USABILITY-UNTERSUCHUNG DER NATÜRLICHSPRACHLICHEN BEDIENUNG EINES SMART-TV

Stefan Hillmann

Quality and Usability Lab, Telekom Innovation Laboratories, Technische Universität Berlin
stefan.hillmann@tu-berlin.de

Kurzfassung: Es werden die Ergebnisse der Usability-Evaluation eines per Sprache bedienten Smart TV präsentiert. Insgesamt haben 40 Teilnehmer im Alter von 20–70 Jahren an der Studie teilgenommen. Es wird gezeigt, dass die Nutzer die Verwendung der Spracheingabe als gut nutzbar und lohnenswert ansehen. Insbesondere dort, wo Spracheingabe einen geringeren Aufwand, als die Nutzung der Fernbedienung, aufweist, wird sie von den Nutzern signifikant bevorzugt. Neben der Bewertung der Interaktion wurden auch die Einstellung gegenüber Technik und die Nutzungshäufigkeit von Sprachsteuerung erhoben. Es zeigt sich, dass nur 22,5% der Teilnehmer regelmäßig die Spracheingabe auf einem Gerät benutzen.

1 Einführung

Hinter dem Begriff Smart Home steht die Idee den Nutzern eines Hauses oder einer Wohnung einen Mehrwert in den Bereichen Kommunikation, Sicherheit, Energie-Management, Entertainment, Gesundheit und Fitness sowie Gerätesteuerung zu bieten. Der Nutzer ist in der Mehrheit der Anwendungsfälle ein Bewohner des Smart Home, aber auch Servicedienstleister, wie z. B. Handwerker, zählen dazu. Im Bereich der Gerätesteuerung liegt der Fokus auf dem Einsatz von verschiedenen Modalitäten (z. B. Sprache oder 2D/3D Gesten), um die Interaktion komfortabler und/oder robuster zu gestalten, als es die taktile Interaktion mittels verschieden ausgeprägter Knöpfe ermöglicht.

In der, in dieser Arbeit vorgestellten, Studie wird der Prototyp eines Smart-TV evaluiert, dessen Bedienung mit natürlicher Sprache als auch mit einer klassischen Fernbedienung erfolgen kann. Die Sprachsteuerung eines Smart-TV dient uns als erster exemplarischer Anwendungsfall für eine natürlichsprachliche Gerätesteuerung im Smart Home.

Im nachfolgenden Abschnitt werden kurz die Funktionalität des Systems als auch die damit durchgeführte Studie (Teilnehmer, Erhebungsmethoden und Durchführung) vorgestellt. Abschnitt 3 präsentiert die Ergebnisse der empirischen Erhebung, welche im Anschluss kurz diskutiert werden (Abschnitt 4).

2 Usability Evaluation eines Smart-TV

2.1 Smart-TV Funktionen

Der in der Studie zu testende Fernseher bietet die Möglichkeit Basisfunktionen des Fernsehers (Kanalwechsel, Lautstärkeregelung) mittels einer klassischen Fernbedingung, als auch per Sprache zu steuern. Weiterhin existiert eine Mediensuche die ausschließlich bei Sprache bedient wird. Hier können das Fernsehprogramm der nächsten 14 Tage (EPG[1]), eine Onlinevideothek (VoD[2])

[1] Elektronischer Program Guide
[2] Video on Demand

	Frauen (20–70)			Männer (26–65)		
	Anzahl	Jahre		Anzahl	Jahre	
Abschluss		$\bar{a}$	$\sigma(a)$		$\bar{a}$	$\sigma(a)$
Hauptschulabschluss	0			1	61,0	
Realschulabschluss	7	42,0	7,07	3	46,0	14,18
Abitur	5	33,8	12,58	4	37,5	18,36
(Fach-)Hochschulabschluss	11	36,5	13,86	9	37,9	15,24
Alle	23	40,6	15,57	17	37,6	11,82

Tabelle 1: Mittelwert ($\bar{a}$) und Standardabweichung ($\sigma(a)$) des Alters (a) aller Teilnehmer. Die Werte sind für die abgefragten Bildungsabschlüsse sowie alle Frauen und Männer angegeben. Für alle Teilnehmer beträgt der Mittelwert 38,85 Jahre und Standardabweichung 13,44 Jahre.

sowie ein Internet-Videoportal (Youtube) nach Sendungen und Filmen durchsucht werden. Beispielhafte Nutzeräußerungen an das System sind „Ton lauter“, „ARD anzeigen“, „Was läuft morgen Abend im ZDF?“ oder „Zeige mir in der Videothek Western aus den siebziger Jahren.“. Insbesondere in Fehlerfällen (z. B. no-match oder zur Anfrage passenden Medieninhalte) gibt das System mittels Sprachsynthese eine entsprechende Fehlermeldung aus. Es findet jedoch kein sich fortsetzender Dialog statt.

Zur Aufnahme des Sprachsignals wird ein Android-basiertes Smartphone verwendet, auf dem eine App installiert ist, die ausschließlich eine Push-to-Talk Funktion zur Verfügung stellt. Das Ende einer Nutzeräußerung wird automatisch erkannt, oder durch den Nutzer, durch nochmaligen Druck auf den Smartphone-Bildschirm, signalisiert. Die, durch eine webbasierte Natrual Language Unit (NLU), erzeugte Interpretation des Sprachsignals wird schließlich an den Fernseher gesendet und dort in eine Aktion umgesetzt. Sowohl der vom Spracherkenner gelieferte Text, als auch die Interpretation werden dem Nutzer, als zusätzliches Feedback zu den gelieferten Inhalten, angezeigt. Diese Information ist bis zum Eintreffen des nächsten Sprachbefehls sichtbar.

2.2 Teilnehmer

Insgesamt nahmen 44 Teilnehmer an der Studie teil, wobei aufgrund von einmaligen technischen Störungen nur bei 40 Teilnehmern vollständige Datensätze erhoben werden konnten. Tabelle 1 zeigt die Verteilung der Teilnehmer hinsichtlich des Alters und Bildungsabschluss. Ein Bachelorabschluss fällt in die Kategorie Hochschulabschluss. Alle Teilnehmer wurden über Kleinanzeigen und ein Webportal zur Probandenakquise rekrutiert.

2.3 Interaktion und Datenerhebung

Im Folgenden wird der Ablauf (vgl. Abbildung 1) des Versuches für einen Teilnehmer kurz beschrieben. Es werden nur die Fragebögen erwähnt, deren Daten im Rahmen dieser Arbeit betrachtet werden. Die Interaktion und Erhebung fanden an einem Termin, mit einer Dauer von 75–90 Minuten statt. Der Teilnehmer erhielt eine Aufwandsentschädigung von 10,00 Euro.

Der erste Teil der Interaktion fand anhand einer Szenariobeschreibung statt, durch die der Versuchsleiter (VL) den Teilnehmer führte. Dabei gab der VL das jeweils nächste Ziel der Interaktion zwischen Nutzer und System, anhand eines Leitfadens, vor. Im zweiten Teil konnte der Teilnehmer die Spracheingabe nach Belieben ausprobieren, es gab keine vorgegebenen Aufgaben oder Aktionen. Insgesamt dauerte die Interaktion (beide Teile) 15–20 Minuten.

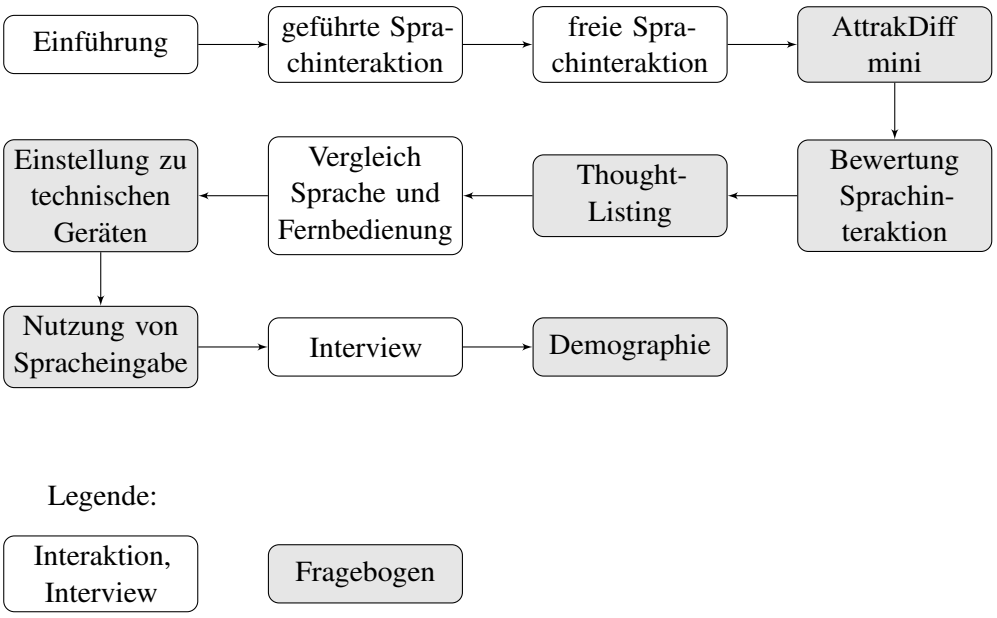

Abbildung 1: Ablauf der Studie für einen einzelnen Teilnehmer.

Das Szenario sah vor, dass sich der Nutzer über das anstehende Fernsehprogramm informiert, anhand verschiedener Kriterien Filme über VoD sucht, Filme in Youtube sucht und die Basisfunktionen (Kanalwechsel, Lautstärke) des Fernsehers per Sprache bedient.

Im Anschluss an die ausschließlich per Spracheingabe gesteuerte Interaktion wurden die Fragebögen *AttrakDiff mini* und ein Fragebogen zu *Beurteilung der Interaktion mittels Spracheingabe* (P.851) [4, 5] beantwortet. Da der P.851 ursprünglich für reine Sprachdialogsysteme ohne visuelle Systemausgaben entworfen wurde, wurden in der aktuellen Studie Items in denen explizit auf ein „Gespräch“ verwiesen wurde angepasst, d. h. die Items wurden so umformuliert, dass sie auf die „Interaktion“ Bezug nehmen (Abbildung 2 zeigt die Items des angepassten Fragebogens, als auch die verwendete Bewertungsskala). Der *AttrakDiff mini* ist eine verkürzte Version des AttrakDiff2 [3, S. 246–250]. Die deutsche Fassung ist in [2, S. 17 ff., S. 55] beschrieben.

Die beiden Fragebögen ermöglichen eine quantitative Auswertung der Nutzerurteile über die Spracheingabe und das System. Um auch qualitative Aussagen zu konkreten Eigenschaften (sowohl positive als auch negative) zu erhalten, wurde zum einen *Thought-Listing* [1, S. 253–255] und zum anderen ein *Leitfaden-Interview* (letzteres zum Abschluss des Termins) durchgeführt. Beim Thought-Listing sollte der Teilnehmer ca. 3–5 Gedanken, die ihm spontan zu der sprachgesteuerten Interaktion einfielen, in Form von Stichwörtern oder -punkten, aufschreiben. Im zweiten Schritt bewertete er jeden Gedanken dahingehend, ob es sich dabei um einen negativen oder positiven Aspekt vom dem System handelt.

Nach der Interaktion mittels Sprache und deren Bewertung fand ein weitere, vom VL geführte, Interaktion statt. Hier wurde der Nutzer aufgefordert eine Aufgabe zunächst per Sprache und anschließend per klassischer Fernbedienung zu erledigen. Direkt danach wurde der Teilnehmer gefragt welche Art der Bedienung er, bei sich zuhause, bevorzugen würde. Dies wurde jeweils für die folgenden Aufgaben durchgeführt: *Lautstärke schrittweise* ändern (von sehr leise kontinuierlich bis zu einer angenehmen Lautstärke), Lautstärke direkt wählen (von sehr leise auf 40 % der maximal möglichen Lautstärke), *Fernsehkanäle schrittweise* durchschalten („zappen“), *Fernsehkanal direkt* aufrufen (durch Nennen des Sendernamens), *EPG durchsuchen* (Für Fernbedienung und Sprache kamen 2 verschiedene Anwendungen zum Einsatz, welche die gleichen Programminformationen lieferten).

Im Anschluss an den Vergleichstest wurde eine neue Version des in [7] vorgestellten Fragebogens

Einstellung zu technischen Geräten (IKT-Fragebogen) sowie die *Nutzung von Spracheingabe*, während der letzten 6 Monate (s. Tabelle 4) abgefragt. Letztere wurde erhoben, um die Vorerfahrung der Nutzer mit sprachgesteuerten Diensten und Geräten zu erfassen. Der IKT-Fragebogen hat eine fünfstufige Skala mit den Polen *stimme überhaupt nicht zu* (1) sowie *stimme völlig zu* (5) und wird in seiner neuen Version demnächst publiziert. Zum Abschluss des Termins wurden demographische Daten (Alter, Geschlecht und höchster Bildungsabschluss) erhoben.

3 Ergebnisse

3.1 Bewertung der Bedienung mittels Sprache

Der mit P.851 erhobene *Gesamteindruck der Interaktion mittels Spracheingabe* (g) hat einen Mittelwert $\bar{g} = 3,64$ und Standardabweichung $\sigma(g) = 0,64$ auf einer Likertskala von 1 (Schlecht) bis 5 (Ausgezeichnet).

Die Bewertung verteilt sich wie folgt über die Skala: *Schlecht* (0 %), *Dürftig* (7,7 %), *Ordentlich* (28,2 %), *Gut* (56,4 %) sowie *Ausgezeichnet* (7,7 %). Die mehrheitlich gute Bewertung spiegelt sich auch in den Antworten auf die restlichen Items des Fragebogens wider, deren Mittelwert, Standardabweichung und prozentuale Verteilung in Abbildung 2 gezeigt sind. Bei den Items *Das System tat nicht das was ich wollte.* (1) und *Das System reagiert nicht immer wie erwartet.* (12) wurde das System am schlechtesten bewertet. Teilweise wurde die richtig erkannte Nutzeräußerung durch die NLU falsch interpretiert, teilweise war die Abbildung der Interpretation auf eine konkrete Funktion nicht nachvollziehbar. *Das System reagierte wie ein Mensch* (16) ist mit 2,52 tendenziell schlecht bewertet. In den Angaben zu den Items 27, 28, 35 und 36 zeigt sich, dass die Interaktion mit System insgesamt durch den Nutzer als gut und lohnenswert betrachtet wird.

Tabelle 2 listet die Ergebnisse des Thought-Listing. Von 207 erfassten Gedanken wurden 185 eindeutig als positiv oder negativ eingeordnet. Aus den 185 Stichwörtern und Kurzsätzen wurden thematische Kategorien gebildet und für jede Kategorie die Anzahl der jeweils positiven und negativen Gedanken berechnet (s. Tabelle 2). Einige der niedergeschrieben Gedanken enthielten 2 Aspekte, die in unterschiedliche Kategorien fallen, weshalb sich in Tabelle 2 eine Summe von 202 Gedanken ergibt. Insgesamt wurden 113 positive (p) und 89 negative (n) Gedanken erfasst. Für Gedanken zu den Aspekten *Spracheingabe* und *Gesamteindruck* überwiegt die Anzahl der negativen Gedanken. Bei den Aspekten *Effektivität*, *Innovation* und *Spaß* sind es die Positiven. Der Unterschied zwischen den Häufigkeiten ist für *Innovation* und *Spaß* signifikant. Alle weiteren Aspekte wurden selten genannt ($p + n < 9$) und könnten zu einer weitergehenden qualitativen Untersuchung herangezogen werden.

Die Mittelwerte der fünf Dimensionen des AttrakDiff mini lauten: 4,8 (*Pragmatische Qualität*), 5,2 (*Attraktivität*), 4,9 (*Hedonische Qualität - Identität*), 4,9 (*Hedonische Qualität - Stimulation*) sowie 4,9 (*Hedonische Qualität*). Diese Werte liegen durchweg auf der positiven Seite der siebenstufigen Skala des Fragebogens. Aus der pragmatischen und hedonischen Qualität leitet sich die Einordnung in der Portfolioansicht (s. Abbildung 3) ab. Dort ordnet sich das System am Schnittpunkt zwischen *neutral* und *begehrt* ein.

3.2 Präferenzen und Einstellung der Nutzer

Abbildung 4 zeigt die Häufigkeitsverteilung zur Angabe der Nutzer ob sie Sprache oder Fernbedienung zur Ausführung einer bestimmten Aktion mit dem Fernseher bevorzugen würden. Sprachsteuerung wird zur direkten Kanalwahl und der Suche nach Sendungen im EPG bevorzugt. Für das schrittweise wechseln der Sender („zappen“) wird die Fernbedienung bevorzugt. In allen 3 Fällen ist der Unterschied in den Häufigkeiten, zwischen Sprache und Fernbedienung, signifikant (Binominaltest, $\alpha = 0,05$, P(Fernbedienung) = $\frac{1}{2}$; $p < 0.01$ in allen Fällen). In allen

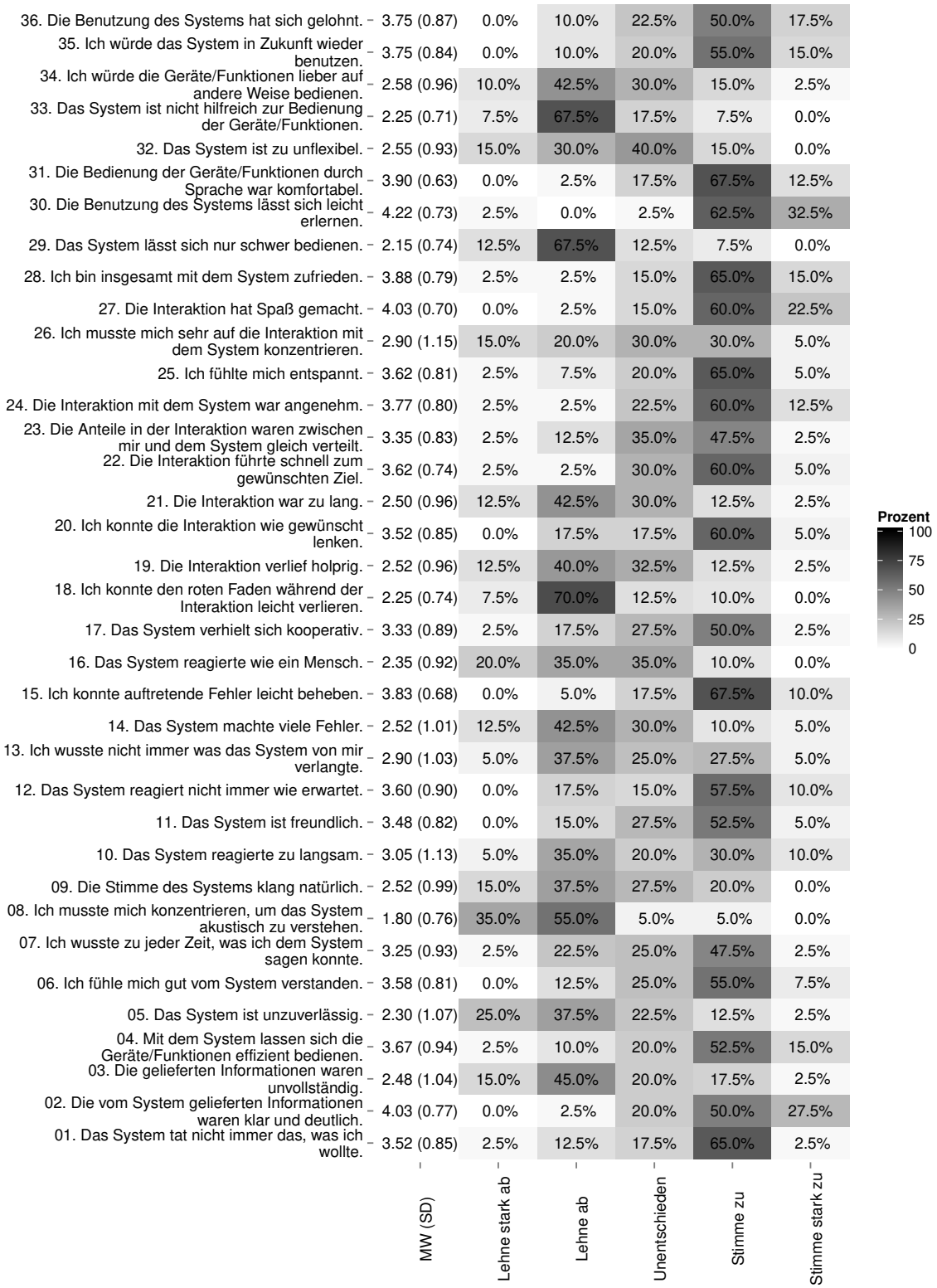

Abbildung 2: Mittelwert (MW) und Standardabweichungen (SD) sowie prozentuale Verteilung der Bewertung, der Interaktion per Spracheingabe, mit dem Fragebogen nach ITU-T Rec. P.851 [4].

Kategorie	p	n	$p+n$	Kategorie	p	n	$p+n$
Spracheingabe	16	28	44	Fernbedienung	1	2	3
Gesamteindruck	8	12	20	Begehrt	2	0	2
Effektivität	10	5	15	Fehler	0	2	2
Innovation*	15	1	16	GUI Eindruck	1	1	2
Spaß*	9	0	9	technische Voraussetzungen	1	1	2
Stimme	1	7	8	Hilfreich	2	0	2
Inhalte	5	3	8	Bildqualität	0	1	1
EPG	3	4	7	Eindruck GUI	0	1	1
Intuitivität	5	1	6	Fernsehen	1	0	1
Interessant	6	0	6	Lautstärke	1	0	1
Funktionsumfang	3	2	5	Smartphone	1	0	1
Internet	5	0	5	Bedienbarkeit	1	0	1
Smartphonenutzung	1	4	5	Angenehm	1	0	1
Kosten	1	3	4	TTS Feedback	0	1	1
Videothek	4	0	4	Modalitätenwahl	0	1	1
Komfort	4	0	4	Erlernbarkeit	1	0	1
Effizienz	0	4	4	Trend	1	0	1
Suchfunktion	3	1	4	Feedback	0	1	1
Smartphone App	0	3	3				
				Summe	113	89	202

Tabelle 2: Anzahl der positiven (p) und negativen (n) Gedanken, sowie die Summe ($p+n$) der beiden Werte. Die Einträge in der Tabelle sind absteigend nach der Größe von $p+n$ sortiert. Die Kategorien wurden aus den, von den Teilnehmern, niedergeschriebenen Gedanken zu der Interaktion gebildet.
*_Innovation_ und _Spaß_ weisen jeweils signifikante Unterschiede zwischen den Häufigkeiten in n und p auf (Binominaltest, $\alpha = 0,05$, P(positiv) = $\frac{1}{2}$).

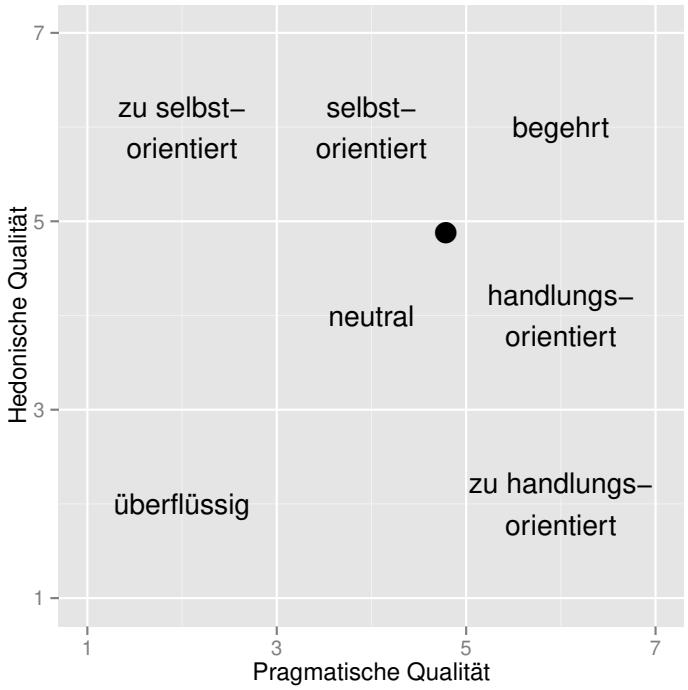

Abbildung 3: Portfolio Darstellung der Bewertung mittels AttrakDiff mini.

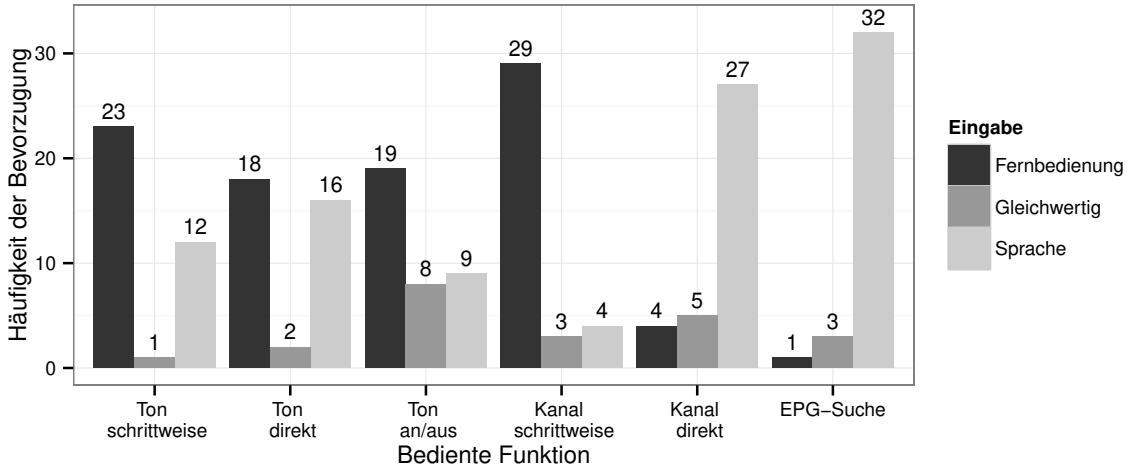

Abbildung 4: Absolute Häufigkeiten der bevorzugten Eingabemethode zu den einzelnen Aufgaben. *Gleichwertig* bedeutet, dass zwischen Sprache und Fernbedienung für den Nutzer kein Unterschied hinsichtlich der Bevorzugung bestand.

	Exploratives Lernen	Technikangst	Interesse an Technik	Markenorientierung	Privatsphäre	Serviceorientierung
$\bar{x}$	4,06	1,42	3,60	3,17	3,15	3,16
$\sigma(x)$	0,59	0,51	0,99	0,48	0,99	1,10

Tabelle 3: Mittelwerte ($\bar{x}$) und Standardabweichung ($\sigma(x)$) der Dimensionen des Fragebogens zur Erhebung der Einstellung gegenüber technischen Geräten.

Fällen zur Steuerung der Lautstärke (Ton), überwiegt die Bevorzugung der Fernbedienung, es kann jedoch kein signifikanter Unterschied gezeigt werden (Binominaltest, $\alpha = 0,05$).
Die Mittelwerte der Dimensionen zur *Einstellung gegenüber technischen Geräten* sind in Tabelle 3 aufgelistet. In den Ergebnissen zeigt sich, dass die Teilnehmer im Mittel gerne neue Geräte und Funktionen ausprobieren (*Exploratives Lernen*) und nur ein geringe *Technikangst* aufweisen.

3.3 Nutzungshäufigkeit von Spracheingabe

Die Befragung der Teilnehmer nach der Nutzung von Spracheingabe/-steuerung während der letzten 6 Monate (ca. 2. Halbjahr 2015) ergab, dass 15 der 40 Teilnehmer mindestens einmal Spracheingabe in dem Zeitraum verwendet haben. Die Anzahl der Nennungen für Gerätearten und geschätzte Häufigkeit der Nutzung zeigt Tabelle 4. Obwohl Mehrfachnennungen möglich waren, wurde immer nur ein Gerät von dem jeweiligen Teilnehmer angegeben. Die am häufigsten genutzte Geräteart ist das Smartphone (14 Nennungen), alle anderen Gerätearten wurden jeweils nur einmal genannt. Eine tägliche oder wöchentlich mehrfache Nutzung mittels Spracheingabe wurde zweimal angegeben.

4 Diskussion und Ausblick

In den Ergebnissen zeigt sich, dass die Steuerung eines Smart-TV per Sprache von den Nutzern akzeptiert wird und als gut sowie als hilfreich und komfortabel eingeschätzt wird. Insbesondere für Aufgaben bei denen die Nutzung von Sprache gleich aufwändig oder weniger aufwändig, im

	Gerät				
Häufigkeit	Smartphone	Tablet	Navi.	Tele.	∑
täglich	1				1
nicht täglich, aber mehmals pro Woche	1	1			2
höchstens einmal pro Woche	7				7
höchstens einmal pro Monat				1	1
seltener als einmal pro Monat	2		1	1	4
∑	11	1	1	1	15

Tabelle 4: Nutzungshäufigkeit von Spracheingabe und Art des verwendeten Gerätes während der letzten 6 Monate.

Sinne von notwendigen Interaktionsschritten, ist, wird Sprache als Eingabemodalität bevorzugt. Dies deckt sich mit früheren Untersuchung von Schaffer et al. [6]. Bei den Teilnehmern fällt auf, dass sie zum einen gerne neue Geräte ausprobieren und zum anderen nur eine geringe Technikangst aufweisen (s. Tabelle 3), allerdings nur 9 Teilnehmer wenigstens einmal im Monat Spracheingabe auf einem Gerät verwenden (vgl. Tabelle 4). Eine zukünftige – bereits in Planung befindliche – Studie, soll auf die Vorerfahrung der Nutzer ein noch stärker geachtet werden.

5 Danksagung

Ich bedanke mich für die Unterstützung bei der Planung und Durchführung der Studie bei meinen Kollegen Patrick Ehrenbrink und Dr. Benjamin Weiss. Die Arbeiten zu der vorgestellten Studie fanden im Rahmen des BMWi-geförderten Verbundprojekts Universal Home Control Interfaces@Connected Usability (UHCI), mit dem Förderkennzeichen 01MG13001G, statt.

Literatur

[1] CACIOPPO, J. T., C. R. GLASS und T. V. MERLUZZI: *Self-statements and self-evaluations: A cognitive-response analysis of heterosocial anxiety*. Cognitive Therapy and Research, 3(3):249–262, 1979.

[2] DIEFENBACH, S. und M. HASSENZAHL: *Handbuch zur Fun-ni Toolbox*. Techn. Ber., Folkwang Universität der Künste, 2011.

[3] HASSENZAHL, M. und A. MONK: *The Inference of Perceived Usability From Beauty*. Human-Computer Interaction, 25(3):235–260, Juli 2010.

[4] ITU-T REC. P.851: *Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems*. Techn. Ber. Supplement 851 to P-Series Recommendations, International Telecommunication Union, Geneva, Switzerland, 2003.

[5] MÖLLER, S., P. SMEELE, H. BOLAND und J. KREBBER: *Evaluating Spoken Dialogue Systems According to De-Facto Standards: A Case Study*. Computer Speech and Language, 21:26–53, Nov. 2007.

[6] SCHAFFER, S., R. SCHLEICHER und S. MÖLLER: *Modeling input modality choice in mobile graphical and speech interfaces*. International Journal of Human-Computer Studies, 75:21–34, März 2015.

[7] WEISS, B., I. WECHSUNG und S. MARQUARDT: *Assessing ICT user groups*. In: *Proceedings of the 7th Nordic Conference on Human-Computer Interaction*, S. 275–283, 2012.