Academic and Industrial Views on the Evaluation of Spoken Dialogue Services

Sebastian Möller et al.

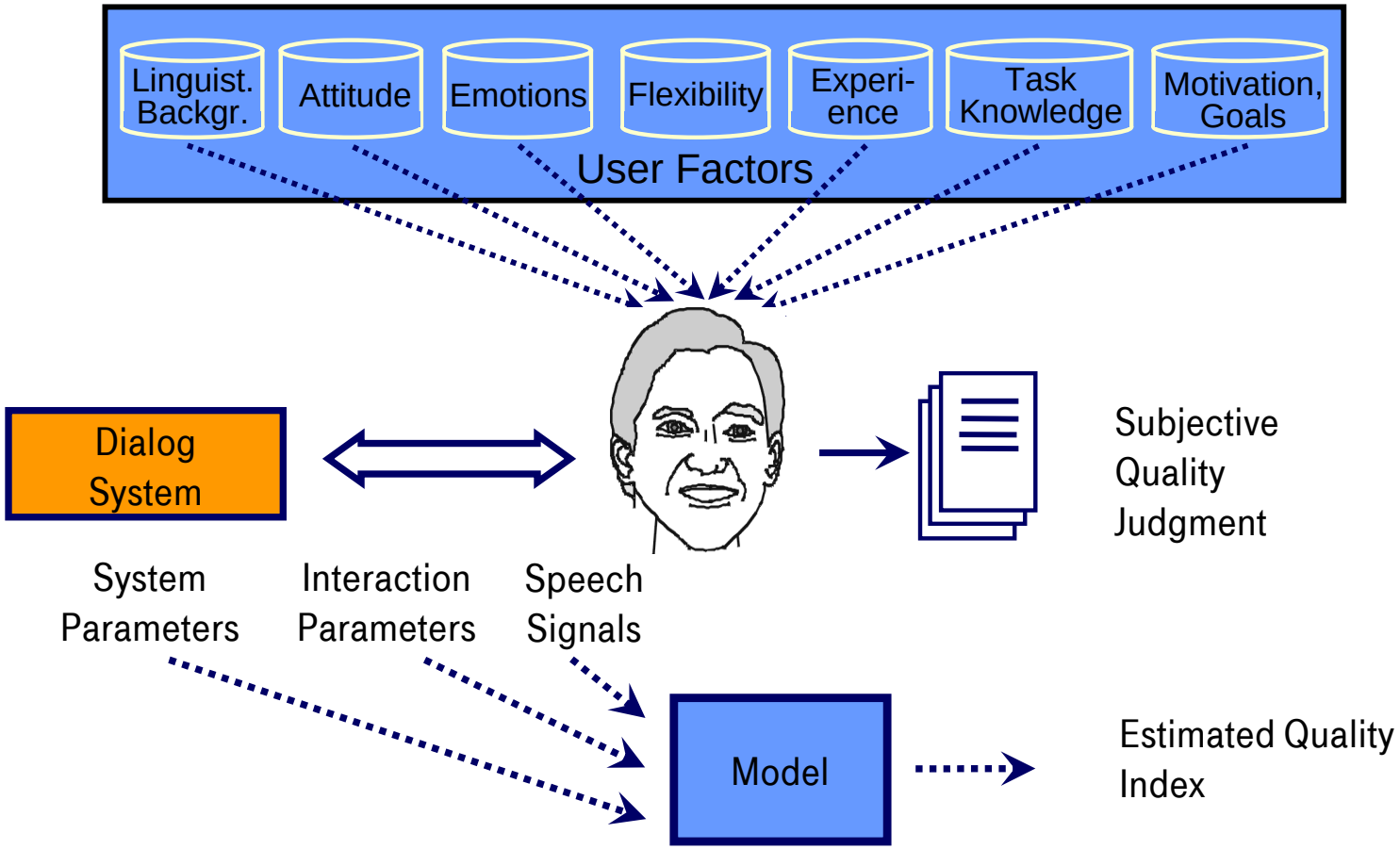
Quality and Usability Lab
Deutsche Telekom Laboratories, TU Berlin

sebastian.moeller@telekom.de



Motivation

Quality assessment and prediction of voice services.



Research questions: What are the perceptions of the users?
How can we avoid costly and time-consuming user tests?

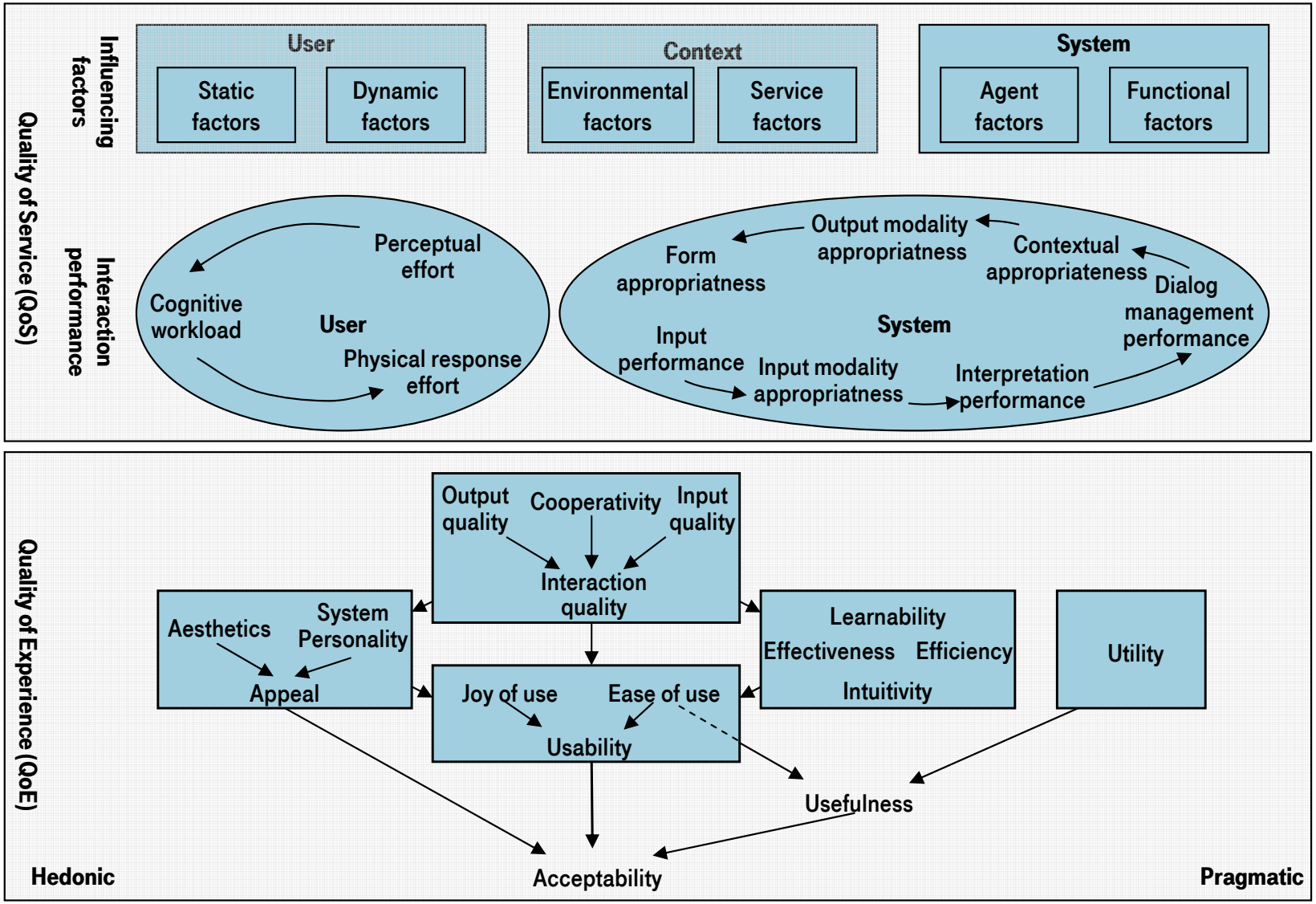
Overview

- **Quality and usability**
  - **Definitions**
  - **Assessment methods**
- Automatic prediction of voice service quality
- Summary and activities @ T-Labs

Quality and Usability
Definitions.

- **User‘s point-of-view:**
  - **Quality**: “Result of appraisal of the perceived composition of the service with respect to its desired composition.”
    (ITU-T Rec. P.851, 2003, following Jekosch, 2000, 2005)
  - **Usability**: “The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.”
    (ISO 9241-11, 1998)
- **System developer‘s point-of-view:**
  - **Aim**: Measure and optimize the characteristics of the system, the usage environment and the user with respect to their impact on perceived quality, usability and acceptability
  - **Performance**: “The ability of a unit to provide the function it has been designed for.” (Möller, 2005)

Quality and Usability
Taxonomy of quality aspects.

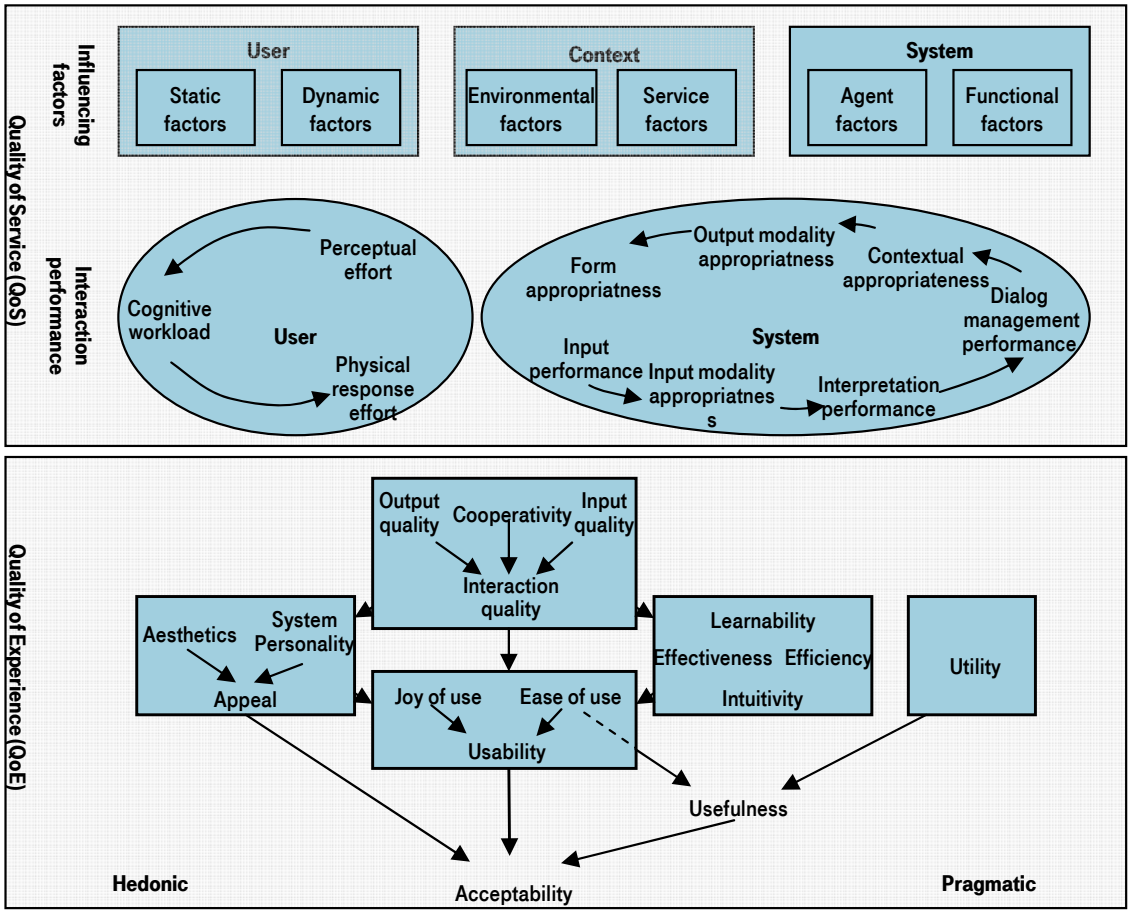


(Möller et al., 2009)

Quality and Usability
Assessment methods of voice services.

Standardized methods:

- **Test interactions** of potential users with the system
- **Questionnaires** covering different quality aspects, e.g. ITU-T Rec. P.851, 2003)
- Quantitative description of user and system behavior via **interaction parameters**, see ITU-T P-Series Suppl. 24 (2005)
- Subjective evaluation of **TTS quality**, see ITU-T Rec. P.85 (1994)
- **Usability evaluation methods**, e.g. ISO 9241-110 (2006)



(Möller et al., 2009)

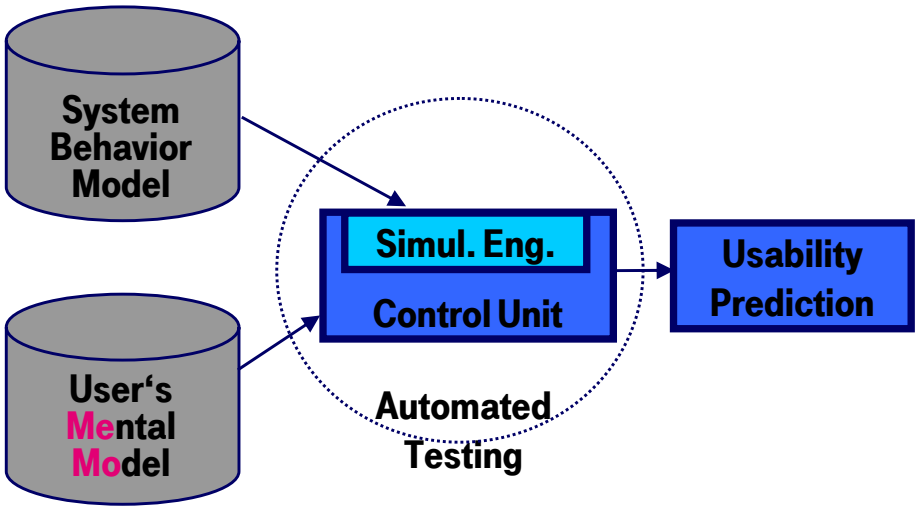
Overview

- Quality and usability
- **Automatic prediction of voice service quality**
  - **MeMo workbench for semi-automatic testing**
  - **SpeechEval workbench**
  - **Prediction of TTS quality**
- Summary and activities @ T-Labs

Automatic Prediction of Voice Service Quality
MeMo workbench for semi-automatic testing

Approach

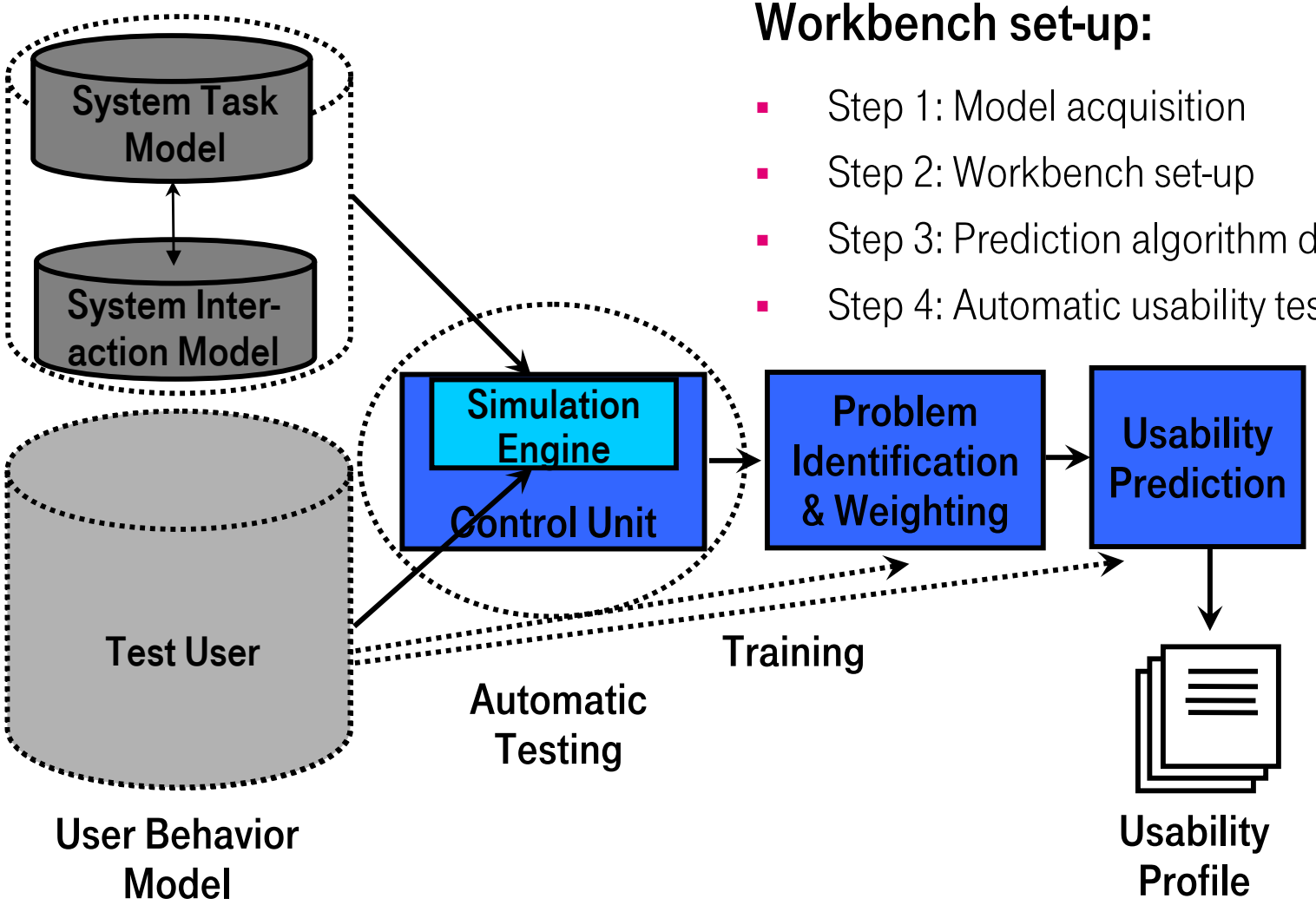
- Make assumptions about (**models of**) the **behavior** of user and application
- Partially **replace the user** in initial evaluations **by a user model**



- Set up a workbench for **automated testing** and **usability prediction**

(Möller et al., 2006)

Automatic Prediction of Voice Service Quality
MeMo workbench for semi-automatic testing.



Workbench set-up:

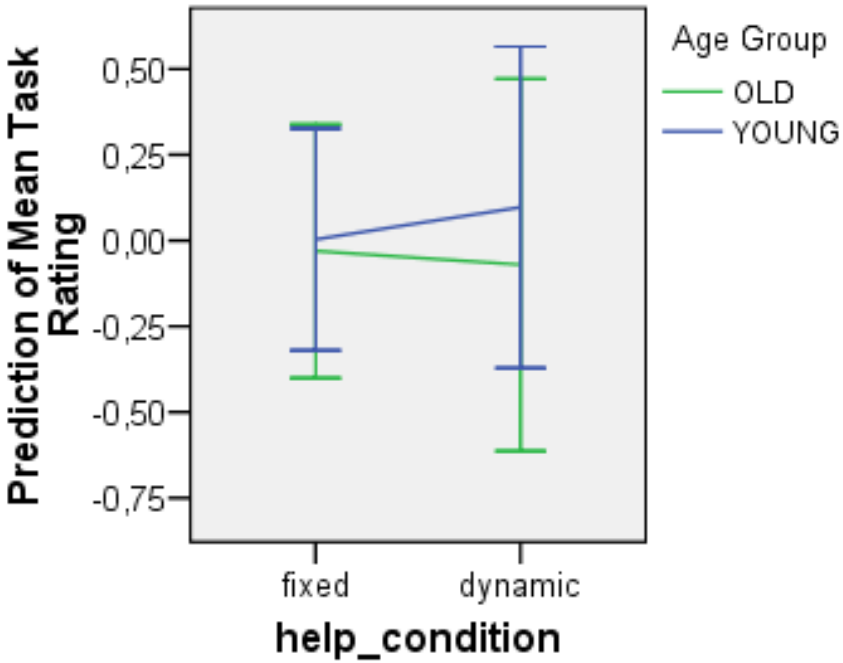
- Step 1: Model acquisition
- Step 2: Workbench set-up
- Step 3: Prediction algorithm derivation
- Step 4: Automatic usability testing

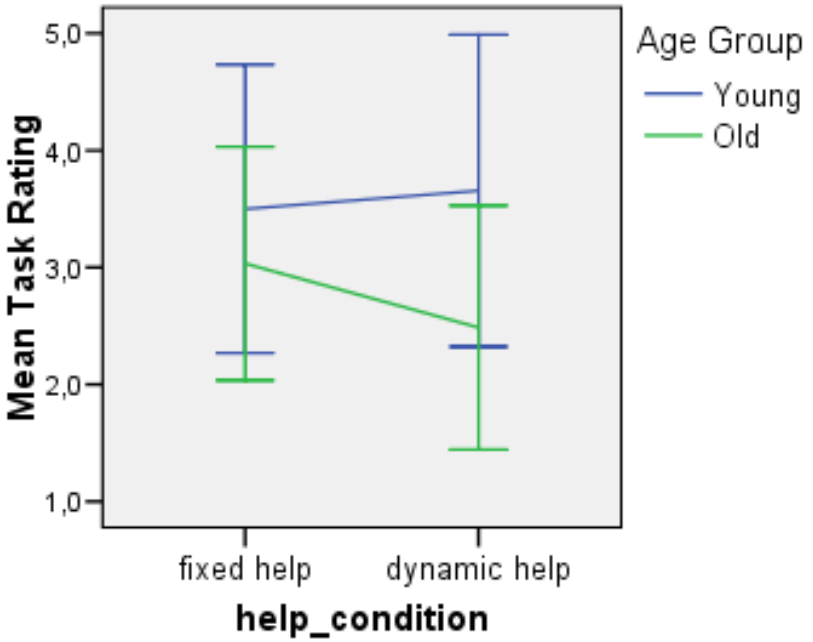
(Möller et al., 2006)

Automatic Prediction of Voice Service Quality
MeMo workbench for semi-automatic testing.

First results

- For **interaction parameters**, relations between configurations are simulated correctly
- **User satisfaction** can roughly be predicted for each configuration

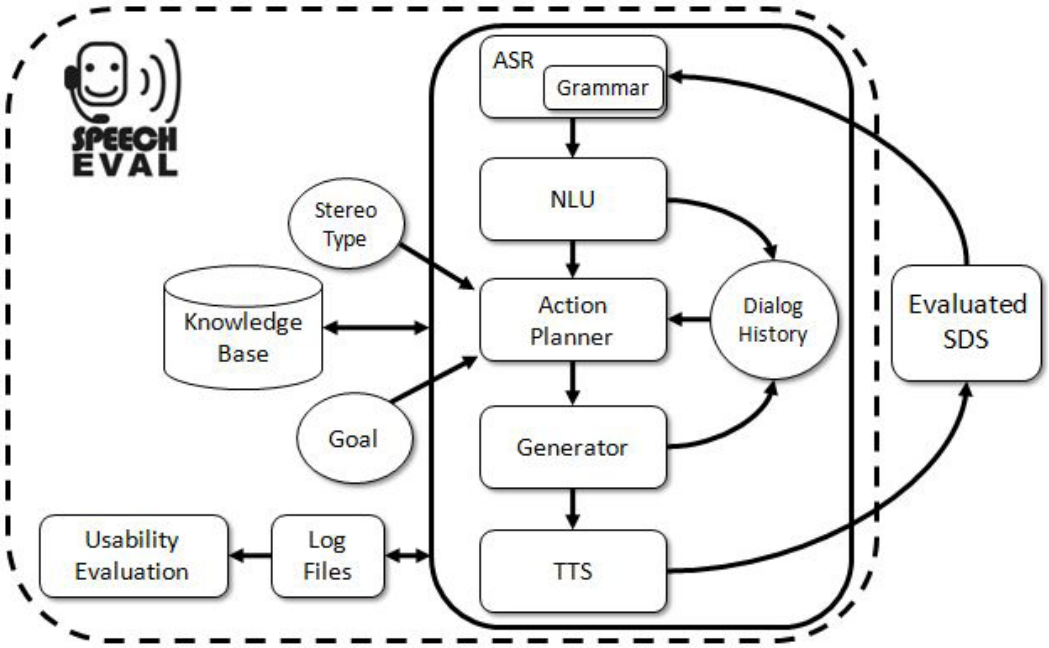




(Engelbrecht et al., 2008 and 2009)

Automatic Prediction of Voice Service Quality SpeechEval Workbench.

Statistical speech-level approach

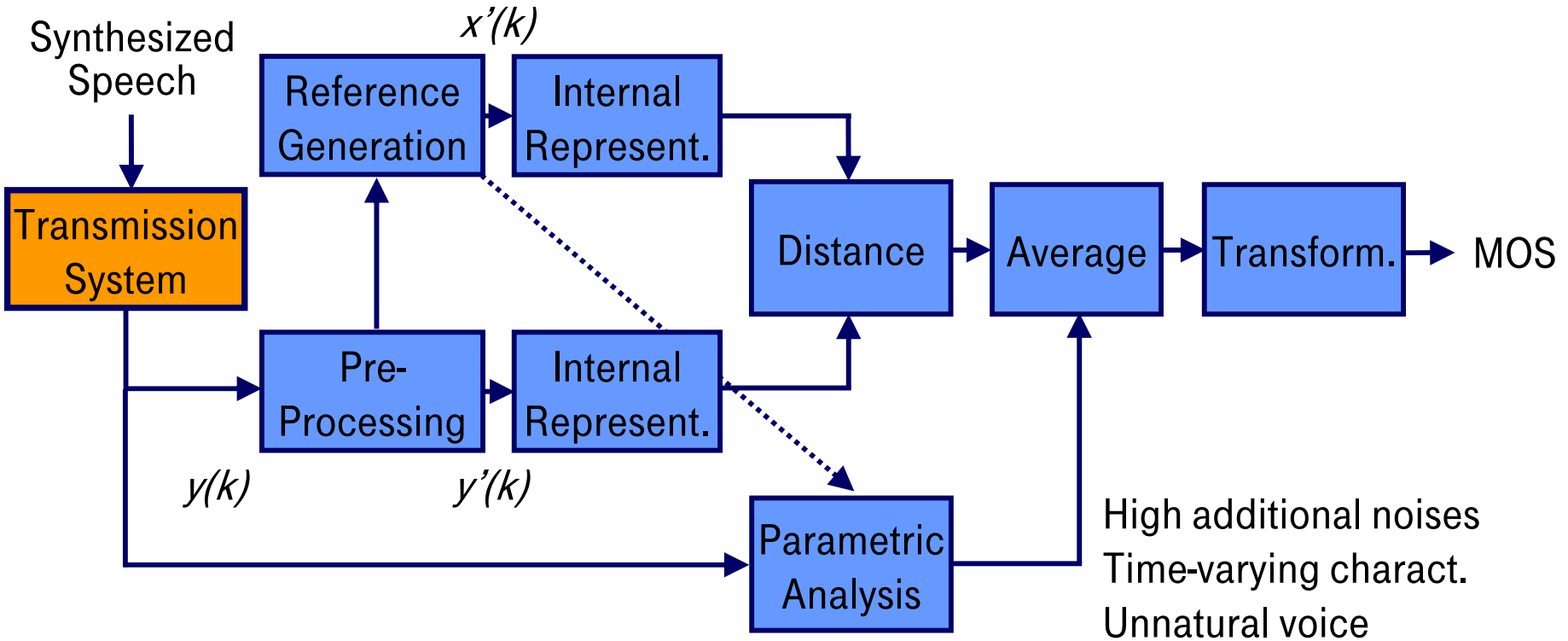


- User behaviour and ratings are **learned from corpora** (i.e. VoiceAward)
- **Domain grammars** to be used in ASR and NLU, and are a basis of domain ontologies
- **Template-based** NLU and NLG
- In the **action planner**, the user model chooses a reply action based on the transition probabilities for similar dialog states observed in the corpus
- **Dialog history** as additional source to determine the optimal next step
- **User stereotypes** (e.g. novice and expert users) for simulating different target behaviors
- **HMM-based** approach for **predicting user satisfaction**

Automatic Prediction of Voice Service Quality
Prediction of TTS quality.

Approaches

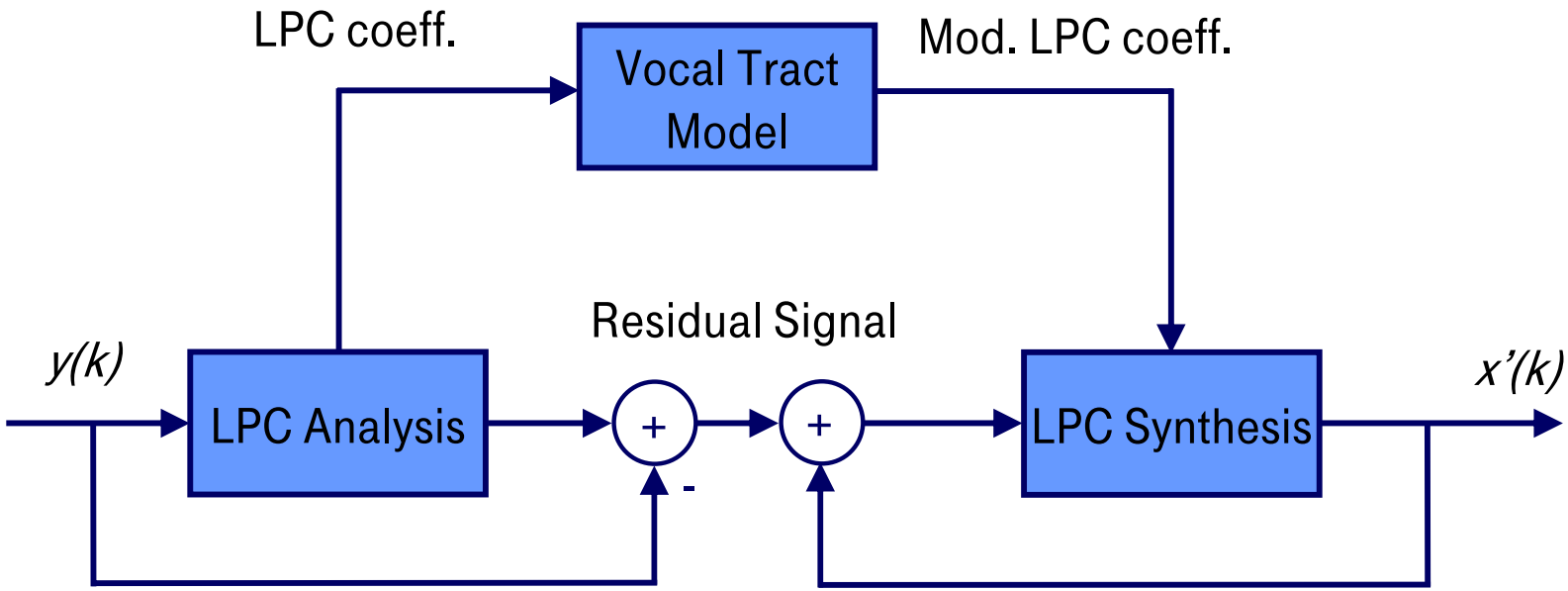
- **Parametric signal-comparison approach**



Automatic Prediction of Voice Service Quality
Prediction of TTS quality.

Approaches

- **Parametric signal-comparison approach**: Reference generation

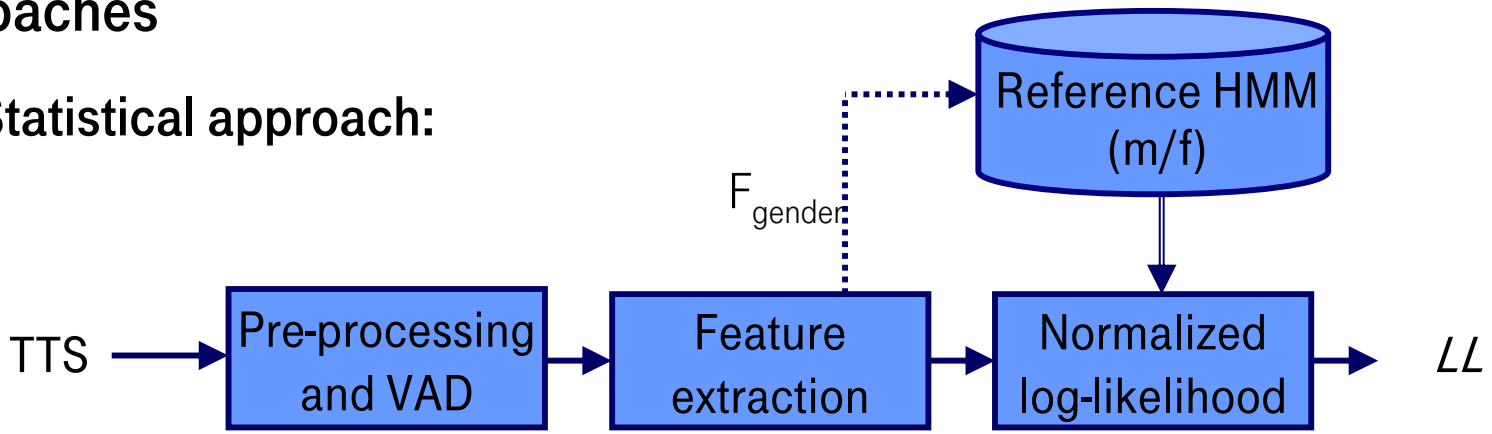


(e.g. ITU-T Rec. P.563, 2004)

Automatic Prediction of Voice Service Quality
Prediction of TTS quality.

Approaches

- **Statistical approach:**



- **Results**: without 3rd-order mapping

Quality dimension	Proposed $\overline{LL}$			ITU-T P.563		
	Male	Female	Overall	Male	Female	Overall
MOS	0.81	0.72	0.77	0.58	-0.05	0.24
LSE	0.72	0.64	0.65	0.50	0.02	0.20
CMP	0.70	0.45	0.54	0.42	-0.11	0.05
ART	0.74	0.47	0.55	0.53	-0.06	0.11
NAT	0.81	0.80	0.81	0.48	-0.06	0.24
PRO	0.54	0.72	0.61	0.28	-0.18	0.12
CFL	0.74	0.81	0.74	0.51	0.06	0.24
ACC	0.65	0.71	0.67	0.35	-0.10	0.15

after 3rd-order mapping

Quality dimension	Proposed $\overline{LL}$		ITU-T P.563	
	Male	Female	Male	Female
MOS	0.83	0.74	0.65	0.05
LSE	0.74	0.70	0.59	0.07
CMP	0.72	0.56	0.48	0.02
ART	0.78	0.57	0.62	0.01
NAT	0.84	0.83	0.59	0.20
PRO	0.61	0.72	0.39	0.07
CFL	0.79	0.82	0.61	0.07
ACC	0.70	0.73	0.47	0.03

(Falk & Möller, 2008)

Overview

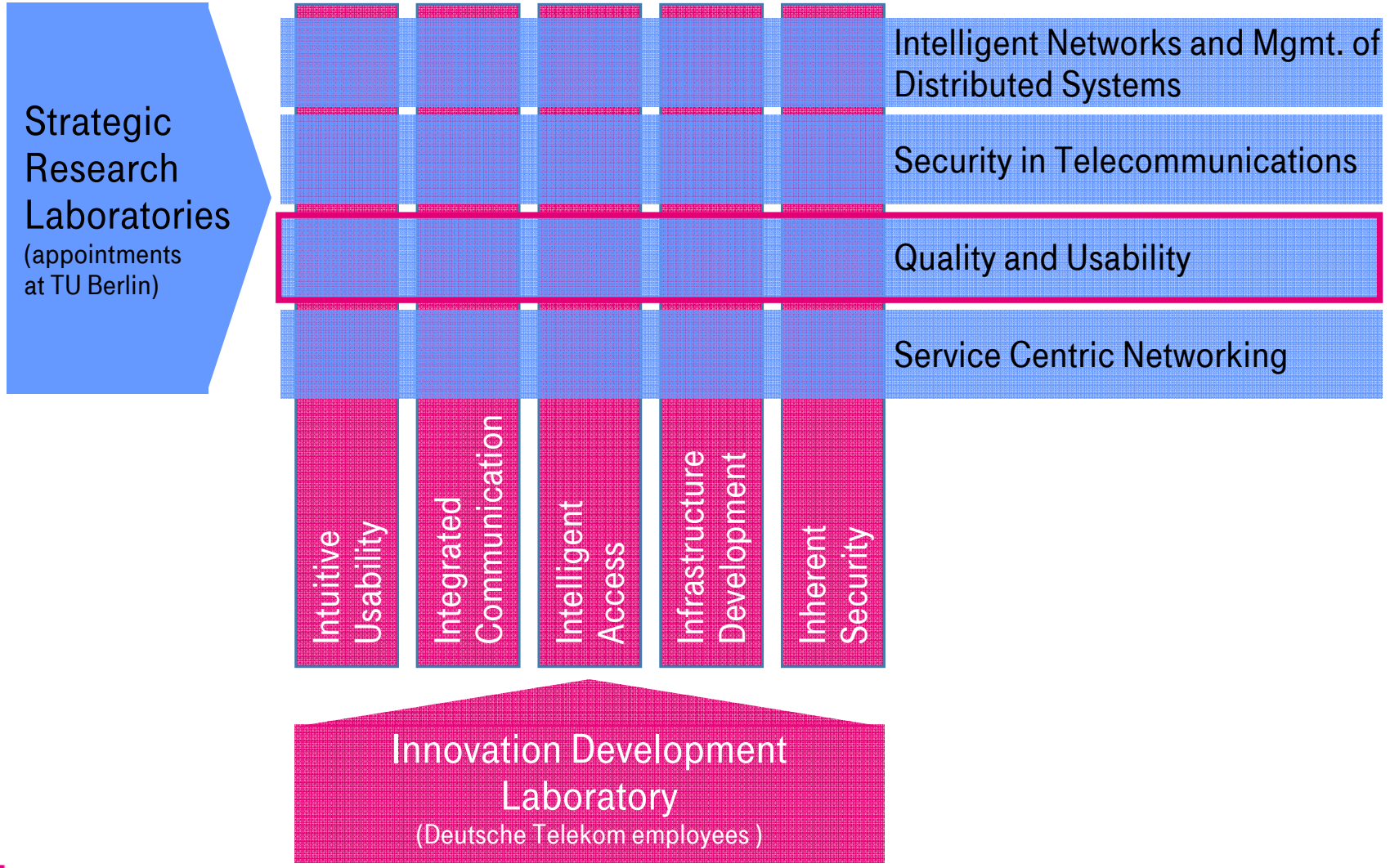
- Quality and usability
- Automatic prediction of voice service quality
- **Summary and activities @ T-Labs**

Summary and Activities @ T-Labs

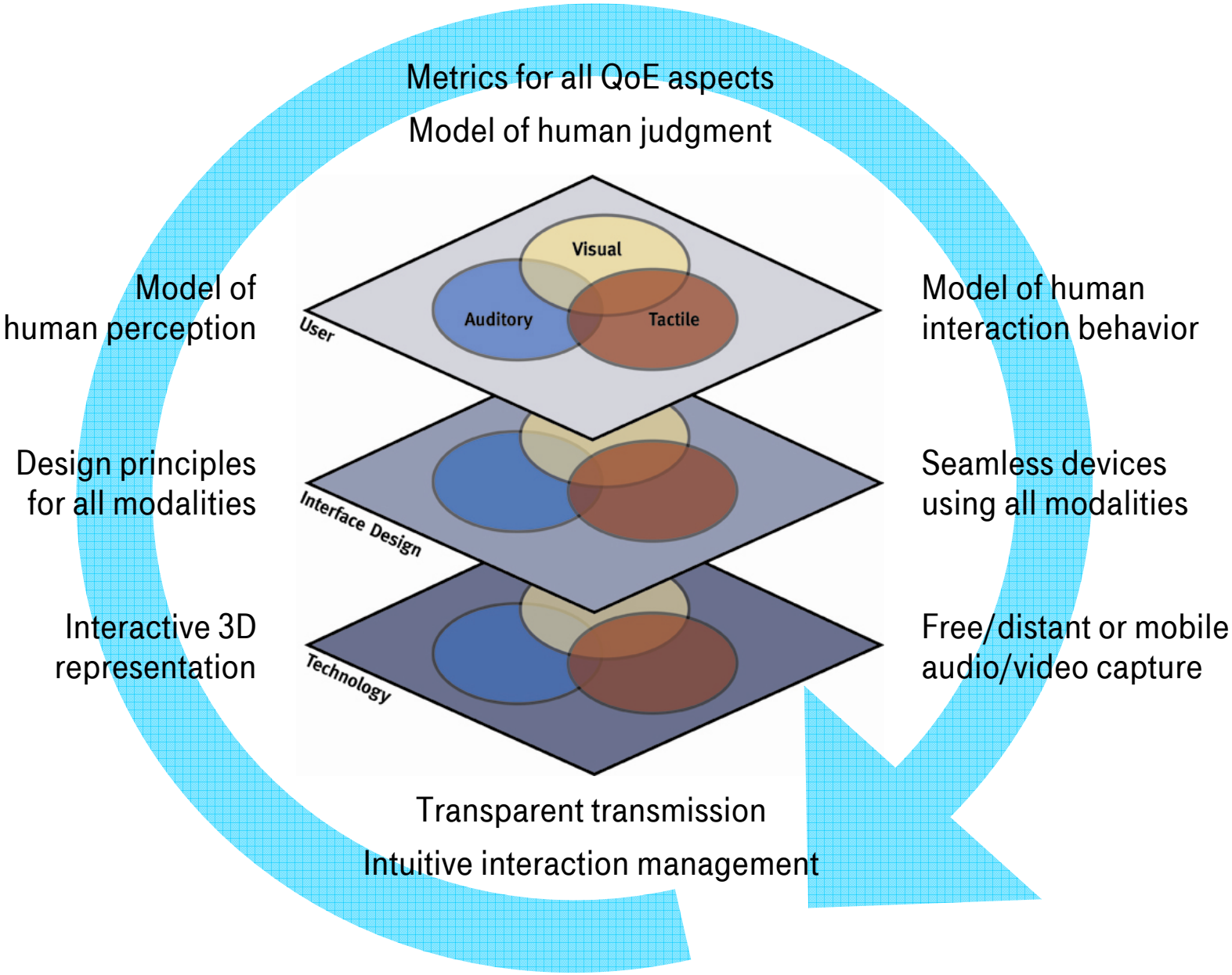
Summary.

- **Assessment** of the **Quality of Experience (QoE)** of spoken dialogue services requires **perception and judgment processes** to take place
  - **Expensive** and **time-consuming** process
  - **Standardized methods** are available, but they cannot yet capture all relevant aspects of usability (cf. taxonomy)
- **Semi-automatic evaluation** principles for **interactive voice services** have been proposed
  - **MeMo workbench**: Rule-based, conceptual level (so far)
  - **SpeechEval**: Corpus-based, speech level
- **Automatic evaluation** of **TTS quality** is underway
  - Parameter-based
  - HMM-based

Summary and Activities @ T-Labs
T-Labs structure.



Summary and Activities @ T-Labs



Summary and Activities @ T-Labs
Core competences.

Subject-based evaluation techniques

Model-based evaluation techniques

Wideband speech perception

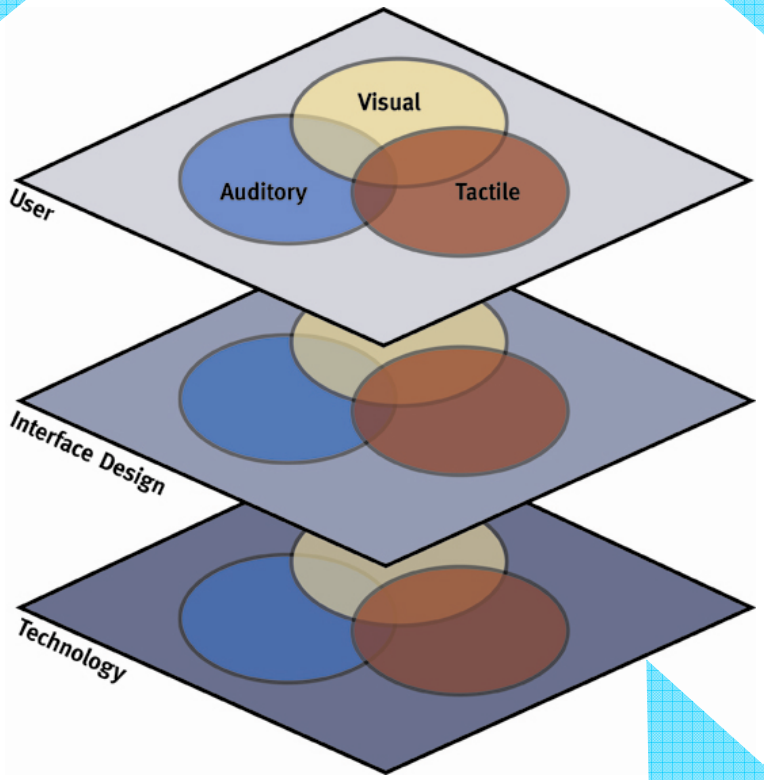
Video perception

Physics-based sound generation

Audio-visual interaction

3D audio rendering

Free viewpoint video rendering



Models for speech interaction

Models for video interaction

Camera-phone interaction

Tactile interaction

Distant speech recognition

Metadata extraction

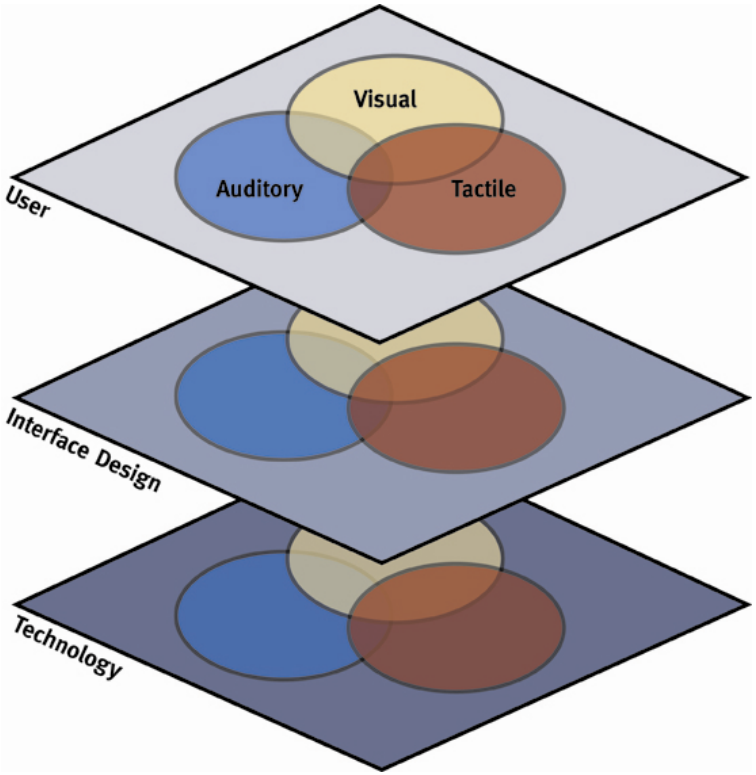
Multi-viewpoint cameras

Speech transmission

Video/image compression

Summary and Activities @ T-Labs
Project groups.

- **Quality:** Perception, judgment and prediction of auditory, visual and tactile signals
- **Mobile & physical interaction:** Developing and testing new forms of interaction
- **Audio technology:** Algorithms and interaction technologies for the auditory channel



- **Usability:** Measuring and optimizing the usability of human interfaces
- **Speech technology:** Identification of speech and speaker characteristics
- **Image & vision computing:** Algorithms and interaction technologies for images and video

Acknowledgements



Thanks to the support from Dmitry Butenkov, Klaus-Peter Engelbrecht, Florian Gödde, Tim Polzehl, Stefan Schaffer, Mathias Schulz and many other colleagues from T-Labs, as well as to DAI-Labor, DFKI, and Michael Wagner (University of Canberra).

Further information can be found under

www.qu.t-labs.tu-berlin.de

Thanks for your attention!

References

Engelbrecht, K.-P., Kruppa, M., Möller, S., Quade, M. (2008). MeMo Workbench for Semi-automated Usability Testing, in: Proc. Interspeech 2008 incorporating SST 2008, AU-Brisbane, 22-26 Sept. 2008, ISCA, 1662-1665.

Engelbrecht, K.-P., Quade, M., Möller, S. (2009). Analysis of a New Simulation Approach to Dialogue System Evaluation, accepted for Speech Communication.

Falk, T. H., Möller, S. (2008). Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems, IEEE Signal Processing Letters 15, 781-784.

ISO 9241-11 (1998). Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability. International Organization for Standardization, Geneva.

ISO 9241-110 (2006). Ergonomics of human-system-interaction – Part 110: Dialogue Principles. International Organization for Standardization, Geneva.

ITU-T Rec. P.85 (1994). A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices, International Telecommunication Union, Geneva.

ITU-T Rec. P.563 (2004). Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications, International Telecommunication Union, Geneva.

ITU-T Rec. P.851 (2003). Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems, International Telecommunication Union, Geneva.

ITU-T Supplement 24 to P-Series Rec. (2005). Parameters Describing the Interaction with Spoken Dialogue Systems, International Telecommunication Union, Geneva.

Jekosch, U. (2005). Voice and Speech Quality Perception. Assessment and Evaluation. Springer, Berlin.

Möller, S. (2005). Quality of Telephone-Based Spoken Dialogue Systems, Springer, New York NY.

Möller, S., Kim, D.-S., Malfait, L. (2008). Estimating the Quality of Synthesized and Natural Speech Transmitted Through Telephone Networks Using Single-ended Prediction Models, Acta Acustica united with Acustica 94, 21-31.

Möller, S., Englert, R., Engelbrecht, K., Hafner, V., Jameson, A., Oulasvirta, A., Raake, A., Reithinger, N. (2006). MeMo: Towards Automatic Usability Evaluation of Spoken Dialogue Services by User Error Simulations, in: Proc. 9th Int. Conf. on Spoken Language Processing (Interspeech 2006 – ICSLP), Pittsburgh PA, 1786-1789.

Möller, S., Engelbrecht, K.-P., Kühnel, C., Wechsung, I., Weiss, B. (2009). A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine Interaction, in: First International Workshop on Quality of Multimedia Experience (QoMEX'09), July 29-31, San Diego CA.