Multimodal User Interfaces 2019

[3] Multimodal Interaction, fusion, fission and systems architecture

Denis Lalanne

05/03/2019

Thanks to Jean Caelen, Laurence Nigay, and Sharon Oviatt for their inputs and inspiring material. Thanks also to Bruno Dumas for all our discussions on the topics.

Outline

- Terminology
- A model of H/M communication
- Advantages and major issues
- CARE and CASE : two conceptual models
- Fusion
- Fission
- Toolkits and formats

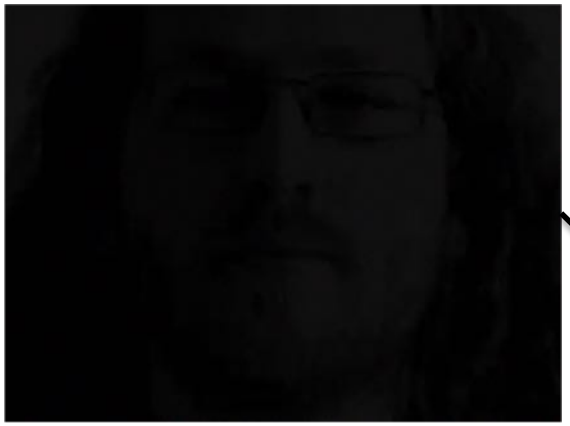
Terminology

- « **Multimodal interfaces** process two or more combined user input modes (such as speech, pen, touch, manual gesture, gaze, and head and body movements)
  - in a coordinated manner
  - with multimedia system output.

  They are a new class of interfaces that aim to recognize naturally occurring forms of human language and behavior, and which incorporate one or more recognition-based technologies (e.g. speech, pen, vision) » (S. Oviatt et al., 2002)

- System-centered definition of multimodality (Nigay & Coutaz 1993). The two features that define a multimodal system are:
  - Fusion of different types of data ;
  - Real-time processing with temporal constraints

Mc Gurk effect: multimodal processing

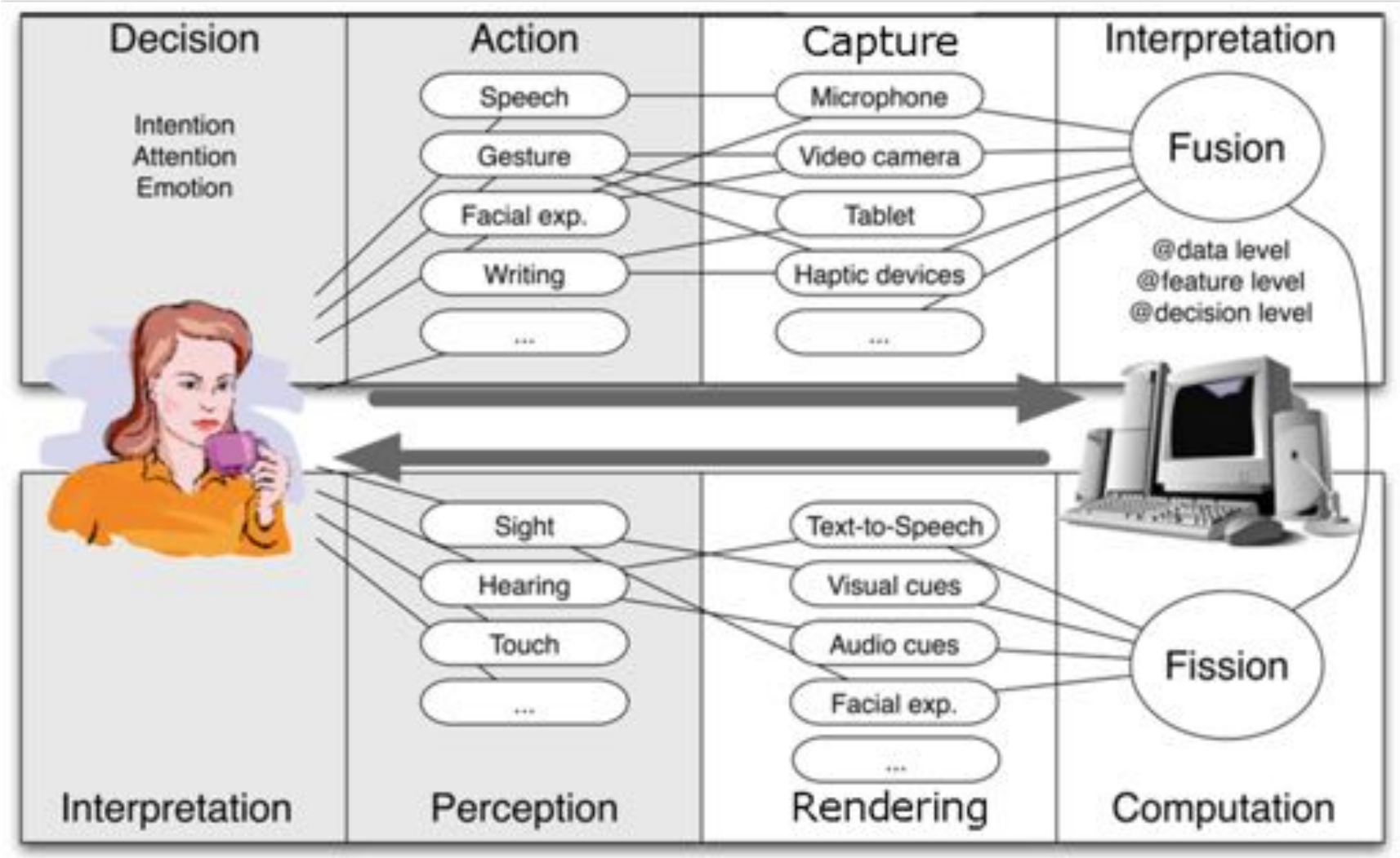


Sound of “ba”

Image of “ga”

Perception of “da”

Multimodal H/M Interaction



Dumas, Bruno, Denis Lalanne, and Sharon Oviatt. "Multimodal interfaces: A survey of principles, models and frameworks." *Human machine interaction*. Springer, Berlin, Heidelberg, 2009. 3-26.

Multimodal Interfaces

Advantages and major issues

Potential benefits

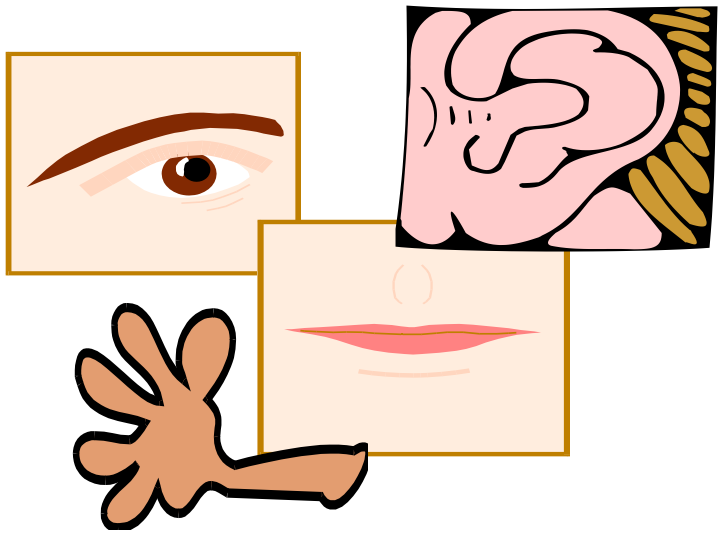
- A list by Maybury and Wahlster [1998, p. 15]:
  - Efficiency
  - Redundancy
  - Perceptability
  - Naturalness
  - Accuracy
  - Synergy
- [Oviatt, 1999a]
  - Improved error handling & efficiency:
    - *36% fewer errors*
    - *10% faster task completion*
  - Greater expressive power
  - Support for users' preferred interaction style
  - Accommodation to diverse users, tasks & usage environments

Multimodal Systems vs differ from GUIs

GUI Interfaces	Multimodal Interfaces
Single input stream	Multiple input streams
Atomic, deterministic	Continuous, probabilistic
Sequential processing	Parallel processing
Centralized architectures	Distributed & time-sensitive architectures

Fondamental problems

- Adequation of tasks to modalities
- Adequation of usage to user profiles
- Fusion of inputs
  - Resolution of the co-reference: match the multimodal referents
- Fission of outputs
  - Resolution of the difference : activate the most suitable referent

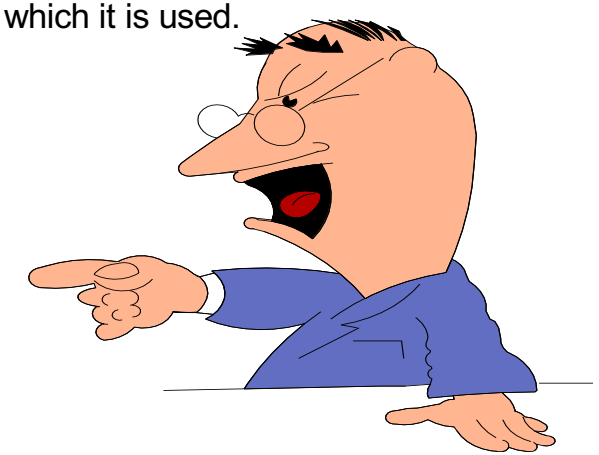


Modalities input/output adequation

- Speech
  - input : commands, macro-commands (isolated words, continuous speech), data inputs
  - output : guides, examples, requests, explanations, relances, etc. (synthesis, sentences to fill)
- Hand-writing
  - input : text, identifiers, numbers (keyboard, pen-based interfaces)
  - output : detailed explanations (screen)
- Gesture
  - input : pointing 2D or 3D (mouse, glove, touch screen), signs (camera)
  - output : force feedback
- Vision
  - input : subject position, facial expression, gesture tracking (image recognition)
  - output : information visualization, virtual reality (image synthesis, animation)

E.g. Some speech properties and difficulties (linguistic & rhetoric)

- Ellipsis
  - Ellipsis refers to any omitted part of speech that is understood; i.e. the omission is intentional.
  - Of verb: Draw a circle. A triangle.
  - Of noun: Destroy the red.
- Anaphora and cataphora
  - Anaphora is an instance of an expression referring to another. The strict definition of anaphora includes only references to preceding utterances. Under this definition, forward references are instead named cataphora, and both effects together are endophora.
  - Regression : Draw a circle. Move it.
  - progression : Destroy it, this circle
- Deixis
  - a deictic expression is an expression that refers to the personal, temporal, or spatial aspect of an utterance, and whose meaning therefore depends on the context in which it is used.
  - Put that there. The circle there
- Repetitions
  - Draw a red circle… No a green
  - Draw a red circle.. No destroy it
- Hesitations



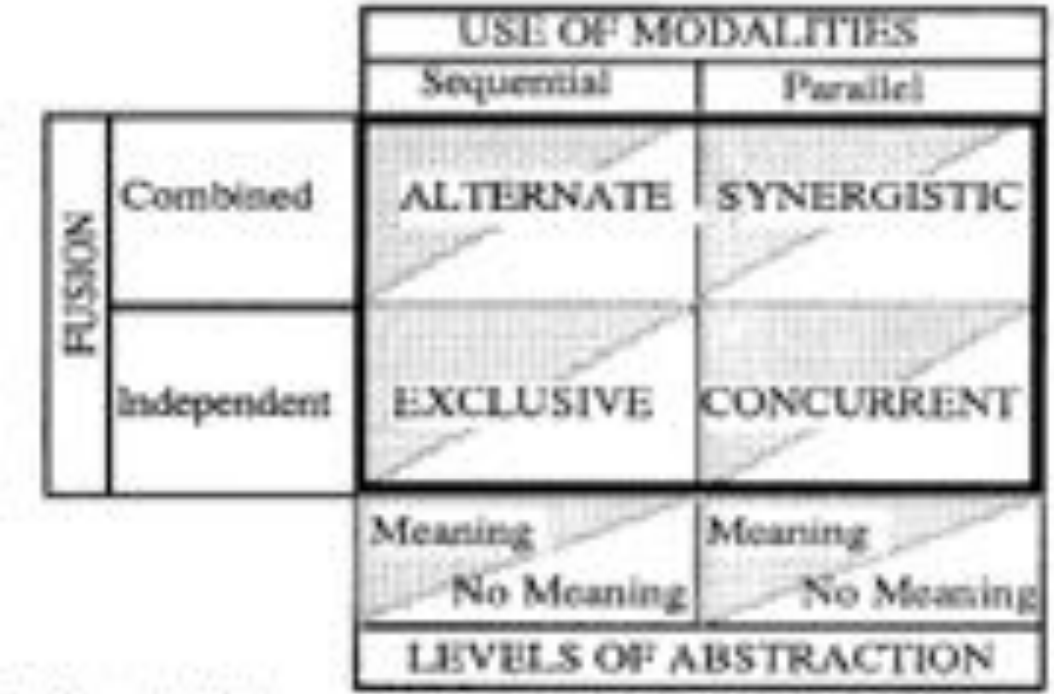
CARE and CASE models

CARE and CASE : Two design spaces [Nigay, Coutaz]

- Need to formalize human/machine multimodal interactions
- Conceptualize the different possible relationships between input and output modalities
- Two conceptual spaces:
  - Multimodal communication types
    Machine-side (CASE)
  - Multimodal systems
    Human-side (CARE) == Usability properties

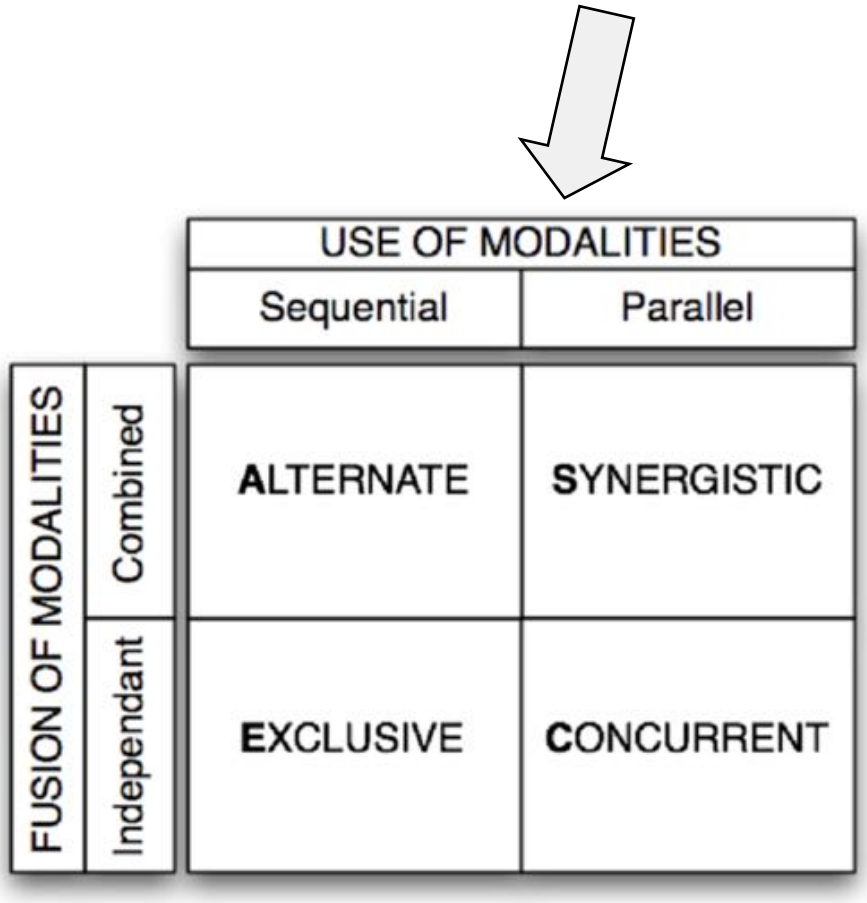
The CASE model: Multimodal systems Communication types

- 3 dimensions in the design space :
  - Levels of abstraction
  - Use of modalities
  - Fusion



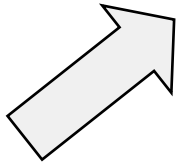
Multimodal communication types Machine-side (CASE)

- Use of modalities:
  - Depends on temporal use
  - Parallel: multiple modalities employed simultaneously.
  - Or sequentially, one at a time.



Multimodal communication types Machine-side (CASE)

- Fusion
  - Possible combination of different types of data
  - Independent: absence of fusion, no coreference.
  - Combined: fusion necessary



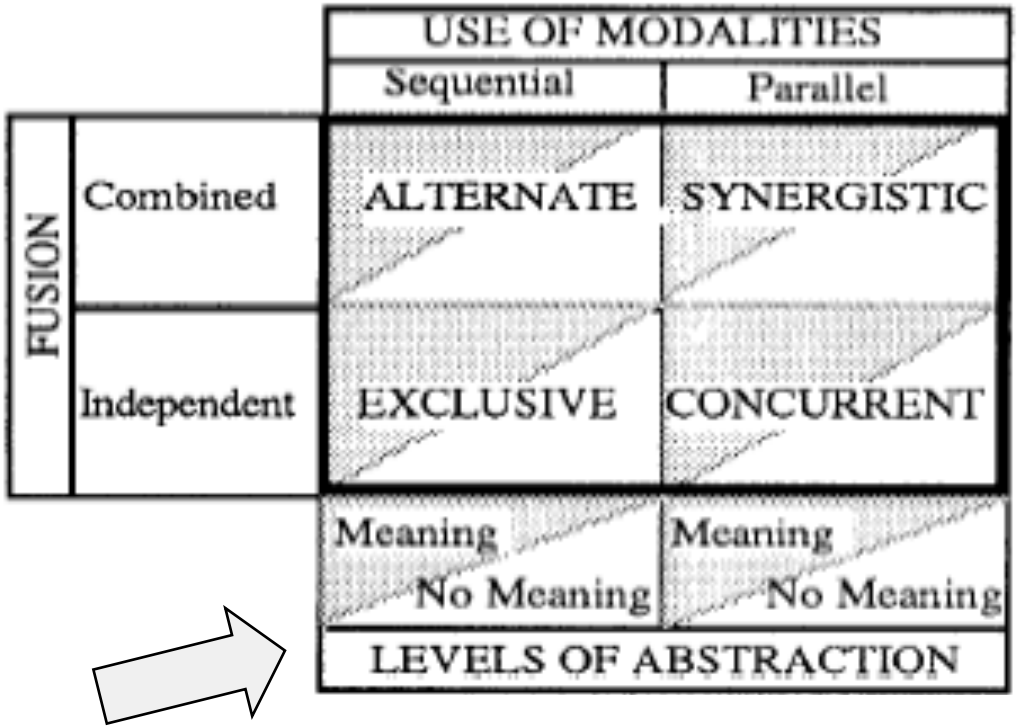
FUSION OF MODALITIES	USE OF MODALITIES: Sequential	USE OF MODALITIES: Parallel
Combined	**A**LTERNATE	**S**YNERGISTIC
Independant	**E**XCLUSIVE	**C**ONCURRENT

Multimodal communication types Machine-side (CASE)

- Levels of abstraction

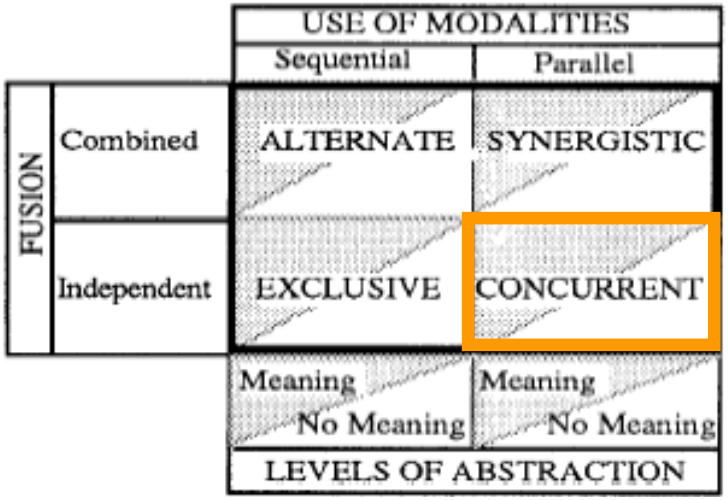
  - Data received from a device can be processed at multiple levels of abstraction;

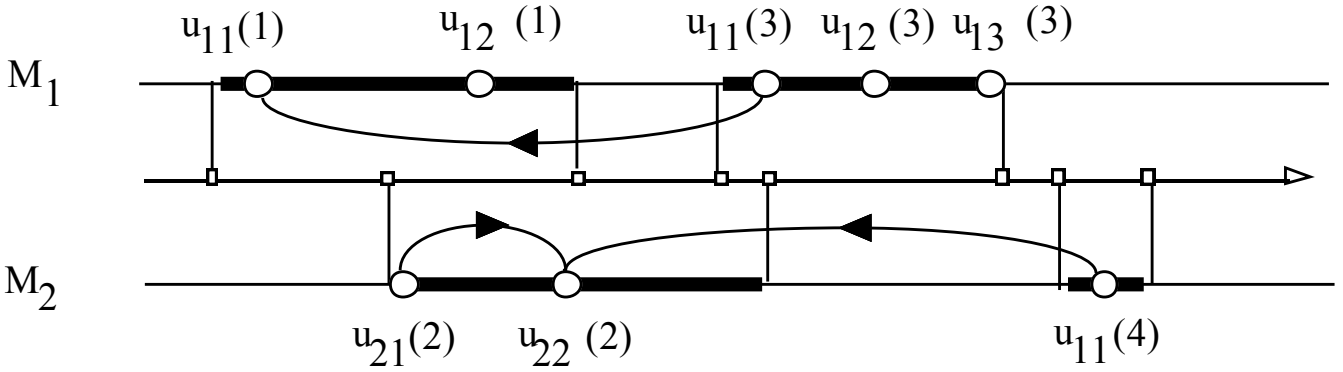
  - Exemple with speech Analysis:
    - Signal level
    - Phonetic level
    - Semantic level



CASE - Concurrent

C = Concurrent,
two distinct tasks in parallel,
No co-reference,
No temporal constraint

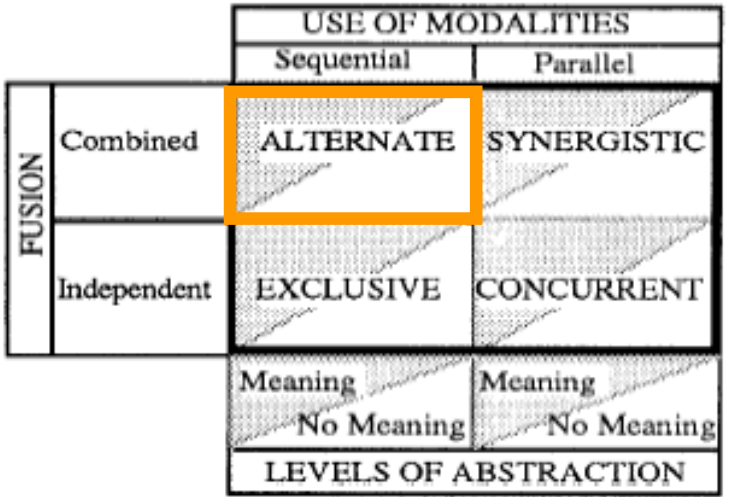


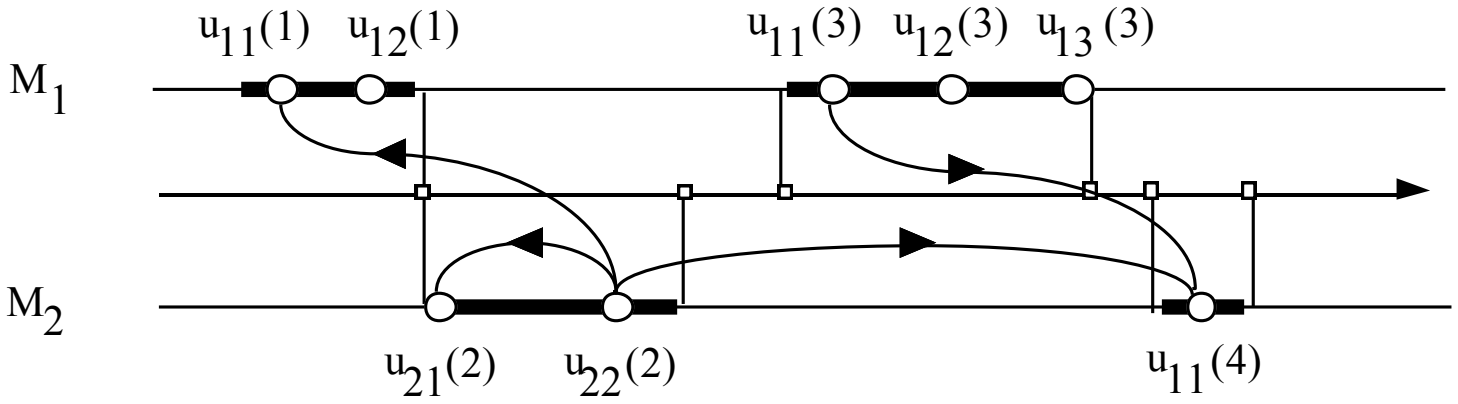


E.g.: *"Draw a circle" + draw a square next to it*

CASE - Alternate

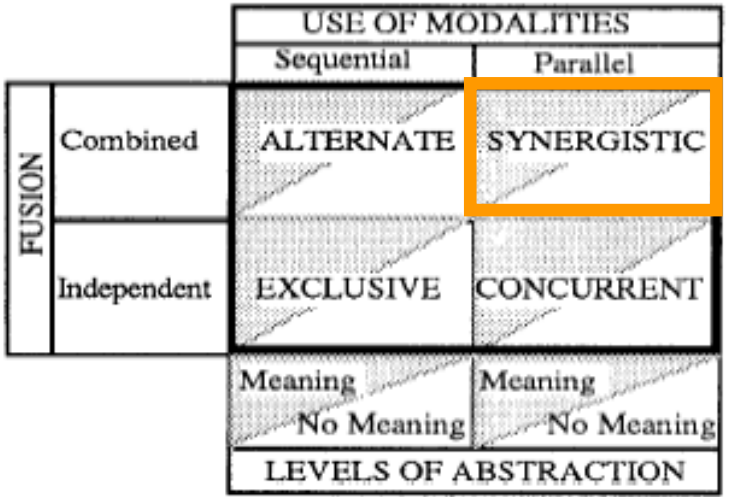
A = Alternate,
A task with temporal alternation of modalities, using coreferences

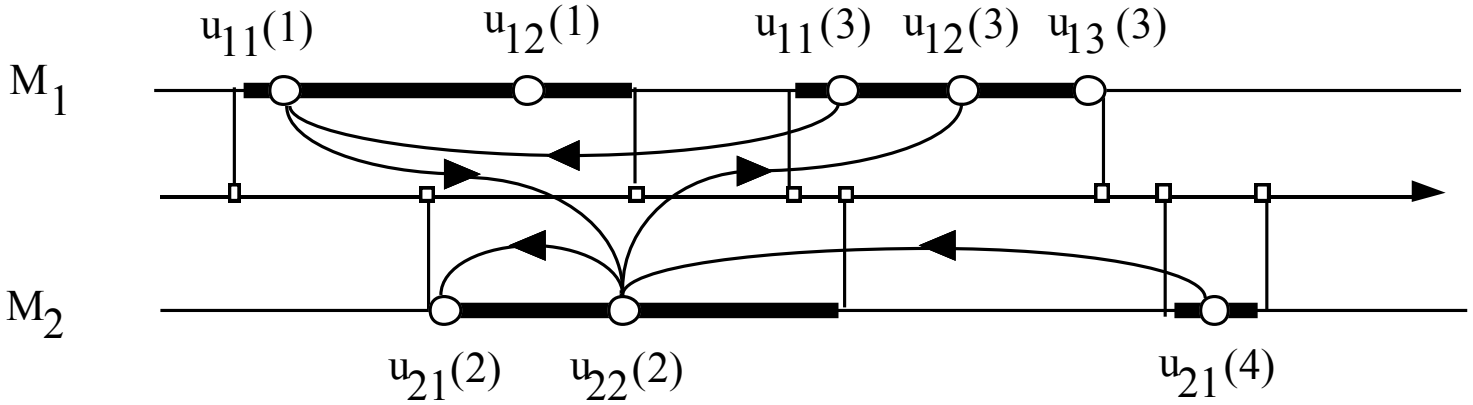




CASE - Synergistic

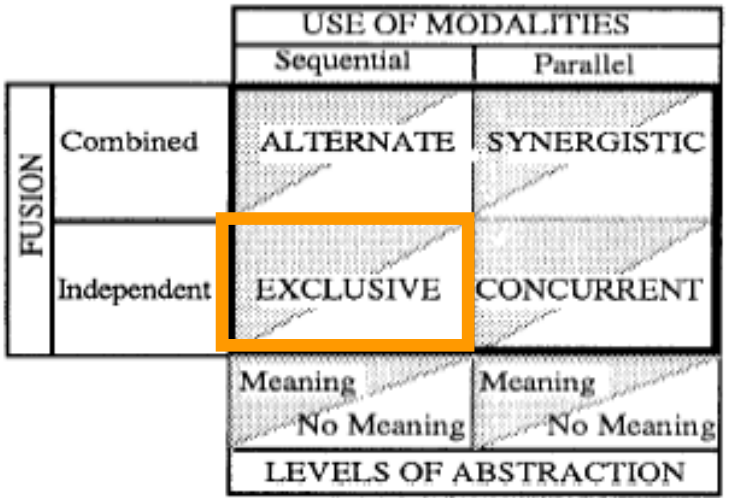
S = Synergistic,
A task, in parallel, using several coreferent modalities

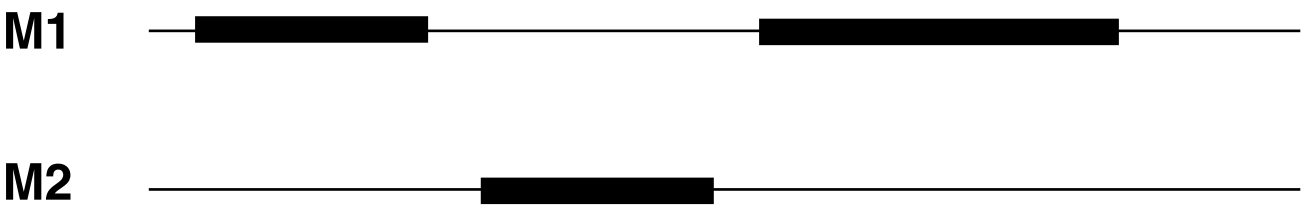




CASE - Exclusive

E = Exclusive,
One task after the other using one modality at a time,
No coreference





The CARE model: Multimodal systems Usability properties

- 4 properties of multimodal interaction in order to characterize and assess usability and fusion in multimodal interaction :
  - **C**omplementarity
  - **A**ssignment
  - **R**edundancy
  - **E**quivalence

- Human side of fusion

Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J., and Young, R. M. Four Easy Pieces for Assessing the Usability of Multimodal Interaction: The CARE Properties. In Proceedings of the 5th International Conference on Human-Computer Interaction (Interact 1995) (Lillehammer, Norway, June 1995), 115–120.

Multimodal systems
Usability properties (CARE)

- Complementarity :
  - If mutliple modalities are to be used within a temporal window to reach a given state
  - No modality taken individually is sufficient to reach the state
  - Can occur sequentially or in parallel

  - E.g.:
    - ✓ « Please give me details about this list » (Command)
    - ✓ and <point at a « list of flights » label > (Position)

Multimodal systems
Usability properties (CARE)

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

  - E.g.:
    - ✓ Movement of mouse to change the position of a window (only gesture efficient, no other modality can help performing this action easely)

Multimodal systems
Usability properties (CARE)

- Redundancy :
  - If multiple modalities have the same expressive power (-> equivalent) and if they are all used within the same temporal window
  - Repetitive behaviour without increasing the expressive power
  - Can occur in parallel or sequentially

  - E.g.:
    - ✓ « Could you show me the list of flights? »
    - ✓ And <clic on « list of flights » button>

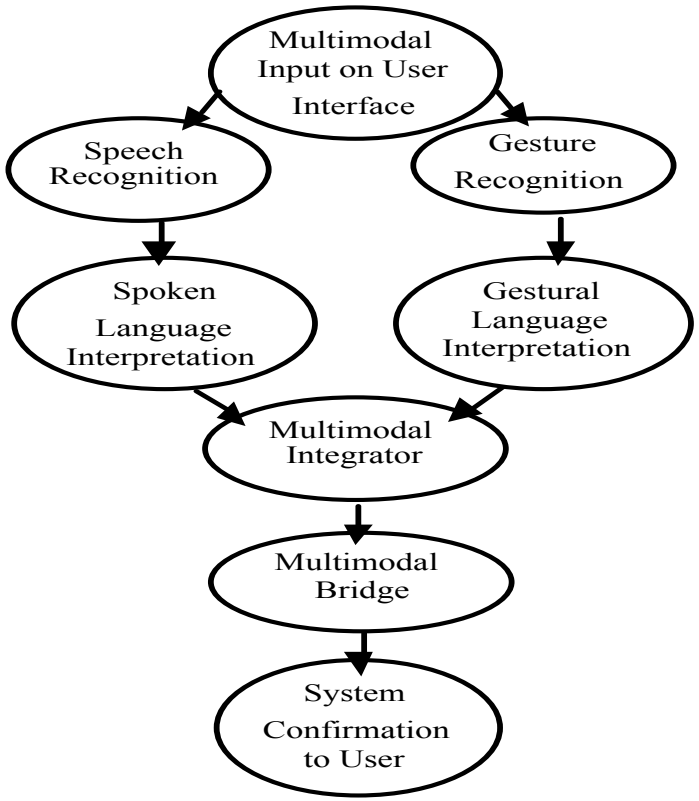
Multimodal systems
Usability properties (CARE)

- Equivalence :
  - Necessary and sufficient to use any one of the available modalities
  - Availability of choice between multiple modalities
  - No temporal constraints

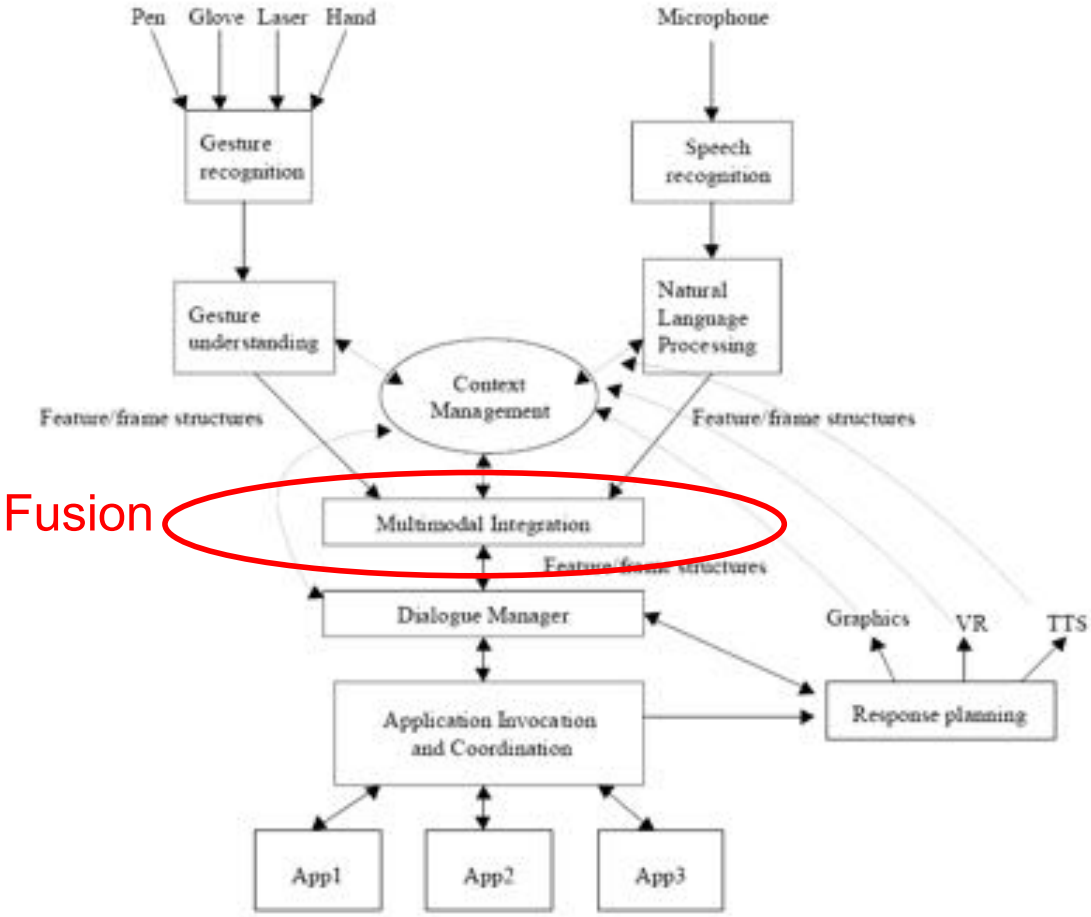
  - E.g.:
    - ✓ « Could you show me the list of flights? »
    - ✓ Or <clic on « list of flights » button>

Multimodal systems
Architecture

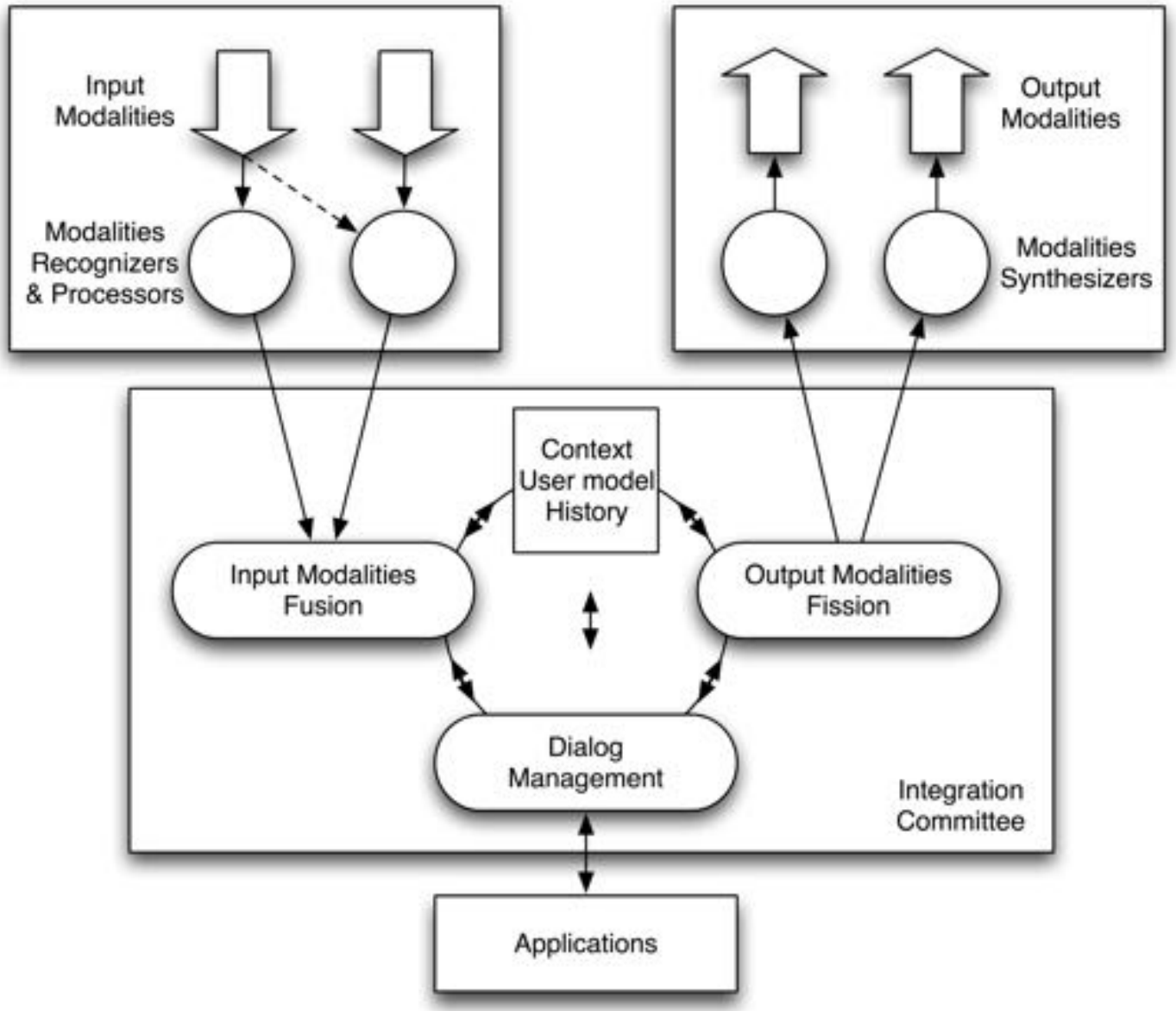
Architecture multimodale



Oviatt

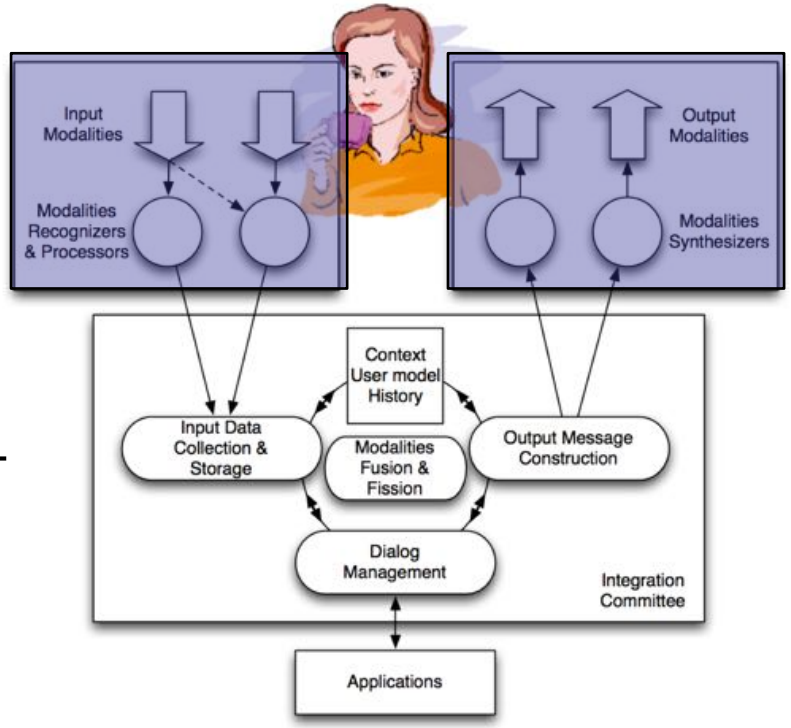


SmartKom


Input
Modalities
Modalities
Recognizers
& Processors
Output
Modalities
Modalities
Synthesizers
Context
User model
History
Input Modalities
Fusion
Output Modalities
Fission
Dialog
Management
Integration
Committee
Applications

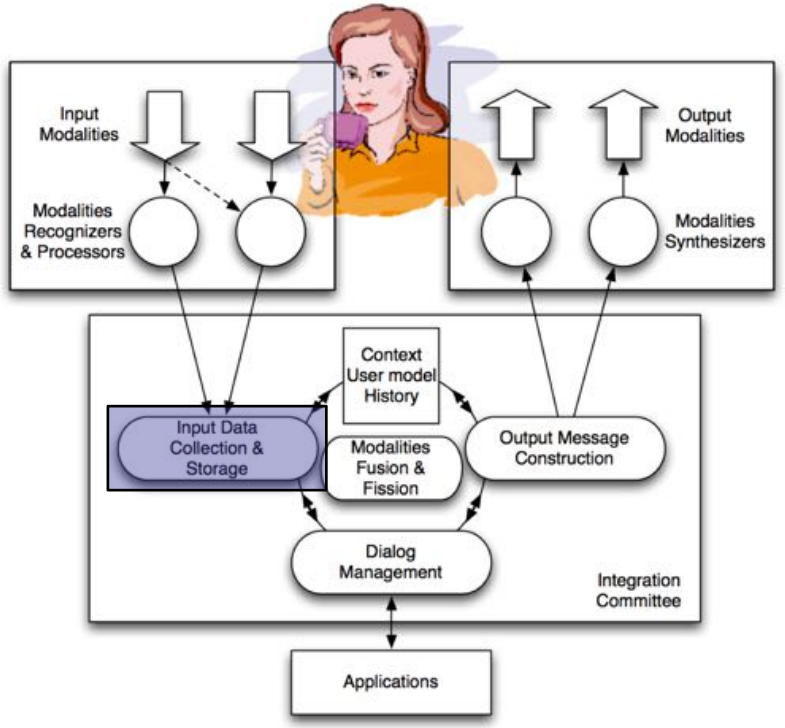
Multimodal Architecture: Input Modalities

- Input data provided by different modalities recognisers
- Big heterogeneity between these recognisers!
  - APIs, programming languages
  - Access through protocols
  - Frequency of input
  - Representation of data
- Need for generic access to these APIs + generic data representation



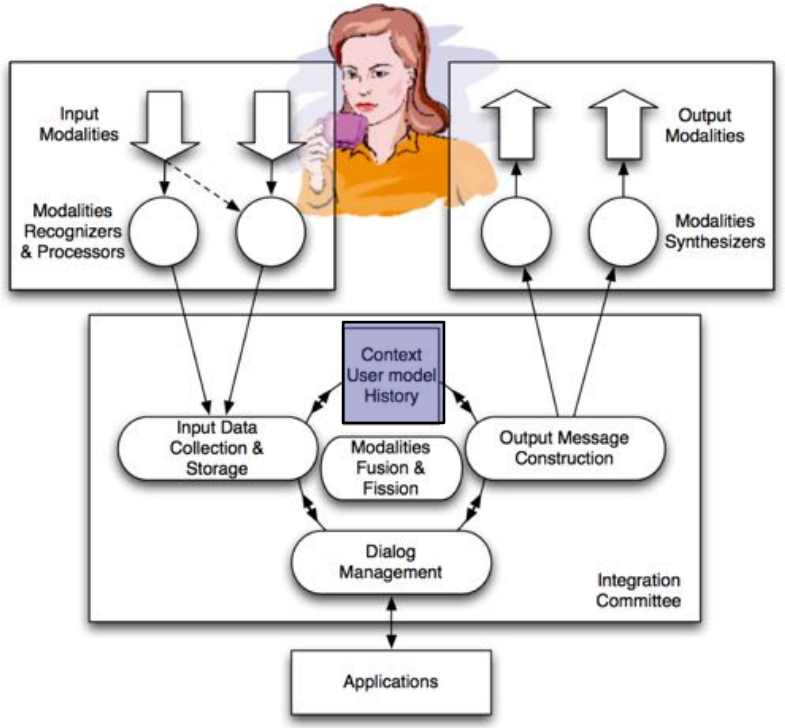
Multimodal Architecture: Storage

- Storage of data collected through the recognisers is not mandatory, but helps when
  - Context has to be taken into account
  - You want to access the history of the interaction
  - You interpret the raw data and want to store intermediate results
- Warning with databases: they can become a strong bottleneck



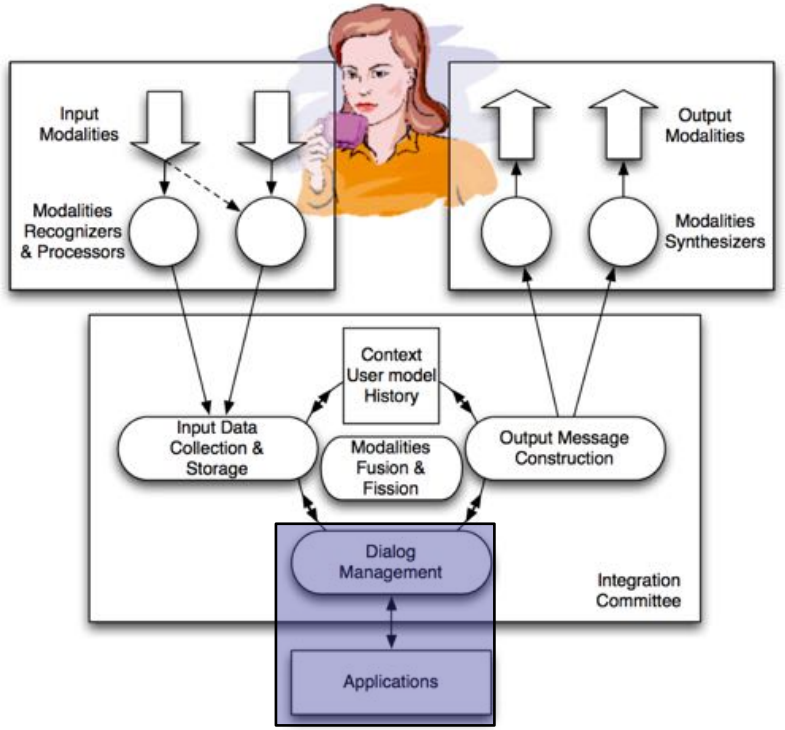
Multimodal Architecture: Context Model

- The context and user models can be used by the system to adapt the application's input and output
  - Example: voice input and output can be used in a car when you have your hands and eyes focussed on driving
- This component is not always present in multimodal systems



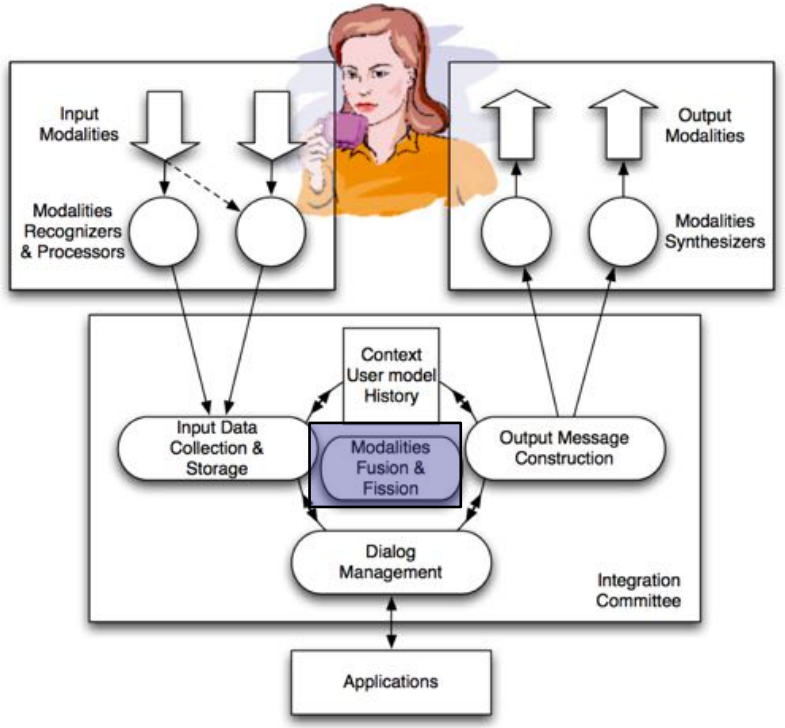
Multimodal Arch.: Dialogue Management

- Representation of the dialogue between the human and the machine
- Links the input/output interface and the application logic
  - Either using a formal representation (e.g. a XML based language, like SMUIML)
  - Or directly embedded in the application



Multimodal Arch.: Fusion and Fission

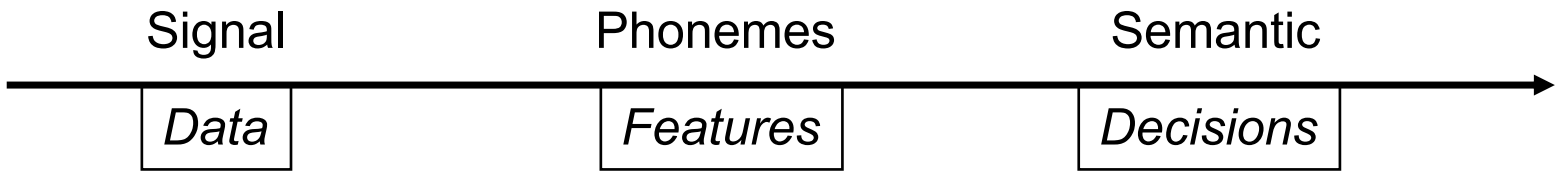
- Fusion of multimodal input
  - Takes the raw data coming from the different modalities and seeks to extract the message from the user
- Fission of output modalities
  - Considers how to best transmit an answer to the user, based on the current context of use and the available output modalities



Fusion

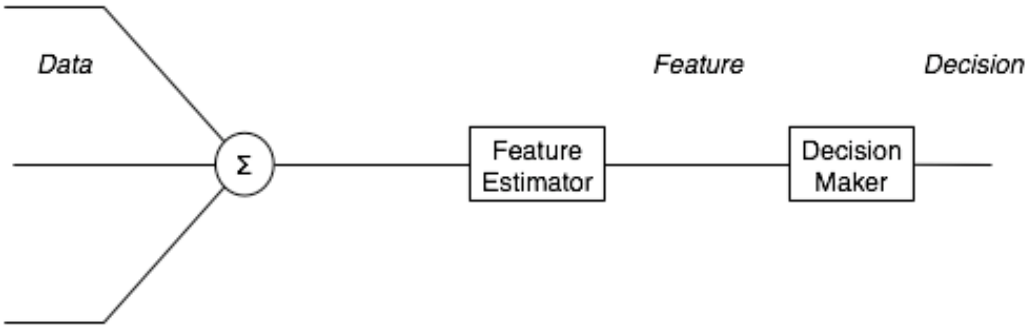
- Definition : « Resolution of the co-reference: match the multimodal referents »
- Theoretically, three types of fusion :
  - Raw Data fusion
  - Features fusion
  - Decision fusion

- Example with speech :



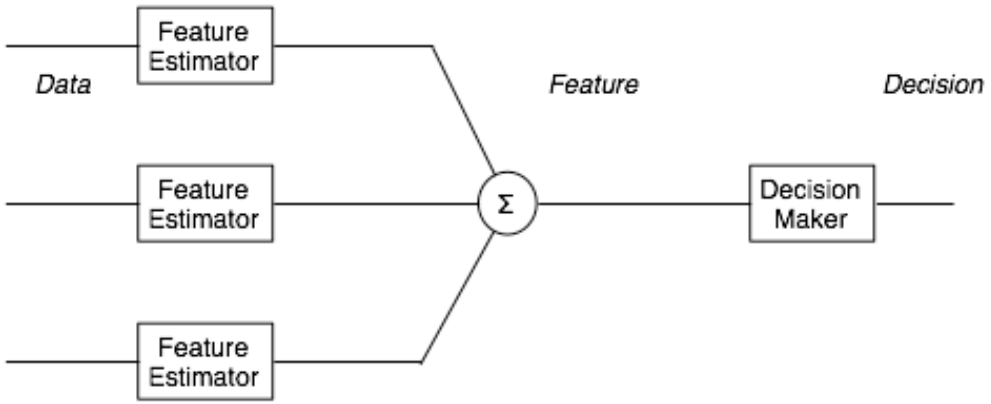
Data-level fusion

- Lowest level of fusion
- Integration of raw observations
- Used to merge data from same type of sensors (2 cameras, e.g.)
- Adds
  - No loss of information
- Cons
  - Too much sensitive to specific nature of sensor to be of real generic use



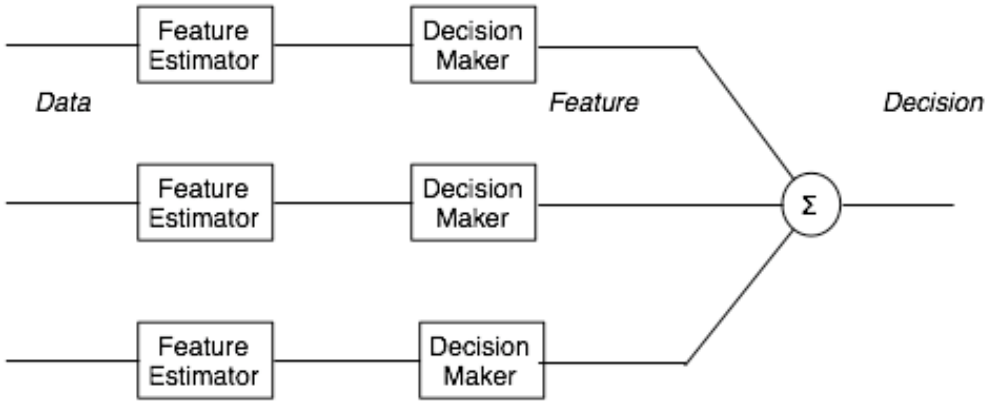
Feature-level fusion

- Also known as « early fusion »
- Each stream of sensory data is first analysed to extract features, then features are fused
- Appropriate for closely coupled and synchronised modalities (e.g. speech and lips)
- Most commonly used methods : ANNs, GMMs, HMMs, etc.



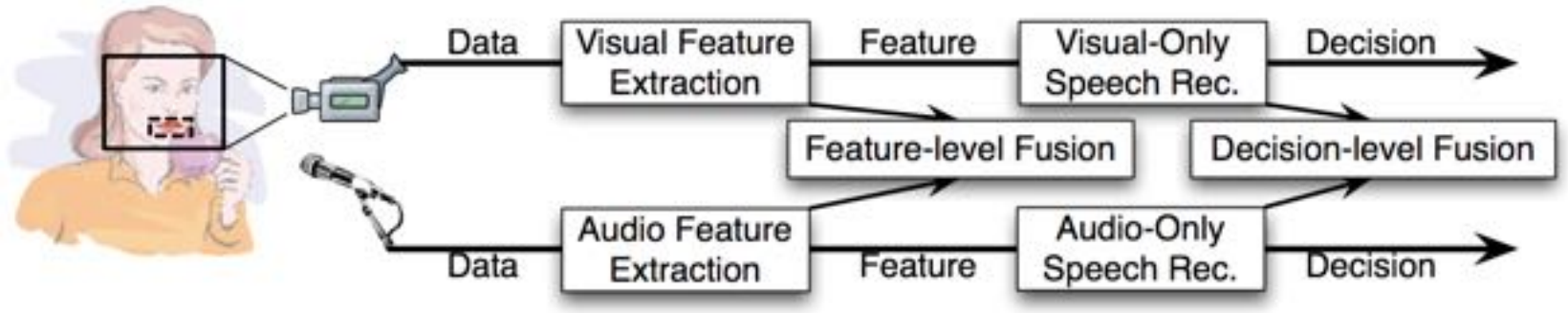
Decision-level fusion

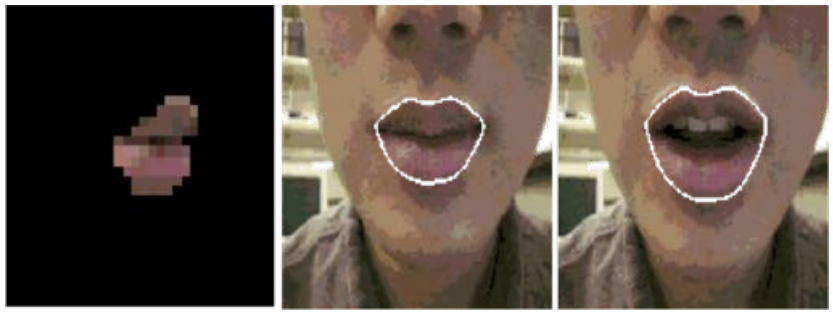
- Also known as « late fusion »
- Most commonly found type of fusion
- Fusion of individual decisions or interpretations (-> works at the semantic level)
  - e.g. speech and gesture
- Robust, but cannot recover loss of information that happened at lower levels



E.g. Augmenting Speech

- Speech Recognition degrades in noisy environments
- Use of Image based modeling of the lips can improve accuracy





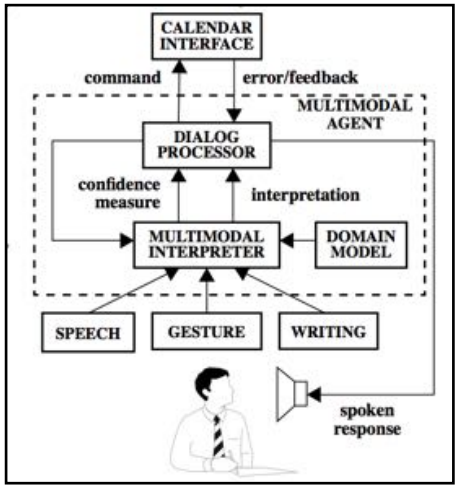
Fusion types comparison

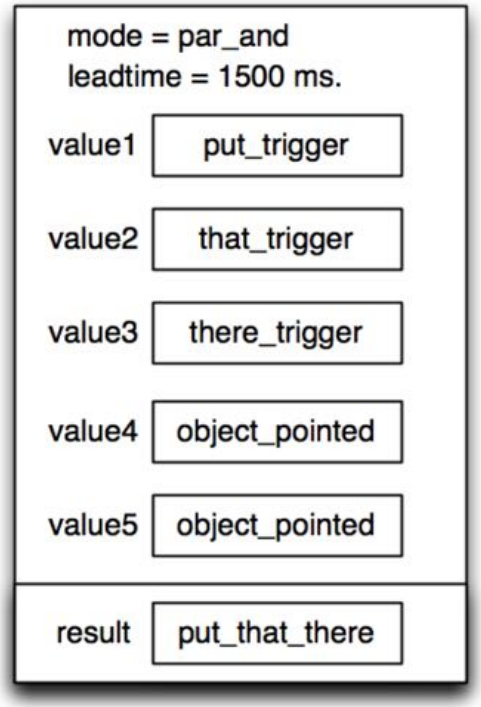
	Data-level fusion	Features-level fusion	Decision-level fusion
Used for…	Raw data of same type	Closely coupled modalities	Weakly coupled modalities
Level of information	Highest level of information detail	Moderate level of information detail	Cannot recover from previous loss of information
Noise/failures sensitivity	Highly susceptible to noise or failures	Less sensitive to noise or failures	Highly resistant to noise or failures
Usage	Not really used for MMI	Used for fusion of particular modes	Most widely used type of fusion

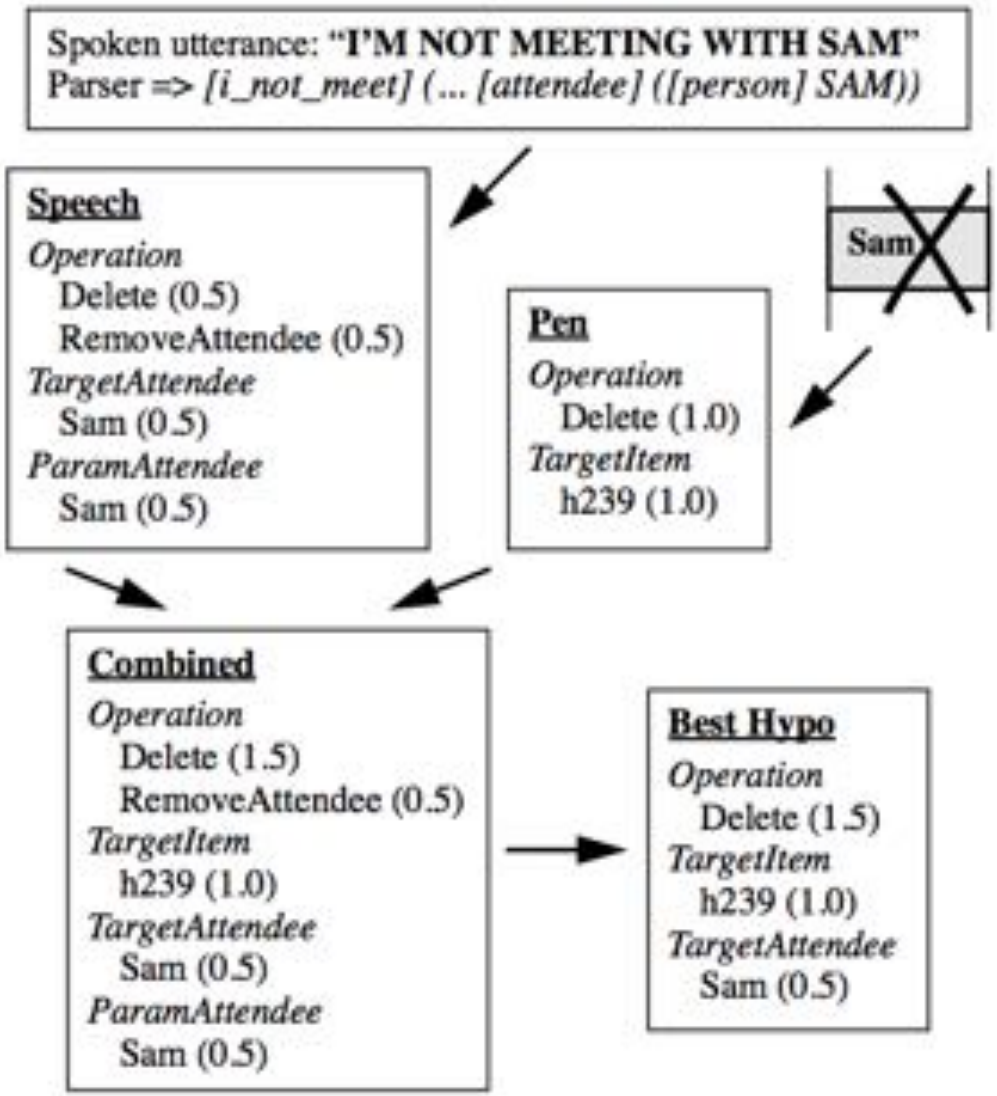
Decision-level fusion methods

- Two main ways to fuse data at decision level :
  - **Frames** : unit of knowledge source describing an object and representing the possible properties of the object's actions or relationships
  - **Unification** : term unification through logic programing mechanisms (rules)
- New hybrid architectures combine symbolic/statistical approach, as an evolution of standard symbolic unification-based approaches

Fusion using meaning frames

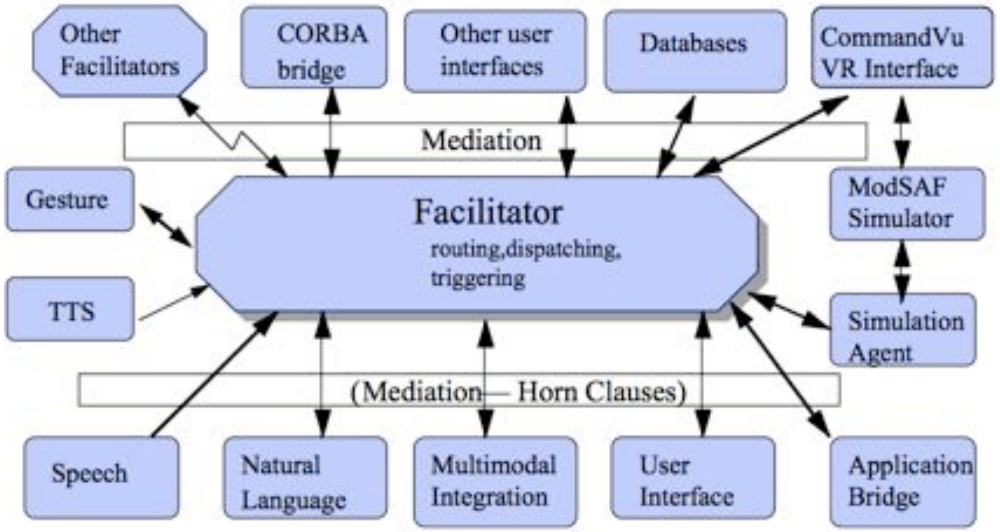


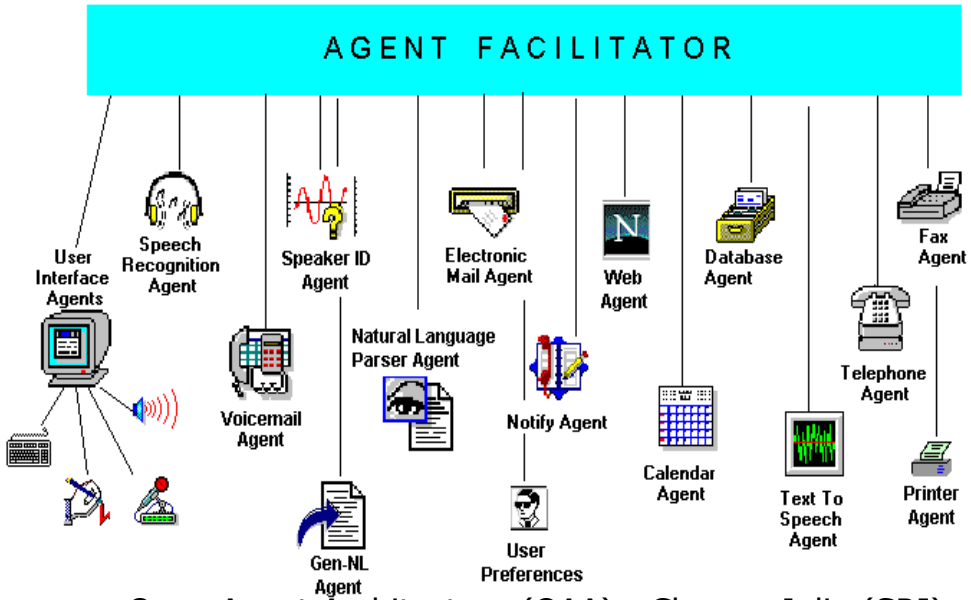




Features of Multimodal Systems

- *Time-sensitive architectures* (need to establish temporal thresholds for time-stamping start & end of each input signal piece)
- *Multi-agent architectures* advantageous for distributing processing & for coordinating many system components (e.g., speech recognition, pen recognition, natural language processing, graphic display, TTS output, application database…)





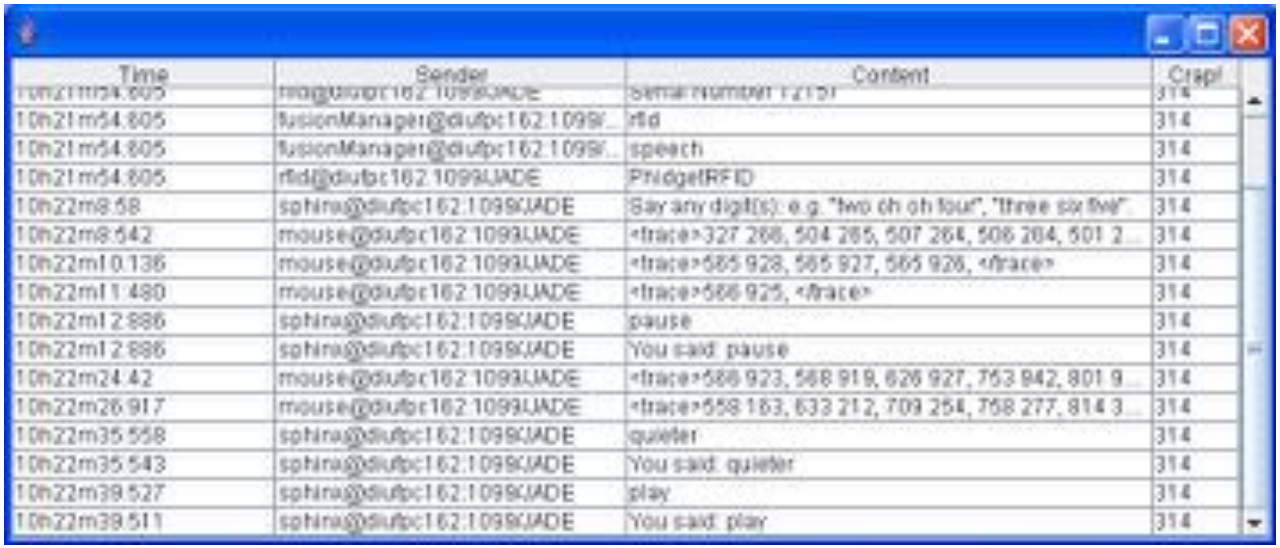
Open Agent Architecture (OAA) - Cheyer, Julia (SRI)
Distributed, Collaborative.

Synchronization

- The time constraint is higly important
- Need to synchronize all the modalities (e.g.: command voice and gesture)
- In which order the command have been entered (interpretation will vary accordingly)?
  - <pointing> Play next track
  - Play <pointing> next track
  - Play next track <pointing> (redondant? synergic?)

+ Delay due to technology (ex.: speech recognition)

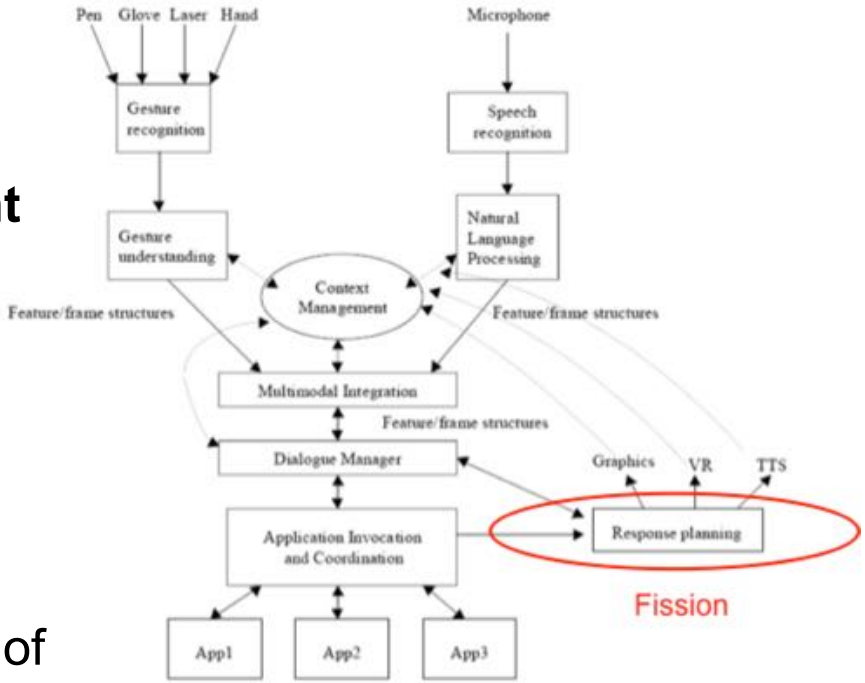
+ Delay due to multimodal system architecture



Time	Sender	Content	Crapl
10h21m54.605	fusionManager@diuftpc162:1099/...	rfid	314
10h21m54.605	fusionManager@diuftpc162:1099/...	speech	314
10h21m54.605	rfid@diuftpc162:1099/JADE	PhidgetRFID	314
10h22m8.58	sphinx@diuftpc162:1099/JADE	Say any digit(s): e.g. "two oh oh four", "three six five"	314
10h22m8.542	mouse@diuftpc162:1099/JADE	<trace>327 266, 504 265, 507 264, 508 264, 501 2...	314
10h22m10.136	mouse@diuftpc162:1099/JADE	<trace>565 928, 565 927, 565 926, </trace>	314
10h22m11.480	mouse@diuftpc162:1099/JADE	<trace>566 925, </trace>	314
10h22m12.886	sphinx@diuftpc162:1099/JADE	pause	314
10h22m12.886	sphinx@diuftpc162:1099/JADE	You said: pause	314
10h22m24.42	mouse@diuftpc162:1099/JADE	<trace>566 923, 568 919, 626 927, 753 942, 801 9...	314
10h22m26.917	mouse@diuftpc162:1099/JADE	<trace>558 183, 633 212, 709 254, 758 277, 814 3...	314
10h22m35.558	sphinx@diuftpc162:1099/JADE	quieter	314
10h22m35.543	sphinx@diuftpc162:1099/JADE	You said: quieter	314
10h22m39.527	sphinx@diuftpc162:1099/JADE	play	314
10h22m39.511	sphinx@diuftpc162:1099/JADE	You said: play	314

Fission

Resolution of the difference - activate the most suitable referent



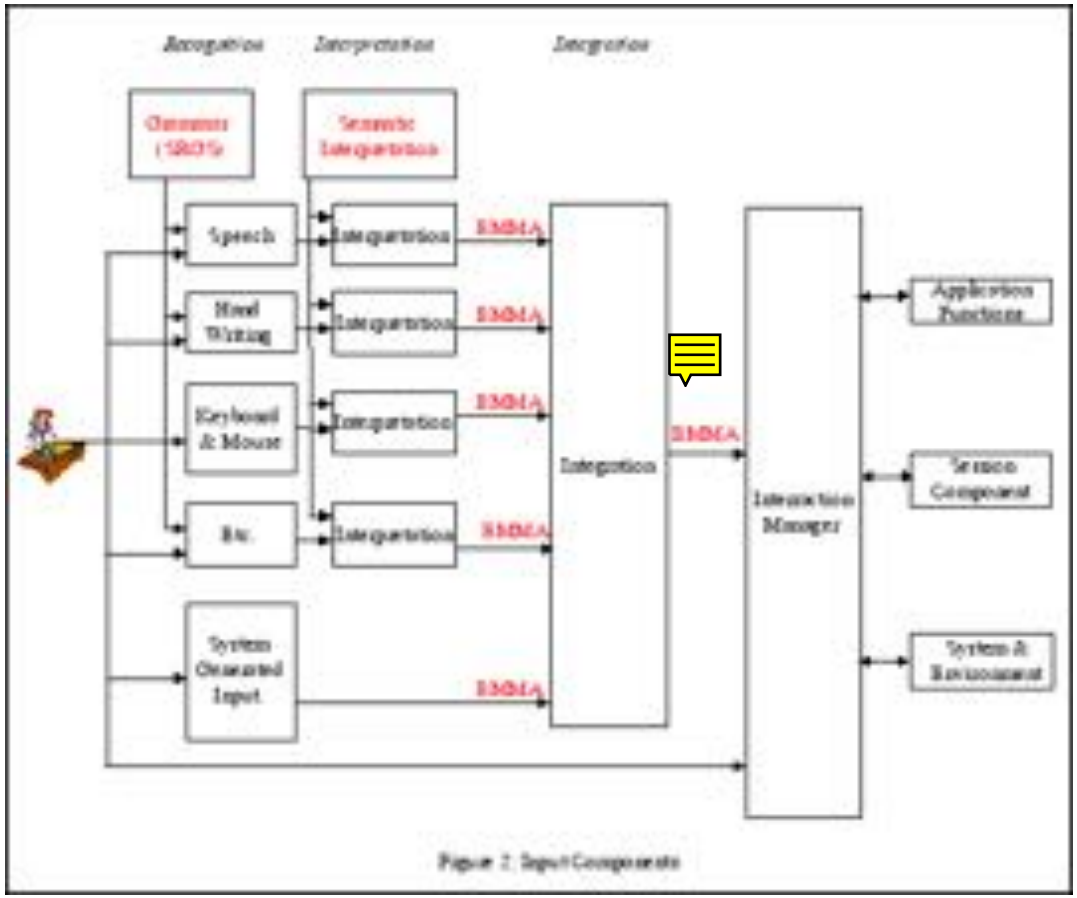
Building an abstract message through combination of channels;

1. Message content selection and structuring;
2. Modality selection;
3. Output coordination: coordination of the output on each channel to form a coherent message.

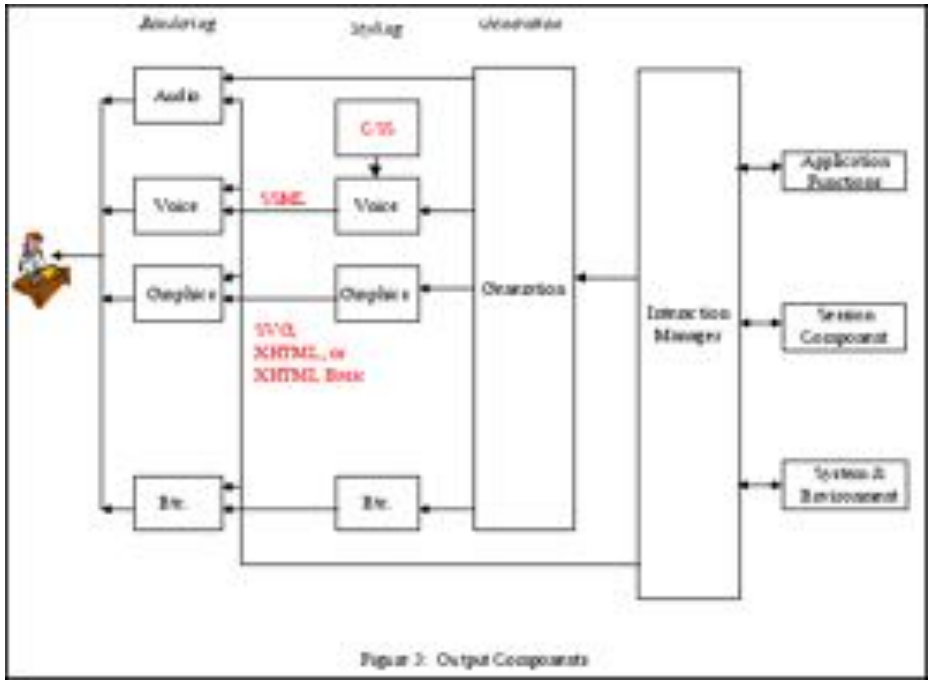
- Classic fission techniques :
  - Screen-based interfaces
  - Visualization
  - Speech synthesis
  - Sonification
  - Embodied conversational agents
  - Taking benefit of human auditory, visual, tactile, etc. senses

State of the art and related technologies and formats

W3C Multimodal Framework



Input-side



Output-side

W3C EMMA

```
<emma:emma version="1.0"
   xmlns:emma:="http://www.w3.org/2003/04/emma"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.w3.org/2003/04/emma
   http://www.w3.org/TR/emma/emma10.xsd"
   xmlns="http://www.example.com/example">
  <emma:one-of id="r1" emma:start="1087995961542" emma:end="1087995963542">
   <emma:interpretation id="int1" emma:confidence="0.75"
   emma:tokens="flights from boston to denver">
     <origin>Boston</origin>
     <destination>Denver</destination>
   </emma:interpretation>
   <emma:interpretation id="int2" emma:confidence="0.68"
   emma:tokens="flights from austin to denver">
     <origin>Austin</origin>
     <destination>Denver</destination>
   </emma:interpretation>
  </emma:one-of>
</emma:emma>
```

EMMA Declaration

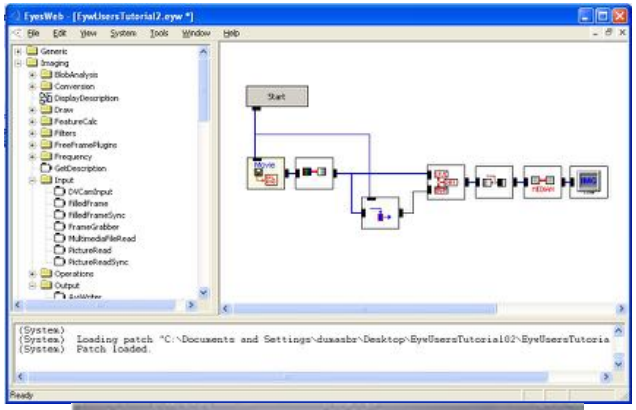
2 interpretations in the same time interval

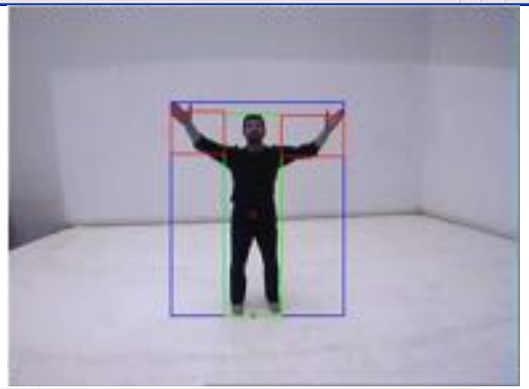
Application-dependant language

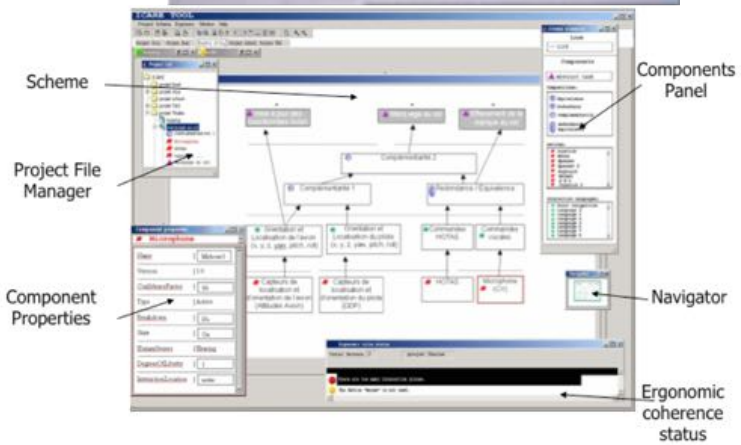
- For more details… http://www.w3.org/TR/mmi-framework/

Toolkits for multimodal interfaces

- Toolkits currently available as open source:
  - *EyesWeb*
    - ✓ Graphical Programmation (GUI)
    - ✓ Mainly used in dance and music research (for the moment?)
    - ✓ Well established community
    - ✓ http://www.eyesweb.org/
  - *Open Interface*
    - ✓ Written in C/C++
    - ✓ Based on software components
    - ✓ http://www.openinterface.org/
  - *HephaisTK*
    - ✓ DIVA Group/University of Fribourg project!
    - ✓ Written in Java, Built upon a software agent middleware
    - ✓ Fusion/fission engines, EMMA support
    - ✓ http://diuf.unifr.ch/diva/projects/hephaistk/
  - *Squidy* (2009, Konstanz, DE)
  - *Mudra* (2010, VUB, BE)

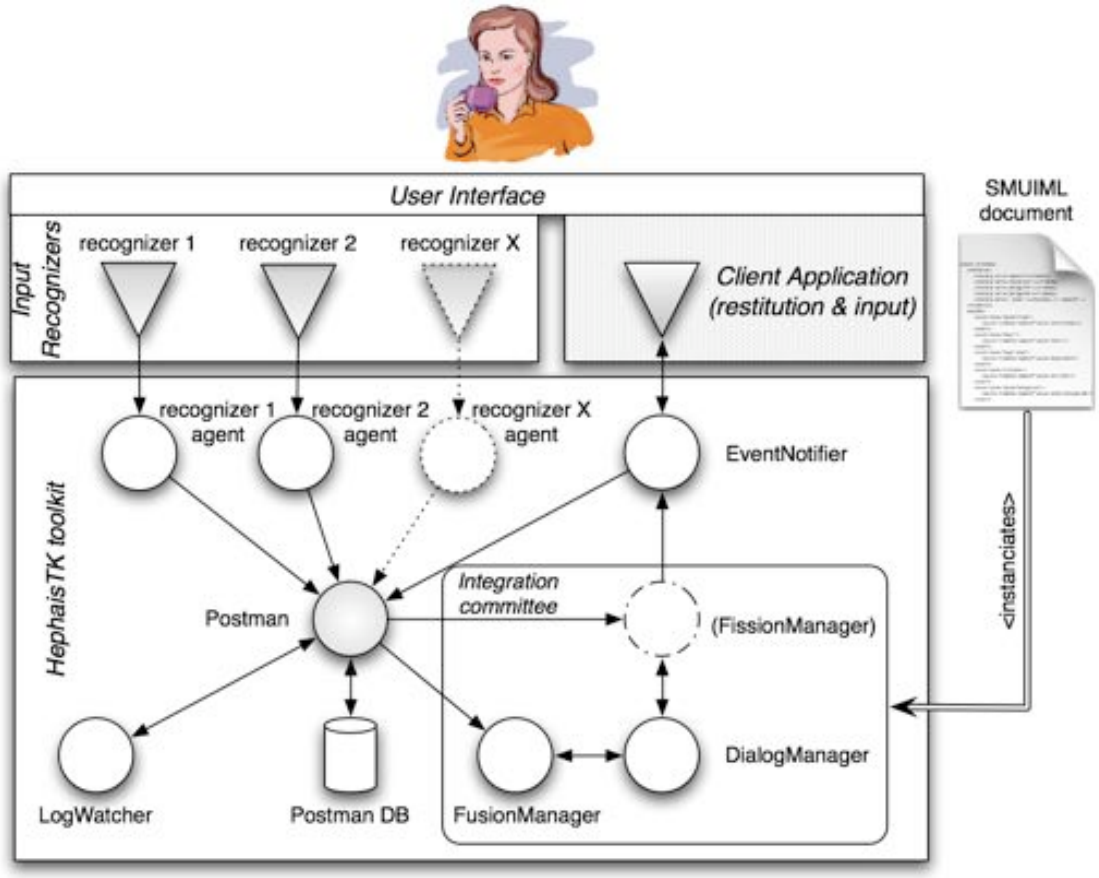






HephaisTK

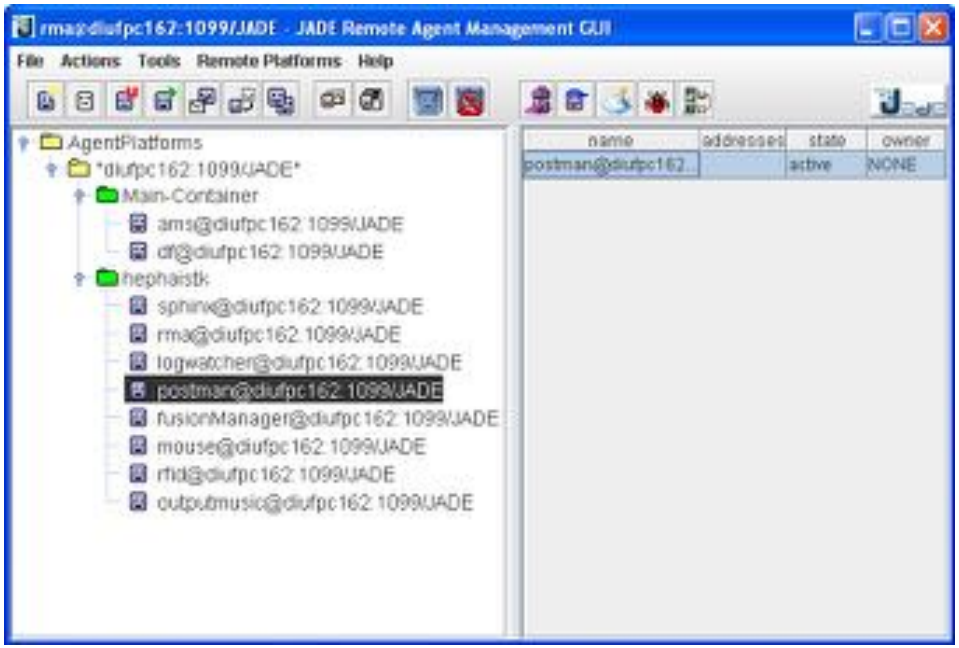
- Software agents-based framework, event based
- Different fusion algorithms integrated
- Description of the dialogue through the SMUIML language



http://www.hephaistk.org

Dumas, B., Lalanne, D., Guinard, D., Ingold, R., Koenig, R., "Strengths and Weaknesses of Software Architectures for the Rapid Creation of Tangible and Multimodal Interfaces". In proceedings of 2nd international conference on Tangible and Embedded Interacton (TEI 2008), pp. 47-54

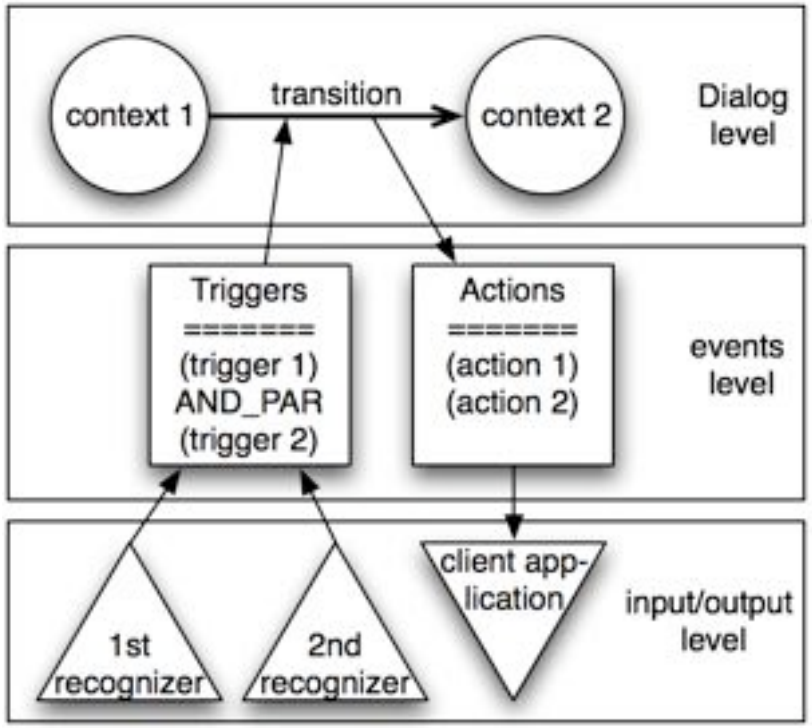
HephaisTK



Time	Sender	Content	Crap!
10h21m54.805	fusionManager@diufpc162:1099/..	rfid	314
10h21m54.805	fusionManager@diufpc162:1099/..	speech	314
10h21m54.805	rfid@diufpc162:1099/JADE	PhidgetRFID	314
10h22m8.58	sphinx@diufpc162:1099/JADE	Say any digit(s): e.g. "two oh oh four", "three six five".	314
10h22m8.542	mouse@diufpc162:1099/JADE	<trace>327 268, 504 265, 507 264, 506 264, 501 2...	314
10h22m10.136	mouse@diufpc162:1099/JADE	<trace>565 928, 565 927, 565 926, </trace>	314
10h22m11.480	mouse@diufpc162:1099/JADE	<trace>566 925, </trace>	314
10h22m12.886	sphinx@diufpc162:1099/JADE	pause	314
10h22m12.886	sphinx@diufpc162:1099/JADE	You said: pause	314
10h22m24.42	mouse@diufpc162:1099/JADE	<trace>586 923, 568 919, 626 927, 753 942, 801 9...	314
10h22m26.917	mouse@diufpc162:1099/JADE	<trace>558 183, 633 212, 709 254, 768 277, 814 3...	314
10h22m35.558	sphinx@diufpc162:1099/JADE	quieter	314
10h22m35.543	sphinx@diufpc162:1099/JADE	You said: quieter	314
10h22m39.527	sphinx@diufpc162:1099/JADE	play	314
10h22m39.511	sphinx@diufpc162:1099/JADE	You said: play	314

SMUIML

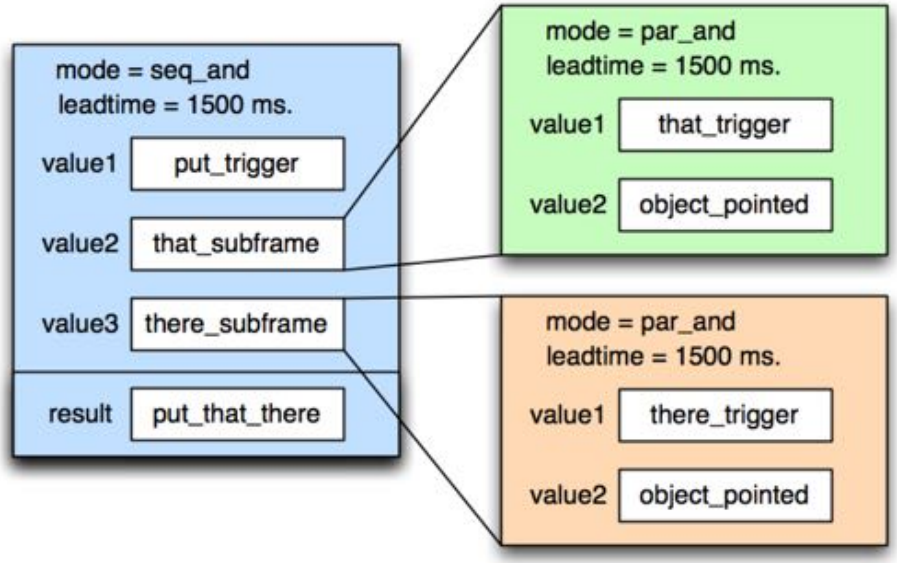
- Synchronized Multimodal User Interface Modeling Language
- Scripting language for HephaisTK
- Wished characteristics:
  - easy-to-read
  - expressive
  - CARE properties
- 3 levels
  - Input/output
  - Events
  - Dialog



Dumas B, Signer B, Lalanne D. (2013). "A Graphical Editor for the SMUIML Multimodal User Interaction Description Language", Science of Computer Programming, Volume 86, 15 June 2014, pp. 30-42.

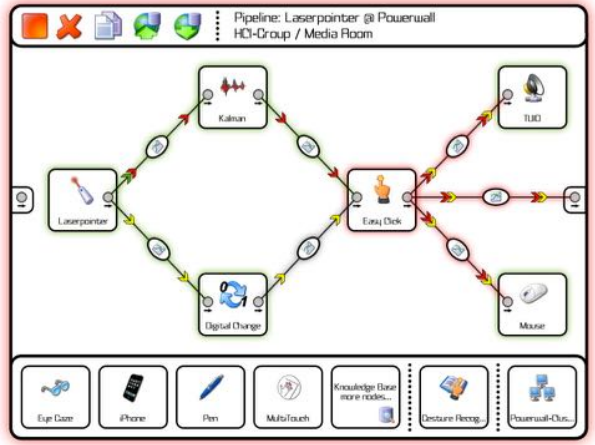
“Put-That-There” Example in SMUIML

```
<context name="start">
  <transition leadtime="1500">
    <seq_and>
      <trigger name="put_trigger" />
      <transition>
        <par_and>
          <trigger name="that_trigger" />
          <trigger name="object_pointed_event" />
        </par_and>
      </transition>
      <transition>
        <par_and>
          <trigger name="there_trigger" />
          <trigger name="object_pointed_event" />
        </par_and>
      </transition>
    </seq_and>
    <result action="put_that_there_action" />
  </transition>
</context>
```

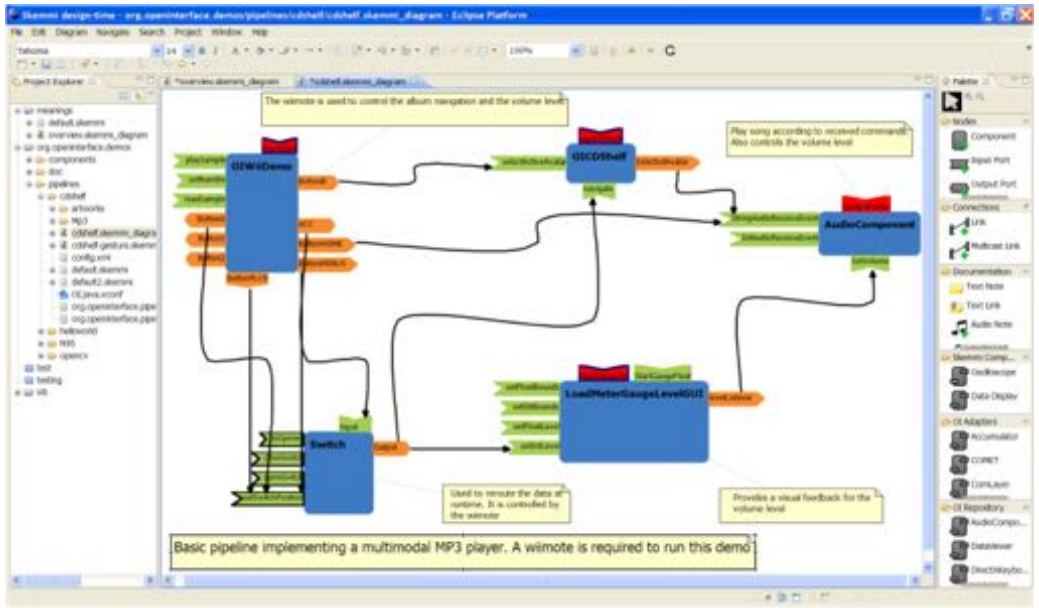


GUIs

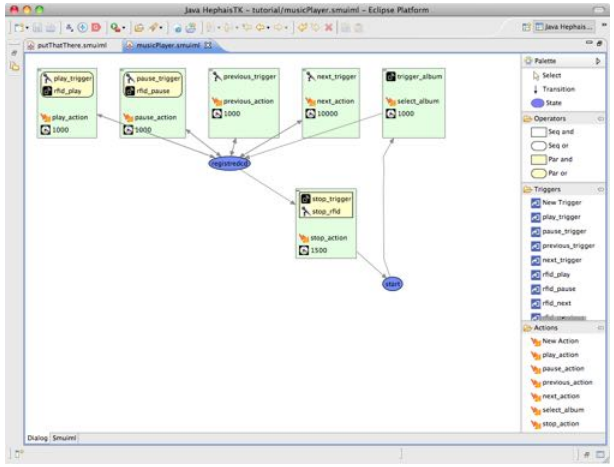
Squidy GUI



openInterface



HephaisTK



References

- Cohen, P., Johnston, M., McGee, D., Oviatt, S., Pittman, J.,?Smith, I., Chen, L. and Clow, J. Quickset: Multimodal interaction for distributed applications. Proceedings of the Fifth?ACM International Multimedia Conference, New York, NY:?ACM Press, 1997, 31-40.
- Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J. and Young, R.: Four Easy Pieces for Assessing the Usability of Multimodal Interaction: The CARE properties, Proceedings of the INTERACT'95 conference, S. A. Arnesen & D. Gilmore Eds., Chapman&Hall Publ., Lillehammer, Norway, June 1995, pp. 115-120.
- Dumas, B., Lalanne, D., Oviatt, S. Multimodal Interfaces: A Survey of Principles, Models and Frameworks. In Denis Lalanne, Jürg Kohlas eds. (2009). Human Machine Interaction, LNCS 5440, Springer-Verlag, Berlin/Heidelberg, pp. 3-27.
- Lalanne, D., Nigay, L., Palanque, P., Robinson, P., Vanderdonckt, J., Ladry, J-F. Fusion Engines for Multimodal Interfaces: a survey. International Conference on Multimodal Interfaces and Workshop on Machine Learning for Multi-modal Interaction (ICMI-MLMI 2009), Cambridge, Massachusetts, USA, ACM, 2009.
- Nigay & Coutaz 1993 : Nigay, L., Coutaz, J. A design space for multimodal interfaces: concurrent processing and data fusion, in Proc. INTERCHI'93 Human Factors in Computing Systems (Amsterdam, April 24-29, 1993), ACM Press, pp. 172-178.
- Sharma, R., Pavlovic, V.I., & Huang, T.S. (1998). Toward multimodal human-computer interface. Proceedings IEEE, 86(5) [Special issue on Multimedia Signal Processing], 853-860.
- Vo, M. T. and C. Wood. 1996. Building an application framework for speech and pen input integration in multimodal learning interfaces. International Conference on Acoustics, Speech, and Signal Processing 1996, Atlanta, GA.
- Wu, Lizhong, Oviatt, S. L., Cohen, P. R., Multimodal Integration-A Statistical View, IEEE Transactions on Multimedia, Vol. 1, No. 4, December 1999, pp. 334-341.

What you should know now…

- How can be represented multimodal communication?
- What are the fundamental problems to solve with multimodal systems?

- What are the CARE and CASE models?
- What are the general architectures of a multimodal system?
- What are the 3 levels associated with multimodal fusion?
- What is multimodal fission?