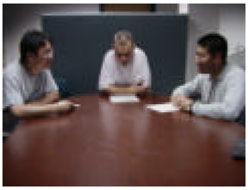
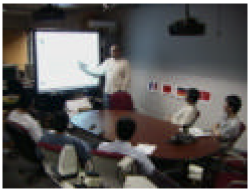
CHIL – Computers in the Human Interaction Loop

Alex Waibel
&
The CHIL Consortium

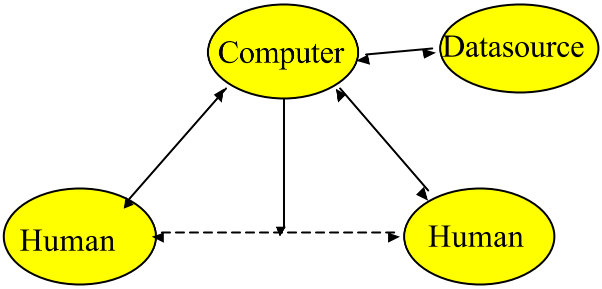
Carnegie Mellon University
Universität Karlsruhe (TH)
http://isl.uka.de, http://chil.server.de,
waibel@cs.cmu.edu





What Problem
Does CHIL Address ?

Different Roles for Humans and Computer



Computer

Datasource

Human

Human

CHIL Services

Four CHIL Services are instantiated (others possible!)

- **Connector**
  - Connects appropriate 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
- **Attention cockpit**
  - Informs the current speaker about interest/boredom of audience
- **Socially Supportive Workspaces**
  - Physically shared infrastructure aimed at fostering collaboration: workspace, portable collaborative devices to carry out joint tasks, common duties, negotiation, agreement…

Interpreting Human Communication

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

Need Recognition and Understanding of Multimodal Cues

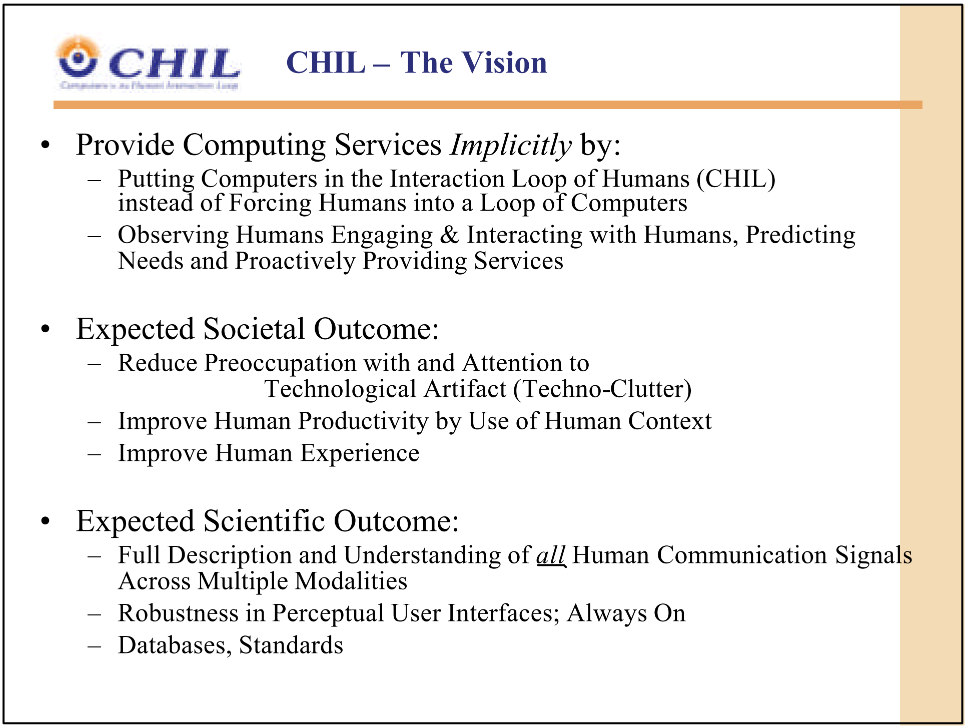
- Verbal:
  - Speech
    - Words
    - Speakers
    - Emotion
    - Genre
  - Language
  - Summaries
  - Topic
  - Handwriting
- Visual
  - Identity
  - Gestures
  - Body-language
  - Track Face, Gaze, Pose
  - Facial Expressions
  - Focus of Attention

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

Project CHIL

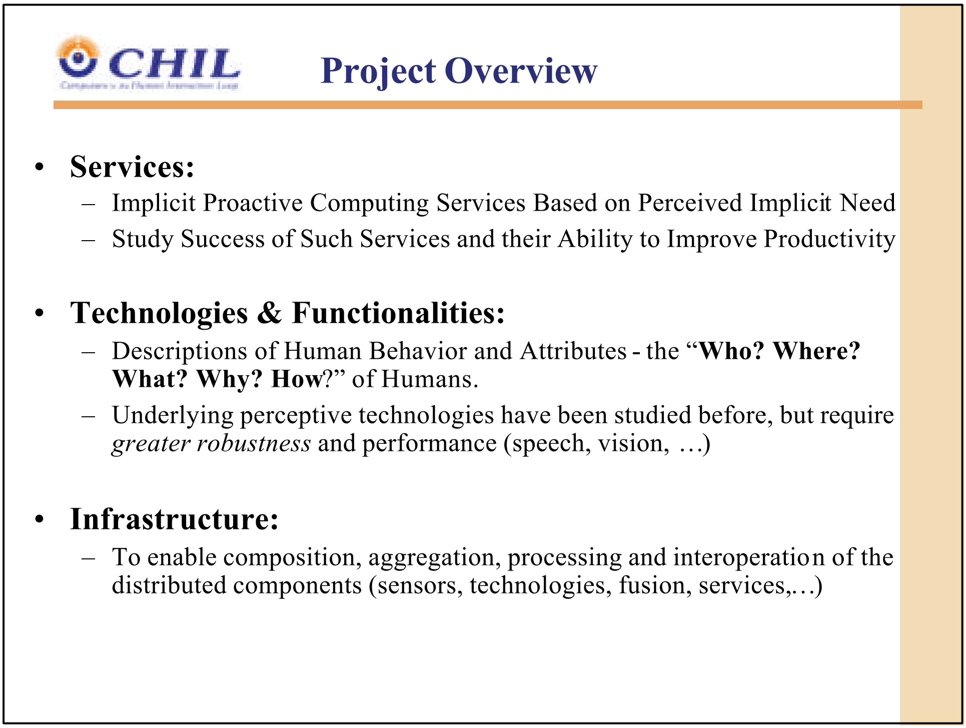
http://chil.server.de

- **Integrated Project** (IP) in 6th 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
  - Possible Follow-On in 2nd Phase



CHIL – The Vision

- Provide Computing Services *Implicitly* by:
  - Putting Computers in the Interaction Loop of Humans (CHIL) instead of Forcing Humans into a Loop of Computers
  - Observing Humans Engaging & Interacting with Humans, Predicting Needs and Proactively Providing Services
- Expected Societal Outcome:
  - Reduce Preoccupation with and Attention to Technological Artifact (Techno-Clutter)
  - Improve Human Productivity by Use of Human Context
  - Improve Human Experience
- Expected Scientific Outcome:
  - Full Description and Understanding of *all* Human Communication Signals Across Multiple Modalities
  - Robustness in Perceptual User Interfaces; Always On
  - Databases, Standards



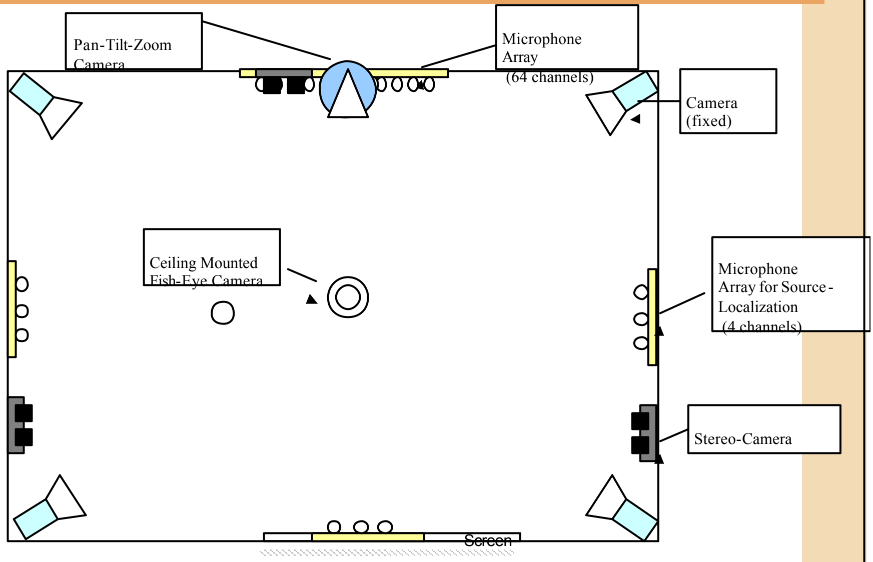
Project Overview

- **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,…)

Technologies & Fusion

- **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

Sensors in the CHIL Room



Pan-Tilt-Zoom Camera

Microphone Array (64 channels)

Camera (fixed)

Ceiling Mounted Fish-Eye Camera

Microphone Array for Source-Localization (4 channels)

Stereo-Camera

Screen

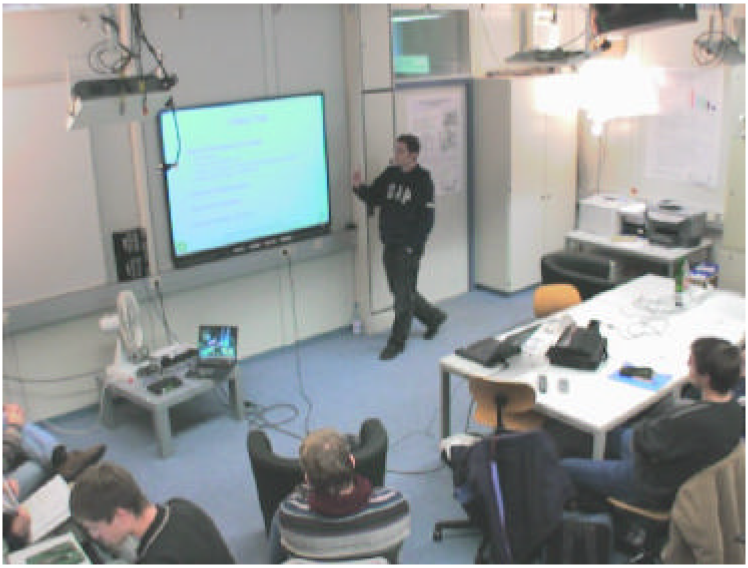
CHIL
Computers in the Human Interaction Loop

Scenario 1: Seminars/Lectures

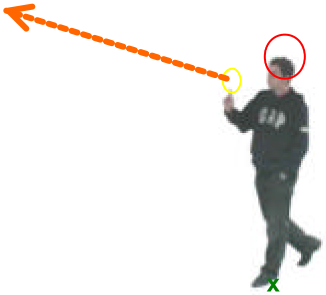
CHIL
Computers in the Human Interaction Loop

Scenario 2: Meetings

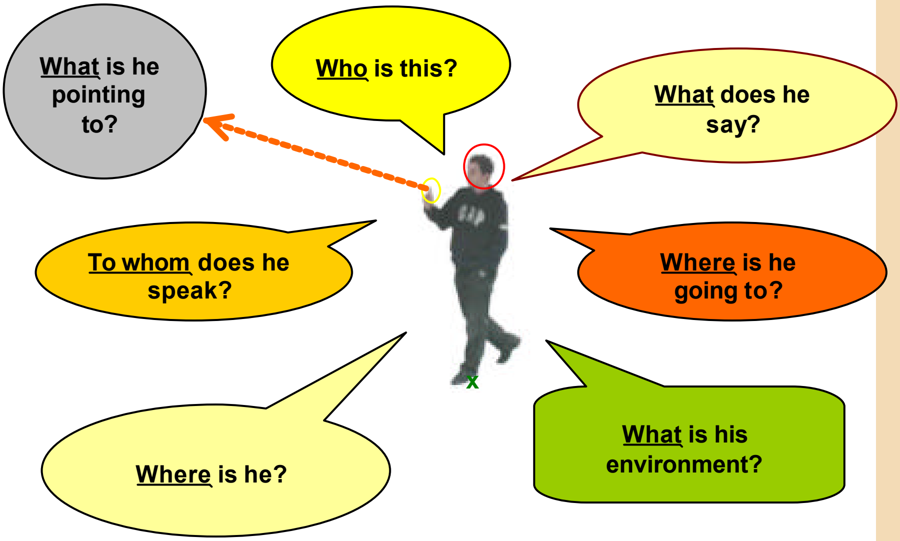
Describing Human Activities



Describing Human Activities



Technologies/Functionalities



What is he pointing to?

Who is this?

What does he say?

To whom does he speak?

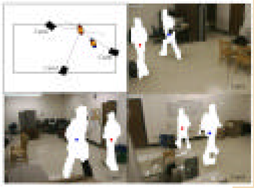
Where is he going to?

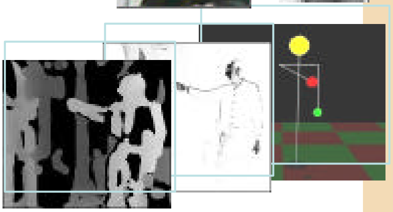
Where is he?

What is his environment?

CHIL Technologies at ISL

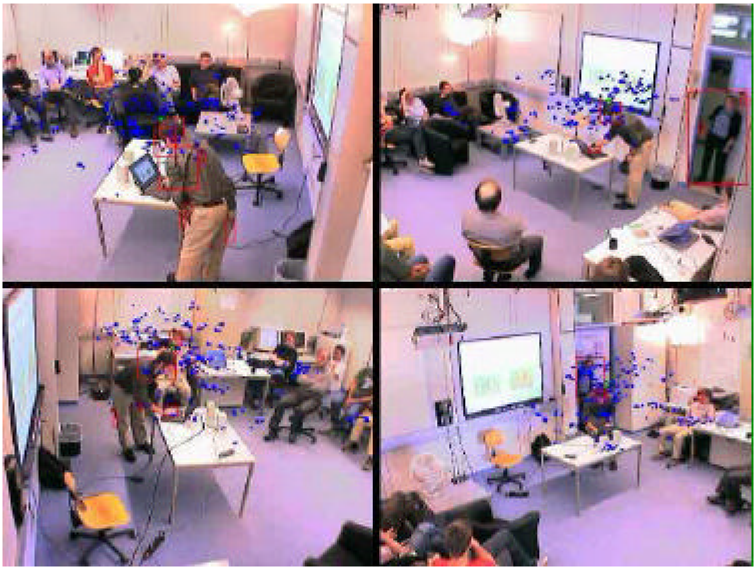
- **Vision**:
  - Person-Tracking
  - Face & Hand Tracking
  - Gesture Recognition
  - Head Pose & Focus of Attention
  - Activity Analysis (AV)
  - Person Identification & Identity Tracking
  - (AV-Speech Recognition)
- **Speech:**
  - Far-field LVCSR
  - Source Localization
  - Speaker ID
  - Topic ID
  - (Dialogue)



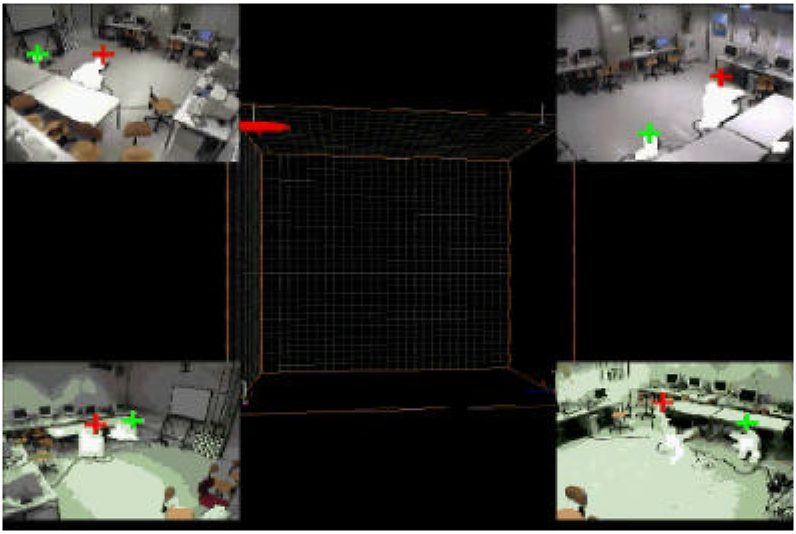


Where ?

Lectures / Seminars



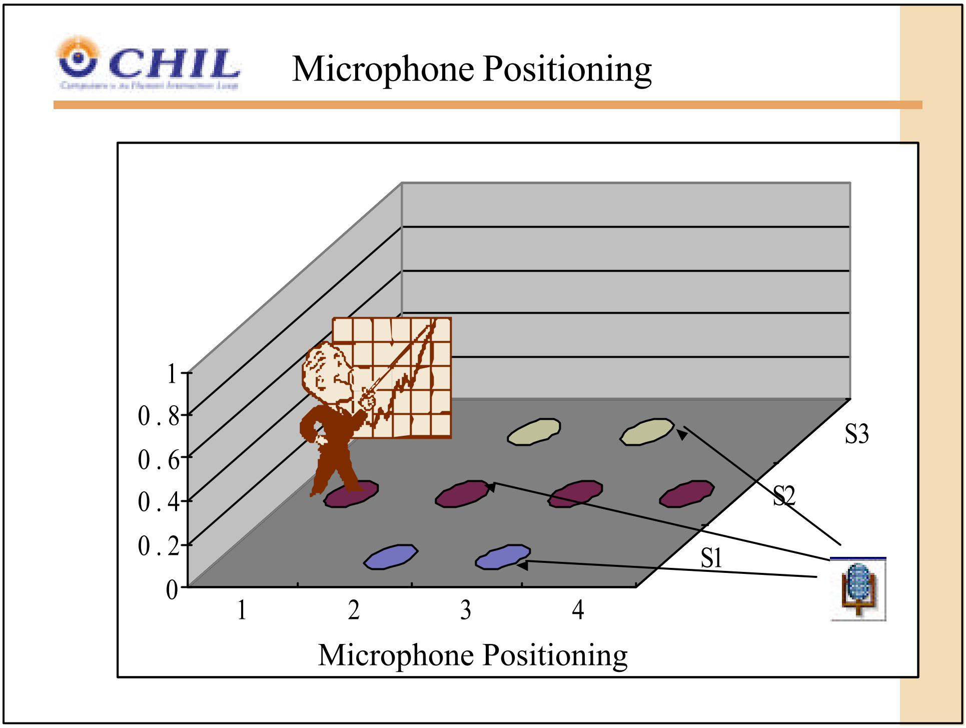
Where ? : 3D Person-Tracking



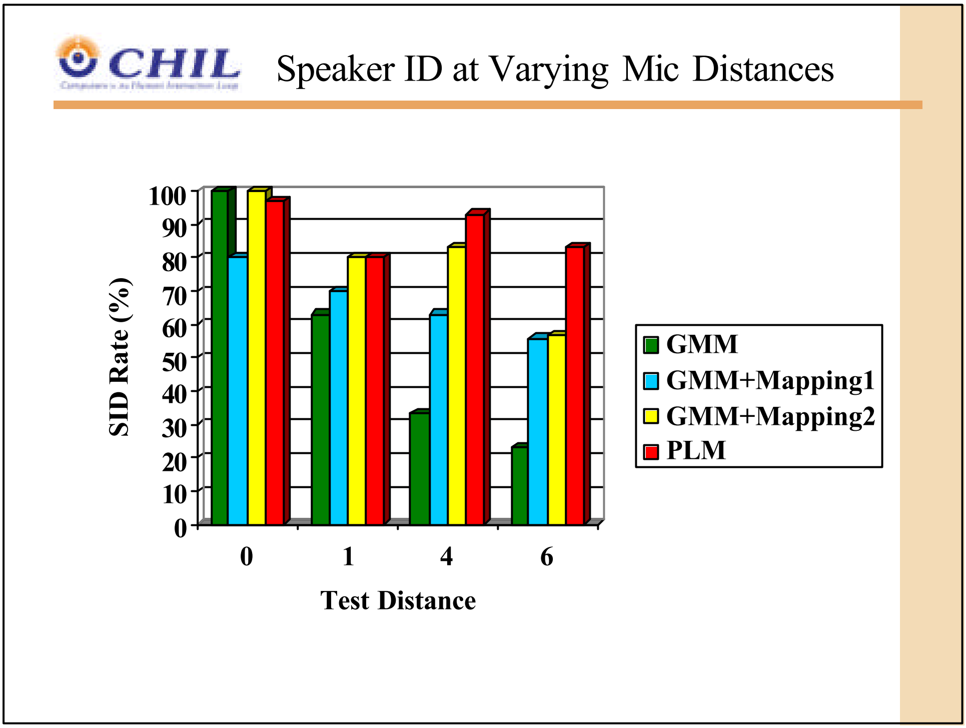
(Focken & Stiefelhagen, ICMI 2002)

Who ?

- Speaker – ID
- Face – ID



Microphone Positioning

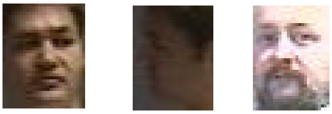


Speaker ID at Varying Mic Distances

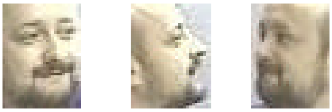
Face Identification

Low quality

Illumination

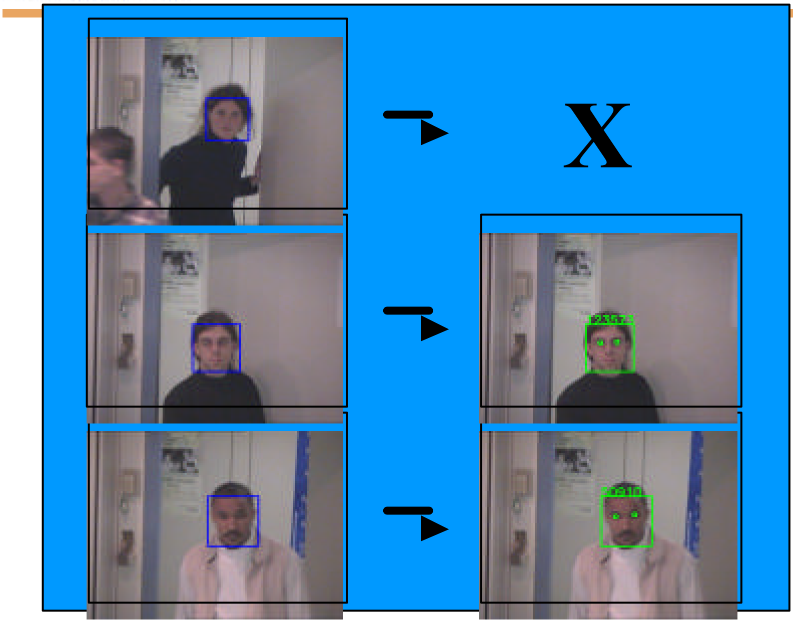


Head pose



Occlusion

Who? Face Identification



Face ID

FID on a Robot Platform

FID (Simulation) – on Robot Platform

Who and Where ?

Integrated Who and Where?

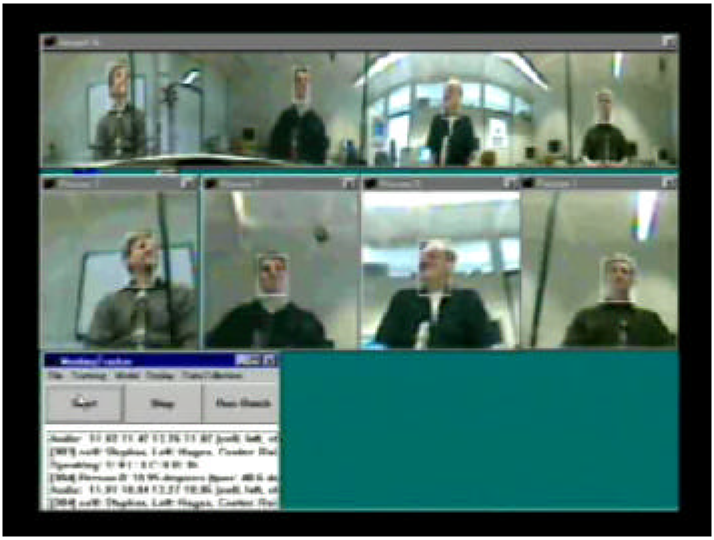
- Tracking *and* Identifying People *over Time*
- Multimodal Integration of Technologies
  - People Localization
    - Person Tracking (Visual)
    - Sound Source Localization (Acoustic)
  - People ID:
    - Face ID
    - Speaker ID

Integrated People Tracking and ID

To Whom ?

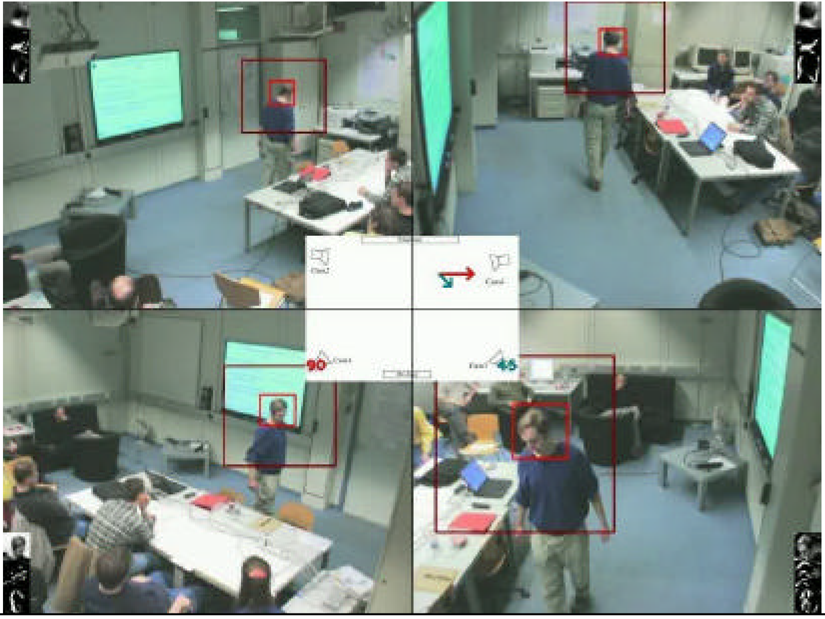
- Focus of Attention

Focus of Attention in Meetings



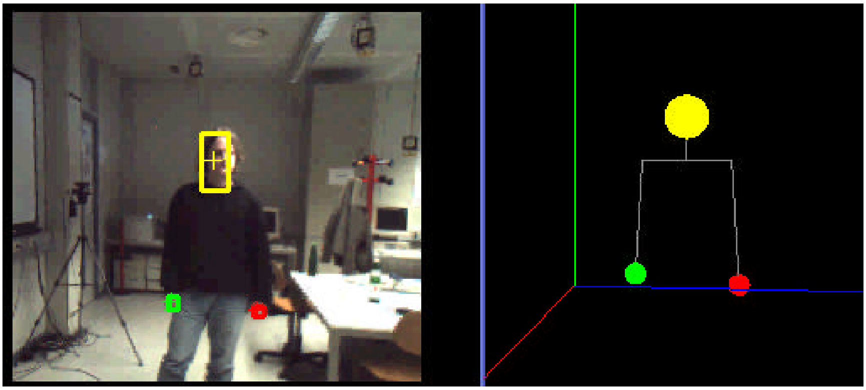
(Stiefelhagen, 2001)

FoA in Open Room Environment



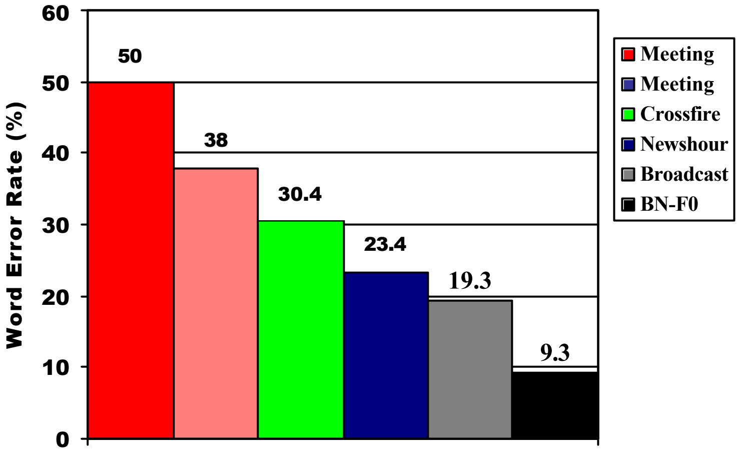
What ?

Who? What? Gesture Recognition



(Nickel & Stiefelhagen, ICMI 2002)

Recognition of Conversational Speech



CHIL

Lecture Speech Reco (WER)

30%
Danish

35%
English

20%
French

40%
Italian

65%
French

45%
Japanese

35%
French

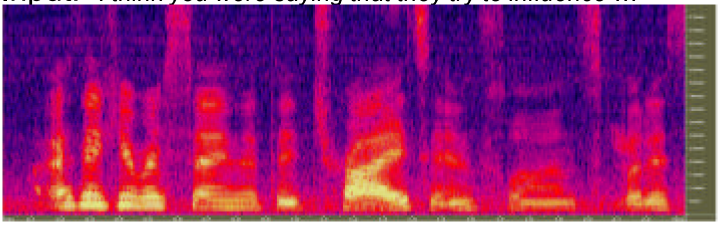
20%
German

CHIL

Sloppy Speech in Meetings/Lectures

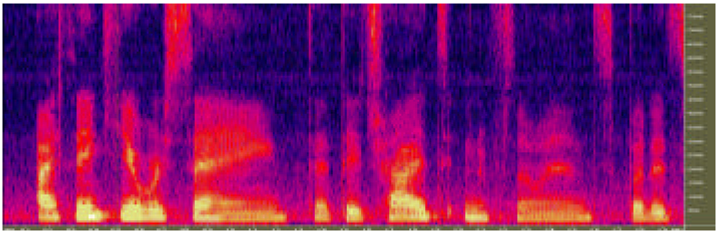
Actual Input: "*I think you were saying that they try to influence …*"

Conver-
Sational
Speech



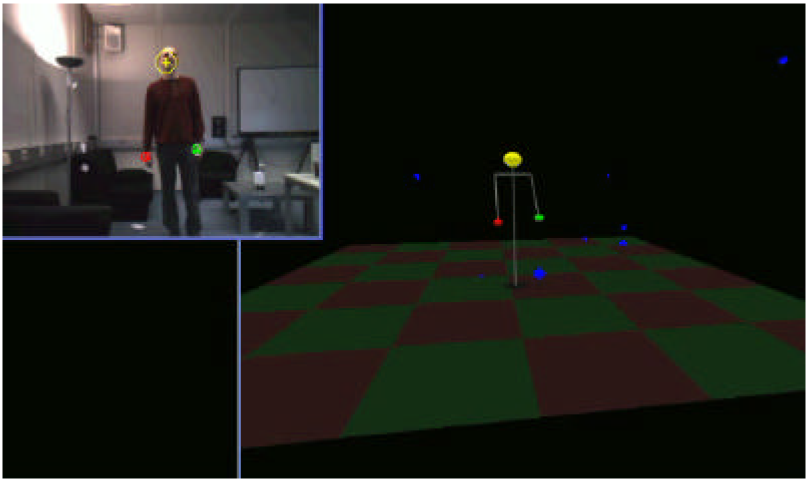
Recognition: "*I think you insanity tries influence …*"

Read
Speech



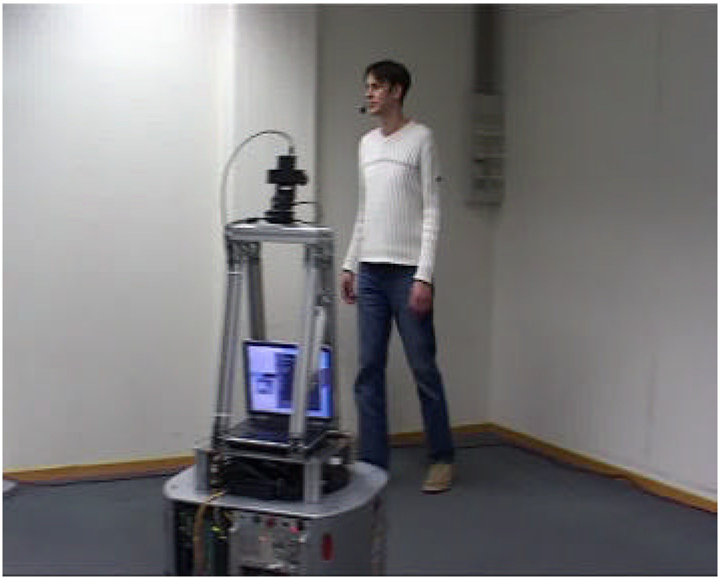
Recognition: "*I think you were saying that they tried to influence …*"

Pointing Gesture Recognition



(Nickel & Stiefelhagen, ICMI 2003, FG2004)

Robot-Human Dialog & Gestures



What? Speech Recognition

- Microphone Likely to be Remote
- Speaking Style Conversational

Non-Verbal Cues for Rich Transcription

Transcript: Onune baksana be adam!

Turkish	**Language ID**
Bus Station	**Acoustic Scene**
Angry	**Emotion ID**
Negotiation	**Discourse Analysis**
Umut	**Speaker ID**
Chemicals	**Topic ID**
Istanbul	**Entity Tracking**

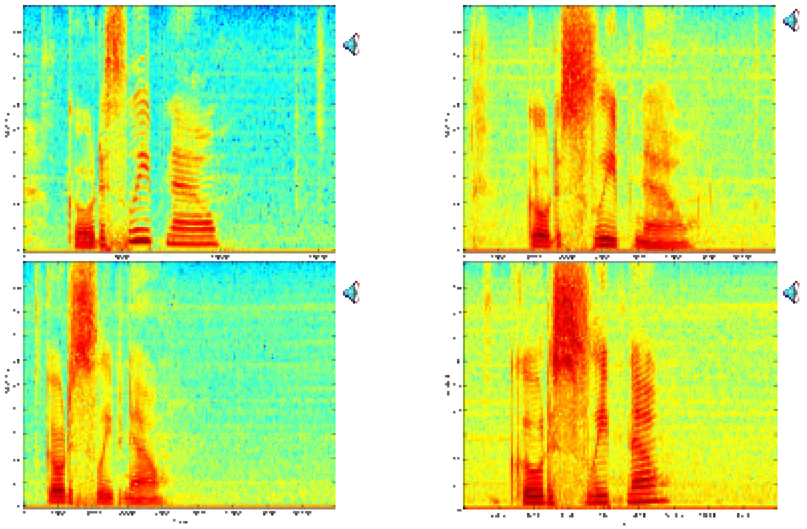
Acoustic Scene Analysis

A day in the life of Rob Malkin, (PhD work, ISL-CMU)

(Acoustic Scene Analysis & Audio Gisting)

How ?

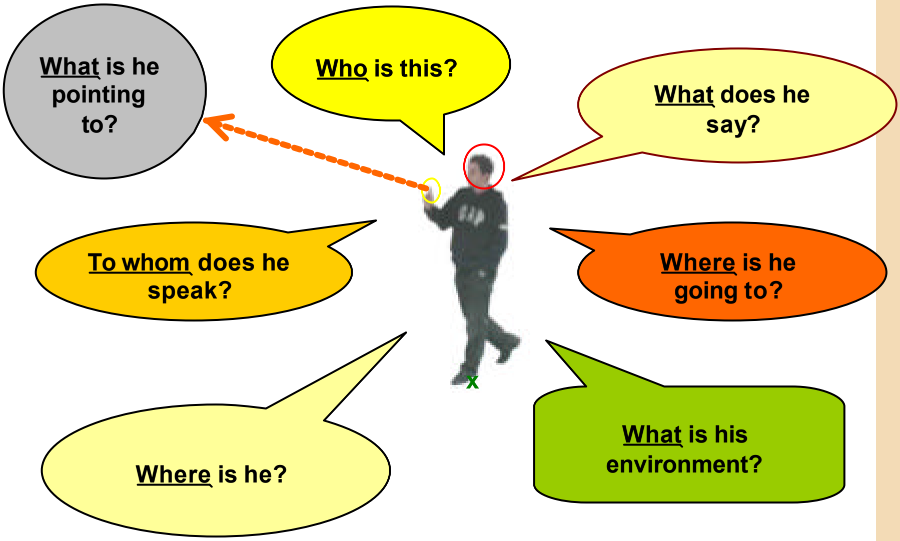
Emotional Speech



Evaluation

- Evaluations are Key to Assessing and Driving Progress
  - Benchmarks, Measures of Performance (MOPs)
  - User Studies, Measures of Effectiveness (MOEs)
- Functionalities & Technologies
  - Working Group in Each Area
  - Define Metrics, Databases and Benchmarks
  - Performance Benchmark Evaluations in Each Area
  - First “Dry-Run” Eval Already Completed in June 2004
  - Second Eval January 2005
    - External Sites Participate for First time
- Services
  - Technologies – Services: Catalog of Technologies
  - Initially Demos, Prototypes
  - Site Visits, Compare and Contrast (Done in November’04)
  - User Studies, Usability Measures
    - Working Group to Develop Metrics

Technologies/Functionalities



What is he pointing to?

Who is this?

What does he say?

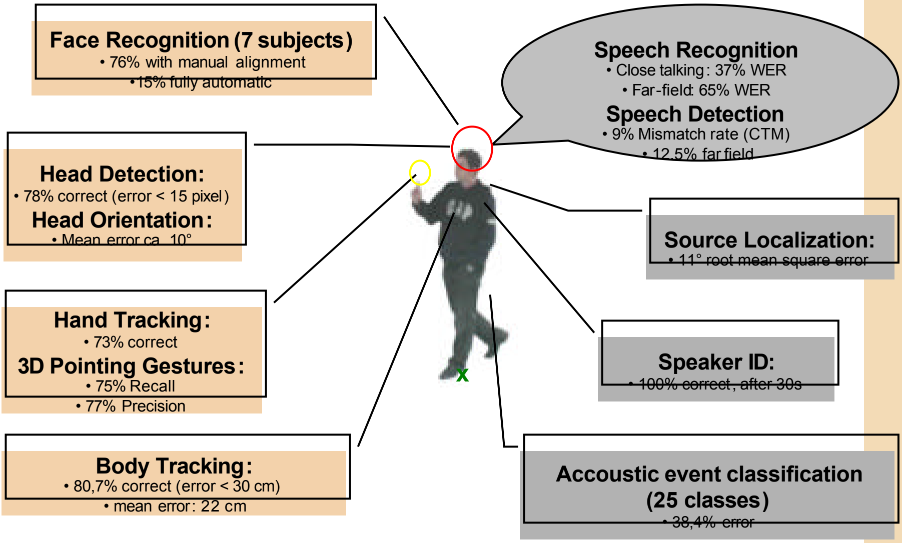
To whom does he speak?

Where is he going to?

Where is he?

What is his environment?

Results, June 2004



Face Recognition (7 subjects)
- 76% with manual alignment
- 15% fully automatic

Speech Recognition
- Close talking: 37% WER
- Far-field: 65% WER

Speech Detection
- 9% Mismatch rate (CTM)
- 12.5% far field

Head Detection:
- 78% correct (error < 15 pixel)

Head Orientation:
- Mean error ca. 10°

Source Localization:
- 11° root mean square error

Hand Tracking:
- 73% correct

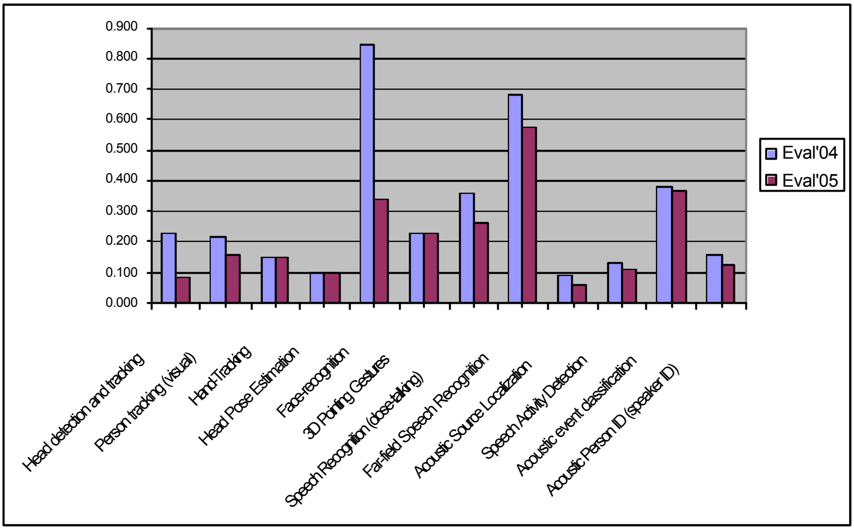
3D Pointing Gestures:
- 75% Recall
- 77% Precision

Speaker ID:
- 100% correct, after 30s

Body Tracking:
- 80,7% correct (error < 30 cm)
- mean error: 22 cm

Accoustic event classification (25 classes)
- 38,4% error

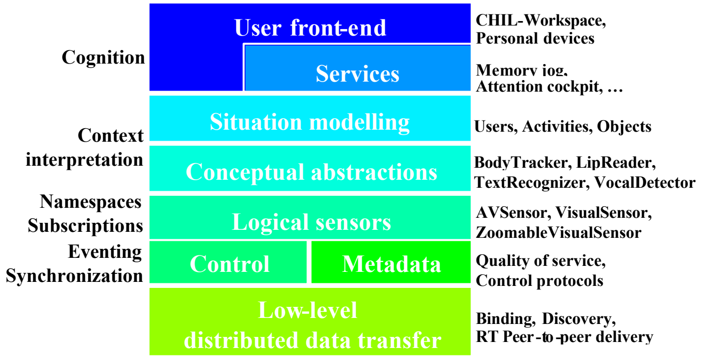
1-Year: Results and Progress



International Collaboration

- NIST and EC Programs Join Forces
  - RT-Meeting'05 – Rich Transcription
    - Emerges from established DARPA activity
    - MLMI Workshops, AMI/CHIL
    - Evaluated Verbal Content Extraction
    - Chair: Garofolo (NIST)
  - CLEAR'05 – 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 13 & 14, Southampton
    - Chair: Stiefelhagen (UKA)

Software Infrastructure



Cognition

User front-end

Services

CHIL-Workspace, Personal devices

Memorv iog, Attention cockpit, …

Context interpretation

Situation modelling

Users, Activities, Objects

Conceptual abstractions

BodyTracker, LipReader, TextRecognizer, VocalDetector

Namespaces

Subscriptions

Eventing

Synchronization

Logical sensors

AVSensor, VisualSensor, ZoomableVisualSensor

Control

Metadata

Quality of service, Control protocols

Low-level distributed data transfer

Binding, Discovery, RT Peer-to-peer delivery

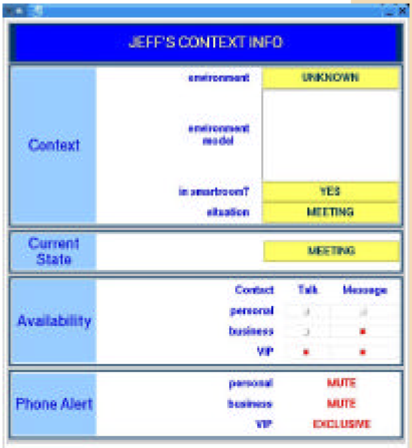
- Layers and APIs have been defined and published
- Implementation has started, first agent-based prototype ready
- NIST smartflow software used for low-level data transfer

Services: How Machines Assist Humans Interacting with Humans

CHIL Connector Video

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



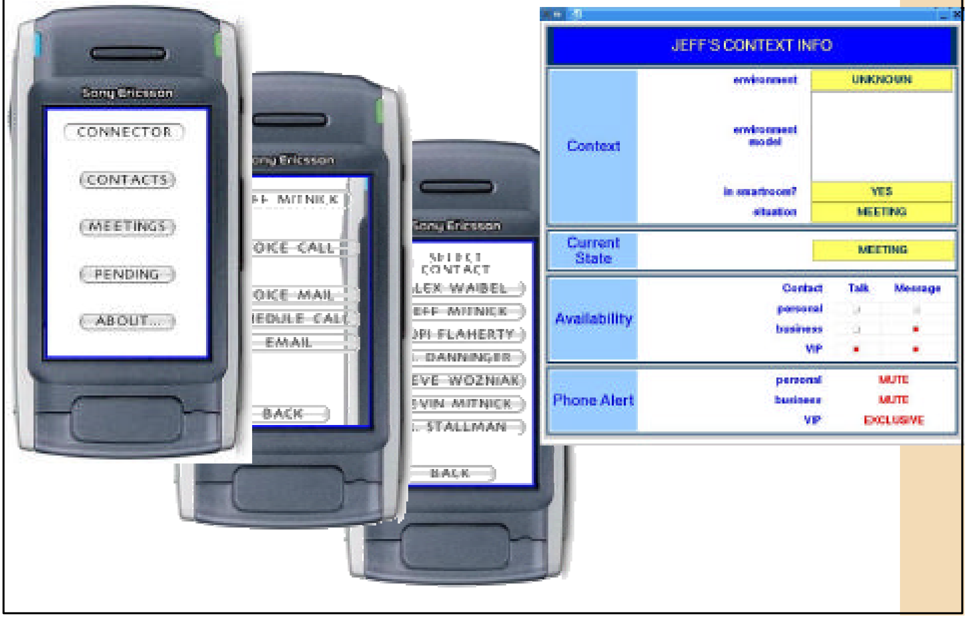
JEFF'S CONTEXT INFO

Context	environment	UNKNOWN
	environment model	
	in smartroom?	YES
	situation	MEETING
Current State		MEETING

Availability	Contact	Talk	Message
	personal	☐	☐
	business	☐	●
	VIP	●	●

Phone Alert	personal	MUTE
	business	MUTE
	VIP	EXCLUSIVE

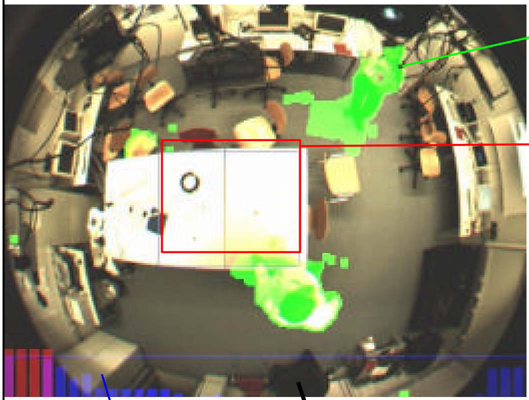
CHIL Connector GUI



JEFF'S CONTEXT INFO

Context	environment	UNKNOWN	
	environment model		
	in smartroom?	YES	
	situation	MEETING	
Current State		MEETING	
Availability	Contact	Talk	Message
	personal	□	□
	business	□	●
	VIP	●	●
Phone Alert	personal	MUTE	
	business	MUTE	
	VIP	EXCLUSIVE	

Meeting Recognizer



Background segmentation: Stauffer algorithm

Region of interest

Meeting Status

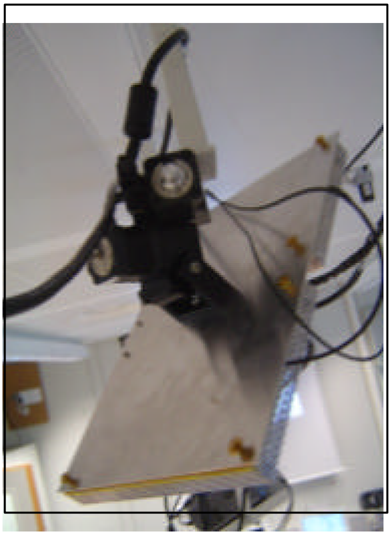
Top Cam View Point Grey Scorpion firewire camera with Fujinon Fisheye lens

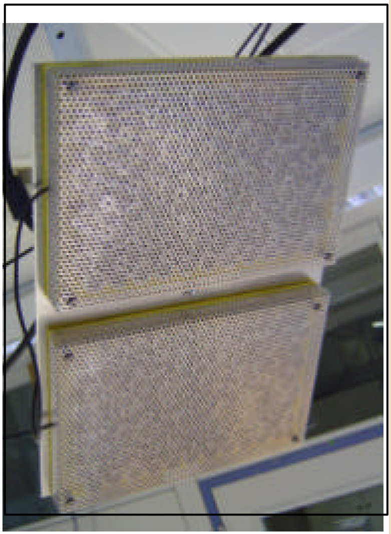
Connector

Private and Public Information Delivery

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

Targeted Audio

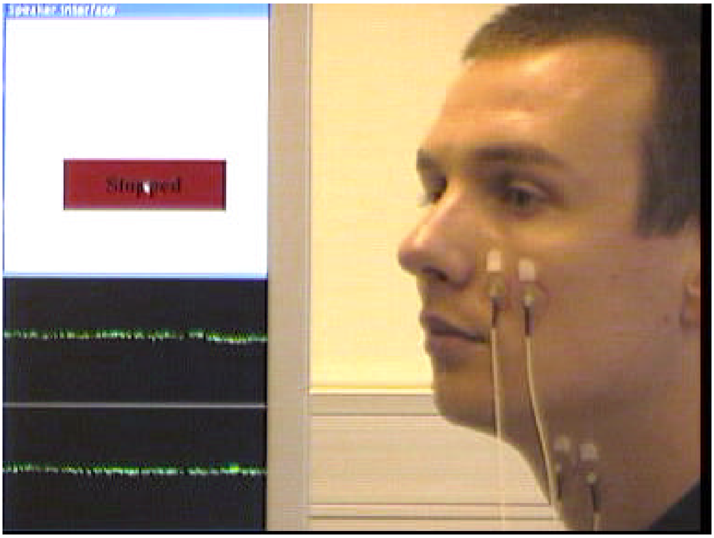




Silent Speech

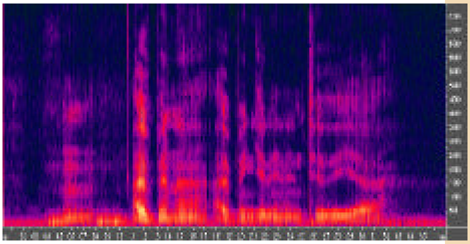
- The Problem of Personalized Input
  - Speech is Problematic
    - Neighbors are Disturbed
    - Neighbors hear what you say
  - Typing is too Slow
- Solution
  - Produce Silent Speech Input
  - EMG – Electrodes Capture Articulator Movement
  - Words are Recognized based on Muscle Activity
  - Speech is transmitted to caller

Silent Speech



Lecture Management System

- Lecture/Meeting May have Been Missed
  - Need Quick Review
- Automatic Records of Lectures, Comments, Q&A…
  - Intelligent Cameraman Records Lecture/Meeting
  - Recall Key Events in Passed Meetings
  - Have Private Memory



Observations: March-April 2004

Collaboration Support

- Documents
  - Drawing and writing
  - Rotations, translation and shrinking
- Agenda
  - Scheduled time
- Minutes
  - Linked to agenda and documents
- Tunneling to import and export documents
- Mirroring to a wall-display

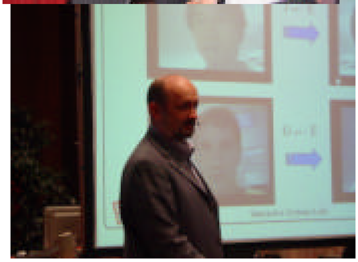
Memory Jog

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

Language Suppo

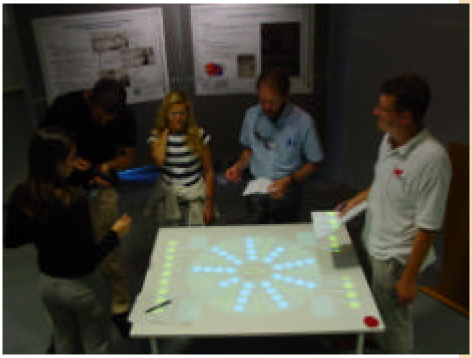
Browsing Lectures

Seminars (Barcelona July ’04)



Transcribe, Index, Summarize Lecture

Retrieve Based on Conversation



Language Tech Day, Barcelona, July ‘04

Language Suppo

PDA Speech Translation in Mobile Scenarios

- Tourism
  - Needs in Foreign Country
  - International Events
    - Conferences
    - Business
    - Olympics '08
- Humanitarian Needs
  - Humanitarian, Government
    - Medical
    - Refugee Registration
    - First Responder
  - Where?
    - USA, Latino Population
    - Third World
    - India 300 Languages, many unwritten

Translation of Lectures and Meetings

Projects:

- TC_STAR (EC FP6)
- STR-DUST (NSF)

Lecture Translation

Our first CHIL IP Review was Translated by Automatic Simultaneous Translation (May ’05)

Project CHIL

http://chil.server.de

- **Integrated Project** (IP) in 6th 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
  - Possible Follow-On in 2nd Phase

Management Approach

Successful Ingredients:

- Distributed Management Functions
  - University of Karlsruhe → Research
  - Fraunhofer IITB → Administration/Financials
- Management Philosophy
  - Evaluation and Coopetition
    - Benchmarks on Technologies
    - Services (more than one!!) Compare & Contrast
    - Workshops to Discuss Results
  - “Market Pull”, not Master Plan
    - Service Developers Compete & Subscribe to Needed Technologies
    - Technology Developers Publish Catalog
- Infrastructure
  - Data Resources and Distribution
  - Benchmarks and Metrics, Ground Rules
  - Architecture for Rapid Prototyping

Future Evaluations

- Services
  - Site Visits, Compare & Contrast (next: Nov. 2006)
  - Usability Tests
- Technologies & Functionalities
  - Open to Outside Participants (Since 2004)
  - New Challenge:
    Multiple People Tracked, Multiple Sites Data
  - In Preparation:
    - Worldwide Multimodal Benchmarking (“Olympic Games”)
    - EC Programs, NIST, … teaming …
    - International Workshops

International Collaboration

- NIST and EC Programs Join Forces
  - RT-Meeting’05 – Rich Transcription
    - Emerges from established DARPA activity
    - MLMI Workshops, AMI/CHIL
    - Evaluated Verbal Content Extraction
    - Chair: Garofolo (NIST)
  - CLEAR’05 –
    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 13 & 14, Southampton
    - Chair: Stiefelhagen (UKA)

CHIL

The CHIL Team

Coordination:

- Scientific Coordinator: Univ. Karlsruhe, Prof. A. Waibel, R. Stiefelhagen
- Financial Coordinator: Fraunhofer IITB, Prof. Steusloff, K. Watson

The CHIL Team:

IITB
Fraunhofer Institut Informations- und Datenverarbeitung

Universität Karlsruhe (TH)

DAIMLERCHRYSLER

ITC irst

TU/e technische universiteit eindhoven

ATHENS INFORMATION TECHNOLOGY

IBM

TALP UPC

KTH

STANFORD UNIVERSITY

CarnegieMellon

INRIA

el Da