Gefördert vom

bmb+f

Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Mensch-Technik-Interaktion
SmartKom

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

–Szenario Home–

Aufnahmezeitraum 01. April 02 bis 31. Juni 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. 13
Dezember 2002

Dezember2002

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 7: Intuitives Arbeiten mit dem Computer: SmartKom Home/Office; AP 7.5 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.Komplexität..13
 6.6. Modalitätenwechsel...13
 6.7.Verarbeitungszeit..14
 7. Ergebnisse der vergleichenden Evaluation TE vs. EE...14
 8. Performanz nach PROMISE...15
5.Fazit...15

1. Evaluationszeitraum

Die End–to–End–Evaluation für das Szenario Home umfaßt den Zeitraum vom 01.04.02 bis einschließlich 31.06.02. Dieses Memo ergänzt Memo 12 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 Home.

2. Anzahl und Art der Aufnahmen

Von den 29 End–to–End–Evaluationsaufnahmen des SmartKom Prototypen Home konnten insgesamt 24 technisch evaluiert werden, wobei 3 Sessions aufgrund technischer Defekte beim Aufnahmesetup für die vergleichende Evaluation von TE und EE (Evaluationstool / PROMISE) ausschieden. Bei 5 der 29 Sessions wurde kein Logfile produziert. Die folgenden Daten beziehen sich damit auf 21 gültige Sessions zur Evaluation des Home–Prototypen (Demonstrator 2.1).

Naivität der Versuchspersonen

Entgegender ursprünglichen Planung der begleitenden Evaluation wurden neben 19 naiven Versuchspersonen auch 10 nicht naive evaluiert. Alle nicht bzw. nur bedingt auswertbaren Sessions wurden mit naiven VPn aufgenommen. Somit enthält die Stichprobe eine ausgewogene Verteilung von 10 nicht–naiven zu 11 naiven VPn.

Silence detection

Um bestmögliche Aufnahmebedingungen zu erhalten, wurde zunächst eine explizite Aufforderung zur Spracheingabe vorgesehen, d.h. die VPn sollten nur bei einem grünen Lichtsignal Audioeingaben machen. Da die erwartete Reaktion auf das Leuchtsignal ausblieb, wurde diese Art der Spracheingabe weiterhin verworfen, da die VPn bereits nach kurzer Zeit nicht auf dieses Signal achteten. 9 der 29 End–to–End–Aufnahmen erfolgten somit ohne 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 Home 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 Antworten gegeben werden. So wurden zum Beispiel die Fernsehprogramme verschiedener Sender kurz hintereinander erfragt mit dem Resultat, daß relativ prompt die vermeintliche Antwort auf die letzte Frage kam. (Beispiel: Frage: „Was läuft auf Pro 7?“ Antwort: z.B. „Hier sehen Sie die Sendungen, die auf ZDF laufen.“).

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

<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="7">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>

<table>
<tr><th colspan="2">Kriterium</th><th>Bedeutung</th></tr>
<tr><td colspan="2">Wechsel von GES nach SMA–Audio</td><td rowspan="8">Wechsel von einem Modus in einen anderen.</td></tr>
<tr><td colspan="2">Wechsel von GES nach USR–Audio</td></tr>
<tr><td colspan="2">Wechsel von Grafik nach SMAAudio</td></tr>
<tr><td colspan="2">Wechsel von Grafik nach USRAudio</td></tr>
<tr><td colspan="2">Wechsel von SMA–Audio nach GES</td></tr>
<tr><td colspan="2">Wechsel von SMAAudio nach Grafik</td></tr>
<tr><td colspan="2">Wechsel von USR–Audio nach GES</td></tr>
<tr><td colspan="2">Wechsel von USRAudio nach Grafik</td></tr>
<tr><td colspan="2">Semantical correctness</td><td rowspan="2">Ist Input/Output passend zum vorangehenden Turn bzw. zum Anwendungsfall?</td></tr>
<tr><td>of input</td><td>of output</td></tr>
<tr><td colspan="2">Output "nicht verstanden"</td><td>Anzahl der "nicht verstanden"–Antworten</td></tr>
<tr><td colspan="2">Syntactical correctness</td><td rowspan="2">Gibt es syntaktische Fehler in den Turns?</td></tr>
<tr><td>of input</td><td>of output</td></tr>
<tr><td colspan="2">Elapsed time Mimikerkennung</td><td>Dauer der Mimikerkennung</td></tr>
<tr><td colspan="2">Elapsed time "Mimikeingabe"</td><td>Dauer der Mimikeingabe</td></tr>
<tr><td colspan="2">Dialogdauer</td><td>Dialogdauer</td></tr>
<tr><td colspan="2">Objekt–Manipulationen</td><td>z.B. Drag–and–Drop–Aktionen</td></tr>
<tr><td colspan="2">Mean user response time</td><td rowspan="2">Mittlere Antwortzeit des Benutzers / Systems (Ende Eingabe–User/Ausgabe–System bis Beginn Ausgabe–System/ Eingabe–User)</td></tr>
<tr><td colspan="2">Mean system response time</td></tr>
<tr><td colspan="2">Appropriate diag. utterances</td><td rowspan="2">Zur Dialogsituation passende/unpassende Äußerungen des Systems</td></tr>
<tr><td colspan="2">Inappropriate diag. utterances</td></tr>
<tr><td colspan="2">Explicit recovery answers</td><td rowspan="2">Explizite (z.B. „Könnten Sie den Film wiederholen?“) /implizite (z.B. „Hier sehen Sie die Sehenswürdigkeiten von Aachen und Aalen“) Reparaturversuche des Systems.</td></tr>
<tr><td colspan="2">Implicit recovery answers</td></tr>
<tr><td colspan="2">Dialog</td><td>system–, usergeführt oder ausgewogen</td></tr>
<tr><td colspan="2">Wortabbrüche</td><td>Abgebrochene Wörter bei der Eingabe</td></tr>
<tr><td colspan="2">Wiederholungen</td><td>Wörtliche Wiederholungen</td></tr>
<tr><td colspan="2">Unverständlich</td><td>unverständliche Eingaben</td></tr>
<tr><td colspan="2">Incremental Compatibility</td><td>sinngemäße Wiederholungen</td></tr>
<tr><td colspan="2">Cooperativity user</td><td rowspan="2">Wie gut geht der jeweilige Dialogpartner auf sein „Gegenüber“ ein?</td></tr>
<tr><td colspan="2">Cooperativity system</td></tr>
</table>

<table>
<tr><th colspan="2">Kriterium</th><th>Bedeutung</th></tr>
<tr><td colspan="2">Gesamtzahl Audio–Turns</td><td rowspan="5">Gesamtzahl der verschiedenen Ein–/Ausgaben auf Tunbasis</td></tr>
<tr><td colspan="2">Gesamtzahl User–Audio–Turns</td></tr>
<tr><td colspan="2">Gesamtzahl System–Audio–Turns</td></tr>
<tr><td colspan="2">Gesamtzahl Gestikturns</td></tr>
<tr><td colspan="2">Gesamtzahl Grafikturns</td></tr>
<tr><td colspan="2">Fehlerrate von Anfragen</td><td>Anzahl falscher Anfragen</td></tr>
<tr><td colspan="2">Fehlerfrequenz von Antworten</td><td>Anzahl falscher Antworten</td></tr>
<tr><td colspan="2">Widersprüche</td><td>bei Ein–/Ausgaben (parallel/sequentiell)</td></tr>
<tr><td colspan="2">Gesamtzahl der Wörter im File</td><td>Gesamtzahl aller Wörter</td></tr>
<tr><td colspan="2">Synchrone Gestik</td><td rowspan="2">Synchronität der Ausgabe (sprachbegleitende Gestik des Wichtels, Lippensynchronität)</td></tr>
<tr><td colspan="2">Synchrone Lippenbewegungen</td></tr>
<tr><td colspan="2">Überlappungen</td><td>Anz. der sich überlappenden Graphik und Synthese</td></tr>
<tr><td colspan="2">Zeitverschiebungen</td><td>Asynchrone Ausgaben</td></tr>
<tr><td colspan="2">Keywords in USR–Turns</td><td rowspan="5">Taskspezifische Keywords bei Benutzer und System sowie aufgegriffene Keywords des jeweiligen Dialogpartners.</td></tr>
<tr><td colspan="2">S–Keywords in USR–Turns</td></tr>
<tr><td colspan="2">Keywords in SMA–Turns</td></tr>
<tr><td colspan="2">USR–Keywords in SMA–Turns</td></tr>
<tr><td colspan="2">USR–Keywords in SMA–Turns Liste</td></tr>
<tr><td colspan="2">Task Complexity</td><td rowspan="5">Die Komplexität eines Anwendungsfalls (meist dasselbe wie Dialogkomplexität, da nur ein Anwendungsfall pro Session vorhanden)</td></tr>
<tr><td colspan="2">% aller Turns einer Task</td></tr>
<tr><td colspan="2">% aller Wörter einer Task</td></tr>
<tr><td colspan="2">% aller Gesten einer Task</td></tr>
<tr><td colspan="2">% aller Grafiken einer Task</td></tr>
<tr><td colspan="2">Task Completion Time</td><td>Die Zeit bis zur Beendigung der Task</td></tr>
<tr><td colspan="2">Prosodie</td><td rowspan="3">Anzahl der gelabelten prosodischen Grenzen und Akzente.</td></tr>
<tr><td colspan="2">Accent</td></tr>
<tr><td colspan="2">Boundary</td></tr>
<tr><td>SMA</td><td>USR</td><td rowspan="3">Gesamtzahl aller Wörter pro Turn/pro Session pro Dialogteilnehmer</td></tr>
<tr><td colspan="2">Anzahl der Wörter im Turn</td></tr>
<tr><td colspan="2">Gesamtzahl</td></tr>
</table>

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

Für den Demonstrator 2.1 (Szenario Home) 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.

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	Gestikerkennung	Prosodie (% | Anzahl | Anzahl)		
	%	%	Erkennung	Accent(Anz)	Boundary
Max	57.78	80	31.43	63	330
Min	–33.33	–350	–1527.08	13	35
Median	22.83	0	0	33	86
Mittelwert	22.35	–12.05	–298.73	32.39	114.17

Alle Erkenner scheinen nicht mit der realitätsnaheren Versuchssituation klarzukommen. So erreicht die Spracherkennung einen Mittelwert von 22.35% über alle 24 Sessions, wobei die maximale Erkennerleistung bei 57.78 % liegt. Das kann unter anderem daran liegen, 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).

Wenn auch die Gestik ein Maximum von 80% Akkuratheit erreicht, ist es durchaus möglich, negative Akkuratheiten zu erhalten, wenn die Anzahl der erkannten Gestik größer als die Anzahl der gelabelten Gestik ist. Es liegt die Vermutung nahe, daß das System wesentlich kürzere Einheiten als Einzelgesten verarbeitet als es von der VP intendiert ist. So könnte sich die Anzahl erkannter Gesten durch Doppel– oder Mehrfachklicks relevant erhöhen.

Auch bei der Prosodieerkennung könnte durch die hohe Anzahl an meist nicht eindeutig definierbaren prosodischen Grenzen bzw. nicht–Phrasen–Akzenten ein sehr niedriger Maximumwert von 31.43% plausibel sein. Negative Akkuratheiten ergeben sich durch einen für neuere Versionen erkannten Bug.

2. Barge In/Offtalk

	BargeIn Vorkommen	Prozentualer BargeIn Anteil (Session)	Offtalk Vorkommen	Prozentualer Offtalk Anteil (Session)
Max	23	4.49%	128	31.37%
Min	0	0.00%	0	0.00%
Median	2	0.01	0	0
Mittelwert	3.83	1.40%	8.87	3.39%

Wie bereits bei den Auswertungskriterien erwähnt, wurde bei der technischen Evaluation der gelabelte Barge In– bzw. Offtalkanteil ermittelt, obwohl diese Funktionalität im Demonstrator 2.1 nicht verfügbar war, d.h. keine gesteuerten Abbrüche erlaubt waren. Die Anzahl nicht gesteuerter Abbrüche für Synthese und Display lassen auf ein situationsbedingtes technisches Problem schließen.

Wie aus den Werten der Tabelle zu BargeIn/Offtalk ersichtlich, hält sich der Mittelwert für beide Kriterien in Grenzen. 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: 31.37%) 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	2	0	1
Min	0	0	0	0	0	0
Median	0	0	0	0	0	0
Mittelwert	0	0	0	0.27	0	0.14

3. Hilfefunktion

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.

	Anzahl der Hilfestellungen pro Aufnahme
Max	123
Min	1
Median	10
Mittelwert	28.78

Wie man den Werten der Tabelle entnehmen kann, tritt in jeder Session die dynamische Hilfe 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.

4. Syntax

	Syntactical correctness of input	Syntactical correctness of output
Max	100.00%	100.00%
Min	78.13%	100.00%
Median	100.00%	100.00%
Mittelwert	96.82%	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	Singl e input	Semantical correctness of input	Semantical correctness of output	Output: „nicht verstanden"
Max	0	0	0	0	83.33%	80.00%	53.84%
Min	0	0	0	0	11.11%	14.28%	10.00%
Median	0	0	0	0	50.00%	35.00%	30.00%
Mittelwert	0	0	0	0	47.14%	36.19%	31.78%

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, was sich ganz klar in den Werten Min, Max und Mittelwert zeigt. Der hohe Anteil an „Ich habe Sie nicht verstanden" Äußerungen (Max: 53,84% und Mean 31.78%) 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		
	Audio				Grafik	Gestik			
	Wörter	Turns	User	System	erfolg.	Turns	Turns	Gestik	TRL	Inputs allg.
Max	449	63	33	30	10	50	25	68.18	20.83	77.27
Min	80	13	6	6	0	5	0	0	0	0
Median	184	33	17	13	0	17	8	14.59	0	20.04

Mittelwert	192.52	32.39	18.74	13.65	1.23	19.87	7.96	17.89	3.26	21.15

Die Anzahl von Wörtern und Turns ist fast ausgewogen für den Audiobereich, für den Graphik–/Gestikbereich liegt das Verhältnis grob bei 5:2.

Die Anzahl der erfolgreichen Inputs liegt bei durchschnittlich 20%.

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	763	21.43%	100.00%	0
Min	60	0.00%	34.48%	0
Median	205.5	5.09%	86.52%	0
Mittelwert	253.41	7.40%	84.94%	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 12 zur ergonomischen Evaluation des Prototypen Home.

	Synchrone Gestik	Keywords in USR–Turns	Keywords in SMA–Turns	USR–Keywords in SMA–Turns
Max	97.22%	34	24	23
Min	37.50%	4	0	0
Median	0.77	11	7	8
Mittelwert	75.53%	12.7	7.91	9.04

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.

6.4. Dialogverhalten

<table>
<tr><td></td><td rowspan="2">Appropriate diag. utterances</td><td rowspan="2">Inappropriate diag. utterances</td><td>Explicit</td><td>Implicit</td><td colspan="3">Dialogführung</td></tr>
<tr><td></td><td colspan="2">recovery answers</td><td>System</td><td>User</td><td>gemischt</td></tr>
<tr><td>Max</td><td>0</td><td>0</td><td>0</td><td>0</td><td colspan="3" rowspan="3">Anzahl Dialoge</td></tr>
<tr><td>Min</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>Median</td><td>0</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>Mittelwert</td><td>0</td><td>0</td><td>0</td><td>0</td><td>21</td><td>3</td><td>0</td></tr>
</table>

Wie aus der Tabelle ersichtlich, versucht das System keinerlei Reparaturmaßnahmen. Der Dialog ist meist systemgeführt.

	Wortab-brüche	Wieder-holungen	Unver-ständlich	Incremental Compatibility	Cooperativity user	Cooperativity system
Max	0	6	19	12	0.50%	100.00%
Min	0	0	0	1	0.00%	34.48%
Median	0	0	1	3	0.23%	86.52%
Mittelwert	0	0.61	2.48	3.79	0.26%	84.94%

Das System (im Gegensatz zum User) verhält sich meist kooperativ. Die Kooperativität des Users wird natürlich durch sinngemäße wie wortwörtliche Wiederholungen und Speech Disfluencies sowie durch häufigen Offtalk sehr eingeschränkt.

6.5.Komplexität

	Komplexität der Gestik (Input)	Komplexität der Gestik (Input)	Komplexität der Grafik (Dialogue Manager)	Komplexität der Grafik (Dialogue Manager)
	% U-Turns	% U-Gesten	% S-Ausgabe	% S-Grafiken
Max	0.36%	1.00%	0.55%	1.00%
Min	0.00%	1.00%	0.24%	1.00%
Median	0.13%	1.00%	0.35%	1.00%
Mittelwert	0.13%	1.00%	0.37%	1.00%

	Komplexität der Sprache (Dialogue Manager)	Komplexität der Sprache (Dialogue Manager)	Komplexität der Sprache (Input)	Komplexität der Sprache (Input)
	% S-Turns	% S-Words	% U-Turns	% U-Words
Max	0.31%	8.80%	0.45%	15.68%
Min	0.10%	2.71%	0.14%	3.11%
Median	0.19%	6.15%	0.28%	5.14%
Mittelwert	0.20%	6.31%	0.29%	5.71%

Die Komplexität des Dialogs ergibt sich aus Eingabe- und Ausgabekomplexität. Beide sind für beide Dialogteilnehmer sehr gering. Die Taskkomplexität weicht für Home nicht von der Dialogkomplexität ab, da es keinerlei Wechsel in andere Anwendungsfälle gab.

6.6. Modalitätenwechsel

Anzahl	GES vs Grafikturns	GES vs SMA-Audio	GES vs USR-Audio	SMA-Audio vs Grafikturns	USR vs SMA-Audio	USR-Audio vs Grafikturns
Max	15	2	7	29	15	52
Min	0	0	0	5	0	4
Median	4	0	0	16	1	19

Mittelwert	4.21	0.38	1	14.92	1.75	21.08

Die hohe Anzahl an Überschneidungen von Gestik und Audio spiegelt einerseits den Wunsch nach Multimodalität auf Seiten des Users wider, andererseits zeigen die Überschneidungen mit der Systemausgabe klar , daß die Dialogverarbeitung zu langsam ist.

Wechsel								
	GES => SMA-Audio	GES => USR-Audio	Grafik => SMA-Audio	Grafik => USR-Audio	SMA-Audio => GES	SMA-Audio => Grafik	USR-Audio => GES	USR-Audio => Grafik
Max	3	4	20	15	2	16	6	24
Min	0	0	1	2	0	0	0	1
Median	1	1	3	8.5	0	5.5	1	7
Mittelwert	1.21	1.25	4.63	8.42	0.46	5.71	1.54	8.33

Die Wechsel in verschiedene Modi zeigen geringfügige Präferenzen für die Audioeingabe.

6.7. Verarbeitungszeit

	Dialog-dauer(ms)	Mean response time (ms)		Elapsed time(ms)			
				Eingabe		Erkennung	
		User	System	Gestik	Sprache	Gestik	Sprache
Max	598851	202141	24543	1403.38	3795.45	771.13	27951.59
Min	204545	3503	4639	0	1573.04	340	2148.13
Median	505251	10211	9295.5	107	1883.91	511	10249.49
Mittelwert	475774.21	23536.38	10274.63	252.17	2237.63	521.63	11465.84

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 Fünffachen 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 in der Regel mit der der technischen Evaluation (184) übereinzustimmen, in 176 Entscheidungsfällen sollte die technische Seite zur Systemverbesserung herangezogen werden, da die VPn teilweise entgegen ihres Dialogverhaltens zu milde gewertet hat, in 99 Entscheidungsfällen ist die Aussage der VPn aussagekräftiger.

Ausgeglichen	TE	EE
184	176	99

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. Kompetenz, Menschlichkeit und Multimodalität sind relativ ausgewogen. Dasselbe gilt für die lange Verarbeitungsdauer. Die häufige Wiederholungsnotwendigkeit wird ebenfalls von beiden Seiten ausgewogen gewertet. Bei der Sprach– und Gestikerkennung sollte eher die weitaus schlechtere technische Seite zur Verbesserung herangezogen werden, obwohl durch die EE auch ganz klar geschlossen werden kann, daß nicht–verstehen (des Systems ausgewogen schlecht bewertet) nicht unbedingt nicht–erkannt heißen muß. Die Dialogkomplexität an sich ist meist sehr niedrig angesetzt.

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 −20.8589.

5. Fazit

Der SmartKom Demonstrator 2.1, Szenario Home, wurde mit 24 bzw. 21 Sessions à 24 (21) Versuchspersonen (jeweils eine Versuchsperson) getestet. Die hier detailliert ausgewerteten technischen Daten und der Vergleich der TE und EE ergaben ein allgemein negatives Bild des Prototypen.

Besonders auffallend sind die 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 Synchronität der ausgabebegleitenden Wichtelgestik mit gut 75% , besonders aber 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.