

# Computers in the Human Interaction Loop

*or: How can Computers Support Human-Human  
Communication*

May 1, 2006

Alex Waibel

Interactive Systems Laboratories

Carnegie Mellon University

University of Karlsruhe

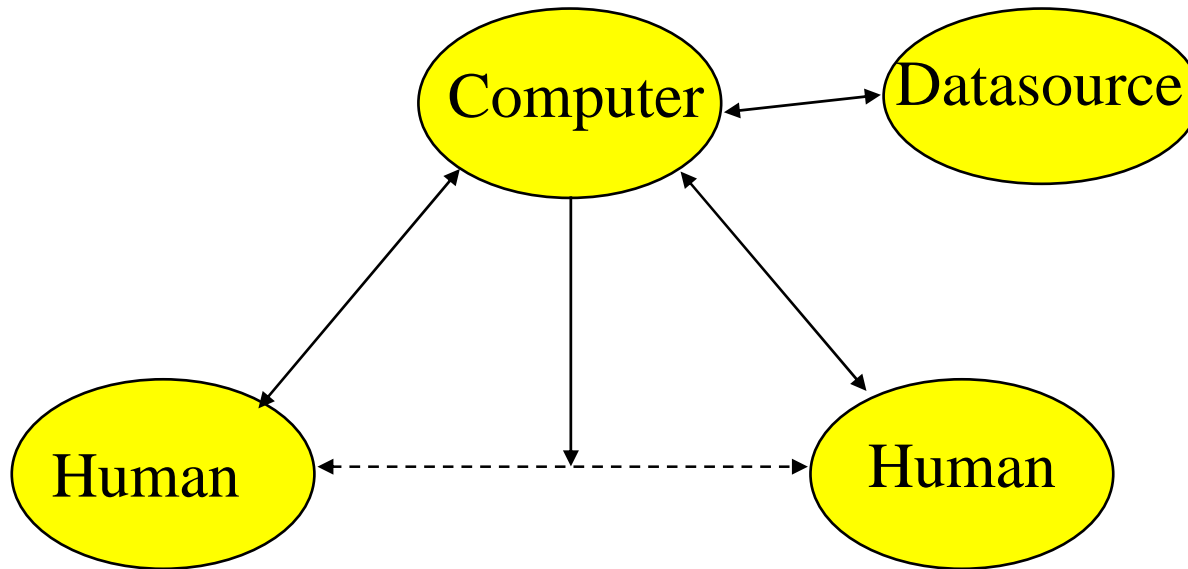
<http://www.interact.cs.cmu.edu>





# Present Human-Computer Interaction





# Interpreting Human Communication

*“Why did Joe get angry at Bob about the budget ?”*

Need Recognition and Understanding of Multimodal Cues

- Verbal:
  - Speech
    - Words
    - Speakers
    - Emotion
    - Genre
  - Language
  - Summaries
  - Topic
  - Handwriting
- Visual
  - Identity
  - Gestures
  - Body-language
  - Track Face, Gaze, Pose
  - Facial Expressions
  - Focus of Attention



We need to understand the: **Who, What, Where, Why and How !**



<http://chil.server.de>

- **Integrated Project (IP)** in 6<sup>th</sup> Framework Program of the EC
  - One of three IP's in the first call Multimodal/Multilingual:
  - CHIL, TC-STAR, AMI
- **International Consortium:**
  - 15 Partners from 9 countries in Europe (12) and the US (3)
- **Coordination:**
  - Research: Prof. A. Waibel – InterACT Center  
Universität Karlsruhe, Carnegie Mellon University
  - Financial: Prof. H. Steusloff - Fraunhofer IITB
- **Term:**
  - 6 Year Goal, Two Phases
  - First (Current) Phase: 3 Years
- **Budget**
  - CHIL: 25 Million Euro Cost Volume for three Years



# The CHIL Project

## The CHIL Team:



Universität Karlsruhe (TH)

Fraunhofer

Institut  
Informations- und  
Datenverarbeitung



DAIMLERCHRYSLER



TU/e technische universiteit eindhoven



Centre de Tecnologies i Aplicacions del Llenguatge i la Parla  
UNIVERSITAT POLITÈCNICA DE CATALUNYA



STANFORD UNIVERSITY  
Carnegie Mellon



# Management Approach

- Goal:
  - Accountability without Stifling Creativity
  - Approach: Coopetition
  - Evaluations, MOPs and MOEs
- Technologies Evaluations
  - Benchmarks CHIL, CLEAR, RT
  - Technology Catalogue
  - Building on and Advancing the State of the Art
- Services
  - Services Built on Tech
    - Architecture, Infrastructure
    - Technology Catalogue
  - Not One Integrator Site, but 4 Service Builder Sites
  - Compare & Contrast Site Visits
  - User Studies, Assess Usability / Effectiveness
  - Creative Surprises Encouraged





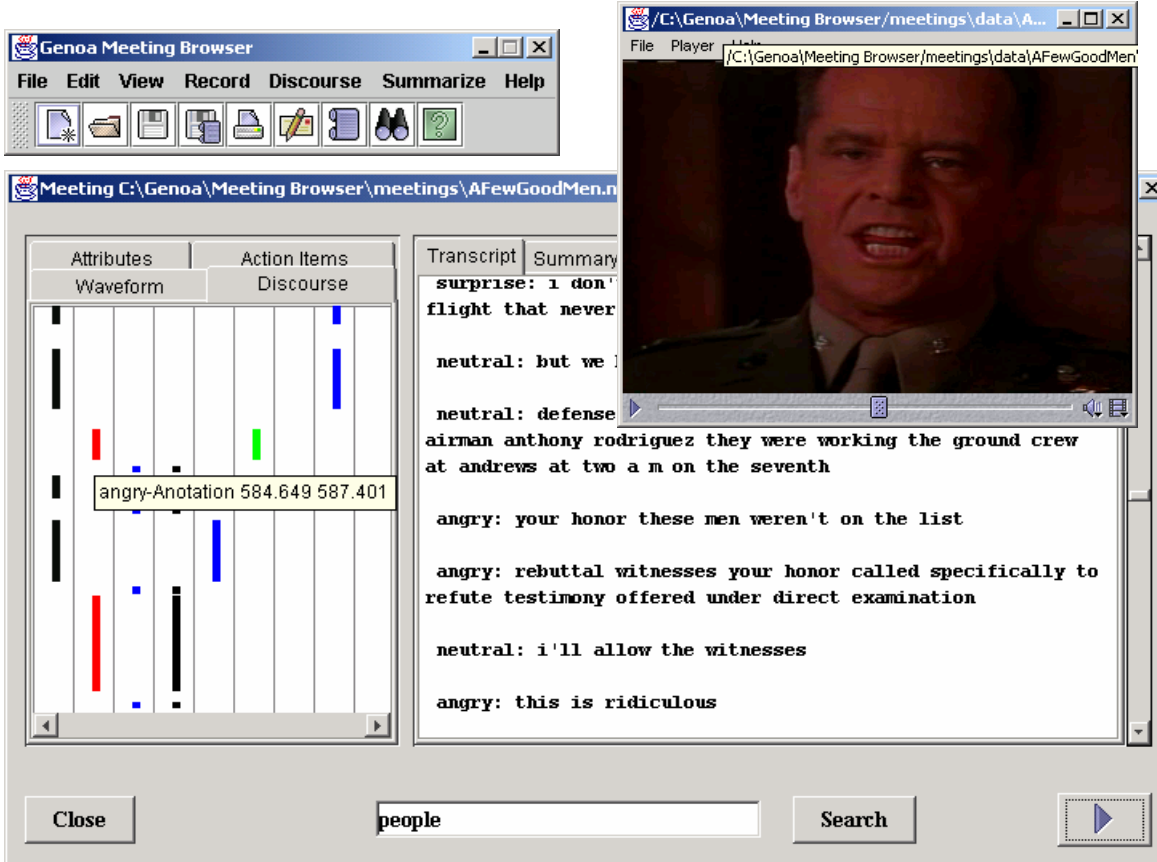
- **Services:**
  - Implicit Proactive Computing Services Based on Perceived Implicit Need
  - Study Success of Such Services and their Ability to Improve Productivity
- **Technologies & Functionalities:**
  - Descriptions of Human Behavior and Attributes - the “**Who? Where? What? Why? How?**” of Humans.
  - Underlying perceptive technologies have been studied before, but require *greater robustness* and performance (speech, vision, ...)
- **Infrastructure:**
  - To enable composition, aggregation, processing and interoperation of the distributed components (sensors, technologies, fusion, services,...)



# CHIL Services



# Retrieval Services: Meeting Browser



- Motivation
  - Projects: Genoa ('97-'00), Fame (01-04)
  - Rapid Access/Review of Meeting Records
- Components:
  - Transcribe Speech
  - Summarization
  - Named Entities
  - Discourse Types, Games, Genres
  - Emotion, Hyperartc.
  - People ID
  - Focus of Attention
  - Speaker Style, Types, Relations

ICASSP'98 – *Experiments in Meeting Recognition*, Yu et al.  
 DARPA BN'98 – *Meeting Browser: Tracking and Summarizing Meetings*, Waibel et al.



# Proactive Services

- **Connector**
  - Connects people through the right device at the right moment
- **Memory Jog**
  - Unobtrusive service. Helps meeting attendees with information
  - Provides pertinent information at the right time (proactive/reactive)
  - Lecture Tracking and Memory
- **Relational Report**
  - Informs the current speaker about interest/boredom of audience
  - Coaches Meetings to be More Effective
- **Socially Supportive Workspaces**
  - Physically shared infrastructure aimed at fostering collaboration
- ***Simultaneous Translation Services***
  - *Detect Language Need and Deliver Services Inobtrusively*
- ... (*and more*)



# The Connector

- Socially Appropriate Connection
  - Connect People when Appropriate by Appropriate Media
- Connecting People depends on:
  - Social Relationship of Parties
  - Space / Environment
  - Activity, User State
  - Urgency of Matter



JEFF'S CONTEXT INFO		
Context	environment	UNKNOWN
	environment model	
	in smartroom? situation	YES MEETING
Current State		MEETING
Availability	Contact	Talk    Message
	personal	<input type="checkbox"/> <input type="checkbox"/>
	business	<input type="checkbox"/> <input checked="" type="checkbox"/>
	VIP	<input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
Phone Alert	personal	MUTE
	business	MUTE
	VIP	EXCLUSIVE

# Memory Jog

....What was his name? ...Where did I meet him? ...What happened at the last meeting?



## Private and Public Information Delivery

- CHIL phone
- Steerable Camera Projector
- Targeted Audio
- Retinal and Heads-Up Displays



# Memory Jog

....What was his name? ...Where did I meet him? ...What did we discuss last time?





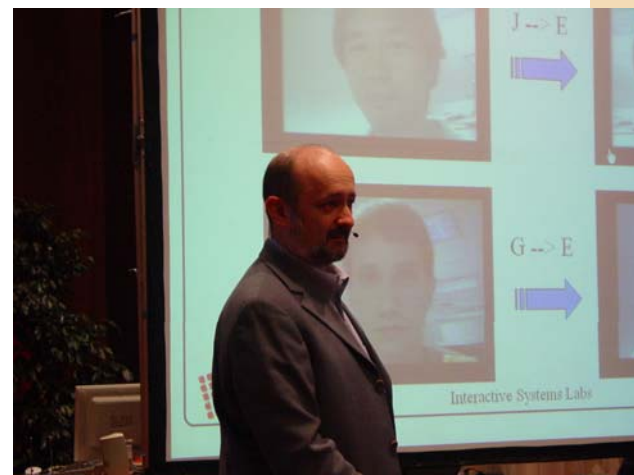
....and what in the world is he saying?



你们的评估准则是什么

# Lecture Translator

- Idea: Translate Domain Unlimited Speeches
- Applications:
  - TV/Radio Broadcast Translation
  - Translation of Lectures and Speeches
  - Parliamentary Speeches (UN, EU,...)
  - Telephone Conversations
  - Meeting Translation
- Technical Difficulty:
  - Open Domain, Open Vocabulary, Open Speaking Style, Spontaneous Speech, Disfluencies, Ill-Formed Sentences
- Research:
  - NSF-ITR STR-DUST, EC-IP TC-STAR
  - Learning, Statistical Learning Algorithms



# TC-STAR

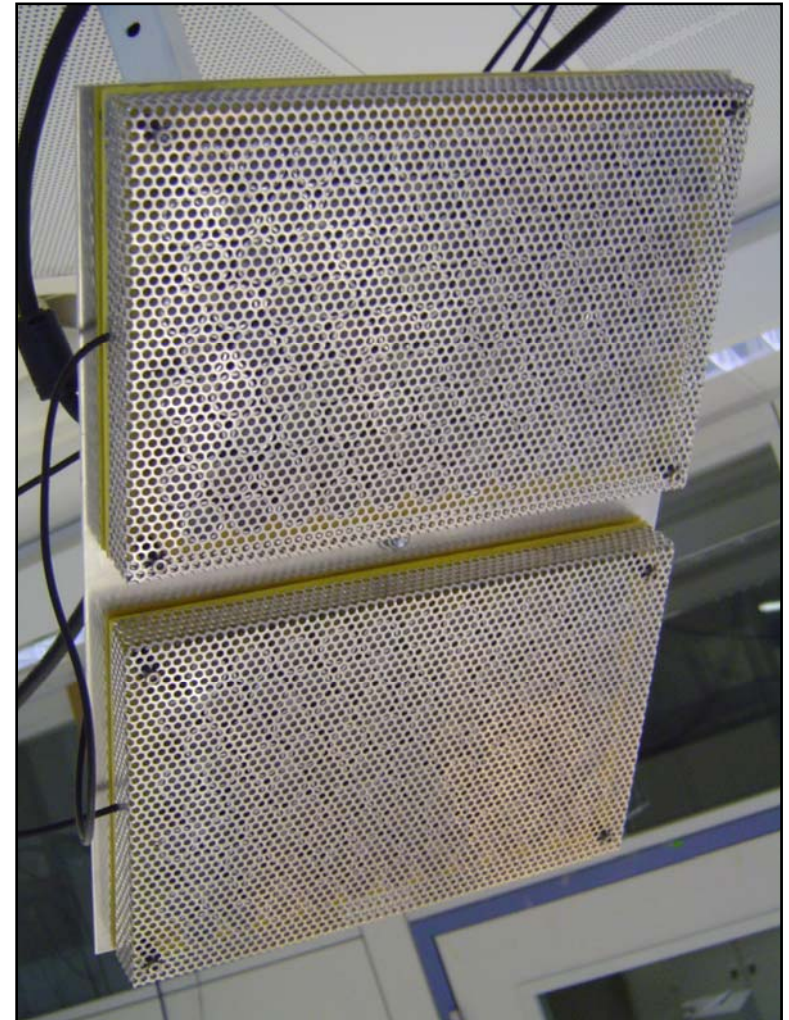


MR PRESIDENT

señor presidente



# Targeted Audio



# Silent Speech based on EMG Signals



# CHIL Technologies



*“Why did Joe get angry at Bob about the budget ?”*

Need Recognition and Understanding of Multimodal Cues

- Verbal:
  - Speech
    - Words
    - Speakers
    - Emotion
    - Genre
  - Language
  - Summaries
  - Topic
  - Handwriting
- Visual
  - Identity
  - Gestures
  - Body-language
  - Track Face, Gaze, Pose
  - Facial Expressions
  - Focus of Attention



We need to understand the: **Who, What, Where, Why and How !**



- **Who & Where ?**

- Audio-Visual Person Tracking
- Tracking Hands and Faces
- AV Person Identification
- Head Pose / Focus of Attention
- Pointing Gestures
- Audio Activity Detection

- **What ? (Input)**

- Far-field Speech Recognition
- Far-field Audio-Visual Speech Recognition
- Acoustic Event Classification

- **What ? (Output)**

- Animated Social Agents
- Steerable targeted Sound
- Q&A Systems
- Summarization

- **Why & How ?**

- Classification of Activities
- Emotion Recognition
- Interaction & Context Modelling
- Vision-based posture recognition
- Topical Segmentation

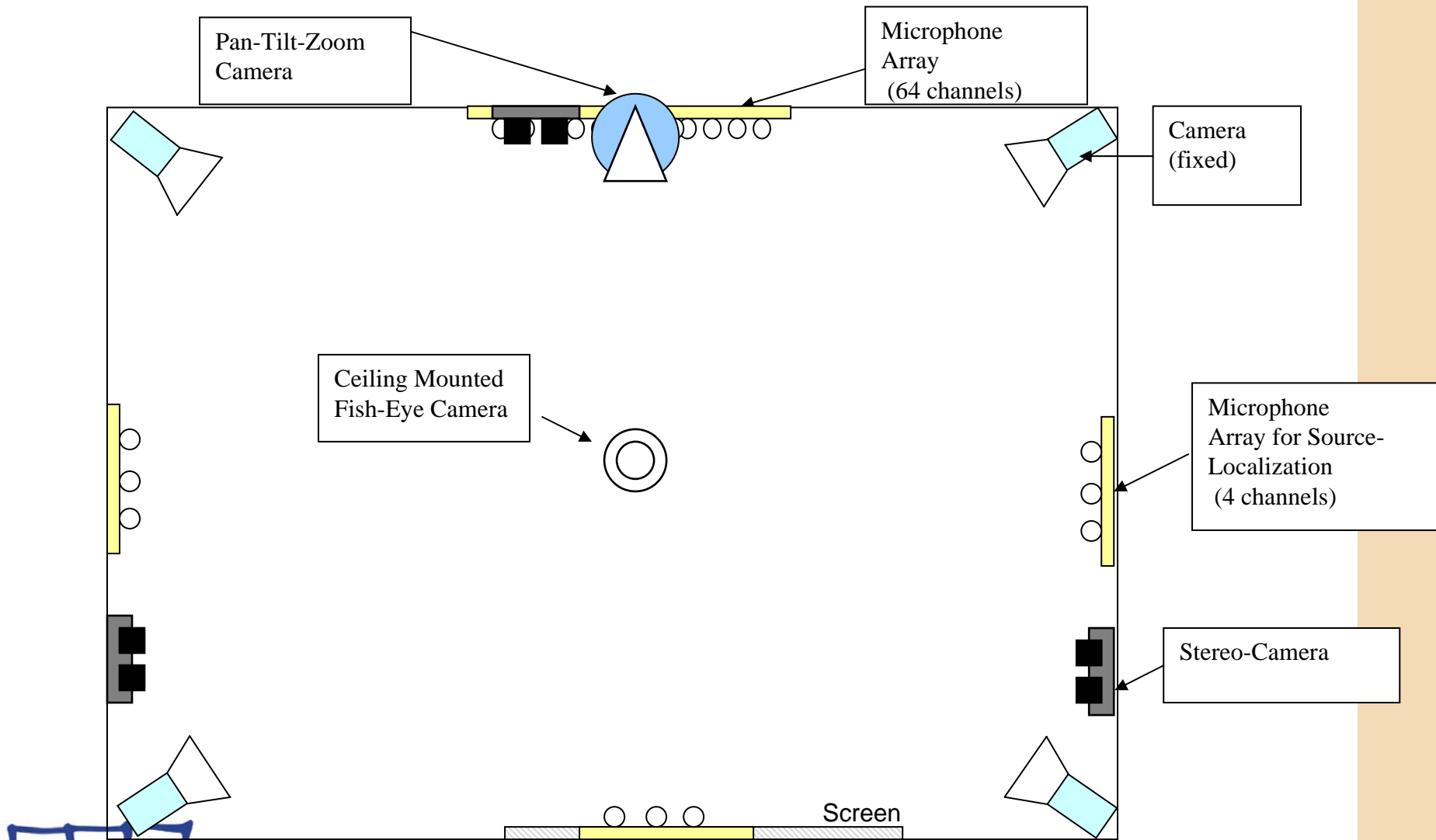




- Require: Performance, Robustness, Realism
  - Distant, Remote Microphones
  - Hands-Free, Always On → Segmentation
  - Sloppy Speech
  - Cross-Talk
  - Noise
  - Disfluencies, Prosody, Structuring Discourse
  - Communication by Other Modalities
  - Other Elements of Speech (Emotion, Direction, Scene Analysis)
  - Multimodal People ID
  - Free People Movement
  - Focus of Attention and Direction
  - Named Entities, OOV's
  - Adaptation and Evolution
  - Summarization



# Sensors in the CHIL Room



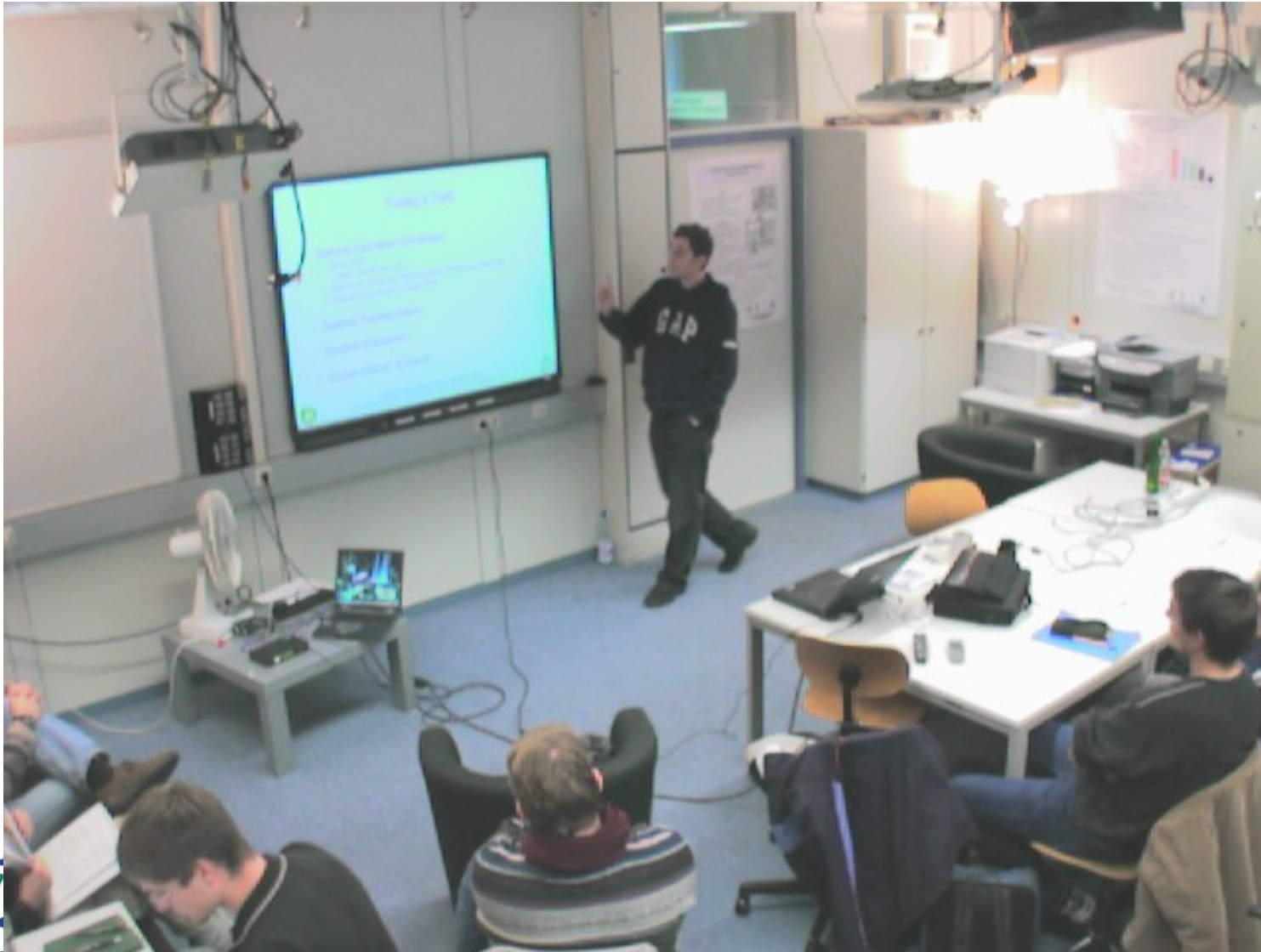
# Scenario 1: Seminars/Lectures



# Scenario 2: Meetings



# Describing Human Activities

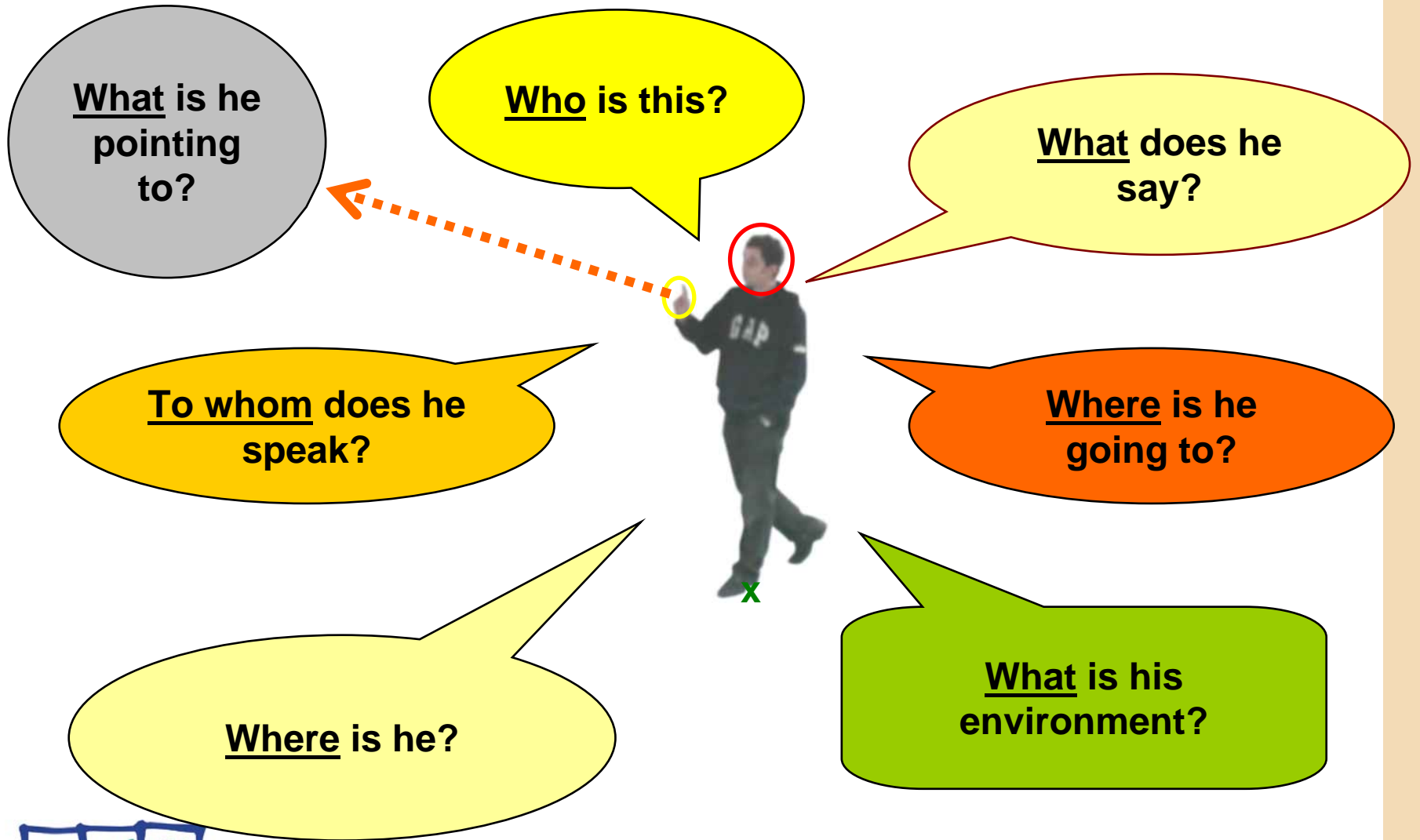


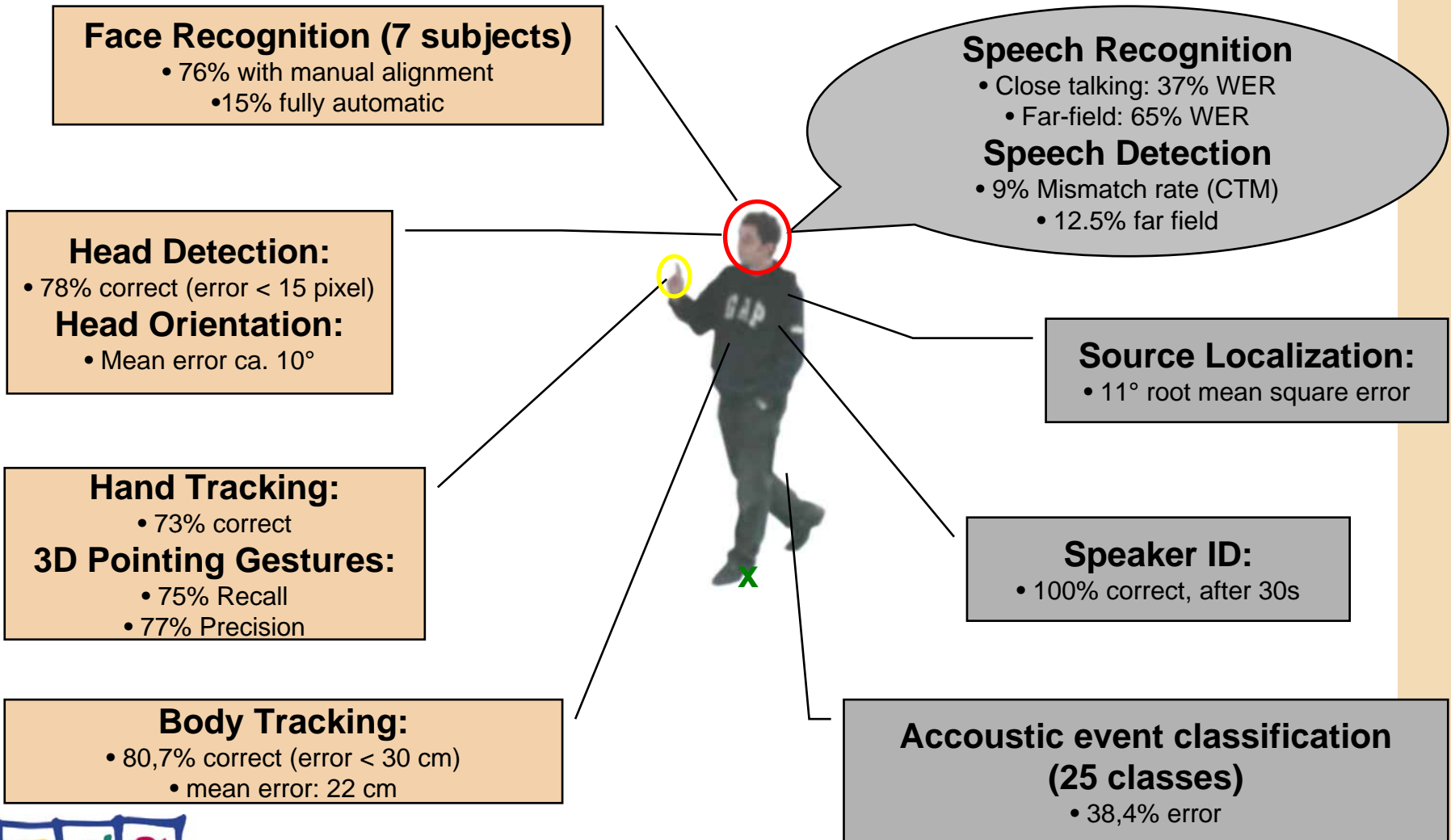
# Describing Human Activities

---



# Technologies/Functionalities







## NIST and EC Programs Join Forces

- RT-Meeting'06 – Rich Transcription
  - Emerges from established DARPA activity
  - MLMI Workshops, AMI/CHIL
  - Evaluated Verbal Content Extraction
  - Chair: Garofolo (NIST)

- CLEAR'06 –

### Classification of Locations, Events, Activities, Relationships

- Emerging from European program efforts (CHIL, etc.) and US-Programs (VACE,..)
- First Joint Workshop to be Held in Europe after Face & Gesture Reco WS, April 6 & 7, Southampton
- Chair: Stiefelhagen (UKA)



# Putting it All Together



# Conclusion

---

- **Human-Human Communication**
  - New Class of Computer Services
  - Supported by Multimodal Perceptual User Interfaces
- **Scientific Challenges**
  - Observing Human-Human Interaction is a New Dimension in Difficulty for Perceptual Processing Technologies
  - Importance of Evaluations and Solid Progress
  - Detect, Understand Human Needs
  - Computer Sciences and Social Sciences Meet
- **Main Products**
  - Instantiated Human-Centered CHIL Services
  - One of the Largest, Most Realistic, Annotated Multimodal Database
  - Benchmarks, Metrics, Evaluation Infrastructure
  - Transferable Perceptual Technologies (Catalogue)
  - Standards & Architecture

