Chapter 6. Data Collection in a Wizard-of-Oz Experiment

in Reinforcement Learning for Adaptive Dialogue Systems
by: Rieser & Lemon.

Course: Autonomous Machine Learning

Hanan Al Rahbi
hanan.alrahbi@snu.ac.kr

Interdisciplinary Program in Cognitive Science
Seoul National University

The Turk

https://youtu.be/RdT4yG8wczQ

Contents

- Introduction
- Experimental Setup
  - Recruited Subjects
  - Experimental Procedure and Task Design
- Noise Simulation
  - Related Work
  - Method
  - Results and Discussion
- Corpus Description
- Analysis
  - Qualitative Measures
  - Subjective Ratings from the User Questionnaires
- Summary and Discussion

What is a WOZ?

- A research method in which a human being simulates the intelligent behavior of a machine
- Before one is able to build a full working system

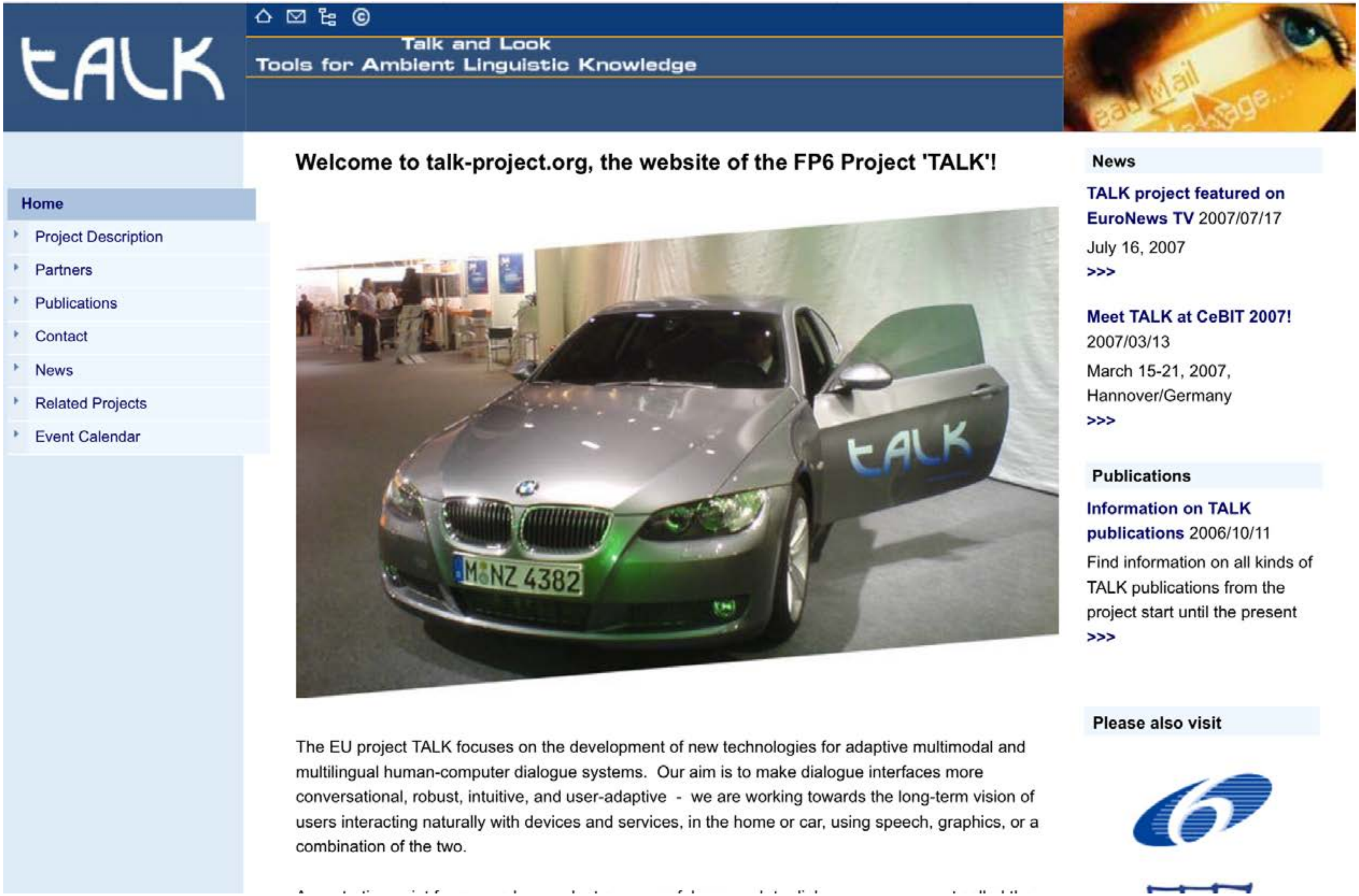
What is it used for?

- Collecting initial data before a system is designed
- Producing more intelligent behavior by current machines

Why is it important?

- Allows data-driven development for domains with no available prototypes
- Helps creating effective policies with the highest rewards
- Saves effort and time

The Experiment



Welcome to talk-project.org, the website of the FP6 Project 'TALK'!

The EU project TALK focuses on the development of new technologies for adaptive multimodal and multilingual human-computer dialogue systems. Our aim is to make dialogue interfaces more conversational, robust, intuitive, and user-adaptive - we are working towards the long-term vision of users interacting naturally with devices and services, in the home or car, using speech, graphics, or a combination of the two.

News

TALK project featured on EuroNews TV 2007/07/17

July 16, 2007

>>>

Meet TALK at CeBIT 2007!

2007/03/13

March 15-21, 2007, Hannover/Germany

>>>

Publications

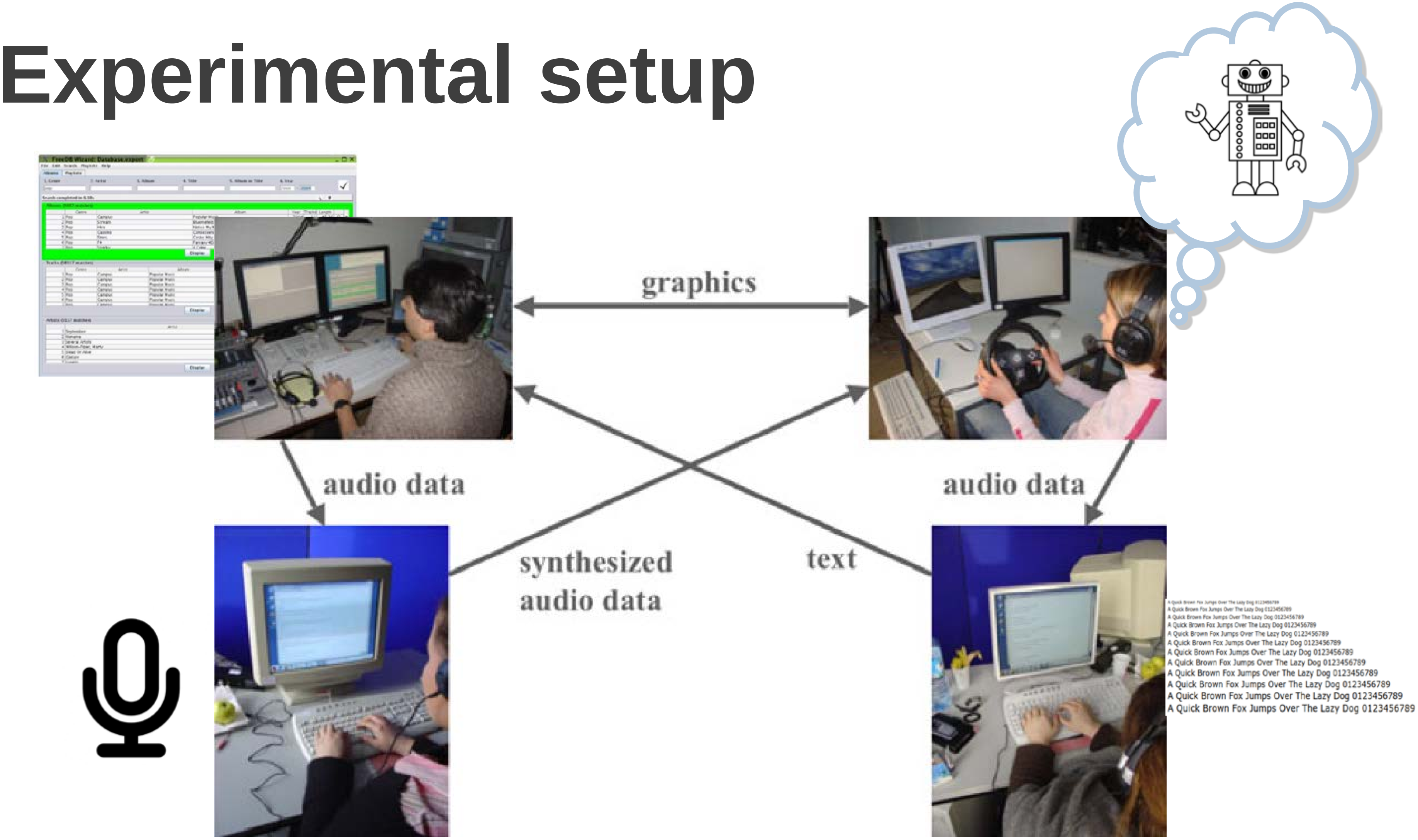
Information on TALK publications 2006/10/11

Find information on all kinds of TALK publications from the project start until the present

>>>

Please also visit

Experimental setup


graphics
audio data
audio data
synthesized audio data
text
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789
A Quick Brown Fox Jumps Over The Lazy Dog 0123456789

Multimodal Wizard-of-Oz data collection setup for an in-car music player application

Recruited Subjects: Wizards & Users

- This experiment focuses on the behavior of users and wizards

	Quota	Age	Language
Wizards	5 (2F, 1M)	20~35	German: Native English: Good

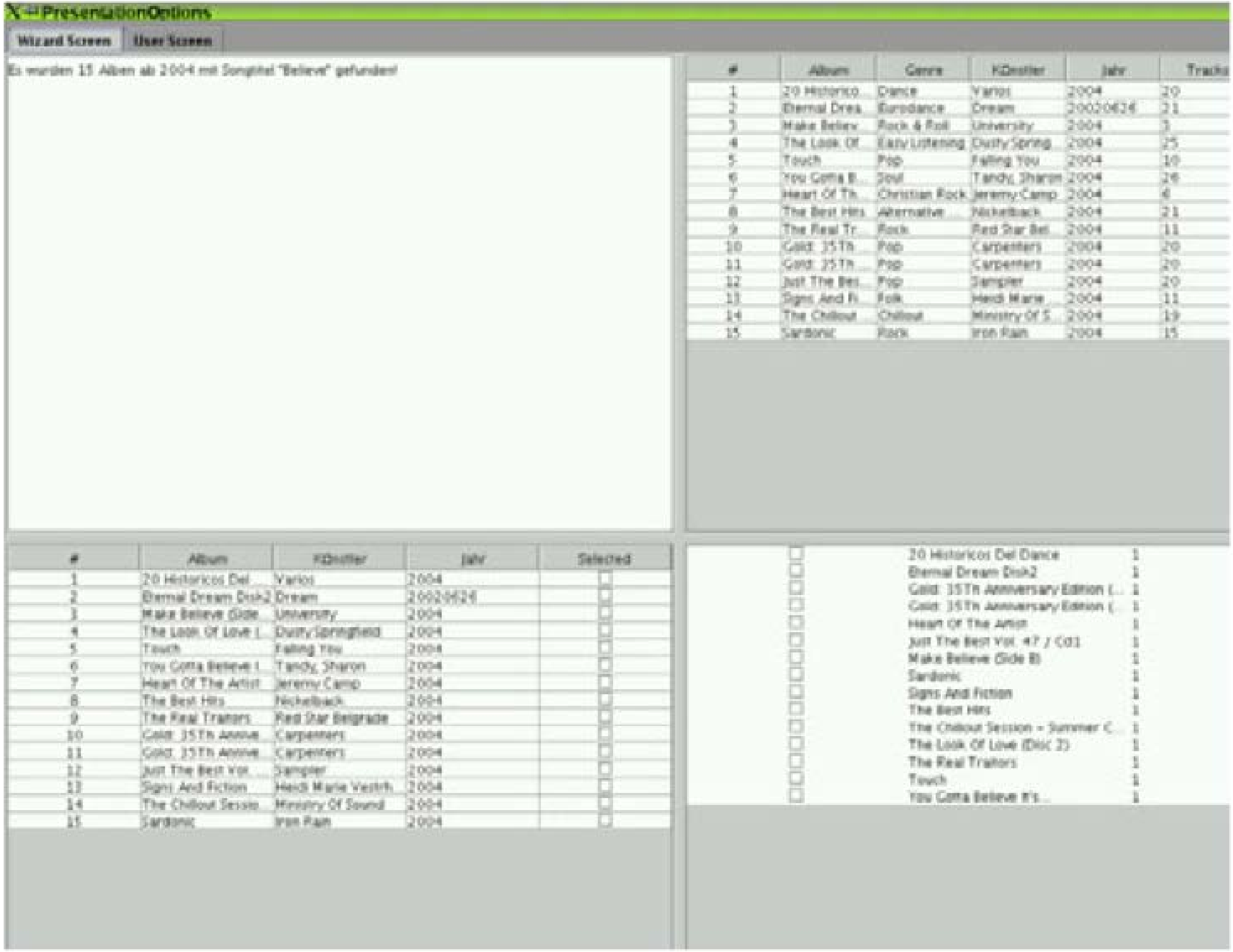
No Experience in dialogue systems

€

	Quota	Age	Field of Study
Users	21 (11F, 10M)	20~30	Social Science:23.8% Languages:23.8% Natural Sciences:28.6% Arts:17%

Experimental Procedure and Task Design

1. Training wizards (database, interaction with users)
2. User and wizard placed in separate rooms
3. User received sheet of instructions upon arrival
4. Introducing the user to the driving simulator (tested)
5. User could solve the tasks in any order they preferred
6. After each task user filled task-specific questionnaire
7. User interviewed by experiment leader



Es wurden 15 Alben ab 2004 mit Songtitel "Believe" gefunden!

#	Album	Genre	Künstler	Jahr	Tracks
1	20 Historico…	Dance	Varios	2004	20
2	Eternal Drea…	Eurodance	Dream	20020626	21
3	Make Believ…	Rock & Roll	University	2004	3
4	The Look Of…	Easy Listening	Dusty Spring…	2004	25
5	Touch	Pop	Falling You	2004	10
6	You Gotta B…	Soul	Tandy, Sharon	2004	26
7	Heart Of Th…	Christian Rock	Jeremy Camp	2004	6
8	The Best Hits	Alternative	Nickelback	2004	21
9	The Real Tr…	Rock	Red Star Bel…	2004	11
10	Gold: 35Th …	Pop	Carpenters	2004	20
11	Gold: 35Th …	Pop	Carpenters	2004	20
12	Just The Bes…	Pop	Sampler	2004	20
13	Signs And Fi…	Folk	Heidi Marie …	2004	11
14	The Chillout …	Chillout	Ministry Of S…	2004	19
15	Sardonic	Rock	Iron Rain	2004	15

#	Album	Künstler	Jahr	Selected
1	20 Historicos Del …	Varios	2004	☐
2	Eternal Dream Disk2	Dream	20020626	☐
3	Make Believe (Side …	University	2004	☐
4	The Look Of Love (…	Dusty Springfield	2004	☐
5	Touch	Falling You	2004	☐
6	You Gotta Believe I…	Tandy, Sharon	2004	☐
7	Heart Of The Artist	Jeremy Camp	2004	☐
8	The Best Hits	Nickelback	2004	☐
9	The Real Traitors	Red Star Belgrade	2004	☐
10	Gold: 35Th Annive…	Carpenters	2004	☐
11	Gold: 35Th Annive…	Carpenters	2004	☐
12	Just The Best Vol. …	Sampler	2004	☐
13	Signs And Fiction	Heidi Marie Vestrh…	2004	☐
14	The Chillout Sessio…	Ministry Of Sound	2004	☐
15	Sardonic	Iron Rain	2004	☐

☐	20 Historicos Del Dance	1
☐	Eternal Dream Disk2	1
☐	Gold: 35Th Anniversary Edition (…	1
☐	Gold: 35Th Anniversary Edition (…	1
☐	Heart Of The Artist	1
☐	Just The Best Vol. 47 / Cd1	1
☐	Make Believe (Side B)	1
☐	Sardonic	1
☐	Signs And Fiction	1
☐	The Best Hits	1
☐	The Chillout Session - Summer C…	1
☐	The Look Of Love (Disc 2)	1
☐	The Real Traitors	1
☐	Touch	1
☐	You Gotta Believe It's…	1

a) Simple text-message conveying how many results were found
b) Output of a list of just the name (album, song, or artist)
c) A table of complete search results
d) A table of complete search results but only displaying a subset of of columns.

Experimental Procedure and Task Design

- Designed 10 task sets
- Every task set was used at least twice
- Each set contains 4 tasks of 2 different types:
  - Search for a specific title/ album
  - Build a playlist

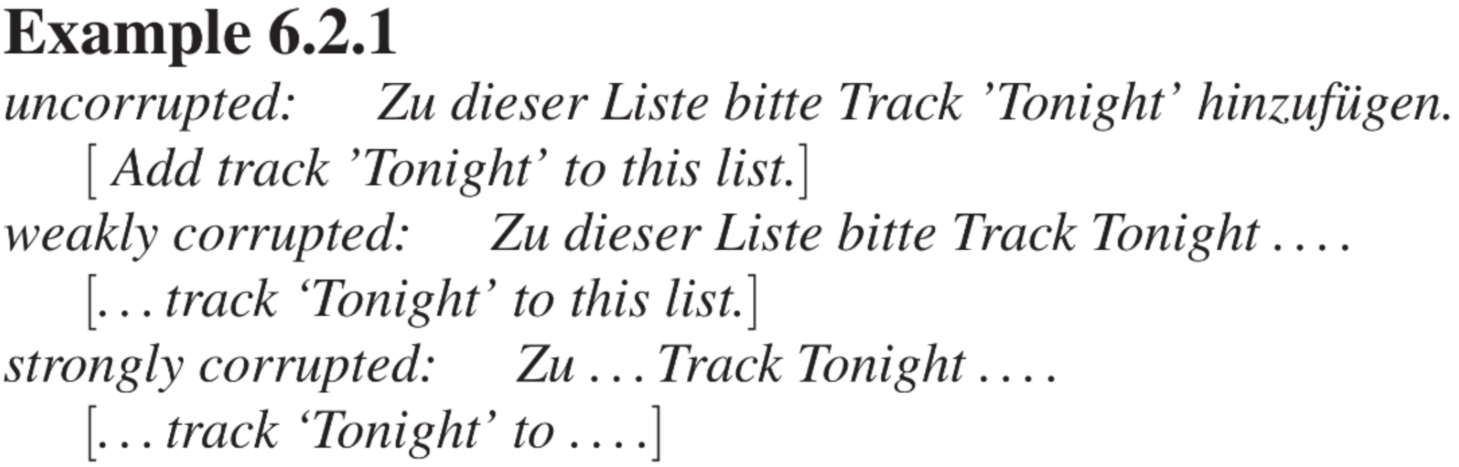
Noise Simulation

- HCI vs. WOZ
- Related work:
  - Skauntze (2003,2005)
  - Stuttle et al (2004); Williams and Young (2004a)
    - Even with high noise wizards are able to interpret the ASR output well and assimilate contextual knowledge about what user actions are likely to follow

Noise Simulation

- Method:
  - To approximate speech recognition errors, a tool was used to randomly delete parts of the transcribed utterances
  - Wizards also build up their own hypotheses about what the user really said (misunderstandings)
  - Word deletion rate of the text varied:
    - 20% weakly corrupted = deletion rate of 20%
    - 20% strongly corrupted= deletion rate of 50%
    - In 60% of the cases the wizard saw the transcribed speech uncorrupted

Noise Simulation



Example 6.2.1

uncorrupted: *Zu dieser Liste bitte Track ’Tonight’ hinzufügen.*

[*Add track ’Tonight’ to this list.*]

weakly corrupted: *Zu dieser Liste bitte Track Tonight* . . . .

[. . . *track ‘Tonight’ to this list.*]

strongly corrupted: *Zu* . . . *Track Tonight* . . . .

[. . . *track ‘Tonight’ to* . . . .]

Results and Discussion

- 30% of the corrupted utterances had a noticeable effect on the interaction
- 7% of all user turns lead to a communication error (much lower than the current WER for spoken dialogue systems {around 30%})
- On the other hand, the error rate is higher than for human-human communication

Results and Discussion

- Shortcomings of the deletion method:
  - Deleting words is a rather crude simulation of real-world acoustic problems (justified)
  - Time delay introduced by transcribing the utterances (both of user and wizard)

- This method is not suitable for studying detailed error, however, it can be sufficient in order to study natural presentation strategies under the presence of noise.

Corpus Description

- 21 sessions, containing 72 dialogue, with about 1600 turns were gathered
- Data for each session includes video and audio recordin, questionnaire data, transcripts, and a log file
- The logging information per session consists of OAA messages in chronological order
- Corpus is marked up and annotated using Nite XML Toolkit (NXT)

Analysis

- Results of corpus analysis for multimodal presentation strategies
- Qualitative measures:
  - 22.3% of the 793 wizard turns were annotated as presentation strategies, resulting in 177 instances for learning
  - 48% screen output
    - 78.6% the table option
    - 17% the list
    - 0.04% text only

  - Verbal presentation only present 1.6 items on average
    - Where wizard summarized the results by presenting the options for the most distinctive feature to the user.

Analysis

- Did the Wizards apply significantly different strategies? It is important to compare! (data will be used for learning)
  - Dialogue length is about the same with very slight differences between wizards
  - Most wizards were equally successful in completing tasks, only one was better with 100% task success, where another one scored 78% task success

Therefore we can say they applied similar strategies (this doesn't mean they react the same way)

However, multimodal behavior of wizards is very limited

  - Only 3 users selected an item by clicking

Analysis

- Subjective Ratings from the User Questionnaire

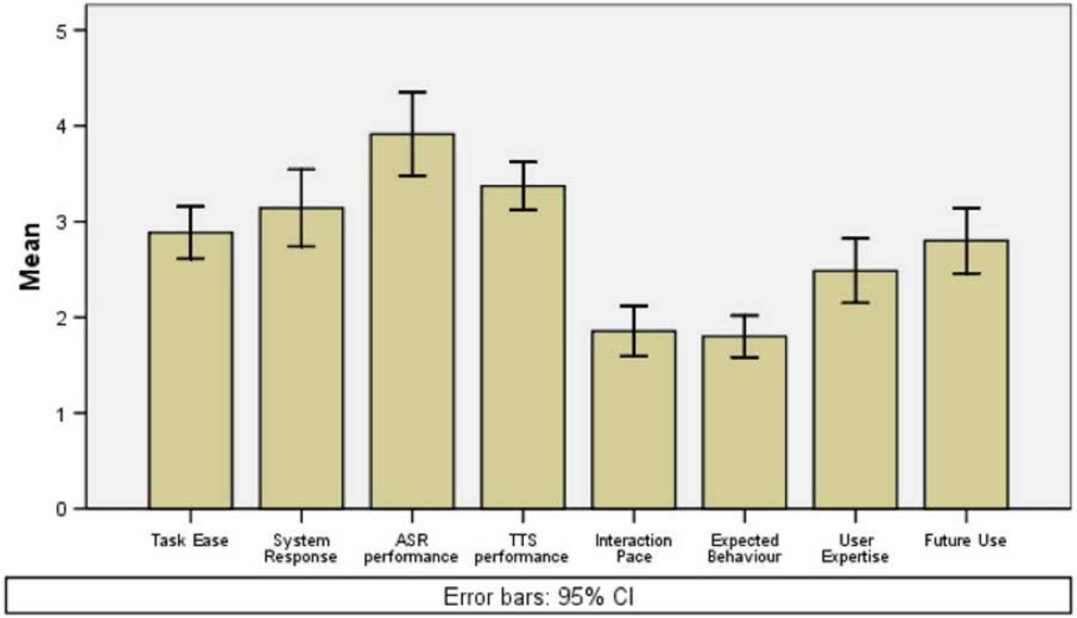
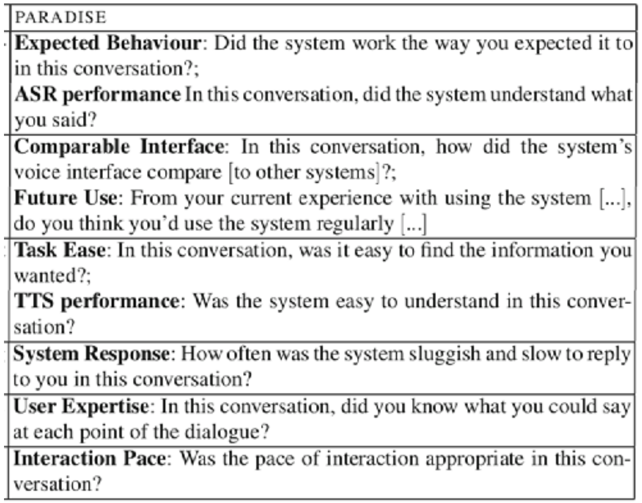


Fig. 6.6 User Ratings on the PARADISE questions on a 5-point Likert Scale



PARADISE
Expected Behaviour: Did the system work the way you expected it to in this conversation?; **ASR performance** In this conversation, did the system understand what you said?
Comparable Interface: In this conversation, how did the system's voice interface compare [to other systems]?; **Future Use**: From your current experience with using the system [...], do you think you'd use the system regularly [...]
Task Ease: In this conversation, was it easy to find the information you wanted?; **TTS performance**: Was the system easy to understand in this conversation?
System Response: How often was the system sluggish and slow to reply to you in this conversation?
User Expertise: In this conversation, did you know what you could say at each point of the dialogue?
Interaction Pace: Was the pace of interaction appropriate in this conversation?

Analysis

- Subjective Ratings from the User Questionnaire

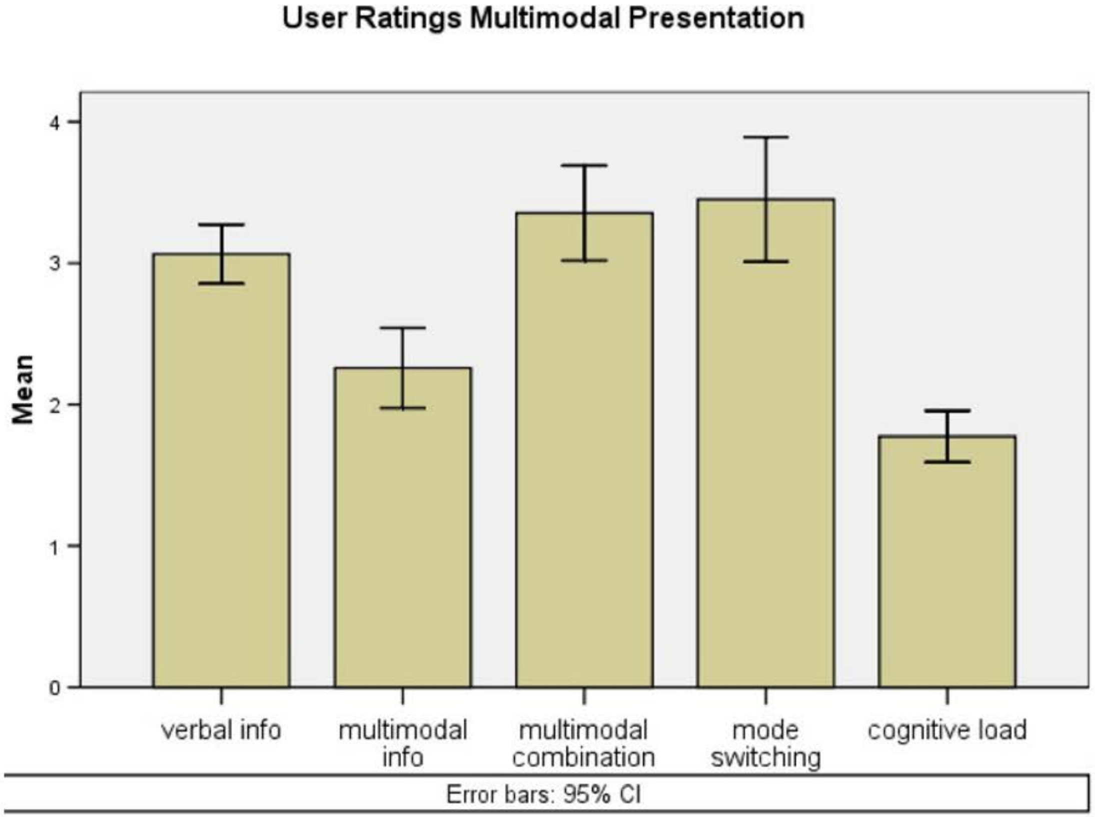


Fig. 6.7 User Ratings on the multimodal presentation questions on a 5-point Likert Scale

Discussion

- Common mistakes (the wizard either):
  - Either wizard displayed too much information On the screen
  - Or fail to present results early enough

Screen output should display appropriate amount of information

- There is a need for a strategy which decides how many database search results to present to the user, when, and which modality to use in an adaptive optimal matter
- Also a strategy to help minimize the large lists displayed, cut the length of the dialogue, as well as the noise
- Include information about users driving performance is very important
- There should be a better and more realistic in-car simulation (the screen size)