

# Multi- and Cross-Lingual Dialog Systems

Alex Waibel and the InterACT Team  
Carnegie Mellon University  
Karlsruhe Institute of Technology  
Mobile Technologies, LLC

[alex@waibel.com](mailto:alex@waibel.com)







PRESS  
TO TEST  
WARN LYS



FULL  
FOR  
QUICK  
REC



N520AN

MINIMUM N<sub>1</sub> SPEED  
STARTING RECOMMENDATIONS  
OAT °C 10 & BELOW 10 TO 71.7 & ABOVE  
N<sub>1</sub>% 12 13 15







**Carnegie Mellon**

**MOBILE**TECHNOLOGIES

**KIT**  
Karlsruhe Institute of Technology





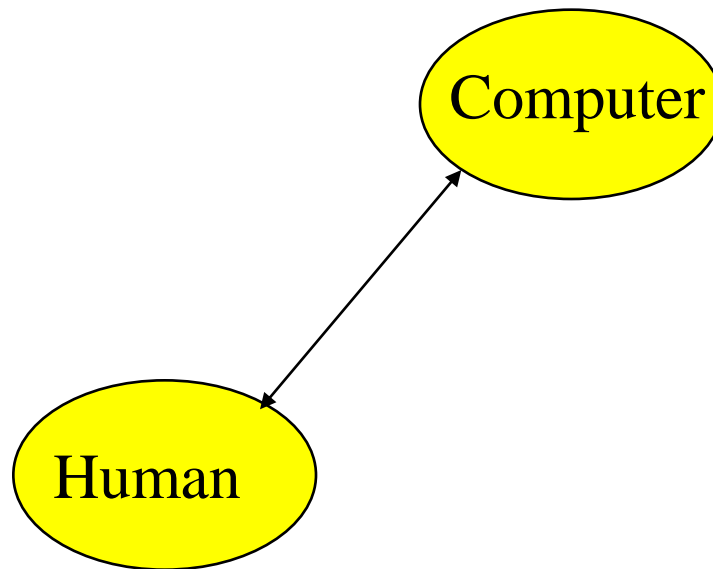














# Present Human-Computer Interaction



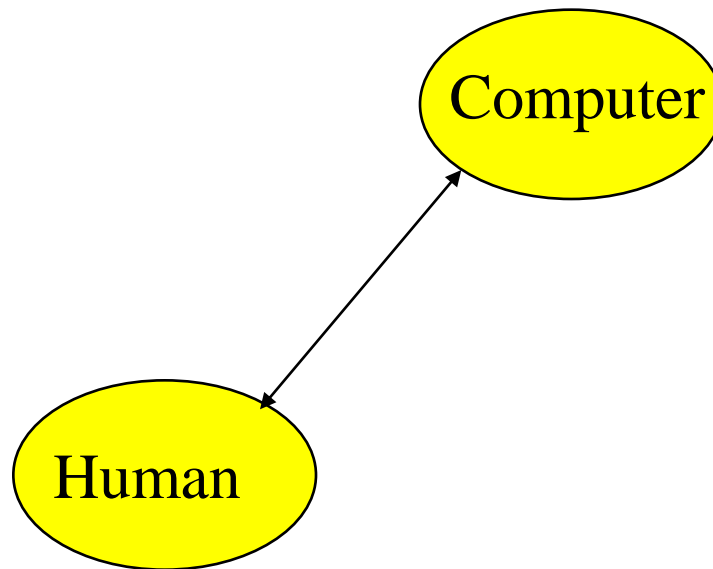


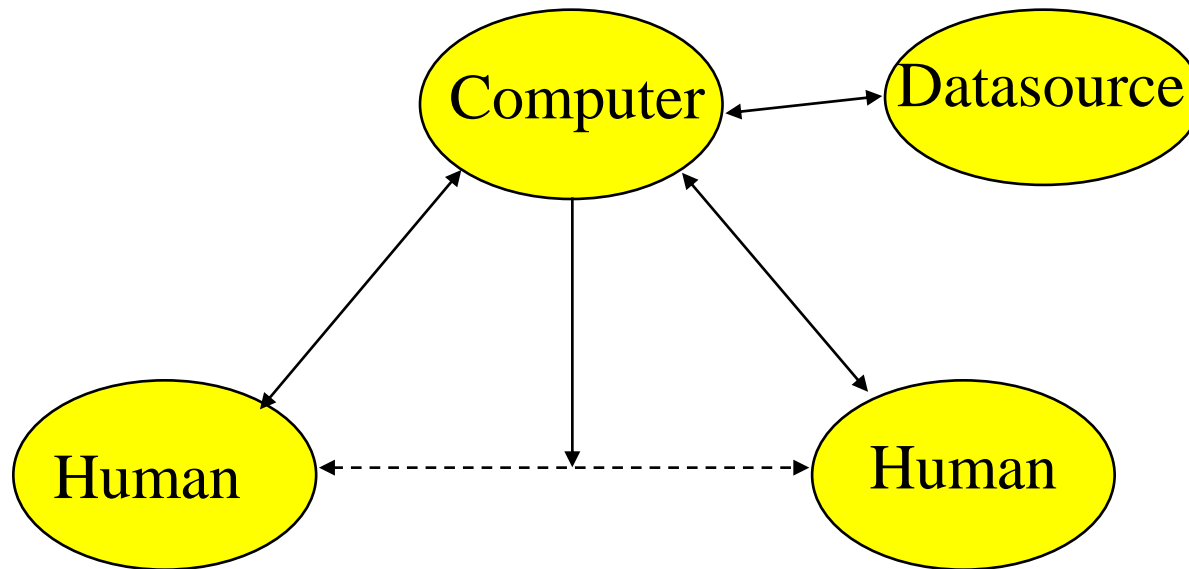
**interACT**

# Humans Interacting With Humans









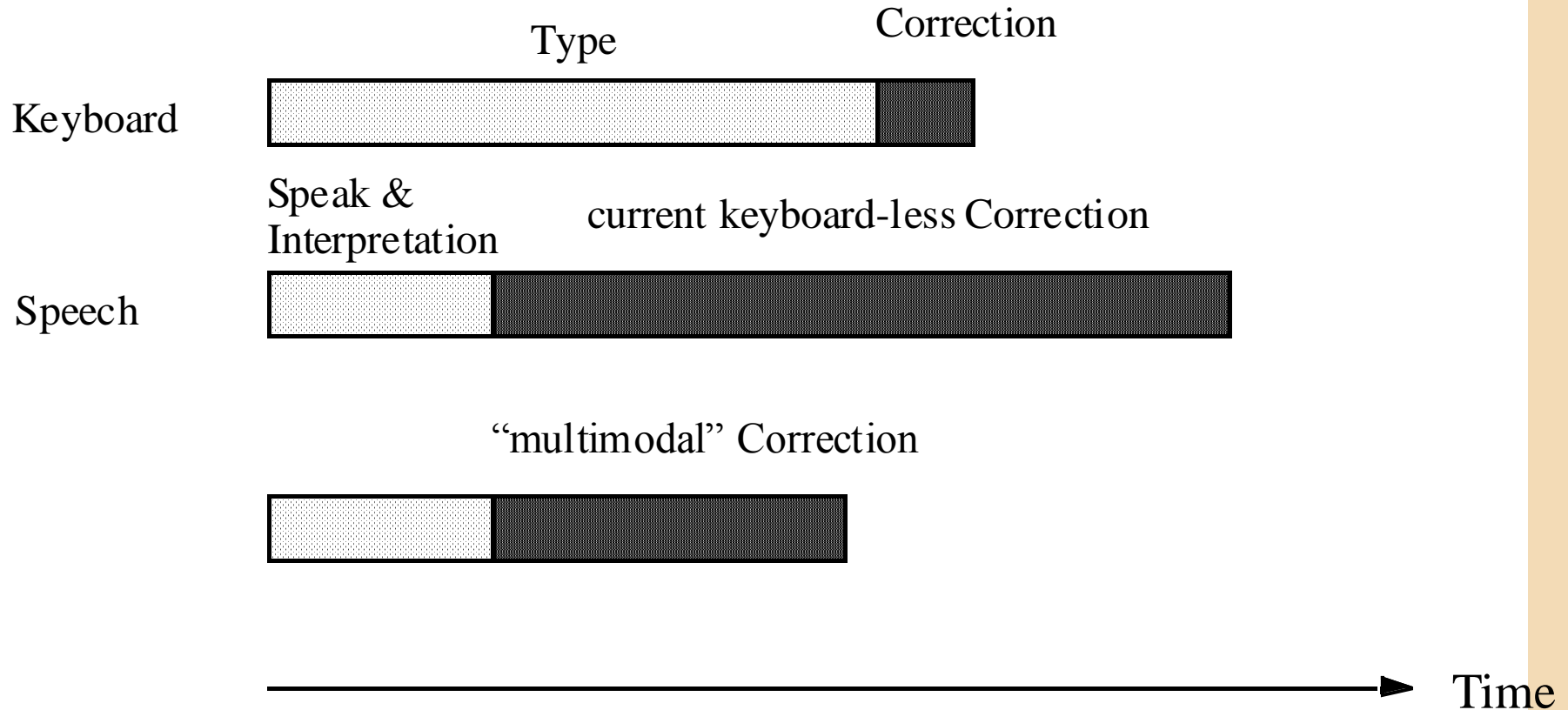


- Human-Machine:
  - Example: Human-Robot Interaction
  - To Err is not only Human
  - Multimodal Dialogs
- Human-Human:
  - Example: Computers in the Human Interaction Loop
  - Context Aware Agents
  - Implicit and Explicit Interaction
- Human-Computer-Human
  - Example: Cross-Lingual Communication
  - Machine as Mediator
  - Consecutive and Simultaneous

# *Human-Machine: Challenges and Lessons Learned*

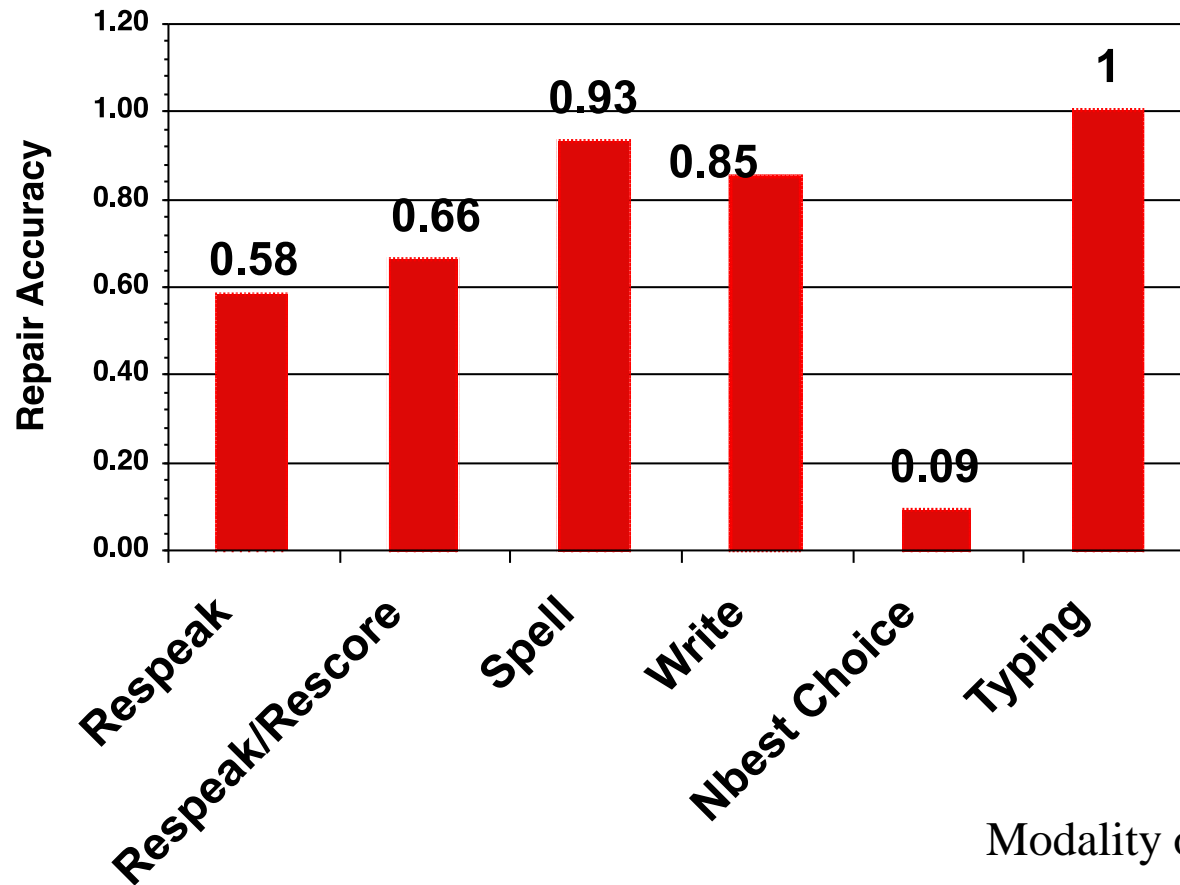


- To Err is Human
- Repair:
  - Repair by Repeating is Singularly Ineffective
  - Error Repair by Dialog
  - Cross-Modal Repair
- Cross-Modal Repair
  - Two Patents

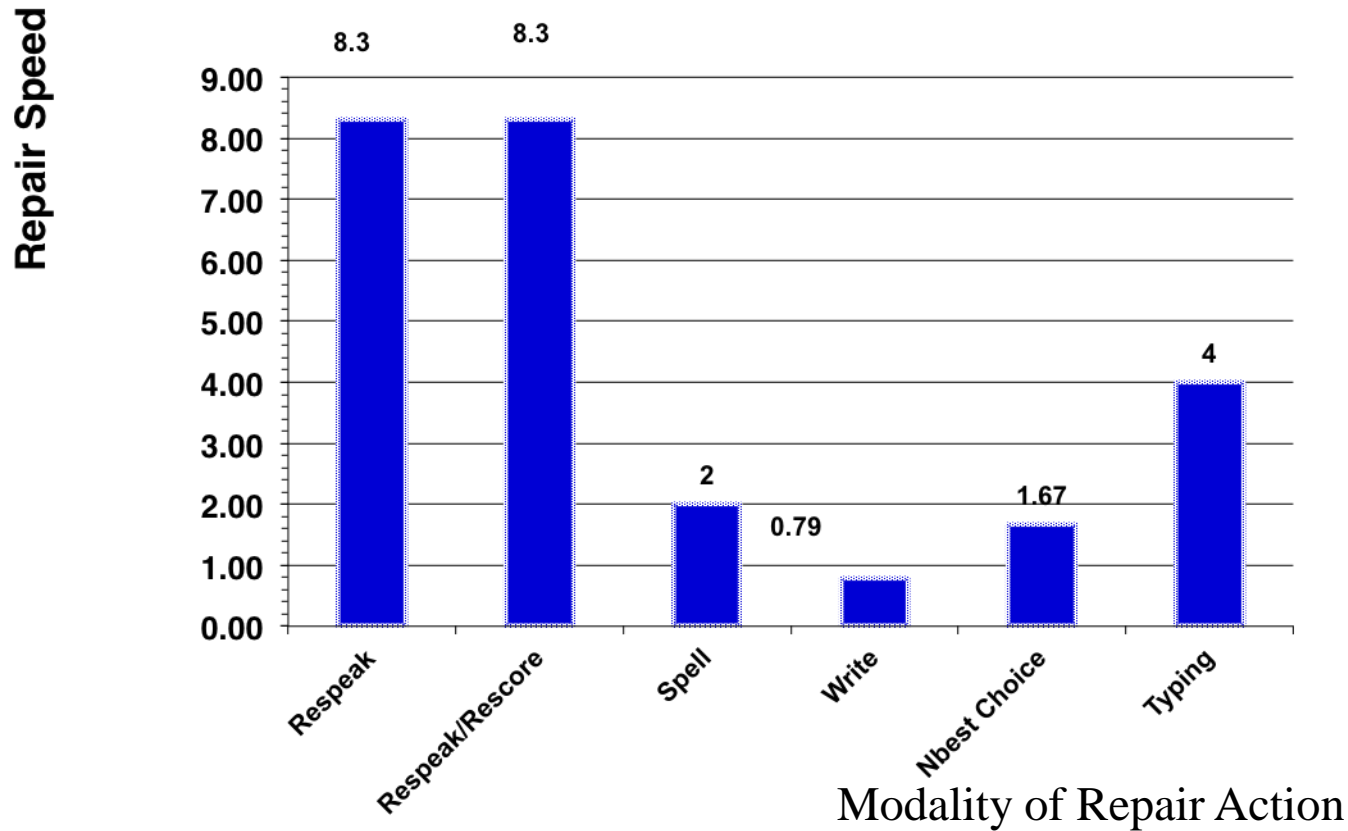




## Accuracy of Repair



## Speed of Repair





# Gestures for Editing and Partial Word Correction


*Delete Words and Characters:*

~~prototype~~ multimodal listening typewriter

~~prototype~~ multimodal listening typewriter

**prototypical** ~~mul~~

*Indicate Cursor Position:*

**prototype**  **multimodal**

**prototype** | **multimodal**

*Partial Word Correction:*

**prototypical** **mul**  

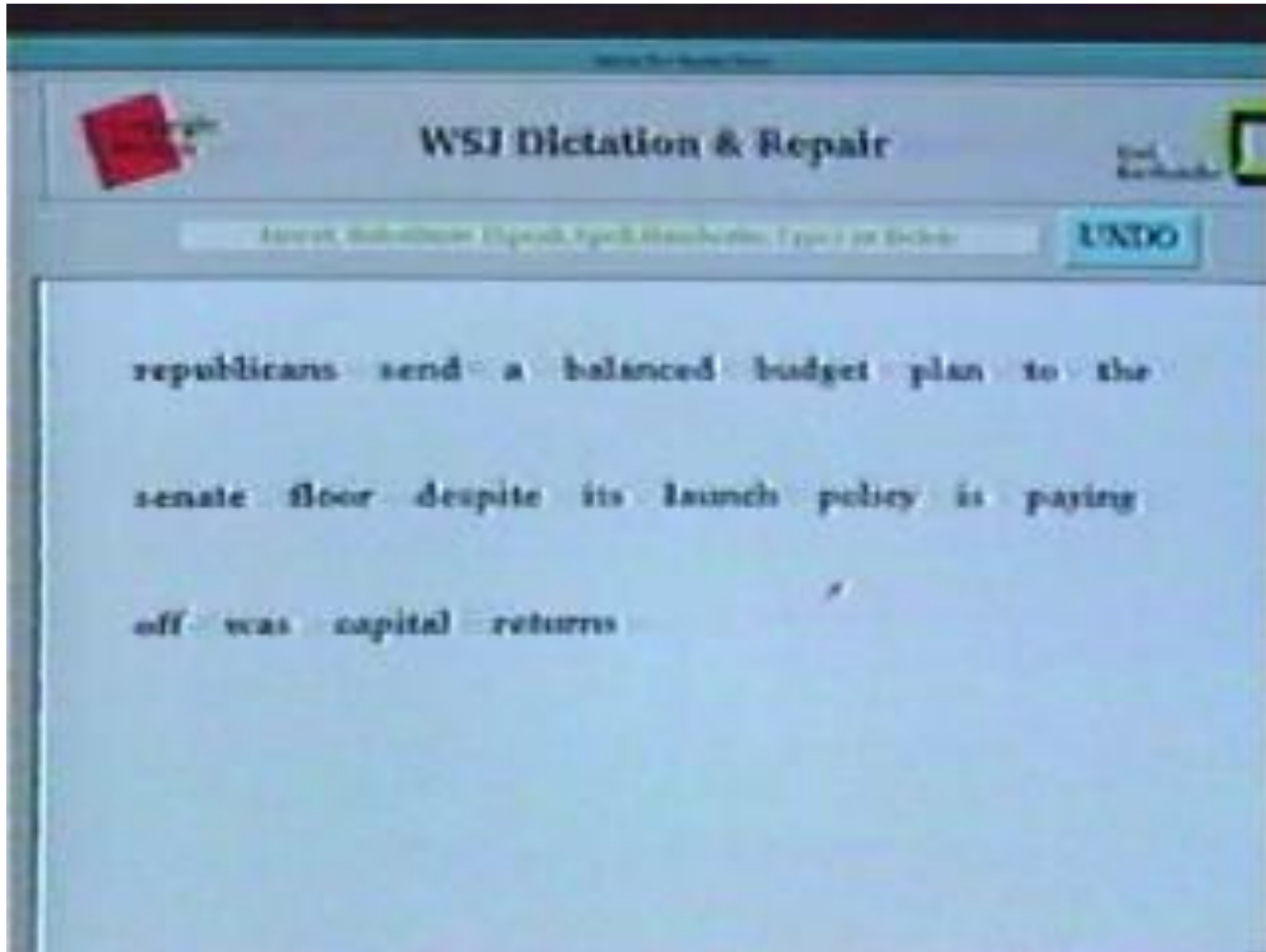

*Select Characters:*

**prototypical** | **mul**



interACT

# Multimodal Repair: (1996-1999 !)

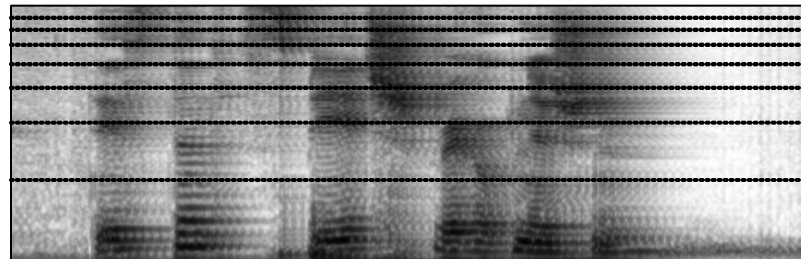
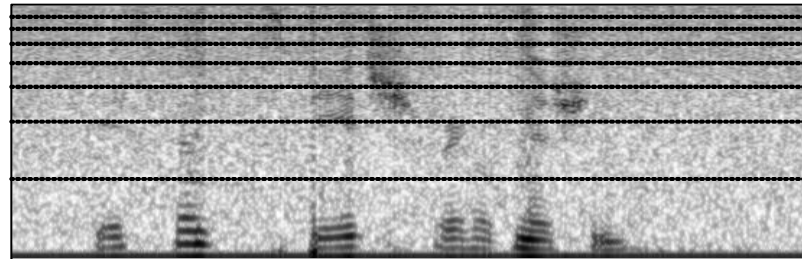
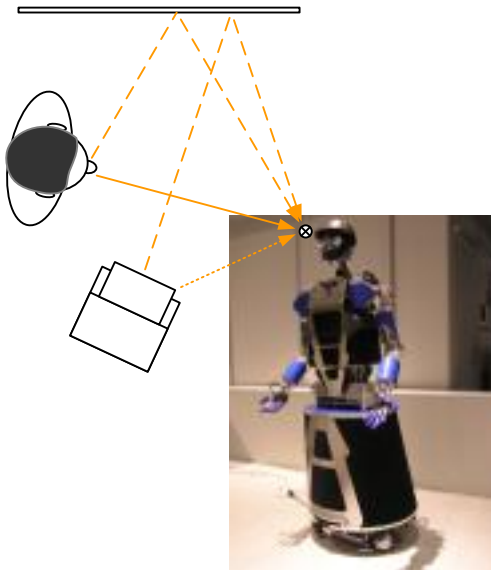
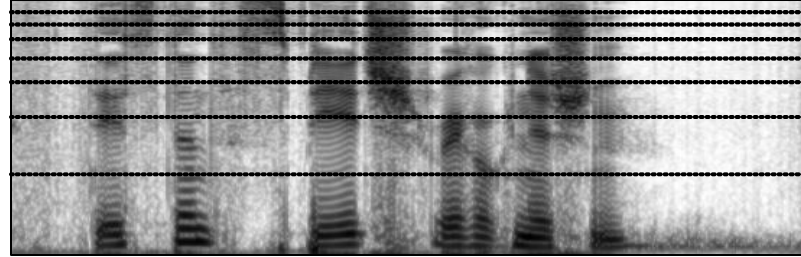
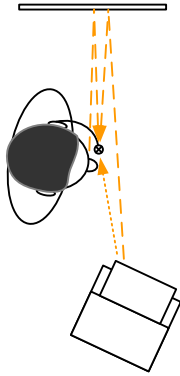


Repair





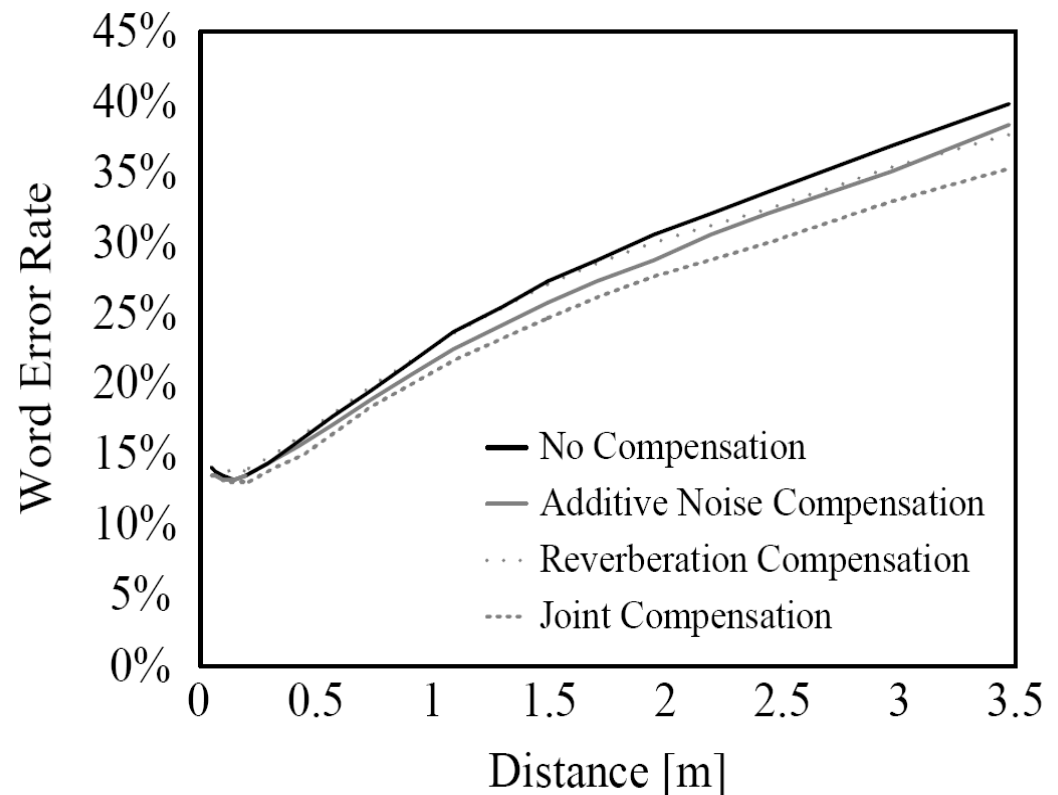
- Humanoid Robots
- SFB-588
  - 10-Year Research Center
- Joint Research:
  - Robotics
  - Multimodal Perception
  - Dialog
  - Planning



## Speech Recognition

### Without Buttons and Close Speaking Mics

- iPhone Input
- Aktive Listening  
→ Robot Turns/Drives Closer
- MicArray  
→ Small Improvements
- Distant Mic Processing
  - Dereverberation
  - Joint Particle Filter







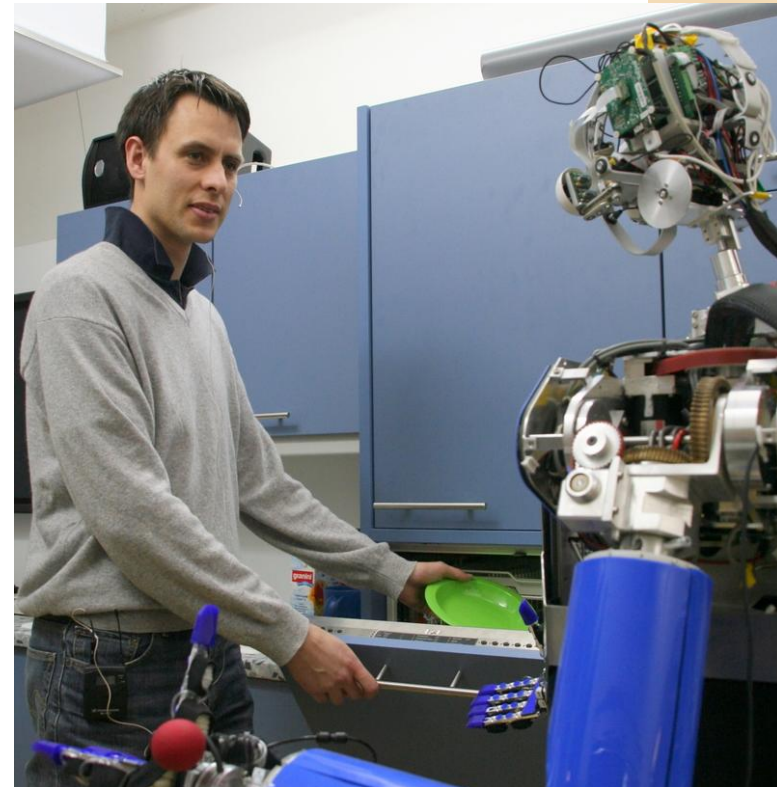


## Adaptation of Dialog Strategy

- Acoustic Adaption
- Dialog Adaptation Situationen, Objekts und Rolls
- Learning of Concepts
- Extended Behavior Network

## Longitudinal Study: One Year 24/7 Operation

- Learning and Forgetting
- New Knowledge is Errorful  
and Needs to be Forgotten

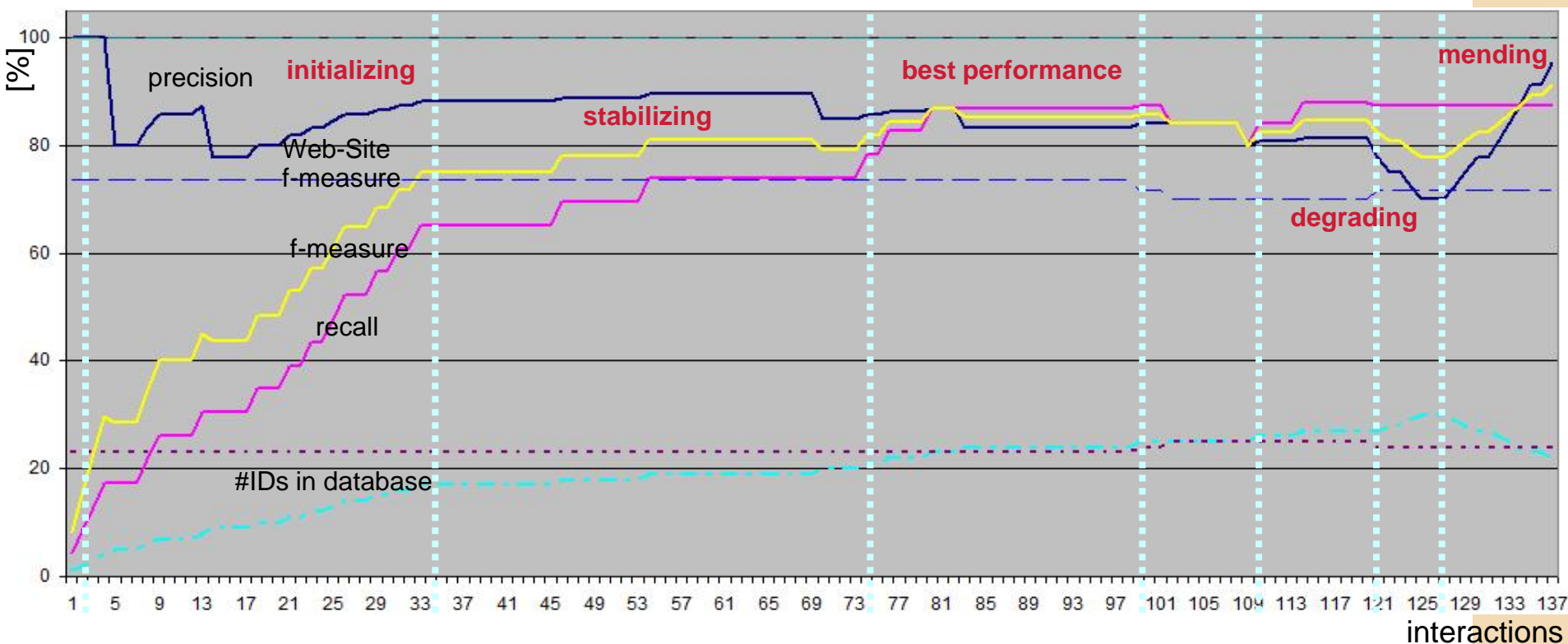


# Knowledge Mending: Personeneinträge in der Datenbank beim Lernen über der Zeit

Evaluation: person ID  
 “dynamic” interACT only  

$$\text{precision} = \frac{\text{correct labels}}{\text{learned labels}}$$

precision	#IDs	first-WER
recall	#real	last-WER
f-measure		Web fmeasure





# *Human-Human Interaction: Challenges and Lessons Learned*



- CHIL – Computer in the Human Interaction Loop
  - Rather than Humans in the Computer Loop
  - Explicit Computing Complemented by Implicit Support
- Implicit Computing Services
  - Support Human-Human Interaction Implicitly
  - Increasingly Powerful Computing Services
  - Implicit Services Observe Context and Understanding
  - Reduction in Attention to Technological Artifact,  
→ Increased Productivity
  - Computer Learns from Human Activity Implicitly



*“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 !**

## Coordination:

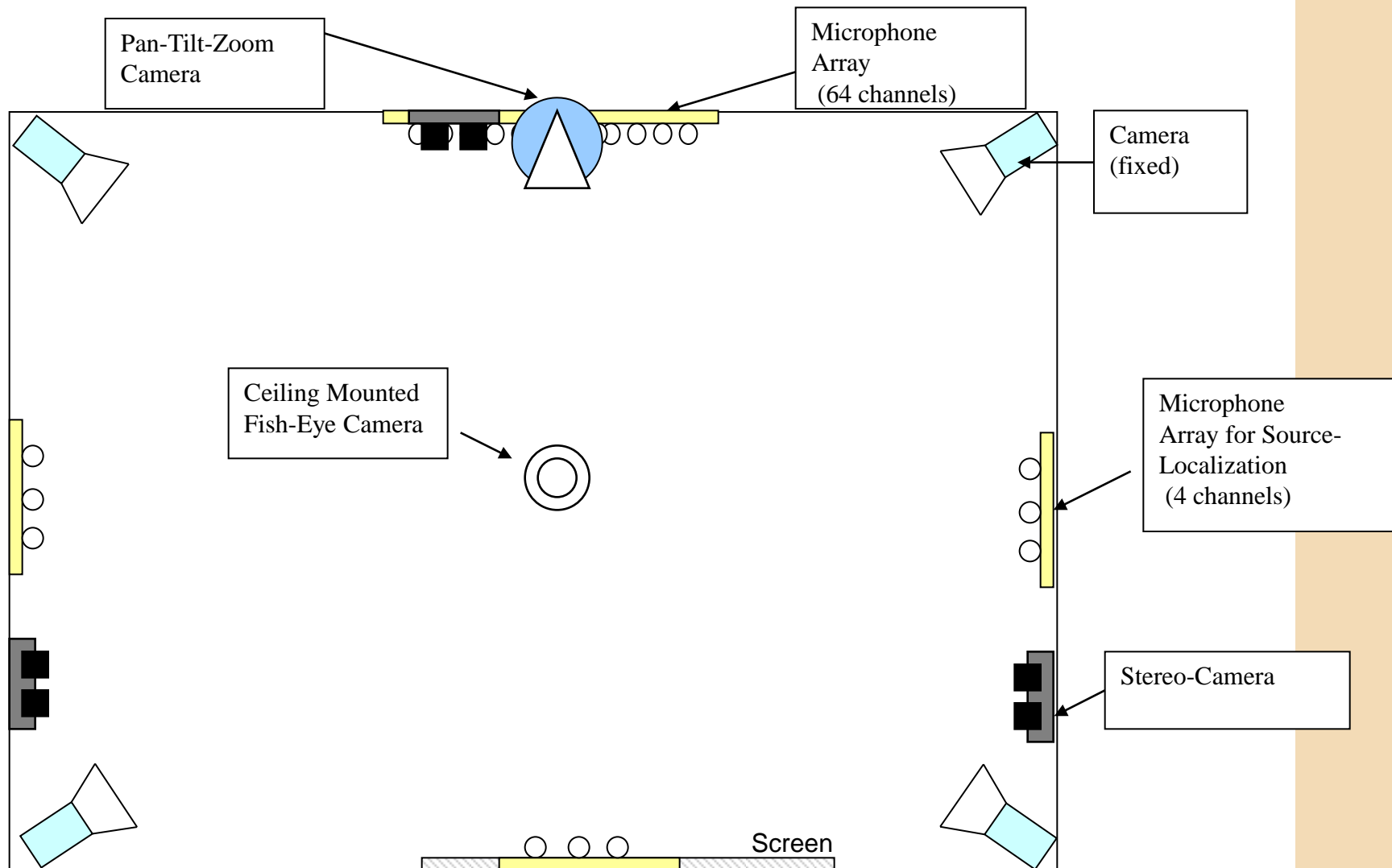
- Scientific Coordinator: Univ. Karlsruhe, Prof. A. Waibel, R. Stiefelhagen
- Financial Coordinator: Fraunhofer IITB, Prof. Steusloff, K. Watson

## The CHIL Team:



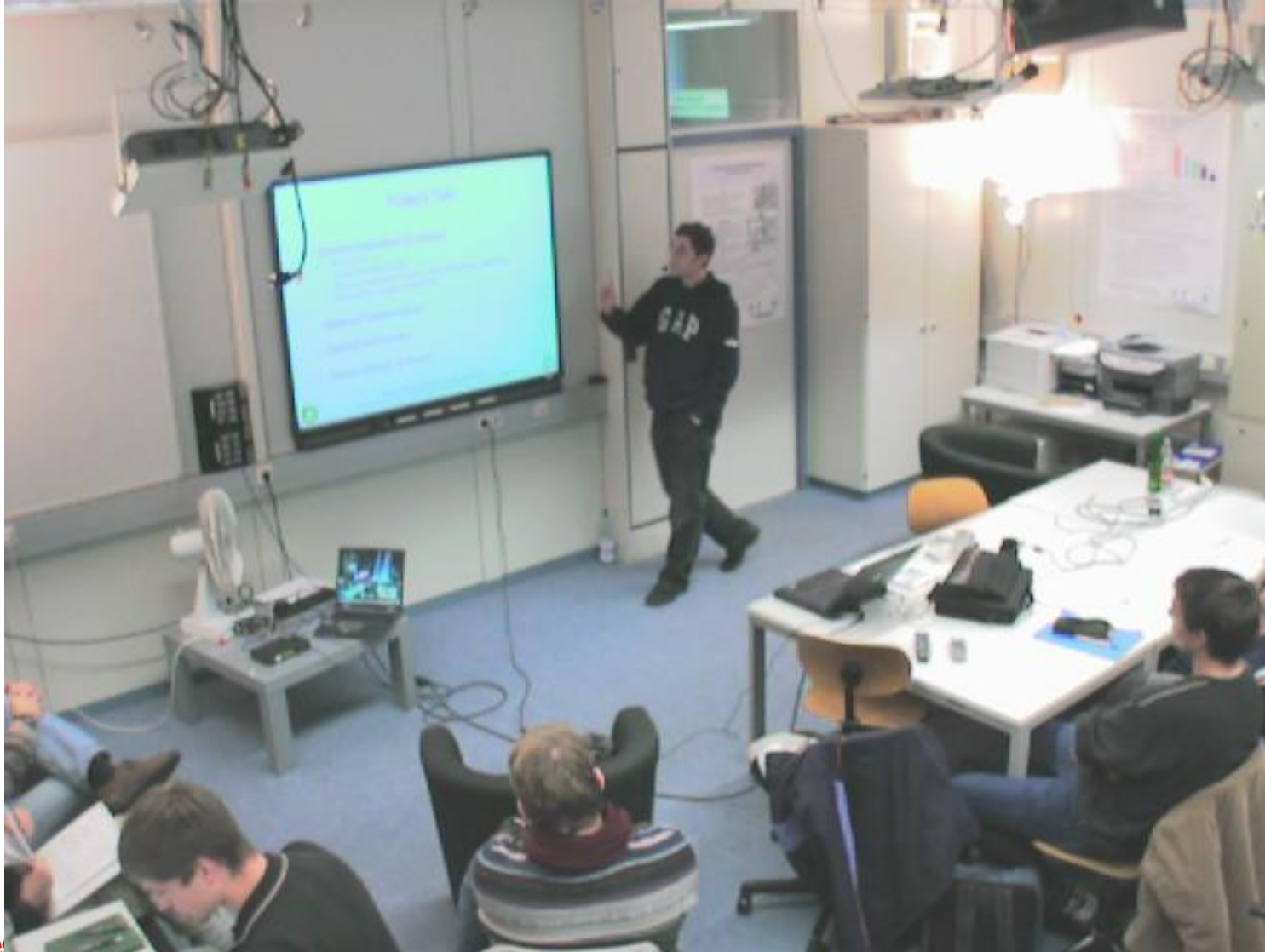


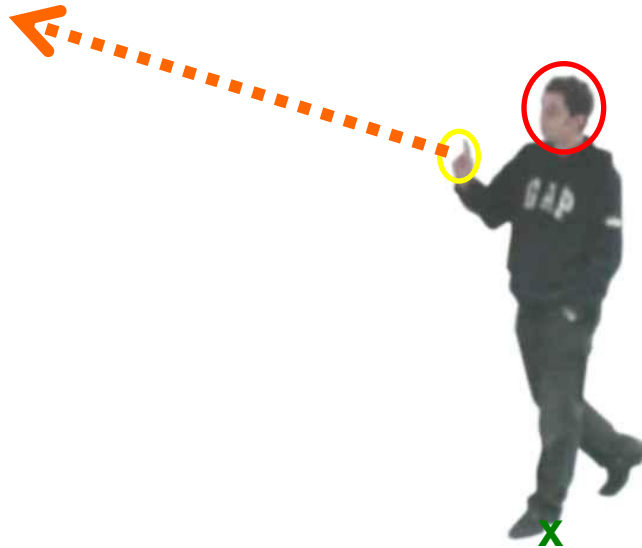
# Sensors in the CHIL Room

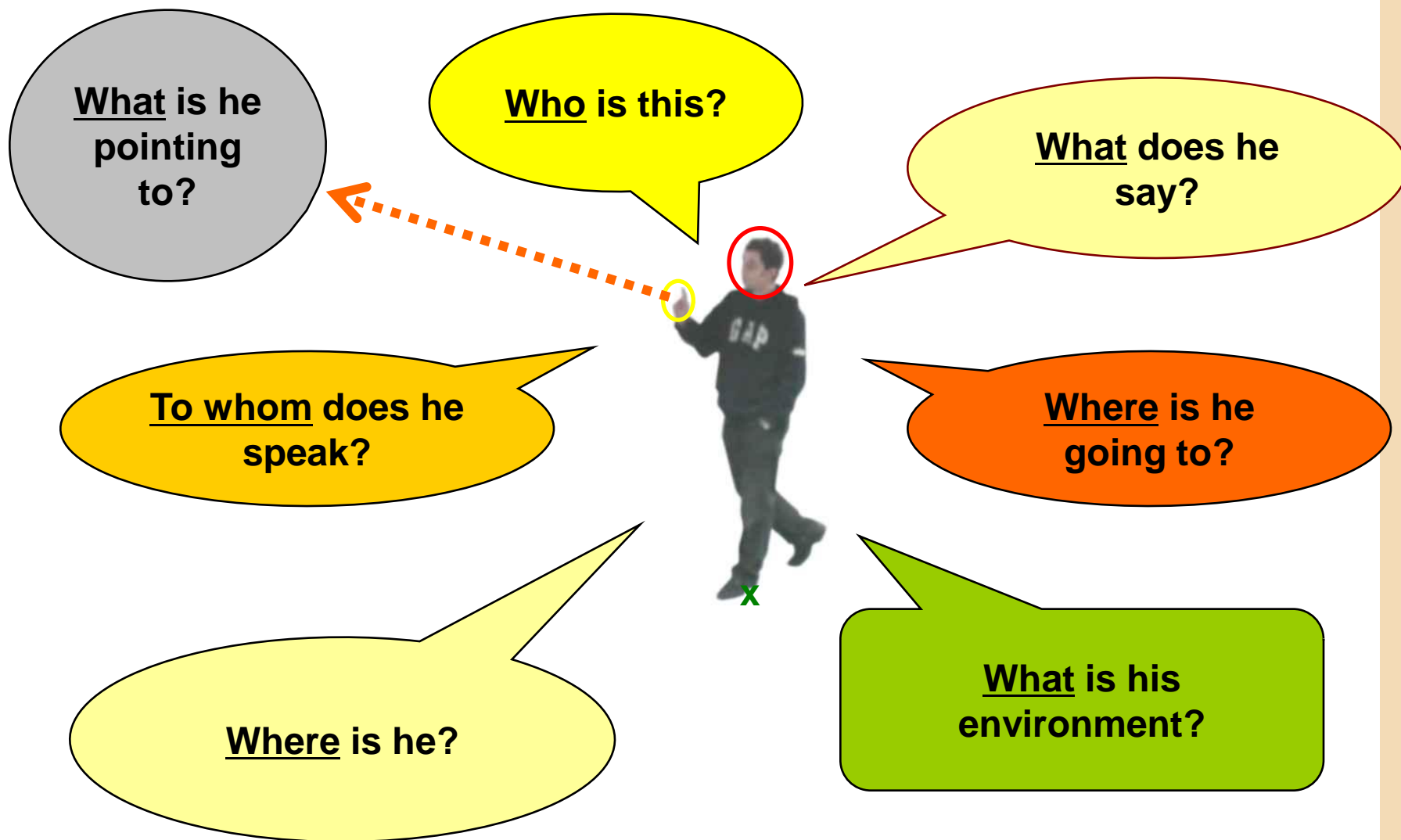




# Describing Human Activities









- **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

Localization



Identification



Tracking & Gesture

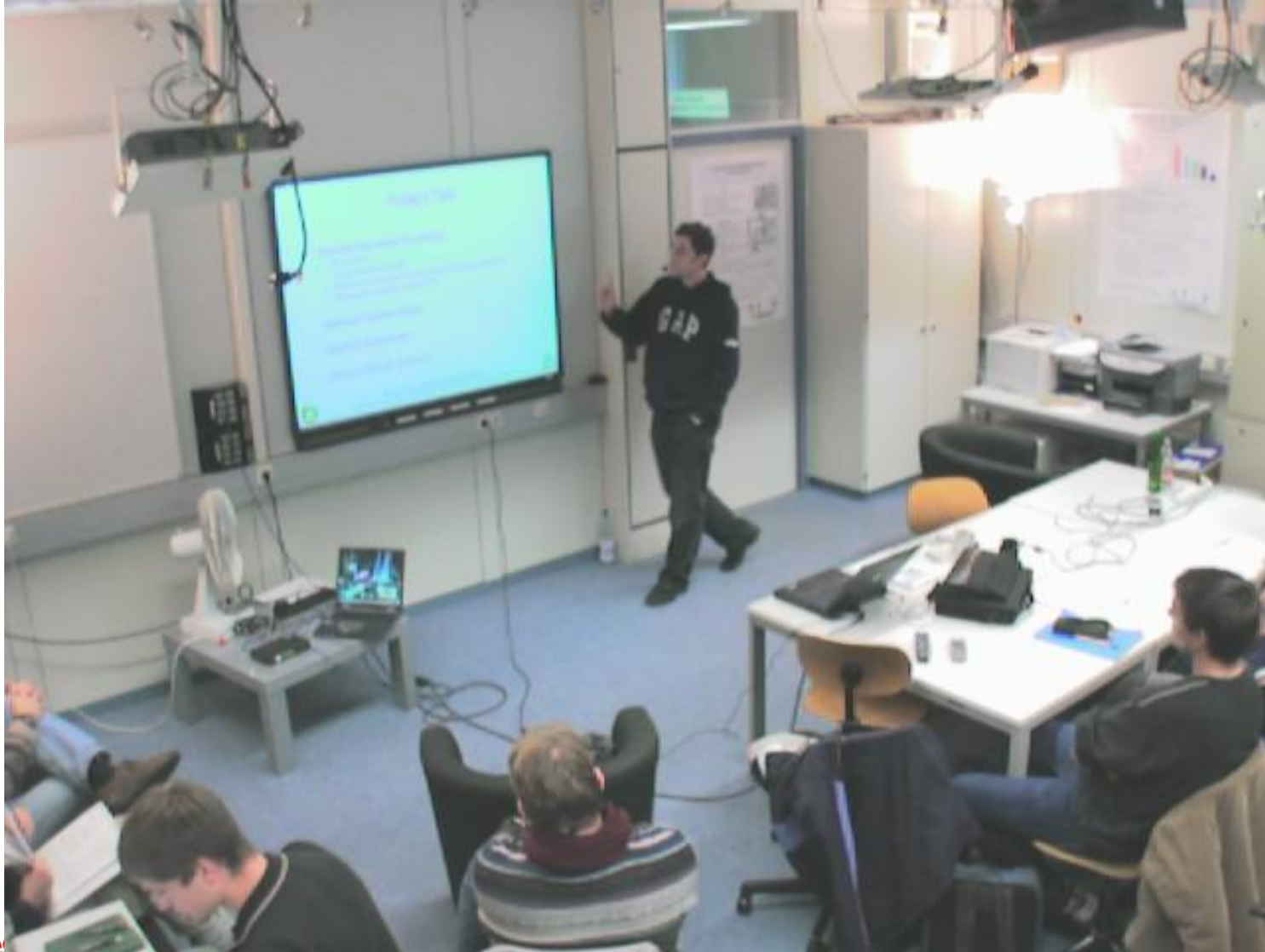


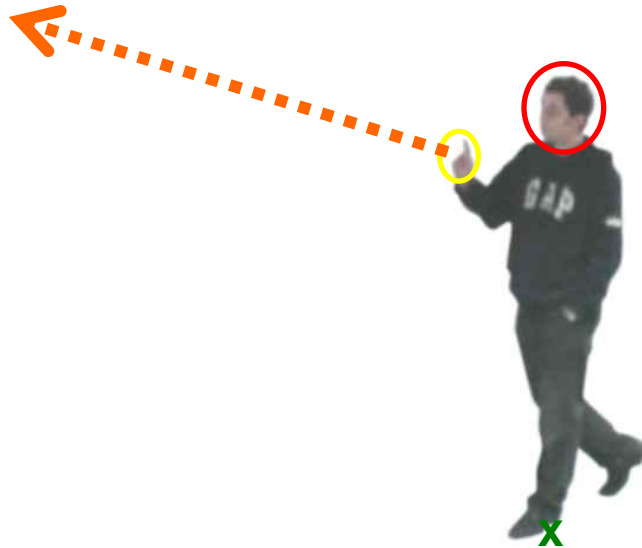
Focus of Attention



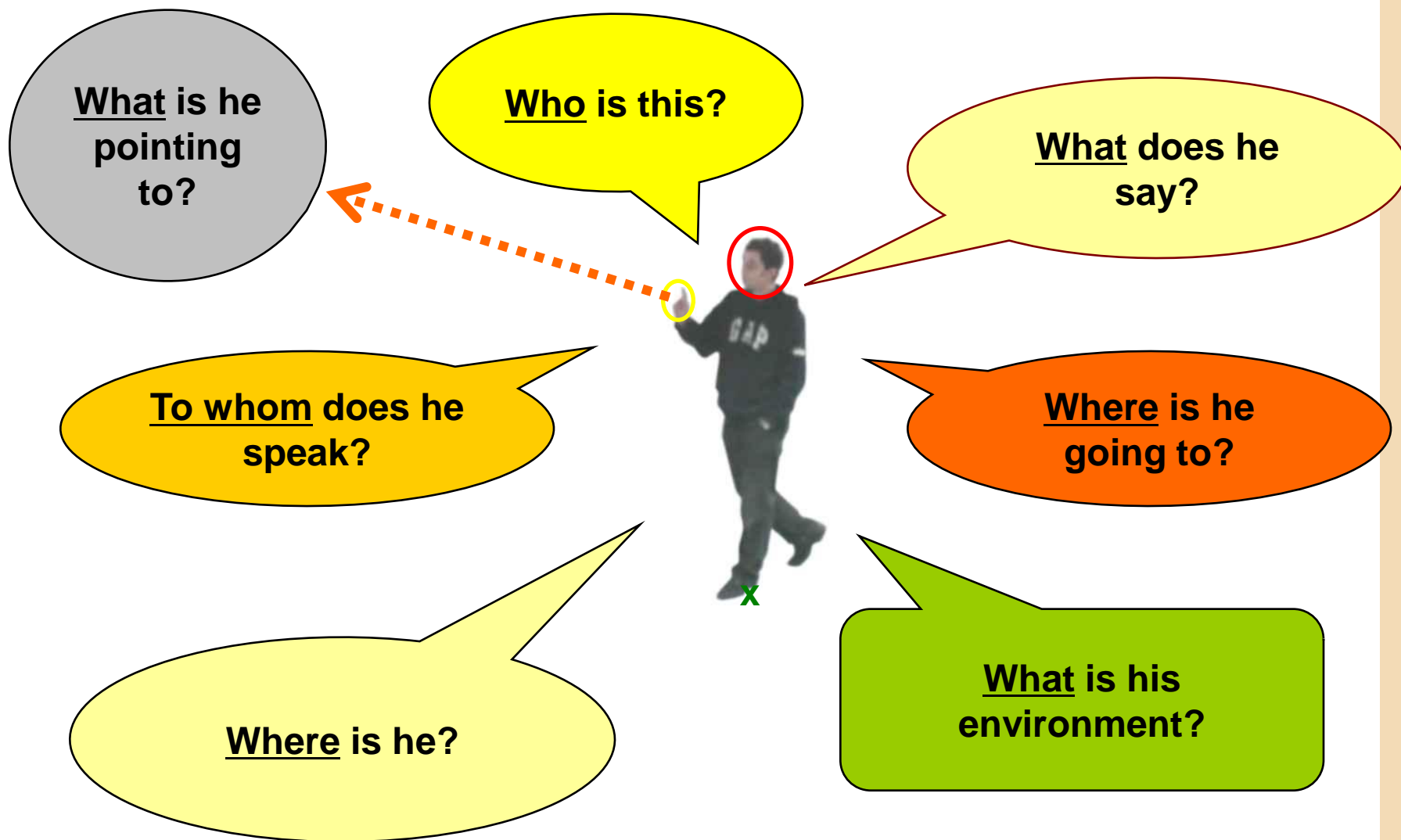


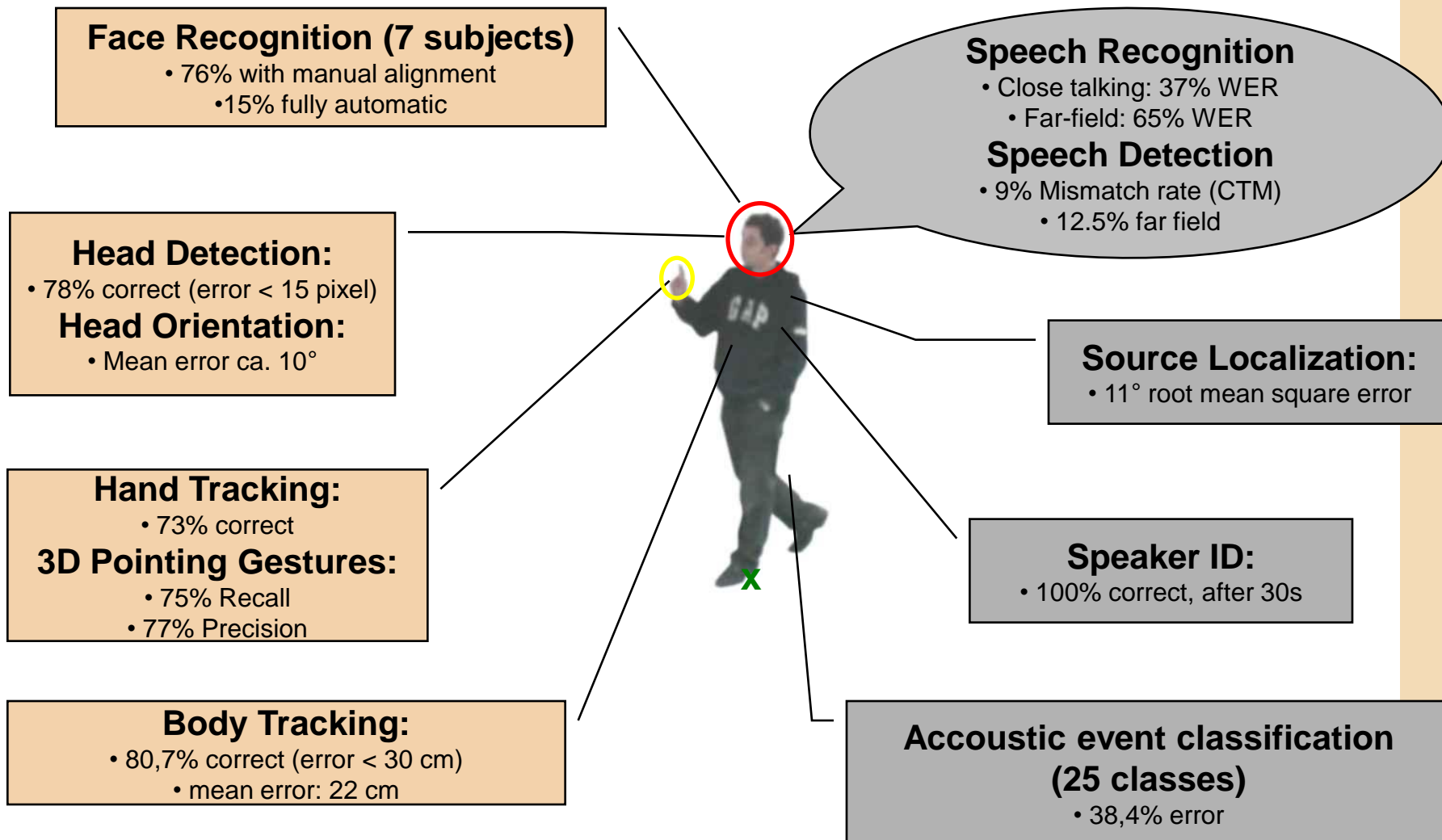
# Describing Human Activities











- 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 13 & 14, Southampton
    - Chair: Stiefelhagen (UKA)



- **Connector**
  - Connects people through the right device at the right moment
- **Meeting Browser**
  - Create Corporate Memory of Events
- **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
- ***Cross-Lingual Communication Services***
  - *Detect Language Need and Deliver Services Inobtrusively*
- ... (and more)



# Phone Calls During Meetings



# Phone Calls During Meetings



# Memory Jog

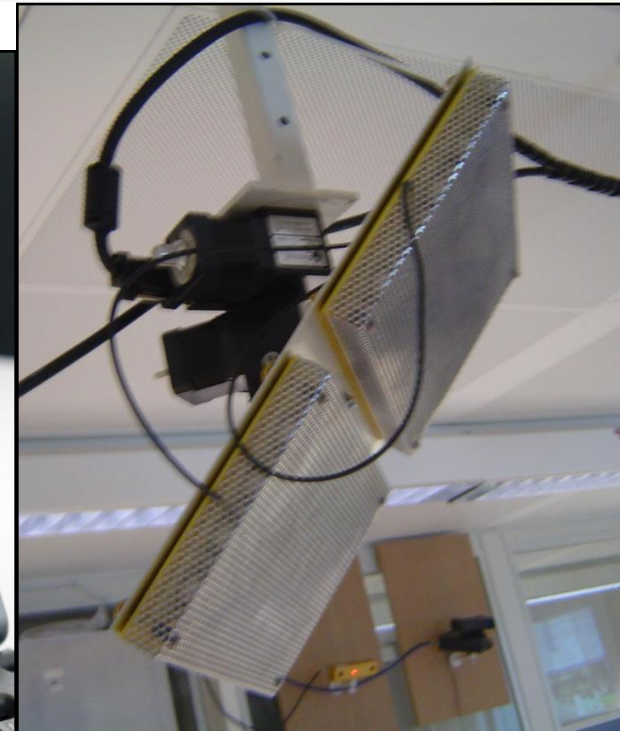
....What was his name? ...Where did I meet him? ...What did we discuss last time?





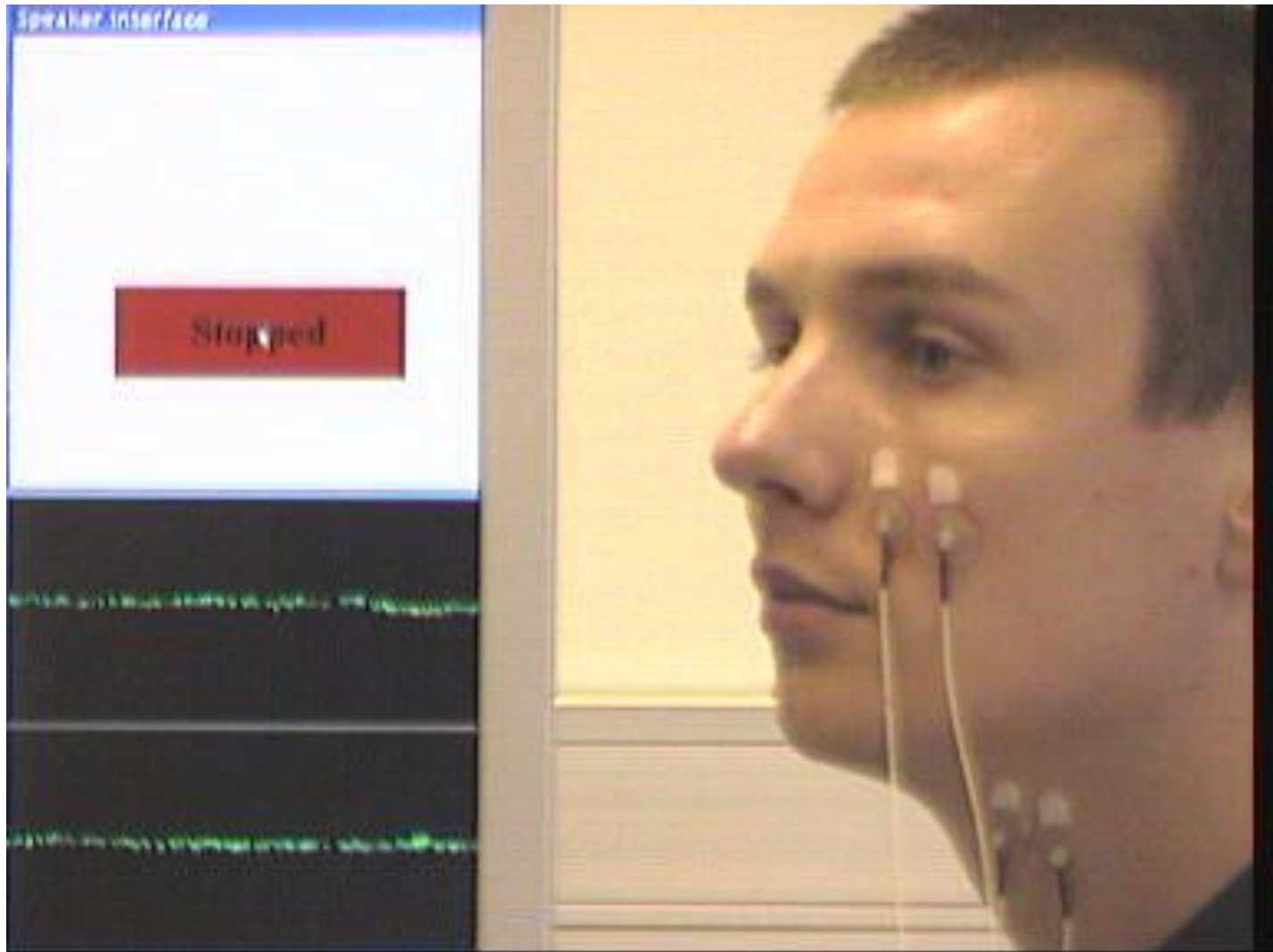
## Private and Public Information Delivery

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





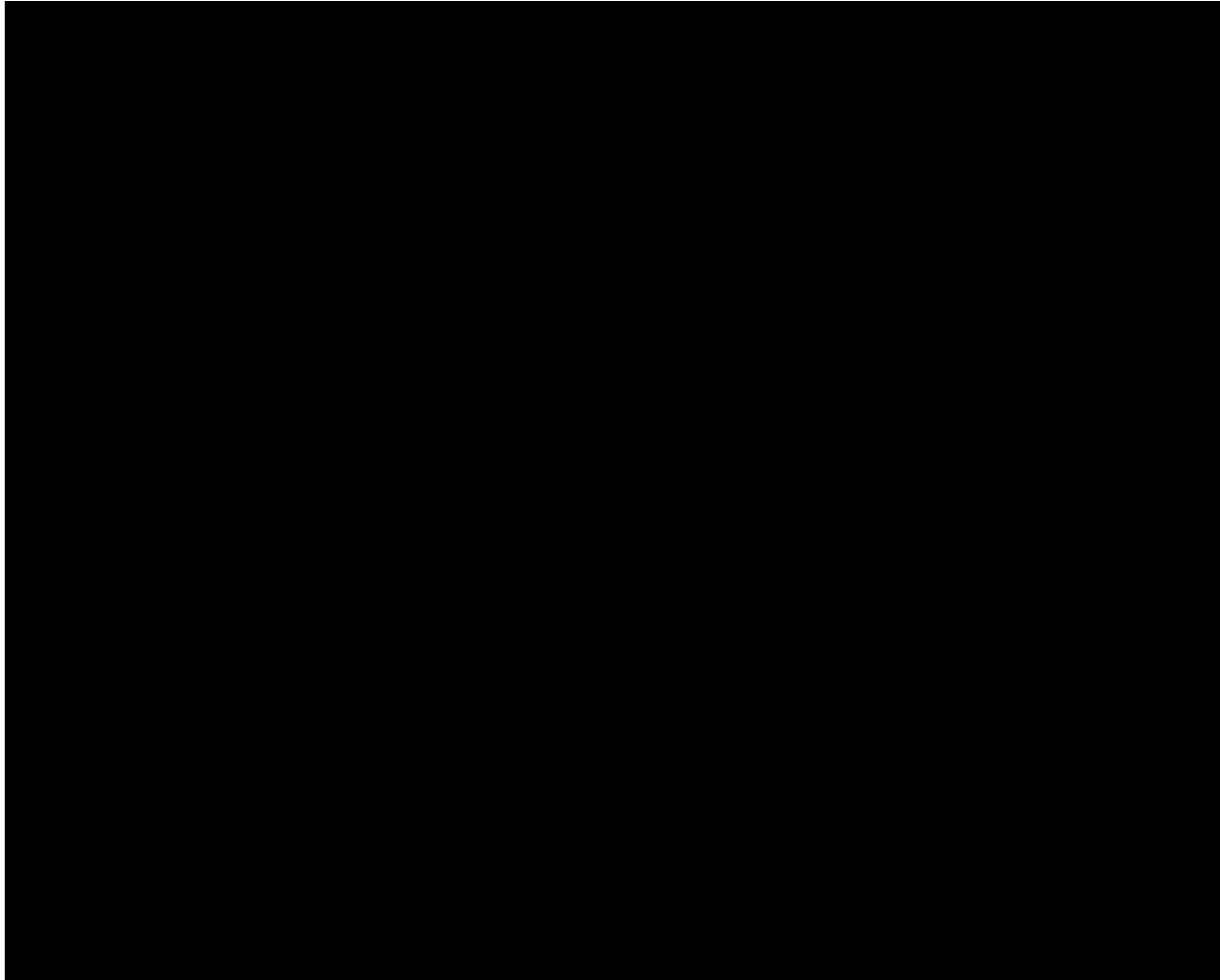
# Silent Speech based on EMG Signals



- 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	<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	



# *Human-Computer-Human Interaction: Challenges and Lessons Learned*



....what is he saying?

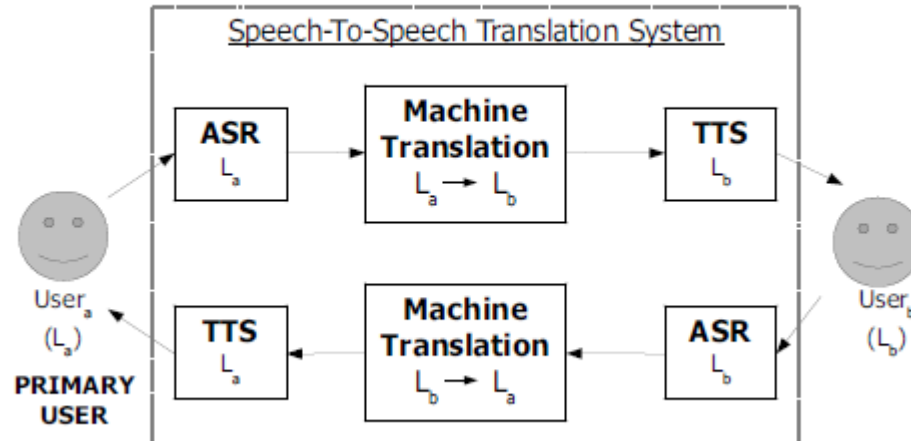


你们的评估准则是什么

- Dilemma:
  - Living in the Global Village
    - Globalization, Global Markets
    - Increased Exchange and Communication
    - European/International Integration
  - Cultural Diversity:
    - Beauty, Identity, Language, Culture, Customs
    - Pride and Individualism
    - Language Ability
  - Challenge:
    - Providing Access to Global Markets and Opportunities  
↔ Maintaining Cultural Diversity/Individuality

# Machine Interpretation

- Is Interpretation by Machine Possible?
  - Yes, and Performance will Continue to Improve
- Is it Replacing Human Interpreters?
  - No! Machine Translation Quality still Worse  
Lacks Human Judgement and Intuition
  - But: Human vs. Machine is Usually not the Choice we have!  
Commonly, it is No Communication or Poor English
  - Language Barriers are Pervasive and a Broad Social Challenge
- The Vision:
  - Multi-Lingual Understanding & Integration for All
  - Europe must Maintain & Nurture its own Diversity and Heritage
  - Europe must Provide for its own Language Support
  - We need to Embrace and Integrate **Both Human and Machine** Support



## To Build a Speech Translator for a New Language

- 6 Component-Engines: Automatic Speech Recognition, Machine Translation, and Text-to-Speech Synthesis
- Each is in Principle Language Independent, but Requires Language Dependent Parameters/Models
- Models are Automatically Trained but Require Large Corpora
- Certain Language Dependent Peculiarities Exist

## Progression of Technologies:

- Domain Limited, Clear Speaking Style (late 80's-91)
  - Janus (first European&US speech-to-speech system)
  - ATT, NEC, ATR
- Domain Limited, Spontaneous ('91-'00)
  - Janus II/III (work on 20 languages),  
Verbmobil, Nespole, Enthusiast,  
C-STAR, ATR, ETRI, NLPR,...
- Fieldable, Maintainable, Spontaneous
  - Transtac, Babylon, Phraselator, Jibbigo, U-STAR
- Domain Unlimited Speech Translation
  - Parliamentary Speeches (TC-STAR)
  - Broadcast News (GALE)
  - Lectures, Seminars (InterACT, STAR-DUST, TC-STAR)



# *Domain Limited Consecutive Translation* Technologies for Cross-Lingual Dialog



JIBBIGO



Hello. Nice to meet you!  
(Hello, nice to meet you.)



Hola, encantado de conocerle.





# Humanitarian Deployment



View Controls Store Advanced Help

iTunes

iPhone sync is complete.  
OK to disconnect.

Search: jibbigio

App Store > Travel > Jibbigio Speech Translator English Spanish

# Jibbigio LLC

**Jibbigio Speech Translator English Spanish**

Category: Travel  
Released Oct 21, 2009  
Seller: Jibbigio LLC  
© 2009 Mobile Technologies LLC  
Version: 1.0  
158 MB


**\$27.99** [BUY APP](#)

**Rated 4+**

## APPLICATION DESCRIPTION

Jibbigio is a bi-directional, natural speech-to-speech translation app that lets you converse with a speaker of another language by naturally spoken sentences.

Jibbigio is not a dictionary and not a phrase book, but a speech translator: You simply speak a sentence in English or Spanish into Jibbigio, and it speaks the sentence aloud in the other language, much like a personal human interpreter would. Jibbigio also shows the recognition and translation in English and Spanish as text on the app screen, so you can be sure your translation is accurate to what you said.



The screenshot shows the Jibbigio app interface. At the top is the JIBBIGIO logo and a navigation bar with icons for speech, text, and settings. The main display area shows the English text "It's nice to meet you. (Nice to meet you.)" next to a US flag. Below this is a red circular button and a play button. At the bottom, the Spanish translation "Encantado de conocerle." is displayed.

## Jibbigio:

- Real-Time Translation
- Spanish - English
- Japanese, Chinese, Arabic, ...
- 40,000 Words
- No Server Necess







# Jibbigo on Apple Commercials



- iTunes & Android App Stores:
  - English, Spanish, French, German, Japanese, Chinese, Korean, Filipino, Iraqi, Thai, Pashto, Dari
- Cost:
  - **Free** Jibbigo Online Translator
  - Off-Line: Freedom from Network
- Outside of App Store:
  - Other Languages in Preparation
  - Enterprise Versions for Special Applications

## Supported Devices

iPhone-OS Compatible Devices  
(First systems commercially deployed in Oct 2009)



iPhone 3GS



iPod Touch



iPad

Android-OS Compatible Devices



Nexus One



Motorola Droid



HTC Droid  
Incredible



Samsung i9000  
Galaxy

## New year. New resolutions. New apps.

**Weight Watchers**  
Free • Resolved to stay fit? Now you can get healthy tips and free recipes from one of the top names in weight loss – all year long. And if you're a Weight Watchers member, you can also track and calculate your points.

**Dragon Dictation**  
Free • Send messages easier and faster, without typing a single word. Dragon Dictation recognizes and transcribes what you say, and lets you paste it directly into an email or text.

**Yelp Monocle**  
Free • Find a new place to eat in a whole new way. Just point your iPhone camera down any street and Yelp Monocle will overlay restaurant ratings, pricing info and even how many miles away they are, right on your screen.

**Ustream**  
Free • Use your iPhone to stream live video to your family and friends, as you shoot it. Through your Ustream channel, you can also share your videos on Twitter, Facebook and YouTube.

**e-Secure**  
Free • Manage your Protection One home alarm system wirelessly from your iPhone. Turn your alarm on or off, receive text alerts, and more – even when you're away from the house.

**RedEye**  
Free • Too many remotes cluttering your home? Turn your iPhone into a handy universal remote that can control your DVR, DVD player and stereo through your RedEye system.

**Green Outlet**  
\$2.99 • Run a more eco-friendly home in 2010. Just enter in what appliances you use and how often each day and Green Outlet will help you identify where you can save energy – and money.

**Master Control**  
Free • Looking to improve your soccer skills this year? This elite Nike training app features instructional videos from FC Barcelona coaches. Get tips on ball control, see drills the pros use and even learn some of their signature moves.

**Jibbigo**  
\$24.99 • Carry a personal Spanish translator in your pocket. Just tap to record any phrase and Jibbigo will play it back in Spanish – or any Spanish phrase back into English.

**Viper SmartStart**  
Free • Turn your iPhone into the ultimate keyless remote. With SmartStart installed in your vehicle, all you need is your iPhone to lock, unlock, or even start your car from just about anywhere.

**ecobee**  
Free • Control your home's ecobee Smart Thermostat even when you're away. Adjust the temperature right from your iPhone so you can make sure you conserve energy costs but still come back to a cozy house.

**New York Subway**  
\$0.99 • There's an amazing new way to find the nearest New York subway station. Just aim your iPhone's camera in any direction and NY Subway will superimpose station information and distances right on your screen.



Jibbigo featured in the Economist magazine 2010

It's a new year with new amazing apps. From live streaming video to augmented reality, there are over 100,000 apps for just about anything. Only on the iPhone and the nation's fastest 3G network.





- How it is Done Now:
  - Human Interpreters
  - Charts, Dictionaries
- Limitations/Problems:
  - Limited Supply!!
  - Fidelity/Trust/Security
  - Number of Languages







# *Unlimited Domain Simultaneous* Speech Translation Technologies

## Domain Unlimited Translators for:

- TV/Radio Broadcast Translation
- Translation of Lectures and Speeches
- Parliamentary Speeches (UN, EU,...)
- Telephone Conversations
- Meeting Translation



你们的评估准则是什么

# Language Barriers





**Studentin aus Tunesien**

**die Vorlesungen waren unverständlich**

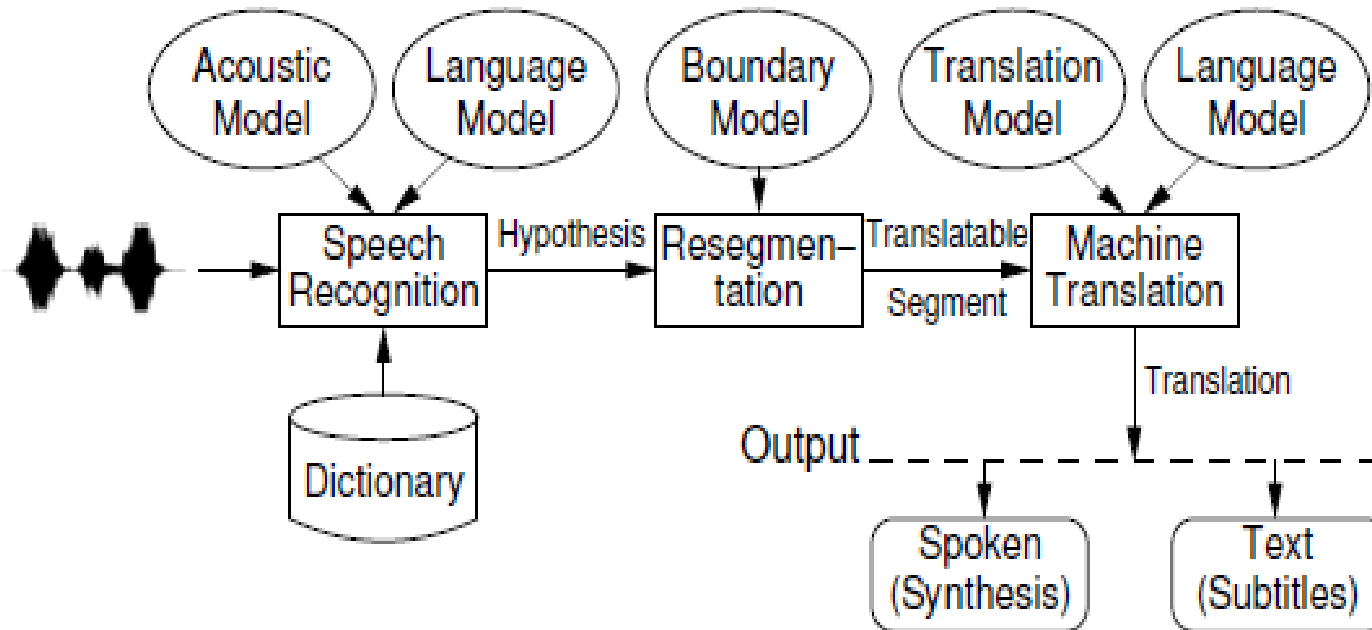
# Translation of Speeches



MR PRESIDENT

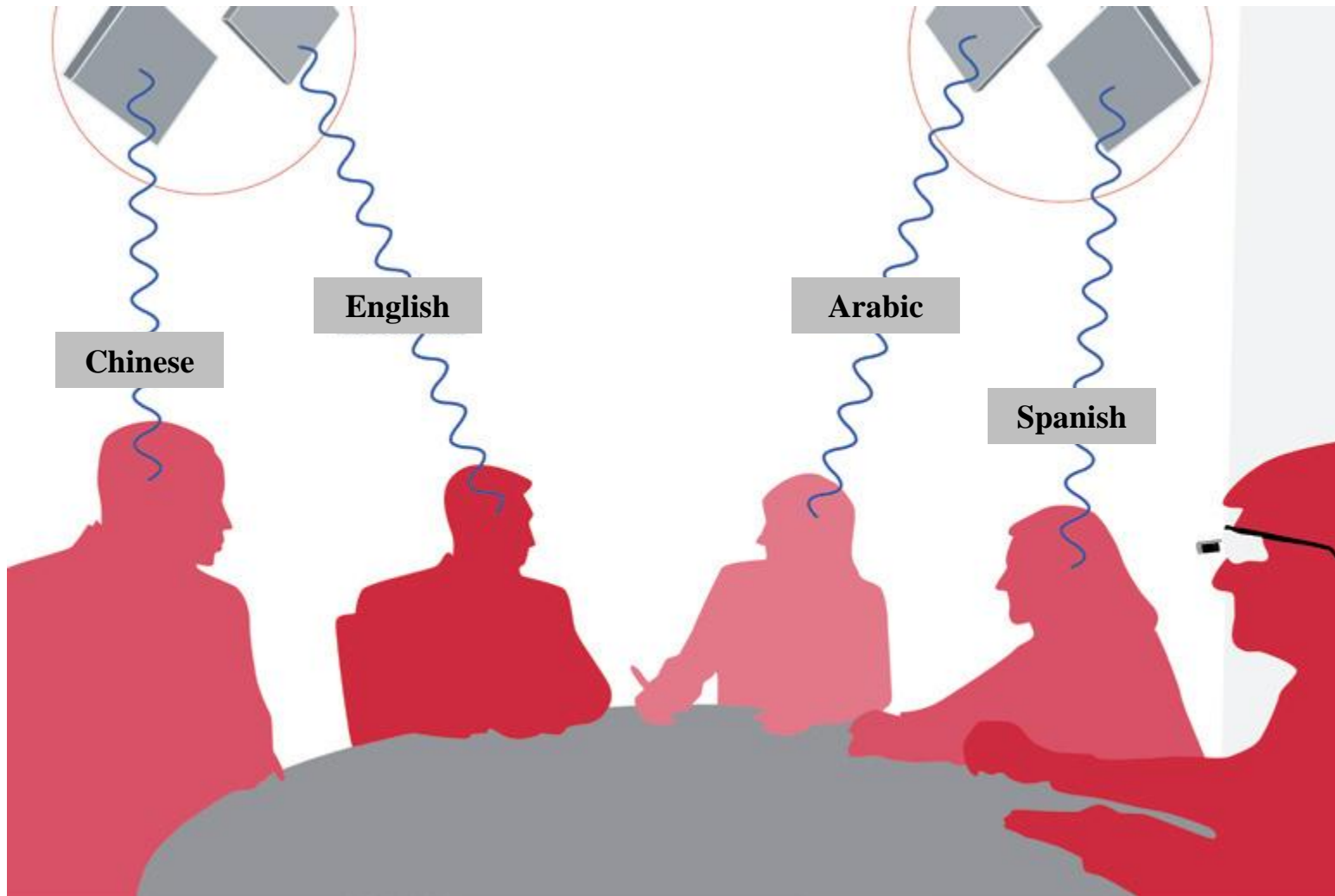
señor presidente







# Meeting of the Future



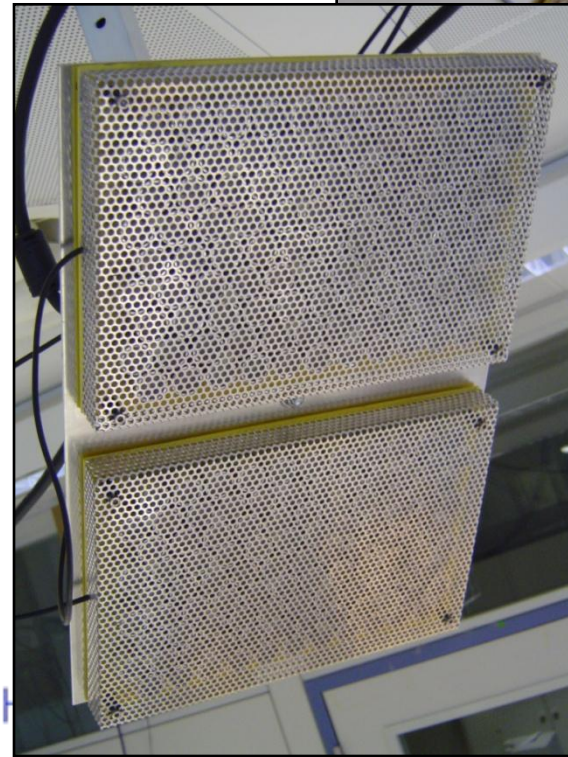
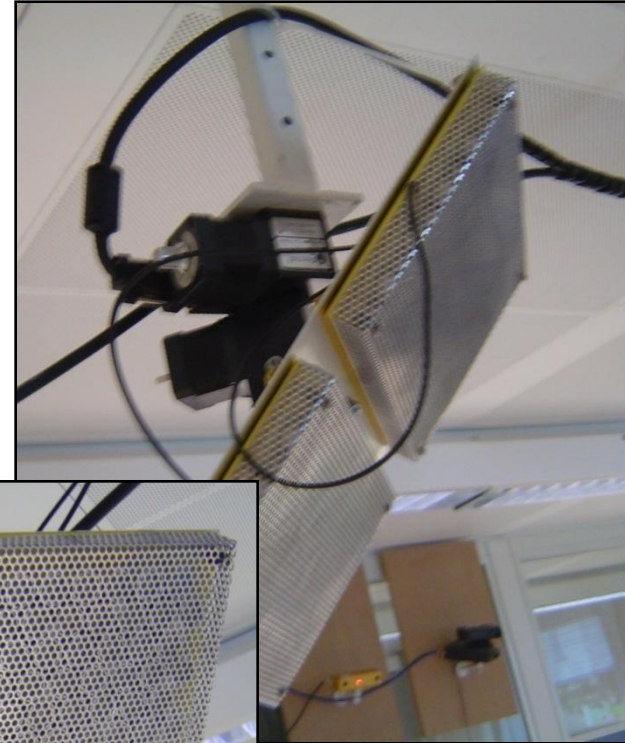
### 3. ---Seeing Personal Translations



- Technology: Heads-up Display Goggles
  - Create Translation Goggles
  - Run Real-Time Simultaneous Translation of Speech
  - Text is Projected into Field of View of Listener
  - Translations are Seen as Text Captions Under Speaker
  - Output: Spanish, German,...

# Hearing Personal Translations

- Technology: Targeted Audio
  - Research under EC Project CHIL  
(Build Inobtrusive Computer Services)
  - Project Partner, Daimler-Chrysler
  - Array of Ultra-Sound Speakers
- Result: Narrow Sound Beam
  - Audible by one Individual Only
  - Others not Disturbed
  - Multiple Arrays Could  
Provide Multiple Languages
  - Steerable
  - Recognize/Track Individual Listener  
and Keep Language Beam on Target





- English->Spanish Lectures  
First Research-Prototype CMU 2005
- German Lectures, KIT '10
- Cloud Based Services, MobileTech, '10
- Transition to a Lecture Service
  - First Beginning: 4 Lectures, 2012
  - EU-BRIDGE





EU★BRIDGE



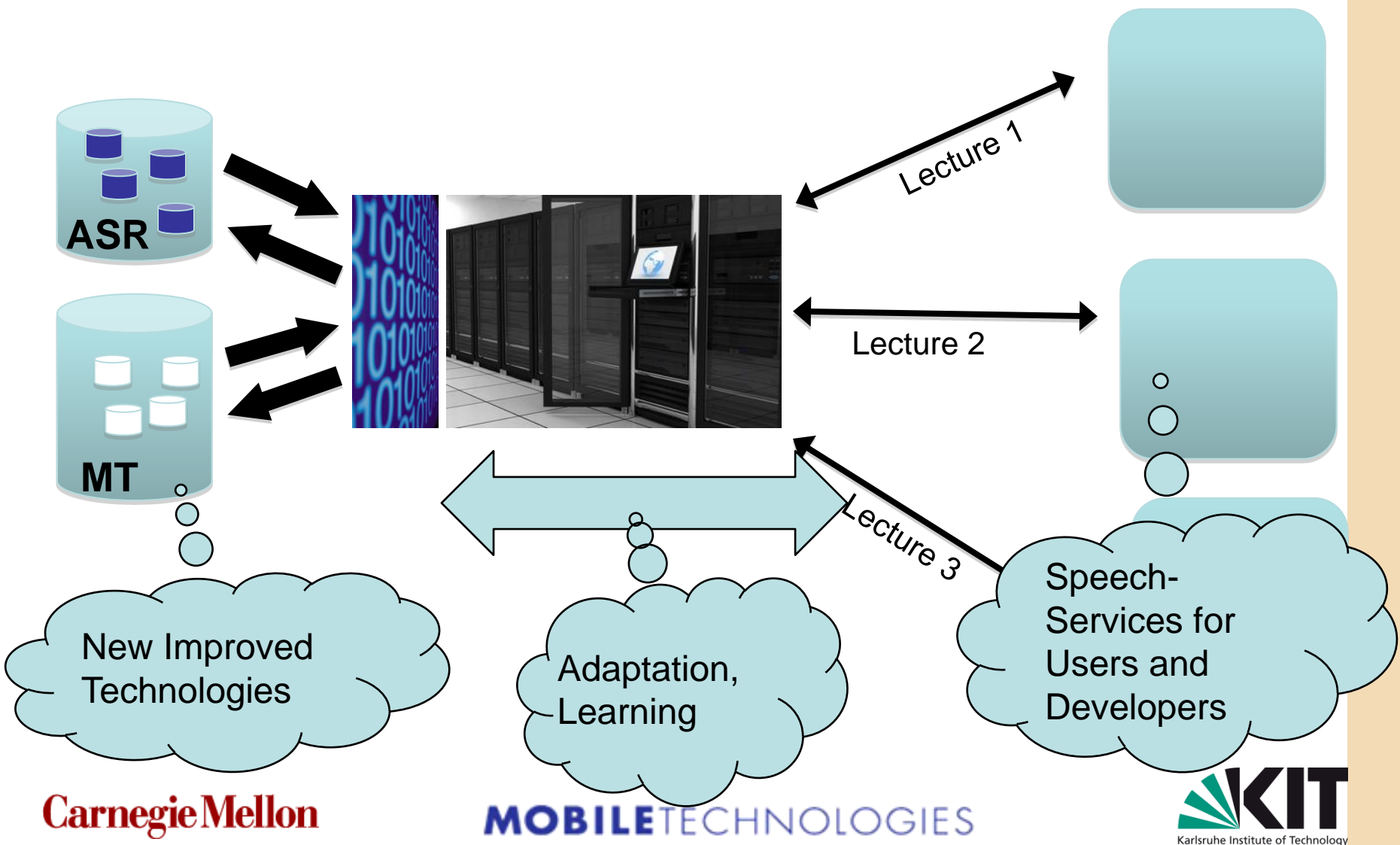
MOBILETECHNOLOGIES



**Components**

**Services**

**Events**







interACT

# Launch, June 11 2012

## Mehr als nur Bahnhof verstehen - Weltweit erster Vorlesungsübersetzer

Karlsruhe. In einer deutschen Vorlesung sitzen und wegen der Sprachbarriere nur Bahnhof verstehen - sowas könnte für ausländische Studenten bald Vergangenheit sein. Der weltweit erste automatische Vorlesungsübersetzer ermöglicht Studierenden künftig, dem Vortrag



Carnegie

von Dozenten auf Englisch zu folgen - schriftlich übersetzt in Echtzeit. Am Montag

Prof. Alex Waibel

MOBILE TECHNOLOGIES


**KIT**  
Karlsruhe Institute of Technology





- Translation of Power Point Slides
- Presentation by Sub-Titles

Minimalwert

 Fakultät für Informatik – Institut für Anthropomatik  
Prof. Dr. U. Hanebeck, Dr.-Ing. T. Asfour

6-8


## 3.1 Formale Grundlagen

Zur Untersuchung und Beschreibung der Eigenschaften und des Verhaltens von logischen Funktionen ist die **Boolesche Algebra** hervorragend geeignet.

Entwickelt wurde sie von dem Mathematiker **George Boole** (1815 –1864) als Algebra der Logik.

It was developed by the mathematician George

3.1 formal basics  
For the study and description of logic functions and the behavior of systems, which is excellently suited.  
It was developed by the mathematician George Boole (1815 1864) as algebra of logic.



- Transcripts useful to Search for Content
  - Slides, and Lectures in the Cloud
  - Search in Lectures and Foils by Way of Search Terms

## LECTURE STORAGE

[Home](#) [Help](#) [Sign in](#)

### List of Lectures

Time	Lecture	Presenter
2012-05-24 14:24:29.544848	<a href="#">Kognitive Systeme - ro24052012</a>	KIT-Dillmann
2012-04-23 14:49:12	<a href="#">KogSys 23.04.2012</a>	KogSys DSP2 Waibel
2012-05-10 15:30:26	<a href="#">Rechner Organisation - RO100512</a>	Asfour
2012-04-17 18:14:07	<a href="#">Rechnerorganisation_17_04_2012</a>	dillmann and asfour are talking
2012-05-07 15:05:01	<a href="#">Kognitive Systeme</a>	Mohr
2012-04-23 15:31:39	<a href="#">Kogsys 23.04.12</a>	NULL
2012-05-09 13:02:38	<a href="#">Kognitive Systeme - KogSys 090512</a>	Waibel
2012-05-23 13:07:13	<a href="#">Kognitive Systeme - kogsys230512</a>	NULL
2012-05-23 13:07:51	<a href="#">Kognitive Systeme - Dozent: Prof. Waibel // Kognitive Systeme 23 05 2012</a>	NULL
2012-05-31 11:32:36.774097	<a href="#">General - Jan MT 310512</a>	NULL
2012-05-08 17:19:57	<a href="#">Kognitive Systeme - MT zweiter Teil</a>	Waibel
2012-04-18 13:00:43	<a href="#">Kognitive Systeme 18.04.2012</a>	NULL
2012-06-04 14:14:30.634523	<a href="#">Kognitive Systeme - ASR 3 und aktuelle Forschung</a>	KIT-Waibel2

[previous](#) [next](#)



## Simultaneous Translation of Lectures

- Continuous Monologue
  - Broadcast News, Speeches, Lectures
- Speaking-Style
  - Fast, spontaneous, fragmentary, and no punctuation!!
  - Noise, Caughing, Singing (!)
- Vocabulary
  - Much larger, Special Vocabularies
- Speed, Realtime
- Service-Infrastructure
  - Many parallel lectures;
  - Automatic, robust assignment of compute power

- MT in **German** Lectures is particularly hard. Why?
- Peculiarities of German:
  - Wordorder:  
*Ich möchte* mich zu der Konferenz über Maschinelle Übersetzung *anmelden*  
→ *I want to register* to the conference on Machine Translation
  - Compounds:  
*Worterkennungsfehlerrate*  
→ Word Recognition Error Rate
  - Inflections and Agreement:  
Zu *der* nächst*en* wichtig*en* interessan*ten* Vorlesung



- Technical Terms  
normally not in ‘normal’ vocabularies
  - Cepstral-Koeffizienten
  - Wälzlagerungen → Roller Bearings
  - Unterraum → Subspace
- Technical Terms  
with special Meanings
  - Klausur → Final Exam (not Retreat)
  - Vorzeichen → Sign (not Omen)
- Formulas:
  - Eff von  $I_x$  →  $f(x)$

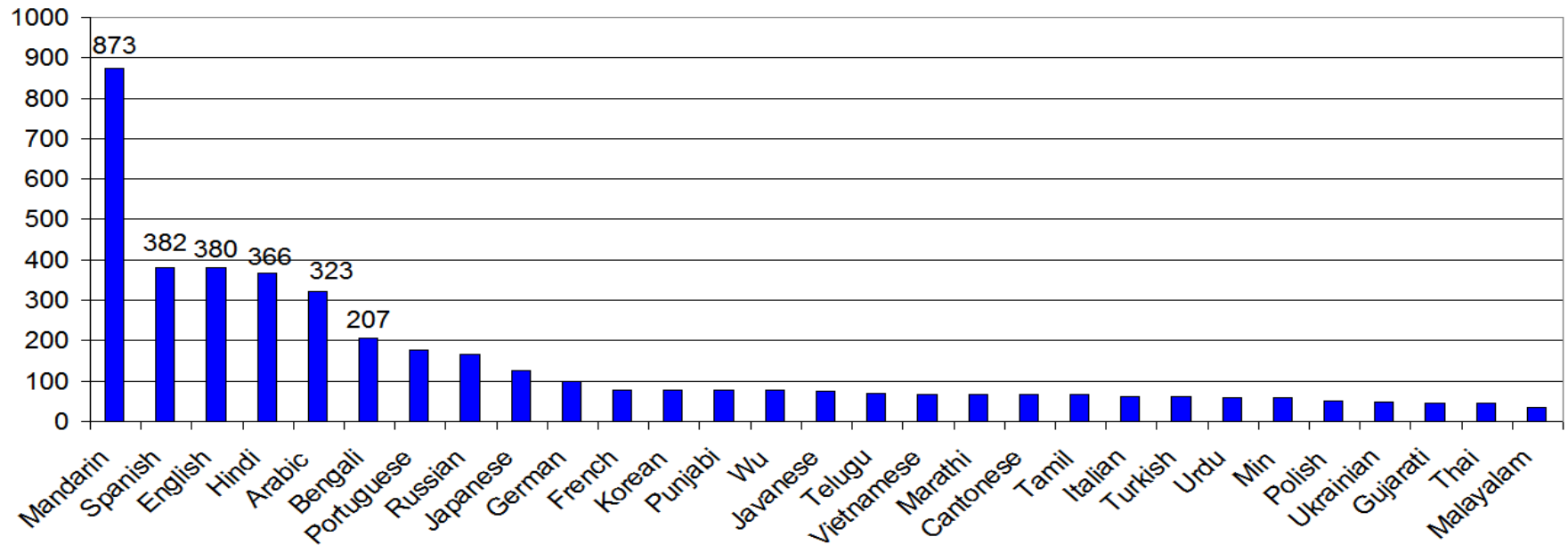


- Foreign Words in German Language
  - Computer Science, English Expressions
  - Political Speeches, Latin Proverbs
- Accent
  - “Würfelkalkül” (Asfour)
- Foreign Words in German Language
  - “Cloud”, “iPhone”, “iPad”, “Laser”
- Inflections & Declinations of these Words
  - Web-ge-casted, down-ge-loaded
- Formation of Compounds:
  - Cloudbasierter Webcastzugriff

# The Long Tail of Language

- Languages:
  - Only a Few Languages are Currently Addressed (<10)
  - Development of Technology Takes Long & Is Expensive
  - Cost more than 1M \$ and DevTime more than 1 Year per Language
  - 6, 000 languages, 36 M potential language pairs, Plus Dialects
  - Technology is Always a Step (or Two!) Behind Deployment

Languages by Million Native Speakers



## Communication between the people of the world

- It is all about Communication
- Multimodality
- Multilinguality