VRIJE
UNIVERSITEIT
BRUSSEL

Next Generation User Interfaces

Multimodal Interaction

Prof. Beat Signer

Department of Computer Science
Vrije Universiteit Brussel

beatsigner.com

Human Senses and Modalities

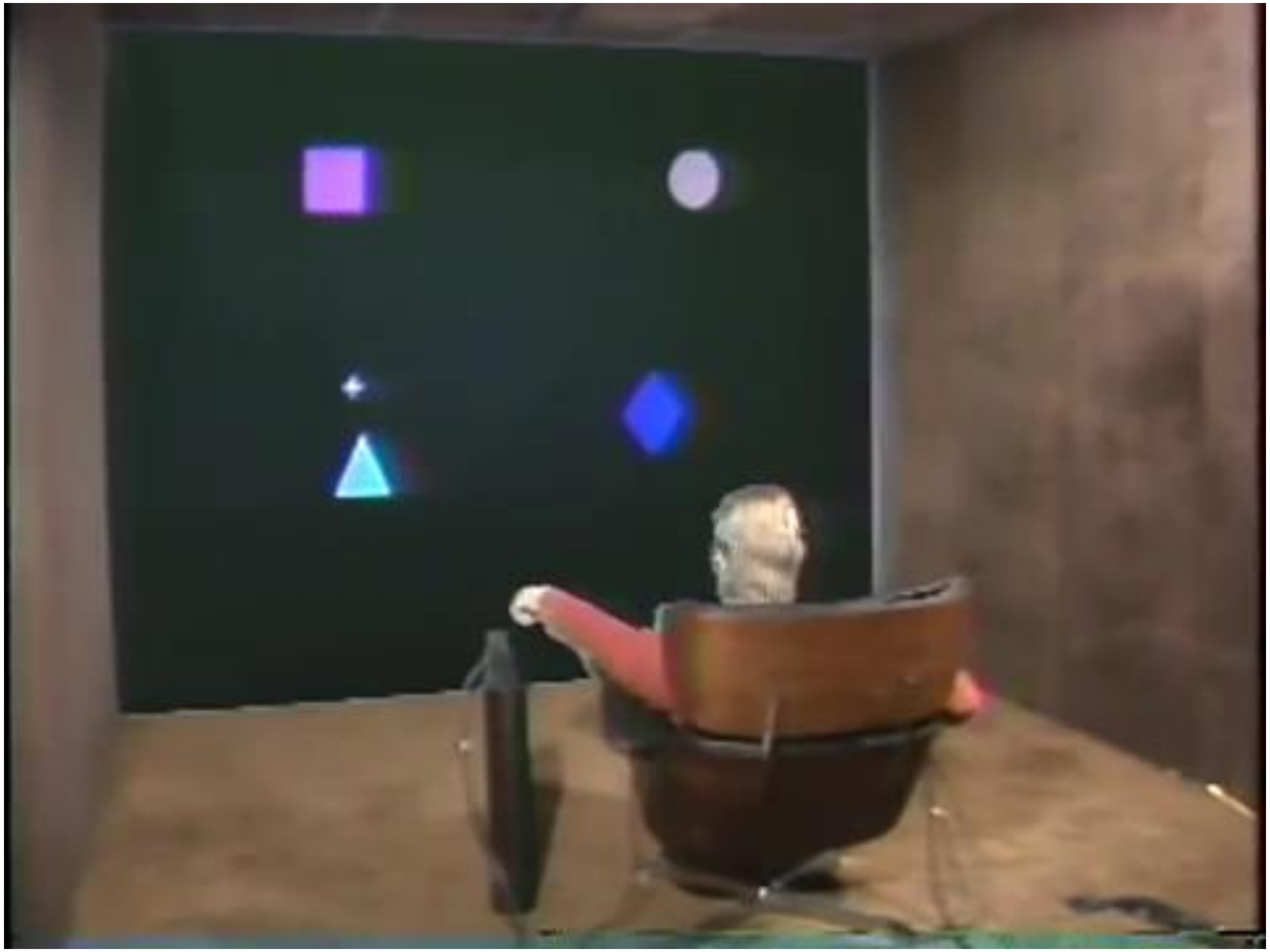
Sensory Perception	Sense Organ	Modality
sight	eyes	visual
hearing	ears	auditory
smell	nose	olfactory
taste	tongue	gustatory
touch	skin	tactile
(balance)	vestibular system	vestibular

> Modality *“refers to the type of communication channel used to convey or acquire information. It also covers the way an idea is expressed or perceived, or the manner an action is performed.”*

Nigay and Coutaz, 1993

- Input as well as output modalities

Bolt's "Put-that-there" (1980)



Bolt, 1980

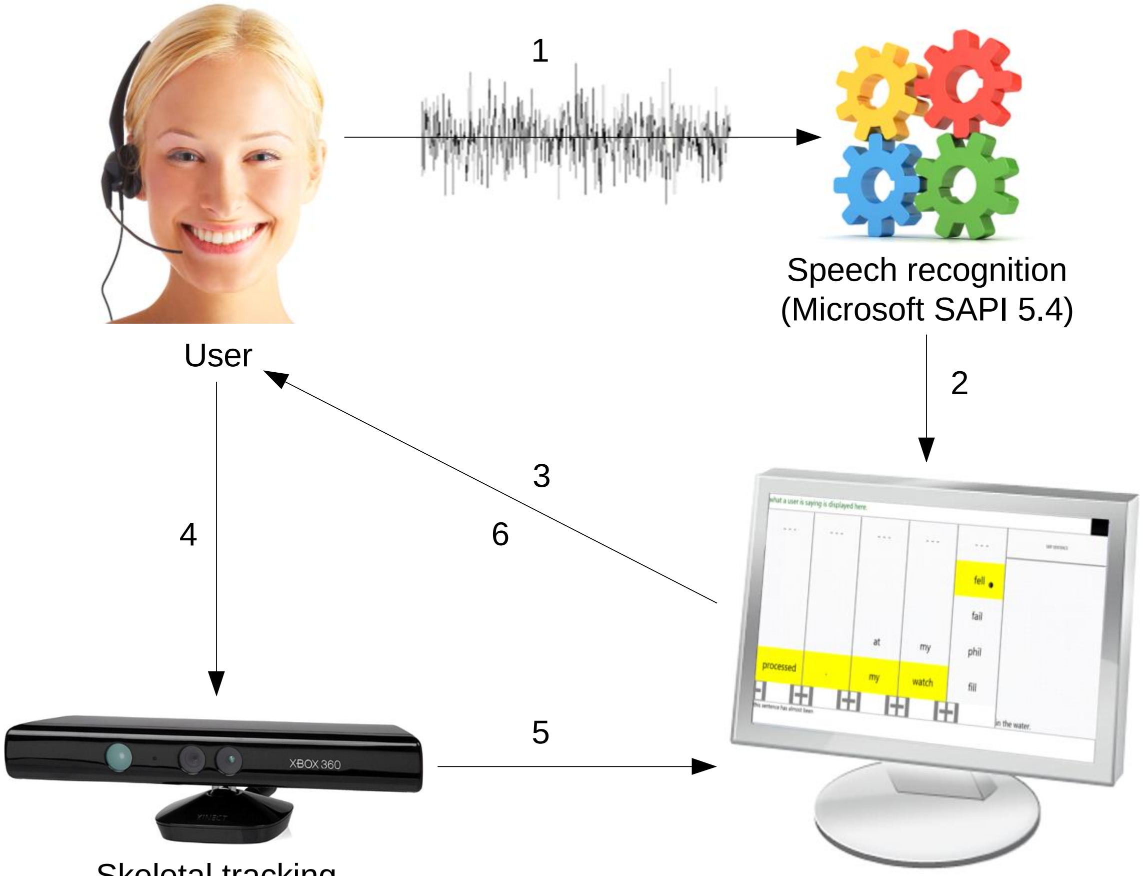
Bolt's "Put-that-there" (1980) ...

- Combination of two input modalities
  - speech is used to issue the semantic part of the command
  - gestures are used to provide the locations on the screen
- *Complementary* use of both modalities
  - use of only speech or gestures does not enable the same commands
- The *"Put-that-there"* interaction is still regarded as a rich case of multimodal interaction
  - however, it is based on an old mouse-oriented metaphor for selecting objects

Multimodal Interaction

- Multimodal interaction is about human-machine interaction involving multiple modalities
  - input modalities
    - speech, gesture, gaze, emotions, …
  - output modalities
    - voice synthesis, visual cues, …
- Humans are inherently multimodal!

Example: SpeeG2



1
2
3
4
5
6
what a user is saying is displayed here.
fell
fail
at
my
phil
processed
my
watch
fill
this sentence has almost been
in the water.

User

Speech recognition
(Microsoft SAPI 5.4)

SpeeG2 GUI

Skeletal tracking
(Microsoft Kinect)

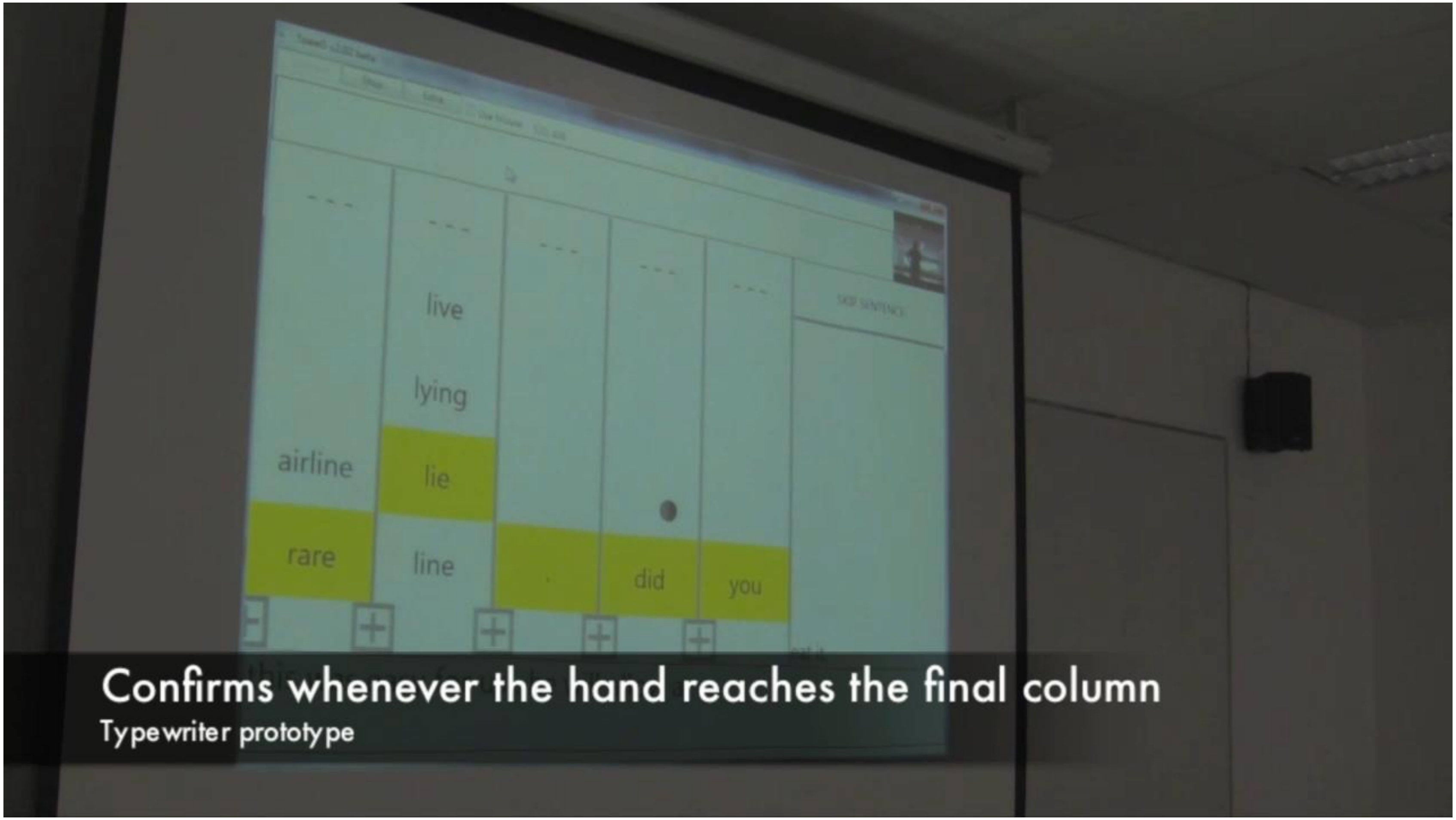
Sven De Kock

Example: SpeeG2 …

- While as user speaks a sentence, the most probable words are shown on the screen
- By using hand gestures the user can correct words (by selection) if necessary
    - feedback might be used for further training
- Improved recognition rate with up to 21 words per minute (WPM)

SpeeG2, WISE Lab

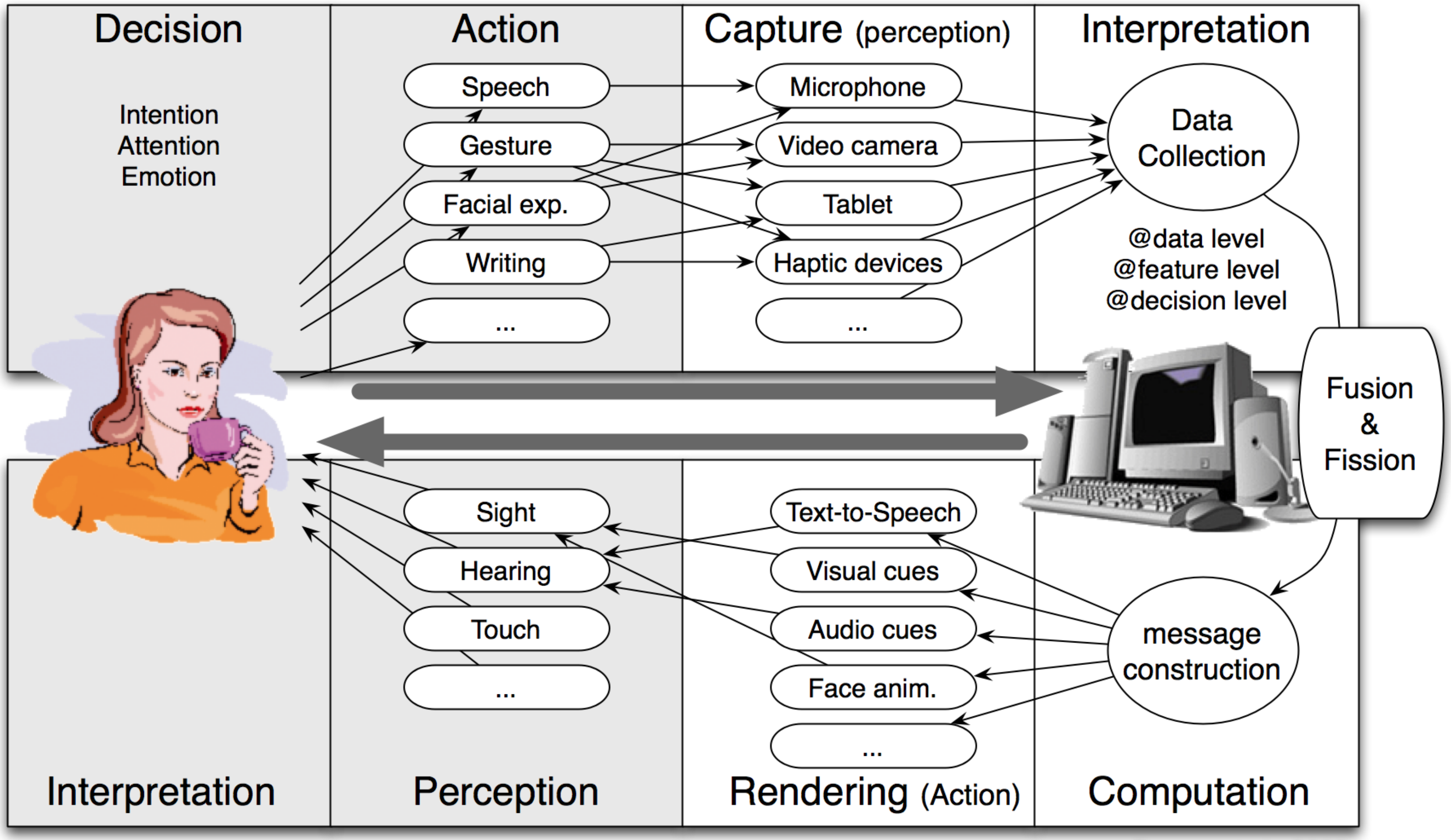
Video: SpeeG2



GUIs vs. Multimodal Interfaces

Graphical User Interfaces	Multimodal Interfaces
Single event input stream that controls the event loop	Typically *continuous* and *simultaneous* input from multiple *input streams*
Sequential processing	Parallel processing
Centralised architecture	Often distributed architectures (e.g. multi agent) with *high computational* and *memory requirements*
No temporal constraints	Requires *timestamping* and *temporal constraints* for multimodal fusion

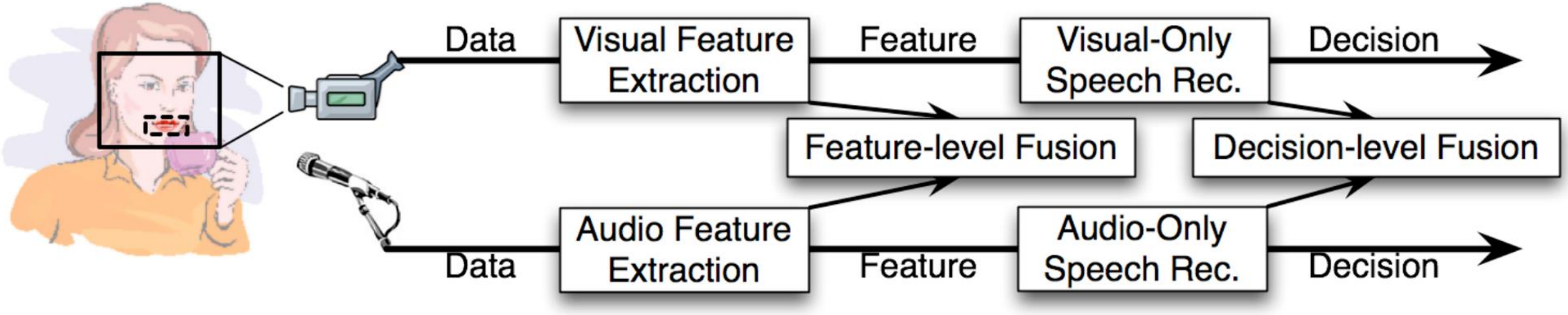
Multimodal Human-Machine Interaction Loop



Dumas et al., 2009

Multimodal Fusion

- Fusion in a multimodal system can happen at three different levels
  - data level
  - feature level
  - decision level



Dumas et al., 2009

Multimodal Fusion …

- Data-level fusion
  - focuses on the fusion of identical or tightly linked types of multimodal data
    - e.g. two video streams capturing the same scene from different angles
    - *low semantics* and *highly sensitive to noise* (no preprocessing)
- Feature-level fusion
  - fusion of features extracted from the data
    - e.g. speech and lip movements
- Decision-level fusion
  - focuses on interpretation based on semantic data
    - e.g. Bolt’s *"Put-that-there"* → speech and gesture
  - fusion of high-level information derived from data- and feature-level fusion
    - highly resistant to noise and failures

Comparison of Fusion Levels

	Data-level Fusion	Feature-level Fusion	Decision-level Fusion
Used for	raw data of the same modality	closely coupled modalities	loosely coupled modalities
Level of detail (data)	highest level of detail	moderate level of detail	difficult to recover data that has been fused on the data and feature level
Sensitivity to noise and failures	highly sensitive	moderately sensitive	not very sensitive
Usage	not frequently used for multimodal interfaces	used to improve the recognition rate	most widely used level of fusion for multimodal interfaces

Advantages of Multimodal Input

- Support and accommodate a user's *perceptual and communicative capabilities*
  - *natural user interfaces* with new ways of *engaging* interaction
- Integration of computational skills of computers in the real world by offering more *natural ways of human-machine interaction*
- *Enhanced robustness* due to the combination of different (partial) information sources
- *Flexible personalisation* by using different modalities based on user preferences and context
- Helps visually or physically impaired users

Multimodal Fission

- Forms a comprehensible answer for the user by making use of multiple output modalities based on the device and context
  - builds an abstract message through a *combination of channels*
- Three main steps
  - selection and structuring of content
  - selection of modalities
  - output coordination
- Coordination of the output on each channel is necessary to form a coherent message

Ten Myths of Multimodal Interaction

1. *If you build a multimodal system, users will interact multimodally*
   - users show a strong preference to interact multimodally but this does not mean that they always interact multimodally
   - unimodal and multimodal interactions are mixed depending on the task to be carried out

Sharon Oviatt

2. *Speech and pointing is the dominant multimodal integration pattern*
   - heritage from Bolt's "Put-that-there" where gesture and speech are used to *select* an object (mouse-oriented metaphor)
   - modalities can be used for more than just object selection
     - written input (e.g. SpeeG2), facial expressions, ...

Ten Myths of Multimodal Interaction ...

3. *Multimodal input involves simultaneous signals*
   - signals are often not overlapping
   - users frequently introduce (consciously or not) a small delay between two modal inputs
   - developers *should not rely on an overlap for the fusion process*
4. *Speech is the primary input mode in any multimodal system that includes it*
   - in human-human communication speech is indeed the primary input mode
   - in human-machine communication speech is *not the exclusive carrier of important information* and does also *not have temporal precedence* over other modalities

Ten Myths of Multimodal Interaction ...

5. *Multimodal language does not differ linguistically from unimodal language*
   - different modalities complement each other
     - e.g. avoid error-prone voice descriptions of spatial locations by pointing
   - in many respects multimodal language is substantially simplified
     - might lead to more robust systems
6. *Multimodal integration involves redundancy of content between modes*
   - early research in multimodal interaction assumed that different redundant modalities could improve the recognition
   - *complementary modalities* are *more important* and designers should not rely on duplicated content

Ten Myths of Multimodal Interaction ...

7. *Individual error-prone recognition technologies combine multimodally to produce even greater unreliability*
   - assumes that using two error-prone input modes such as speech and handwriting recognition results in greater unreliability
   - however, *increased robustness* due to *mutual disambiguation* between modalities can be observed
8. *All users' multimodal commands are integrated in a uniform way*
   - different users use different integration patterns (e.g. parallel vs. sequential)
   - system that can detect and adapt to a user's dominant integration pattern might lead to increased recognition rates

Ten Myths of Multimodal Interaction ...

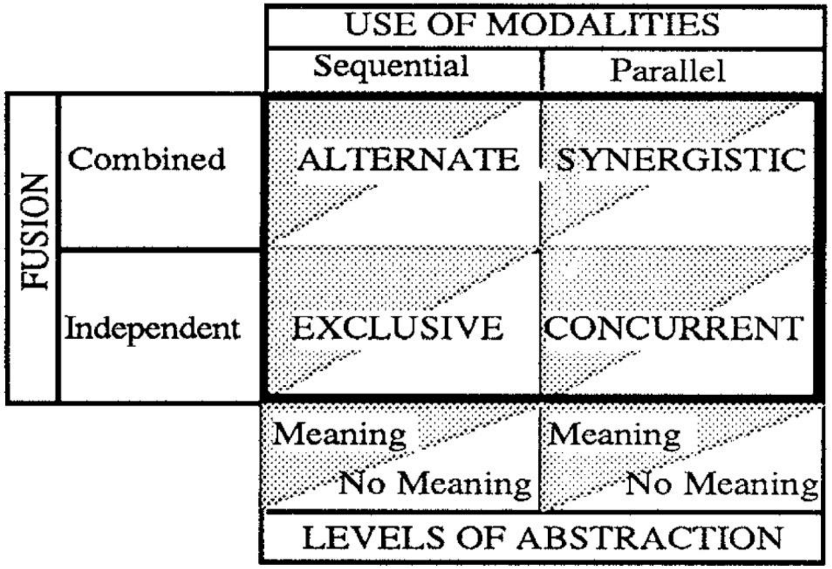
9. *Different input modes are capable of transmitting comparable content*
   - some modalities can be compared (e.g. speech and writing)
   - however, each modality has its own very distinctive gestures
     - e.g. gaze vs. a pointing device such a the Wiimote
10. *Enhanced efficiency is the main advantage of multimodal systems*
    - often there is only a minimal increase in efficiency
    - however, many other advantages
      - less errors, enhanced usability, flexibility, mutual disambiguation, …

Formal Models for Combining Modalities

- Formalisation of multimodal human-machine interaction
- Conceptualise the different possible relationships between input and output modalities
- Two conceptual spaces
  - CASE: use and combination or modalities on the *system side* (fusion engine)
  - CARE: possible combination of modalities at the *user level*

CASE Model

- 3 dimensions in the CASE design space
  - levels of abstraction
  - use of modalities
  - fusion

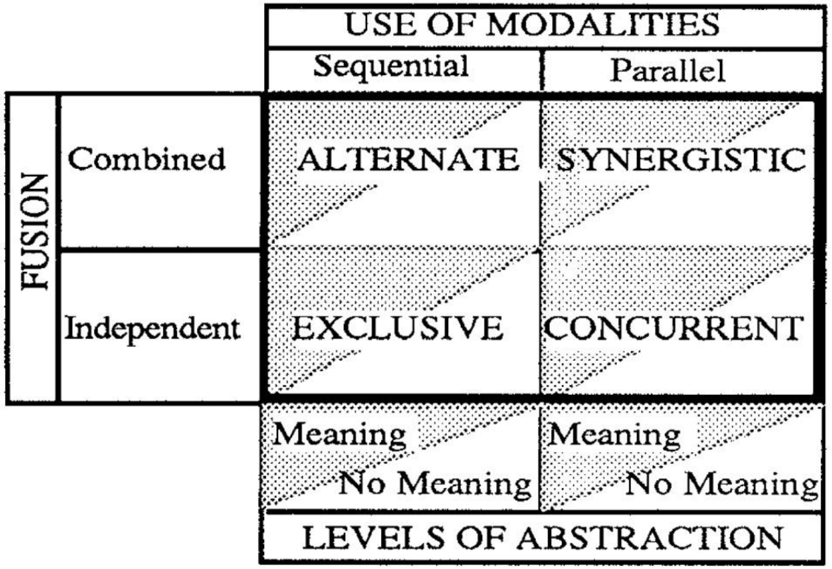


		USE OF MODALITIES	
		Sequential	Parallel
FUSION	Combined	ALTERNATE	SYNERGISTIC
	Independent	EXCLUSIVE	CONCURRENT
		Meaning No Meaning	Meaning No Meaning
		LEVELS OF ABSTRACTION	

Nigay and Coutaz, 1993

CASE Model: Levels of Abstraction

- Data received from a device can be processed at multiple levels of abstraction
- Speech analysis example
  - signal (data) level
  - phonetic (feature) level
  - semantic level
- A multimodal system is always classified under the meaning category

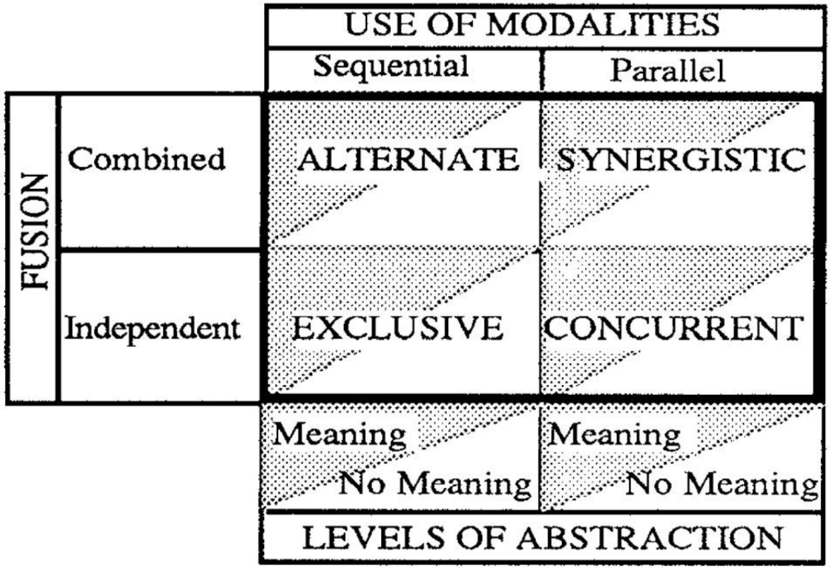


		USE OF MODALITIES	
		Sequential	Parallel
FUSION	Combined	ALTERNATE	SYNERGISTIC
	Independent	EXCLUSIVE	CONCURRENT
		Meaning No Meaning	Meaning No Meaning
		LEVELS OF ABSTRACTION	

Nigay and Coutaz, 1993

CASE Model: Use of Modalities

- The use of modalities expresses the temporal availability of multiple modalities
  - *parallel use* allows the user to employ multiple modalities simultaneously
  - *sequential use* forces the user to use the modalities one after another

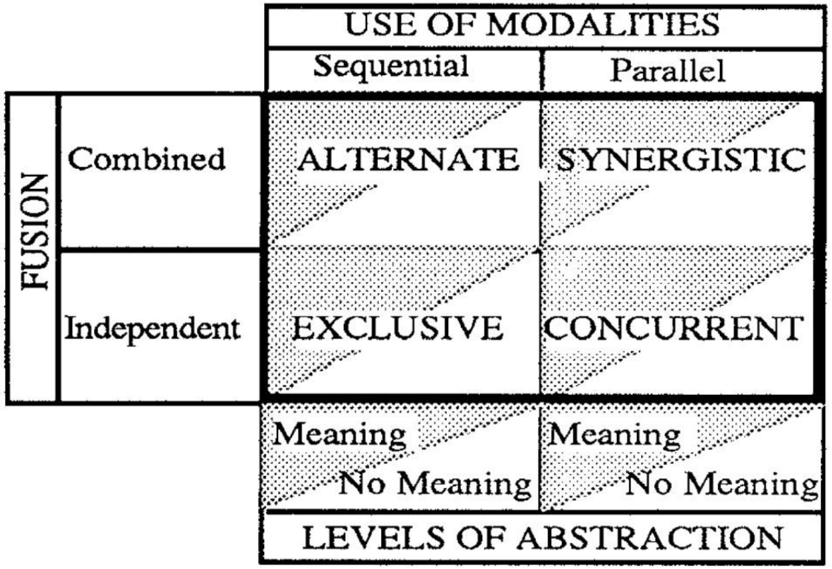


		USE OF MODALITIES	
		Sequential	Parallel
FUSION	Combined	ALTERNATE	SYNERGISTIC
	Independent	EXCLUSIVE	CONCURRENT
		Meaning No Meaning	Meaning No Meaning
		LEVELS OF ABSTRACTION	

Nigay and Coutaz, 1993

CASE Model: Fusion

- Possible combination of different types of data
  - *combined* means that there is fusion
  - *independent* means that there is an absence of fusion

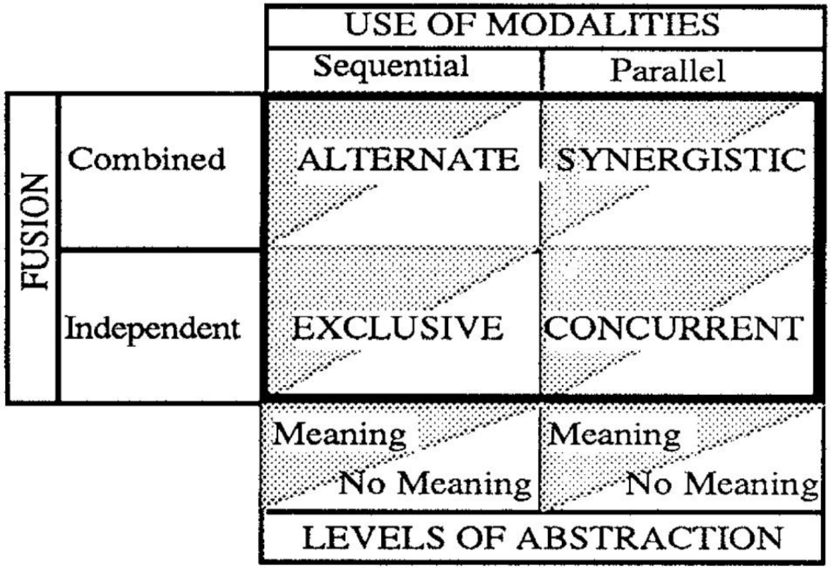


		USE OF MODALITIES	
		Sequential	Parallel
FUSION	Combined	ALTERNATE	SYNERGISTIC
	Independent	EXCLUSIVE	CONCURRENT
		Meaning No Meaning	Meaning No Meaning
		LEVELS OF ABSTRACTION	

Nigay and Coutaz, 1993

CASE Model: Classification

- Concurrent
  - two distinctive tasks executed in parallel
  - e.g. draw a circle and delete a square via a voice command
- Alternate
  - task with temporal alternation of modalities
  - e.g. *"Draw a circle there"* followed by pointing to the location

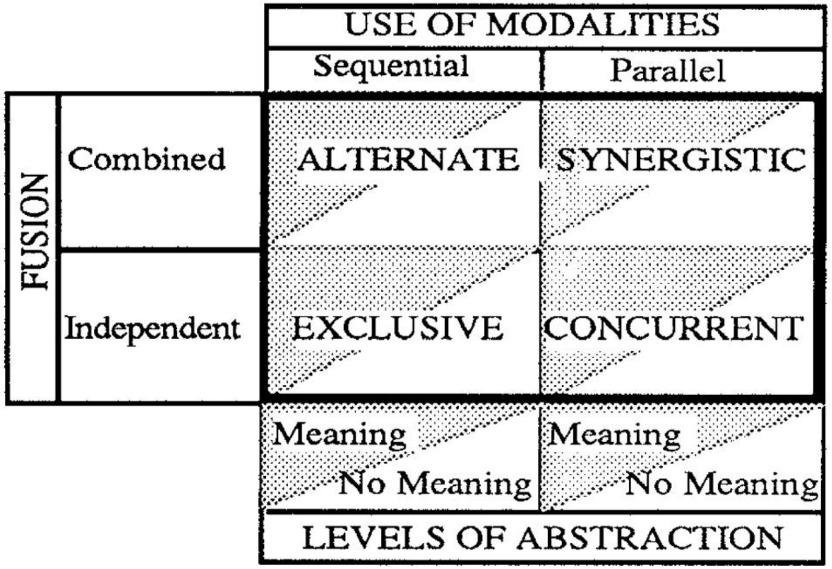


		USE OF MODALITIES	
		Sequential	Parallel
FUSION	Combined	ALTERNATE	SYNERGISTIC
	Independent	EXCLUSIVE	CONCURRENT
		Meaning No Meaning	Meaning No Meaning
		LEVELS OF ABSTRACTION	

Nigay and Coutaz, 1993

CASE Model: Classification

- Synergistic
  - task using several coreferential modalities in parallel
  - e.g. *"Draw a circle here"* with concurrent pointing to the location
  - *synergistic multimodal systems are the ultimate goal*
- Exclusive
  - one task after the other with one modality at a time (no coreference)



		USE OF MODALITIES	
		Sequential	Parallel
FUSION	Combined	ALTERNATE	SYNERGISTIC
	Independent	EXCLUSIVE	CONCURRENT
		Meaning No Meaning	Meaning No Meaning
		LEVELS OF ABSTRACTION	

Nigay and Coutaz, 1993

CARE Properties

- Four properties to characterise and assess aspects of multimodal interaction in terms of the *combination of modalities at the user level* [Coutaz et al., 1995]
  - *C*omplementarity
  - *A*ssignment
  - *R*edundancy
  - *E*quivalence

Complementarity

- Multiple modalities *are to be used within a temporal window* in order reach a given state
- *No single modality* on its own is *sufficient* to reach the state
- Integration (fusion) can happen *sequentially* or *in parallel*

- Example
  - *“Please show me this list”*
    and <pointing at ‘list of flights’ label>

Assignment

- Only one modality can be used to reach a given state
- Absence of choice

- Example
  - Moving the mouse to change the position of a window (considering that the mouse is the only modality available for that operation)

Redundancy

- If multiple modalities have the same expressive power (equivalent) and if they are used within the same temporal window
- Repetitive behaviour without increasing the expressive power
- Integration (fusion) can happen *sequentially* or *in parallel*

- Example
  - “Could you show me the list of flights?”
    and <click on ‘list of flights’ button>

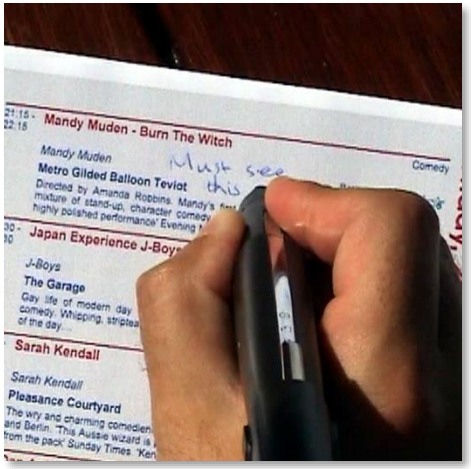
Equivalence

- Necessary and sufficient to *use any single of the available modalities*
- Choice between multiple modalities
- *No temporal constraints*

- Example
  - “Could you show me the list of flights?” or <click on ‘list of flights’ button>

Example for Equivalence: EdFest 2004

- Edfest 2004 prototype consisted of a digital pen and paper interface in combination with speech input and voice output [Belotti et al., 2005]
- User had a choice to execute some commands either via *pen input* or through *speech input* (equivalence)

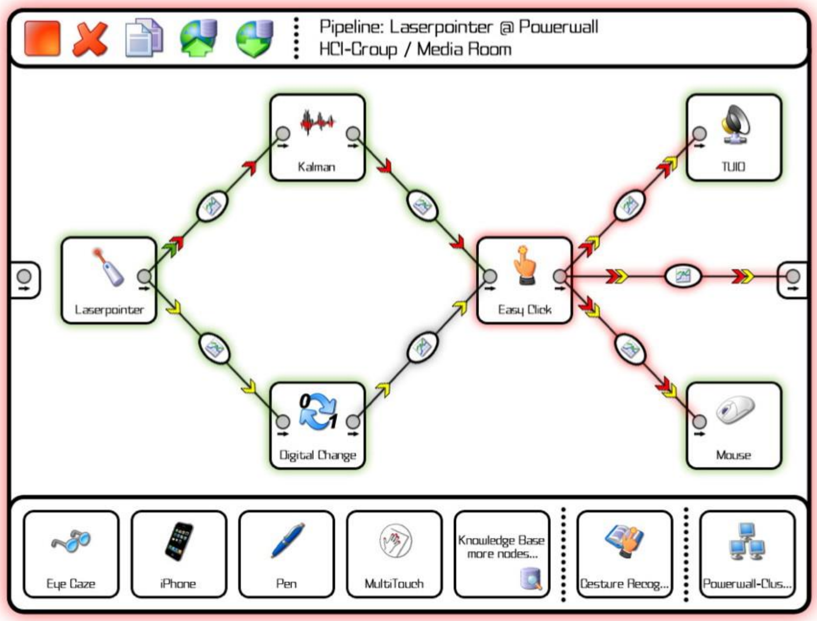


Multimodal Interaction Frameworks

- Software frameworks for the creation of multimodal interfaces
  - Squidy
  - OpenInterface
  - HephaisTK
  - Mudra
- Three families of frameworks
  - stream based
    - Squidy, OpenInterface
  - event based
    - HephaisTK
  - hybrid
    - Mudra

Squidy

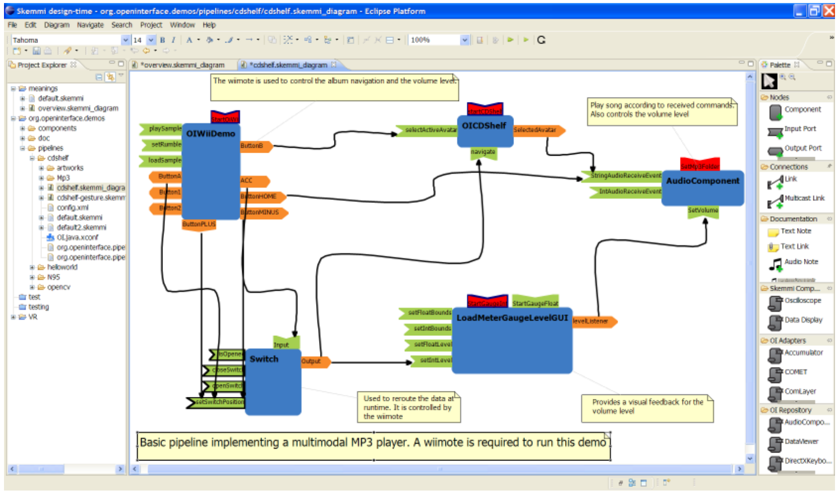
- Integrated tool with a GUI
- Interfaces are define by pipelining components and filters
- Online knowledge base of device components and filters
- Semantic zooming
  - e.g. zooming on a filter shows a graph of the data passing through the filter



http://www.squidy-lib.de

OpenInterface Framework

- European research project (2006–2009)
- Component-based framework
- Online component library with components for numerous modalities and devices
- Design an application by pipelining components



The wiimote is used to control the album navigation and the volume level.

Play song according to received commands. Also controls the volume level

Used to reroute the data at runtime. It is controlled by the wiimote

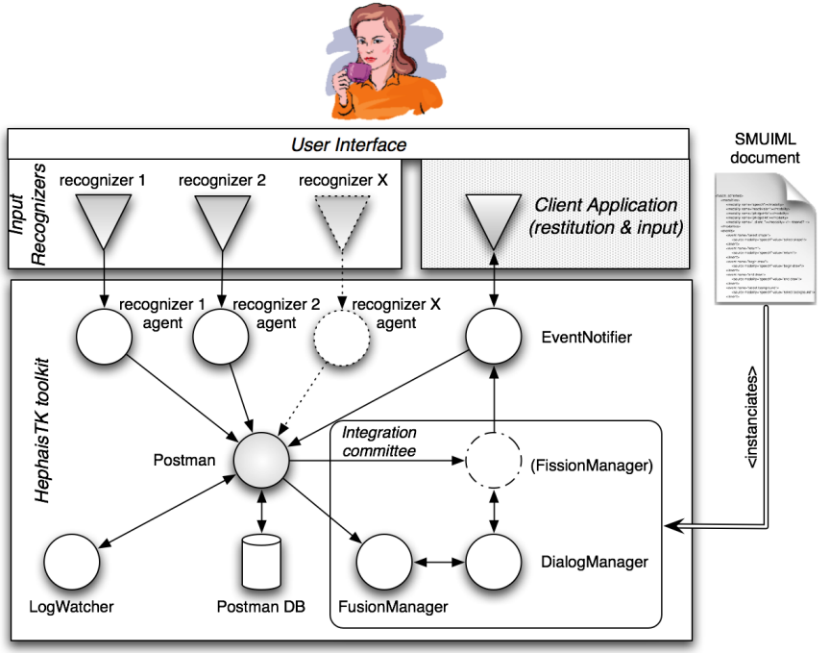
Provides a visual feedback for the volume level

Basic pipeline implementing a multimodal MP3 player. A wiimote is required to run this demo

http://www.openinterface.org

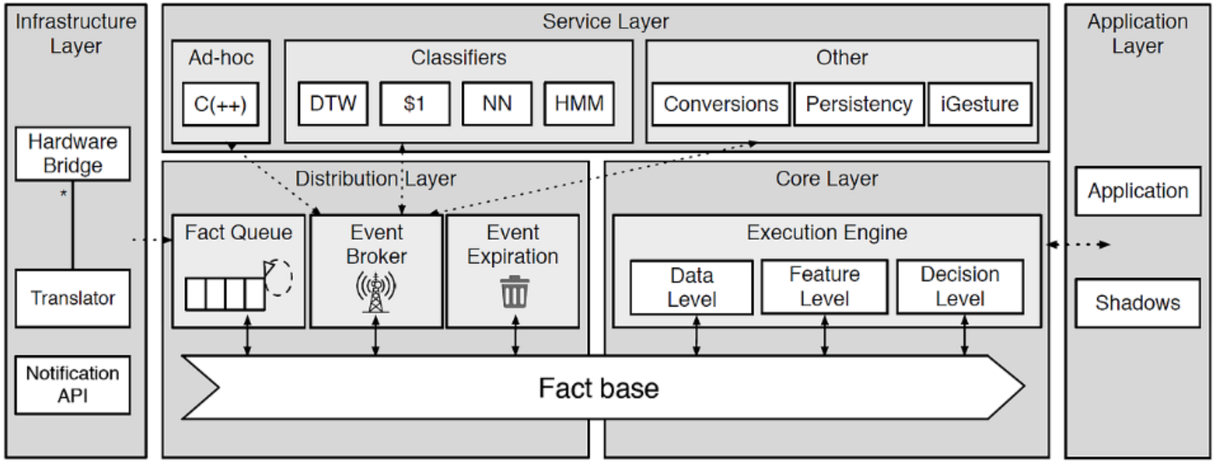
HephaisTK

- Software agents-based framework
- Focus on events
- Different fusion algorithms
- Description of the dialogue via the SMUIML language



http://www.hephaistk.org

Mudra



Hoste et al., 2011

- Fusion across different levels of abstraction
  - unified fusion framework based on shared fact base
- Interactions defined via declarative rule-based language
- Rapid prototyping
  - simple integration of new input devices
  - integration of external gesture recognisers

References

- R.A. Bolt, *"Put-that-there": Voice and Gesture at the Graphics Interface*, In Proceedings of Siggraph 1980, 7th Annual Conference on Computer Graphics and Interactive Techniques, Seattle, USA, July 1980
  - https://dx.doi.org/10.1145/965105.807503
- L. Nigay and J. Coutaz, *A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion*, In Proceedings of CHI 1993, 3rd International Conference on Human Factors in Computing Systems, Amsterdam, The Netherlands, April 1993
  - https://dx.doi.org/10.1145/169059.169143

References …

- J. Coutaz et al., *Four Easy Pieces for Assessing the Usability of Multimodal Interaction: The CARE Properties*, In Proceedings of Interact 1995, 5th International Conference on Human-Computer Interaction, Lillehammer, Norway, June 1995
    - https://dx.doi.org/10.1007/978-1-5041-2896-4_19
- B. Dumas, D. Lalanne and S. Oviatt, *Multimodal Interfaces: A Survey of Principles, Models and Frameworks*, Human Machine Interaction, LNCS 5440, 2009
    - https://dx.doi.org/10.1007/978-3-642-00437-7_1

References …

- S. Oviatt, *Ten Myths of Multimodal Interaction*, Communications of the ACM 42(11), November 1999
  - https://dx.doi.org/10.1145/319382.319398
- Put-that-there demo
  - https://www.youtube.com/watch?v=RyBEUyEtxQo
- R. Belotti, C. Decurtins, M.C. Norrie, B. Signer and L. Vukelja, *Experimental Platform for Mobile Information Systems*, Proceedings of MobiCom 2005, 11th Annual International Conference on Mobile Computing and Networking, Cologne, Germany, August 2005
  - https://beatsigner.com/publications/belotti_MobiCom2005.pdf

References …

- L. Hoste and B. Signer, *SpeeG2: A Speech- and Gesture-based Interface for Efficient Controller-free Text Entry,* Proceedings of ICMI 2013, 15th International Conference on Multimodal Interaction, Sydney, Australia, December 2013
  - https://beatsigner.com/publications/hoste_ICMI2013.pdf
- L. Hoste, B. Dumas and B. Signer, *Mudra: A Unified Multimodal Interaction Framework*, Proceedings of ICMI 2011, 13th International Conference on Multimodal Interaction, Alicante, Spain, November 2011
  - https://beatsigner.com/publications/hoste_ICMI2011.pdf

References …

- L. Hoste, *A Declarative Approach for Engineering Multimodal Interaction*, PhD thesis, Vrije Universiteit Brussel (VUB), Brussels, Belgium, June 2015
  - https://beatsigner.com/theses/PhdThesisLodeHoste.pdf
- A. Stanciulescu, Q. Limbourg, J. Vanderdonckt, B. Michotte and F. Montero, *A Transformational Approach for Multimodal Web User Interfaces Based on UsiXML*, Proceedings of ICMI 2005, 7th International Conference on Multimodal Interfaces, Trento, Italy, October 2005
  - https://dx.doi.org/10.1145/1088463.1088508

References …

- SpeeG2 demo
  - https://www.youtube.com/watch?v=ItKySNv8l90

Next Lecture

Pen-based Interaction