Multimodality in Multimedia Interfaces, AAAI Press (1997)
- Nigay, L., Coutaz, J. The CARE Properties and Their Impact on Software Design. Intelligence and Multimodality in Multimedia Interfaces, (1997) *http://iihm.imag.fr/en/publication/*
- Oviatt, S. "Ten myths of multimodal Interaction", Comm. of the ACM, 42, 11 (1999), 74-81
- Turk, M., Robertson, G. Eds, Perceptual user Interfaces. Comm. of the ACM, 43, 3 (2000), 32-70
- ICS :
  - See supporting documents

54

Readings

- ACM SIGCHI: ACM's Special Interest Group on Computer-Human Interaction
  - http://www.sigchi.org/
- ICMI conference
  - International Conference on Multimodal Interfaces
- CHI conference
  - Computer Human Interface
- UIST conference
  - User Interface Software and technology
- MobileHCI conference
  - Human Computer Interaction with Mobile Devices and Services

55