

Traps and Tricks when Designing Good Speech Interfaces

Lars Bo Larsen

Multimedia Information and Signal Processing

Dept. of Electronic Systems

Aalborg University, Denmark

Outline

- Success criteria for speech interfaces – what do people *really* want? – a view on past successes and failures
- What speech is good and *not* good for
- Some design recommendations for speech interfaces
- The future belongs to multi modality!
 - Distributed Speech Recognition (DSR) as a paradigm for speech based interaction on mobile devices
 - Examples of services we're currently developing at Aalborg University
- Conclusions and questions

Factors for successful speech services

In 2000, the COST249 project organised a workshop with the title:

“Voice Operated Telecom Services: Do they have a bright future?”

- The purpose was (obviously) to discuss and assess whether this was the case.
- One of the invited keynote speakers was Hugh Cameron, from Learnout and Hauspie. In his talk*, he presented an investigation of a large number of commercial speech driven services in the U.S. and Canada
- His success criteria were simple:
 - Was the service commercially viable?
 - Is it still active?
 - Cameron identified the business goals, the factors leading to success or failure and gave each service Smileys accordingly

*) Hugh Cameron: **“SPEECH AT THE INTERFACE”**, in proc of: **COST249 Workshop: “Voice Operated Telecom Services: Do they have a bright future?”**, Ghent, Belgium 2000

Very Successful Speech Services

Service	Users	Business objective	Success factors	Impediments	Result
Third party and collect call billing	general public	reduce operator workload	operator backing, provable business case, involuntary use		😊😊
Public directory assistance	general public	reduce DA operator workload and stress	operator backing (North America), provable business case		😊😊
Package tracking	general public	reduce agent workload for delivery company	provable business case, possible human operator backing		😊😊
Auto attendant	general public	reduce receptionist workload, save telco number charges	use is involuntary, ability to have human operator or DTMF backing, no user admin / training load		😊😊
Desktop/mobile dictation	professionals	reduce typist / transcriber workload	provable business case	recognizer training by user	😊😊

Successful Speech Services

Service	Users	Business objective	Success factors	Impediments	Result
Voice activated voice mail	voice mail users	product differentiation	eyes free, hands free operation, possible DTMF backing		☺
Voice activated wireless dialing	general public (mobile users)	substitute for remembering phone numbers, eyes free dialing	eyes free operation	directory creation by user, poor noise robustness	☺
Stock quotations	brokerage customers	reduce broker workload	provable business case, human operator backing	100% automation intent (foregoes upsell opportunities)	☺
Travel information	general public	reduce agent workload for travel retailer	possible human operator backing	unclear service boundaries	☺
Travel reservations	travel agents	reduce agent workload for travel wholesaler	closed user group, provable business case, human operator backing	user training required, complex transactions	☺
Mobile dictation	Professional	substitute for typing	portable device, not tied to desk	recognizer training by user, output editing effort	☺

Somewhat Successful Speech Services

Service	Users	Business objective	Success factors	Impediments	Result
Personal agent	mobile professionals	Remote control of inbound call handling, voice mail, calendar	output voice quality (recorded prompts)	high configuration load, unclear service boundaries, decreased accessibility	☺☹
Desktop dictation	professionals	substitute for typing	visual feedback	recognizer training by user, output editing effort	☹
Voice activated email	(corporate) email users	service differentiation, increase airtime revenue	not tied to desk / PC	output voice quality, slow for scanning, limited composing capability	☹
Telephone banking	general public	improve transaction speed, reduce agent workload	measurable business value, human operator backing	lack of persistent feedback, over-confirmation	☹

Unsuccessful Speech Services

Service	Users	Business objective	Success factors	Impediments	Result
Voice activated premier dialing	general public	substitute for looking up business phone numbers	no user configuration / training effort	unclear service boundary, difficult to prove value to subscribing businesses	☹
Desktop command & control	PC users (professionals)	substitute for mouse/keyboard input	visual feedback	disturbs privacy in offices (noise), unclear service boundary	☹
Voice activated wire line phone dialing	general public	substitute for remembering phone numbers		directory creation by each user, interference with modem use, various marketing mistakes	☹☹
Calling card voice activated dialing	calling card users	calling card customer loyalty		explicit caller authentication, lack of privacy, directory creation by user	☹☹

Cameron's' conclusions:

Based on his survey, Cameron recommended to use:

- Use human operator or DTMF fallback
- Avoid voice output of lists or tabular information
- Target few, specific tasks
- High voice output quality
- Visual feedback whenever possible

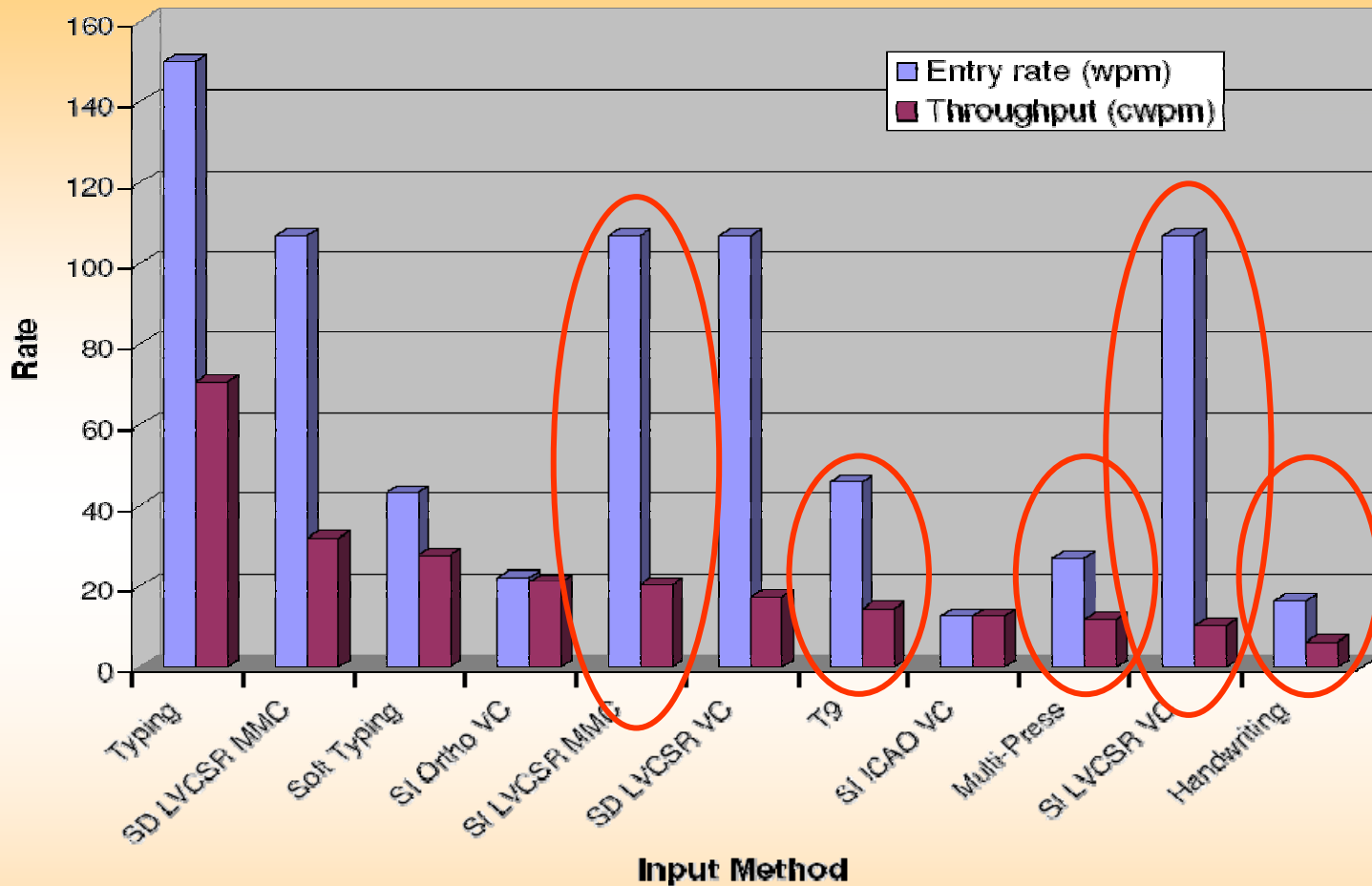
- Avoid configuration management by users
- Avoid (explicit) checking of user identity
- Save identifiable costs (and measure them)

Cameron's' conclusions:

So, when will people use speech services?

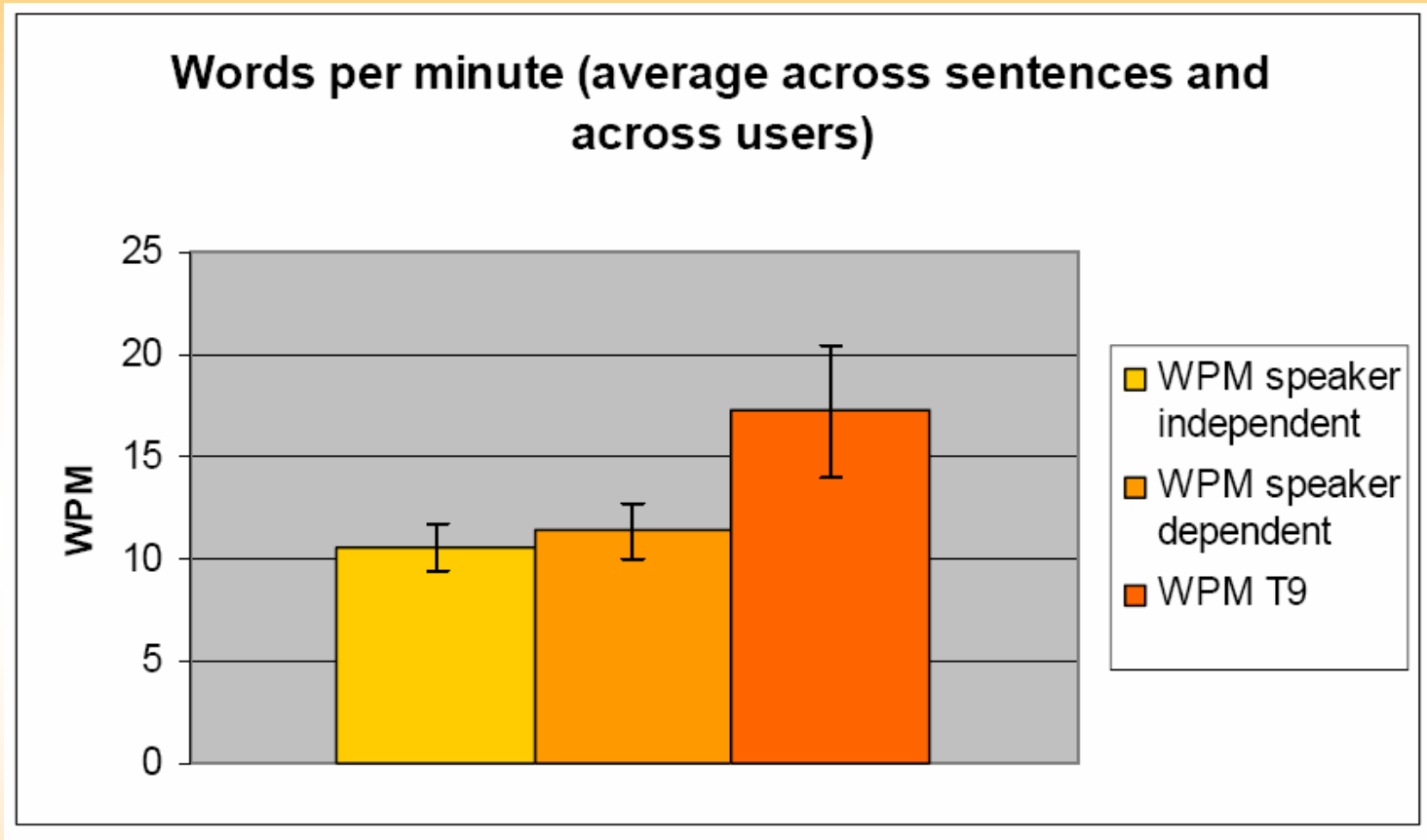
- When they are offered no choice
- When it corresponds to the privacy of their surroundings and the task at hand
- When their hands and eyes are busy on another task
- When it's quicker than any alternative

Comparison of predicted entry rates for SMS composition



From: Research Challenges in the Automation of Spoken Language Interaction
Roger K. Moore, *ASIDE 2005 workshop, Aalborg*. Used with permission.
<http://www.dcs.shef.ac.uk/~roger/publications/RKM%20Keynote%20ASIDE2005.pdf>

SMS dictation entry rates for Nokia 6630

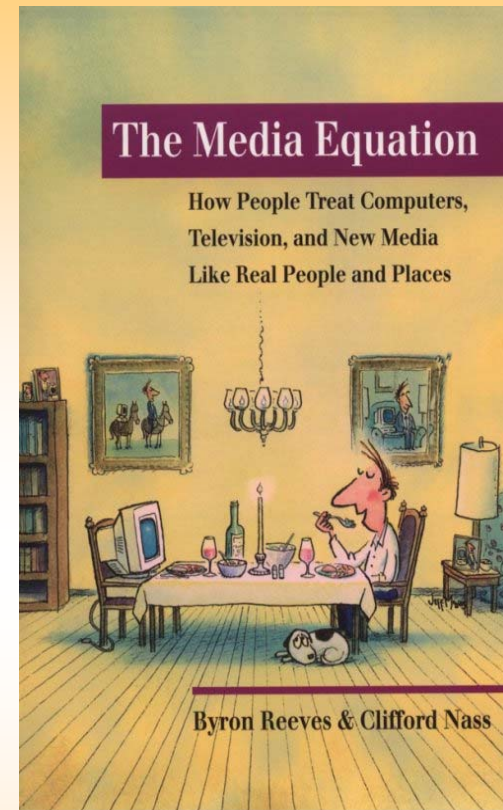


From: "SHORT MESSAGE DICTATION ON SYMBIAN SERIES 60 MOBILE PHONES"
E. Karpov et al, in Proceedings of the SIMPE workshop, Helsinki, S2006. Used with permission

Cognitive Implications of Speech Interfaces

Reeves and Nash: “The Media Equation”:

- Users attribute human qualities to systems mediated by speech
- When a service violates a rule of politeness, the violation is viewed as social incompetence and it is offensive
- Users respond more positively to clearly identifiable personalities than to ambiguous or inconsistent personalities
- A bigger image or louder voice have more emotional impact
- Negative messages are more easily remembered than positive ones



Cognitive Implications of Speech Interfaces

- A female voice is perceived as a more trustworthy source of information about relationships, while a male voice is considered more reliable when dispensing technical information (stereotypes - not only gender, but also age, social class)
- In terms of users' attention, memory and evaluation of content, voice and audio quality are more important than the image quality
- Flattery, even by machines, works!

Cognitive Implications of Speech Interfaces

- An overly loud voice will make users even less likely to cooperate with the system
- Over-confirmation will be considered annoying to users, as a form of unwelcome redundancy
- A system response preceded by a significant pause is perceived to be less trustworthy than a quick one
- An irrelevant response (which is perceived as impolite) is more annoying to users than an incorrect, but relevant one
- Voice and audio fidelity, absence of spurious “cuts” correlate positively to users perception of the quality of the information

Speech vs. Visual displays

Speech is:

- Effective for conveying emotion, narrative thread and crescendo
- Compact for communicating ambiguity
- Difficult to scan - serialised
- Quick for denoting choice, selecting, or asking questions
- Slow for enumerating choices
- Objects are not visible
- Loud, omni directional (public)
- Non-persistent

Visual Displays are:

- Inefficient at conveying deliberate ambiguity, emotion, nuance
- Easily scanned or searched - parallel
- Quick for presenting choices, patterns, structure
- Objects are visible
- Silent and unidirectional (private)
- Persistent

Some Design Recommendations:

Speech vs. Graphical displays

- Use speech for input and graphics for output, whenever possible, and no special circumstances (such as hands-busy/eyes-busy) are present
- Be very careful with speech output – pay close attention to the quality and consistency of speech output
- Use text input as backup for speech (c.f. DTMF)

General

- Avoid configuration management by users

Some Design Recommendations:

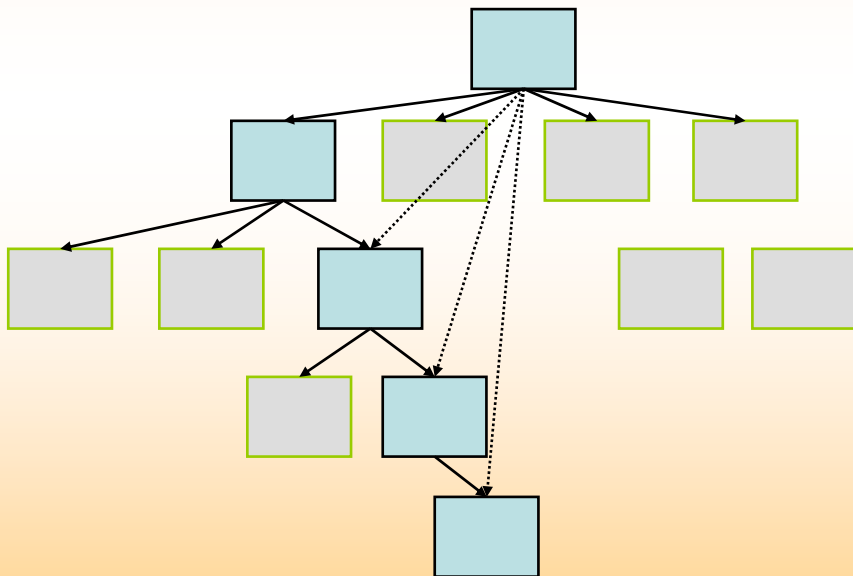
Fundamental problem: How to solve the problem of accommodating novice and experienced user simultaneously:

1. Novice users require guidance:
 - Provide system-directed dialogue, closed questions, etc
 - Provide training (tutor) / user profile, if applicable
 - Do not provide very long instructions or prompts
2. Experienced users require direct access to functions
 - Provide user-directed dialogue, open questions, etc.
 - Provide barge-in and control of the pace
 - Provide help on request only

A Solution: Pseudo Sub-Menus:?

Introduce pseudo menu hierarchy to aid novice users while retaining underlying flat structure for experienced users

- offers guidance (structure) for novice users
- leaves option for efficient navigation for experienced users



Problem: How do the experienced users discover the options?



Some indications that they can find out for themselves – at least for simple cases

Multi Modal User interfaces on Mobile Devices

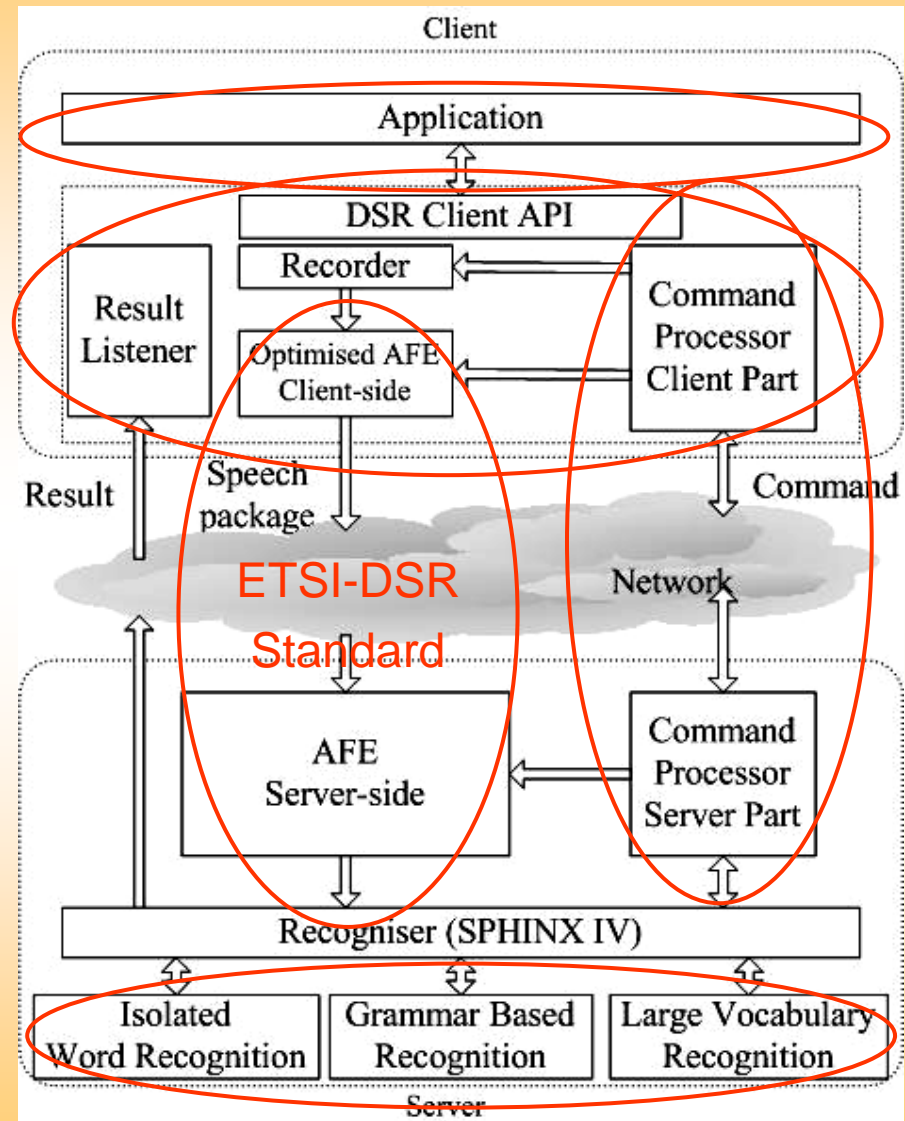
“Speech Centric” multi modal user interfaces on mobile devices have been very much in focus in the research community since 2000:

- Numerous projects and labs have developed prototypes to investigate how best to design such interfaces. However, no commercial services have emerged yet (to my knowledge)
- A central usability problem is how best to integrate the serial, non-persistent speech modality with the parallel persistent visual modality
- A more technical problem is that, while speech-only, telephone-based services relies on a standard telephone connection, multi-modal applications demands both voice and data connections. Furthermore, most research is done using PDA's

Distributed Speech Recognition (DSR)

Client-server architecture:

- Enables high-complexity speech recognition tasks on portable devices
- Based on the ETSI Aurora advanced front-end standard for DSR
- Real-time implementation and integration into network environment
- Low data rates (5.6 kbit)
- Robust to packet loss, delays, due to robust signalling protocol
- Allows simultaneous voice and data communication on mobile devices



Flexible server architecture

- supports multi-user simultaneously accessing with
 - Different speech recognition tasks
 - Different character set

Noise robustness and error concealment

- ETSI-DSR advanced front-end (fixed-point version)

Speech recogniser

- SPHINX 4 (open source)

Data streams and network load

- Data channel (speech and result): 5.6kbps
- Control channel (command): negligible

AAU Real-time DSR implementation

Flexible server architecture:

- Supports multi-user simultaneously accessing with:
 - Different speech recognition tasks
 - Different character set
 - Little signalling overhead

We Use the CMU Sphinx4 Decoder as basis for our recognition engine:

- A recognizer designed to be flexible (java).
- Live mode and batch mode speech recognizers, capable of recognizing discrete and continuous speech.
- Generalized pluggable language model architecture, supporting for example BNF-style grammars and N-gram models.
- Open source, LM Toolkit, community support, etc.

Applications:

POSH Project 9:30

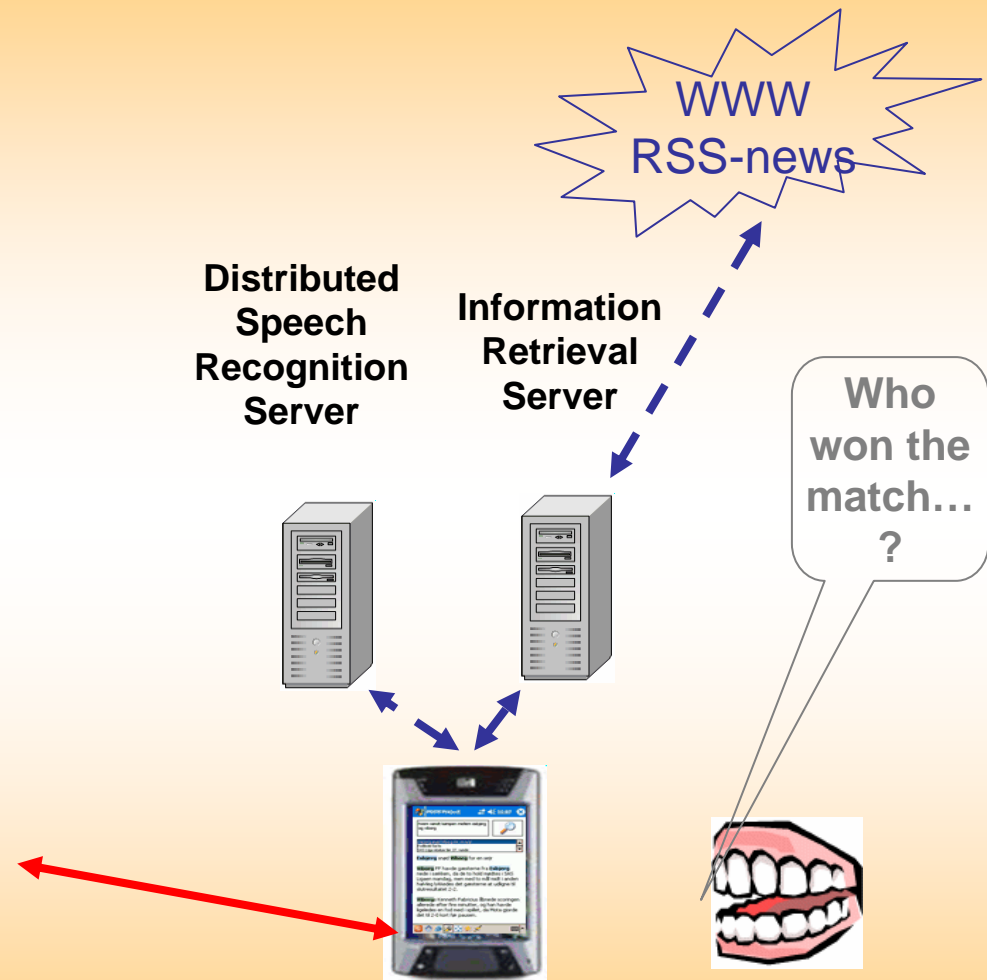
hvem vandt kampen mellem esbjerg og viborg

100% - Esbjerg snød Viborg for en sejr
85% - Esbjerg på vej med klar sejr
50% - SAS Liga-status, 26. spillerunde

Esbjerg snød **Viborg** for en **sejr**

Viborg FF havde gæsterne fra **Esbjerg** nede i sækken, da de to hold mødtes i SAS Ligaen mandag, men med to mål midt i anden halvleg lykkedes det gæsterne at udligne til slutresultatet 2-2.

Viborgs Kenneth Fabricius åbnede scoringen allerede efter fire minutter, og han havde ligeledes en fod med i spillet, da Mota gjorde det



The Soccer news IR demonstrator

Goal: To demonstrate mobile information search using spoken queries and to act as proof-of-concept.

- The demonstrator does not yet use a Trigram language model. Instead we use a rule-grammar with a vocabulary of about 700 words, of which player and team names accounts for 400.
 - The grammar basically recognises typical questions about players, teams and matches:
 - Examples:
 - Who plays for Aalborg?
 - When did Brøndby last play a tie/loose/win?
 - Who was the referee in the match between Aalborg and Viborg
 - Who won the Champions League match between FC Copenhagen and Manchester United?
 - A group of students are currently building an English version incorporating dialogue handling

The Car Rental service

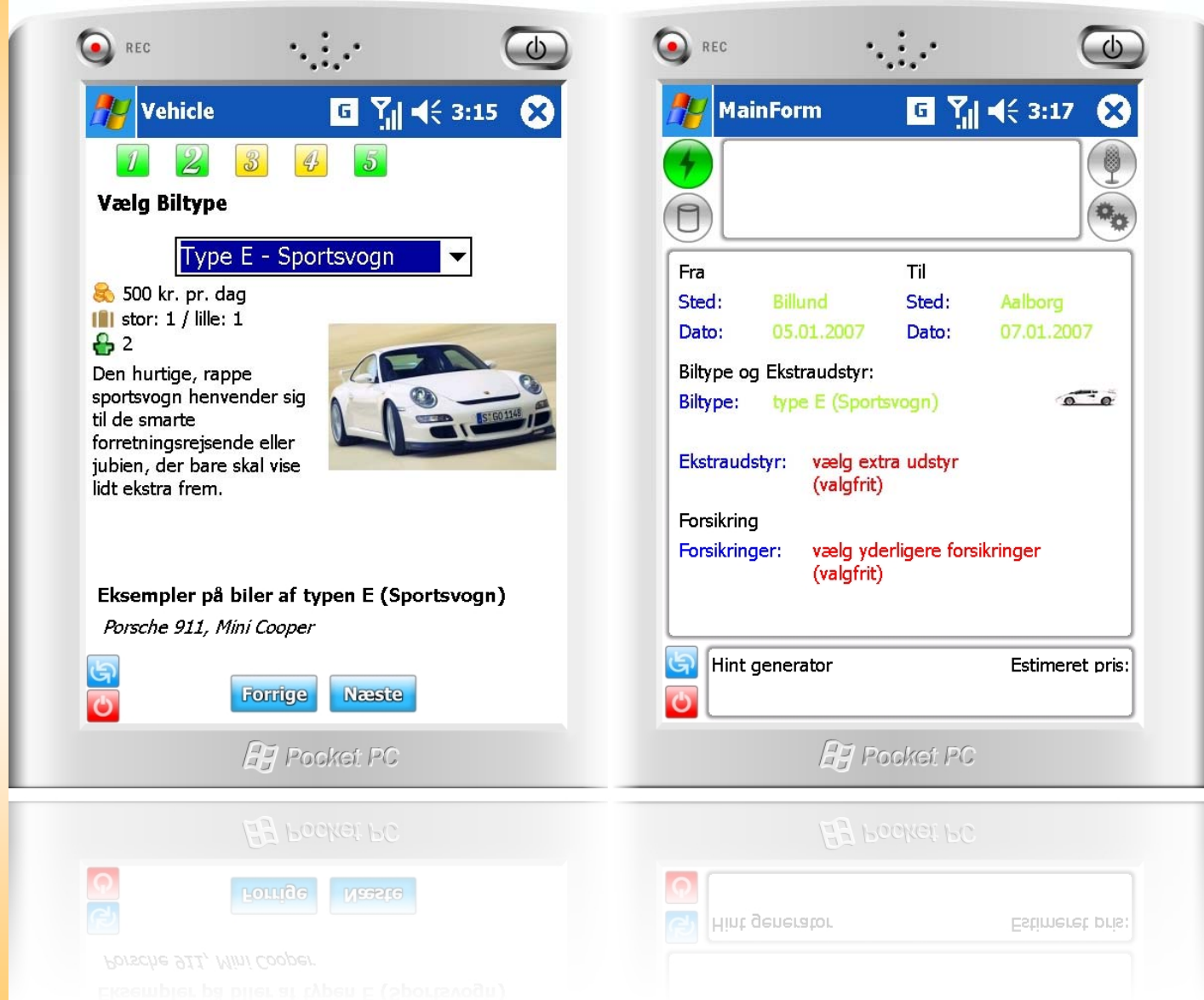
Goal: To enable comparative usability studies of the speech based design vs. a traditional GUI design and evaluate architecture of distributed services

Functions: Enables a travelling user to rent cars in a similar manner as e.g. Avis and Hertz websites

- GUI Design based on interviews and paper prototyping
- Two versions: Speech based and GUI based for comparative testing
- Progressive dialogue handling
- Design not fully finalised for the speech version GUI

Car Rental GUIs

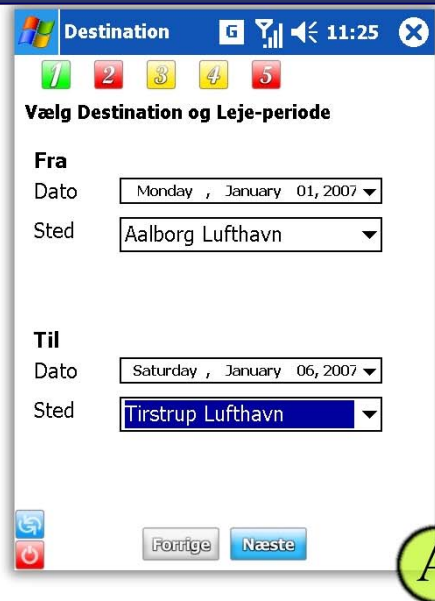
The left-hand screen is the non-speech version, the right-hand side is the speech based version (colour design is not finalised)



Stylus-GUI

The Stylus GUI follows a typical “Wizard-like” progression, guiding the user through the selections.

The design intentionally resembles websites of car rental companies and includes traditional widgets, such as check-boxes, drop-down menus, etc.



Destination 11:25

1 2 3 4 5

Vælg Destination og Leje-periode

Fra
Dato: Monday, January 01, 2007
Sted: Aalborg Lufthavn

Til
Dato: Saturday, January 06, 2007
Sted: Tirstrup Lufthavn

Forrige Næste

A



Vehicle 11:56

1 2 3 4 5

Vælg Biltype

Type E - Sportsvogn

500 kr. pr. dag
stor: 1 / lille: 1
2

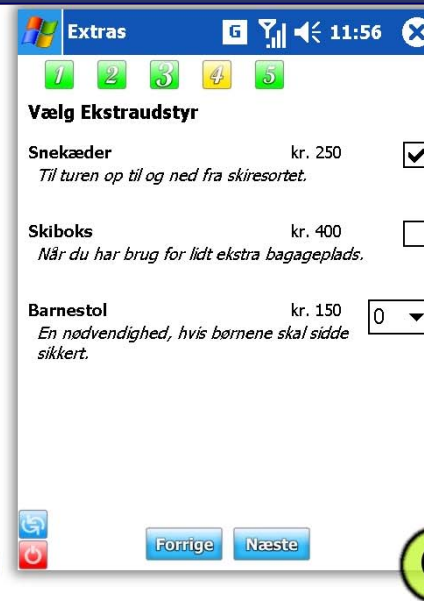
Den hurtige, rappe sportsvogn henvender sig til de smarte forretningsrejsende eller jubien, der bare skal vise lidt ekstra frem.



Eksempler på biler af typen E (Sportsvogn)
Porsche 911, Mini Cooper

Forrige Næste

B



Extras 11:56

1 2 3 4 5

Vælg Ekstraudstyr

Snekæder kr. 250
Til turen op til og ned fra skiresortet.

Skiboks kr. 400
Når du har brug for lidt ekstra bagageplads.

Barnestol kr. 150
En nødvendighed, hvis børnene skal sidde sikkert.

Forrige Næste

C



Insurance 11:56

1 2 3 4 5

Vælg Forsikringer

Kasko kr. 500
En kaskoforsikring.

Ansvar kr. 500
Ansvarsforsikringen er lovpligtig.

Udvidet ansvar kr. 500
Forsikringen for de ekstra uheldige.

Tyveri kr. 500
Et godt tilvalg, hvis vognen parkeres i højrisikoområde.

Forrige Næste

D



OverviewAndCo 11:57

1 2 3 4 5

Overblik og Reservation

Periode
Fra d. 01.01.2007
Til d. 06.01.2007

Destinationer
Afhentning: Aalborg Lufthavn
Afl levering: Tirstrup Lufthavn

Bil: type E - Sportsvogn 500 pr. dag
Eksempler: Porsche 911, Mini Cooper

Valgt ekstraudstyr: 250
Snekæder

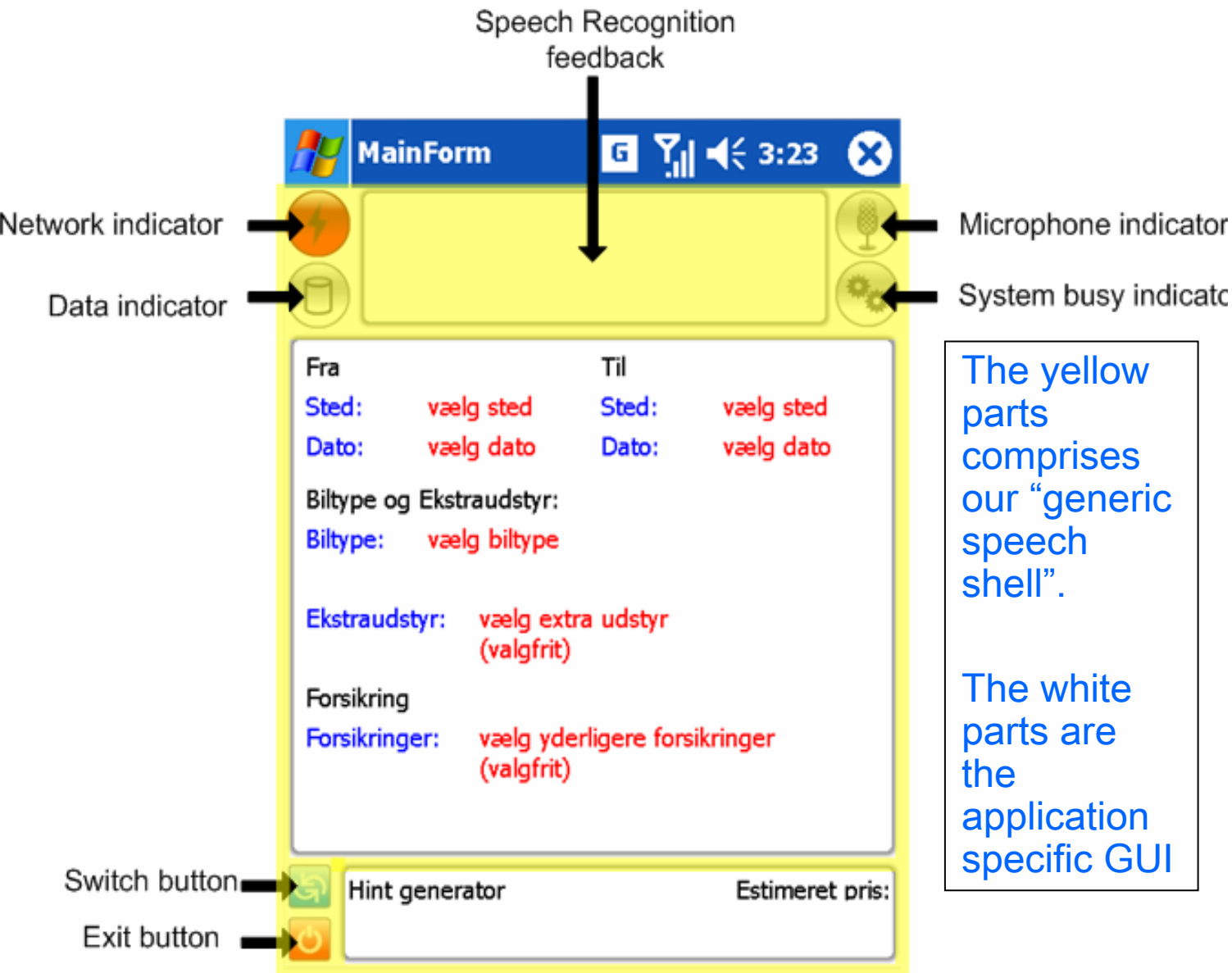
Valgte forsikringer: 1000
Ansvar, Tyveri

Total lejepris 14000

Forrige Næste Lej bil

E

Design of speech GUI



The Speech GUI design intentionally avoids the traditional widgets and affords speech input instead

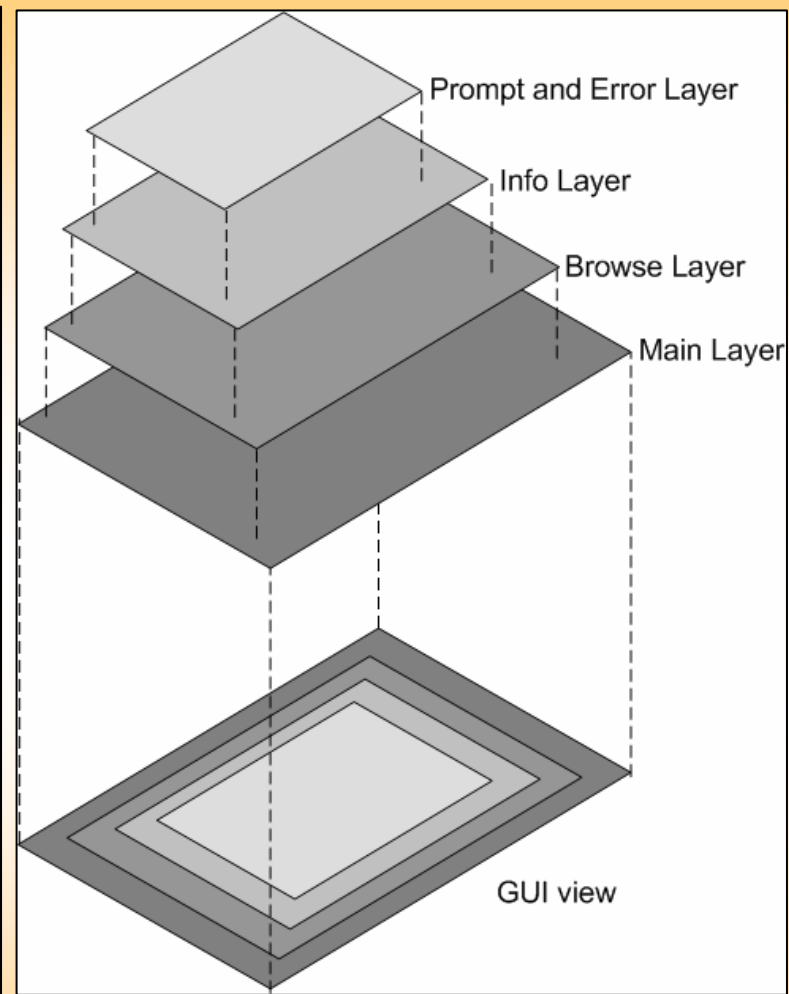
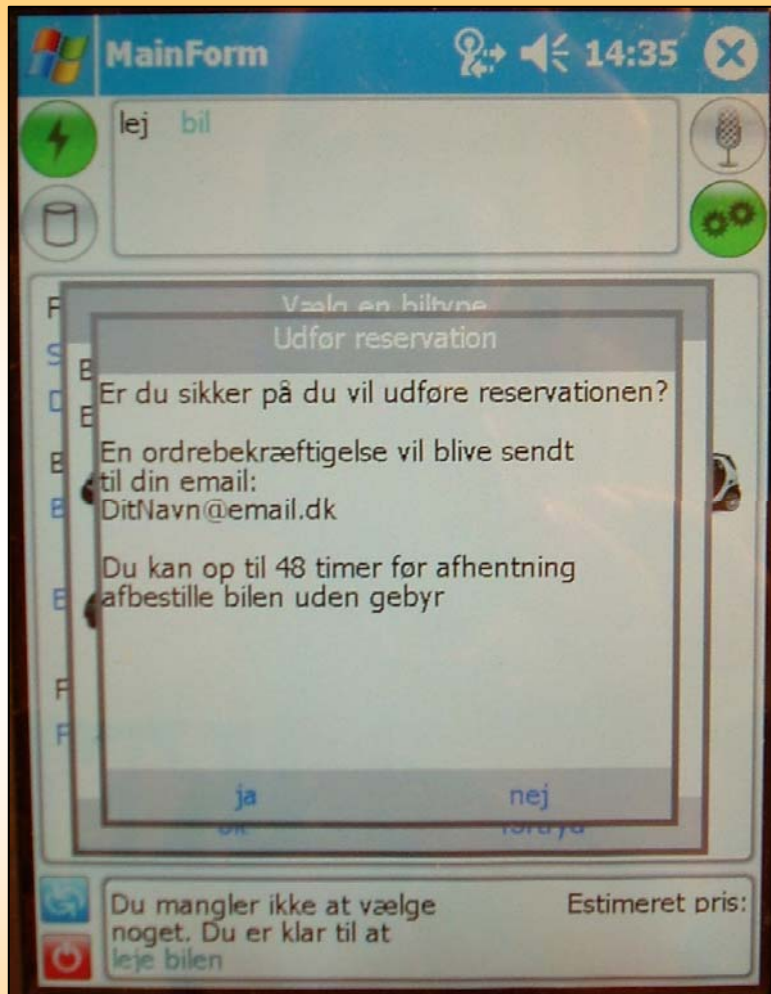
The yellow parts comprises our “generic speech shell”.

The white parts are the application specific GUI

A “hint generator” guides the user and suggest possible next inputs

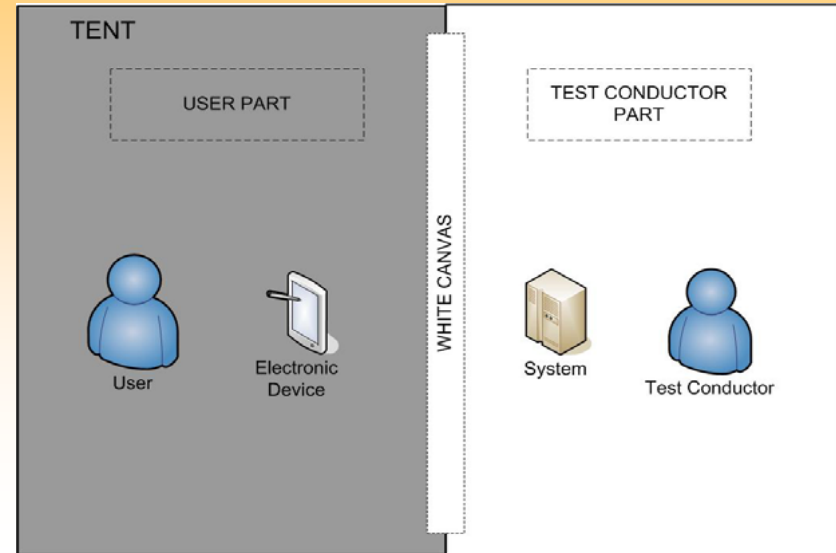
All output is graphical, we do not use TTS

Handling Sub-Dialogues and Confirmations



- We represent sub-dialogues as pop-ups instead of e.g. tabbed panes to force the user's attention on resolving this immediately

Usability lab for mobile services



Idea: To create an immersive environment for “controlled field experiments”

- We simulate different network contexts and user environments by placing the user in a “tent” where s/he is isolated from the real world.
- We simulate physical environments by back projecting Video shot from a “first person perspective” on a wall close in front of the user
- We simulate network context using a heterogeneous wireless network emulator

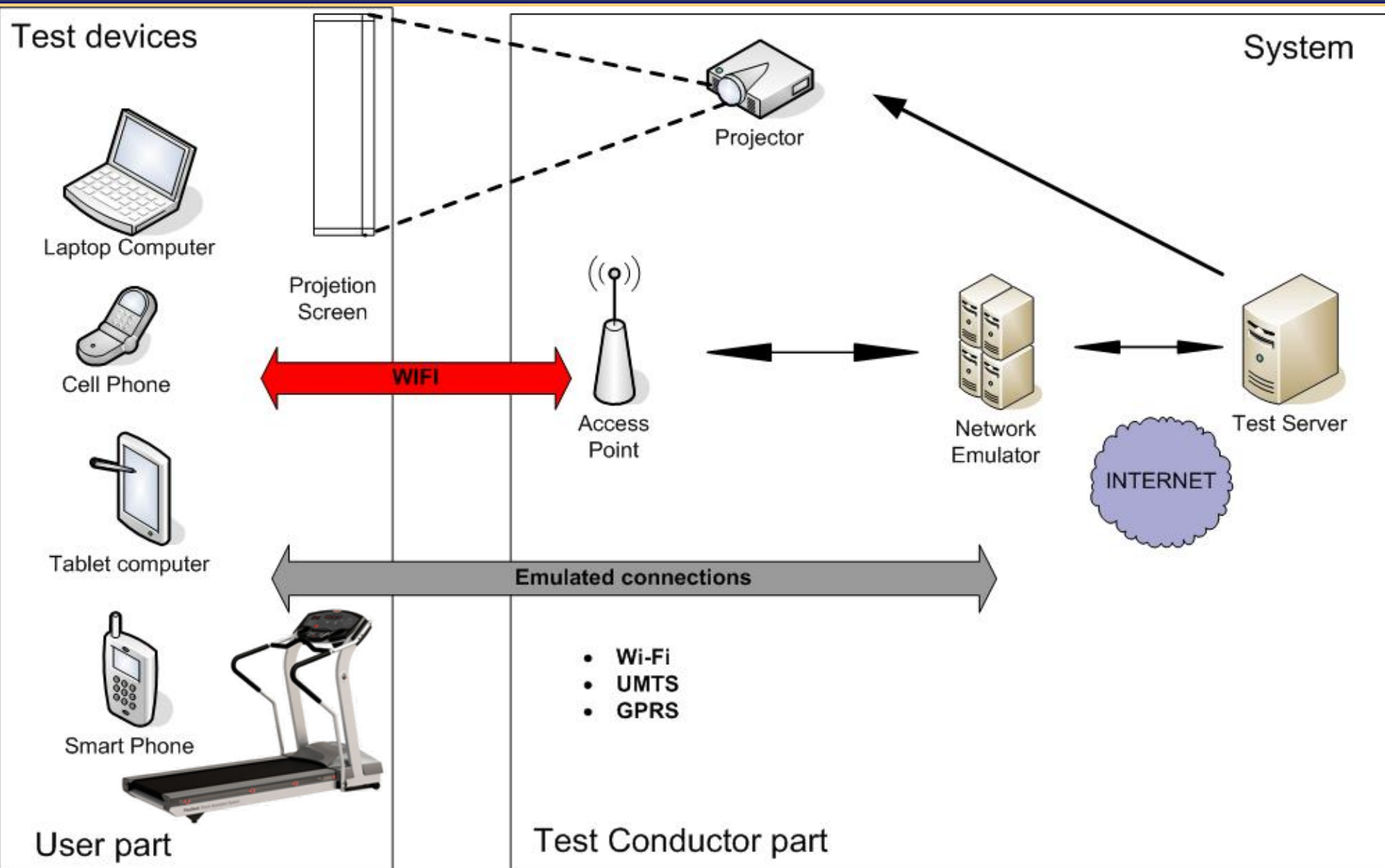
Usability lab for mobile services



Workshop, Tampere, November 2006

Lars Bo Larsen: "Traps and Tricks..."

Architecture



Our Next Steps

Design and conduct end user tests to investigate users' preferences of our design ideas, their performance, etc

- Speech-GUI vs. Stylus-GUI
- The tests will (at least partly) be carried out in the “Tent” to verify this as a viable setup

Strengthen our DSR platform:

- Port front-end to mobile phone (at least ETSI Basic F.E.), either Symbian (e.g. Nokia 60 series) or Win mobile
- May include TTS, support for dialogue engine (VoiceXML), etc.

Summing up

Since the 1980'ies, speech has been predicted an imminent breakthrough.

However, something always seem to get in the way:

- Speech recognition turned out to be much harder than first believed. TTS and speaker identification likewise
- The WWW came along and “stole” the killer applications
- The special usability issues of speech interfaces were not properly understood, even when performance started to become sufficient for realistic services
- Speaking to machines is maybe not as natural as some believed
- Speech needs backup from other modalities if not reduced to very simple applications
- Speech + Graphics seems ideal for small mobile devices, which are becoming ubiquitous and powerful (and not bigger)
- So, maybe this time, finally?