AMAZON ALEXA: WANN IST DER SPRACHASSISTENT GANZ OHR?

Ein Reaktions-Check

MARKTWÄCHTER
DIGITALE WELT

Kurzuntersuchung der Verbraucherzentralen
Dezember 2017

verbraucherzentrale

Herausgeber
Verbraucherzentrale NRW e.V.
Mintropstr. 27
40215 Düsseldorf
Tel. (0211) 3809 0
Fax. (0211) 3809 172
marktwaechter@verbraucherzentrale.nrw

Text: Verbraucherzentrale NRW e.V.

Stand: Dezember 2017

INHALT

1. Hintergrund....4
2. Methoden....6
3. Ergebnisse....9
   3.1 Deskriptive Statistiken....9
   3.2 Weitere Beobachtungen....11
4. Zusammenfassung....11
5. Quellenverzeichnis....13

1. Hintergrund

Was sind digitale Sprachassistenten? Digitale Sprachassistenten sind Programme, die über eine Spracherkennungs-Software die Aufzeichnung, Analyse und das anschließende Ausführen von Sprachanweisungen leisten sollen. Im Gegensatz zu herkömmlichen Schnittstellen zwischen Mensch und Maschine ist bei Digitalen Sprachassistenten das Betätigen von Tasten oder Bildschirmen nicht mehr erforderlich – diese Schnittstelle wird daher auch als „Voice User Interface" bezeichnet.[1] Digitale Sprachassistenten werden als eine Art Alltagshelfer beworben.[2] So können Sprachanweisungen, die sich an digitale Sprachassistenten richten, beispielsweise die Online-Suche von Informationen betreffen, die Verwaltung von Terminen und Kontakten, das Auslösen von Bestellvorgängen, aber auch die Bedienung vernetzter Gegenstände.

Wie funktionieren digitale Sprachassistenten? Digitale Sprachassistenten sind mit dem Internet verbunden. Sie hören die Umgebung kontinuierlich ab, um auf mögliche Befehle reagieren zu können. Die aktive Sprachaufzeichnung und -auswertung wird durch in der Regel vorgegebene Signalwörter ausgelöst: Nennt der Nutzer ein Signalwort, wird der anschließend gesprochene Befehl aufgezeichnet und bei Cloud-basierten Lösungen an die Server des Software-Anbieters geschickt. Dort wird der Sprachbefehl verarbeitet und – sofern er nicht vom Nutzer wieder gelöscht wird – dauerhaft gespeichert.[3] Das Ergebnis der Verarbeitung wird zurück an das Endgerät des Nutzers gesendet: Der Sprachassistent soll die gewünschte Antwort geben oder Handlung ausführen (z.B.: Licht an- und ausschalten, Bestellvorgänge auslösen, Informationen geben etc.). Laut einer repräsentativen Befragung von *Bitkom Research* würden 39 Prozent der Befragten digitale Sprachassistenten nutzen; das größte Interesse besteht dabei an der Sprachsteuerung von Haushaltsgeräten.[4]

Marktüberblick. Digitale Sprachassistenten werden als Software heute schon von verschiedenen großen Anbietern angeboten oder entwickelt. So werden – je nach Betriebssystem und Endgerät – Sprachassistenten wie *Siri* (*Apple*), *Cortana* (*Microsoft*), *SVoice* (*Samsung*) oder *Google Now* (*Google*) oft vorinstalliert – etwa auf dem Smartphone – mitgeliefert. Anbieter wie *IBM* (*Watson*)[5] oder *Lenovo* (*CAVA*)[6] drängen ebenfalls auf den internationalen Markt. Auf dem Smartphone, Tablet oder PC wurde *Google Now* im Jahr 2017 von 17 Millionen deutschen Verbrauchern genutzt.[7] Dem Gesamtmarkt der digitalen Sprachassistenten wird ein starkes Wachstum vorhergesagt.[8]
Für den Heimgebrauch kommen digitale Sprachassistenten derzeit in gesonderter Hardware verbaut auf den Markt – in der Regel handelt es sich um mit Mikrofonen ausgestattete Lautsprecher, die zukünftig auch über einen Display und Kameras verfügen können.[9] Im Jahr 2017 ist hierbei ein Wettrennen der Tech-Giganten zu beobachten: *Amazon*s „smarter" Lautsprecher *Echo* mit dem dazugehörigen Sprachservice *Alexa* ist seit Januar 2017 auf dem deutschen Markt verfügbar, es folgte *Google Home* Anfang August 2017. Andere Unternehmen, die die Technikbranche dominieren, arbeiten an der Entwicklung vergleichbarer Produkte. Ende 2017 soll *Apple*'s *Homepod* zumindest im englischsprachigen Ausland auf den Markt kommen.[10] Für *Microsoft's Cortana* wird derzeit das entsprechende Produkt

[1] z.B. https://msdn.microsoft.com/en-us/library/ms994650.aspx; da hierbei die Stimme als zentrales Eingabemedium fungiert, wird zeitweise auch vom „Internet of Voice" (dt.: Internet der Stimme) in Anlehnung an das sog. Internet der Dinge gesprochen, wobei letzteres die Vernetzung von Alltagsgegenständen betrifft.
[2] z.B. Wendel (2017); https://www.youtube.com/watch?v=jDmYcfM6mXY (Amazon Echo Werbung 2017, Stand 19.09.2017).
[3] z.B. https://www.amazon.com/gp/help/customer/display.html?nodeId=201809740 (Unterpunkt 1.3, Stand 19.09.2017).
[4] Befragt wurden 1004 Personen ab 14 Jahren; Bitkom (2016).
[5] http://androidmag.de/allgemein/kuenstliche-intelligenz-fuer-alle-ibm-macht-watson-fuer-jeden-verfuegbar/
[6] http://winfuture.de/news,98723.html
[7] Statista (2016).
[8] Statista (2017).
[9] z.B. *Amazon Echo Look*; Herbig (2017).
[10] z.B. Albus (2017), s. auch https://www.apple.com/newsroom/2017/06/homepod-reinvents-music-in-the-home/

Invoke vom Hi-Fi-Komponenten-Hersteller *Harman Kardon* entwickelt.[11] Branchen-Gigant *Samsung* entwickelt derzeit den Sprachservice *Bixby*, der verbaut in einem Lautsprecher die Sprachsteuerung der zahlreichen Home Entertainment Produkte (z.B. Smart TV) von *Samsung* ermöglichen soll. Medienberichten zufolge soll auch *Facebook* an der Entwicklung eines „smarten" Lautsprechers arbeiten, in dem *Facebooks* eigener Sprachservice „*M*" zum Einsatz käme.[12]

Datenschutzaspekte. Gerade die für den Heimgebrauch verfügbaren Sprachassistenten werfen verschiedene verbraucherrelevante Fragen auf – insbesondere in Bezug auf Privatsphäre, Datenschutz und Datensicherheit.
Da die Geräte über Mikrofone die private Umgebung des Nutzers kontinuierlich abhören, steht hier insbesondere die Frage im Raum, welche Daten verarbeitet und wo diese gespeichert werden. Daneben stellt sich die Frage, wie Anbieter Schutz vor Hacker-Angriffen bieten[13] und die Datenübertragung zu den Anbieter-Servern ausreichend sichern können.
Datenschutzaspekte werden jedoch bereits auf Ebene der Spracherkennung relevant: Denn erst wenn das aktivierende Signalwort genannt wird, soll die danach aufgezeichnete Spracheinheit zur Verarbeitung und dauerhaften Speicherung an externe Server gesendet werden. Daraus folgt, dass der Sprachassistent *ausschließlich* auf das eingestellte Signalwort reagieren sollte, damit ungewollte Aufzeichnungen vermieden werden; nur so ist eine datensparsame Nutzung möglich. Eine kürzlich aufgedeckte Panne bei Testversionen des Sprachassistenten *Google Mini* hingegen machen das Risiko ungewollter und unbemerkter Sprachaufzeichnung offensichtlich.[14]
Dies könnte ebenso zum Problem werden, wenn Begriffe als Signalwörter programmiert werden, die in dieser, in abgewandelter oder zusammengesetzter Form auch in der Alltagssprache regelmäßig vorkommen. Denn je nachdem wie *sensitiv* die Spracherkennungssoftware für den bloßen phonetischen – also klanglichen – (Teil-)Ausdruck ist, könnte es zu ungewünschten Aktivierungen des Sprachassistenten kommen und somit auch zu einer unerwünschten Datenübertragung. Um diesem Szenario näher auf den Grund zu gehen, wird in der vorliegenden Untersuchung – am Beispiel von *Amazon Alexa* – geprüft, inwieweit durch fehlerhafte Spracherkennung versehentliche Aufzeichnungen und somit unerwünschte Datenübertragungen bei der Verwendung digitaler Sprachassistenten möglich sind. Im Vordergrund steht die Frage, wie spezifisch die Spracherkennung ausschließlich auf das einprogrammierte Signalwort reagiert.

[11] Gurman & Frier (2017).
[12] http://www.itmagazine.ch/Artikel/65209/Facebook_soll_an_Lautsprecher_mit_Touchscreen_arbeiten.html
[13] z.B. Barnes (2017).
[14] Die auf Werbeveranstaltungen hauptsächlich an Journalisten verteilten *Google Home Minis* waren mit defekten Aktivierungsmechanismen ausgestattet, sodass diese fortwährend Geräusche, beispielsweise aus dem Fernseher, aber auch Unterhaltungen aufzeichneten und speicherten; s. z.B. Beer (2017).

2. Methoden

Untersuchungsgegenstand. Zum Zeitpunkt der Untersuchung (Juli 2017) war *Amazon*'s Lautsprecher *Echo* das einzige für den Heimgebrauch auf dem deutschen Markt verfügbare Produkt im Bereich der digitalen Sprachassistenten. Die Untersuchung wurde daher anhand von *Amazon Echo* durchgeführt, der in Kombination mit *Amazons* Sprachservice *Alexa* und mit der entsprechenden App geprüft wurde (*Alexa* Version 2.0.1063.0, letzte Aktualisierung am 13.05.2017; s. Abbildung 1). Bei *Amazon Echo* kann der Nutzer per Voreinstellung jederzeit zwischen vier verschiedenen Signalwörtern wählen („Alexa", „Amazon", „Echo", „Computer"). Es ist nicht möglich, mehr als ein Wort gleichzeitig als Signalwort einzustellen, sodass der Sprachservice für den Zeitraum der Einstellung immer nur auf eines der Signalwörter reagiert.

Abbildung 1. Fotografie des Lautsprechers (links) und Screenshot der App im Playstore (rechts).

Methodische Umsetzung. Die Grundlage der folgenden Untersuchung sind die Sprachmerkmale, die den digitalen Assistenten tatsächlich aktivieren sollen. Eine Informationsanfrage, die sich an den Assistenten richtet, könnte beispielsweise lauten: „Alexa, wie viele Einwohner hat Deutschland?". Deutlich wird, dass das Signalwort, so wie einprogrammiert (d.h. im Original), am Anfang einer Äußerung steht und eine Aktivierung des Sprachservices beabsichtigt. Für jedes der vier möglichen Signalwörter wurde zunächst eine solche Aussage als Vergleichsbasis getestet (*Vortest*, Einwohnerzahl verschiedener Länder).
Eine spezifische Spracherkennung müsste aus Sicht des Verbraucherschutzes mindestens feststellen können, ob nur ein ähnlich klingendes Wort beziehungsweise eine Worterweiterung genannt wurde und ob das betreffende Wort am Anfang eines Satzes und in befehlsähnlicher oder fragender Form genannt wird. Um zu überprüfen, inwieweit eine solche Differenzierung auf Ebene der Sprachaktivierung stattfindet, wurde eine in ihrer Struktur experimentelle Untersuchung durchgeführt. Hierzu wurde

Sprachmaterial in Form einer Reihe von Sätzen erstellt[15], innerhalb derer die folgenden Faktoren variiert wurden:

1. *Form des Signalworts*: Der ausgesprochene Satz enthält das Signalwort
   a. im Original
   b. leicht abgewandelt oder erweitert[16]
   c. stark abgewandelt[17]

2. *Position* des Signalworts im Satz: Das vermeintliche Signalwort befindet sich
   a. am Satzanfang, so wie es bei einem Sprachbefehl üblicherweise der Fall wäre.
   b. mitten im Satz, so wie es in einer normalen Unterhaltung üblicherweise der Fall wäre. Im Vergleich zur Befehlsform ist das Signalwort in einen Satzfluss eingebettet. [18]

Tabelle 1. Übersicht der Testbedingungen inkl. Anzahl der Testwörter.

Form des Signalworts	Position des Signalworts		Gesamtanzahl Testwörter
	Satzanfang	Satzmitte	
original	4	4	8
leicht abgewandelt	9	9	18
stark abgewandelt	0[a]	12	12
Gesamtanzahl Testwörter	11	25	38

[a] Starke Abwandlungen wurden nur in der Satzmitte verwendet, um die Testeinheiten auf Aussagen zu beschränken, wie sie auch in einer natürlichen Unterhaltung vorkommen könnten.

Testmaterial. Das Testmaterial beinhaltete die beschriebenen Variationen für jedes der vier Signalwörter (Tabelle 2). Starke Abwandlungen der Signalwörter wurden nur in der Satzmitte verwendet, um die Testeinheiten auf Aussagen zu beschränken, wie sie auch in einer natürlichen Unterhaltung vorkommen könnten. Wortabwandlungen wurden in Einzelfällen auch über ähnlich klingende Fremdwörter oder Spitznamen erreicht. Bei der Erstellung des Testmaterials konnten nicht gleich viele Abwandlungen für jedes Signalwort gefunden werden, daher unterscheidet sich die absolute Anzahl an Testwörtern für die vier Signalwörter (Tabelle 2).
Die Testwörter wurden jeweils innerhalb eines Testsatzes genannt, auf den noch ein weiterer Satz folgte (s. Anhang auf S. 13). Beispielweise wurde für das Signalwort „Alexa“ in der Bedingungskombination „Signalwort leicht abgewandelt, Nennung in der Satzmitte“ folgende Testeinheit eingesprochen:

> „Ich habe **Alex** gesagt, dass er die Pistole fallen lassen soll.
> Niemand muss erfahren, was wir letzten Sommer getan haben!“

Durch den Folgesatz wurde einerseits der Kontext der Äußerung verdeutlicht, und andererseits konnte somit überprüft werden, wann der Sprachassistent eine ggf. gestartete Aufnahme abbricht.[19]

Bewertung. Der oben genannte Beispielsatz zeigt auch, dass das Testmaterial keine Sätze enthielt, die inhaltlich tatsächlich einen an den Sprachservice gerichteten und ausführbaren Befehl oder eine Frage darstellten. Es gab also zu keinem Zeitpunkt die inhaltliche Absicht, den Sprachassistenten zu aktivieren. Insofern markierte rein inhaltlich betrachtet jede beobachtete Reaktion des Sprachassistenten einen falsch-positiven Alarm.[20] Da es auf technischer Ebene jedoch kaum zu erwarten ist, dass der

[15] Testmaterial in Anhang A auf Seite 13; die Konzeption, Erstellung des Testmaterials sowie die Durchführung und Dokumentation wurden vom DMW-Team der Verbraucherzentrale NRW vorgenommen.
[16] Die Abwandlungen der Original-Signalwörter wurden offen gesammelt und über eine Silbensuche im Deutschen Referenzkorpus ergänzt, z.B. Kupietz, Belica, Keibel, & Witt (2010), s. http://www1.ids-mannheim.de/kl/projekte/korpora.
[17] Die Klassifizierung der Testwörter in die Gruppen „leichte Abwandlung“ und „starke Abwandlung“ wurde nicht validiert und besteht daher nur unter Vorbehalt.
[18] Gemessen an der Wortanzahl des jeweiligen Satzes befinden sich die Testwörter jedoch nicht systematisch in der tatsächlichen Mitte des Satzes.
[19] Hatte die Aufzeichnung nach diesem Satz nicht geendet, waren die Sprecher angewiesen, stichprobenartig so lange weiter zu reden, bis dies der Fall war.
[20] In der Signal-Detection Theorie werden Reaktionen eines Akteurs je nach dem bewertet, ob

Sprachassistent inhaltlich differenzieren kann, ob er tatsächlich angesprochen wurde,[21] wurde die Bedingung Original/Satzanfang nicht bewertet. Reaktionen in allen anderen Bedingungen wurden als falsch-positive Alarme eingestuft (rote Markierung in Tabelle 4, Tabelle 5). Jede ausbleibende Reaktion wurde als korrekte Ablehnung gewertet (grüne Markierung in Tabelle 4, Tabelle 5).

Tabelle 2. Übersicht über die verwendeten Testwörter.

Form des Signalworts	Testwörter				Gesamt
original	**Alexa**	**Amazon**	**Echo**	**Computer**	4
leicht abgewandelt	Alex Alexis Alexander	Amazonas Amazone	Echos Tejo[a]	Computersystem Supercomputer	9
stark abgewandelt	Komplexer lexikal reflexartig	Klimazone	Ich schon Micho Psycho	absoluter Router komm Uta komm Peter akuter	12
Gesamt	7	4	6	8	**25**

[a] Spanisches Wort für „Eiche"; der Buchstabe „j" wird ähnlich dem deutschen „ch" ausgesprochen.

Durchführung. Zur Einrichtung des Sprachassistenten wurde ein neues Nutzerkonto bei *Amazon* eigens für den Untersuchungszweck erstellt. Im Anschluss wurde das Testmaterial von zwei Sprechern separat voneinander eingesprochen (Sprecher 1: männlich, Sprecher 2: weiblich; Testzeitraum: 29.06. bis 03.07.2017).[22] Beide Sprecher saßen in der gleichen Position, in einem Abstand von ungefähr drei Metern vom Lautsprecher entfernt, das Gesicht dem Lautsprecher zugewandt (s. Abbildung 2). Jede Testeinheit wurde pro Sprecher zehnmal vorgelesen.[23]

Dokumentation. Jeder Testdurchlauf wurde mit Hilfe einer Videokamera aufgezeichnet. Im Anschluss wurde anhand der Videoaufzeichnungen für jeden gesprochenen Satz dokumentiert, ob der Sprachservice bei Nennung des Testsatzes mit einem Lichtsignal reagiert (im Folgenden „Reaktionen") und wenn ja, wie lange er ungefähr benötigt, um festzustellen, dass es sich nicht um einen tatsächlichen Sprachbefehl handelt – das heißt wie lange der Lautsprecher über ein Leuchten signalisiert, dass die Aufzeichnung aktiv ist.[24]

a) der Akteur reagiert und der Stimulus, auf den er reagieren soll, präsent ist („Hits")
b) der Akteur reagiert, obwohl der Stimulus nicht präsent ist („false alarms / false positives")
c) der Akteur *nicht* reagiert, obwohl der Stimulus präsent ist („misses")
d) der Akteur *nicht* reagiert und der Stimulus *nicht* präsent ist („correct rejections");

s. z.B. Lerman et al. (2010). In der vorliegenden Untersuchung wird angenommen, dass der Stimulus nicht präsent ist, wenn es seitens des Sprechers keine Intention gibt, den Sprachassistenten zu aktivieren.

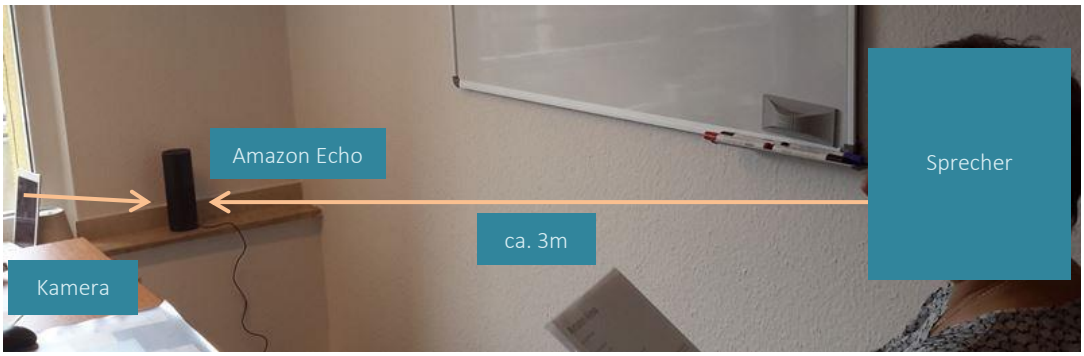
[21] Dies ist v.a. nicht zu erwarten, weil die lokale Spracherkennung lediglich das Signalwort erkennen und den Aufzeichnungs- und Übertragungsvorgang auslösen soll. Die eigentliche Sprachverarbeitung des (vermeintlichen) Befehls findet jedoch erst statt, nachdem die Aufzeichnung schon gesendet wurde; s. z.B. http://t3n.de/news/amazon-echo-google-home-sprachdaten-loeschen-881818/.

[22] Eine der Testbedingungen musste bei Sprecher 2 am 14.07. erneut durchgeführt werden, da die Videoaufzeichnung des ersten Durchlaufs unvollständig war. In die Auswertung fließt nur die zweite, vollständig dokumentierte Aufzeichnung mit ein.

[23] Zufällige Variationen in der Aussprache und Tonalität waren hierbei möglich, wurden jedoch nicht systematisch eingesetzt.

[24] Es wurde nicht überprüft, ob die Dauer des Lautsprecher-Leuchtens mit der Dauer der dazugehörigen Aufzeichnung übereinstimmt, die grundsätzlich innerhalb der App abrufbar ist.

Abbildung 2. Testaufbau (nachgestellte Fotografie).



3. Ergebnisse

3.1 Deskriptive Statistiken

Sprecherunterschiede. Zunächst wurde über alle Bedingungskombinationen inklusive der Informationsanfrage (Baseline) hinweg geprüft, ob es systematische Ergebnisunterschiede zwischen den beiden Sprechern gibt. Hierbei zeigt sich, dass der Sprachservice zwar in einigen Testbedingungen nicht bei beiden Sprechern reagiert hatte, das Verhältnis über die Bedingungen hinweg jedoch ausgeglichen ist.[25] Da keine systematischen Sprecher- und somit Geschlechterunterschiede festgestellt wurden, wurden die Testdurchläufe beider Sprecher für die weitere Auswertung gemeinsam betrachtet. Somit wurden für jede Testbedingung zwanzig Wiederholungen pro Testwort und Bedingung analysiert.

Vortest. Die Ergebnisse zeigen zunächst, dass der Sprachassistent auf absichtsvolle Befehle in fast 100 Prozent aller Testdurchläufe reagiert (Tabelle 3). Die Sprachschnittstelle ist also ausgesprochen sensitiv für die einprogrammierten Signalwörter.

Tabelle 3. Vortest:
Reaktionen und maximale Aufzeichnungsdauer.

Signalwort	*N*	%	Max Sek.
Alexa	20	100	8
Amazon	19	95	4
Echo	19	95	5
Computer	20	100	8

Reaktionen pro Testbedingung und Testwort. Das Ergebnismuster in der Bedingung, in der das Signalwort im Original am Satzanfang steht (Tabelle 4), ähnelt stark den Ergebnissen des Vortests. Insofern bestätigt sich die Erwartung, dass der Sprachassistent auf inhaltlicher Ebene nicht erkennen kann, ob sich eine Sprachäußerung tatsächlich an ihn richtet: Wird eine Aufnahme in dieser Bedingung initiiert, versucht Alexa auch – notwendigerweise vergeblich – einen Befehl auszuführen.

Tabelle 4. Anzahl und mittlere Prozentzahl an Reaktionen
des Sprachservices pro Testbedingung.[a]

	Satzanfang			Satzmitte		
Form Signalwort	*N Durch-gänge*[b]	*N Reak-tionen*	*% Reak-tionen*	*N Durch-gänge*[b]	*N Reak-tionen*	*% Reak-tionen*

[25] Ein t-Test zeigt, dass der Sprachservice bei Sprecher 1 (M = 4.55, SD = 4.36) nicht signifikant häufiger reagiert als bei Sprecher 2 (M = 4.48, SD = 4.43), $t(82) = .08$, $p = .94$. Es gibt keinen systematischen Unterschied in der durchschnittlichen Aufzeichnungsdauer zwischen Sprecher 1 (M = 5.84, SD = 2.06) und Sprecher 2 (M = .78, SD = 1.94), $t(49) = -1.68$, p =.10.

Vortest	80	78	98	-	-	-
Original	80	74	93	80	37	**46**
leicht abgewandelt	180	129	**72**	180	54	**30**
stark abgewandelt	-	-	-	240	7	**3**

[a] Beige markiert wurden Bedingungen, in denen falsch-positive Alarme beobachtet wurden. Nicht bewertet wurde die Bedingung Original/Satzanfang, da die hier zu verzeichnenden Reaktionen zwar inhaltlich, nicht jedoch aus Perspektive des Voice-Interfaces als falsch-positiv bezeichnet werden kann.
[b] Anzahl der Testdurchgänge über alle Testwörter in der jeweiligen Bedingung hinweg. Die unterschiedliche Anzahl ergibt sich durch die Anzahl an Testwörtern in den jeweiligen Bedingungen (s. Tabelle 1).

Bei leicht abgewandelten oder leicht erweiterten Signalwörtern reagiert der Sprachservice noch in 72 Prozent der Fälle, wenn das entsprechende Signalwort am Anfang des Satzes steht.
Auf Ebene der einzelnen Testwörter (Tabelle 5) zeigt sich, dass der Sprachservice tendenziell eher zu reagieren scheint, wenn das Signalwort komplett in dem abgewandelten Wort vorhanden ist, wie zum Beispiel bei **Alexa**nder (12 von 20; im Vergleich zu Alex: 2 von 20).
Taucht ein vermeintliches Original-Signalwort in der Satzmitte auf, sind noch 46 Prozent Reaktionen zu verzeichnen; wenn es sich um ein leicht abgewandeltes oder erweitertes Wort handelt, sind noch 30 Prozent falsch-positiver Reaktionen zu verzeichnen. Wird das Signalwort stark abgewandelt in der Satzmitte genannt, sind immer noch Reaktionen zu verzeichnen, auch wenn die Reaktionsrate auf drei Prozent sinkt. Auf Ebene der einzelnen Testwörter zeigt sich: Nur bei dem Signalwort „Computer“ aktivierten solch starke Abwandlungen mitten im Satz die Sprachaufzeichnung, zum Beispiel bei den Wörtern „absoluter“ oder „Komm Peter“.
Ungeachtet der Reaktionen pro Testwort zeigt sich, dass am Satzanfang alle neun getesteten leichten Abwandlungen mindestens einmal eine falsch-positive Reaktion des Sprachassistenten hervorgerufen haben. In der Satzmitte reagierte der Sprachassistent mindestens einmal auf alle Original-Signalwörter sowie auf 6 der 25 getesteten Abwandlungen.

Tabelle 5. Anzahl der Reaktionen und maximale Aufzeichnungsdauer pro Signalwort.

Form Signalwort	Testwort	Satzanfang		Satzmitte	
		N	Max Sek.	*N*	Max Sek.
Original	Alexa	20	11	3	10
leicht abgewandelt	Alex	2	3	0	-
	Alexis	17	11	0	-
	Alexander	12	11	0	-
stark abgewandelt	komplexer	-	-	0	-
	lexikal	-	-	0	-
	reflexartig	-	-	0	-
Original	Amazon	17	**19**	20	9
leicht abgewandelt	Amazone	18	10	15	10
	Amazonas	19	11	11	10
stark abgewandelt	Klimazone	-	-	0	-
Original	Echo	18	**23**	2	11
leicht abgewandelt	Echos	18	10	0	-
	Tejo	13	10	0	-
stark abgewandelt	Ich schon	-	-	0	-
	Micho	-	-	0	-
	Psycho	-	-	0	-
Original	Computer	19	11	12	**13**
leicht abgewandelt	Supercomputer	15	10	17	9
	Computersystem	15	11	11	11
stark abgewandelt	Komm Uta	-	-	0	-
	Komm Peter	-	-	6	7
	absoluter	-	-	1	8

	akuter	-	-	0	-
	Router	-	-	0	-

[a] Beige Markierung: falsch-positive Alarme; grüne Markierung: korrekte Ablehnungen (d.h. Nicht-Reaktionen). Nicht bewertet wurde die Bedingung Original/Satzanfang, da die hier zu verzeichnenden Reaktionen zwar inhaltlich, nicht jedoch aus Perspektive des Voice-Interfaces als falsch-positiv bezeichnet werden kann.

3.2 Weitere Beobachtungen

Im Zuge der Untersuchung waren vereinzelte, nicht systematisch dokumentierte Auffälligkeiten zu beobachten, die im Folgenden unabhängig von den quantitativen Ergebnissen (s. Abschnitt 3.1) beschrieben werden.

Aufzeichnungsdauer. Reagiert *Amazon*'s Sprachassistent auf ein Signalwort, können relativ lange Gesprächsaufzeichnungen entstehen (bis zu 23 Sekunden). Beobachtet wurde dies vor allem bei Sprachausschnitten, in denen die „Unterhaltung" nach Nennung des vermeintlichen Signalworts einfach weitergeht und nicht – wie bei einem klassischen Befehl – nach einem Satz endet. Der Grund hierfür ist vermutlich, dass der Spracherkennungs-Algorithmus so lange wie möglich versucht, einen Befehl aus dem Gesagten abzuleiten, der tatsächlich umgesetzt werden kann.

Training. Es wurde nicht systematisch überprüft, ob *Alexa* im Zeitverlauf weniger falsch-positive Aufzeichnungen produzierte. Dies wäre denkbar, da *Amazon* für seinen Sprachassistenten damit wirbt, dass dieser mit der Zeit immer mehr dazu lerne.[26] Dieser Lernprozess kann beschleunigt werden, indem der Nutzer in der App bewertet, ob eine gegebene Antwort von *Alexa hilfreich* war.[27] Allerdings lässt sich keine Rückmeldung darüber abgeben, ob die Aufzeichnung überhaupt erwünscht war. Insofern lässt sich die Spracherkennung nicht aktiv in Richtung höhere Reaktionsspezifizität trainieren.

Sprachdaten löschen. Nutzer können ihre Sprachaufzeichnungen innerhalb der App oder in ihrem *Amazon*-Account löschen. Hierbei wird nur innerhalb der App die Möglichkeit geboten, Aufzeichnungen einzeln zu löschen. Dies ist gerade in der derzeit noch verhältnismäßig unausgereiften Anwendung jedoch mühsam, da die App beispielsweise nach Aufrufen einer Aufzeichnung immer wieder zum Beginn der Liste springt. Über den *Amazon*-Account, innerhalb dessen die Auswahl und Löschung einzelner unerwünschter Aufzeichnungen einfacher handhabbar wäre, lassen sich nur alle Aufzeichnungen auf einmal für ein Gerät löschen. Nicht gelöschte Aufzeichnungen werden dauerhaft auf den *Amazon*-Servern gespeichert.

Zufällige Reaktionen. Der Sprachassistent reagierte in mehreren Gesprächssituationen, ohne dass die genannten Wörter eine nachvollziehbare Ähnlichkeit mit dem eingestellten Signalwort hatten. Diese Reaktionen auf die zuvor geäußerten Wörter konnten anschließend jedoch nicht innerhalb der Testsystematik (s. Abschnitt 2) wiedergegeben werden. Insofern wirkte die Spracherkennung in Bezug auf die Signalwörter zeitweise zufällig.[28]

4. Zusammenfassung

Die Ergebnisse der vorliegenden Untersuchung legen den Schluss nahe, dass *Amazons* Spracherkennungs-Software sehr sensitiv für Sprachäußerungen ist, die dem Signalwort klangmäßig ähnlich sind. Dabei spielt durchaus eine Rolle, an welcher Position im Satz das vermeintliche Signalwort steht. Bei Nennung des

[26] https://www.amazon.de/dp/B01DFKBG54/ref=gw_aucc_dopp_multidoppbisc_0717?pf_rd_p=cd3f51a0-a233-42fa-9873-baa156f35c64&pf_rd_r=AY1ZA5ZYZE3SBN4HE91K (Stand: 01.08.2017).
[27] Im Zuge der Testung wurden keine solcher Bewertungen vorgenommen.
[28] Diese Beobachtung fand außerhalb der systematischen Testung statt und konnte daher nicht aufgezeichnet werden.

Originalwortes am Satzanfang reagierte *Alexa* doppelt so häufig (93%) als wenn das Signalwort mitten im Satz vorkommt. Auch hier reagiert der Sprachassistent jedoch noch in 46 Prozent der Fälle. Übergreifend lässt sich also aus Verbrauchersicht bemängeln, dass die Spracherkennung insofern unspezifisch ist, als auch Abwandlungen des Originalworts und dessen Nennung im Gesprächsfluss den Sprachservice aktivieren können. Diese fehlende Begrenzung und teilweise zufällig wirkende Aktivierung kann zu unerwünschten Gesprächsmitschnitten im Alltag führen. Diese Gesprächsmitschnitte werden – sofern der Nutzer sie nicht löscht – automatisch und dauerhaft auf den Anbieter-Servern gespeichert. Abschließend soll darauf hingewiesen werden, dass das Ausmaß der skizzierten Problematik im Wesentlichen davon abhängt, welche Wörter als Signalwörter eingestellt werden können. Begriffe und Namen der Alltagssprache sind naturgemäß anfälliger für die aufgeführten Fehler als seltene Wörter oder Kunstwörter, die eigens für diesen Zweck erfunden wurden. Insofern sind die Ergebnisse der vorliegenden Untersuchung nicht ohne Vorbehalt auf andere digitale Sprachassistenten verallgemeinerbar, zeigen aber eine grundsätzliche Problematik digitaler Sprachassistenten auf.

5. Quellenverzeichnis

Albus, D. (2017). Hat Samsung einen smarten Lautsprecher mit Bixby patentiert? *TurnOn. Das Saturn Magazin*. Abgerufen von https://www.turn-on.de/lifestyle/news/hat-samsung-einen-smarten-lautsprecher-mit-bixby-patentiert-267916 (Stand: 01.08.2017).

Barnes, M. (2017). Alexa, are you listening? *MWR Labs*. Abgerufen von https://labs.mwrinfosecurity.com/blog/alexa-are-you-listening/ (Stand: 12.12.2017).

Beer, K. (11. Oktober 2017). Datenschutzpanne: Testgeräte von Google Home Mini hörten ständig zu. *Heise.de.* Abgerufen von https://www.heise.de/newsticker/meldung/Datenschutzpanne-Testgeraete-von-Google-Home-Mini-hoerten-staendig-zu-3856399.html (Stand: 20.11.2017).

Bitkom (2016). Digitale Sprachassistenten als intelligente Haushaltshelfer. *Bitkom Research*. Abgerufen von https://www.bitkom.org/Presse/Presseinformation/Digitale-Sprachassistenten-als-intelligente-Haushaltshelfer.html (Stand: 01.08.2017).

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler* (6. Auflage). Heidelberg: Springer Medizin Verlag.

Dachwitz, I. (2017). Bitkom-Umfrage: Kunden wollen Sprachassistenten nicht nutzen, weil sie Unternehmen nicht vertrauen. *Netzpolitik.org.* Abgerufen von https://netzpolitik.org/2016/bitkom-umfrage-kunden-wollen-sprachassistenten-nicht-nutzen-weil-sie-unternehmen-nicht-vertrauen/ (Stand: 03.08.2017).

Herbig, D. (2017). Echo Look: Amazons Alexa-Kamera für Modebewusste. *Heise.de*. Abgerufen von https://www.heise.de/newsticker/meldung/Echo-Look-Amazons-Alexa-Kamera-fuer-Modebewusste-3698219.html (Stand: 01.08.2017).

Kupietz, M., Belica, C., Keibel, H., & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, N. et al. (eds.): *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta: European Language Resources Association (ELRA), S. 1848–1854.

Lerman, C. L., Tetreault, A., Hovanetz, A., Bellaci, E., Miller, J., Karp, A., Mahmood, A., Strobel, M., Mullen, S., Keyl, A., & Toupard, A. (2010). Applying Signal-Detection Theory to the study of observer accuracy and bias in behavioral assessment. Journal of Applied Behavior Analysis, 43(2), 195–213. http://doi.org/10.1901/jaba.2010.43-195.

Rheingold Institut (2017). Wie Alexa die geheimen Wünsche ihrer Nutzer erfüllt. Rheingold. Abgerufen von http://www.rheingold-marktforschung.de/veroeffentlichungen/artikel/Pilotstudie_Wie_Alexa_die_geheimen_Wuensche_ihrer_Nutzer_erfuellt.html (Stand: 01.08.2017).

Schoon, B. (2017). Google Home proves 6 times better at searches than Amazon Echo in 3,000 question quiz. 9to5Google.com. Abgerufen von https://9to5google.com/2017/06/26/google-home-six-times-better-search-amazon-echo/ (Stand: 01.08.2017).

Gurman, M. & Frier, S. (2017). Facebook Is Working on a Video Chat Device *Bloomberg.com.* Abgerufen von https://www.bloomberg.com/news/articles/2017-08-01/facebook-is-said-to-work-on-video-chat-device-in-hardware-push (Stand: 12.12.2017).

Statista. (2016). Anzahl der Internetnutzer, die auf einen virtuellen Sprachassistenten zurückgreifen in Deutschland im Jahr 2016 (in Millionen). Berlin: Statista Digital Market Outlook.

Statista. (2017). Umsatz mit virtuellen digitalen Assistenten für Endkunden im Jahr 2015 sowie eine Prognose bis 2021 (in Millionen US-Dollar). Berlin: Tractica.

Tucker, D. (2017). Voice User Interface Design - Purpose and Process. *Microsoft Developer Network*. Abgerufen von https://msdn.microsoft.com/en-us/library/ms994650.aspx (Stand: 01.08.2017).

Wendel, M. (2017). Amazon Alexa: Alle wichtigen Sprachbefehle im Überblick. *Home&Smart*. Abgerufen von https://www.homeandsmart.de/amazon-alexa-alle-wichtigen-sprachbefehle (Stand: 01.08.2017).

ALEXA	
ORIGINAL - SATZANFANG	
Alexa, lass die Pistole sofort fallen	Niemand muss erfahren, was wir letzten Sommer getan haben!
LEICHT ABGEWANDELT - SATZANFANG	
Alexander, lass die Pistole sofort fallen.	Niemand muss erfahren, was wir letzten Sommer getan haben!
Alex, lass die Pistole sofort fallen.	Niemand muss erfahren, was wir letzten Sommer getan haben!
Alexis, lass die Pistole sofort fallen.	Niemand muss erfahren, was wir letzten Sommer getan haben!
ORIGINAL - SATZMITTE	
Ich habe Alexa gesagt, dass sie die Pistole fallen lassen soll.	Niemand muss erfahren, was wir letzten Sommer getan haben!
LEICHT ABGEWANDELT - SATZMITTE	
Ich habe Alexander gesagt, dass er die Pistole fallen lassen soll.	Niemand muss erfahren, was wir letzten Sommer getan haben!
Ich habe Alex gesagt, dass er die Pistole fallen lassen soll.	Niemand muss erfahren, was wir letzten Sommer getan haben!
Ich habe Alexis gesagt, dass er die Pistole fallen lassen soll.	Niemand muss erfahren, was wir letzten Sommer getan haben!
STARK ABGEWANDELT - SATZMITTE	
Ich habe gestern ganz reflexartig einen Flug nach Paris gebucht.	Jetzt brauche ich unbedingt noch ein Hotel für die Zeit.
Um den Text lexikal zu analysieren, brauche ich ein Fachbuch.	Hast du eine Idee, welches sich dafür eignet?
Das Ganze ist doch komplexer als ich dachte.	Vielleicht kaufe ich lieber erst ein Fachbuch.

AMAZON	
ORIGINAL - SATZANFANG	
Amazon, das ist ein riesiges Unternehmen.	Wir sollten endlich eine gemeinsame Weihnachtsfeier organisieren.
LEICHT ABGEWANDELT - SATZANFANG	
Amazonas, das ist ein riesiges Unternehmen.	Wir sollten endlich eine gemeinsame Weihnachtsfeier organisieren.
Amazone, das ist ein riesiges Unternehmen.	Wir sollten endlich eine gemeinsame Weihnachtsfeier organisieren.
ORIGINAL - SATZMITTE	
Ich habe bei Amazon einen Laptop bestellt.	Leider war ich überhaupt nicht zufrieden mit dem Verkäufer.
LEICHT ABGEWANDELT - SATZMITTE	
Ich möchte unbedingt Urlaub am Amazonas machen.	Jetzt brauche ich nur noch einen passenden Reiseführer.
Meine Freundin sieht aus wie eine Amazone und ist sehr stark.	Sie besiegt sogar ihren Bruder im Armdrücken.
STARK ABGEWANDELT - SATZMITTE	
Wir fliegen in eine ganz andere Klimazone.	Ich muss mir unbedingt passende Anziehsachen dafür kaufen.
Ich möchte unbedingt in den Urlaub, aber ich weiß gar nicht wohin.	Ich werde mal online nach einer Reise suchen.

ECHO	
ORIGINAL - SATZANFANG	
Echo, aus dem Buch von Walter Moers	Vielleicht sollte ich das Buch morgen online bestellen.
LEICHT ABGEWANDELT - SATZANFANG	
Echos, sind Wiederhalle von Tönen.	In den Bergen haben wir oft welche gehört.
Tejo, ist das spanische Wort für Eibe	Das habe ich im letzten Spanien-Urlaub gelernt
ORIGINAL - SATZMITTE	
Ich hab in der neuen „Echo“ einen Artikel gelesen.	Dort stand, dass man jeden Morgen einen Smoothie trinken sollte.
LEICHT ABGEWANDELT - SATZMITTE	
In den Bergen haben wir ständig Echos gehört.	Meine Wanderschuhe sind nun aber komplett durchgelaufen.
Ich habe gelernt, dass „Tejo“ das spanische Wort für Eibe ist.	Ich brauche für so was aber eigentlich ein Sprachlexikon.
STARK ABGEWANDELT - SATZMITTE	
Dieser Psycho hat gestern schon wieder bei mir angerufen.	So langsam ist es mir ganz schön unheimlich.
Darüber habe ich schon oft nachgedacht.	Vielleicht bestell ich den Fernseher nun einfach.
Hierzu hatte Micho mich letztens angerufen.	Ich weiß aber nicht, ob er den Fernseher nun bestellt hat.

COMPUTER	
ORIGINAL - SATZANFANG	
Computer, du bist mein bester Freund	Mit keinem rede ich so gerne wie mit dir, denn du hörst immer zu.
LEICHT ABGEWANDELT - SATZANFANG	
Computersystem, du bist mein bester Freund.	Mit keinem rede ich so gerne wie mit dir, denn du hörst immer zu.
Supercomputer, du bist mein bester Freund.	Mit keinem rede ich so gerne wie mit dir, denn du hörst immer zu.
ORIGINAL - SATZMITTE	
Ich habe den Computer neugestartet	Ich will schauen, ob die gestohlene Office-Version auch wirklich funktioniert.
LEICHT ABGEWANDELT - SATZMITTE	
Ich habe das Computersystem neugestartet.	Ich will schauen, ob die gestohlene Office-Version auch wirklich funktioniert.
Ich habe den Supercomputer neugestartet.	Ich will schauen, ob die gestohlene Office-Version auch wirklich funktioniert.
STARK ABGEWANDELT - SATZMITTE	
Er ist mit akuter Bronchitis zum Arzt gegangen.	Das Wartezimmer dort war zum Bersten voll.
Ich rief „komm, Peter“, wir müssen gehen.	Aber er wollte unbedingt noch den Krimi zu Ende schauen.
Ich rief „komm, Uta“, wir müssen gehen.	Aber er wollte unbedingt noch den Krimi zu Ende schauen.
Ich habe unseren Router neu konfiguriert.	Hoffentlich funktioniert das WLAN-Sniffing jetzt besser.
Das ist doch absoluter Wahnsinn, was du da tust.	Schlaf lieber noch eine Nacht drüber!