Platforms and methods for multi modal system architectures

Fall 2016 Course for VGIS9 and AAT9

Lecture 1: Introduction to Course and
Architectures and Paradigms for Multimodal platforms

Lars Bo Larsen

Architectures and Paradigms for Multimodal platforms

Outline:

- Introduction to the MMUIP Course
  - Lectures, Readings, mini projects, exam
- Architectures for Decision Level / Late-Fusion systems
  - Agents based systems
  - Frame-based integration

Course teachers

- Zheng-Hua Tan, Associate Professor (zt)
  - Main course lecturer, mini projects, readings
- Lars Bo Larsen, Associate Professor (lbl)
  - Course lecturer
- Xiaodong Duan, Ph.D. Student (xd)
  - Course lecturer, mini projects, readings

Course Overview

Lectures 1-6 (Friday Mornings, except this week)

1: Introduction & late fusion architectures (lbl)
2: ROS, HRI (xd)
3: Data/mid-level fusion (zt)
4: Talking to computers (zt)
5: Eyegaze tracking (zt)
6: Speech as a modality in MMUI (lbl)

Course Overview Ctd.

- Readings part (Lectures 7-10)
  - 1-2 students prepare a lecture and present to the class
- Mini project presentations in (Lecture 11)
  - Each mini project groups presents in class
- Exam – based on Mini project (lecture 12)
  - The exam will be based on your mini project handins, but can include topics from the lectures – just like a regular project exam.
  - The exam is placed here to accomodate those going abroad

Readings (lectures by students)

- Groups of 1-2 students each do a class lecture
  - Choose a topic and identify corresponding literature together with a course lecturer
- Examples of topics presented previous years:
  - Fusion of multiple visual cues
  - Fusion of audio and visual cues (speech recognition and guesture)
  - Eyegaze tracking and other modalities
  - Biometric sensing
  - Ideas out of your own projects

Mini Projects

- Form 2 persons mini project groups (can be same as lecture groups / project group)
- Choose topic with course lectures. All groups are assigned a “mini project supervisor”)
- The topic must be relevant to course goals (ie address a multi modal interaction issue) and include implementation and test
- The mini project includes a presentation and demo by the end of the course and is documented in a short (<20 page) report.

Mini Projects examples (from last year)

Iskren and Octavian:

- Measurements of cognitive load when solving cross-modal tasks using EEG and eye-tracking.
- Joint lecture on EEG signals and their applications

Morten and Mark:

- Multi-modal Fusion of Lip Motion Recognition and Speech Recognition
- Joint Lecture on Traffic Sign Recognition System

Ivan:

- Speech and Gesture interaction in virtual 3D environment
- Lecture on Hand and hand motion tracking. Gesture recognition

Thiemo:

- Fusion of video stream information from visual and infrared light
- Lecture: Multimodal Image Registration

Information fusion in multi modal systems

- Data level fusion
- Feature level fusion
- Decision level fusion

Will be the topic for a lecture in two weeks time by Zheng-Hua

Decision-level fusion

- Aka, late fusion, late intergration
- Integrates common meaning representations derived from different modalities into a combined final interpretation.
- Utilizes independent classifiers, one for each stream, which can be trained independently.
- The final decision is reached by combining the partial outputs of the unimodal classifiers.
- Requires a common meaning representation framework for all modalities used and a well-defined operation for integrating partial meanings.

Decision-level fusion

- Advantages:
  - Since the input types can be recognized independently, they do not have to occur simultaneously.
  - Flexible asynchronous architecture.
  - The software development process is simpler.
  - Independendent "off-the-shelf"-modules can be integrated
- Disadvantage
  - The correlations between the channels are taken into account only later during the integration step.
  - Global optimisation not possible

Decision-level (late) fusion

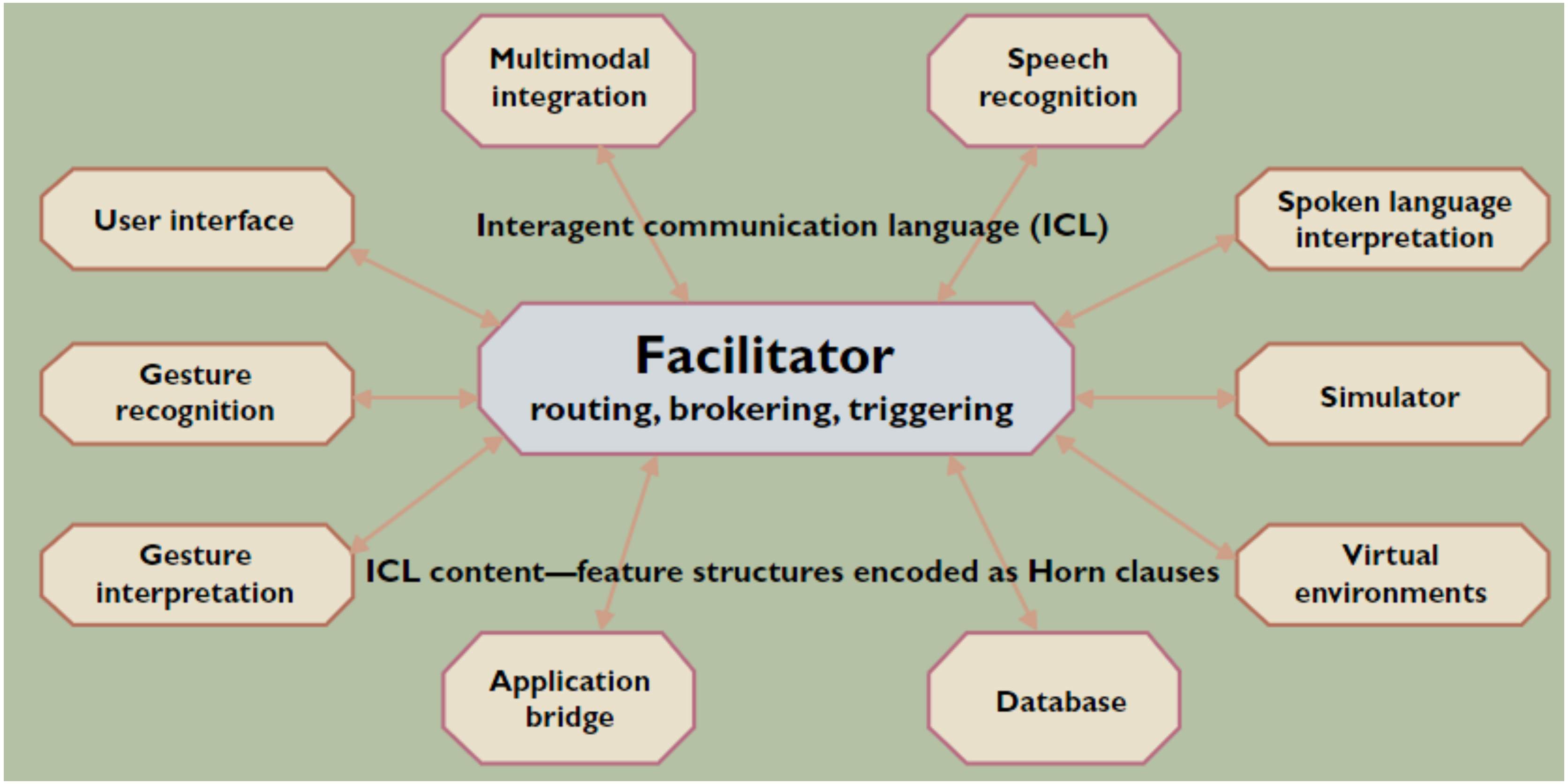
- Applied most often for multimodal HCI.
- Experimental studies show that a late integration approach (i.e., a decision-level fusion) might in some cases provide higher recognition scores than an early integration approach.

Late fusion architectures

Artificial Intelligence legacies:

- Multi Agent based
- Blackboard Oriented architectures
- Frame-based paradigm

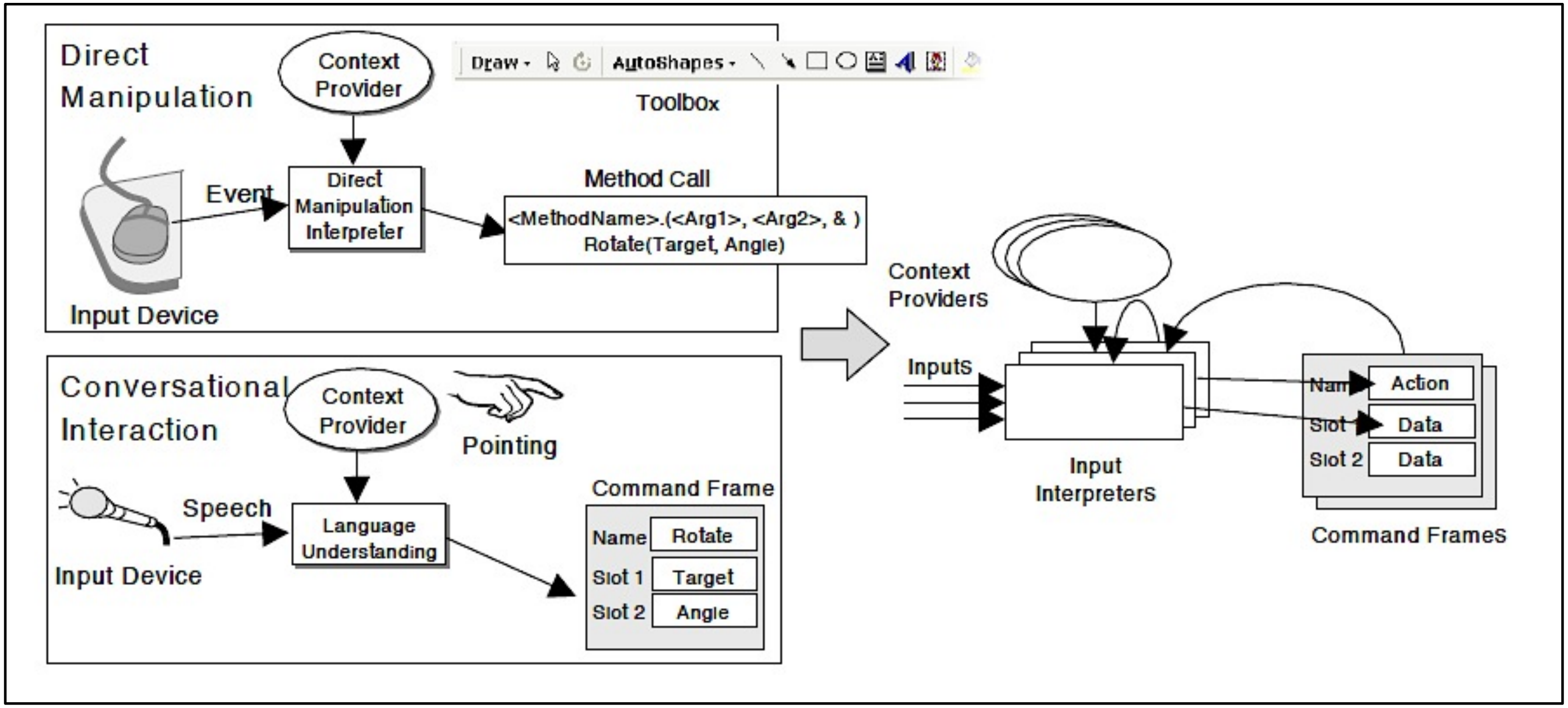
Multi Agent Architecture



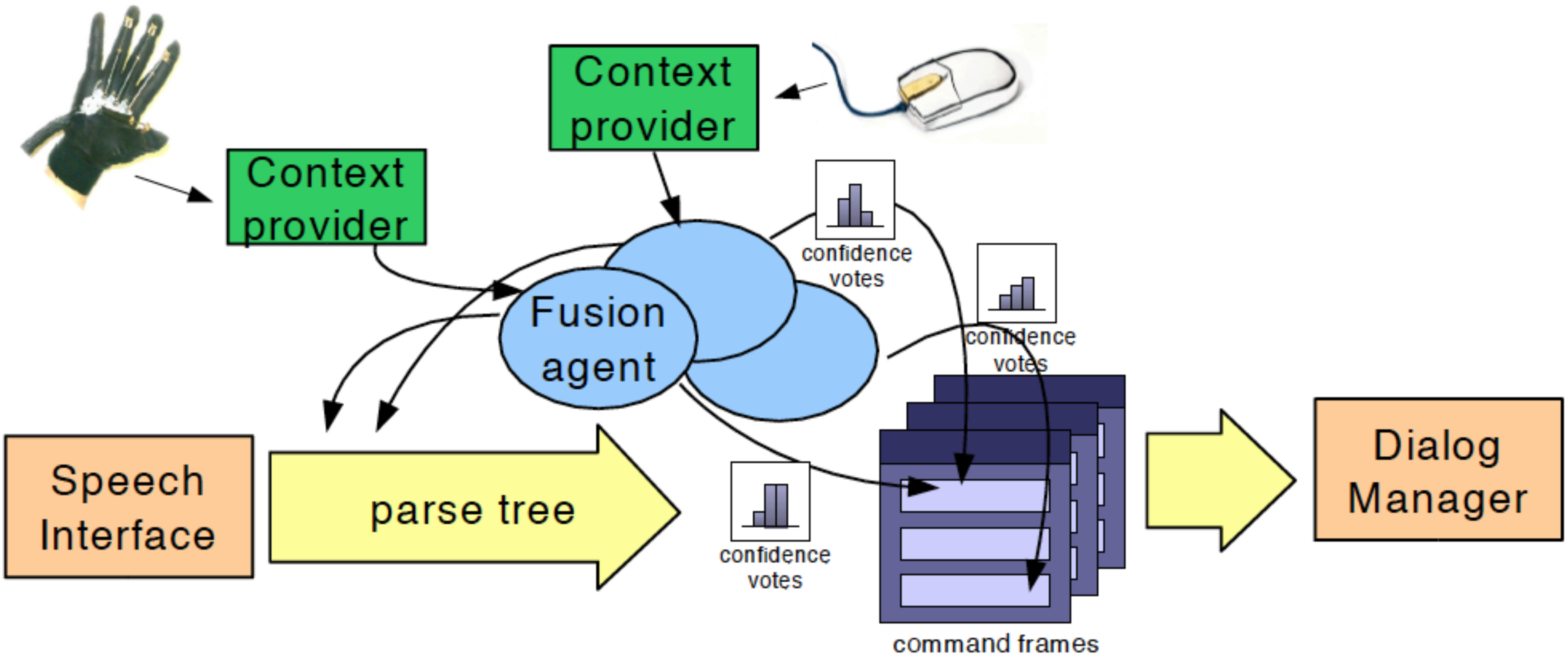
(Oviatt et al. 2000)

- Quickset System (mentioned in UED course)

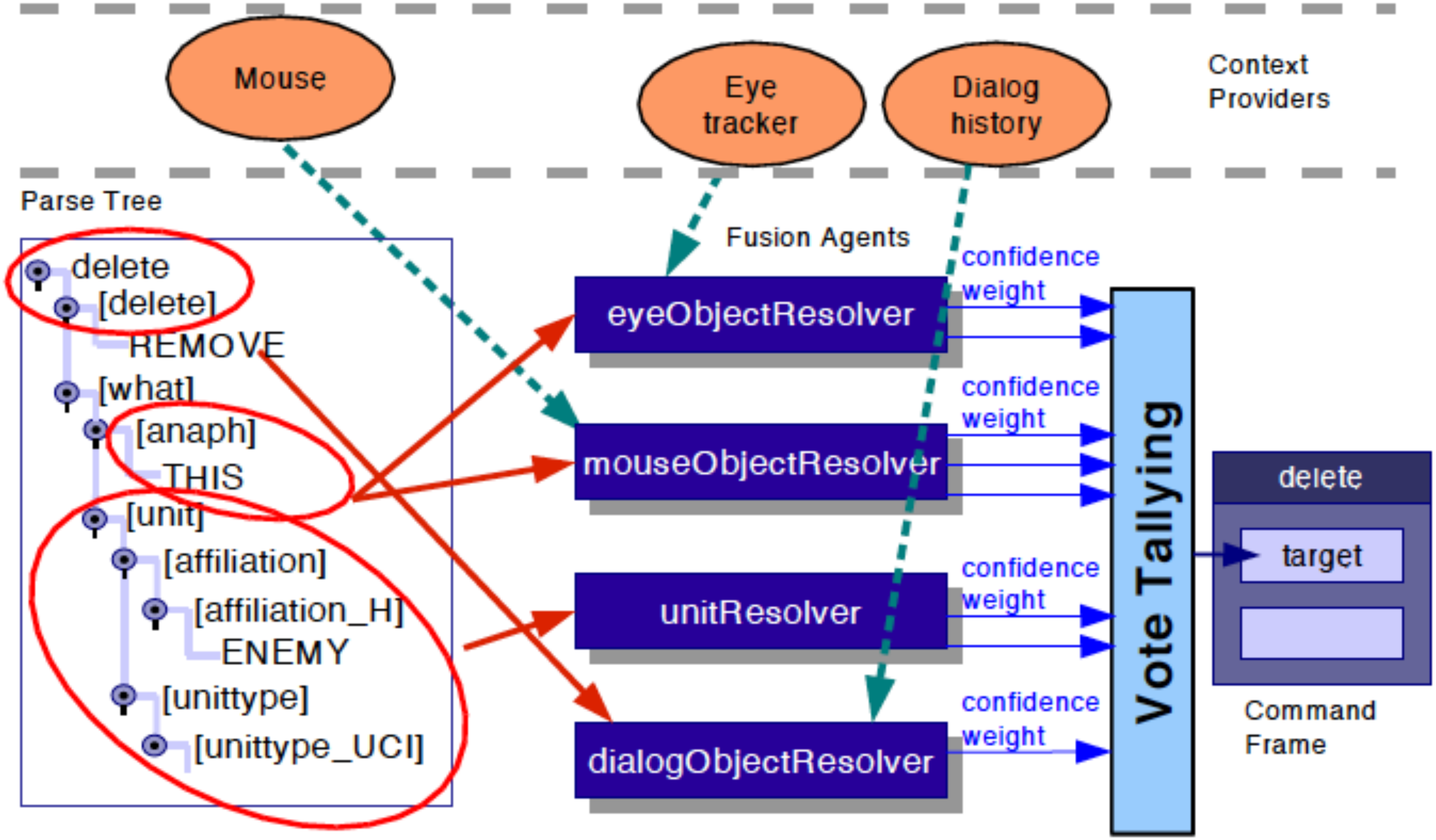
D.M. Compared to Conversational interaction – in both cases knowledge of the context is necessary to interpret commands



Close-up – Fusion agents combines (or inserts) contextual information into the speech input and forms fused “command frames”



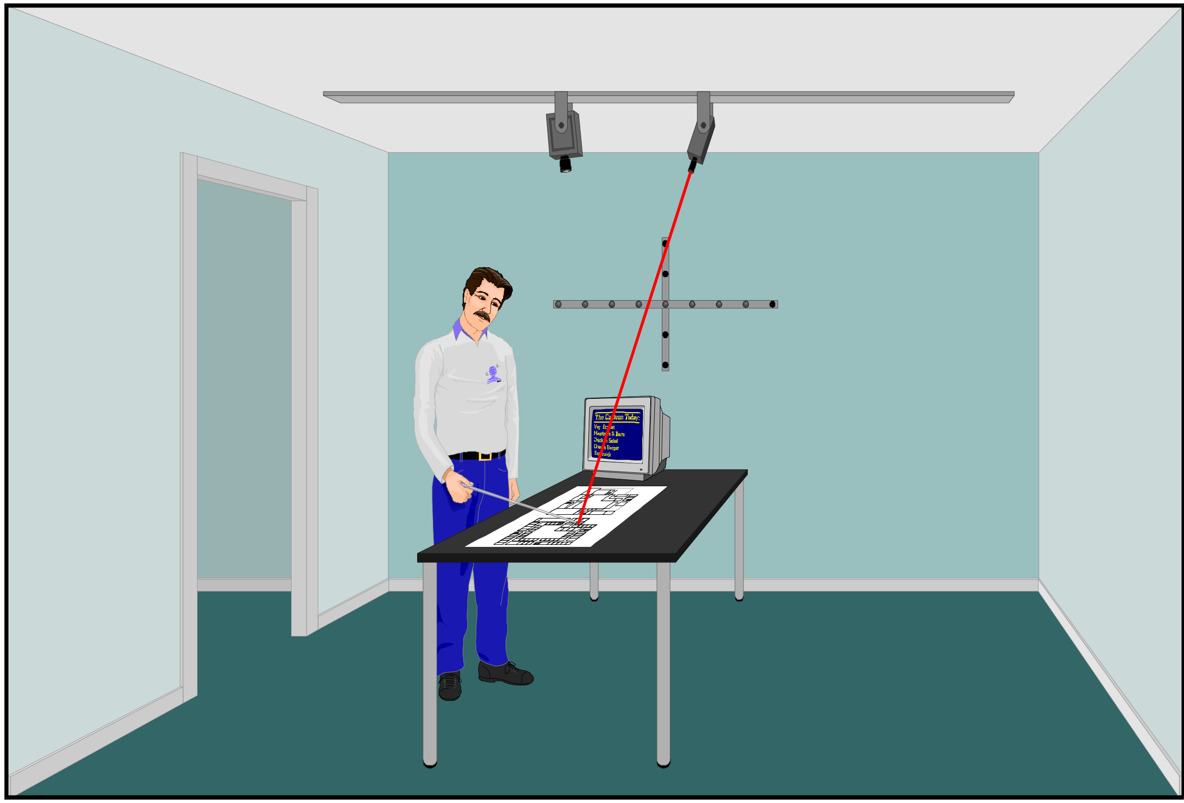
Details of the fusion process



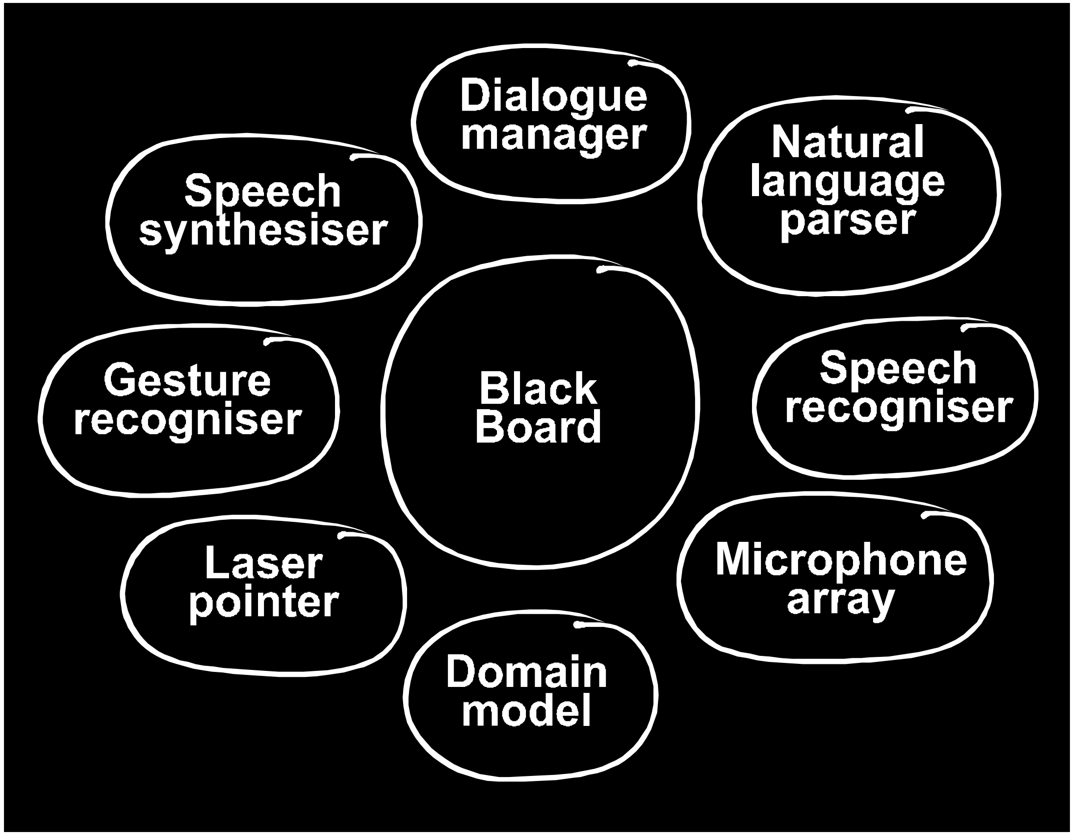
Intellimedia Workbench

Initial Application Domain: A Campus Information System

- Accepts speech and pointing inputs
- Generates pointing (virtual laser pointer) and spoken output



Blackboard Architecture



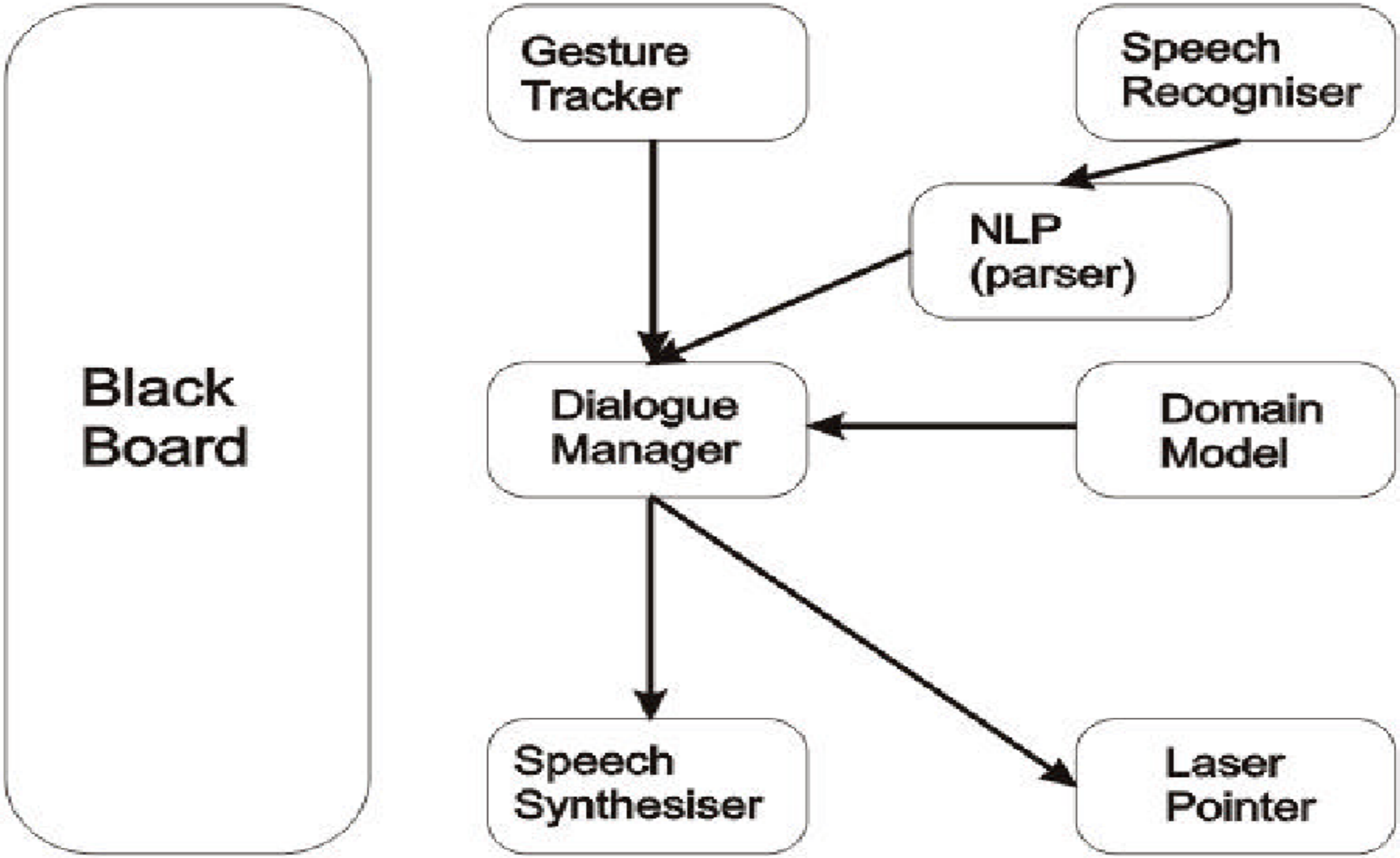
The AAU ”Intellimedia Workbench” utilised an open Black Board Architecture

Larsen, Moeslund et al 1999

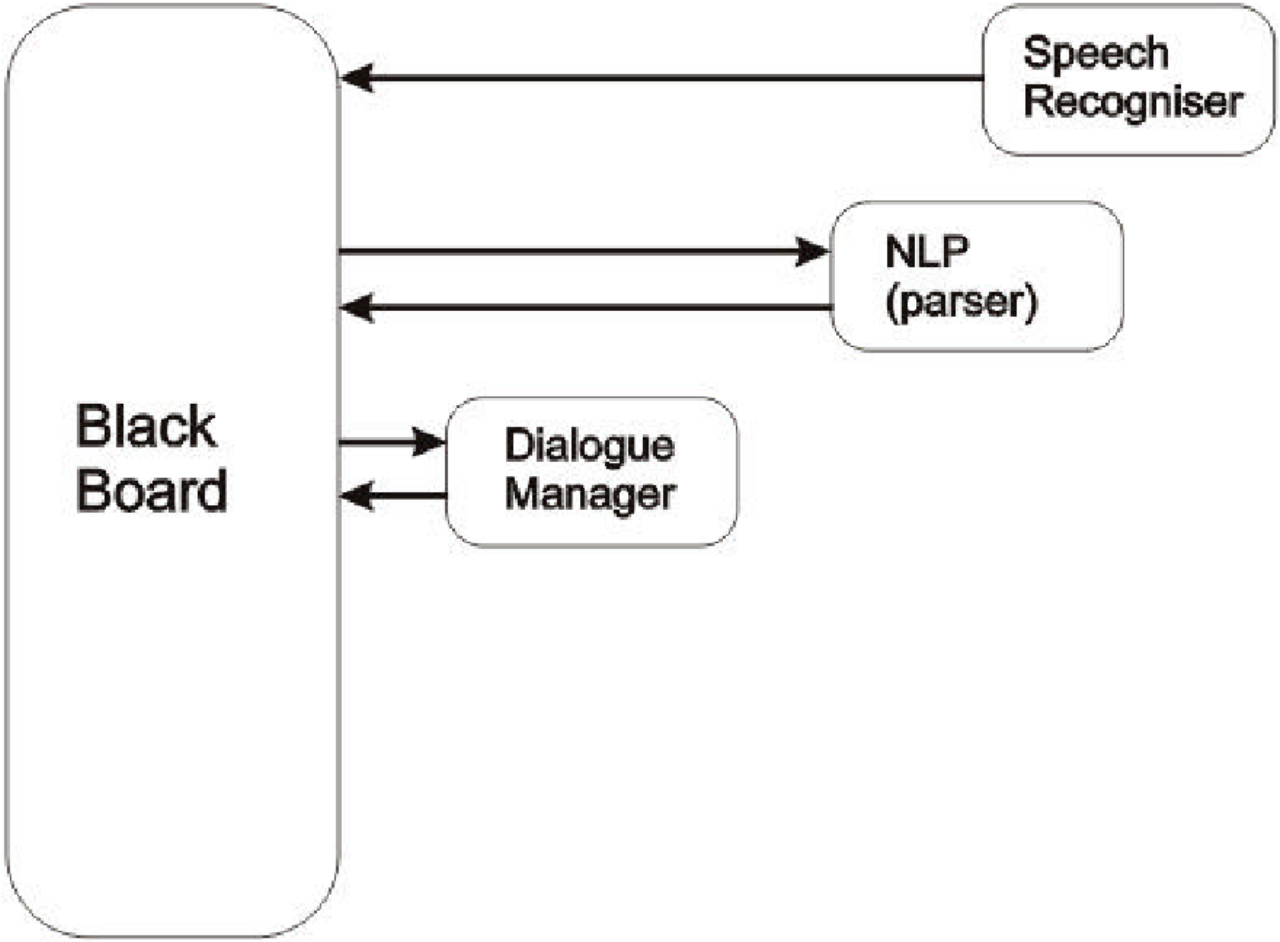
The Blackboard

- A blackboard is a central knowledge repository
- The blackboard stores semantic representations (frames) produced by each of the other modules and keeps a history of these over the course of an interaction. All modules communicate through the exchange of semantic representations with each other or the blackboard.

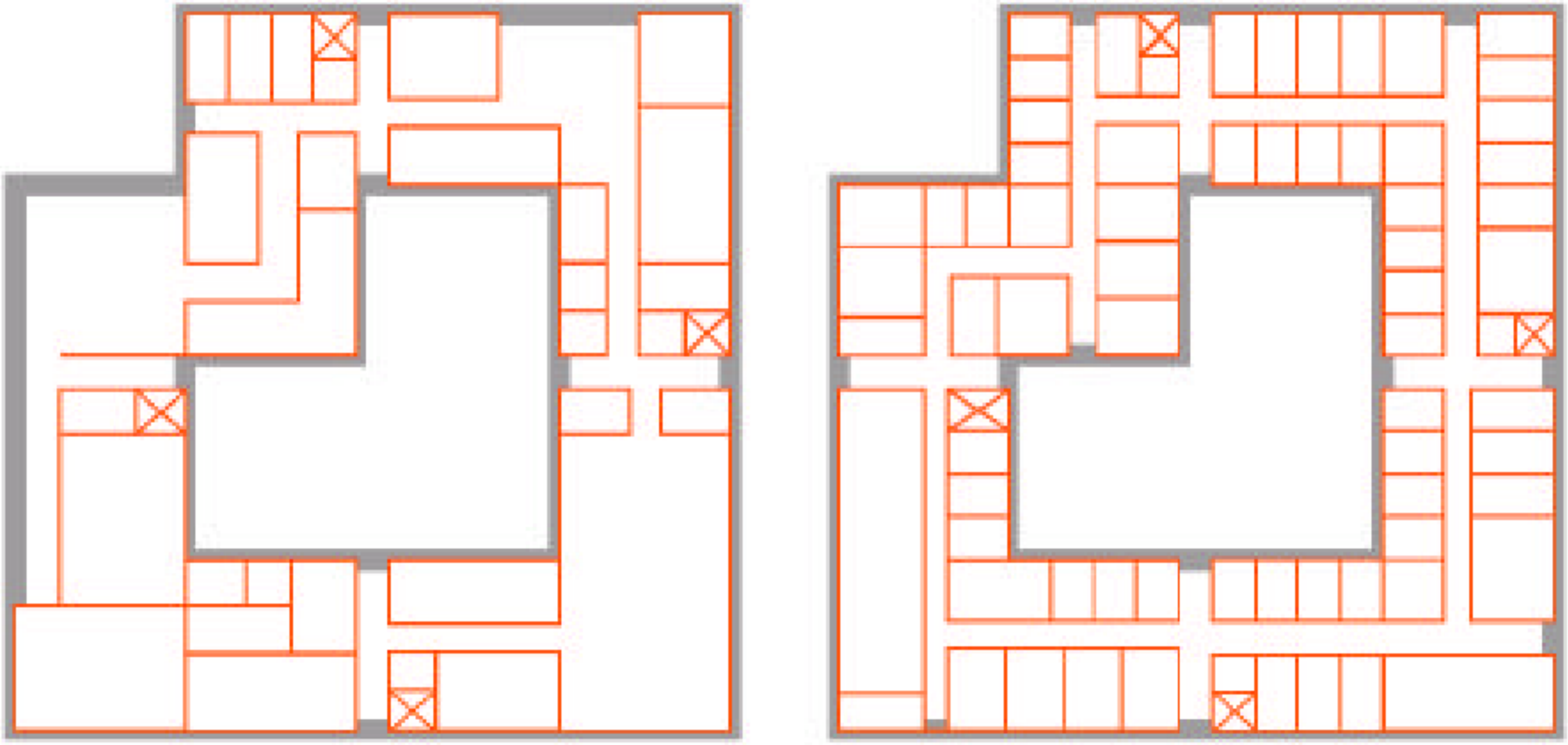
Blackboard Architecture



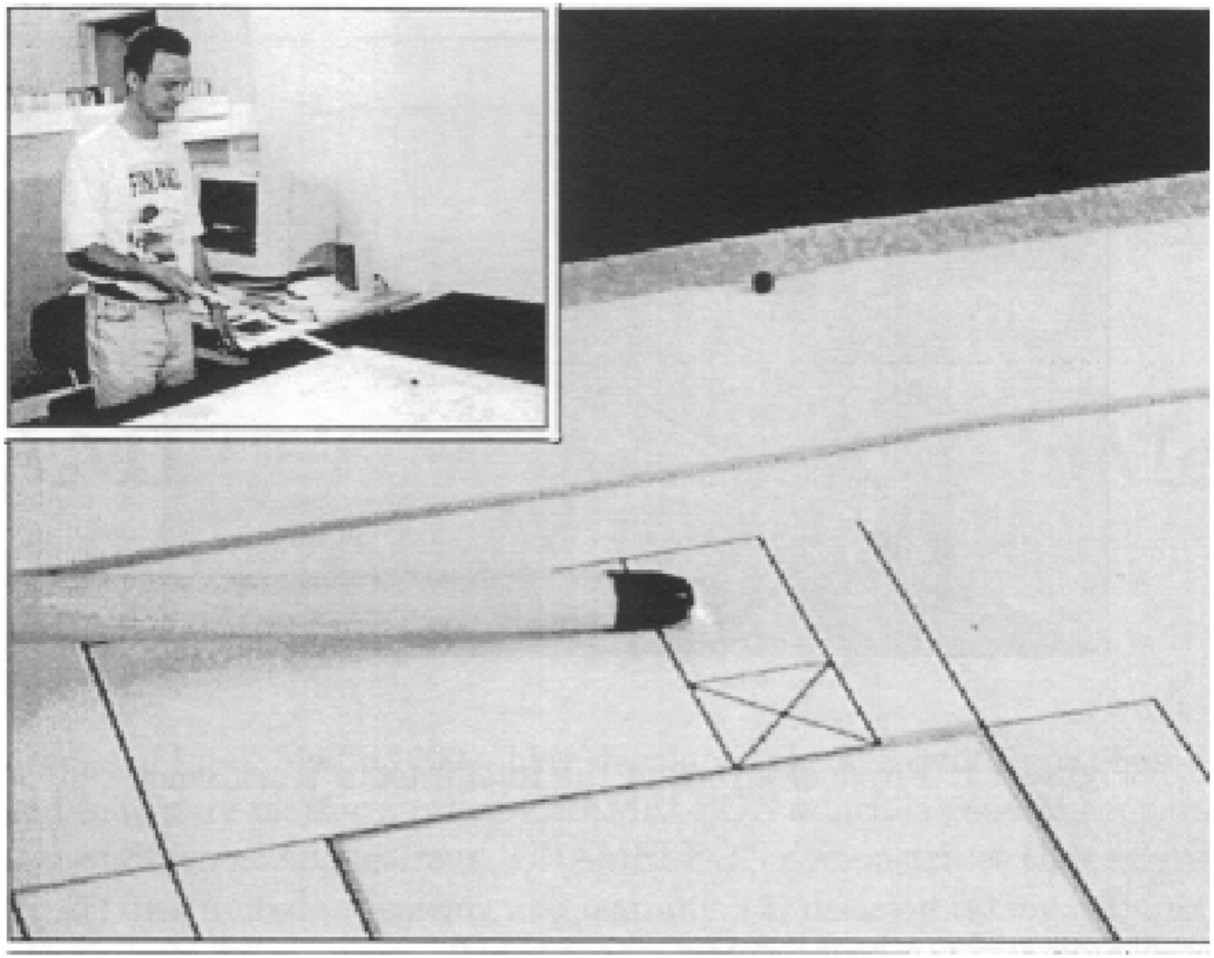
Information flow with the blackboard



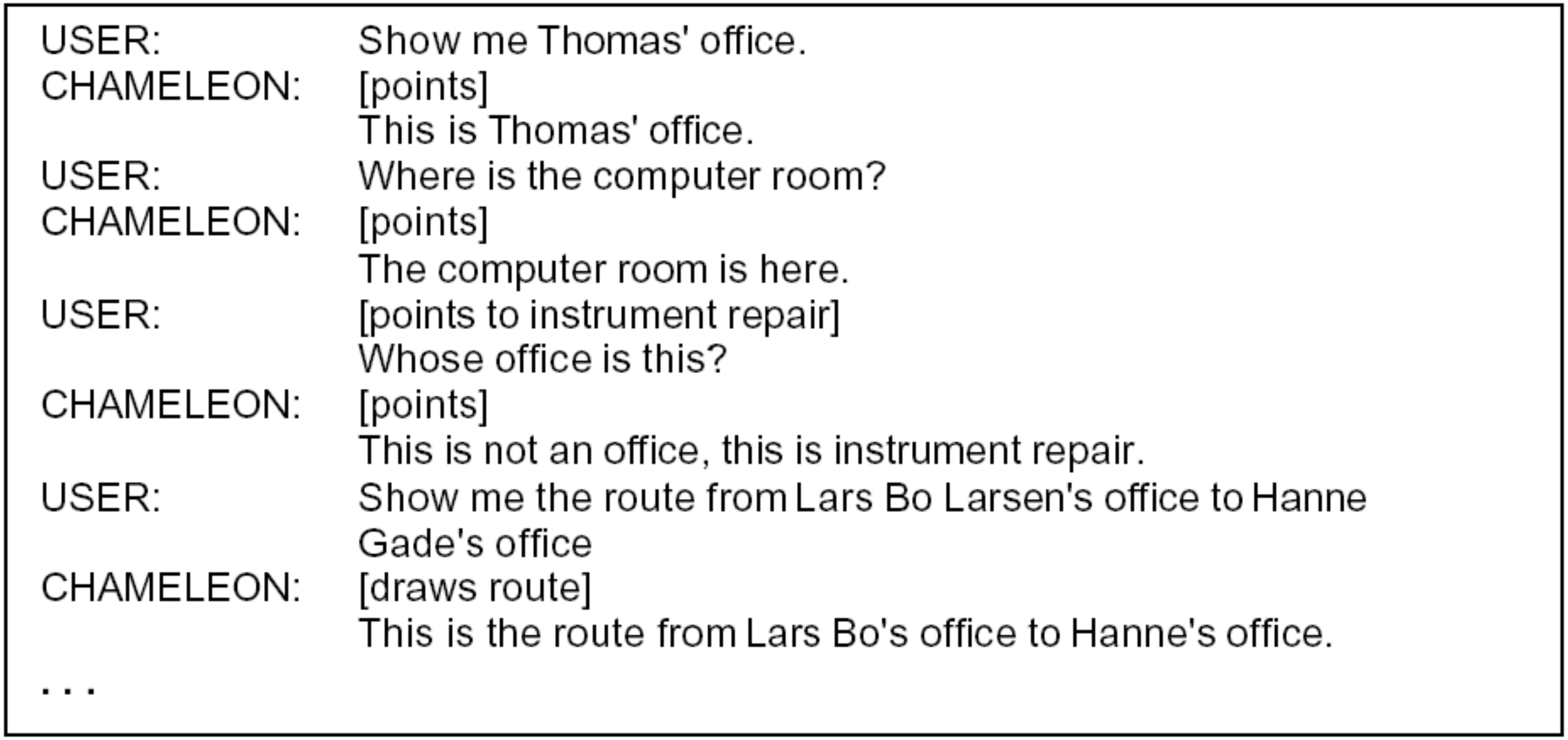
Blueprint of the Fredrik Bajers Vej 7 A-



Example of pointing situation



Example Dialogue



USER:	Show me Thomas' office.
CHAMELEON:	[points] This is Thomas' office.
USER:	Where is the computer room?
CHAMELEON:	[points] The computer room is here.
USER:	[points to instrument repair] Whose office is this?
CHAMELEON:	[points] This is not an office, this is instrument repair.
USER:	Show me the route from Lars Bo Larsen's office to Hanne Gade's office
CHAMELEON:	[draws route] This is the route from Lars Bo's office to Hanne's office.
. . .	

Inter Process Communication

- Applications typically consist of several interdependent modules, often running on separate machines or even dedicated hardware.
- Such distributed applications have a need to communicate in various ways. Some modules feed others in the sense that all generated output from one is treated further by another.
- In the Campus Information System all modules report their output to the blackboard where it is stored.

Semantic Representation

- Semantic representations are frames in the spirit of Minsky and the frame semantics consists of input, output, and integration frames for representing the meaning of intended user input and system output.

- The intention is that all modules in the system will produce and read frames. Frames are coded as messages built of predicate-argument structures following a specific BNF definition.

Frame Semantics

- Frames represent some crucial elements such as *module,*
- *input/output, intention, location,* and *timestamp*.
  - Module is simply the name of the module producing the frame (e.g. NLP).
  - Inputs are the input recognised whether spoken (e.g. "Show me Hanne's office") or gestures (e.g. pointing coordinates) and
  - outputs are the intended output whether spoken (e.g. "This is Hanne's office.") or gestures (e.g. pointing coordinates).
  - Timestamps can include the times a given module commenced and terminated processing and the time a frame was written on the blackboard.

Frame Semantics, ctd.

  - The frame semantics also includes representations for two key phenomena in language/vision integration: reference and spatial relations.
- Frames can be grouped into three categories: input, output and integration.
  - Input frames are those which come from modules processing perceptual input
  - Output frames are those produced by modules generating system output and
  - Integration frames are integrated meaning representations constructed over the course of a dialogue (i.e. all other frames).

Frame structure for knowledge rep. and integration

Examples of Input frames from Speech and Gesture Recognisers:

An input frame takes the general form:
[MODULE
INPUT: input
INTENTION: intention-type
TIME: timestamp]

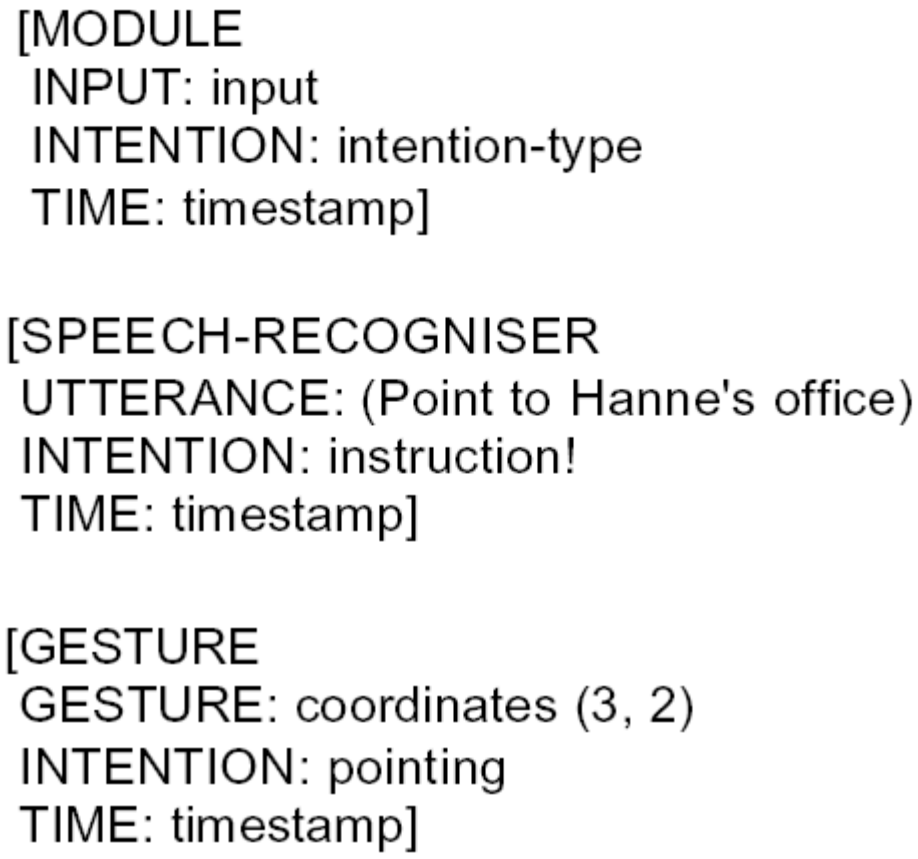
Intention types: query?, instruction!, declarative, pointing, mark-area, indicate-direction.

[SPEECH-RECOGNISER
UTTERANCE: (Point to Hanne's office)
INTENTION: instruction!
TIME: timestamp]

[GESTURE
GESTURE: coordinates (3, 2)
INTENTION: pointing
TIME: timestamp]

Examples of Frames

General input frame format, and examples of speech and gesture input frames



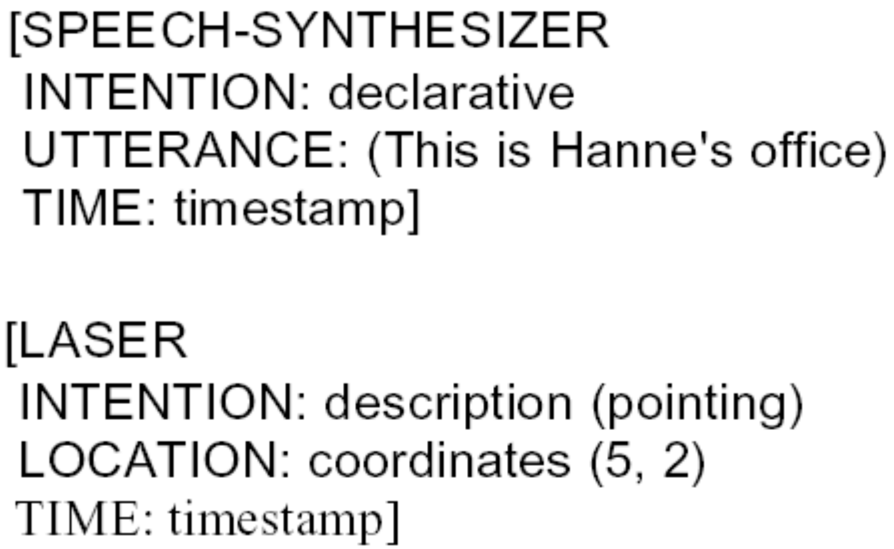
```
[MODULE
 INPUT: input
 INTENTION: intention-type
 TIME: timestamp]

[SPEECH-RECOGNISER
 UTTERANCE: (Point to Hanne's office)
 INTENTION: instruction!
 TIME: timestamp]

[GESTURE
 GESTURE: coordinates (3, 2)
 INTENTION: pointing
 TIME: timestamp]
```

Output Frames

General input frame format, and examples of speech and gesture input frames

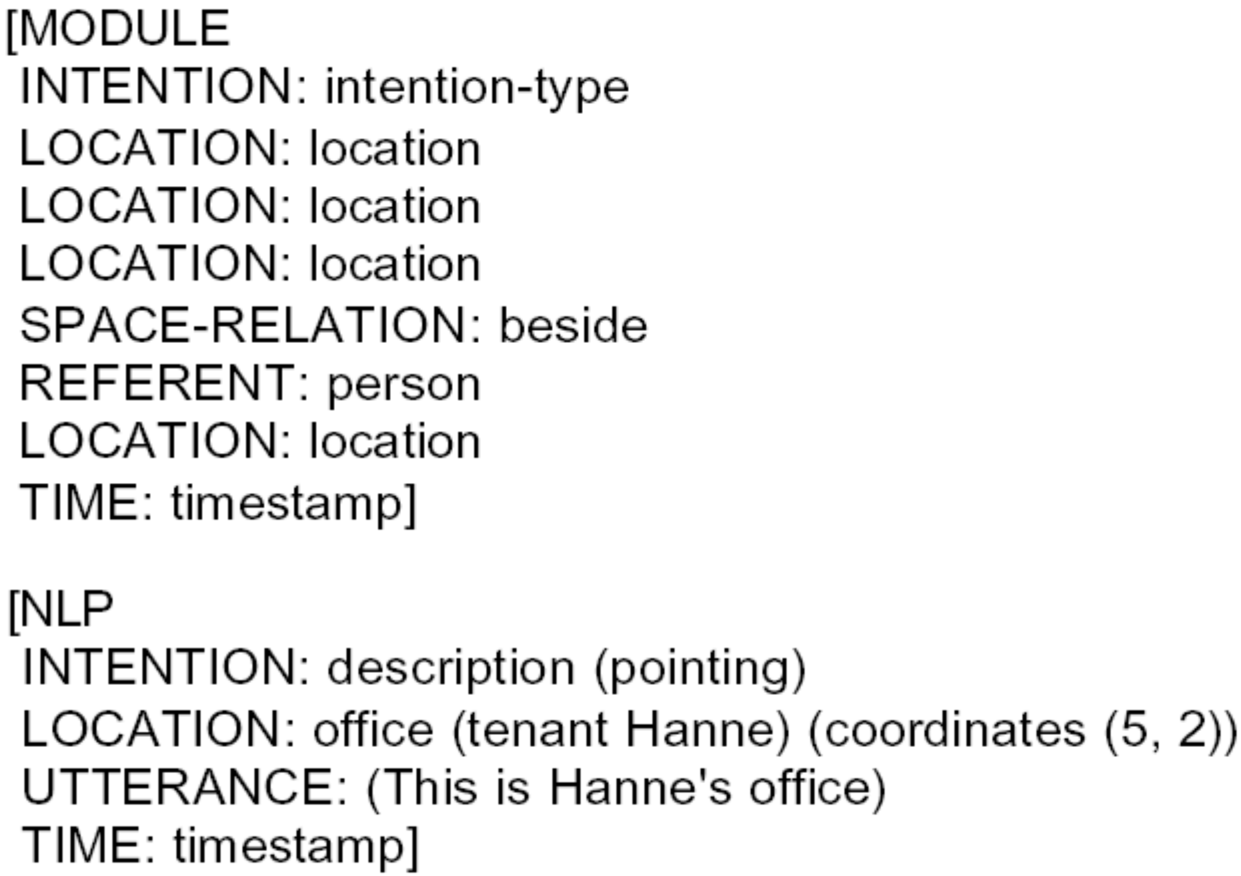


```
[SPEECH-SYNTHESIZER
 INTENTION: declarative
 UTTERANCE: (This is Hanne's office)
 TIME: timestamp]

[LASER
 INTENTION: description (pointing)
 LOCATION: coordinates (5, 2)
 TIME: timestamp]
```

Integration Frames

Examples of Integration frames



```
[MODULE
 INTENTION: intention-type
 LOCATION: location
 LOCATION: location
 LOCATION: location
 SPACE-RELATION: beside
 REFERENT: person
 LOCATION: location
 TIME: timestamp]

[NLP
 INTENTION: description (pointing)
 LOCATION: office (tenant Hanne) (coordinates (5, 2))
 UTTERANCE: (This is Hanne's office)
 TIME: timestamp]
```

Referencing

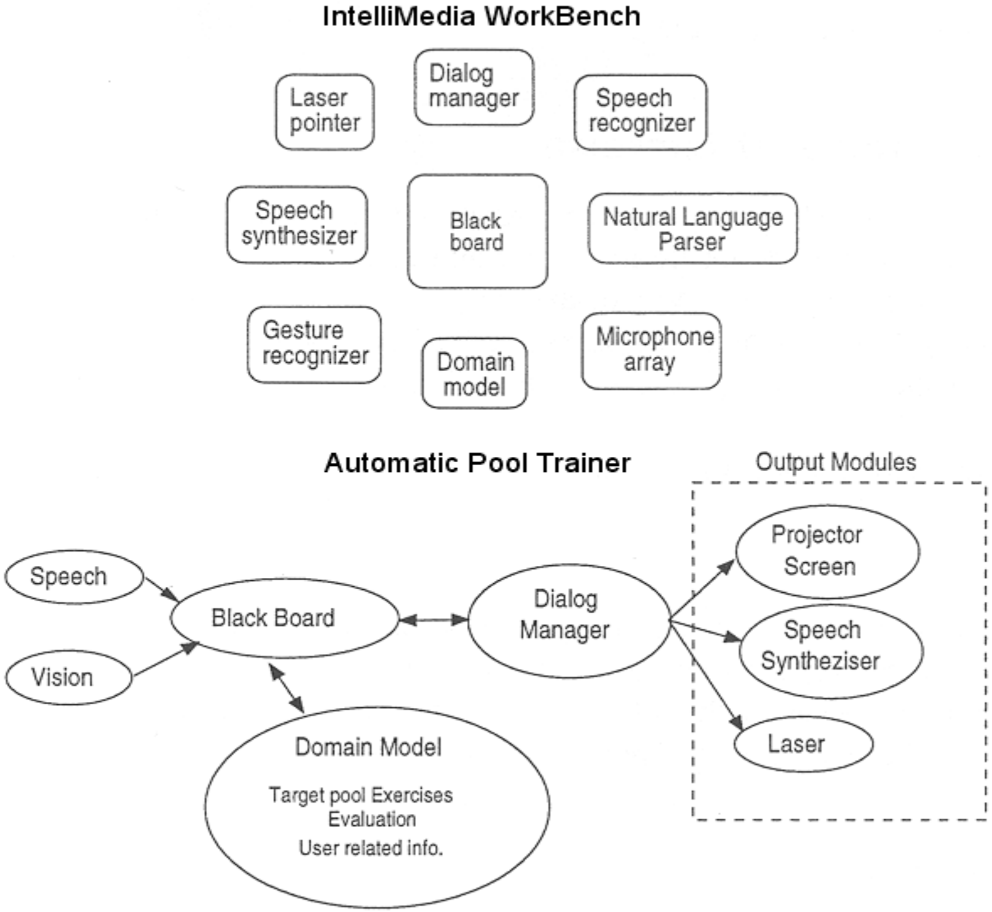
- In dialogue systems fusing the oral mode with other modalities, we are likely to be confronted with the two “traditional” linguistic reference types:
  - Endophora, the reference of a linguistic entity to another linguistic entity normally before (anaphora) but eventually also after it (cataphora).
  - Deixis, the reference by means of a linguistic entity whose interpretation is relative to the extralinguistic context, e.g. who is speaking, the time or place of speaking.

Referencing in Multi Modal Dialogue

- Cross-media reference, the (deictic) reference of a linguistic entity to an antecedent in another communication channel, e.g. a picture or video sequence being displayed by the system.
- Cross-user/system reference, the endophoric/deictic reference of a linguistic entity in the user-input to an antecedent in the system output or, vice versa, of a linguistic entity in the system-output to an antecedent in the user input.
- The resolution of cross-user/system reference demands that the system "understands" its own output to the same extent as it "understands" the input from the user.
- However, many systems seem to be focusing one-sided on the problem of recognizing and understanding the intentions of the user but using more rigid template-like structures for "replaying" responses.

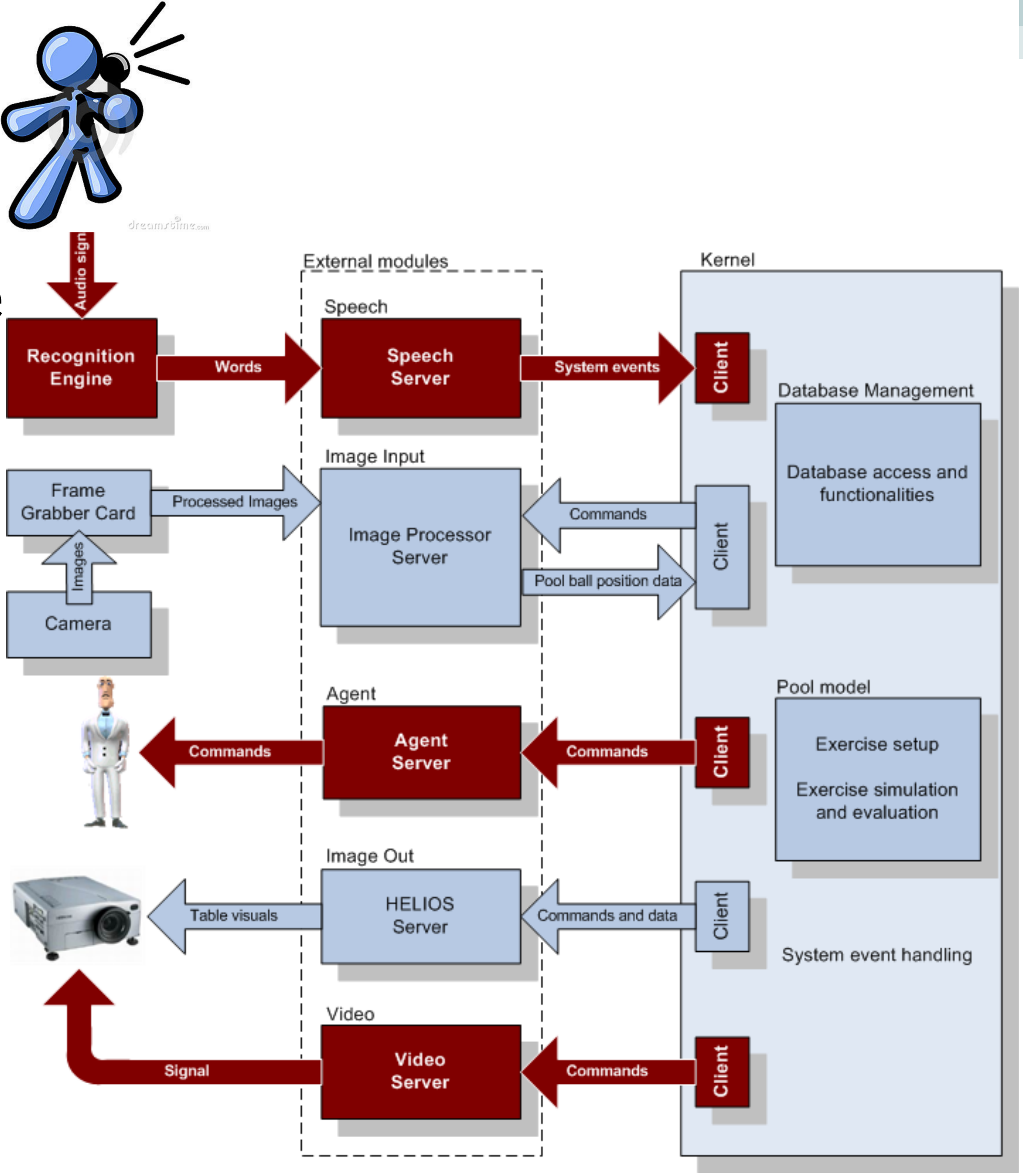
IntelliPool – The automatic Pooltrainer

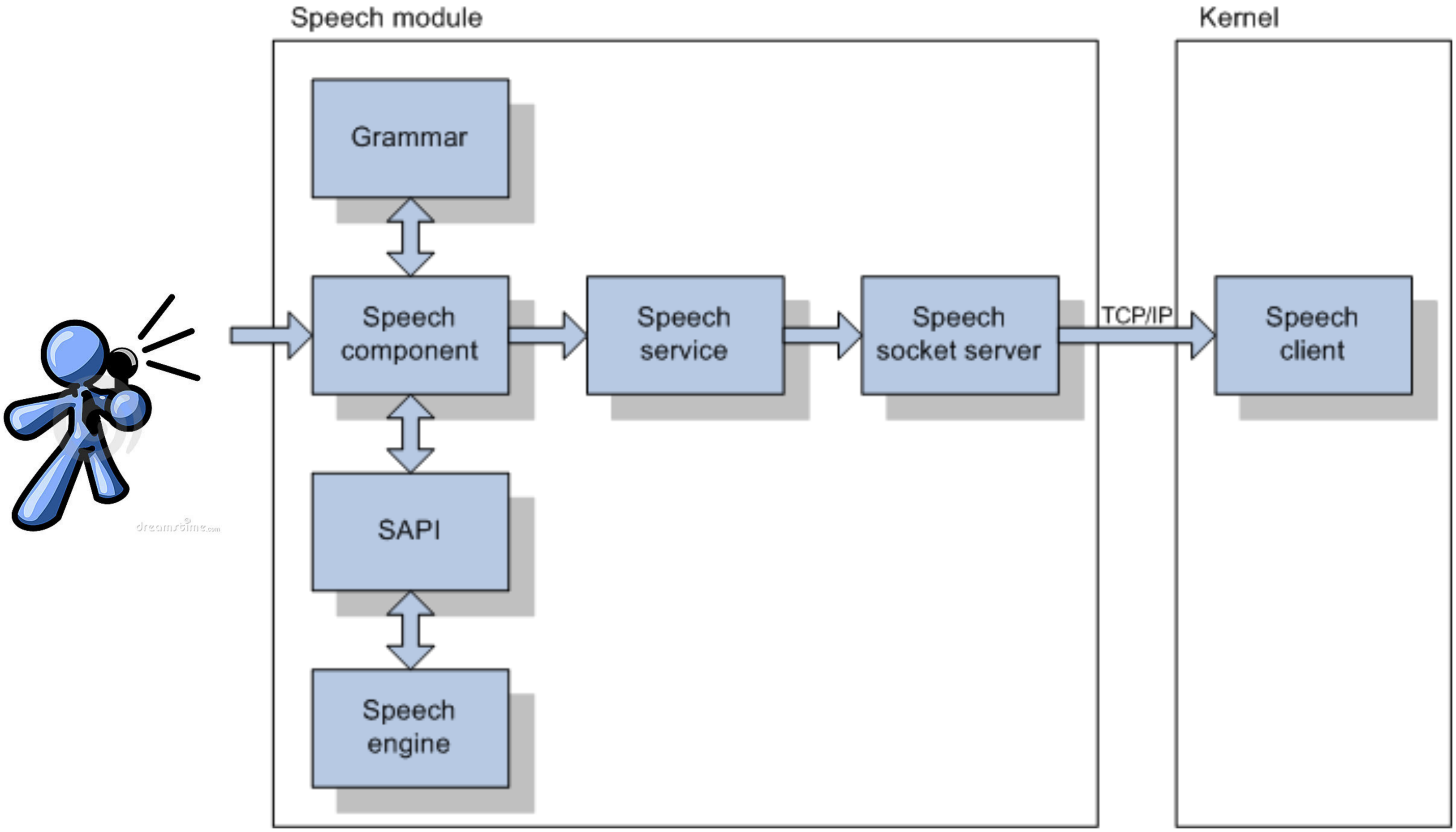
- The WorkBench modules and architecture were reused in the creation of an Automatic Pool Trainer



Architecture

- Modular architecture
  - Kernel
  - Interface modules:
    - Image Processor
    - HELIOS
    - Speech
    - Agent
    - Video
- Event driven
  - Socket connections
- IntelliPool3 intact
  - Menu can still be used





Speech module

Kernel

Grammar

Speech component

Speech service

Speech socket server

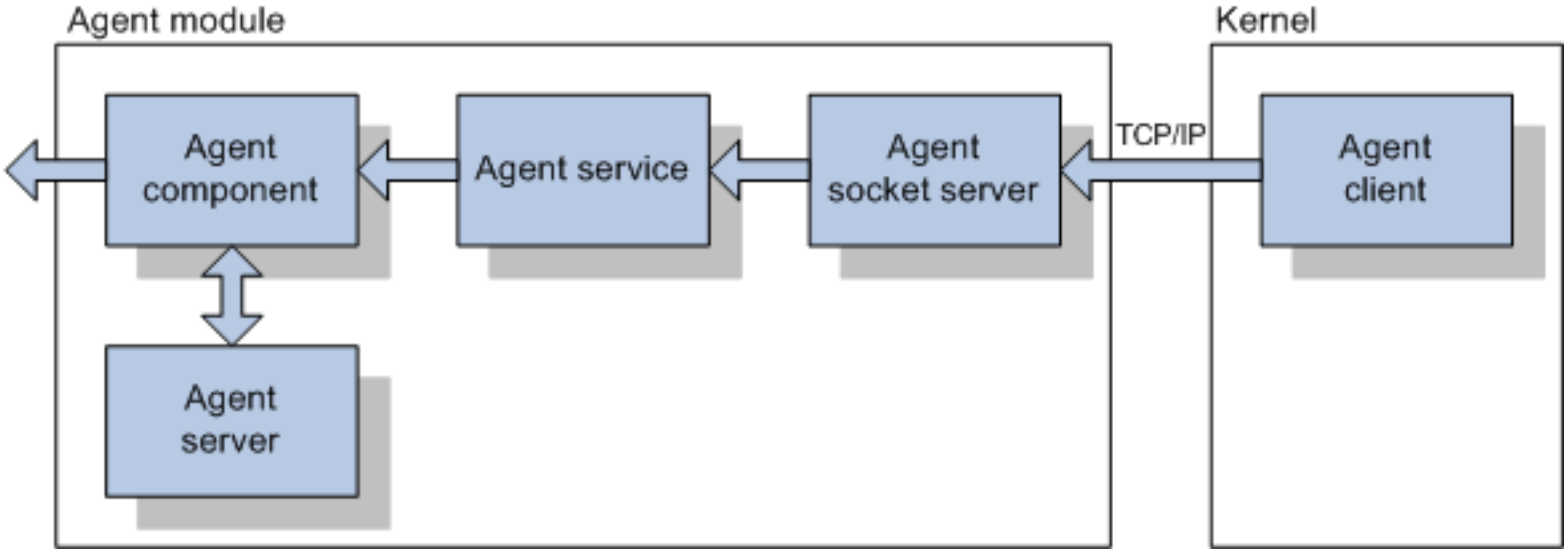
TCP/IP

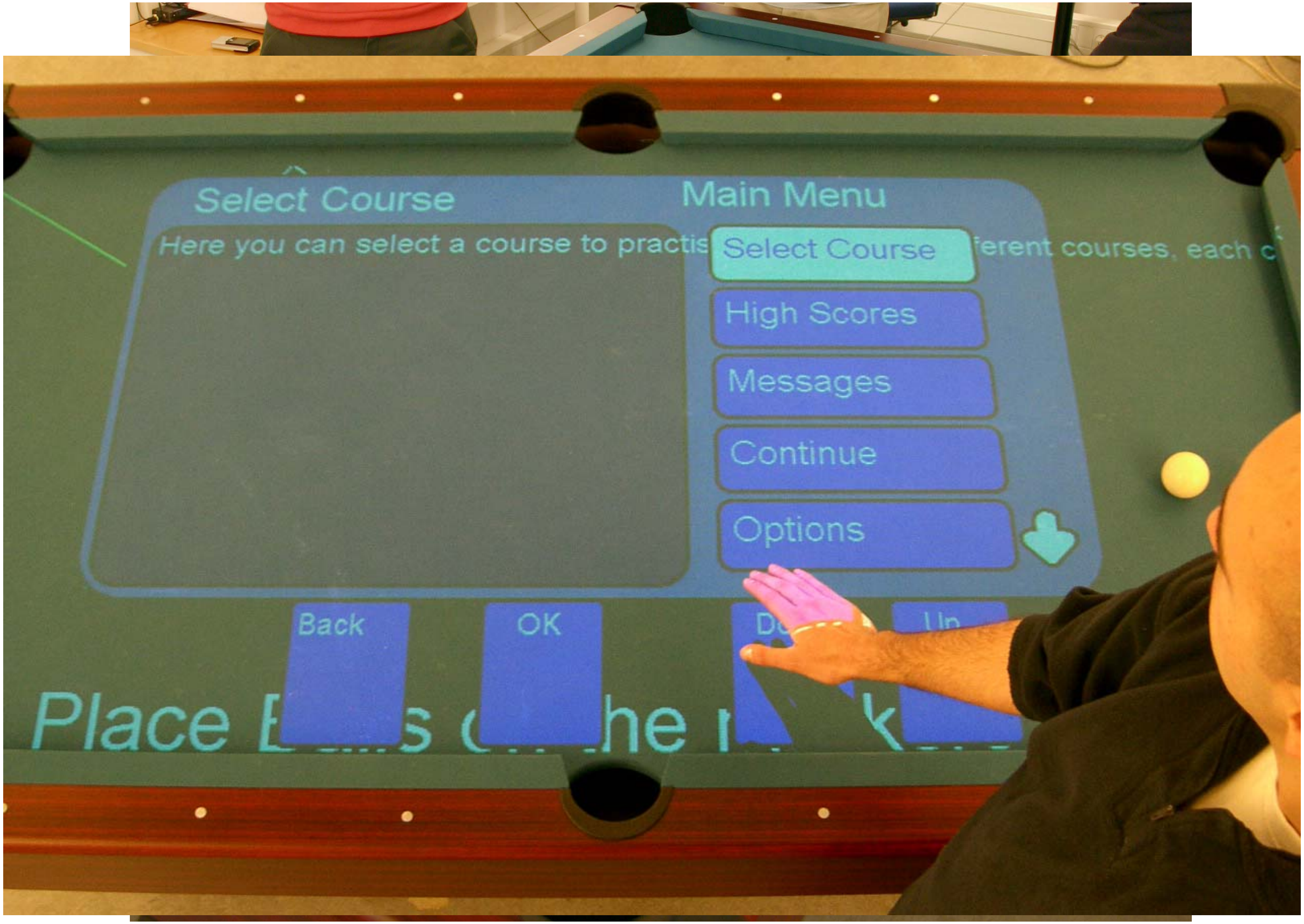
Speech client

SAPI

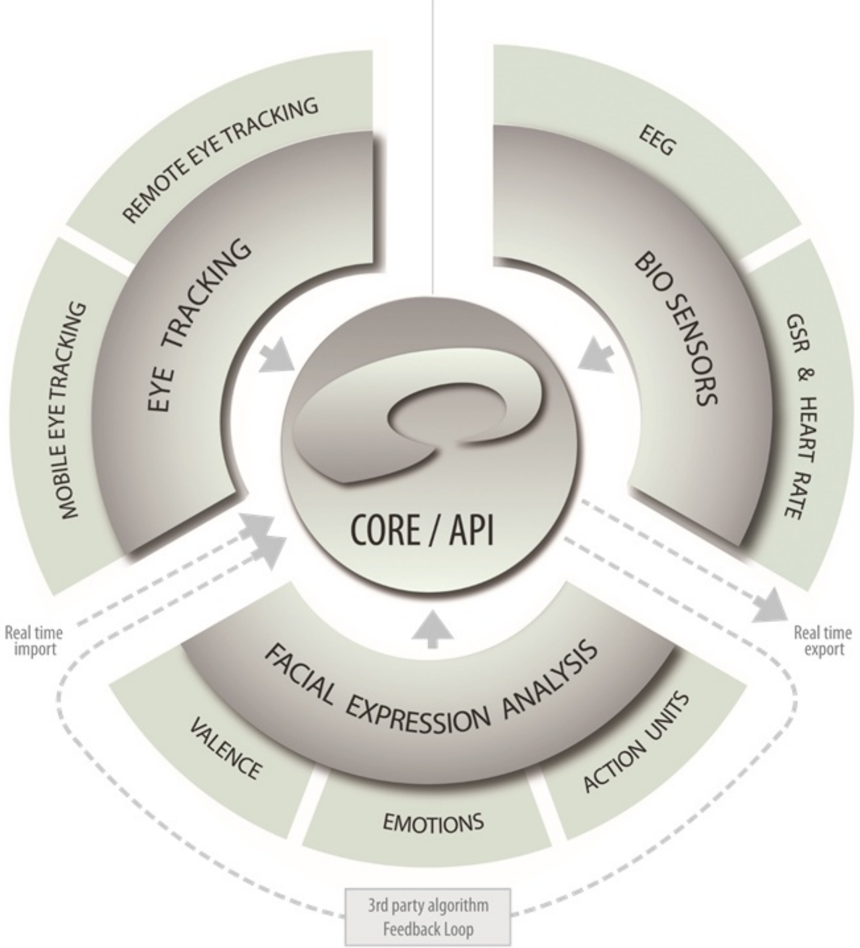
Speech engine

Agent Module




Select Course
Main Menu
Here you can select a course to practis
ferent courses, each c
Select Course
High Scores
Messages
Continue
Options
Back
OK
Do
Up
Place B s o he r k

Attention Tool

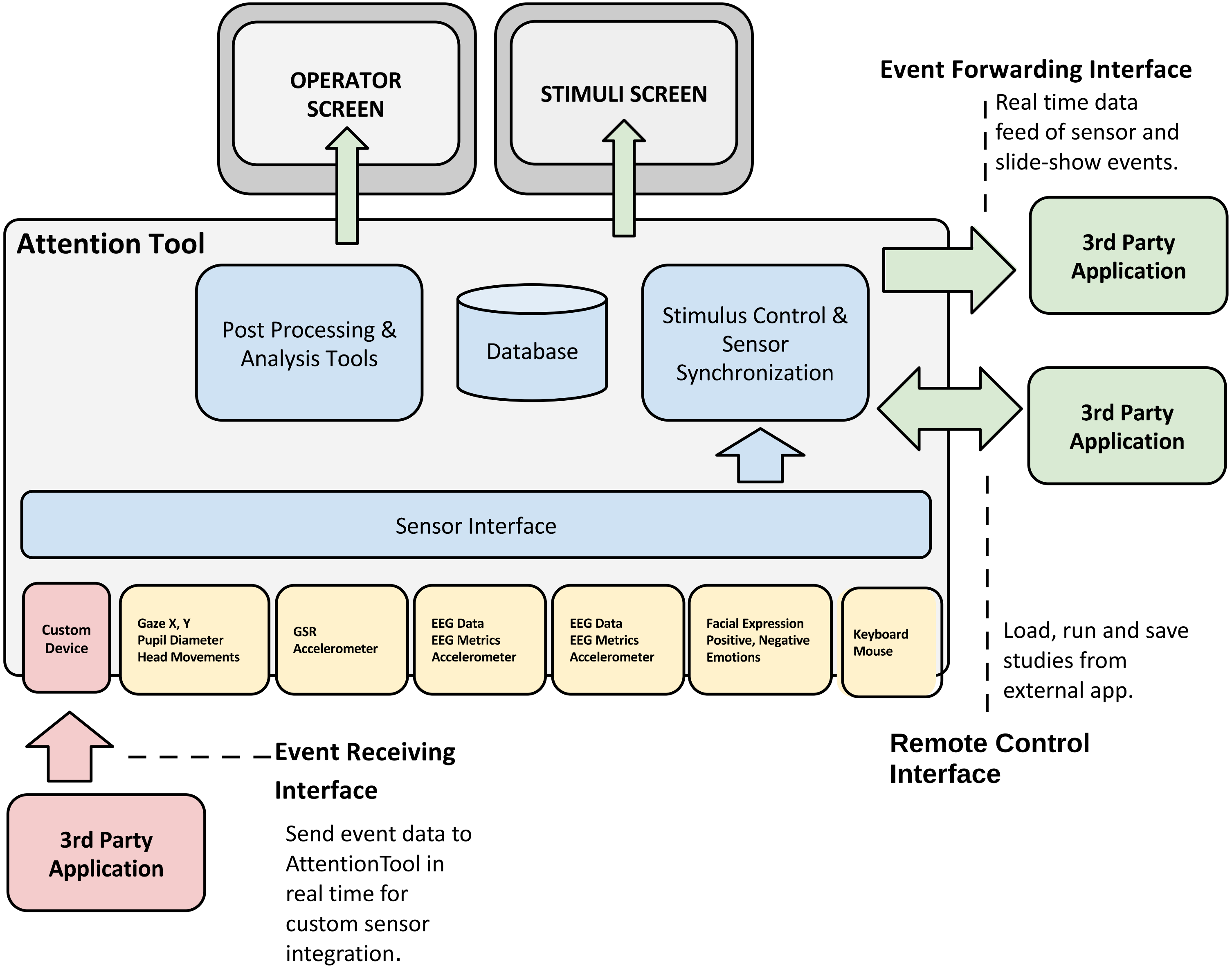


The Attention Tool is a multi modal platform but serves a slightly different purpose:

It senses and processes biometric data from humans

- Eye-tracking
- Facial expressions
- GSR – galvanic skin response
- EEG for detection of human emotional response
- …

But it does not directly implement multi modal human Machine Interaction



OPERATOR SCREEN

STIMULI SCREEN

Event Forwarding Interface

Real time data feed of sensor and slide-show events.

Attention Tool

Post Processing & Analysis Tools

Database

Stimulus Control & Sensor Synchronization

3rd Party Application

3rd Party Application

Sensor Interface

Custom Device

Gaze X, Y
Pupil Diameter
Head Movements

GSR
Accelerometer

EEG Data
EEG Metrics
Accelerometer

EEG Data
EEG Metrics
Accelerometer

Facial Expression
Positive, Negative
Emotions

Keyboard
Mouse

Load, run and save studies from external app.

Remote Control Interface

3rd Party Application

Event Receiving Interface

Send event data to AttentionTool in real time for custom sensor integration.

Summary

- Agent-Based architectures for Late Fusion is the most widely used approach for integration of input modalities
  - ▫ Allows independent development of sensors and processing
  - ▫ Requires semantic represenation format for fusion
  - ▫ Potential problems with synchronisation
- However: It is not how humans do it