Evaluation multimodaler Systeme: Ist das Ganze die Summe seiner Teile?

Ina Wechsung[1], Klaus-Peter Engelbrecht[1], Julia Seebode[2], Stefan Schaffer[2], Florian Metze[3], Sebastian Möller[1]

Deutsche Telekom Laboratories, TU Berlin[1]
Graduiertenkollegs prometei, TU Berlin[2]
interACT center, Carnegie Mellon University[3]

Zusammenfassung

Multimodale Systeme stellen dem Nutzer mehrere Kommunikationskanäle zur Verfügung. Bei der Evaluation müssen daher sowohl die einzelnen Modalitäten als auch deren Kombination in Betracht gezogen werden. Dieser Beitrag untersucht, ob, und wenn ja, wie Bewertungen für einzelne Modalitäten mit der Bewertung des Gesamtsystems zusammenhängen. Dazu wurde ein Raummanagement- und Informationssystem in zwei unimodalen und einer multimodalen Version getestet. Multiple lineare Regressionen zeigten, dass hinsichtlich Gesamt- und Globalurteilen die Bewertungen einzelner Modalitäten sehr gute Prädiktoren für die Bewertungen des Gesamtsystems sind. Für einzelne Faktoren waren die Vorhersagen allerdings weniger genau.

1 Einleitung

Da menschliche Kommunikation von Natur aus typischerweise multimodal ist, wird erwartet, dass auch multimodale Systeme vorteilhaftere Interaktionen ermöglichen (Chen 2006). Durch das Angebot mehrere Kommunikationskanäle könne, so die Annahme, auf unterschiedliche kognitive Ressourcen zurückgegriffen werden. Dies sollte die menschliche Informationsverarbeitung unterstützen (Wickens 2002). Doch zeigt der gegenwärtige Forschungsstand, dass das Hinzufügen einer Modalität nicht zwangsläufig zu einer Verbesserung führt (Oviatt 1999). Mehr Freiheitsgrade in der Interaktion können auch in einer höheren, kognitiven Belastung resultieren (Schomaker et al. 1995). Außerdem sind Interferenzen bzw. Synchronisierungsprobleme zwischen den Modalitäten möglich (Schnotz, Bannert, & Seufert 2002). Die Evaluation multimodaler Schnittstellen ist demnach ein komplexes Problem. Berücksichtigt werden müssen die einzelnen Modalitäten sowie deren Kombinationen. Etablierte Evaluationsmethoden beschränken sich üblicherweise auf unimodale Systeme und sind daher zur Bewertung multimodaler Systeme nicht ohne Einschränkungen geeignet. Darüber hinaus ist die Gewichtung und Kombination der Bewertungen einzelner Modalitäten zu einem Gesamturteil schwierig (Beringer et al. 2002). In dieser Studie wurde daher untersucht ob und wenn ja, inwieweit Bewertungen unimodaler Systemversionen mit dem Gesamturteil der multimodalen Systemversion zusammenhängen.

2 Methode

Ingesamt nahmen 36 deutschsprachige Personen (17 männlich, 19 weiblich) im Alter zwischen 21 und 39 Jahren an der Studie teil. Keiner der Probanden besaß Vorerfahrung mit dem System. Getestet wurde ein Raummanagement- und Informationssystem, das über eine grafische Nutzerfläche mit Touchinput, Sprachinput oder einer Kombination von beiden gesteuert werden kann. Die Probanden bearbeiteten 6 Aufgaben mit dem System: Navigation, Suchen, Anzeigen und Buchen von Räumen, Anzeigen von Veranstaltungen und Suchen von Mitarbeitern. Der Test lief in drei Blöcken ab. Zunächst wurden die Probanden gebeten die Aufgaben mit einer vorgegebenen Modalität zu bearbeiten und danach zu bewerten. Dies wurde für die jeweils andere Modalität wiederholt. Die Reihenfolge der unimodalen Testblöcke wurde variiert. Anschließend wurden die Aufgaben ein weiteres Mal präsentiert. Die Probanden konnten diesmal die Modalität zur Aufgabenbearbeitung frei wählen. Auch eine Kombination der Modalitäten war möglich. Wieder erfolgte eine Bewertung des vorher absolvierten Versuchsblocks. Subjektive Bewertungen der 3 Systemversionen wurden über den AttrakDiff-Fragebogen (Hassenzahl, Burmester, & Koller 2003) erfasst. Neben den Skalen Pragmatik (d.h. Usability) und Attraktivität (d.h. Gesamturteil) misst dieser hedonische Qualitäten, die für die Bewertung von Systemen zunehmend bedeutsam werden. Zudem wurde eine Gesamtskala basierend auf den Mittelwerten aller 28 AttrakDiff-Items gebildet. Weiterhin wurde für den multimodalen Testblock annotiert, welche Modalität von den Teilnehmern jeweils gewählt wurde. Basierend darauf wurden die Prozentsätze der Modalitätennutzung berechnet.

3 Ergebnisse

Bewertungen der unterschiedlichen Systemversionen: Auf allen AttrakDiff Skalen zeigten sich Unterschiede zwischen den drei Systemversionen: Auf der Skala *Pragmatische-Qualität* (PQ) wurde die Version mit Touchsteuerung am besten und die mit Sprachsteuerung am schlechtesten bewertet ($F(2;66)=93,79$; $p=.000$; $part.Eta^2=.740$). Auf beiden hedonischen Skalen erhielt die multimodale Version die besten Bewertungen. Hinsichtlich der Skala *Hedonische Qualität-Stimulation* (HQ-S /($F(2;68)=12,84$; $p=.000$; $part.Eta^2=.274$) erhielt die Version mit Sprachsteuerung die schlechteste Beurteilung, auf der Skala *Hedonische Qualität–Identität* (HQ-I /$F(1,65; 55,99)=15,35$; $p=.000$; $part.Eta^2=.311$) dagegen die Version mit Touchsteuerung. Bezüglich der globalen Skala des AttrakDiff, der Skala *Attraktivität* (ATT), waren die schlechtesten Bewertungen für die Version mit Sprachsteuerung und die besten für die Version mit Touchsteuerung zu beobachten ($F(1,51; 51,22)=47,53$; $p=.000$; $part.Eta^2=.583$). Auf der Gesamtskala erhielt die Version mit Sprachsteuerung schlechtere Bewertungen als die beiden anderen Versionen ($F(2;66)=38,38$; $p=.000$; $part.Eta^2=.538$).

Zusammenhang zwischen Beurteilungen unimodaler Systemversionen und Beurteilungen des multimodalen Systems. Um zu untersuchen wie die Bewertungen der unimodalen Systemversionen mit den Bewertungen des multimodalen Systems zusammenhängen, wurde

für jede (Unter-) Skala eine multiple lineare Regression berechnet. Die Beurteilungen der unimodalen Versionen dienten dabei als Prädiktorvariablen, die Beurteilungen des multimodalen Systems als Zielvariable. Die Ergebnisse zeigen, dass für die Skala Attraktivität sowie für die Gesamtskala die Beurteilungen der unimodalen Versionen sehr gute Prädiktoren für die Beurteilung des multimodalen Systems sind. Für beide Skalen waren die Beta-Gewichte für die Beurteilungen der Version mit Touchsteuerung höher als die der Version mit Sprachsteuerung. Dies entspricht den Nutzungsdaten des multimodalen Systems: Die Touchsteuerung wurde weitaus häufiger genutzt. Ein größerer Einfluss der Bewertungen der Touchsteuerung auf die Bewertungen des multimodalen Systems ist daher plausibel. Für die drei anderen Skalen (*HQ-I, HQ-S, PQ*) lag die Varianzaufklärung zwischen 61 und 69 Prozent. Für beide Hedonische Skalen waren die Beta-Gewichte der Regressionsgleichungen dabei höher für die Bewertungen des Systems mit Sprachsteuerung. Demnach hatte hier, entgegen der Nutzungsdaten, die Bewertung des Systems mit Sprachsteuerung einen größeren Einfluss auf die Bewertung des multimodalen Systems als die Bewertung der Version mit Touchsteuerung.

Skala	**Touch**				**Sprache**				**R^2**	**RMSE**	**F (df)**
	B	**SE B**	**β**	**t (df)**	**B**	**SE B**	**β**	**t (df)**			
Gesamt	.805	.112	.566	0.21* (32)	.680	.098	.546	6.91* (32)	.829	.370	74.94* (2,31)
ATT	.845	.094	.684	9.02* (32)	.478	.087	.419	5.52* (32)	.837	.411	81.99* (2,32)
PQ	.797	.174	.537	4.57* (31)	.468	.130	.421	3.59* (31)	.628	.703	26.19* (2,31)
HQ-S	.689	.134	.521	5.13* (32)	.633	.119	.536	5.31* (32)	.693	.508	36.05* (2,32)
HQ-I	.282	.106	.331	2.66* (32)	.661	.144	.572	4.60* (32)	.612	.527	25.24* (2,32)

*Table 1. Ergebnisse der multiplen linearen Regression auf Basis aller Daten (*p<.01)*

Zur Überprüfung von Overfitting-Effekten wurde eine 10fache Kreuzvalidierung durchgeführt. Für die Skalen *Attraktivität* sowie für die Gesamtskala zeigten sich die Modelle stabil. Für die anderen Skalen ergaben sich größere Overfitting Effekte. Abgesehen von der Skala *PQ* zeigten sich demnach die Modelle, die konsistent mit der tatsächlichen Modalitätennutzung waren, als zuverlässiger.

	Gesamt	**ATT**	**PQ**	**HQ-S**	**HQ-I**
R^2	.799	.805	.539	.607	.391
RMSE	.384	.431	.754	.572	.615

Table 2. Ergebnisse der multiplen linearen Regression nach 10facher Kreuzvalidierung

4 Diskussion und Schlussfolgerungen

Zielstellung dieses Beitrags war es den Zusammenhang zwischen subjektiven Bewertungen für unimodale Systemversionen und multimodaler Systemversion zu untersuchen. Es zeigte sich, dass Gesamt- und Globalbewertungen der einzelnen Modalitäten gute Prädiktoren für die Bewertungen des multimodalen Systems sind. Weiterhin war zu beobachten, dass für die

stabilen Vorhersagen, die Modalität, die während der Interaktion mit dem Gesamtsystem häufiger genutzt wird, von höherem Einfluss auf die Bewertungen für das Gesamtsystem ist. Zudem konnte übereinstimmend mit Oviatt (1999) gezeigt werden, dass das bloße Hinzufügen einer Modalität nicht zwangsläufig zur Verbesserung eines Systems führt. Für diese Untersuchung und das getestete System bedeutet dies, dass das Ganze (multimodale Systemversion) tatsächlich der Summe seiner einzelnen Teile (unimodale Versionen) entspricht. Allerdings gilt diese Schlussfolgerung nicht für spezifischere Qualitätsfaktoren (hedonische und pragmatische Qualitäten). Hier waren stabile Vorhersagen der Urteile für die multimodale Systemversion auf Basis der Bewertungen der unimodalen Versionen nicht möglich. Weiterhin ist anzumerken, dass diese Studie nur Ergebnisse eines Fragebogens berücksichtigt. Es ist demnach offen, ob ähnliche Ergebnisse auch für andere Fragebögen oder Performanzdaten zu beobachten sind. Außerdem ist unklar inwieweit das getestete System und das Testdesign die Ergebnisse beeinflusst haben. So waren bei der multimodalen Systemversion Interferenzen zwischen den einzelnen Modalitäten möglich (z.B. reagierte der Spracherkenner manchmal auf nicht an das System gerichtete Sprache). Außerdem wurde die multimodale Systemversion immer am Ende getestet. Möglicherweise versuchten die Teilnehmer konsistent zu urteilen, in dem sie ihre Bewertungen für die unimodalen Versionen zur Bewertung des Gesamtsystems addierten. In einer Anschlussuntersuchung soll deshalb die Reihenfolge der Systemversionen geändert werden.

Literatur

Beringer, N., Kartal, U., Louka, K., Schiel, F., & Türk, U. (2002). PROMISE: A Procedure for Multimodal Interactive System Evaluation. In *Proceedings of the Workshop Multimodal Resources and Multimodal Systems Evaluation*, S. 77-80.

Chen, F. (2006). *Designing Human Interface in Speech Technology*. New York: Springer.

Hassenzahl, M., Burmester, M. & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Ziegler, J. & Szwillus, G. (Hrsg.): *Mensch & Computer 2003, Interaktion in Bewegung*. Stuttgart: B.G. Teubner. S. 187-196.

Oviatt, S. (1999). Ten myths of multimodal interaction. *Communications of the ACM, 42*(11), 576-583.

Schnotz, W., Bannert, M. & Seufert, T. (2002). Towards an integrative view of text and picture comprehension: Visualization effects on the construction of mental models. In Otero, J., Graesser A. & Leon, J. A. (Hrsg.): *The Psychology of Science Text Comprehension*. Mahwah: Erlbaum, S. 385-416.

Schomaker, L., Nijtmans, J., Camurri, A., Lavagetto, F., Morasso, P., Benoît, C., Guiard-Marigny, T., Le Goff, B., Robert-Ribes, J., Adjoudani, A., Defée, I., Münch, S., Hartung, K., & Blauert, J. (1995). *A Taxonomy of Multimodal Interaction in the Human Information Processing System.* Nijmegen: NICI.

Wickens, C.D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*, 159-177.

Kontaktinformationen

ina.wechsung@telekom.de