Gefördert vom

bmb+f

Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie

Mensch-Technik-Interaktion

SmartKom

Gestaltung der End–to–End–Evaluation in SmartKom

Version 1.0

Nicole Beringer, Marion Libossek

Ludwig–Maximilians–Universität München

Technisches Dokument Nr. 13

März 2001

März 2001

Nicole Beringer

IPSK LMU München
Schellingstr. 3
80799 München

Tel.: (089) 2180−5751
FAX: (089) 2800362

E−Mail: beringer@phonetik.uni−muenchen.de

Marion Libossek
IPSK LMU München
Schellingstr. 3
80799 München

Tel.: (089) 2180−5753
FAX: (089) 2800362

E−Mail: marion@phonetik.uni−muenchen.de

Dieses Technische Dokument gehört zu Projektfeld II: Anwendungsszenarien

Das diesem Technischen Dokument zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01 IL 905 gefördert. Die Verantwortung für den Inhalt liegt bei den Autoren.

Inhaltsverzeichnis

1 Einleitung....................5
2 Begriffsklärung....................6
2.1 Benutzbarkeit....................6
2.2 Effizienz....................6
2.3 Akzeptanz....................7
2.4 Konzepte....................7
3 Leistungsumfang....................7
3.1 Unterschiedliche Benutzergruppen....................7
3.2 E2E parallel zu WOZ–Aufnahmen....................8
3.3 Verschiedene Modalitäten....................8
4 Aufteilung der Evaluation....................8
4.1 Technische Evaluation (TE): Benutzbarkeit, objektive Effizienz....................8
4.2 Ergonomische Evaluation (EE): subjektive Effizienz, Akzeptanz....................9
5 Was soll evaluiert werden?....................10
5.1 Fragenkatalog zur technischen Evaluation....................10
5.2 Fragenkatalog zur ergonomischen Evaluation....................10
6 Offene Fragen....................15
7 Literatur....................15

1 Einleitung

Als End−to−End−Evaluation definieren wir die »Evaluation der Aufnahme−Prototypen SmartKom Public / Mobile / Home mittels Benutzerexperimenten.« (lt. Projektantrag [SKantrag]).

Aufgabe der E2E−Evaluation ist die Untersuchung der Benutzbarkeit, Effizienz und Akzeptanz der zugrunde liegenden **Konzepte** der Prototypen SmartKom−P / M / H für die multimodale Kommunikation parallel zur Datenerhebung.

Unter Konzept ist hier der Einsatz und die Kombination verschiedener Modalitäten für die Ein− und Ausgabe, nicht jedoch die aktuelle Ausarbeitung der einzelnen Anwendungsfälle zu verstehen.

Die Parallelität der E2E−Aufnahmen zu den normalen Wizardaufnahmen trägt eventuell zu einer Weiterentwicklung der Prototypen aufgrund der Erkenntnisse der Evaluationsexperimente bei.

Um zu gewährleisten, dass sich die VP bei ihrer Bewertung nicht zu stark von Schwächen und (groben) Fehlern der Demonstratoren beeinflussen lassen, werden für die E2E−Experimente Module die nicht funktionieren oder einem noch festzulegenden Ausschlußkriterium »zu« schlecht funktionieren durch Wizardeeingriffe ersetzt: Schnittstellenausgaben der »Zuliefermodule« abfangen, evtl notwendige manuelle (Wizard)−Eingaben veranlassen und so umwandeln, daß die Ergebnisse (wenn möglich) an die (funktionierenden) nachfolgenden Module weitergeben werden können. Zu prüfen bleibt hier noch, inwieweit dies möglich ist.

Um dies zu gewährleisten, gilt es, Module die nicht funktionieren, zu ersetzen (Schnittstellenausgaben der »Zuliefermodule« abfangen, evtl notwendige manuelle (Wizard)−Eingaben veranlassen und so umwandeln, daß die Ergebnisse (wenn möglich) an die (funktionierenden) nachfolgenden Module weitergeben werden können).

2 Begriffsklärung

2.1 Benutzbarkeit

Darunter verstehen wir hauptsächlich das Funktionieren bzw Nicht−Funktionieren der Module und deren Koordination.
Herauszufinden gilt, ob der Benutzer alle Funktionen des Systems aus technischer Sicht nutzen kann.
Methode:
- indirekt durch standardisierte Experimente,
- direkt durch »Trockentest« ohne VP (isoliert bzw in Zusammenarbeit (Schnittstellen))

(s. hierzu Technische Evaluation).

2.2 Effizienz

Die Betrachtung der Effizienz teilt sich auf in:

a) objektive Effizienz:
- Wie leicht / schwer ist das System für den Benutzer zu bedienen?
- Wieviele Fehler / Nachfragen etc. treten bei der Bearbeitung von definierten Standardaufgaben auf?
- Wie verhält sich das System in Bezug auf Reaktionszeiten bei der Bearbeitung der Anfrage / Ausgabe?

Interessant könnte hier auch eine objektive Bewertung des Umgangs des Benutzers mit dem System sein, also z.B. wieviel Off−Talk tritt auf, wenn das System auch auf Äußerungen des Benutzers reagiert, die nicht an das System gerichtet waren oder auch die Reaktion auf Verhalten des Systems bei Barge−In.
Das Auswerten von standardisierten Experimenten gehört ebenfalls dazu.

Für die Bewertung der objektiven Effizienz werden die System−Protokolle der Aufnahmen analysiert.

b) subjektive Effizienz:
- Wie bewertet der Benutzer (über Interview / Fragebogen / evtl Gruppendiskussion) die Arbeit mit dem System?
- Welchen Eindruck hinterläßt die Dateneingabe / Benutzerführung / Datenausgabe?
- Wie werden Wartezeiten / Fehler / Nachfragen aus der Sicht des Benutzers bewertet? (Beispiel: lenkt ein animierter Wichtel von langen Wartezeiten ab?)

Dazu werden Fragebögen und standardisierte Experimente ausgewertet.

Fragebögen sind dynamisch zusammenstellbar, je nachdem welche Module bewertet werden können (da lauffähig) oder sollen (da neue Version vorhanden). Außerdem können die VP natürlich nur für Anwendungen bzw Szenarien, die sie getestet haben, eine Bewertung durchführen.

Es bietet sich an, parallele Tests für die objektive und subjektive Effizienz durchzuführen und danach unter den Aspekten:

- Tatsächliche Leistung des Systems (technisch orientiert) im Gegensatz zur
- empfundenen Leistung (aus ergonomischer Sicht)

auszuwerten.

2.3 Akzeptanz

Darunter verstehen wir die Einstellung der Versuchspersonen gegenüber dem System, NACHDEM sie das System verwendet haben (man könnte hier ja auch zusätzlich davor nachfragen).
Die Informationen werden über Fragebögen / Interviews / Gruppendiskussionen / standardisierte Experimente verschiedene Aspekte (Sicherheit, Privatsphäre, Nutzen) abklären.

Beim Experiment könnte z. B. die Akzeptanz der verschiedenen Biometrie-Methoden (entweder für ein und dieselbe VP oder verschiedene VP-Gruppen) verglichen werden; Akzeptanz der Kamera könnte durch einen Knopf »Kamera aus« auf dem Display getestet werden, wobei die VP in (z. B.) zwei Gruppen geteilt werden, die, denen erklärt wird, wozu die Kamera dient, und die, die diese Information nicht erhalten.

2.4 Konzepte

»Zielstellung:
Im Rahmen dieses Arbeitspakets soll untersucht werden, inwieweit die zugrunde liegenden Konzepte – realisiert in Aufnahme-Szenarien – des Prototypen SmartKom-? von Benutzern aus unterschiedlichen Benutzergruppen (naiv bis Experte) akzeptiert werden (subjektive Evaluation mittels Interview). Ergebnisse sollen den Design-Prozess positiv beeinflussen.«

Was definiert werden muss ist, bis zu welchem Grad das »Konzept« des Systems mit den echten Modulen dargestellt wird,
Hier könnten wir folgende Punke als Anhaltspunkte nehmen:

- Wir können nur vorhandene Module betrachten und diese nur, wenn sie technisch funktionsfähig sind
- Wird das System durch das Verhalten eines Moduls »unbenutzbar« (extrem lange Wartezeiten, Fehler z. B. in mehr als 3 / 4 aller Anfragen / Ausgaben), werden Module ausgeschlossen und durch Wizards ersetzt.

3 Leistungsumfang

3.1 Unterschiedliche Benutzergruppen

Im Idealfall soll die Rekrutierung der Versuchspersonen eine breite Streuung von unbedarften Computerlaien bis hin zu Profis abdecken (Rekrutierungsprofil (in %)).

Die Verteilung der VP sollte auch in Anlehnung an statistischen Daten über die Bevölkerungszusammensetzung Deutschland (zu klären: zu welchem Anteil sollen Ausländer befragt werden) gesteuert werden.

Die LMU wird sich hauptsächlich auf einige mögliche Benutzergruppen konzentrieren (offen für technische Neuerungen, unbedarft gegenüber neuer Technik, Geschäftsleute, Privatpersonen usw).
Für die ergonomische Evaluation des Konzepts können außerdem vorhandene Fragebögen der WOZ−Aufnahmen mit verwendet werden, für spezielle Evaluationsaufnahmen mit einer festen Aufgabe scheint es sinnvoll, neue VP zu nehmen, da die Gefahr besteht, dass durch die verschlechterte Leistung der »Trick« bei den ersten Aufnahmen auffliegt.

3.2 E2E parallel zu WOZ−Aufnahmen

Neben den WOZ−Aufnahmen sollen für die einzelnen Szenarien Evaluationen anlaufen. Wichtig ist, daß mindestens alle relevanten Module für die Ausgabe nicht simuliert werden. Andere Module können simuliert werden. Aber wenn z.B. getestet werden soll, wie gut das System mit einer »echten« synthetischen Stimme ankommt (die u.U. nicht ganz so stark verzerrt ist, wie bei unseren Aufnahmen), kann dies nicht simuliert werden. Des weiteren werden schwere Fehler bei den ersten Schritten der Eingabe die Experimente weitestgehend sinnlos machen, da die meisten Aktionen der folgenden Module dann ebenfalls durch Wizards ersetzt werden müßten. Es wird z. B. nicht möglich sein schwere Fehler in der Spracherkennung abzufangen.
Ansonsten kann in den (neben den) normalen Aufnahmen hauptsächlich die Akzeptanz bewertet werden, da wir sonst nur unsere Wizards bewerten.
Die Ergebnisse − besonders von Benutzbarkeit und Effizienz − werden, da sie im Entwicklungszeitraum der Prototypen entstehen, sofort wieder an die Partner weitergegeben. Somit ist gewährleistet, daß die Module einen unmittelbaren Nutzen daraus ziehen können.

3.3 Verschiedene Modalitäten

Es müssen für die Fragebögen und die objektive technische Auswertung Gestik, Mimik, Akustik berücksichtigt werden (z. B. Adäquatheit der Wichtelausgabe, persönliche Bewertung der Erleichterung bzw des Komforts bei der Bedienung durch verschiedene Modalitäten)

4 Aufteilung der Evaluation

»Dabei sollen neben technischen Aspekten (Funktionalität, Effizienz) auch ergonomische Gesichtspunkte (Bedienbarkeit, intuitive Bedienung, etc.) untersucht werden.«

Die End−to−End−Evaluation wird auf folgende zwei Bereiche aufgeteilt.
Als Grundlage der Evaluation wird das Testbed mit den zu verschiedenen Zeiten verfügbaren Modulen verwendet.

4.1 Technische Evaluation (TE): Benutzbarkeit, objektive Effizienz

Bei der technischen Evaluation wird die Funktionalität des Systems in ihren Einzelheiten (s. Kapitel 5) untersucht. Ist das System funktionstauglich, wenn nicht, bei welchen Modulen

gibt es Probleme? (Benutzbarkeit) Daneben werden objektiv bewertbare technische Daten des Systems (Wie schnell, wie viele Fehler, Echtzeitfaktor bei der Spracherkennung usw.) ausgewertet (objektive Effizienz) bzw. zur Bewertung an geschulte Experten der ergonomischen Evaluation weitergegeben.
Die TE teilt sich wiederum in 3 Bereiche:

1. »Trockentest«
   Für die Effizienz der E2E−Evaluation ist es wichtig, daß die jeweiligen Endergebnisse aller Module bewertet werden. Das soll allerdings nicht heißen, daß die einzelnen Module auf spezifische Fehler hin analysiert werden, sondern, daß die Kommunikationshäufigkeit bzw. −erfolge (nur »hat funktioniert«, »hat nicht funktioniert, weil Schleife / tot«) analysiert wird.
2. Evaluation der Schnittstelle Mensch − System
   Das betrifft alle Module oberhalb der Linie Interpretation/Analyse − Generierung (s. Graphik). Sie werden bezüglich ihrer Benutzbarkeit und Effizienz analysiert.
3. Eine Konsistenz der Beurteilungen durch die Experten wird durch die Bereitstellung der Daten durch Wizards in einem Auswahlkatalog mit vorgegebenen Evaluationskriterien erreicht.
   Zu klären ist, inwieweit Module simuliert werden sollen! Sollen Fehler bewußt produziert werden?

4.2 Ergonomische Evaluation (EE): subjektive Effizienz, Akzeptanz

Die ergonomische Evaluation bewertet den subjektiven Bereich der Effizienz, sowie die Akzeptanz und wird dies über Benutzerexperimente und −interviews durchführen.

Dabei sollen die VPn das System anhand der Ausführung einer / zwei Aufgaben lösen. Die Bewertung erfolgt anschließend einmal allgemein für die Interaktion der VP mit dem System (Eingabemodule / Ausgabemodule / Biometrie) − somit wird hauptsächlich das Konuzpt bewertet − und zum anderen soll kurz die ausgeführte Anwendung evaluiert werden. Die Bewertung der Anwendungen wird knapp gehalten, da diese prinzipiell »austauschbar« sind und Smartkom hauptsächlich eine verbesserte MM−Kommunikation bieten möchte.

- Subjektive Effizienz:
  Wie bewertet die VP die Effektivität des Gerätes zur Lösung von bestimmten Problemen insgesamt?
  Wie werden die Eingabemedien bzw Ausgabemedien, sowie der Dialog MM allgemein bewertet?
  Wie bewerten die VP die Leistung speziell für den durchgeführten Fall? (Anwendungen, Szenario)
- Akzeptanz:
  Was ist die Einstellung der VP dem System gegenüber allgemein?
  Wie werden die Eingabemedien bzw die Ausgabemedien, sowie die verschiednen Methoden der Biometrie empfunden?

Zu welchen Personen würde dieses System passen, bzw wie sehen die VP das System (z. B. durch ein Übertragen auf Fahrzeugtypen)

Beim Interview gelten grundsätzlich dieselben Aspekte wie beim Fragebogen, allerdings sind einige Punkte besser in einem Fragebogen zu beantworten (kommt die VP gut mit dem System zurecht oder nicht), während andere Punkte durch ein persönliches Interview besser ausgeleuchtet werden können.

5 Was soll evaluiert werden?

5.1 Fragenkatalog zur technischen Evaluation

- Welche Module funktionieren wie gut? (Trockentest & Schnittstelle Mensch−Module)
- Bearbeitungszeiten
- Interaktionswege (mit welchem Modus antwortet das System auf verschiedene Eingabemodi des Anwenders):
  - Spracheingabe −> Sprachausgabe
  - Gesteneingabe −> Sprachausgabe
  - Spracheingabe −> Displayausgabe
  - Gesteneingabe −> Displayausgabe
  - usw

- Fehleranfälligkeit / Schwere der Fehler
- Frequenz der wiederholten Anfragen (wegen System, wegen User)
- Frequenz von erkanntem Offtalk
- Frequenz von Barge−In
- Bearbeitungszeiten von Eingabe bis Ausgabe
- Welche Hilfestellung bietet das System dem Anwender?
- Korrektheit der Ausgaben: Überwiegend falsche / korrekte Antworten (Oberfläche)

5.2 Fragenkatalog zur ergonomischen Evaluation

I. Effizienz:

e1. Allgemeines

1.1 **Allgemeine Effizienz des Systems:** Unterstützt die Software die Erledigung Ihrer Arbeitsaufgaben, ohne Sie als Benutzer unnötig zu belasten? (FRAG / INT)

1.2 **Selbstbeschreibungsfähigkeit:** Gibt Ihnen die Software genügend Erläuterungen und ist sie in ausreichendem Maße verständlich?

1.3 **Schnelligkeit des Systems** (FRAG / INT / TECH)

1.4 **Erlernbarkeit:** Ist das System so gestaltet, dass Sie sich ohne grossen Aufwand einarbeiten konnten und bietet es auch dann Unterstützung, wenn Sie neue Funktionen lernen möchten? (FRAG / INT)

1.5 **Intuitivität:** Kommt die Software durch eine einheitliche und verständliche Gestaltung Ihren Erwartungen und Gewohnheiten entgegen?(FRAG / INT)

1.6 Fehler

1.6.1 Fehlermeldungen (hilfreich, verständlich, verwirrend) (FRAG)

1.6.2 Fehleranfälligkeit (wie oft treten Fehler auf?) (FRAG / TECH)

1.6.2.1 Fehlertoleranz: Bietet Ihnen die Software die Möglichkeit, trotz fehlerhafter Eingaben das beabsichtigte Abeitsergebnis ohne oder mit geringem Korrekturaufwand zu erreichen?

1.7 **Hilfestellung**

1.7.1 Menge ok? (FRAG / TECH)

1.7.2 Effizienz (haben sie geholfen) (FRAG)

1.7.3 Hilfeangebot

e2. **Eingabe**

2.1 Allgemein

2.2 **Sprache**

2.3 **Gestik**

2.4 **Mimik / Emotionskamera**

2.5 **Bewertung je nach Anwendung** (die Anwendungen, die die VP im Experiment verwendet hat insgesamt in bezug auf die Eingabemedien bewerten) (FRAG / INT)

2.6 **Bewertung je nach Szenario** (das Szenario, das die VP im Experiment verwendet hat insgesamt in bezug auf die Eingabemedien bewerten) (FRAG / INT)

e3. **Ausgabe**:

3.1 **Allgmein**

3.2 **Sprache**

3.3 **Display**

3.4 Wichtel

3.5 **Bewertung je nach Anwendung** (die Anwendungen, die die VP im Experiment verwendet hat insgesamt in bezug auf die Eingabemedien bewerten) (FRAG / INT)

3.6 **Bewertung je nach Szenario** (das Szenario, das die VP im Exeriment verwendet hat insgesamt in bezug auf die Eingabemedien bewerten) (FRAG / INT)

e.4 **Dialog**

4.1 Das System versteht mich (INT)

4.2 Ich verstehe das System (INT)

4.3 Das System geht auf mich ein (ändert seine Ausgabe) (INT)

4.4 Natürlichkeit des Dialogs (ich kann unterbrechen) (INT)

4.5 Überwiegend falsche / korrekte Antworten (FRAG / INT)

4.6 Fehlreaktionen des Systems (z. B. wegen Off-Talk) (FRAG / INT)

II **Akzeptanz**

a1. **Allgemein**

1.1 Gesamteindruck: wie gefällt das System? (FRAG)
1.2 Würde so ein System verwendet werden (allgemein)? (FRAG)
1.3 Für welche Anwendungen würde so ein System verwendet? (FRAG / INT)
1.4 Für welche Szenarien würde so ein System verwendet? (FRAG / INT)
1.5 Wie wird der Neuheitsgrad des Systems eingeschätzt? (FRAG)
1.6 Wieviel würde man für so ein System ausgeben? (FRAG)
1.7 Wie fühlt man sich bei der Bedienung? (FRAG / INT)
1.8 Wie »menschlich« wirkt das System (FRAG / INT)
  1.8.1 Speziell durch welche Elemente (FRAG / INT)

a2. **Eingabemedien**

2.1 Spracheingabe
2.2 **Gestikeingabe**
2.3 Stifteingabe
2.4 **Mimik / Emotionskamera**

a3. Ausgabemedien

3.1 **Sprachausgabe**
3.2 **Displayausgabe**
3.3 Wichtelausgabe

a4. Biometrie (FRAG)

4.1 allgemein: erwünscht / unerwünscht
4.2 Unterschrift
4.3 Stimmidentifikation
4.4 Handkontur
4.5 Weitere Methoden?

a5. Benutzergruppen (Für welche Gruppen denken die VP ist SK besonders geeignet?)

5.1 geschätztes Alter der Hauptanwender (FRAG / INT)
5.2 Geschlecht der Hauptanwender (FRAG / INT)
5.3 Nationalität (FRAG)
5.4 Erfahrenheit im Umgang mit Computern (FRAG / INT)
5.5 Berufsgruppen (FRAG / INT)
5.6 Wäre SK ein X (z. B. Auto), welches? (FRAG)

6 Offene Fragen

- »Zusätzlich werden die Datenspender in eine detailierte Profildatenbank aufgenommen, die für die Evaluation von grundlegenden Biometrie−Experimenten erforderlich wird.«
  - Inwieweit sind Biometrieexperimente grundlegend für die Evaluation?
  - Ist damit die Forschung bei Siemens oder die Datensammlung gemeint?

- Inwieweit sollen Module von Wizards ersetzt werden? Fehlerproduktion?
- Wie oft soll evaluiert werden? (Vorschlag: Prototyp und laufende Integration)
- Wie viel soll evaluiert werden? (Vorschlag: 10 – 15 Vps pro Meilenstein)
- Wie detailliert sind die Logfiles?

7 Literatur

[SKantrag] DFKI et al.: Leitprojekt SmartKom – Dialogische Mensch–Technik–Interaktion durch koordinierte Analyse und Generierung multipler Modalitäten

[E2EVM00] S.Jekat, L. Tessiore: End–to–End Evaluation of Machine Interpretation Systems: A Graphical Evaluation Tool. In: Proceedings of LREC 2000, Athen Mai 2000.