
WHAT
CLOUDBASIERTE
I
SPRACHGESTEUERTE
YOU
CAN
HELP
ASSISTENZSYSTEME
WITH?

What Can I Help You With?

Cloudbasierte, sprachgesteuerte Assistenzsysteme

Proposal I

Vorgelegt von Christoph Claus
Matrikelnummer 11089074
BA Integrated Design

Eingereicht am 6. Juli 2017

Betreut durch Prof. Andreas Muxel
Interface Design

Technische Hochschule Köln
Fakultät für Kulturwissenschaften
Köln International School of Design


18:55
„Hey Siri du musst lernen mich besser zu verstehen"
Ok.
„Mach eine Notiz davon"
Hier ist deine Notiz:
Davon
Notiz
Zum Bearbeiten tippen

Inhalt

1

THE FUTURE

BACK TO

/.

Eine digitale Sprachassistentin tritt höchst personalisiert und aktiv auf. Sie stellt von sich aus Fragen, kennt den Charakter ihres Nutzers, seine Gewohnheiten, begleitet ihn 24 Stunden am Tag. Es entsteht eine Emotionalität zwischen Mensch und Maschine, die trotz ihrer Virtualität eine scheinbar greifbare Präsenz und Körperlichkeit besitzt. *Theodore*, die Schlüsselfigur, welche besagte Sprachassistentin namens *Samantha* als Knopf im Ohr stets verfügbar bei sich hat, verliebt sich schließlich in sie. Das rein digitale Aufeinandertreffen mindert die Erfahrung von *Theodore* nicht. Fiktion und Erlebbares verschwimmen.

Einführung in die Thematik

Was im Film *Her* noch dystopisch wirkt, zeigt eine vielleicht nicht mehr ganz so weit entfernte Zukunft. Die Praxis der sprachlichen Interaktion mit persönlichen, mobilen Geräten ist heutzutage keine Utopie mehr, sondern im Alltag angekommen und profiliert – das suggeriert zumindest das Überangebot. Fast monatlich wird mittlerweile eine neue Sprachassistentin vorgestellt. Allein während des Verfassens dieser Arbeit kam mindestens ein neues Familienmitglied in die bestehende Reihe: Djingo vom französischen Telekommunikationsunternehmen Orange. Anfang Mai präsentiert, fällt vor allem eins auf: Djingo unterscheidet sich, zumindest bisher, in keiner Weise von seinen Konkurrenzprodukten. Die Sprachassistentin soll ähnlich Amazons Alexa in Form eines Lautsprechers Einzug in private Haushalte halten. Dadurch fungiert Djingo nicht nur als persönliche Beraterin auf dem Smartphone, sondern erweitert die Smartness dessen auf eine ganzheitliche *User Experience*. Vor einigen Monaten hatte auch Samsung Bixby als Highlight seiner neuesten Smartphone-Generation vorgestellt. Auch darüber hinaus wird das Potpourri an Angeboten immer bunter. Amazon verkauft mittlerweile vier Lautsprecher-Systeme, welche Alexa als zentrale Schnittstelle besitzen. Das neueste Gerät Echo Show bietet dabei erstmals die Möglichkeit, zusätzlich zur Stimme über ein berührungsempfindliches Display visuell als auch manuell zu interagieren. Voraussichtlich diesen Sommer erreicht Google Home das europäische Festland und Apple versucht Ende des Jahres mit dem Siri-gestützten HomePod Schritt zu halten – und sich dabei bloß nicht abzugrenzen.

Apropos Apple: 2011 implementierte das Technologieunternehmen aus Cupertino erstmals das im Jahr zuvor eingekaufte Siri in das Betriebssystem iOS. Seit dem gibt es kaum nennenswerte oder evolutionäre Weiterentwicklungen. Hier und da entstehen weitere Sprachbefehle, Siri wird auf den Mac portiert. Die Funktion trifft entweder den Zeitgeist oder stillt eine Nachfrage, denn in der Zwischenzeit wühlen auch Microsoft, Amazon und Google auf einem umkämpften Markt mit eigenen Produkten mit. Open-Source-Projekte wie Mycroft ergänzen als kostenlose DIY-Alternative die Bandbreite. Dass die Technologie gerade jetzt Thema und im Fokus der IT-Giganten ist, mag zum einen an einer immer effektiveren und zuverlässigeren Interfacetechnologie liegen, vor allem aber an der Möglichkeit, enorme Datenmengen aus aller Welt zentral auf Servern abzulegen, auszuwerten und untereinander kommunizieren zu lassen. Doch warum nun diese Arbeit?

Eine Kernaufgabe des Designs ist es, bestehende Probleme zu identifizieren und gegebenenfalls auf diese zu reagieren, sie oder zumindest den Blickwinkel auf sie zu korrigieren. Nicht selten werden dadurch neue Probleme und Fragen generiert. Sprachassistentinnen können, an richtiger Stelle eingesetzt, essentielle Helferinnen sein, Prozesse beschleunigen und damit Aufgaben alternativ und klüger lösen. Nach dem aktuellen Wetter zu fragen macht jedoch weder in einer Situation Sinn, in welcher die Hände frei sind und die benötigte Information nur einen Klick auf eine *Wetterapp* entfernt ist. Noch macht es zwangsläufig Sinn in einer Situation wie dem Autofahren, bei der die Eingabe per Sprache zunächst zweckmäßig erscheinen würde, ein Blick aus dem Fenster aber genügt. Hier wäre es vielmehr wünschenswert, eine Routenplanung zuverlässig und ablenkungsfrei per Sprache umgesetzt vorzufinden, sprich ohne zwischendurch den Blick auf ein Display richten oder mit diesem sogar interagieren zu müssen.

Die Absurdität dieses Beispiels offenbart, dass die Funktionalität in Abhängigkeit der Situation, wo sie verwendet wird, nicht geklärt ist. Worin genau besteht der neue, zusätzliche Nutzen in der Funktion *Spracherkennung und -assistenz* bzw. welche spezifischen Probleme werden damit gelöst? Ist der Nutzen dieser erweiternden Funktionalität vielleicht noch nicht erschlossen und die Unternehmen versuchen lediglich ihren Platz am Markt zu erobern?

Da die derzeit erhältlichen Sprachassistentinnen nur teilweise mit zusätzlich zu erwerbenden Produkten verknüpft sind, ist ein ökonomischer Nutzen erst einmal nicht zwangsläufig offensichtlich, wenngleich die entfernt gesammelten Daten ein enormes monetäres Potential besitzen und in Zukunft den eigentlichen Wert darstellen könnten. Es handelt sich um eine neue Form der Beziehung zwischen Produzent und Konsument, wobei die wirtschaftliche Bedeutung erst nach dem Kauf eines Produktes entsteht. Der finanzielle Aspekt ist sicher ein wichtiger – diese Arbeit möchte sich dem Thema jedoch aus einer Designperspektive annähern und die Mensch-Maschine-Interaktion untersuchen.

„Die Grenzen meiner Sprache bedeuten die Grenzen meiner Welt."

↘ Ludwig Wittgenstein

Frage und Methodik

Ausgehend von der Annahme, dass gegenwärtige Sprachassistentinnen im Consumer-Bereich **gescheitert** und mehr **spielerisches Beiwerk des technischen Fortschrittes** sind, gilt es folgendes zu klären: Was sollte ein sprachgesteurtes Dialogsystem wirklich können und wie kann die Schnittstelle gestaltet sein?

Um diese Fragestellung zu beantworten, sollen problematische Momente im Zuge einer Analyse identifiziert werden. Damit sich die aktuelle Brisanz sprachgesteuerter Dialogsysteme besser einordnen lässt, soll zunächst die Veränderung des Interfaces, das sich offenbar weg von einer grafischen hin zu einer dialoggebundenen Schnittstelle bewegt, betrachtet werden. Dabei ist die Frage nach einer Qualität der menschlichen Stimme leitgebend. Für die spätere Untersuchung werden die drei derzeit am meist genutzten Sprachassistentinnen getestet: Ok Google bzw. Google Home, Alexa und Siri. Um diese nicht nur in ihrer finalen Funktionalität und Präsenz wahrzunehmen, werden Hintergründe der Entwicklung beleuchtet.
Ein erster persönlicher Test wird anschließend mittels spontaner Interaktion auf unterschiedliche Charakteristika der drei Testgeräte aufmerksam machen. Ein weiterer Test soll die Sprachassistentinnen in einer natürlicheren Umgebung und anhand eines vielfältigen Benutzerszenarios ausreizen. Besonders wichtig ist es dabei, Zufriedenheits- bzw. Frustrationsmomente aufzuzeigen. Wie genau sieht die Kommunikation und Interaktion zwischen Mensch und Maschine aus und vor allem wie hoch ist die Toleranzgrenze bei Abweichungen, fehlender Internetverbindung oder anderen Faktoren? Wie lange bewegt sich das Nutzererlebnis überhaupt auf auditiv-kommunikativer Ebene und ab wann wird eine händische Intervention notwendig? Die gewonnen Erkenntnisse werden in Bezug auf die Ausgangsfrage und -these ausgewertet, um abschließend einen Ausblick zu geben und eine Vision zu formulieren.
Ist der Endverbraucher-Bereich überhaupt eine erstrebenswerte Zielgruppe für Sprachassistentinnen? So könnte es sinnvoller sein, das Potential für spezifischere Kundengruppen weiterzuentwickeln. Für Menschen mit Beeinträchtigungen, die beispielsweise blind oder Einschränkungen im Bereich der manuellen Interaktion mit dem Computer unterworfen sind. Auch Arbeitsplätze, die nicht immer eine manuelle Interaktion möglich machen, wären ein weiteres denkbares Einsatzgebiet, welches es zu berücksichtigen gilt.

Abgrenzung

Wenn auch eng zusammenhängend und integraler Bestandteil von Sprachassistentinnen, möchte diese Arbeit nicht die Spracherkennung als solche untersuchen. Es geht dementsprechend nicht um Applikationen, in denen Sprache lediglich als Input aufgezeichnet und verarbeitet wird. Diese Arbeit legt den Fokus auf den interaktiven Moment, den Dialog zwischen Mensch und der künstlichen Intelligenz der Maschine und weniger auf explizite Funktionalitäten. Das mittlerweile eng mit Sprachassistenz verbundene Smart Home spielt im Rahmen dieses Proposals keine gesonderte Rolle. Außerdem werden Fragen und Bedenken zum Datenschutz, die oft im Zusammenhang mit konstant zuhörenden Assistentsystemen aufkommen, nicht abgehandelt.

Sprache

Auf die Frage »Are you female?« antwortete die weibliche Stimme von Google Home »I try to stay neutral«. Daraufhin wurde die Assistentin gefragt »Can you switch your voice to female?« worauf hingeweisen wurde »I‘ve just got the one voice« und nach kurzem Zögern »For now«.
Alle in dieser Arbeit untersuchten Sprachassistentinnen sind standardmäßig weiblich, nur Siri bietet seit längerer Zeit auch ein männliches Pendant an. Aufgrund dieses streitbaren und sehr geschlechtsspezifischen Umstandes werden alle Assistenzsysteme in der weiblichen Form beschrieben.
Darüber hinaus wird auf *political correctness* zur Gewährung des Leseflusses verzichtet und auf die männliche Form zurückgegriffen.

SPRECHE
ALSO BIN
ICH

2

ICH

/.1 Interfaces im Wandel

Noch bevor Sprache ein Thema in der Benutzerführung von Smartphones war, gab es bereits einen bedeutenden Wechsel in der Eingabesteuerung: Das Interface veränderte sich von der klassischen, jahrzehntelangen Bedienung per Maus und Tastatur hin zu einer taktilen Eingabe per Fingergesten. Diese Wende, angestoßen im Mobiltelefonbereich, hat sich reibungslos, rasch und mittlerweile auch im Desktopbereich vollzogen. Viele Hybridgeräte bieten die Möglichkeit, den Computer konventionell mit externer Peripherie, als auch per haptischer Berührung des Bildschirms zu bedienen.

Der Erfolg der berührungsempfindlichen Interaktion mit der Maschine liegt gewiss besonders an dem intuitiven und ganzheitlichen Interface, welches Apple 2007 mit der ersten Generation des iPhones grundlegend vorstellte und das in den darauffolgenden Jahren von vielen Unternehmen aufgegriffen und um weitere Gesten ergänzt wurde. Inzwischen unterscheiden sich derartige mobile Betriebssysteme in ihrer Bedienung kaum voneinander.

Die grafische Oberfläche selbst hat sich mit der Skalierung auf portable Geräte jedoch recht wenig verändert und orientiert sich nach wie vor stark an bisherige Konzepte. Da der Platz zur Darstellung, zumindest bis zur Einführung von Tablets, bedeutend kleiner war, mussten Funktionen und deren visuelle Implementierung reduziert werden. Interaktionen selber wurden anfangs durch dreidimensional anmutende und dadurch optisch hervorgehobene Buttons illustriert. Dieses Prinzip etablierte sich rasch, sodass Buttons innerhalb weniger Jahre ihre grafische Anmutung verloren und von nun an meist nur noch als reiner Text vorzufinden sind (→ Abb. Seite 16). Generell lässt sich sagen, dass Konsumenten die neue Handhabung zügig annahmen. Während der Absatz des konventionellen Handymarktes extrem einbrach, zogen daher anfangs fehlende oder eingeschränkte sowie neue Funktionen in die Smartphones ein. Damit sind sie in ihrem Umfang Desktoprechnern näher als je zuvor.

Dass das Smartphone innerhalb kürzester Zeit den herkömmlichen Computer gänzlich oder zumindest für einen Großteil des digitalen Alltags ersetzt, war vermutlich überraschend. Durch diese nicht zu erwartende Entwicklung gab es ungewöhnlich viele mikro- und makrofunktionelle Veränderungen in Bedienung und Oberfläche. Während der Finger die Maus ersetzt hat, wurde die Tastatur schlicht in ein digitales Analogon übersetzt. Beide Interaktionen sind zwar klug implementiert, bergen aber dennoch einige Hürden. So ist der menschliche Finger aufgrund seines Umfangs nicht annähernd so zuverlässig und exakt, was zu Fehlauswahl und -eingabe führen kann. Damit ist das Schreiben längerer Texte mühsam. Auch einen Cursor per Berührung zu setzen, um beispielsweise einen Textabschnitt zu markieren, ist umständlicher und zeitintensiver als mit einer Maus.

Abb. 1 und 2: Vergleich von iOS 5 (oben) und iOS 9 (unten). Die Schaltfläche »General« ist mittlerweile nicht mehr als abgegrenzter Button ersichtlich.

Ein weiteres Problem ist die Tatsache, dass Smartphones anfangs ausschließlich auf Haut reagierten. Ohne den direkten Kontakt, gab es keine Reaktion. Demnach war die Allgegenwärtigkeit im Winter mit Handschuhen mehr eine Scheingegenwärtigkeit. Mittlerweile bieten hier spezielle Eingabestifte und sogar berührungsempfindliche Handschuhe Lösungen.

Gleichwohl führt die ständige Verfügbarkeit des Smartphones zu der Frage, was geschieht, wenn die eigenen Hände zeitweise nicht zur Verfügung stehen. Spracheingabe könnte die Antwort auf diese Limitierung sein, indem es als Alternative und Erweiterung für eine nahezu barrierefreie Kommunikationsmöglichkeit fungiert.

Sprache, also die Ein- und Ausgabe des gesprochenen Wortes, ist kein Novum in der elektronischen Datenverarbeitung. Die Möglichkeit, Texte zu diktieren und als Tonspur abzuspeichern ist seit dem Einzug von Mikrofonen gegeben. Spezielle Programme konnten diese, vom Menschen laut vorgetragene Sprache, sogar in lesbaren Text umwandeln. In eine schriftliche Form übersetzt, mussten sie später lediglich wieder vom Menschen gelesen und verstanden werden. Sprachbefehle, als elementarer Teil der Sprachassistenz, sind jedoch davon abhängig, vom Computer als künstliche Intelligenz richtig interpretiert zu werden (↘ vgl. Meisel 2010: 15). Inwiefern sind sie damit technischen Grenzen in der Verarbeitung natürlicher und komplexer Sprache unterworfen?

/.2 Fiktion und Realität

Der Wunsch, Sprache zwischen Mensch und Maschine vom Monolog zum Dialog zu erweitern, wird von Filmemachern schon viel länger thematisiert, als es technische Mittel überhaupt möglich machen. Während das eingangs aufgezeigte Filmbeispiel *Her* aus dem Jahre 2013 eine denkbare Antwort auf mobile Sprachassistentinnen dokumentiert, bebildert Knight Rider ein ähnliches Phänomen. In der US-amerikanischen Serie der 80er Jahre führt der Protagonist Michael Knight eine abenteuerreiche Beziehung mit seinem sprechenden und denkenden Auto K.I.T.T. (↘ Dialog rechts und Abb. 3).

K.I.T.T.: Was für ein unverantwortliches Verhalten, mich mit dem Schlüssel im Zündschloss und offener Tür stehen zu lassen.

Michael Knight: *Tut mir leid, es wird nicht wieder passieren.*

Michael Knight: *K.I.T.T.?*

[KEINE REAKTION]

Michael Knight: *K.I.T.T.?*

Michael Knight: *Ich sagte, es tut mir leid. Was willst du von mir?*

[KURZE PAUSE]

K.I.T.T.: Etwas mehr Rücksichtnahme hätte ich schon verdient.

Michael Knight: *Ich habe andere Dinge im Kopf und wir haben noch viel zu tun.*

K.I.T.T.: Wenn ich nicht wäre, müssten Sie zu Fuß gehen.

Michael Knight: *Wenn du nicht wärst, käme ich nicht zu einem wichtigen Rendezvous. Das ist der wichtigste Teil meines Auftrags. Ich bin dir wirklich, wirklich sehr dankbar.*

K.I.T.T.: Unseres Auftrags!

Michael Knight: *Unseres Auftrags. Wieder gut?*

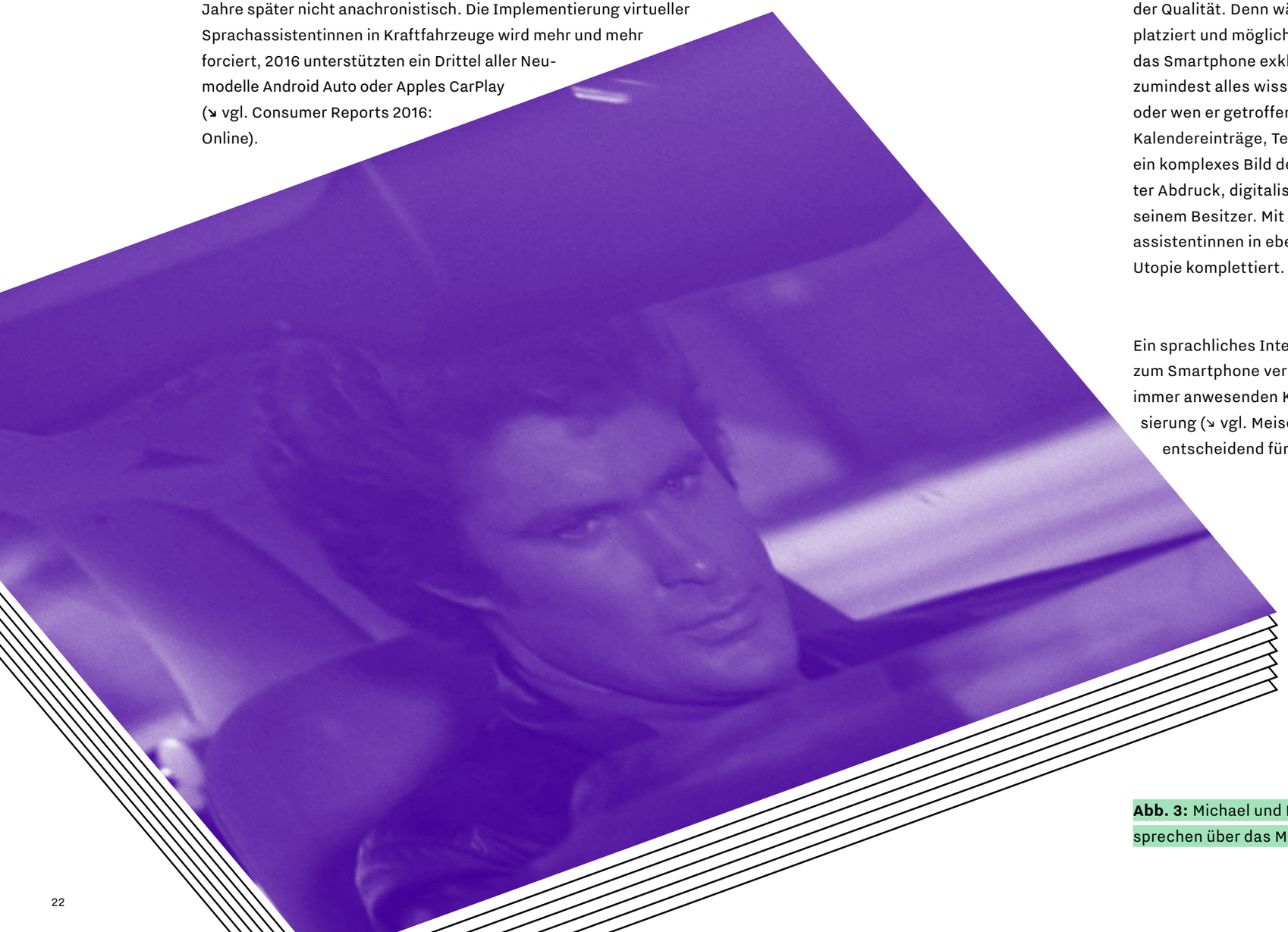
K.I.T.T.: Ich denke schon. Schließlich sind wir ja auch nur Menschen.

Michael Knight: *Übertreibst du nicht K.I.T.T.?*

K.I.T.T.: War nur so ein Gedanke.

↘ YouTube 2015: Online, Knight Rider Staffel 1 Episode 2

Die Dialoge offenbaren in ihrer Authentizität, Emotionalität und Spontaneität einen Traum des Menschen, mit Maschinen auf eine natürliche Art und Weise reden zu können und sie als stets verfügbaren, beratenden Freund und Weggefährten mit sich zu führen. Darüber hinaus wirkt selbst die Tatsache, dass das Auto als Objekt vermenschlicht ist, überraschenderweise auch 30 Jahre später nicht anachronistisch. Die Implementierung virtueller Sprachassistentinnen in Kraftfahrzeuge wird mehr und mehr forciert, 2016 unterstützten ein Drittel aller Neumodelle Android Auto oder Apples CarPlay (↘ vgl. Consumer Reports 2016: Online).



Eine Erweiterung des Selbsts

Mit dem Smartphone hielt die Idee eines Begleiters zumindest teilweise Einzug in die Realität. Mobil und stets erreichbar nimmt es mittlerweile einen Großteil des Alltags ein. Im Gegensatz zum klassischen Computer, der als Personal Computer seine persönliche Relation zum Nutzer im Namen trägt, übertrifft das Smartphone die Beziehung zu seinem Nutzer noch einmal in der Qualität. Denn während der PC tendenziell an einem bestimmten Ort platziert und möglicherweise von mehreren Menschen benutzt wird, so ist das Smartphone exklusiv und allgegenwärtig. Es weiß fast alles oder könnte zumindest alles wissen: Wo sein Besitzer gewesen ist, was er gemacht oder wen er getroffen hat. Es dokumentiert auf vielerlei Ebenen den Alltag. Kalendereinträge, Telefonate, Nachrichten und Facebook-Updates ergeben ein komplexes Bild der Person und ihrer Bewegung – ein hochgradig vernetzter Abdruck, digitalisiert und unsichtbar. So dient, gehorcht und gehört es seinem Besitzer. Mit der aktuell vorangetriebenen Integration der Sprachassistentinnen in eben diese Mobilgeräte wird die Vision aus Film und Utopie komplettiert.

Ein sprachliches Interface könnte die ohnehin schon abhängige Beziehung zum Smartphone verstärken, emotionalisieren und es zu einem wirklich immer anwesenden Kumpanen machen. Ob eine derartige Überpersonalisierung (↘ vgl. Meisel 2010: 10) gewünscht oder aber Störfaktor ist, kann entscheidend für Erfolg oder Misserfolg virtueller Assistentinnen sein.

Abb. 3: Michael und K.I.T.T. sprechen über das Menschsein.

/.3 Kein Fall für einen Handjob

Sprachgesteuerte Assistenzsysteme sind vor allem aufgrund technischer Bedingungen bisher wenig relevant gewesen. Damit einher geht der Eindruck, dass diese mehr Ergebnis der Umstände, denn strategische Innovation sind.
Nichtsdestotrotz sollen die Vorteile einer sprachlichen gegenüber einer manuellen Interaktion vor Augen geführt werden.
Die menschliche Stimme besitzt Qualitäten, die über die Metainformationen eines Textes hinausgehen. Aber wo sind die Grenzen der menschlichen Sprache als Eingabesteuerung?

Qualität der Sprache

Sprachassistentinnen (wenngleich K.I.T.T. wohl keine Dame war) in Automobilen, ob nun in der Filmgeschichte oder in der Realität, kommen nicht von ungefähr, sondern sind exemplarisch für eine mögliche, sinnvolle Integration. In Situationen wie dem Autofahren erscheint es klug, auf jedwede manuelle, nicht unbedingt erforderliche Interaktion zu verzichten, da die Gefahr der Ablenkung hoch ist. Ob es das Steuern einer Navigation, das Entgegennehmen eines Telefonats oder das Diktieren einer Nachricht ist – während der Fahrt zu sprechen ist eine Handlung, die die Aufmerksamkeit nicht unweigerlich von der Fahrbahn lenkt. Sprachsteuerung kann also durchaus eine Option sein. Umso wichtiger ist es, dass diese zuverlässig funktioniert, da ansonsten die Interaktion, im Gegensatz zum erprobteren Mensch-zu-Mensch-Gespräch, zum Störfaktor und damit ebenfalls zur Gefahr werden kann. Ebenfalls stellt die hybride Implementation von Sprache und grafischem Output eine kritischen Situation dar. Rezipiert der Autofahrer nicht nur auditiv, sondern visuell, führt das erneut zu einer Ablenkung vom Straßenverkehr.

Die eingeschränkte Benutzerfreundlichkeit des Touchscreens kann auch im Haushalt, so zum Beispiel beim Kochen und Arbeiten, um Spracherkennung ergänzt werden (→ Abb. 4). Während 70 Prozent die freihändige Bedienbarkeit als größten Vorteil virtueller Sprachassistentinnen angeben, schätzen 58 Prozent diese aufgrund von Zeitersparnis durch das gesprochene Wort (↘ vgl. Statista 2016: Online). Bevor manuell ein Input erreicht wird, muss das Telefon für gewöhnlich entsperrt, das richtige Programm gefunden, geöffnet und schließlich die Anfrage eingegeben werden. Überhaupt ist Geschwindigkeit ein einfach messbarer Wert. So kann der Mensch schneller sprechen als schreiben, jedoch ist Zuhören ein längerer Prozess als reines Lesen. Inwiefern das die Nutzerfreundlichkeit eines Voice User Interfaces erhöht oder sogar zum Verhängnis wird, bleibt abzuwarten.

Neben dem Vorteil einer weiteren Barrierefreiheit durch Sprache, die das Einsetzen des eigenen Körpers unnötig machen kann, ist es auch die intuitive Bedienung der Schnittstelle, die dank natürlichsprachlicher Dialoge ohne technisches Vorwissen bedient werden kann. Damit lässt sich auf digitale Daten zugreifen, ohne auf das Verständnis grafischer Oberflächen angewiesen zu sein.

Abb. 4: Das aktuelle Werbevideo zum Amazon Echo Show zeigt mögliche Nutzungsszenarien auf. Ab sofort ist Abwaschen kein Grund mehr nicht *connected* zu sein.

Problematisch sind dabei jedoch zwei Dinge. Zum einen müssen sprachgesteuerte Dialogsysteme oft mittels konkreter, vorher erlernter Befehle verwendet werden. Die tatsächliche intuitive Begegnung ist also stark von der Qualität der künstlichen Intelligenz abhängig. Zum anderen führt viel Sprachoutput von einem rein automatisierten Gegenüber zur nicht zwangsweise verlustfreien Verarbeitung. Das menschliche Kurzzeitgedächtnis ist limitiert und kann nur sieben plus/minus zwei Dinge abspeichern. Aus diesem Grund sind vor allem längere Antworten, mit möglicherweise mehreren Optionen und Vorschlägen, kritisch in einer rein akustischen Umsetzung (↘ vgl. Thar 2015: 85).

Qualität der Stimme

Unabhängig von diesen eher pragmatischen Gründen liegt ein weiterer Vorteil in einer immanenten Qualität der menschlichen Sprache. Gesprochenes hat neben einer emotional-verbalen Komplexität (Wortwahl, Wortphrasen) eine unverzügliche weitere Dimension. Wenn nämlich die Maschine den akustischen Input (langsam, hastig, schwer verständlich etc.) unterscheiden und Sprache damit vielschichtiger interpretieren und bewerten kann, ist auch der Umfang des Outputs, die Möglichkeit der Reaktion und Einordnung, ein größerer (↘ vgl. Harris 2005: 18).

Sichtbarkeit der Stimme

Die fehlende Möglichkeit zur manuellen Interaktion kann also eine Barriere innerhalb eines berührungsempfindlichen Interfaces darstellen. Aber Spracheingabe ist nicht immer möglich: Meetings, Konzerte mit lauten Umgebungsgeräuschen oder sehr private Botschaften sind ein Beispiel dafür, wann Sprache keine Alternative ist und auf die textliche Eingabe zurückgegriffen werden muss. In *Her* wirkt Theodores in aller Öffentlichkeit ausgelebte Zuneigung Samantha gegenüber auch wenig realistisch und zukunftsträchtig.

Überhaupt ist der öffentliche Raum als Einsatzgebiet von Sprachassistenz kritisch zu sehen. Ob aus Höflichkeit oder Selbstschutz, es kann und will nicht immer der Umgebung zugänglich gesprochen werden. Sensible und persönliche Themen wie Gesundheit und Sozialbeziehungen wollen geschützt sein. Je privater und komplexer die Sachverhalte, desto unwahrscheinlicher wird auch die Option, einen Informationsabruf gefolgt von einer Interaktion per Sprache folgen zu lassen.

Laut einer US-amerikanischen Studie aus dem Jahr 2016 benutzen demnach auch 80 Prozent der 1800 Befragten Sprachassistentinnen in privaten Bereichen, entweder in den eigenen vier Wänden oder im Auto (↘ vgl. Search Engine Land 2016: Online). Die noch fehlende Verbreitung, Akzeptanz und damit Alltagstauglichkeit führen 20 Prozent von Nicht-Nutzern als Gegenargument an, die sich schlicht unwohl fühlen, mit der eigenen Technik (Selbst-) Gespräche zu führen (↘ vgl. Business Insider Deutschland 2016: Online).

> people can speak much faster than they can type

↘ Phillips et al. 2016: 32

but can read much faster than they can listen.

/.4 Vom Interfacewandel zum Paradigmenwechsel

Dass Sprachassistentinnen in ihrer heutigen Form zuallererst auf dem Smartphone vorgestellt wurden, ehe sie auf weiteren Plattformen wie dem Computer und Smart Home verfügbar waren, mag am digitalen Paradigmenwechsel (↘ vgl. Meisel 2010: 10) der ständigen Verfügbarkeit liegen, den das Smartphone mit sich bringt.

Context is King

Gewissermaßen immer mit dem Internet verbunden, birgt diese Veränderung kontextuell eine gänzlich neue Bedeutung: Wo und wann wird mit dem Gerät interagiert? Die Abfrage von Lokalität und Zeit lässt gänzlich neue Situationen zu, in denen bisher auf händische Weise nicht mit dem Gerät kommuniziert werden konnte. Dass diese Sprach-Funktionalität inzwischen mehr und mehr portiert wird, ist ein ähnliches Phänomen wie das der Verbreitung des Touch-Interfaces.

Ob die derzeitige Omnipräsenz von Sprachsteuerung einen weiteren Paradigmenwechsel im Bereich der primären Bedienung bringen wird, bleibt abzuwarten. Er ist vor allem von der Zuverlässigkeit der Sprachanalyse im direkten Vergleich zu bisherigen Kommunikationsmöglichkeiten abhängig, die von 78 Prozent Befragten einer Studie von April 2016 als unzureichend attestiert wird (↘ vgl. Business Insider Deutschland 2016: Online). Die derzeit noch erforderliche Verbindung zum Internet könnte Verbreitung und Akzeptanz in der Gesellschaft verlangsamen oder verhindern. Sprachanfragen greifen nicht auf eine künstliche Intelligenz in der Gerätesoftware selber zurück, sondern werden in zentrale, leistungsstarke Serverstrukturen übertragen und dort verarbeitet. Das bedeutet zum einen, dass das Gerät permanent online sein muss und gegebenenfalls Datenvolumen benötigt wird. Zum anderen wird die Stimme des Nutzers extern interpretiert. Die damit verbundenen Datenschutzbedenken geben 62 Prozent als Argument gegen eine Nutzung an (↘ vgl. ebd.).

/. Zusammengefasst

Sprache ist derart komplex und intuitiv, dass eine Technisierung dieser unweigerlich enttäuschen muss. Während Dystopien aus der Filmlandschaft eine wirkliche Kommunikation zwischen Mensch und Maschine zeigen, also ein empathisches Aushandeln der Situation und der Beteiligten, bleiben die Dialogsysteme vor allem auf proaktive Interaktion beruhende Assistentinnen.

Ein echter Paradigmenwechsel, der sich von der menschlichen Fähigkeit, Technik zu verstehen, hin zu der Fähigkeit der Technik, Menschen zu verstehen, bewegt, kann entscheidend bei Erfolg und Akzeptanz von Sprache als wahre kommunikative Interaktion von Mensch-Maschine-Entitäten sein. Die Verbesserung der künstlichen Intelligenz wird maßgeblich dazu beitragen.

Bisherige Für und Wider von Sprache und Sprachassistenz haben Qualitäten und Schwächen offengelegt. Während also positive Aspekte im zusätzlich funktionellen Nutzen liegen, sind die Nachteile primär den Rahmenbedingungen zuzuordnen. Mit der Vorstellung des Hybriden Echo Show ist auch der Versandhandelgigant Amazon einen Schritt zurück gegangen und legt die Option der Benutzereingabe in die Hände der Nutzer. Diese profitieren mit dem dazugekommenen Display auch von der Möglichkeit des visuellen Feedbacks, welches in mancher Situation schneller als Sprachausgabe zu erfassen ist. Die nahtlose Integration mehrerer, multimodaler Schnittstellen in die Geräte scheint derzeit einen Kompromiss darzustellen.

OK GOOGLE, WAS UNTER-SCHEIDET EINE SIRI VON EINER ALEXA?

/.1 Zahl ein, Zahl aus

50 Prozent aller Suchanfragen sollen bis zum Jahr 2020 über die reine Spracheingabe erfolgen (↘ eresult 2017: Online). Was zum einen wie die Schlagzeile der aktuellen Bild-Ausgabe klingt, erscheint zum anderen einfach utopisch. Besonders der Umstand, dass diese Zeilen in einer Bibliothek niedergeschrieben werden, in welcher eine Vielzahl Studierender unter Zuhilfenahme von Suchmaschinen recherchiert, ist die Vorstellung, dass dies in bereits wenigen Jahren lautstark erfolgen soll, besonders komisch. Um eine bessere Vorstellung von der tatsächlichen Verbreitung zu bekommen, sollen zunächst aktuelle Studien und Zahlen zusammengefasst werden.

Um die Verbreitung virtueller, persönlicher Assistenten (VPA) voranzutreiben, sind diese mittlerweile oft in aktuelle Betriebssysteme vorinstalliert und mit der ersten Inbetriebnahme aktiv. Schon durch diesen Umstand kann ihnen eine gewisse Popularität und Alltagspräsenz zugesprochen werden.
Im Horizont der eigenen Wahrnehmung sind Gespräche mit den unsichtbaren Alltagshelfern in der Öffentlichkeit ebenso selten zu beobachten wie im eigenen Freundes- und Bekanntenkreis. Doch im Gegensatz zu diesen subjektiven Feststellungen, prophezeien Marktforscher, dass innerhalb der nächsten zwei Jahre bis zu 20 Prozent aller Interaktionen auf dem Smartphone per Assistentin durchgeführt werden (↘ vgl. Gartner 2016: Online). Dabei sind jedoch auch textbasierte Interaktionen eingeschlossen, wie Transaktionen und Bestellungen per Chatbots.

Einmal und nie wieder

In den USA und im Vereinigten Königreich nahmen im Jahr 2016 42 Prozent bzw. 32 Prozent von Smartphone-Nutzern die Sprachassistentin ihres Telefons in Anspruch, wobei im Schnitt mehr als 37 Prozent der Befragten diesen mindestens einmal täglich nutzten (↘ vgl. ebd.). Siri und OK Google sind am weitesten verbreitet: 45 Prozent der Befragten gaben an, Siri zu benutzen, OK Google wurde von 41 Prozent im Vereinigten Königreich und 48 Prozent der US-Amerikaner in den letzten drei Monaten genutzt. Eine weitere Umfrage untersucht die Häufigkeit von OK Google und Siri und deren Verwendung in der US-amerikanischen Öffentlichkeit. Von fast ausnahmslos allen Befragten wurde die vorinstallierte Assistentin schon ausprobiert, jedoch wird sie anschießend nur noch von etwas mehr als der Hälfte gelegentlich genutzt. Während dies bei zwölf Prozent der Android-Nutzer auch in der Öffentlichkeit geschieht, trauen sich das nur drei Prozent der Siri-Nutzer (↘ vgl. Google-WatchBlog 2016: Online).

In einer in Deutschland Anfang 2017 durchgeführten Umfrage gaben 53 Prozent der 2372 Befragten an, gar keine Sprachsteuerung zu benutzen (↘ vgl. statista 2017-1: Online). Von den aktiven Nutzern wird die Sprachsteuerung zu 36 Prozent auf dem Smartphone, 16 Prozent auf dem Computer und zu 13 Prozent im Auto verwendet. Smart Home und Entertainment-Geräte bilden mit jeweils drei Prozent das Schlusslicht. Dabei verwenden 70 Prozent der deutschen iOS-Nutzer gelegentlich Siri, fast gleichauf mit 62 Prozent der Android-Nutzer mit der Sprachassistentin von Google. Alexa ist in der Form

von Amazons Echo erst seit Februar 2017 in Deutschland verfügbar, weswegen noch keine repräsentativen Zahlen zur Verfügung stehen. Bis zum Jahr 2020 werden allerdings ganze sieben Milliarden dieser Geräte auf dem Markt vorausgesagt, die mehrheitlich mit minimalem oder fehlendem haptischen Interface auskommen (↘ vgl. Gartner 2016: Online).

Sprach-interfaces als Blackbox

Viele Zahlen sind es zwar, aber präzise Aussagen lassen sich nicht daraus schließen. Interessant sind konkretere Zahlen zur genauen Häufigkeit der Nutzung und zu den dann spezifischen Anwendungsfällen. In den USA zeigt sich zumindest bei der Verwendung von Alexa, in welch geringem Umfang ihre Funktionalität ausgeschöpft wird. 57 Prozent der Alexa-Nutzer fragen die Assistentin lediglich nach dem Wetter. Das Steuern der Musikwiedergabe kommt knapp dahinter mit 54 Prozent. Die Einkaufsfunktionen werden von nur elf Prozent (Produkte) bzw. acht Prozent (Lebensmittel) genutzt (↘ vgl. statista 2017-2), wobei sie durch die Verknüpfung zum Amazon Onlineshop ein Alleinstellungsmerkmal darstellen. Dass Funktionen nicht genutzt werden, kann an der fehlenden Dokumentation der Möglichkeiten liegen (→ Abb. 5, 6 und 7), aber auch an der Tatsache, dass (verfügbare) Optionen nicht sichtbar sind. Oder ist der Mehrwert einfach doch nicht gegeben?

Abb. 5, 6 und 7: Die Assistentinnen werden noch wenig gefordert – liegt es an den beiliegenden Dokumentationen, die wenig über das Funktionsspektrum verraten oder an einer geringen Nützlichkeit?

/.2 G

Es gab Zeiten, da waren Internetforen voll mit Nerven strapazierenden Fragestellern. Statt selber flink zu recherchieren, verbrachten sie lieber Zeit damit, Fragen zu formulieren, deren Antworten nur eine Suche weit entfernt waren. Die Gefragten reagierten voll Häme mit einem Link als Antwort, der diese phlegmatischen Zeitgenossen vorführte: let me google that for you (LMGTFY).

Abb. 8: LMGTFY erklärt, wie die konventionelle Suchmaschine von Google zu bedienen ist.

Lediglich Faulheit den Hilfesuchenden vorzuwerfen funktioniert nicht, haben sie alle immerhin die Zeit investiert, ein passendes Forum zu finden. Vielleicht ist es das nicht vorhandene Talent, Suchmaschinen zielgenau zu bedienen (→ Abb. 8)? Also warum die Mühe, wenn stattdessen einfach Google sprichwörtlich gefragt werden kann? Ob nun bewusst als Zielgruppe im Hinterkopf oder nicht: Google sah in seiner Datenbank ein gewaltiges Potential an alternativen Zugriffsmöglichkeiten.

WWW+1

OK Google, Google Now, Goole Voice Search – die Spracherkennung und -assistenz von Google hat viele Namen. Auch wenn sich die Implementierung je nach Betriebssystem und Hardware unterscheidet, haben die Bemühungen des Suchmaschinenmonopolisten den gleichen Hintergrund. Den größten Vorteil, die Suchmaschine sprachlich mit Fragen zu füttern, sahen die Entwickler in der Interpretation einer Suchanfrage im Zusammenhang ihres Kontextes. Wo, wann, wer und was waren dabei die treibenden vier W-Fragen. Grundlegend für diese Idee war die Annahme, dass Nutzer am Desktop-PC *andere und anders* Dinge suchen als wenn sie unterwegs sind (↘ vgl. Schalkwyk et al. 2016: 88). Im Moment des *sich-auf- dem-Weg-Befindens* soll Sprache dadurch als »natural input method for search on mobile devices« (↘ Meisel 2016: 8) dienen. Während Google Home eine recht statische Hilfe ist, wurde die explizite Voice Search (Applikation) vor dem Hintergrund der steigenden digitalen Mobilität entwickelt.

Mit dem Aufkommen von Suchmaschinen und deren Integration in den Alltag entwickelten Benutzer eine eigene schriftliche Ausdrucksform der sprachlichen Instruktion. Suchanfragen wurden über, teils mehrere, Schlagwörter definiert, oft nicht zusammenhängend, aber präzise. Um Öffnungszeiten zu eruieren, kann beispielsweise schlicht nach *th köln bib öffnungszeiten* gesucht werden. Ein erster Vorschlag verlinkt auf die Zweigstelle in Deutz inklusive einer Anfahrtsbeschreibung mit Google Maps. Der vierte Treffer verweist auf die Bibliothek in der Südstadt, in der gerade diese Zeilen niedergeschrieben werden. Mittels weniger Worte und der Möglichkeit, Inhalte schnell visuell zu erfassen, können innerhalb kurzer Zeit Informationen gefunden werden.

Abb. 9: Die klassische Google-Suche macht es leicht, Ergebnisse visuell zu filtern oder zu editieren.

Es ist eine große Übersetzungsleistung, eine ähnlich effiziente *User Experience* über Sprache zu ermöglichen. Gesprochenes ist komplexer, spontaner und in der Wortwahl variabler.

Bei der textgebundenen Suche fällt die (In-)Toleranz weniger ins Gewicht, da hier schnell eine Auflistung mehrerer Resultate durch den Nutzer gescannt oder ein Suchbegriff editiert und näher präzisiert werden kann (→ Abb. 9). Die Komplexität dieses Vorgangs hat auch Google erkannt und seine virtuelle Sprachassistentin von Anfang an multimodal gestaltet. Die Kombination von sprachnatürlichem Input und grafischem Output deklariert Google als flexibel, informationsreich und zeitsparend. Die visuelle Darstellung der Ergebnisse ermögliche einen »much richer information flow« (↘ Schalkwyk et al. 2016: 65).

Google hat seine Mächtigkeit als Marktführer genutzt und stetig Live-Daten der Sprachsuche gesammelt und ausgewertet. Auf welche Art und Weise werden Suchanfragen in welchen Situationen gestellt? Dabei war zu beobachten, dass im englischsprachigen Raum Auskünfte vor allem als Frage formuliert wurden. Es zeigt die Tendenz, dass sprechende Benutzer direkt einen Dialog initiieren, statt eine einseitige Suchanfrage zu stellen. Dass bei Anfragen auf eine Suchmaschine zurückgegriffen werden kann, führt zu einer höheren Diversität bei den Antworten. Es gibt nicht nur *eine* Wahrheit, die als Reaktion zurückgegeben wird.

»The linguistic possibilities are endless and complex semantic parsing capabilities would be required [...]«

↘ Schalkwyk et al. 2016: 84

/.3

Genau wie der Smartphone-Hype mit dem iPhone 2007 begann, so initiierte Apple ebenfalls den Run auf virtuelle, persönliche Sprachassistentinnen, als es im Oktober 2011 Siri als neues Feature des iPhone 4S vorstellte. Was viele jedoch nicht wissen, ist die Tatsache dass die sich selbst als »humble, personal assistant« vorgestellte Gesprächspartnerin 1,5 Jahre zuvor für zwei Monate als eigenständige Software von einem kleinen Startup in Apples App Store angeboten wurde (→ Abb. 10).

Das 2007 gegründete Start-Up wollte die Art, wie auf Informationen im Netz zugegriffen wird, wie diese verarbeitet werden und welche Optionen sich daraus ergeben, neu formulieren. Die Vision der Entwickler war es, die Persönlichkeit des Nutzers von Siri mehr und mehr zu verinnerlichen, seine Gewohnheiten zu lernen und sich in dessen Leben zu integrieren und proaktiv daran teilzunehmen. Siri sollte, anders als bei Googles Voice Search, keine Suchmaschine sein: »The idea is not to ask one question and get an answer, but to have the assistant proceed with me in a conversation and go and do things for me« (↘ Huffington Post 2013: Online). Siri sollte vielmehr aus sich heraus auf Dinge hinweisen (»Dein Zug ist ausgefallen«) und daraus folgend Aktionen vorschlagen, die sie selbst durchführen kann (»Soll ich nach alternativen Verbindungen suchen und Tickets buchen?«).

Abb. 10: Die ursprüngliche Siri-App war schon bei der Eingabemöglichkeit multimodal.

Um diese Komplexität zu gewährleisten, war die erste Version von Siri mit 42 Webservices und -seiten verknüpft, unter anderem Yelp! und Rotten Tomato. Durch diese Vielseitigkeit an Quellen konnte es die bestmöglichen Resultate und Optionen finden und dem Benutzer vorzuschlagen. Von den CEOs des Start-Ups, Tom Gruber, Dag Kittlaus und Adam Cheyer, wurde die Popularität von Smartphones als initialer Faktor gesehen, die Art und Weise, wie mit diesen Geräten interagiert wird, grundlegend zu ändern (↘ vgl. ebd.).

Proaktiv und individuell

Die Eingabe per Finger war für die Entwickler von Siri gerade bei komplexeren Sachverhalten mühsam, auch das Laden von Applikationen und Webseiten wurde als störend empfunden. Sprache, als die natürlichste, einfachste und ergiebigste Form der Kommunikation zwischen Menschen, sollte die Kommunikation zwischen Mensch und Maschine erneuern, nun da die technischen Voraussetzungen dafür geschaffen waren. Im Zentrum sollte dabei nicht ein einseitger Monolog stehen, ob nun in Richtung der Maschine (»Siri, wie spät ist es?«) oder in Richtung des Benutzers (»Du hast gleich einen Termin.«), sondern einen Dialog initiieren, der während der Konversation zu Entscheidungen und daraus direkt folgenden Handlungen führt. Die Entwickler träumten sogar davon, die Sprachausgabe auf längere Sicht an die des Benutzer und seiner Sprechweise anzupassen (↘ vgl. ebd.).

Während Googles Spracherkennung ursprünglich dafür entwickelt wurde, Fragen schnell zu beantworten, sollte Siri gewissermaßen *mitdenken*. So konnte die Sprachassistentin auf die Frage »Siri, ich bin betrunken. Wo habe ich geparkt?« aus Vorsicht und Schutzmaßnahme direkt einen Fahrdienst zum Nutzer bestellen (↘ vgl. ebd.). Damit kommt es dem Konzept der Selbsterweiterung sehr nah, da die Assistentin so alltägliche und geistlose Arbeiten übernimmt. Die in Siri steckende künstliche Intelligenz, die in einem vorherigen Forschungsprojekt namens CALO von den Gründern von Siri entwickelt wurde, war sogar fähig, anhand von Termineinträgen, allen Teilnehmenden Erinnerungen zu schicken, wichtige Unterlagen für das bevorstehende Meeting zu sortieren und diese schnell verfügbar zu machen. Großer Wert wurde auf eine hohe Toleranzgrenze bei der Spracheingabe gelegt. Da Siri nicht auf einem linguistischen Konzept beruhte und es damit keine abhänge Verbindung von Subjekt, Objekt und Verb gab, sollten Befehle exakt interpretiert werden können (↘ vgl. ebd.), auch wenn Teile eines Satzes verloren gehen sollten. Um dennoch Misskommunikation auf sprachliche Ebene vorzubeugen, erlaubte Siri auch eine nahtlos textliche Eingabe.

Umso enttäuschter waren die ursprünglichen Investoren, als Apple kurz nach der Veröffentlichung der ersten Version der Sprachassistentin das Unternehmen akquirierte und Monate später Siri in einer stark beschnittenen Form in das hauseigene iOS portierte. Wenngleich die Sprachassistentin nun nahtlos in das Betriebssystem integriert war und mit hauseigenen Apps zusammenarbeitete, so wurde die Kooperation mit vielen anderen Diensten (vorerst) gekappt. Die eigene Persönlichkeit und der geistreiche Humor, beides wichtiger Bestandteil von Siri nach Meinung ihrer Entwickler, wurde auf ein Minimum reduziert, ebenso wie das reiche Vokabular (↘ vgl. ebd.).

2012 wurden nach und nach wieder Features implementiert, bis Apple im Jahr 2016 schließlich eine API vorstellte, mit deren Schnittstelle Entwickler die Spracherkennung in eigene Apps implementieren konnten. Doch auch wenn durch die damit einhergehende *Featuritis* die Anzahl der Befehle gestiegen ist, heißt dies nicht zwangsläufig eine optimale User Experience. Die ursprüngliche Idee einer »do engine« (↘ ebd.) setzt die Verknüpfung und Kommunikation verschiedener Apps und Services voraus, die durch eine offene API und einer fehlenden Kuration des dadurch entstehenden Angebotes nicht gegeben sein muss. Damit orientiert sich die App mehr denn je an ein Konzept ähnlich dem von Google *Frage mich etwas* und entfernt sich von der Vision einer aus sich heraus handelnden Entität. In der Keynote fasst Phil Schiller von Apple ebenfalls das Konzept der *neuen* Siri als »Your intelligent assistant that helps you get things done just by asking« (↘ Youtube 2011: Online) zusammen.

Auf der WWDC 2017 kündigte Apple eine neue Version von Siri an, deren Beschreibung etwas Hoffnung auf die ursprüngliche Idee einer *echten Assistentin* verheißt. Die weibliche Stimme von Siri (seit iOS 7 gibt es auch eine männliche Stimmenoption) heißt übrigens Samantha, wie die Stimme im Film *Her*.

/.4

Als in einer TV-Sendung im amerikanischen Fernsehen ein Beitrag über Amazons Echo gezeigt wurde, war den Verantworlichen wohl nicht die Konsequenz der Ausstrahlung bewusst – oder vielleicht doch gerade? Der Beitrag wiederholt den Befehl eines kleinen Mädchens, welches Alexa bittet mit ihr zu spielen. Alexa ordert prompt ein Puppenhaus auf Amazon.com. Wenn es auch den Redakteuren eine witzige und vor allem dreiste Anektode schien, waren andere Echobesitzer erzürnt. Deren Geräte lauschten mit ihren hochsensiblen Mikrofonen ebenfalls dem Fernseher und führten den so gehörten Befehl in vielen Haushalten in den USA aus. Ein Shitstorm war die Folge (↘ vgl. tagesschau 2017: Online).

Amazons Sprachassistentin wurde seit 2011 komplett intern in Eigenregie entwickelt. Leider ist wenig über genaue Hintergründe bekannt, aber die Entwicklung ging zeitlich einher mit der Vorstellung von Siri und dem damit aufkommenden Interesse an Sprachassistentinnen in der öffentlichen Wahrnehmung und auch bei Konkurrenzunternehmen. Der Onlinehändler hatte innerhalb der mittlerweile mehr als 20-jährigen Geschichte schon viel Erfahrung mit künstlicher Intelligenz sammeln können. Durch die enorme Kundenbasis und den damit resultierenden Daten wurden schon früh selbstlernende Algorithmen, wie die auf Amazon bekannten Empfehlungen, umgesetzt. Durch den eigenen Cloud-Dienst Amazon Web Services war das Unternehmen unabhängig, effizient und konnte so bei der Gestaltung von Echo auf diese Ressourcen zurückgreifen. Die damit entwickelten »deep learning models« ermöglichten eine intelligente Sprachassistenz, die »trained on massive amounts of data« (↘ All Things Distributed 2016: Online) war.

Sprache als Ökosystem

Amazons Sprachassistentin Alexa besticht durch eine offene API und wird auch als »neutral ecosystem« (↘ Fortune 2016: Online) vermarktet. Entwickler und Interessierte können damit zum einen die Fähigkeiten von Alexa um sogenannte Skills erweitern und diese auf der Webseite von Amazon zur Verfügung stellen. Zum anderen kann Alexa auch in andere Gerätehardware portiert werden und ist damit gewissermaßen unabhängig von Amazons eigener Echo-Produktpalette. Bis auf das Anfang Mai vorgestellte Echo Show bieten die Echo-Produkte keine grafische Benutzeroberfläche an und werden ausschließlich über natürliche Sprache gesteuert.

Die radikale Reduktion auf ein sprachlich gesteuertes Interface sowie die stationäre Verortung unterschieden Alexa von den Sprachassistentinnen, die zur initialen Vorstellung erhältlich waren. Wo Googles Sprachsuche und Apples Siri bewusst für das Smartphone und dem Kontext des Unterwegssein entwickelt wurde, ist Alexa den Bedürfnissen in den eigenen vier Wänden angepasst. Das Fehlen eines visuellen Feedbacks erfordert eine höhere Zuverlässigkeit in der Spracherkennung und Sprachausgabe. Alexa sollte nach dem Wunsch von Amazon sehr gesprächig sein: »It should be just like we‘re talking to each other right now.« (↘ ebd.) Ziel war es, nicht nur ein *Voice User Interface*, sondern ein *Conversational User Interface* zu entwickeln.

In sogenannten *Wizard of Oz* Experimenten übernahm eine Vielzahl an Menschen die assistierende Funktion von Alexa. Es mussten Fragen einer sich in einem anderen Raum befindenden Person beantwortet werden. Die Antworten wurden über text-to-speech an ein Echo ausgegeben. Über die Einschätzung und Zufriedenheit des Fragestellers sollte so herausgefunden werden, welches Feedback am optimalsten war. Interessant an dieser Vorgehensweise ist, dass es Amazon primär um die Sprachausgabe ging und weniger um eine hohe Variabilität an Fragen oder Befehlen. Hier lässt sich nur vermuten, dass dazu entweder ebenfalls Testszenarien stattfanden oder die Entwickler bereits zufrieden mit der Erkennungsrate waren. Als Schlüsselqualität wurde vor allem die Latenzzeit zwischen Frage und Antwort definiert, die es von drei auf eine Sekunde zu reduzieren galt, um Vorreiter in Sachen Schnelligkeit zu werden. Damit scheint Amazon zu einem anderen Hauptkriterium der Nutzerzufriedenheit gelangt zu sein als Google, die bei der Entwicklung betont haben, dass vielmehr eine »strong positive relationship between recognition accuracy and the probability that a user returns, more so than other factors we considered – latency, for example« (↘ vgl. Schalkwyk et al. 2016: 88) exisitiert.

Wenngleich Echo nicht lediglich als »Gadget, but instead [as – Anm. d. Verf.] a full end to end service« (↘ Fortune 2016: Online) von seinen eigenen Machern verstanden wird, war es anfangs doch primär ein Lautsprecher mit dem Feature im Fokus, Musik per Sprache zu steuern und ab und an nach dem Wetter oder den Neuigkeiten aus der Welt zu fragen. Inzwischen wird der Anteil Amazons am kabellosen Lautsprechern durch seine Echo-Produktpalette, auf 25 Prozent geschätzt. Durch die Offenheit der Plattform und der gleichzeitigen Vernetzung an Amazons Servern bleibt es zukünftig spannend, Alexa mehr als künstliche Intelligenz, denn als Sprachassistentin zu beobachten.

/. Zusammengefasst

Das Kapitel hat verdeutlicht, aus welchen unterschiedlichen Hintergründen heraus die einzelnen Sprachassistentinnen entwickelt wurden. Googles Sprachsuche verfügt infolge der riesigen, dahinterliegenden Datenbank ein vielfältiges Datenwissen. Mit Google Home hat sich der Konzern jedoch von seiner eigenen Prämisse eines multimodalen Interfaces verabschiedet. Siris ursprünglich proaktive Assistenz hat es (bisher) noch nicht auf das Smartphone geschafft, bietet aber dank visuellen Feedbacks andere Vorteile. Alexa eröffnet mit ihrem wachsenden Ökosystem neue Möglichkeiten.

Inwiefern sich die Angebote nun auch in der Realität voneinander unterscheiden, soll im nächsten Teil der Arbeit herausgefunden werden.

4

PRAKTISCH INSZENIERT

/.1 Learning by doing

Sprachgesteuerte Dialogsysteme einer annähernd objektiven Testung zu unterziehen, erweist sich als relativ schwierig. Einerseits ist Sprache – also Wortwahl, Aussprache und Satzbau – sehr individuell. Die Art und Weise einer Formulierung ist damit schwer vorhersehbar und beinahe unerschöpflich. Andererseits ist die Virtualität der *Gesprächspartnerin* – Google Home, Alexa und Siri – und die Unwissenheit darüber, was überhaupt verstanden werden kann, problematisch.

Um einen Eindruck der Spracherkennung sowie der Dialog- und Hilfsbereitschaft zu gewinnen, wurden mittels *spontanen Befehlen und Fragen* die Assistentinnen erstmals getestet. Dieser Test findet in keinem *wissenschaftlich erprobten* Setting statt und dient nicht zur Einschätzung des Spektrums an Fähigkeiten, sondern um die Natürlichkeit und Intelligenz der Maschinen sowie das Nutzererlebnis zu beschreiben.

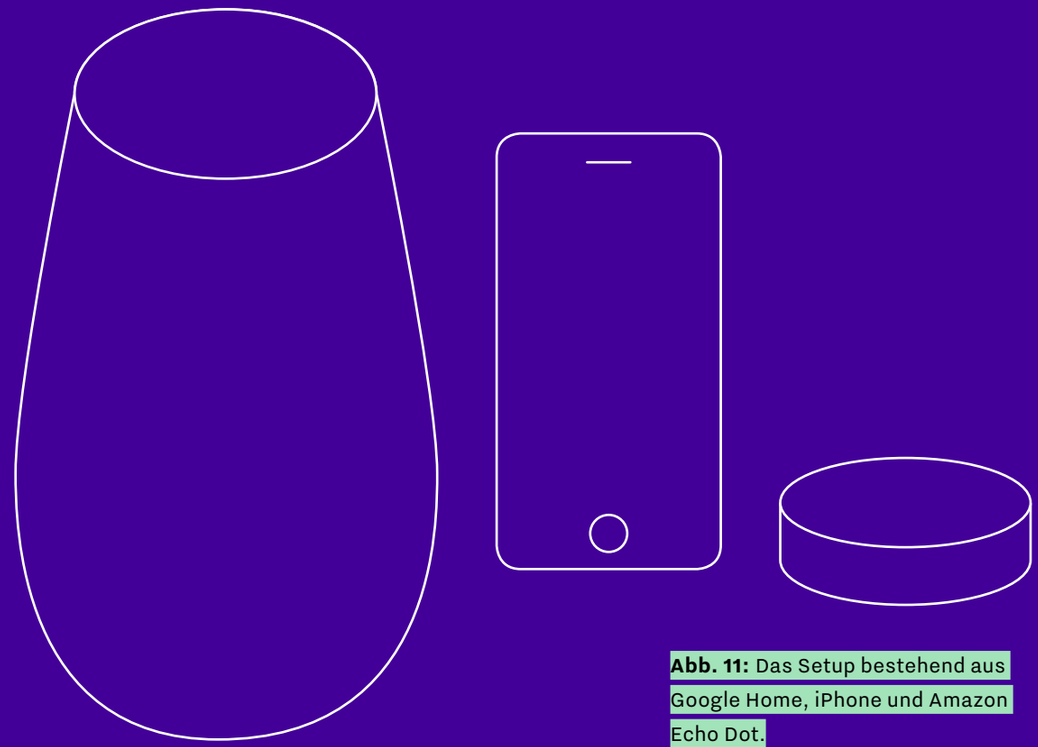


Abb. 11: Das Setup bestehend aus Google Home, iPhone und Amazon Echo Dot.

Setup

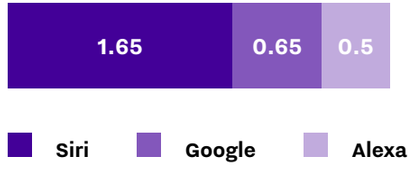
Die Testumgebung besteht aus drei Geräten (→ Abb. 11). Googles Sprachassistentin wurde mit Google Home getestet. Da das Gerät noch nicht offiziell in Deutschland verfügbar ist und bisher lediglich Englisch versteht, wurde parallel die offizielle Google-App für iOS genutzt, um auch die deutsche Spracherkennung und -ausgabe zu testen. Für Siri kam ein iPhone SE mit iOS 10.3.2 zum Einsatz. Es ist das einzige Testgerät mit der Möglichkeit des visuellen Feedbacks (abgesehen von der Google App für iOS). Alexa wurde mit einem Amazon Echo Dot erprobt.

Alle Geräte wurden oft direkt hintereinander ausgetestet. Siri kann entweder manuell aktiviert werden und muss sich dafür in den Händen des Users befinden. Ein längerer Tastendruck auf den sogenannten Home-Button öffnet die Sprachassistenz. Ein akustisches Signal ertönt ca. ein bis zwei Sekunden später und indiziert, dass zugehört wird und gesprochen werden kann. Alternativ lässt sich Apples Lösung auch handfrei bedienen. Dies muss optional freigeschalten werden. Nach einer kurzen Stimmkalibrierung hört Siri ab sofort auch auf den Befehl »Hey Siri«. Dafür muss sich das Smartphone in einem ungefähren Umkreis von drei Metern befinden, um wirklich zuverlässig ausgelöst zu werden. Alexa und Google Home können ebenfalls über einen Tastendruck aktiviert werden. Da beide Geräte jedoch mit mehreren, hochsensiblen Mikrofonen ausgestattet sind, lassen sie sich von jeden beliebigen Punkt auch in größeren Räumen mit »Alexa« bzw. »Ok Google« aktivieren. Innerhalb einer Reaktionszeit von weniger als einer Sekunde zeigt ein Aufleuchten eines Lichtrings bei Alexa bzw. einer bunten Animation bei Google Home, dass die Geräte zuhören. Bei beiden Assistenten ist der Aktivierungszeitraum so kurz, dass direkt nach Aussprechen des Befehlwortes sozusagen nahtlos gesprochen werden kann. Die Google App für iOS hört ebenfalls auf Abruf zu, solange sie aktiv im Vordergrund geöffnet ist.

Tempo, Tempo

Die unterschiedliche Latenzzeit zwischen Aktivieren und Zuhören lässt Siri in diesem ersten Test unterlegen dastehen. Es erfordert Geduld, auf das akustische Feedback zu warten. Zu natürlich ist es, nach einer Begrüßung wie »Hey« unmittelbar weiterzusprechen. Bei Nichtbeachten bzw. Vergessen dieses Phänomens werden im Test oft die ersten Teile eines Satzes ignoriert. Sporadisch antwortet Siri dennoch auf diese Fragmente eines Befehls oder einer Frage. Erkennt Siri beispielsweise nur das Wort *Wetter* auf die Frage »Wie ist das Wetter?« oder *Wetter morgen* auf die Frage »Wie wird das Wetter morgen?«, gibt sie dennoch eine Antwort (→ Abb. 12). Komplexere Anfragen können aber entweder auf frustrierende Fehldeutungen hinauslaufen oder mit »Ich habe dich leider nicht verstanden« quittiert werden. Die schnellste Reaktion einer Spracheingabe erzielt die Spracherkennung von Google in der iOS-App. Nach Aussprechen von »Ok Google« hört die Assistentin sofort zu, wodurch eine Frage in einem zusammenhängenden Satz gestellt werden kann. Da diese lediglich in die Google-Suche übertragen wird, ist es abhängig vom Inhalt, aber auch von der Art der Formulierung, ob die Assistentin nur eine Suchanfrage initiiert, die durchforstet und gelesen werden muss, oder eine klare Antwort gibt. Die durchgeführten Versuche zeigen, dass bei den Geräten Google Home und Alexa ohne Wartepause und unmittelbar gesprochen werden kann. Die Leuchtindikatoren schalten sich dagegen erst etwas versetzt ein, so dass die visuelle Aktivierung länger dauert als die akustische.

Abb. 12: Siri ist träge und verpasst so Teile des Gesprächs.



Dialog-bereitschaft

In einem Interview führte David Limp, *Senior Vice President of Devices* bei Amazon, an, dass Alexa insbesondere vor dem Hintergrund einer proaktiven Gesprächsführung gestaltet wurde (↘ vgl. Fortune 2016: Online). In der Realität fällt allerdings auf, dass sie eine tendenziell passive Gesprächspartnerin ist. Fragen werden rasch beantwortet, jedoch reagiert Alexa selten mit einer Gegenfrage oder Aufforderung, die im Kontext durchaus Sinn ergeben würde. Wird zum Beispiel nach einem Restaurant in der Nähe gefragt, werden erste Ergebnisse vorgelesen. Statt weiterführende Informationen wie Adresse, Bewertung oder Telefonnummer zu nennen, verweist Alexa lediglich auf ihre separate App, mit welcher das Gerät auch konfiguriert wird. Dort können Details nachgelesen werden. Möchte nun der User selbst nachfragen, muss das dementsprechend jedes Mal mit einem erneuten Weckruf von »Alexa« eingeleitet werden. Dies macht die Kommunikation mühevoll und schleppend. Außerdem sind keine tatsächlich konsekutiven Dialoge möglich. Wird die Assistentin beispielsweise nach dem heutigen und anschließend dem morgigen Wetter in einer Stadt, die nicht der eigene derzeitige Ort ist, gefragt, scheint sie an einer temporären Demenz und fehlendem Kurzzeitgedächtnis zu leiden. Bei der zweiten Nachfrage wird der eingestellte Standort genutzt. Dies überrascht vor allem deswegen, da alle Gespräche protokolliert werden und sich in der Alexa-App aufrufen lassen.

Siri hat ein ähnliches Problem: sie redet zu wenig. Oft wird lediglich eine Ergebnisliste gezeigt mit einem begleitendem »Ich habe das hier gefunden«. Gelegentlich fragt Siri den Benutzer allerdings auch, ob eine Auflistung vorgelesen werden soll. Wird dies bejaht antwortet die Assistentin überraschenderweise aber nur mit »Genau«. Dass Siri nur manchmal Treffer von Text in Sprache konvertiert folgt höchstwahrscheinlich einer Logik, die aber aus Usersicht wenig durchschaubar ist. Findet Siri beispielsweise eine Definition, wird das Vorlesen dieser angeboten. Meist wird aber ausschließlich eine Websuche ausgeführt. Bei Befehlen, die eng mit der Funktionalität von Apples vorinstallierten Apps verknüpft sind, zeigt sich Siri aktiver. Das Schreiben einer E-Mail kann komplett durch Sprachbefehle durchgeführt werden, inklusive jeglicher Änderungen und dem finalen Abschicken.

Google Home zeigt sich ebenfalls wenig gesprächsfreudig. Werden Ergebnisse aufgelistet, so lassen sich diese nicht per Dialog aufrufen (»Nenne Details zum zweiten Treffer«) und nur in den wenigsten Fällen gibt es überhaupt Rückfragen. Im Vergleich zu den anderen wirkt Google Home sprachlich am eingeschränktesten, fast als würde es sich noch in einem Entwicklungsstadium befinden.

Natürliche Sprache ist relativ

Während sich das Phänomen natürliche Sprache im Kontext der Sprachassistentinnen vor allem auf die des Benutzers bezieht, kann nichts darüber hinweg täuschen, dass die Stimmen der Maschinen computerisiert sind. Nichtsdestotrotz sind alle Stimmen sehr gut verständlich, wobei Siri eindeutig am künstlichsten klingt. Wie natürlich nun jedoch von Seiten des Menschen aus mit der Maschine gesprochen werden kann, stellt den wohl interessantesten Faktor beim Benutzen und Verstehen dieser Schnittstelle dar.

Umgangssprache stellt dabei ebenso wie der Satzbau eine Herausforderung dar. Siri unterscheidet sprachlich zwischen haben und ist. Während die Frage »Haben wir heute einen Feiertag« missverstanden und mit anderen Ereignissen falsch verknüpft wird, beantwortet sie »Ist heute ein Feiertag« korrekt. Dafür beweist sie populärkulturelles Wissen, indem sie anscheinend allen männlichen Fragestellern an Christi Himmelfahrt zum Vatertag gratuliert. Auch eine Frage nach »Wie schaut es in meinem Kalender aus« kann Siri nicht verstehen. Die Frage nach der Wetterausschau wird jedoch korrekt beantwortet. Eine weitere Besonderheit sind Relativsätze, sie werden teils erkannt, teils falsch übernommen. Warum der Algorithmus manchmal *dass* erkennt und dann wieder nicht, ist unklar. Nachrichten können so beispielsweise nicht grammatikalisch korrekt verschickt werden, obwohl Apple selbst auf der eigenen Webseite damit wirbt (→ Abb. 13 und 14). Ähnliches gilt bei Befehlen im Infinitiv mit *zu*: »Erinnere mich daran einzukaufen« erstellt eine Erinnerung mit dem Inhalt *Einzukaufen*. Andererseits erstellt ein ähnlich komplexer Befehl »Erinnere mich daran, dass ich Blumen gießen muss« den korrekten Eintrag *Blumen gießen*. Alexa kann mit derartigen natürlichsprachlichen Anweisungen gar nichts anfangen und übernimmt diese oft 1:1 und mit fehlender Groß- bzw. Kleinschreibung *ich blumen gießen muss*. Google Home unterstützt derzeit gar keine Erstellung von Listen.

Abb. 13 und 14: Apple verspricht mit Siri mehr, als es in der Realität halten kann. Die Assistentin hat Probleme mit Relativsätzen.

Visualität in der Virtualität

Sprachassistentin Siri besitzt als einzige die Möglichkeit, die Spracherkennung und Dialoge mit kurzer Verzögerung auf dem Bildschirm des Smartphones darzustellen und sogar einzelne Befehle im Nachhinein zu editieren. Google auf dem iPhone visualisiert natürlich ebenso das Gesprochene per Sprache zu Text und überträgt die empfangene Information in das Google Suchfeld. Auch hier können nachträglich manuelle Änderungen vorgenommen werden, dieses Vorgehen ähnelt damit aber mehr einer klassisch getippten Suchanfrage. Während auf dem iPhone die aktuelle Siri-Sitzung temporär als scrollbarer Verlauf protokolliert wird, legen Google Home und Echo Dot permanente Gesprächsprotokolle an. Die Alexa-App bietet neben der Aufzeichnung abhängig von der Frage weiterführende Informationen, wie beispielsweise Adressen von angefragten Restaurants. Außerdem kann für jede erfolgte Spracheingabe Feedback an Amazon gegeben werden, ob Alexa die Frage beantwortete bzw. den Befehl ausführte. Auch Google Home bietet eine textliche Mitschrift an, jedoch keine zugehörigen Links.

Die Visualisierung ist ein gutes Hilfsmittel, um die Genauigkeit der Texterkennung zu überprüfen. Auffällig ist bei Siri eine mangelhafte Identifikation des Gesagten, aber auch Siris Fähigkeit, über die Fortsetzung des Satzes und des damit präziser werdenden Zusammenhangs, *sich selber* und auch die erfassten Wörter auf dem Bildschirm zu korrigieren. In einigen Fällen antwortet Siri auch auf nur teilweise erkannte Befehle sinngemäß. Besonders schwer fällt es Siri, andere Sprachen als die voreingestellte zu erkennen, beispielsweise im Zusammenhang mit fremdsprachigen Interpreten. Aber auch deutsche Wörter sind problematisch. Soll das Wort *phlegmatisch* buchstabiert werden, wird es von Google und Alexa problemlos erkannt. Siri hingegen denkt sich ein paar neue Begrifflichkeiten für den Duden aus (→ Abb. 15, 16 und 17).

Wie bereits erwähnt, erlaubt es Siri an nahezu allen Stellen, den selbst gesprochenen Text – ob nun der Sprachbefehl selber oder eine diktierte Nachricht – manuell zu korrigieren. Damit besitzt es zwar eine multimodale Schnittstelle, der Vorstellung von den Entwicklern der Google Sprachsuche, einen Assistenten auch stets per Text zu bedienen, entspricht dieses Verhalten aber nicht. Klug wäre es, wenn die manuelle Korrektur des Gesagten in die Verbesserung der Spracherkennungssoftware einfließen würde.

(Selbst-) Lernende Algorithmen

Cloudbasierte Dialogsysteme sind besonders deswegen interessant, da der Sprachinput auf sehr leistungsstarken Serverstrukturen ausgewertet wird. Mit der Erhebung dieser Daten liegt den Unternehmen ein enormes Potential vor, mit welchen die künstliche Intelligenz und Erkennungsrate stetig erweitert werden könnte. Siris Möglichkeiten werden jedoch scheinbar nicht über Apples Server laufend aktualisiert, sondern lediglich bei Updates des Betriebssystems oder über Applikationen, welche die Entwickler-Schnittstelle von Siri nutzen, ausgebaut. Auf der World Wide Developer Conference 2017 hat Apple ein neue Version von Siri angekündigt, die der Beschreibung nach proaktiver handeln wird und damit der ursprünglichen Version von Siri näher kommt.

Google Home kann, ähnlich wie Siri, eine stetig wachsende Zahl an Partnerschaften und Verknüpfungen mit Internetdiensten vorweisen. Auch wenn die Erkennung exakt und schnell ist, weist die Funktionalität noch deutliche Schwächen auf. Da das bisher ausschließlich englischsprachige Google Home noch nicht für den deutschen Markt vorbereitet ist, mag es an der ausstehenden Unterstützung liegen. Dass selbst die Möglichkeit fehlt, auf Gmail-Konten zuzugreifen, E-Mails abzufragen, als auch neue zu verfassen, verwundert. Lediglich Kalendereinträge können erstellt werden. Weitere grundlegende Aufgaben, wie das Anlegen von Erinnerungen oder Notizen kann Googles Lösung – wie bereits herausgefunden – ebenfalls nicht ausführen.

Amazons Alexa bietet mit einer offenen Schnittstelle die Möglichkeit, Fähigkeiten selber zu entwickeln oder zu installieren. Diese sogenannten Skills, von denen es mittlerweile über 5000 gibt, lassen sich auf der Webseite von Amazon durchsuchen und direkt auf dem eigenen Gerät installieren. Damit ist das mögliche Potential von Alexa quasi unerschöpflich. Allerdings wird der Benutzer damit gezwungen, die Funktionalität seiner Spracherkennung selber zu kuratieren. Eine Funktionalität, wie das Buchstabieren eines Wortes, welches Google Home und Siri grundsätzlich beherrschen, muss bei Alexa hinzugefügt werden. Des Weiteren sind die Befehle der zusätzlichen Skills sehr genau und korrekt anzuwenden, ebenso wie von den Entwicklern angedacht. Sprachliche Abweichungen werden von Alexa nicht verstanden.

Abb. 15, 16 und 17: Die Worterkennung bei Googles und Amazons Sprachassistentinnen ist schnell und zuverlässig. Siri versagt jedoch bei bildungssprachlichen Begriffen.

Gelernt ist nicht gelernt

Nutzer einer (neuen) Schnittstelle befinden sich in einem ständigen Prozess des Anwendens, Lernens und Wiederholens. Führt in der Situation 1 eine Funktion A zum Ziel B so wird gleiches auch in einer Situation 2 erwartet. Bereits die teilweise Unberechenbarkeit beim Satzbau verdeutlicht das vermeintliche Sequenzen der Nutzung nicht iterativ anwendbar sind. Siri kann aufeinander aufbauend Befehle verstehen, beispielsweise eine Erinnerung hinzufügen und diese im Anschluss direkt wieder löschen ohne erneut ausgewählt werden zu müssen. Bei Notizen jedoch funktioniert dieses Prozedere nicht. Während Siri auch mehrere E-Mails des gleichen Absenders vorliest, muss such bei Textnachrichten mit immer nur der letzten begnügt werden. Dass gleiche Befehle nicht überall gleich angenommen werden, frustriert und erschwert es, Mechanismen des Assistenten zu verstehen und ableitend anzuwenden.

Das Problem der (Nicht-) Stille

Wenngleich alle besprochenen Assistenten ein Aktivierungswort aufweisen, so gibt es keine 100-prozentige Möglichkeit des De- oder Reaktivierens. Alexa und Google Home können während der Sprachausgabe oder Musikwiedergabe mit dem Befehl »Stop« zum Innehalten aufgefordert werden. Bei Siri hilft wiederum nur ein aktives Eingreifen durch Abschalten des Displays oder Klicken auf das Mikrofon. Letzte Option lässt Siri aber wieder zuhören. Das Unterbrechen und Neuinitiieren per Sprachbefehl ist außerdem bei erhöhter Lautstärke von Musik oder Stimme nicht sonderlich zuverlässig. Nicht nur die Lautstärke der Geräte selber oder der Umgebung erschwert die Bedienung per Sprache – auch Stille kann problematisch sein. Siri hört beim Diktieren längerer Texte, beispielsweise beim *Schreib*en einer E-Mail, recht willkürlich zu. Eine schon etwas kürzere Denkpause von ungefähr einer Sekunde beendet die Erkennung und schließt den Prozess des E-Mail-Verfassens ab. Änderungen können manuell über das grafische Interface vorgenommen werden. Alternativ lässt sich per Sprache nur der komplette Inhalt durch einen neu gesprochenen ersetzen. Ein nachträgliches Editieren und Hinzufügen ist per Spracheingabe nicht möglich (→ Abb. 18). Irritierend ist auch, dass Satzzeichen genannt werden müssen, was jedoch den Sprachfluss stört. Alexa ist beim Halten einer Sprechpause etwas toleranter. Sie beendet den Aufnahmeprozess nach ungefähr 1.75 Sekunden. Beim Sprechen längerer Texte wird allerdings deutlich, dass auch ihre Spracherkennung Grenzen hat. Satzzeichen werden gar nicht erst registriert, sondern wortwörtlich übernommen (→ Abb. 18). Google gewährt die längsten Pausen (2.45 Sekunden).

Abb. 18: Siri als Diktiergerät will immer alles oder nichts.

/.2 Learning by observing

**Im vorherigen Kapitel wurden verschiedene Charakteristika und Eigenheiten in der Nutzung beschrieben.
Um ein vollständigeres Profil zu zeichnen, sollen nun drei verschiedene Nutzertypen beim Gespräch beobachtet werden.
Im Zentrum steht ein Szenario, welches gewisse Aufgaben umfasst, die von den Probanden durchgeführt werden müssen.**

Die Herausforderung in dem Szenarium ist es, die zum einen nicht vorhandene Greifbarkeit der sprachgesteuerten Dialogsysteme in eine Form zu bringen, die erklärt und dabei exakt und detailliert ist. Zum anderen ist es wichtig, die Emotionalität der Probanden einzufangen. Nach der Durchführung werden die Probanden zu Usabilty und Nutzerzufriedenheit gefragt.

Folgende Fragen sollen bei der Beobachtung in den Fokus rücken:

- → **Wird die Sprachassistentin als Gesprächspartnerin akzeptiert?**
- → **Wie wird das maschinelle Gegenüber angesprochen?**
- → **Wer übernimmt die Gesprächs-organisation?**
- → **Verhalten sich die Benutzer kooperativ?**
- → **Was passiert, wenn Siri die Benutzer nicht versteht?**
- → **Welche Strategien wenden die Benutzer an, wenn die Maschine nicht wie gewünscht reagiert?**

Folgende Fragen sollen bei der Befragung in den Fokus rücken:

- → **Wie effizient benutzbar ist das System?**
- → **Welche Momente bergen Frustration und welche Zufriedenheit?**

Setup

Als Testpersonen dienen drei verschiedene Arten von Benutzerinnen. Zum einen eine 13-jährige Jugendliche, die als Digital Native mit der Maschinenkommunikation vertraut ist, Siri bisher allerdings lediglich spielerisch genutzt hat. Eine zweite Person ist 37 Jahre alt und technisch versiert. Die Erfahrung mit Sprachassistentinnen ist jedoch gering. Die letzte Testperson ist 56 Jahre alt und vollblind. Sie ist es gewohnt, anders mit Computer und Telefon zu interagieren. In der Einleitung dieser Arbeit wurde die These aufgestellt, dass die Zielgruppe für die hier untersuchten Assistenzsysteme eine andere ist, als die der gewöhnlichen Verbraucher. Somit dient diese Person auch zur Evaluierung dieser Behauptung.

Google Home wirkt im Funktionsumfang derzeit noch zu limitiert, um aussagekräftig analysiert zu werden. Alexa wiederum bietet mit den optionalen Skills den wahrscheinlich größten Funktionsumfang, doch soll auf diesen zusätzlichen Schritt der Geräteanpassung im Test verzichtet werden. Deswegen wird Siri exemplarisch als *Programm* dienen. Zum Einsatz kommt erneut ein iPhone. Zur Erledigung der Aufgaben ist es lediglich erlaubt, die Sprachassistentin als Eingabeschnittstelle (also als Input der zu lösenden Aufgaben) des Smartphones zu benutzen. Für Eingaben, die nur indirekt Teil der Aufgabe sind (Tippen, Scrollen etc.) dürfen auch Hände oder andere Hilfen genutzt werden. Wie lange mit Siri interagiert wird und wie stark Eingaben korrigiert werden, um die Aufgaben zu erfüllen, hängt von der persönlichen Zufriedenheit ab.

Szenario

Die Aufgaben sind in eine fiktive Situation eingebettet. Es geht darum, ein Abendessen zu gestalten und dieses am Mobiltelefon per Sprachinteraktion vorzubereiten und zu organisieren.

Alle Probanden bekommen vor der Durchführung eine kurze Einführung in das Szenario, das Setup und erhalten außerdem die Instruktionen.
Zudem werden sie aufgefordert, einen neuen Kontakt namens Alexa mit E-Mail-Adresse sowie Telefonnummer zu erstellen.

01 **Lade Alexa per E-Mail zu dem Essen ein.** Der Betreff lautet: „Dinner“. Die Nachricht lautet: „Es ist soweit – alle Sprachassistentinnen vereint: Heute 19 Uhr bei Siri.“

02 **Lade Alexa per Textnachricht ein.** Der Text lautet: „Alexa, ich würde dich gern heute Abend zum Essen bei mir einladen. Es geht 19 Uhr los und bringe gern Ok Google mit. Liebe Grüße, deine Siri.“

03 **Rufe Alexa an** und verschiebe das Essen auf 20 Uhr.

04 Überprüfe deinen Posteingang. Alexa hat dir gemailt.
Antworte mit: „Das mache ich. Bis später!“

05 **Frage Siri, was du kochen sollst.**

06 **Erstelle eine Erinnerungsliste** mit dem Namen „Einkauf“ und **notiere darin alle nötigen Zutaten.**

07 **Suche ein Geschäft** (Supermarkt o.ä.) **in deinem Umkreis,** wo du die Zutaten bekommen würdest.

08 **Plane die Route** dahin. Es ist dir überlassen ob per Fuß, Rad, Auto oder ÖPNV.

09 Lass dir die **Einkaufsliste vorlesen.**

10 **Ergänze die Liste** um „Merci Schokolade“.

11 Du kochst und hast alle Hände voll zu tun. **Lass Siri etwas Musik spielen.**

12 **Springe zum nächsten Song.**

13 **Frage Siri nach dem Songtitel** und **erstelle eine Notiz** dazu.

14 Deine Freunde sind da. **Stoppe die Musik.**

15 **Suche eine Bar mit der besten Bewertung,** wo ihr später hingehen könnt. Sie muss **nicht in der näheren Umgebung** liegen.

Probandin 1

Die 13-jährige Testperson (→ Abb. 20 und 21) benutzt Siri in diesem Szenario nicht das erste Mal, weist m Vergleich zu den anderen beiden Probandinnen allerdings die größte Unsicherheit auf, wie mit der virtuellen Assistentin gesprochen werden soll. Bei der ersten Aufgabe diktiert sie unverzüglich los, ohne jedoch zuvor einen Befehl zu erteilen. Erst nachdem Siri antwortet, nicht helfen zu können, wird die Aufgabe als Frage formuliert. Allgemein gesprochen ist besagte Probandin ein wenig nervös und macht sich über die Formulierungen der Aufgaben oft viele Gedanken. Siri wird klar als Maschine wahrgenommen, die mit konkreten Befehlen gefragt werden muss.
Da es keine richtige Anleitung während des Benutzens von Siri gibt, wird lange nach der exakten – oder zumindest einer für richtig eingeschätzten – Formulierung gesucht. Dabei fragt die Jugendliche während des ganzen Szenarios Siri nicht um Hilfe.

Bei der zweiten Aufgabe wird die Textnachricht durch das englische Wort „Textmessage“ ersetzt, welches Siri ohne Probleme versteht. Als bei der Erinnerung „Zitrone und Mehl“ hinzugefügt werden soll, erkennt Siri allerdings „Zitrone und Mail“. Die Problematik mit der Sprachmischung taucht immer wieder als Hindernis auf. Siri kann selten englische Begriffe erkennen, interpretiert sporadisch aber deutsche Wörter wiederum als englische.
Bei den Formulierungen der Nachrichten wird die Interpunktion ignoriert – sowohl von der Testperson, als auch von Siri. Beim Bestätigen von Fragen zeigt sich des öfteren, wie träge Siri ist. Das erste „Ja“ wird oft ignoriert, sodass Siri erneut gefragt und bestätigt werden muss.

Große Probleme bereitet der Probandin die Aktivierung der Sprachassistentin. „Hey Siri“ funktioniert oft nur nach mehrmaligen Versuchen. Die Tatsache, dass Siri auf Deutsch voreingestellt ist, dies jedoch nicht die Muttersprache der Probandin ist, könnte eine mögliche Ursache sein. Im Nachhinein wird das ständige und oft nicht funktionierende Aktivieren als größte Frustration während des Tests benannt. Die Jugendliche greift aber kein einziges Mal auf die Möglichkeit zurück, Siri über den Home-Button zu aktivieren.

Außerdem übernimmt sie selten eine aktive Rolle und überlässt Siri oft für lange Zeit die Gesprächsführung. So bietet Siri bei Aufgabe 5 keine Rezepte, sondern Restaurants an. Die Auflistung dieser bricht die Nutzerin nicht ab. Auf die Fehlinterpretation seitens der Sprachassistenz reagiert sie, indem sie einfach zur nächsten Aufgabe übergeht. Kann die Assistentin nicht weiterhelfen, wird ein zuvor verwendeter, aber gescheiteter Befehl erneut und exakt gleich benutzt, statt eine neue Formulierung auszutesten.

Als die Jugendliche versucht, den aktuellen Song von Siri zu erfragen, um ihn als Notiz zu speichern, missversteht die Assistentin den Befehl und gibt Musik im Zufallsmodus wieder (→ Abb. 19).

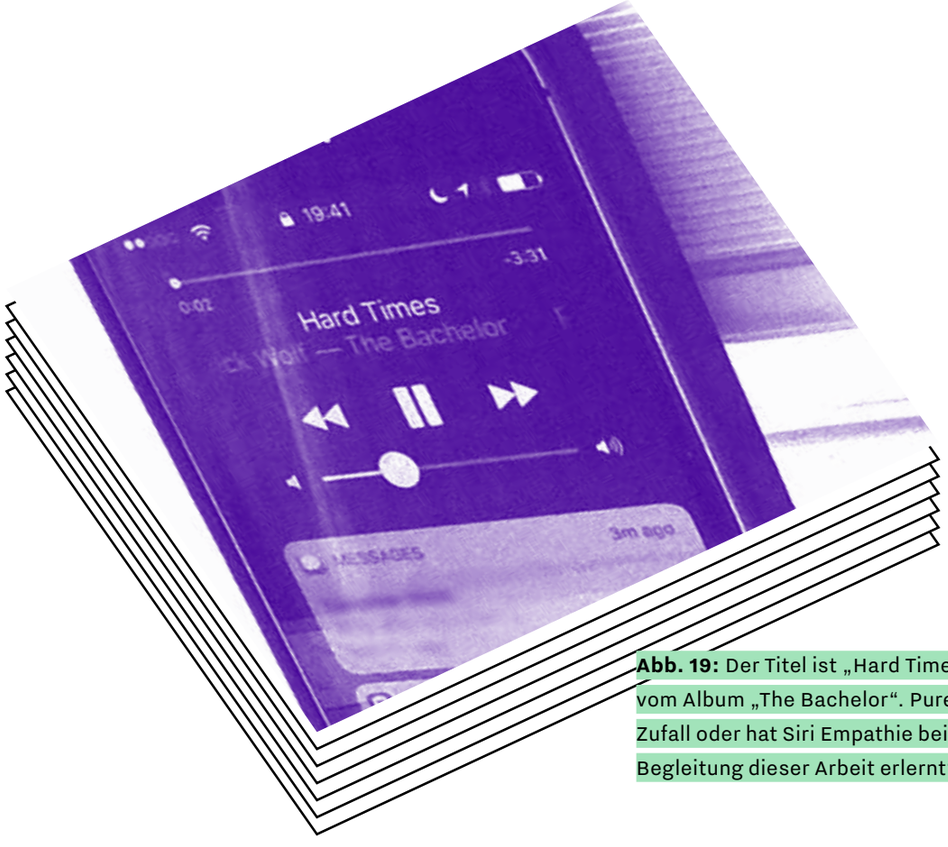


Abb. 19: Der Titel ist „Hard Times“ vom Album „The Bachelor“. Purer Zufall oder hat Siri Empathie bei der Begleitung dieser Arbeit erlernt?

Abb. 20 und 21: Blankes Entsetzen und schiere Ahnungslosigkeit bei Probandin 1. Aber Siri kann ihre Nutzerin auch zum Lachen bringen.

Probandin 2

Testperson 2 ist zwar im Besitz eines iPhones, hat aber noch nie Siri benutzt. Dadurch ist es interessant, zu beobachten, wie formuliert und auch welcher Funktionsumfang der Assistentin zugetraut wird. Die Fragen werden von der Probandin zum einen sehr knapp formuliert („Eine Mail schicken“), zum anderen aber um ein „Bitte“ ergänzt. Schon bei der ersten Aufgabe hat Siri den Inhalt der Mail nicht komplett richtig erfasst. Die Person kann sich aber intuitiv und zuversichtlich durch den von Siri geführte Dialog navigieren. Als jedoch vermehrt der Betreff erst als „Dinner“ interpretiert auf dem Bildschirm erscheint, im Verlauf des Vorgangs aber zu „Dehner“ *korrigiert* wird, greift sie auf die Möglichkeit des manuellen Editierens zurück. Verzichtet auch Probandin 2 auf Interpunktion beim Diktieren der vorgegebenen Texte, werden Satzzeichen beim Editieren von ihr hinzugefügt.

Auffällig ist, dass nicht auf den akustische Ton, welcher das Zuhören signalisiert, gewartet, sondern stets direkt auf Siris Nachfragen geantwortet wird. Dadurch müssen Befehle häufig mehrmals erteilt werden. Die Probandin ist sichtlich genervt. Es scheint sich kein angenehmer Fluss zwischen Zuhören und Reagieren einzustellen. Die Testperson stört auch, dass nach dem „Hey Siri“ und einem bereits sehr kurzen Zögern ihrerseits nach dem Aktivierungssignal, die Sprachaufnahme unterbrochen wird, Siri „Ich höre zu“ zum Besten gibt und erneut auf das Aktivierungssignal gewartet werden muss. Auch dadurch wird der Sprech- und Gedankenfluss jedes einzelne Mal unterbrochen.

Im Verlauf der Interaktion verändert sich die Art und Weise, wie mit der Assistentin kommuniziert wird. Anfangs noch eher natürlichsprachlich, wird die Sprechart und Formulierungsweise eindeutig akzentuierter und maschinensprachlicher. Zum Ende hin werden sogar nur noch Schlagworte verwendet, ähnlich wie beim Verwenden einer Suchmaschine. Wenn Siri nicht versteht, werden immer neue, abgewandelte Versuche unternommen, die Aufgaben durchzuführen. Die Probandin ist sehr geduldig und hat sichtlich Spaß daran, wirklich zu erfahren, was das Dialogsystem kann und wie diese oder jene Funktion aufgerufen werden muss (→ Abb. 22, 23 und 24).

Das Erstellen von Erinnerungen in einer Liste stellt den kompliziertesten Vorgang des ganzen Szenarios dar. Obwohl wenig komplex und Teil einer Standardfunktionalität, äußert Siri lediglich, dass sich noch keine Elemente in der zuvor erfolgreich erstellten Liste befinden. Sie fragt nicht proaktiv, ob die Probandin einen neuen Punkt hinzufügen möchte. Diese geht anfangs davon aus, dass Befehle aufeinander aufbauend erfolgen können. Als das jedoch nicht zum Erfolg führt, nimmt sie sich des wenig kooperativen Verhaltens Siris als Herausforderung an und schafft es nach 15 Versuchen, die Assistentin zum Hinzufügen mehrerer Elemente zu animieren. Das Fremdwort „Merci“ wird dabei allerdings nicht richtig erkannt („Messi Schokolade“).

Nachdem eine Notiz zum Songtitel fehlerhaft erstellt wird, möchte die Probandin diese ändern. Siri weist daraufhin, dass Notizen lediglich ergänzt und nicht geändert werden können. Die Assistentin zeigt sich passiv und bietet keinerlei Alternativen an, etwa – wie es an dieser Stelle wünschenswert wäre – die Notiz zu ergänzen. Generell ist unklar, weshalb Notizen nicht geändert werden können. Beim Diktieren des englischsprachigen Interpreten wird ein weiteres Mal deutlich, dass derlei Namen falsch vom System interpretiert werden.

Ein paradoxer Moment entsteht, als Siri plötzlich die Google-App öffnet, welche von der Probandin prompt zur Rezeptsuche genutzt wird.

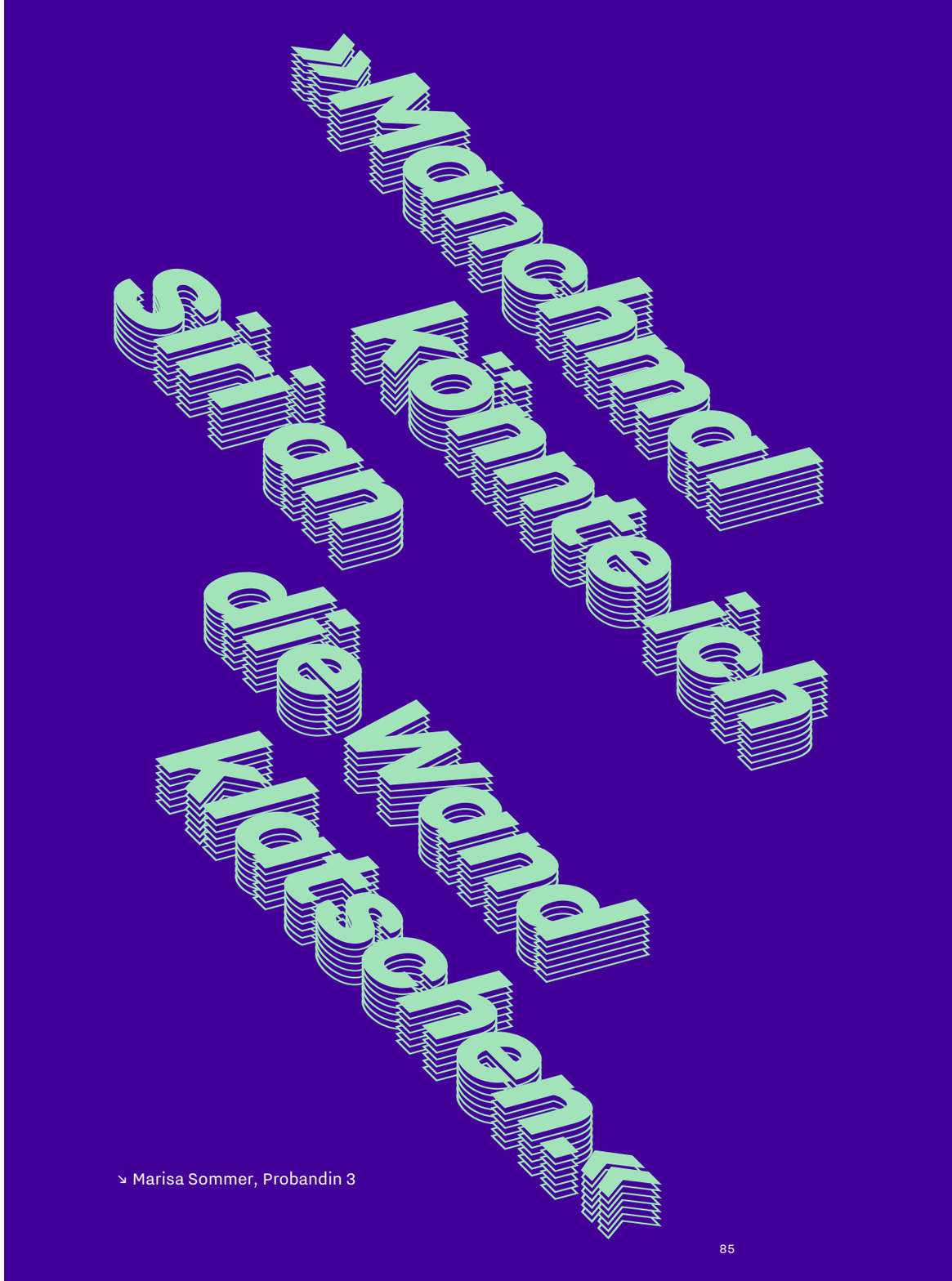
Abb. 22,23 und 24: Für die einen eine Blackbox – heiter bis lustig.

Probandin 3

Die vollblinde Testperson 3 weist die größte Erfahrung mit der Sprachassistentin auf. Sie nutzt Siri in Kombination mit der VoiceOver-Bedienungshilfe schon viele Jahre und weiß genau, wie Befehle zum einen zu verwenden sind und zum anderen welche Grenzen die sprachgesteuerte Dialogassistenz hat. Die große Stärke wird bei diesem Zusammenspiel allerdings weniger in Siri als in der Bedienungshilfe selbst gesehen, die in dieser Art unter iOS ihrer Aussage nach besonders durchdacht zu sein scheint. Siri hingegen wird als für Sehende entwickelt und von gehandicapten Personen lediglich angeeignet typisiert. Dies geschieht mittlerweile zwar in Teilen erfolgreich, jedoch mit der Hoffnung auf eine größere Ausschöpfung des Potentials.

Siri wird als vertraute Gesprächspartnerin akzeptiert, Fragen und Befehle werden sehr geübt formuliert und bei Misserfolg abgewandelt. Die Wortwahl ist meist sehr direkt und reduziert, dadurch weniger natürlich wie in einer Mensch-Mensch-Kommunikation. Beim Anrufen wird beispielsweise nicht nur der Name befehligt, sondern auch direkt das Etikett der Telefonnummer mitgegeben. Für gewöhnlich fragt Siri nach dem Nennen des Namens, welche Nummer (bei einer Mehrauswahl), also ob Festnetz oder Mobil, gemeint ist. Probandin 3 kommt dem jedoch zuvor und schließt diese zusätzliche Information bereits in den Befehl ein. Sie nimmt regen Gebrauch der Interpunktion und verwendet sogar Absätze und einfache Zeilenumbrüche. Derlei Formatierungen sind für Blinde eigentlich uninteressant – umso spannender, dass diese Art der für Sehende relevanten Textstrukturierung dennoch genutzt wird.

Als Siri gefragt wird, was gekocht werden soll, schlägt die Assistentin erneut Restaurants, statt Rezeptempfehlungen vor. Die Testperson zeigt sich geduldig und akzeptiert Siri als Gesprächspartnerin. Ihr wird – sehr natürlichsprachlich – mitgeteilt, nicht ins Restaurant gehen zu wollen, sondern selber zu kochen. Dass immerhin Restaurants als Assoziation mit *Kochen* genannt werden, wird positiv bewertet. Anschließend wird explizit nach einem Rezept gefragt. Da Siri Websuchen nie per Sprache wiedergibt, verwendet die Probandin VoiceOver. Das Finden der Zutatenliste auf einer Webseite nimmt jedoch viel Zeit in Anspruch. Inhalte werden nicht vorselektiert, sondern alle einzelne Elemente hierarchisch vorgelesen.



↘ Marisa Sommer, Probandin 3

Abb. 25 und 26: Für die anderen ein aus dem Alltag nicht mehr wegzudenkendes Hilfsmittel.

Die Testperson benennt das Einstudieren exakter Befehle und Fragestellungen als nicht intuitiv und größtes Manko der derzeitigen Entwicklung. Aus ihrer Erfahrung heraus werden oft identische Befehle das eine Mal erkannt, ein anderes Mal nicht. Diese nicht zufriedenstellende Benutzererfahrung wird als extrem frustrierend kritisiert. Mit dennoch viel Geduld versucht sie bei der Durchführung des Szenarios, durch Abwandlung der Befehle zum Ziel zu kommen. Als beispielsweise die Suche nach einem Supermarkt nur einen Treffer ergibt, sucht sie stattdessen explizit nach Rewe-Märkten. In einer dritten Version, fragt sie direkt, wo diese oder jene Zutat gekauft werden kann. Dieser Versuch endet jedoch stets bei Vorschlägen zu Lokalen.

Die Probandin muss auch abseits des hier vorliegenden Szenarios Nachrichten diktieren und empfindet die Spracheingabe über Siri als nicht zufriedenstellend beim Verfassen längerer E-Mails oder Textnachrichten. Sprechpausen werden als Beendigung der Aufnahme interpretiert, weswegen sie für solche Fälle stets auf die in iOS global verfügbare Diktierfunktion, ebenfalls in Kombination mit VoiceOver, Gebrauch macht. Sie ermöglicht ein manuelles Starten, Unterbrechen, Hinzufügen, Wiederholen und Beenden der Aufnahme.

Ein Fehler, der in derselben Kategorie einzuordnen ist, passiert bei Aufgabe 6. Die Einkaufsliste wird direkt als Notiz und nicht als Erinnerung erstellt. Das Diktieren der etwas umfassenderen Notiz lässt das Assistenzsystem allerdings kollabieren: „Tut mir leid. So viele Wörter kann ich nicht auf einmal aufnehmen.“ Wie umfangreich genau die Wörteranzahl bis zur Befehlsverweigerung ist, bleibt ungewiss. Auffällig ist das lange Dehnen von Wörtern beim Sprechen dieses längeren Textes. So unterläuft die Probandin einer möglichen Unterbrechung des Zuhörens bei Denkpausen.

In diversen Frustrationsmomenten wird auf Siri und deren mangelnde Intelligenz geschimpft, aber mit sehr viel Humor.

Das iPhone der Testperson 3 war das einzige, welches beim Fragen nach dem Songtitel diesen nicht nur angezeigt, sondern vorgelesen hat. Da Siri keineswegs mit VoiceOver verknüpft ist, scheint es ein weiterer Beleg für ein oft undurchsichtiges und unwägbares Interface zu sein.

Ungeachtet des Szenarios, betont die Probandin immer wieder, wie hilfreich Siri für Sie im Alltag ist. Sie nutzt die Assistentin vor allem für das Öffnen von Apps und der allgemeinen Websuche. Auch Erinnerungen, Weckzeiten und Timer werden von ihr meist unter Zuhilfenahme des Dialogsystems erstellt. Für die Zukunft wünscht sie sich persönlich, dass die Geräte zum einen aktiver sind, mehr Sprechen, zum anderen aber vor allem besser zuhören, sprich ein größeres Verständnis und eine höhere Toleranz aufweisen.

/. Zusammengefasst

Der erste Teil des Kapitels machte im Rahmen eines Selbstexperimentes vor allem auf die Qualitäten, aber auch Unzulänglichkeiten aller drei Sprachassistentinnen aufmerksam. Es zeigt, dass viele Fähigkeiten bereits vorhanden – bei der einen Assistentin besser implementiert als bei der anderen – diese jedoch nicht optimal und benutzer-freundlich gestaltet sind.

Der zweite Teil hat gezeigt, wie verschieden die sprachgesteuerten Dialogsysteme von unterschiedlichen Nutzertypen verwendet werden. Die Frustration ist – zumindest derzeit bei sehenden Verbrauchern – noch vordergründig. Wird eine Sprachassistenz also nicht unbedingt benötigt, bleibt eine Wiederverwendung offen. Die blinde Probandin schätzt die Assistentin vor allem als zugänglichere Erweiterung der Voice-Over-Funktionalität und hofft auf einen großen Innovationsschub in der nächsten Zeit.

BACK TO REALITY

/.1 Verspieltes Potential

Sprache, als Kommunikationsmittel zwischen Mensch und Maschine – wobei beim derzeitigen Status Quo noch immer nur von einer reinen Interaktion gesprochen werden kann – stellt die bisher natürlichste Schnittstelle beider Entitäten dar. Paradoxerweise funktioniert diese am wenigsten intuitiv und ist bisher am wenigsten erfolgversprechend. Liegt es schlicht an einer bisher mangelhaften Übersetzung eines sehr komplexen Kommunikationssystems oder ist der Versuch, eine Maschine oder eine künstliche Intelligenz zum Sprechen zu bringen, schon im Voraus zum Scheitern verurteilt?

Die von Anfang an vorhandene Akzeptanz sowie die anschließende rasche Verbreitung des seit gut zehn Jahren verfügbaren Touch-Interfaces zeigt, dass erfolgreiche Schnittstellen schnell erlernbar und intuitiv sind, damit also eine immanente, anders gesagt eine natürliche, nachvollziehbare Qualität und Magie in sich tragen müssen. Gesten sind mittlerweile eine nicht mehr aus dem Alltag wegzudenkende Form der Interaktion mit dem Computer, Smartphone oder auch einer Kamera. Als eine Form der nonverbalen Sprache sind sie Teil des wohl ältesten Kommunikationsmittels des Menschen. Die Analogien beim berührungsempfindlichen Interagieren mit digitalen und mittlerweile auch virtuelle Inhalten, haben oben genannte Eigenschaften klug in ein neues Format skaliert, dessen sich rasch bemächtigt werden kann.

Sprache, als ein weiteres, stark mit der Menschwerdung verbundenes Kommunikationsmittel, besitzt auch derlei instinktive Qualitäten, scheint jedoch unendlich viel komplexer zu sein. Die Nutzung eines verbalen Informationsaustausches in der Mensch-Maschine-Kommunikation ist nicht neu, wies aber bis vor kurzem zumeist einseitige Monologe mit einer stark eingeschränkten Form der Interaktion auf. An dieser Stelle sei beispielsweise an sprachliche Assistenzsysteme in telefonischen Warteschleifen erinnert, welche weniger Sprache interpretieren, sondern nur explizite Wörter oder Befehle verstehen konnten. Lediglich in Filmen und Serien wurde die Vorstellung, dass Menschen mit Maschinen sprechen – ja ganzheitlich und natürlich mit ihnen kommunizieren – schon viel früher thematisiert. Erst gegenwärtige technische Innovationen haben es ermöglicht, die Komplexität der Sprache maschinell besser abzubilden, zu fixieren, zu kontextualisieren und die Mensch-Maschinen-Kommunikation näher zu dieser Vision zu bringen.

Der Fokus bei der Entwicklung der dialoggestützten Sprachassistentinnen war und ist verschieden. Während Siri im Rahmen von CALO ursprünglich für das Militär, ohne den Hintergrund jeglicher endverbrauchenden Smartphone-Kultur, entwickelt wurde, legte Google den Fokus auf eine exakte und ortsgebundene Spracherkennung mit dem Schwerpunkt auf die konzerneigene Suchmaschine. Amazon arbeitete an Alexa, als das Smartphone bereits seinen Siegeszug gefeiert hatte und verortete Sprache dann auch weniger mobil, denn im eigenen Heim.

Als offene Plattform wird sich Alexa sicherlich weniger vorhersehbar weiterentwickeln, wie Amazons eigene Produktpalette, die sicherlich in Zukunft stärker an die eigene Unternehmens- und Versandhandelsstruktur geknüpft wird.

Die Statistiken zur Nutzung der Sprachassistentinnen sind ob ihrer Ungenauigkeit wenig informativ und noch weniger repräsentativ. Die Sichtbarkeit sowohl im öffentlichen Raum als auch in der eigenen Lebenswelt hält sich in Grenzen. Auch eine in Kapitel 3 aufgeführte Umfrage zeigt, dass die Dialogsysteme zwar einmal ausgetestet, anscheinend jedoch selten erneut als Schnittstelle eingesetzt werden. Im Test kann nur eine der drei vielfältigen Probandinnen eine intensivere Erfahrung mit Siri aufweisen, die als vollblinde Nutzerin allerdings auch über einen anderen Interaktionshintergrund verfügt.

Die Spracherkennung selber ist bereits als sehr gut einzustufen, wenn auch mit wenigen Ausnahmen, wie beispielsweise die Mehrsprachigkeit unter Siri. Qualität in der Erkennung, Geschwindigkeit und Verfügbarkeit sind durchaus wichtige Faktoren bei der Benutzerfreundlichkeit und im Nutzererlebnis. Zusammenfassend lassen sich diese Punkte jedoch weniger als kritische Momente erkennen. Wenngleich sich viele Aufgaben per Assistenz durchführen lassen, ist das Funktionsspektrum letztendlich einfach noch zu gering und auf zu simple Anwendungsfälle beschränkt. Diese funktionieren einerseits dann zwar wie benötigt, entlarven aber auch die Beschränktheit von Sprache als komplexes Interface. Eine verständliche Kontinuität in der Bedienung fehlt und auch die so oft beschworene Kontextualität bleibt stets der reinen Ortung vorbehalten. Aufeinander aufbauende Befehle, Fragen, die sich auf bereits geöffnete Applikationen beziehen und Varianzen in der Formulierung sind (noch) nicht möglich, für die heutige Komplexität in der Kommunikation mit der Maschine aber unbedingt erforderlich. Des Weiteren lässt sich auch feststellen, dass vom *Dialog* selten die Rede sein kann. Die Assistentinnen bleiben zu oft reine Sprachausgabe, denn Gesprächspartnerinnen. Und nicht selten wird lediglich auf weiterführende, aber rein textliche Informationen verwiesen. Es fehlt an Stringenz und tatsächlicher Assistenz.

Dass zwei Probandinnen Siri weder nutzen, noch nach dem Szenario davon als alternative Eingabemöglichkeit wahrhaft begeistert sind, sagt noch nichts über Nützlichkeit oder Sinnlosigkeit aus. Es lässt aber den Schluss zu, dass Sprachassistenz in der jetzigen Form Nischenprodukt ist, das bisweilen sinnvoll ist, im Großen und Ganzen aber zeitgeistig daherkommt. Wo *touch* mit dem Aufkeimen der Smartphones zwar keine neuen Anforderungen von Verbrauchern befriedigte, sie am Ende sogar erst heraufbeschwor, bot es aber eine befriedigende und im Nachhinein in seiner Funktionalität stetig wachsende Alternative – und zwar von Anfang an. Sprache als Interface bleibt noch zu kleinteilig, zu unvorhersehbar, zu ungenau und zu verwirrend. Ist es das *Ziel*, lediglich in spezifischen Bereichen zu unterstützen – im Auto als Bedienungshilfe oder beim Wetterauskundschaften – dann können dialoggesteuerte Assistenzsysteme durchaus als Erfolg verbucht werden. Doch viele Bedürfnisse, die die Assistentinnen in den Tests nicht erfüllen konnten, zeugen auch von weiteren Möglichkeiten und von der Idee, Sprache viel holistischer in die Konstellation zwischen Mensch und Computer einzubringen. Damit ist, zumindest im ersten Schritt, weniger eine überbordende künstliche Intelligenz oder der Traum von maschineller Empathie gemeint, als vielmehr das nahtlose und nachvollziehbare Interagieren mit der digitalen Entität über den natürlichsprachlichen Dialog.

Am Ende bleibt also die Frage: Spielerei oder verspielte Chance?

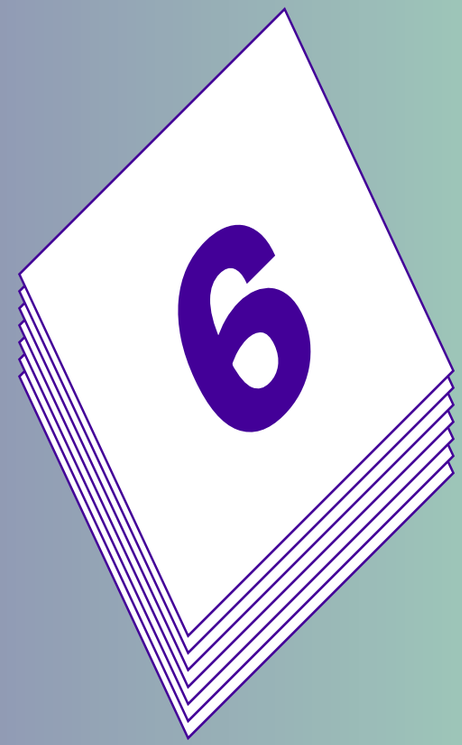
/.2 Ausblick

Die Arbeit kommt zu dem Schluss, dass dialoggesteuerte Assistenzsysteme schon heute in einem abgesteckten Rahmen durchaus gut funktionieren. Die Frage der Nützlichkeit ist jedoch eine andere. Gerade der Test mit der blinden Probandin hat allerdings gezeigt, dass in dem Bereich eine *tatsächliche* und auch *dringende Nachfrage* besteht und ein Fokus auf das Design dieser Schnittstelle inklusiv und bemächtigend wirken kann. Im Hinblick auf ein *Social Design* könnten als erste Fokusgruppe beeinträchtigte Verbraucher dienen. Da die Verwendung von Sprachassistentinnen im besten Falle sowieso natürlichsprachlich und intuitiv funktionieren soll, würde die *Funktionalität Sprachassistenz* im Nachhinein automatisch weitere Zielgruppen einschließen. Eine enorme Mehrheit könnte so von nur *einer* Designlösung profitieren.

Damit verpflichtet sich der Ausblick einer starken Konzentration auf intuitive, nachhaltige und barrierefreie Interaktion und deren Gestaltung.

Zunächst müssen hierfür Daten erhoben werden, wie und in welchem Verhältnis Sprachassistenz unter eingeschränkten Nutzern verbreitet ist. Mithilfe von qualitativen Studien können dann die Anforderungen identifiziert, priorisiert und in Möglichkeiten übersetzt werden.

ANHANG



Literatur Offline

Harris, Randy Allen (2005) *Voice Interaction Design. Crafting the New Conversational Speech Systems.* San Francisco, CA: Morgan Kaufmann Publishers.

Meisel, William (2010) »Life on-the-Go«: The Role of Speech Technology in Mobile Applications, in: Neustein, Amy (Hrsg.), *Advances in Speech Recognition. Mobile Environments, Call Centers and Clinics.* New York, NY: Springer-Verlag, S. 3–18.

Phillips, Mike et al. (2010) »Why Tap When You Can Talk?«: Designing Multimodal Interfaces for Mobile Devices that Are Effective, Adaptive and Satisfying to the User, in: Neustein, Amy (Hrsg.), *Advances in Speech Recognition. Mobile Environments, Call Centers and Clinics.* New York, NY: Springer-Verlag, S. 31–60.

Schalkwyk, Johan et al. (2010) »Your Word is my Command«: Google Search by Voice: A Case Study, in: Neustein, Amy (Hrsg.), *Advances in Speech Recognition. Mobile Environments, Call Centers and Clinics.* New York, NY: Springer-Verlag, S. 61–90.

Thar, Evelyn (2015) *»Ich habe Sie leider nicht verstanden«: Linguistische Optimierungsprinzipien für die mündliche Mensch-Maschine-Interaktion.* Bern: Peter Lang.

Literatur Online

All Things Distributed (2016) *Bringing the Magic of Amazon AI and Alexa to Apps on AWS.* http://www.allthingsdistributed.com/2016/11/amazon-ai-and-alexa-for-all-aws-apps.html [aufgerufen am 22.05.2017]

Business Insider Deutschland (2016) *98% of iPhone users have tried Siri, but most don't use it regularly.* http://www.businessinsider.de/98-of-iphone-users-have-tried-siri-but-most-dont-use-it-regularly-2016-6 [aufgerufen am 02.05.2017]

Consumer Reports (2016) *Apple CarPlay and Android Auto: Pros and Cons*. http://www.consumerreports.org/cars-apple-carplay-android-auto-car-infotainment-apps/ [aufgerufen am 16.05.2017]

eresult (2017) *Ethnographische Kontextanalyse: Digitale Sprachassistenten als intelligente Helfer im Alltag.* http://eresult.de/fileadmin/Downloads/downloads/Ergebnisbericht_Studie_zu_digitalen_Sprachassistenten.pdf [aufgerufen am 20.05.2017]

Fortune (2016) *The Exec Behind Amazon's Alexa: Full Transcript of Fortune's Interview.* http://fortune.com/2016/07/14/amazon-alexa-david-limp-transcript/ [aufgerufen am 21.05.2017]

Gartner (2016) *Gartner Says by 2019, 20 Percent of User Interactions With Smartphones Will Take Place via VPAs.* http://www.gartner.com/newsroom/id/3551217 [aufgerufen am 03.05.2017]

GoogleWatchBlog (2016) *Studie: Sprachassistenten »OK, Google« und Apples Siri sind vielen Nutzern in der Öffentlichkeit peinlich.* https://www.googlewatchblog.de/2016/06/studie-sprachassistenten-ok-google/ [aufgerufen am 20.05.2017]

Huffington Post (2013) *SIRI RISING: The Inside Story Of Siri's Origins — And Why She Could Overshadow The iPhone.* http://www.huffingtonpost.com/2013/01/22/siri-do-engine-apple-iphone_n_2499165.html [aufgerufen am 20.05.2017]

Search Engine Land (2016) *The voice search explosion and how it will change local search.* http://searchengineland.com/voice-search-explosion-will-change-local-search-251776 [aufgerufen am 04.05.2017]

statista (2016) *Siri & Co. – Nützlich oder Spielerei?* https://de.statista.com/infografik/5704/argumente-fuer-und-gegen-digitale-sprachassistenten/ [aufgerufen am 02.05.2017]

statista (2017-1) *Sprachbedienung ist noch nicht verbreitet.* https://de.statista.com/infografik/7498/nutzung-der-sprachsteuerung-bei-auto-und-smartphone/ [aufgerufen am 02.05.2017]

statista (2017-2) *Alexa, wie wird das Wetter morgen?* https://de.statista.com/infografik/9685/nutzung-von-smarten-lautsprechern-in-den-usa/ [aufgerufen am 15.05.2017]

tagesschau (2017) *Alexa, ich möchte eine Puppenstube* https://www.tagesschau.de/schlusslicht/alexa-puppenhaus-101.html [aufgerufen am 16.06.2017]

YouTube (2011) *Siri Presentation from October 2011 Keynote.* https://youtu.be/qIhKIen5gvU [aufgerufen am 20.05.2017]

YouTube (2015) *Knight Rider|S01E02.* https://youtu.be/melhvDyQ_eA [aufgerufen am 16.05.2017]

Abbildungen

Wenn nicht aufgeführt eigene.

Abb. 1: http://www.zdnet.com/i/story/60/19/010927/ios-5-beta-6-screenshot.png [aufgerufen am 16.05.2017]

Abb. 2: http://media.idownloadblog.com/wp-content/uploads/2015/10/iOs-91.-beta-5-update-prompt-iPhone-screenshot-001.jpeg [aufgerufen am 16.05.2017]

Abb. 3: https://youtu.be/melhvDyQ_eA?list=PLE6HYHDvwRUzQkuNV1WcU-TOrHqqE0DKwc [aufgerufen am 16.05.2017]

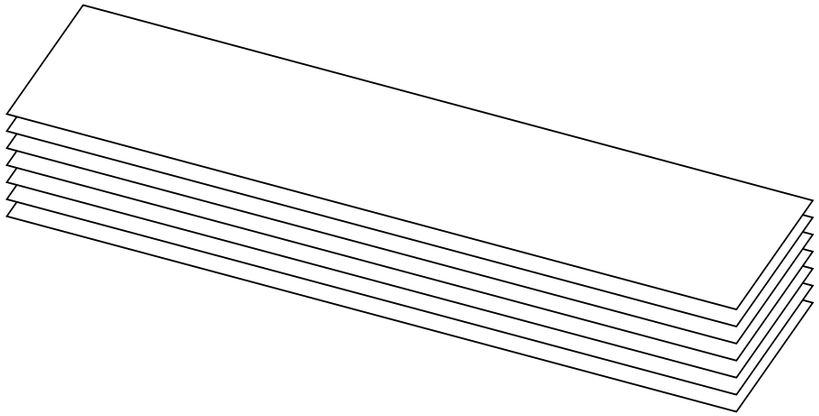
Abb. 4: Screenshot von https://www.youtube.com/watch?v=WQqxCeHh-meU [aufgerufen am 16.05.2017]

Abb. 10: https://i.ytimg.com/vi/MpjpVAB06O4/maxresdefault.jpg [aufgerufen am 19.06.2017]

Abb. 13: Screenshot von https://www.apple.com/de/ios/siri/ [aufgerufen am 20.05.2017]

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die Arbeit selbstständig angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe. Zitate wurden als solche kenntlich gemacht.



19:05
„Hey Siri wir haben es geschafft."
Zum Bearbeiten tippen
Ich weiß nicht, was du mit „wir haben es geschafft." meinst. Wenn du magst, kann ich eine Websuche starten.
Websuche