Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Auswertung der End–to–End Evaluation des SmartKom Prototypen (technische Evaluation, Evaluationstool, PROMISE)

–Szenario Mobil–

Aufnahmezeitraum 01.August 02 bis 31. August 02

N. Beringer, H. Mögele, I. Jacobi, C. Beiras–Cunqueiro, L. Wang, U. Türk, G. Kouam–Wotchung, M. Mozul, U. Kartal

Ludwig–Maximilians–Universität München

Memo Nr. 15
Februar 2003

Februar 2003

Auswertung der End–to–End Evaluation des SmartKom Prototypen (technische Evaluation, Evaluationstool, PROMISE)

–Szenario Mobil–

Aufnahmezeitraum 01.August 02 bis 31. August 02

N. Beringer, H. Mögele, I. Jacobi, C. Beiras–Cunqueiro, L. Wang,
U. Türk, G. Kouam–Wotchung, M. Mozul, U. Kartal

Ludwig Maximilans Universität München
Schellingstr. 3
80799 München

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

E–Mail: smart–eval@phonetik.uni–muenchen.de

Dieses Memo gehört zu Teilprojekt 6: Intuitives Arbeiten mit dem Computer: SmartKom Mobil; AP 6.5 Ergonomische Evaluation

Das diesem Memo 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.Evaluationszeitraum....4
2.Anzahl und Art der Aufnahmen....4
3.Auswertungskriterien....4
4.Ergebnisse der technischen Evaluation....9
1.Erkennerergebnisse....9
2. Barge In/Offtalk....10
3. Hilfefunktion....10
4. Syntax....11
5. Semantik....11
6. Dialog....11
6.2. Anzahl 11
6.3. Kompetenz/ Natürlichkeit....12
6.4. Dialogverhalten....12
6.5.Modalitätenwechsel....13
6.6.Verarbeitungszeit....13
7. Ergebnisse der vergleichenden Evaluation TE vs. EE....14
8. Performanz nach PROMISE....14
5.Fazit....15

1. Evaluationszeitraum

Die End–to–End–Evaluation für das Szenario Mobil umfaßt den Zeitraum vom 01.08.02 bis einschließlich 31.08.02. Dieses Memo ergänzt Memo 14 zur Ergonomischen Evaluation (EE) und beschreibt die Ergebnisse der Technischen Evaluation (TE), der vergleichenden Evaluation (TE vs. EE) sowie die PROMISE–Performanz des SmartKom Prototypen Mobil.

2. Anzahl und Art der Aufnahmen

Von den 18 End–to–End–Evaluationsaufnahmen des SmartKom Prototypen Mobil konnten alle technisch evaluiert werden. 17 Aufnahmen konnten vergleichend evaluiert und die PROMISE –Performanz berechnet werden. Eine der 18 Sessions enthielt einen leeren Fragebogen. Die folgenden Daten beziehen sich damit auf 18 (TE) bzw. 17 (TE vs. EE) gültige Sessions zur Evaluation des Mobil–Prototypen (Demonstrator 3.1).

Naivität der Versuchspersonen

Entgegender ursprünglichen Planung der begleitenden Evaluation wurden neben 12 naiven Versuchspersonen auch 6 (TE) bzw. 5(TE vs. EE) ‚davon eine VPn doppelt, nicht naive evaluiert. Da es auch mit den nicht naiven VPn zu massiven technischen Problemen während der Aufnahmen kam (kein Bild, falsches Szenario, Hängenbleiben der Dialogverarbeitung), mußte die Mobil–Evaluation vorzeitig abgebrochen werden.

Technisches Setup

Alle 18 End–to–End–Aufnahmen erfolgten mit dem Demonstrator 3.1. Die Audioeingabe erfolgte mit dem Sennheiser–Head–set und Silence–Detection.

Sonstige Abweichungen der End–to–End Aufnahmen zum Beta–Test der Systemgruppe

Im Gegensatz zu Demo–Situationen und dem Beta–Test der Systemgruppe hat der Benutzer **keine** Möglichkeit, den Verarbeitungslauf des Systems parallel nachzuvollziehen. Die VPn bekommen ausschließliches Feedback vom System über die Graphik–/Sprachausgabe. Wie die Aufnahmen mit dem Prototyp Mobil gezeigt haben, kommt es durch das Fehlen eines sichtbaren Verarbeitungszustands oft zu Mehrfacheingaben durch die VP. Das hat zur Folge, daß der User sich sozusagen selbst überholt und dementsprechend verwundert ist, wenn auf seine Anfragen völlig bzw. zum Teil unpassende oder überhaupt keine Antworten gegeben werden. Es ist nicht nachvollziehbar, warum das System von sich aus bei fast allen Aufnahmen einen Szenariowechsel (Beispiel: System: „Wollen Sie das Fernsehprogramm sehen?“) vornahm, ohne diesen jedoch durchführen zu können.

3. Auswertungskriterien

Die nachfolgende Tabelle gibt einen Überblick über die zur TE–Evaluation vorgesehenen Kriterien und ihrer Bedeutung.

Kriterium	***Bedeutung***
Spracherkennung	Akkuratheit der Spracherkennung
Gestikerkennung	Akkuratheit der Gestikerkennung

Kriterium		***Bedeutung***
Mimikerkennung		Akkuratheit der Mimikerkennung
Prosodieerkennung		Akkuratheit der Prosodieerkennung
Userstateerkennung		Akkuratheit der Userstateerkennung
Barge–In		Anzahl der gelabelten Bargelns Prozentualer Anteil zum Gesamt–User–Audio
BargeInVorkommen		
Prozentualer Barge In Anteil		
Offtalk		Anzahl aller als Offtalk definierten Events Prozentualer Anteil zum Gesamt–User–Audio
Offtalk Vorkommen		
Prozentualer Offtalk Anteil		
Komplettabbruch		Gesamtzahl aller vom System erkannten Abbrüche Übereinstimmung mit den gelabelten Abbrüchen Gesamtzahl aller nicht vom User intendierten Abbrüche
Gesamtzahl erkannt		
Übereinstimmung		
Anz. nicht gesteuerter Abbrüche		
Syntheseabbruch		Nur Sprachausgabe: Gesamtzahl aller vom System erkannten Abbrüche Übereinstimmung mit den gelabelten Abbrüchen Gesamtzahl aller nicht vom User intendierten Abbrüche
Gesamtzahl erkannt		
Übereinstimmung		
Anz. nicht gesteuerter Abbrüche		
Displayabbruch		Nur Graphikausgabe: Gesamtzahl aller vom System erkannten Abbrüche Übereinstimmung mit den gelabelten Abbrüchen Gesamtzahl aller nicht vom User intendierten Abbrüche
Gesamtzahl erkannt		
Übereinstimmung		
Anz. nicht gesteuerter Abbrüche		
Keyword nicht erkannt		Abbruchwunsch des Users
No.of senses		Anzahl der Mehrdeutigkeiten pro Satz /pro Eingabe
Sentence	Single input	

<table>
<tr><th colspan="4">Kriterium</th><th>Bedeutung</th></tr>
<tr><td colspan="4">Anzahl erfolgreiche TRL–Turns</td><td rowspan="6">Anzahl und prozentuale Verteilung von Audio– und Gestikeingaben gesamt und im Hinblick auf Erfolg.</td></tr>
<tr><td colspan="4">Anzahl TRL–Turns</td></tr>
<tr><td colspan="4">Gestikeingaben im File</td></tr>
<tr><td colspan="4">Prozentualer Anteil erfolgreicher Gestikeingaben</td></tr>
<tr><td colspan="4">Prozentualer Anteil erfolgreicher TRL</td></tr>
<tr><td colspan="4">Prozentualer Anteil erfolgreicher Inputs allgemein</td></tr>
<tr><td colspan="4">Fehlermeldungen</td><td>Fehlermeldungen des Systems</td></tr>
<tr><td colspan="4">Anzahl der Hilfestellungen pro Aufnahme</td><td>Anzahl der sichtbaren Hilfeversuche der dynamischen Hilfe</td></tr>
<tr><td colspan="4">Effizienz der Hilfestellung</td><td>Führt die dyn. Hilfe zu besserem Erfolg?</td></tr>
<tr><td colspan="4">doppeldeutige Ausgaben</td><td>ambige Bedeutung paralleler Ausgaben</td></tr>
<tr><td colspan="4">doppeldeutige Eingaben</td><td>ambige Bedeutung paralleler Eingaben</td></tr>
<tr><td colspan="4">Komplexität der Gestik Input</td><td rowspan="9">Die Komplexität von Audioeingabe, Gesten, Sprachausgabe, und Graphikausgabe wird durch die Länge und Häufigkeit der Modalität gegeben. Je länger eine Ein–/Ausgabe bzw. je mehr parallele Ein–/Ausgaben stattfinden, desto komplexer wird der Dialog.</td></tr>
<tr><td colspan="2">%USR–Turns</td><td colspan="2">%USR–Gesten</td></tr>
<tr><td colspan="4">Komplexität der Grafik Dialogue Manager</td></tr>
<tr><td colspan="2">%SMA–Ausgabe</td><td colspan="2">%SMA–Grafiken</td></tr>
<tr><td colspan="4">Komplexität der Sprache</td></tr>
<tr><td colspan="2">Dialogue Manager</td><td colspan="2">Input</td></tr>
<tr><td>%SMA–Turns</td><td>SMA–Words</td><td>%USR–Turns</td><td>%USR–Words</td></tr>
<tr><td colspan="4">GES vs Grafikturns</td><td rowspan="6">Anzahl der sich zeitlich überlappenden Modi.</td></tr>
<tr><td colspan="4">GES vs SMA–Audio</td></tr>
<tr><td colspan="4">GES vs USR–Audio</td></tr>
<tr><td colspan="4">SMA–Audio vs Grafikturns</td></tr>
<tr><td colspan="4">USR vs SMA–Audio</td></tr>
<tr><td colspan="4">USR–Audio vs Grafikturns</td></tr>
</table>

Kriterium		***Bedeutung***
Wechsel von GES nach SMA–Audio		Wechsel von einem Modus in einen anderen.
Wechsel von GES nach USR–Audio		
Wechsel von Grafik nach SMAAudio		
Wechsel von Grafik nach USRAudio		
Wechsel von SMA–Audio nach GES		
Wechsel von SMAAudio nach Grafik		
Wechsel von USR–Audio nach GES		
Wechsel von USRAudio nach Grafik		
Semantical correctness		Ist Input/Output passend zum vorangehenden Turn bzw. zum Anwendungsfall?
of input	of output	
Output "nicht verstanden"		Anzahl der "nicht verstanden"–Antworten
Syntactical correctness		Gibt es syntaktische Fehler in den Turns?
of input	of output	
Elapsed time Mimikerkennung		Dauer der Mimikerkennung
Elapsed time "Mimikeingabe"		Dauer der Mimikeingabe
Dialogdauer		Dialogdauer
Objekt–Manipulationen		z.B. Drag–and–Drop–Aktionen
Mean user response time		Mittlere Antwortzeit des Benutzers / Systems (Ende Eingabe–User/Ausgabe–System bis Beginn Ausgabe–System/ Eingabe–User)
Mean system response time		
Appropriate diag. utterances		Zur Dialogsituation passende/unpassende Äußerungen des Systems
Inappropriate diag. utterances		
Explicit recovery answers		Explizite (z.B. „Könnten Sie den Film wiederholen?“) /implizite (z.B. „Hier sehen Sie die Sehenswürdigkeiten von Aachen und Aalen“) Reparaturversuche des Systems.
Implicit recovery answers		
Dialog		system–, usergeführt oder ausgewogen
Wortabbrüche		Abgebrochene Wörter bei der Eingabe
Wiederholungen		Wörtliche Wiederholungen
Unverständlich		unverständliche Eingaben
Incremental Compatibility		sinngemäße Wiederholungen
Cooperativity user		Wie gut geht der jeweilige Dialogpartner auf sein „Gegenüber“ ein?
Cooperativity system		

Kriterium	***Bedeutung***
Gesamtzahl Audio–Turns	Gesamtzahl der verschiedenen Ein–/Ausgaben auf Tunbasis
Gesamtzahl User–Audio–Turns	
Gesamtzahl System–Audio–Turns	
Gesamtzahl Gestikturns	
Gesamtzahl Grafikturns	
Fehlerrate von Anfragen	Anzahl falscher Anfragen
Fehlerfrequenz von Antworten	Anzahl falscher Antworten
Widersprüche	bei Ein–/Ausgaben (parallel/sequentiell)
Gesamtzahl der Wörter im File	Gesamtzahl aller Wörter
Synchrone Gestik	Synchronität der Ausgabe (sprachbegleitende Gestik des Wichtels, Lippensynchronität)
Synchrone Lippenbewegungen	
Überlappungen	Anz. der sich überlappenden Graphik und Synthese
Zeitverschiebungen	Asynchrone Ausgaben
Keywords in USR–Turns	Taskspezifische Keywords bei Benutzer und System sowie aufgegriffene Keywords des jeweiligen Dialogpartners.
S–Keywords in USR–Turns	
Keywords in SMA–Turns	
USR–Keywords in SMA–Turns	
USR–Keywords in SMA–Turns Liste	
Task Complexity	Die Komplexität eines Anwendungsfalls (meist dasselbe wie Dialogkomplexität, da nur ein Anwendungsfall pro Session vorhanden)
% aller Turns einer Task	
% aller Wörter einer Task	
% aller Gesten einer Task	
% aller Grafiken einer Task	
Task Completion Time	Die Zeit bis zur Beendigung der Task
Prosodie	Anzahl der gelabelten prosodischen Grenzen und Akzente.
Accent	
Boundary	
SMA / USR	Gesamtzahl aller Wörter pro Turn/pro Session pro Dialogteilnehmer
Anzahl der Wörter im Turn	
Gesamtzahl	

Kriterium	*Bedeutung*
Elapsed time Gestikeingabe	Dauer der Gestik–/Spracheingabe/–erkennung.
Elapsed time Spracheingabe	
Elapsed time Gestikerkennung	
Elapsed time Spracherkennung	

Für den Demonstrator 3.1 (Szenario Mobil) waren folgende Funktionalitäten nicht vorgesehen, wurden aber zum Teil ergonomisch erfragt bzw. konnten während der Aufnahmen beobachtet werden:

Objekt–Manipulationen, Mimikerkennung (Userstateerkennung), Elapsed time Mimikerkennung, Elapsed time "Mimikeingabe", Barge–In, Offtalk, Abbrüche, Synchrone Lippenbewegungen. Diese werden an entsprechender Stelle eingeschränkt ausgewertet. Aufgrund des mobilen Einsatzes der Oberfläche konnte keine Gestikannotation als Referenz erstellt werden.

Die Hilfefunktion ist für TE und EE unterschiedlich. Wichtig hierbei ist aufzuzeigen, daß der Benutzer explizite Hilfe wünscht und die durch das System bereitgestellte Hilfe in der Regel nicht erkannt wird.

4. Ergebnisse der technischen Evaluation

1. Erkennerergebnisse

	Spracherkennung	Prosodie (% | Anzahl | Anzahl)		
	%	Erkennung	Accent(Anzahl)	Boundary(Anzahl)
Max	85.48	0	61	420
Min	–84.31	–1720	6	42
Median	37.03	–123.42	36	112
Mittelwert	32.21	–321.41	33.94	159.47

Im Vergleich zum Home–Prototypen erreicht die Spracherkennung einen Mittelwert von 32.21% über alle 18 Sessions, wobei die maximale Erkennerleistung bei 85.48% also knapp 28% über dem Prototypen Home liegt. Negative Werte (MIN) lassen sich dadurch erklären, daß es sich bei allen Versuchspersonen, naiv wie nicht naiv, nicht um SmartKom–geschulte Benutzer handelt, d.h. die Versuchspersonen weichen, wenn auch geringfügig, von den vorgegebenen Demodialogen ab oder machen zusätzliche, das System irritierende Geräusche (z.B. durch Blätterrascheln, Spielen mit dem Eingabestift) bzw. durch die allgemeinen Probleme in der Dialogverarbeitung (Selbsterkennung des Systems, Fehlinterpretation des Umgebungsgeräuschs etc.).

Die Gestik konnte aufgrund schlechter szenariobedingter Labelbedingungen und daher fehlender Referenzannotationen nicht bzgl. ihrer Akkuratheit bewertet werden.

Auch bei der Prosodieerkennung ist insgesamt eine Verschlechterung zum

Prototypen Home erkennbar.

2. Barge In/Offtalk

	Bargeln Vorkommen	Prozentualer Bargeln Anteil (Session)	Offtalk Vorkommen	Prozentualer Offtalk Anteil (Session)
Max	48	11.11%	90	20.92%
Min	0	0.00%	0	0.00%
Median	6	2.00%	7	2.00%
Mittelwert	11.35	2.94%	15.29	4.77%

Wie aus den Werten der Tabelle zu Bargeln/Offtalk ersichtlich, hält sich der Mittelwert für beide Kriterien in Grenzen, er liegt aber um ca. 50 – 100% über dem jeweiligen Home–Wert. Durch die gesteuerte Versuchssituation ist auch der maximale Anteil an Barge In sehr gering, da die VPn angehalten wurden, die Systemausgaben / Eingabeaufforderungen abzuwarten.Der relativ hohe Anteil an Offtalk (MAX: 20.92%) ist allerdings zu beachten, da das System diesen Offtalk als Spracheingabe interpretieren könnte und somit nicht nur Spracherkennung und Prosodieerkennung, sondern die gesamte Weiterverarbeitung in ihrer Performanz leidet.

Abbrüche						
Komplett			Synthese		Display	
	Gesamtzahl erkannt	Anz. nicht gesteuerter	Gesamtzahl erkannt	Anz. nicht gesteuerter	Gesamt–zahl erk.	Anz. nicht gesteuerter
Max	0	0	0	0	0	0
Min	0	0	0	0	0	0
Median	0	0	0	0	0	0
Mittelwert	0	0	0	0	0	0

Wie bereits bei den Auswertungskriterien erwähnt, wurde bei der technischen Evaluation der gelabelte Barge In– bzw. Offtalkanteil ermittelt, obwohl diese Funktionalität im Demonstrator 3.1 nicht verfügbar war, d.h. keine gesteuerten Abbrüche erlaubt waren.

3. Hilfefunktion

	Anzahl der Hilfestellungen pro Aufnahme
Max	195
Min	0
Median	74
Mittelwert	76.06

Wie ebenfalls schon im Abschnitt zu den Auswertungskriterien erwähnt, wird die dynamische Hilfe zwar ausgewertet, entspricht aber in keinster Weise der von den Benutzern erwarteten Hilfe.

Wie man den Werten der Tabelle entnehmen kann, tritt die dynamische Hilfe mit vielen Hilfestellungen in Aktion. Wie im Abschnitt zur vergleichenden Evaluation noch aufgezeigt werden wird, ist diese dynamische Hilfe in den meisten Fällen ohne Bedeutung, ja dringt nicht einmal bis zur Systemausgabe vor. Es ist offensichtlich, daß das System genau erkennt, daß der Dialog falsch

verläuft, es sich selbst aber durch das Ignorieren der dynamischen Hilfefunktion selbst schadet.

4. Syntax

	Syntactical correctness of input	Syntactical correctness of output
Max	100.00%	100.00%
Min	83.33%	100.00%
Median	97.00%	100.00%
Mittelwert	95.70%	100.00%

Die Auswertung der syntaktischen Korrektheit zeigt optimale Werte für den Output. Da es sich beim Input um spontansprachliche Äußerungen des Benutzers handelt, können diese teilweise von einer idealen Syntax des Deutschen abweichen.

5. Semantik

Bei der Bewertung des semantischen Gehalts der Ein– und Ausgabe wurden folgende Berechnungen erstellt:

			No. of senses				
	doppel–deutige Ausgaben	doppel–deutige Eingaben	Sen–tence	Single input	Semantical correctness of input	Semantical correctness of output	Output: „nicht verstanden"
Max	0	0	0	0	70.00%	50.00%	90.90%
Min	0	0	0	0	11.11%	2.94%	0.00%
Median	0	0	0	0	28.00%	15.00%	74.00%
Mittelwert	0	0	0	0	33.68%	17.53%	65.78%

Aus der Tabelle wird ersichtlich, daß es weder doppeldeutige Eingaben, noch Ausgaben gibt und daß es auch keinerlei Satz– oder Eingabeambiguitäten gibt. Die semantische Korrektheit der Aus– und Eingabe berechnet sich aus der inhaltlichen Reaktion auf eine vorangegangene Äußerung bzw. um taskspezifische Eingaben. Je schlechter der prozentuale Anteil an semantisch korrekter Eingaben, desto schlechter ist auch die semantische Korrektheit der Ausgabe zu erwarten. Der, im Vergleich zum Home–Szenario noch höhere Anteil an „Ich habe Sie nicht verstanden" Äußerungen (Max: 90.90% und Mean 65.78%) sollte ebenfalls Anreiz zu Systemverbesserungen sein.

6. Dialog

Die folgenden Unterabschnitte geben einen Einblick in die Dialogsituation. Zunächst wird die Anzahl von Wörtern, Gesten und Graphikausgaben pro Session, bei der Audio–Ein–/Ausgabe zusätzlich die Anzahl der Turns bestimmt.

6.2. Anzahl

Gesamtzahl							Prozentualer Anteil erfolgreicher in %		
	Audio					Grafik			
	Wörter	Turns	User	System	erfolg.	Turns(Events)	Gestik	TRL	Inputs allg.
Max	497	61	32	34	7	49	68.18	17.07	17.07

Min	51	6	6	2	0	4	0	0	0
Median	235	36	20	16.5	0	28	14.59	0	0
Mittelwert	254.12	33.94	19.18	15.69	1.29	28.47	17.89	3.05	3.05

Die Anzahl von Wörtern und Turns ist fast ausgewogen für den Audiobereich, für den Graphik–System–Audiobereich liegt das Verhältnis grob bei 1:2.

Die Anzahl der erfolgreichen Inputs liegt bei durchschnittlich 3%!

Im Vergleich zum Home–Szenario hat sich die Anzahl der Audio– und Grafikturns bzw. der Wortanteil bei Audio kaum verändert.

6.3. Kompetenz/ Natürlichkeit

Die Kompetenz und Natürlichkeit des Systems spiegelt sich nicht ausschließlich in nachfolgenden Kriterien wieder. Sie werden logischerweise durch die nachfolgenden Unterpunkte Dialogverhalten, Komplexität und Modalitätenwechsel ergänzt.

	Fehlermeldungen	Fehlerrate von Anfragen	Fehlerfrequenz von Antworten	Wider–sprüche
Max	677	21.43%	100.00%	0
Min	0	0.00%	100.00%	0
Median	273	5.86%	100.00%	0
Mittelwert	298.59	7.36%	100.00%	0

Wie sich aus der hohen Anzahl von fehlerhaften Antworten erkennen läßt, scheint das System nicht sehr kompetent mit der gestellten Aufgabe umgehen zu können. Diese Aussage bestätigt ebenfalls das Memo 14 zur ergonomischen Evaluation des Prototypen Mobil.

	Synchrone Gestik	Keywords in USR–Turns	Keywords in SMA–Turns	USR–Keywords in SMA–Turns
Max	58.54%	16	4	0
Min	0.00%	0	0	0
Median	0.00%	2	1	0
Mittelwert	14.09%	3.24	1.59	0

Eine relativ hohe Natürlichkeit der Persona ergibt sich durch die gute Koordination sprachbegleitender Gesten der Persona. Auch das Aufgreifen von taskspezifischen Keywords zeugt von größerer Natürlichkeit. Im Vergleich zum Szenario Home büßt die Natürlichkeit bei Mobil deutlich ein (MAX für synchrone Gestik: 58.54%, Mittelwert 14.09%)!

6.4. Dialogverhalten

	Appropriate diag. utterances	Inappropriate diag. utterances	Explicit	Implicit	Dialogführung		
			recovery answers		System	User	gemischt

Max	0	0	0	0	Anzahl Dialoge		
Min	0	0	0	0			
Median	0	0	0	0			
Mittelwert	0	0	0	0	12	4	0

Wie aus der Tabelle ersichtlich, versucht das System keinerlei Reparaturmaßnahmen. Der Dialog ist meist systemgeführt, zwei Dialoge sind nicht auswertbar.

	Wortab–brüche	Wieder–holungen	Unver–ständlich	Incremental Compatibility	Cooperativity user	Cooperativity system
Max	0	3	15	6	0.50%	100.00%
Min	0	0	0	1	0.00%	100.00%
Median	0	0	2	2	0.33%	100.00%
Mittelwert	0	0.63	4.25	2.71	0.34%	100.00%

Das System (im Gegensatz zum User) verhält sich kooperativ, was aber den Dialogverlauf nicht maßgeblich beeinflußt. Die Kooperativität des Users wird natürlich durch sinngemäße wie wortwörtliche Wiederholungen und Speech Disfluencies sowie durch häufigen Offtalk sehr eingeschränkt.

6.5.Modalitätenwechsel

Anzahl	SMA–Audio vs Grafikturns	USR vs SMA–Audio	USR–Audio vs Grafikturns
Max	19	14	32
Min	0	0	0
Median	0	2	0
Mittelwert	2.18	3.82	3.94

Die hohe Anzahl an Überschneidungen von User– und Systemausgabe zeigt klar , daß die Dialogverarbeitung zu langsam ist.

Wechsel aus/in Grafikmodus				
	G.=>SMA–Audio	G.=>USR–Audio	SMA–Audio=>G.	USR–Audio=>G.
Max	8	9	7	11
Min	0	0	0	0
Median	0	0	0	1
Mittelwert	0.65	1.24	0.71	2.12

Die Wechsel aus bzw. in den Grafikmodus sind relativ gering und nicht besonders aussagekräftig, da ein Vergleich mit der Gestik fehlt.

6.6.Verarbeitungszeit

	Dialog–dauer(ms)	Mean response time (ms)	Elapsed time(ms)	
			Eingabe	Erkennung

		User	System	Gestik	Sprache	Gestik	Sprache
Max	621049	16241	8854	1308.52	2242.27	2713.42	17621.41
Min	176000	2999	2255	281.84	1485.77	323.94	2809
Median	513000	7644	4388	437.22	1864.02	1518.68	6458.53
Mittelwert	477209.2	8157.67	4724.4	675.86	1864.02	1518.68	7586.39

Die Verarbeitungzeit ist auffallend langsam. Andere Systemeinschätzungen von seiten der VPn sind auf „prompte" Antworten bei Turnüberholungen zurückzuführen. Negative Eingabe– und Antwortzeiten ergeben sich durch eben diese Turnüberholungen.

Die Verarbeitungszeit der Gestik ist immer unter der Perzeptionsschwelle von 3 sec., die der Spracherkennung in der Regel weit darüber. Die durchschnittliche Echtzeit der Spracherkennung liegt bei ca. dem Dreifachen der Eingabezeit, die natürlich auch meist wesentlich länger als die Spracheingabe ist.

7. Ergebnisse der vergleichenden Evaluation TE vs. EE

Der Evaluationsvergleich entstand, um den Modulentwicklern eine tendentielle Orientierungsrichtlinie für nötige Verbesserungsmaßnahmen zu geben. Insgesamt gesehen scheint die Aussage der VPn bei den 17 evaluierten Sessions in der Regel mit der der technischen Evaluation (112) übereinzustimmen, in 78 Entscheidungsfällen sollte die technische Seite zur Systemverbesserung herangezogen werden, da die VPn teilweise entgegen ihres Dialogverhaltens zu milde gewertet hat, in 110 Entscheidungsfällen ist die Aussage der VPn aussagekräftiger. In 36 Fällen konnte weder die TE noch die EE im Gesamteindruck bestätigt werden; im technischen Bereich läßt sich das v.a. durch die fehlende Gestikreferenz erklären.

Ausgeglichen	TE	EE	weder TE noch EE
112	78	110	36

Es gilt zu beachten, daß der Score **nicht** gleichbedeutend mit einer **positiven** Wertung ist! Insbesondere bei der „Hilfe" ist ersichtlich, daß das Fehlen einer expliziten Hilfe stark gewertet wurde. Im Gegensatz zur Evaluation des Prototypen Home ist bei Mobil die Kompetenz, Menschlichkeit und Multimodalität wenig ausgewogen. Für die lange Verarbeitungsdauer, Offtalk und semantische Korrektheit gelten ähnliche Bewertungen auf beiden Seiten. Die häufige Wiederholungsnotwendigkeit wird ebenfalls von beiden Seiten ausgewogen gewertet. Bei der Spracherkennung gilt zu beachten, daß sowohl TE als auch EE relativ ausgewogen sind bzw. sogar eine gute Tendenz bei der TE liegt. Dies zeigt auf, daß die „Verständnisprobleme" des Systems wohl eher nach der eigentlichen Spracherkennung stattfinden. Die Dialogkomplexität an sich ist meist sehr niedrig angesetzt, die Fehlerrate der Antworten ist dagegen relativ hoch.

8. Performanz nach PROMISE

Die Performanz nach PROMISE ergibt in etwa das gleiche Bild wie der unter 5.7 beschriebene Evaluationsvergleich durch menschliche Evaluierer.

Die Performanz nach PROMISE liegt bei –40.5661.
Im Vergleich zu Home hat sich die Performanz um 100% verschlechtert.

5. Fazit

Der SmartKom Demonstrator 3.1, Szenario Mobil, wurde mit 18 Sessions à 17 Versuchspersonen (eine VP machte eine zweite Session) getestet, wobei für den Evaluationsvergleich eine Session aufgrund unzureichender ergonomischer Daten nicht herangezogen werden konnte. Die hier detailliert ausgewerteten technischen Daten und der Vergleich der TE und EE ergaben ein allgemein negatives Bild des Prototypen.

Besonders auffallend sind die immer noch schlechte Performanz aller Erkenner, insbesondere der Prosodie, der fehlende korrekte Umgang mit Offtalk und Barge In, die lange Wartezeit zwischen Befehlseingabe und Reaktion des Systems und die damit verbundenen Turnüberholungen, die hohe Fehlerrate bei der Beantwortung der Userfragen bzw. die niedrige Anzahl erfolgreicher Eingaben sowie die Diskrepanz zwischen dynamischer Hilfe und deren Effizienz. Trotz allgemein negativer Bewertung des Systems gibt es dennoch ein paar positive Aspekte. So ist die syntaktische Korrektheit der Sprachausgabe (100%) hervorzuheben.

Es gilt zu hoffen, daß die essentiellen Dialogverarbeitungsmodule des Basisystems im weiteren Verlauf des Projekts maßgebliche Verbesserungen durchführen, wie das bereits bei der Spracherkennerversion für den Demonstrator 3.1 vonstatten ging.