Future Spoken Dialog Systems: Multimodal, Multilingual, Multiparty, Multitask

Wolfgang Wahlster

German Research Center for Artificial Intelligence
DFKI
Saarbrücken, Kaiserslautern, Bremen, Berlin
e-mail: wahlster@dfki.de
WWW:http://www.dfki.de/~wahlster
13 Trends for Spoken Dialog Systems

1. From **Unimodal** to **Multimodal Dialogs**
2. From **Monolingual** to **Multilingual Systems**
3. From **Single Task** to **Multitask Dialogs**
4. From **Dyadic Dialogs** to **Multiparty Conversations**
13 Trends for Spoken Dialog Systems

5. From Close Speaking to Microphone Arrays for Distant Speaking

6. From Cooperative Speech to Spontaneous Speech

7. From Stationary to Mobile Spoken Dialog Systems

8. From Hosted Voice Portals to Cloud-based Speech Solutions
13 Trends for Spoken Dialog Systems

9. From **Client-Server Spoken Dialog Systems** to **Embedded Systems**

10. From **Database Transactions** to **Problem Solving Dialogs**

11. From **Access to the Web of Information** to **the Internet of Services**

12. From **Generic** to **Personalized Voice User Interfaces**

13. From **Human-Machine** to **Human-Environment-Interaction**
Multimodal Dialog Systems

Graphical User interfaces

Spoken Dialogue

Video Input

Physical Action

Gestural Interaction

Haptic Interaction
Overlapped, non-native accented and spontaneous Speech

Multiparty Speaker Diarization and Tracking
Who spoke When, with Whom, Where about What?
Just-in-time Access to Relevant Documents or Fragments of Past Recorded Meetings

Killer App for Call Centers: Just-in-Time Answer Retrieval during the Conversation between an Agent and a Customer by Parallel Speech Understanding
SuVi: The Generation of Meeting Summaries as Story Boards in Cartoon Style

Still pictures extracted from video capture, cartoon-style speech balloons for spoken dialog contributions and text boxes for the results of topic detection
SuVi: The Generation of Meeting Summaries as Story Boards in Cartoon Style

Still pictures extracted from video capture, cartoon-style speech balloons for spoken dialog contributions and text boxes for the results of topic detection
DICIT (Distant-talking Interfaces for Control of Interactive TV) EC project

- Coordinated by FBK
- Goal: voice control of TV and related devices
- Robustness against noise and audio interferences
- Smart processing also including a real-time multi-speaker localization
- Understanding of voice input queries
- Multimodal spoken dialog management

For more details: http://dicit.fbk.eu
Multilingual Access to a Electronic Program Guide (EPG) with Distant Speech
Multiparty Dialog between Virtual & Human Football Experts: Discussing the UEFA EURO 2016 in France

Multilingual Virtual Moderator

N > 2 Virtual Experts

n > 2 Human Football Fans from Different EU Member States Speaking their Mother Tongues
Discussing the Best of European Football in Your Mother Tongue 2016

1. on your mobile with football fans from all over Europe

2. with spontaneous speech translation, diarization, simultaneous cross-lingual multimodal content linking

3. 24 languages of 24 European teams

4. quiz and game shows, defining your own teams, virtual coaching

multimodal, multilingual, multiparty, multitask
The Basic Architecture of the 4M EURO 2016 System

Ontologies for
- Football
- Dialog Model
- Games
- Videos

Narration and Quiz and Game Engine

ASR, NLP, Dialog
- Spoken Input Analysis
- Conversational Dialog Engine
- Gesture
- Emotion Engine
- Action Encoder
- Scheduler
- Simultaneous Translation
- Diarization
- Summarization

Character Modeling and Animation

NL Generation and Visualization

Platforms
3D Monitors
SmartPhone
Dashboard
WebPads

Player
- Emotional Speech Synthesis
- 3D Character Rendering
- Video Streaming
Who won the World Football Championship in 2006?

Italy

Personal guide for the FIFA world cup
Monolingual Multiparty Football Quiz and Game Show at DFKI
Multitask Games with Multimodal Dialogs
- Weight sensors and microphones in the car take measures / capture speech on the respective seats
- Values and speech features are broadcasted and received by personal (nomadic devices)
- Speaker models and weights are stored on personal device.
- Personal devices “decode” the position information and decide, which service is allowed to use it
Multiparty Conversation and Speaker Identification in the Car
There Are Many Open Problems for the Next 6 Years:

- Integrating **top-down context and dialog knowledge** into low-level speech recognition processes.

- Exploiting more knowledge about **human communication and translation strategies** including psycho- and neurolinguistic inspirations.

- Avoiding **expensive data collections and cognitively unrealistic training data for machine learning.**
10 Years after Verbmobil + 5 Years after SmartKom/SmartWeb

15 and 16 November 2010, Saarbrücken: 10 Years Verbmobil
Looking Back and Looking Ahead
Football Tournaments Create Emotions: Emotional Speech, Emotional Facial Expressions
Realistic Facial Expressions combined with Emotional Speech Synthesis

Jules: the Robotic Head by Hanson Robotics used by a Team at the University of Bristol
Thank you very much for your attention.