Aufbau einer unabhängigen natürlich-sprachlichen Mensch-Roboter-Interaktion

---

Martin Eisoldt

30.10.2019

Technische Universität Dresden

Fakultät Informatik
Institut für Angewandte Informatik
Professur für Mensch-Computer-Interaktion

Masterarbeit

Aufbau einer unabhängigen natürlich-sprachlichen Mensch-Roboter-Interaktion

Martin Eisoldt

Hochschullehrer Prof. Dr. rer. nat. habil. Gerhard Weber
Professur für Mensch-Computer-Interaktion
Technische Universität Dresden

Betreuer David Gollasch, M. Sc.

Eingereicht am: 30.10.2019

Martin Eisoldt
Aufbau einer unabhängigen natürlich-sprachlichen Mensch-Roboter-Interaktion
Masterarbeit, 30.10.2019
verantwortlicher Hochschullehrer: Prof. Dr. rer. nat. habil. Gerhard Weber
Fachbetreuer: David Gollasch, M. Sc.
Technische Universität Dresden
Fakultät Informatik
Institut für Angewandte Informatik
Professur für Mensch-Computer-Interaktion
01062 Dresden

Erklärung

Ich erkläre, dass ich die vorliegende Arbeit mit dem Titel *Aufbau einer unabhängigen natürlich-sprachlichen Mensch-Roboter-Interaktion* selbstständig unter Angabe aller Zitate angefertigt und dabei ausschließlich die aufgeführte Literatur und genannten Hilfsmittel verwendet habe.

Dresden, 30.10.2019

Martin Eisoldt

Abstrakt

Im Zusammenhang mit der zunehmenden Alterung der Gesellschaft scheint es attraktiv, gewisse unterstützende Tätigkeiten durch Assistenzroboter durchführen zu lassen. Um mit diesen interagieren zu können, bietet sich gesprochene, natürliche Sprache an. Da kommerzielle Sprachassistenten allerdings häufig zu Bedenken im Hinblick auf den Datenschutz führen, sollte dafür vorzugsweise ein Open-Source System eingesetzt werden. Bei dem Vergleich verschiedener Assistenzsystem hat sich *Mycroft Ai* als das geeignetste System erwiesen hat. Auf Basis der Analyse von Einsatzszenarien von Assistenzrobotern konnte ein allgemeingültiges Konzept für die Interaktion erstellt werden, das es Menschen ohne Vorkenntnisse erlaubt, mit dem Roboter zu interagieren. Diese Interaktion fühlt sich dabei für den Nutzer weitestgehend natürlich an, während zeitgleich durch das System der Datenschutz gewährt wird. Diese Aussagen wurde auf durch Aussagen von Probanden einer Pilotstudie gestützt.

Abstract

As modern society grows older, it seems useful to use service robots for certain tasks. To keep interaction intuitive and suitable for people with different capabilities, using spoken language promises the best results. As the concerns about privacy with commercial voice assistants are always present, an open-source solution is fitting better. By comparing different voice assistants, *Mycroft Ai* turned out to be the best suiting. Based on the analysis of different use cases for service robots, a universal concept for Human-Robot interaction could be created. That concepts allows people to interact with the robot, regardless of their knowledge whilst keeping a certain natural feeling and guaranteeing privacy. These statements could be backed up by the responses of probands in a pilot study.

Zielstellung

Kontext Verschiedentliche Motivationen begründen einen sich erhärtenden Bedarf an smarter,assistiver Technologie in Form autonomer Assistenzroboter. Ein gesellschaftlicher Druck geht dabei besonders von der alternden Bevölkerung in Verbindung mit dem angespannten Pflegesektor aus. Hier sind die Hoffnungen groß, das Leben von alten Menschen sowie von Pflegekräften deutlich zu erleichtern, indem Assistenzroboter zukünftig ein breites Spektrum von Unterstützung in unterschiedlichen Alltagssituationen bieten können. Zwar sind die Entwicklungen in der Robotik rasant schnell, aber reale Anwendungsfälle beschränken sich bislang eher auf industrielle Fertigungsroboter. Ein grundlegendes Problem im Bereich der Assistenzroboter ist der Mangel an Funktionsvielfalt, welche einen Roboter überhaupt erst interessant macht. Ein Lösungsansatz ist hier das Bereitstellen einer gemeinsamen, erweiterbaren Entwicklungsplattform. Der Segway Robotics „Loomo“ bietet hierfür eine Basis bestehend aus Hardware- und Softwareplattform. Die Hardware besteht aus einem Segway/Ninebot Self-Balancing Vehicle in Kombination mit einer Intel-Atom-basierten Recheneinheit und verschiedentlichen Sensoren und Aktuatoren zur Wahrnehmung und Interaktion mit der Umgebung. Die Softwareplattform bildet Android in Verbindung mit einem SDK zur Ansteuerung der Sensorik und Aktorik. Die Implementierung von Funktionen für spezifische Anwendungsfälle erfolgt in Form von Android-Apps mit Zugriff zum Steuerungs-SDK. Weiterhin erlaubt Loomo auch die Erweiterung der Hardware mittels Erweiterungskupplung bestehend aus belastbarer Metallaufhängung, USB-Anbindung und zusätzlicher Stromversorgung.

Projektziel Den Kern einer guten Mensch-Roboter-Interaktion bildet eine leistungsstarke und überzeugende Sprachinteraktion. Vor dem Hintergrund des Einsatzes im Umfeld älterer Menschen ist insbesondere die natürlich-sprachliche Interaktion wichtig, sodass statische Befehle eine unzureichende Umsetzung bildeten. Die Integration bestehender Sprachassistenten (speziell Amazon Alexa) ist ein Weg, der gegangen werden kann. Kommen jedoch datenschutzrechtliche Erwägungen hinzu, ist ein alternativer Weg zu beschreiten. Welche Möglichkeiten es hierzu gibt und welche Grenzen solche Systeme haben, soll Gegenstand dieser Arbeit sein. Konkret könnte das Open-Source-Projekt MyCroft.ai von Interesse sein, von Anbietern wie Amazon

etc. unabhängig zu sein und dennoch eine gute natürlich-sprachliche Interaktion zu ermöglichen.
Ziel dieser Arbeit soll es speziell auch sein, die Machbarkeit und Leistungsfähigkeit von MyCroft.ai im Umfeld der Assistenzrobotik zu untersuchen. Hierzu soll auch eine prototypische Implementierung erfolgen.

Schwerpunkte

- Einarbeitung in die folgenden Themengebiete inklusive Analyse des aktuellen Forschungsstandes:
  - Sprachassistenten im Allgemeinen. Verfahren der Erkennung der Nutzerintention.
  - Entwicklung von Steuerungs-Apps für Segway Robotics Loomo
  - Implementierung von MyCroft.ai in einer Android-Umgebung
- Analyse der beschriebenen Einsatzszenarien hinsichtlich der Machbarkeit und Eignung zur Implementierung einer unabhängigen Sprachinteraktion mit einem Assistenzroboter
- Entwicklung eines systematischen Konzepts zur Umsetzung der Sprachinteraktion mit MyCroft.ai
- Umsetzung des Konzepts und Implementierung von einfachen Steuerungsbefehlen für den Roboter
- Evaluation und Auswertung des erarbeiteten Verfahrens auf angemessene, wissenschaftliche Weise
- Dokumentation der Ergebnisse in geeigneter, wissenschaftlicher Form

Inhaltsverzeichnis

1 Einleitung

Die aktuelle Entwicklung der Technik weist klar die Tendenz auf, dass in immer mehr Bereichen des des Alltags Roboter zum Einsatz kommen. Diese können zunehmend mehr Aufgaben des täglichen Lebens übernehmen und entwickeln sich somit zu echten Assistenzrobotern. Jedoch variieren diese Aufgaben mit jedem Tag. So ist es nicht ausreichend, dass die Roboter zu vorgegebenen Zeiten die immer gleichen Tätigkeiten durchführen. Vielmehr müssen sie mit dem Nutzer interagieren und flexibel auf die Anforderungen reagieren. Eine solche Interaktion lässt sich beispielsweise über gesprochene Sprache umsetzen. Jedoch ist es dafür notwendig, dass die gesprochenen Befehle durch die Maschine korrekt erkannt werden.
Aktuell sind große Unternehmen bereits damit beschäftigt, Sprachassistenten zu erstellen. Zu diesen gehört beispielsweise Amazon Alexa oder Google Assistant, welche bereits in zahlreichen Produkten der Hersteller zum Einsatz kommen. An dieser Stelle stellen sicher aber immer auch datenschutzrechtliche Bedenken, da durch die Helfer prinzipiell jedes Geräusch in der Umgebung mitgeschnitten werden kann. Für europäische Nutzer ist auf den ersten Blick auch nicht zu erkennen, welche Gesetze für die Datenverarbeitung zur Geltung kommen, da die großen IT-Konzerne ihren Sitz häufig in den USA haben und dortige Gesetze durchaus von den europäischen abweichen. So berichtet Pfeifle [Pfe18] darüber, dass Strafverfolgungsbehörden in den USA ohne Zustimmung durch den Nutzer Zugriff auf Aufzeichnungen erhalten haben, die von Alexa erstellt wurden.
Für die Verarbeitung einer Nutzereingabe durch kommerzielle Anbieter wird diese an eine Cloud geschickt. Dabei handelt es sich zumeist um eine unternehmenseigene Infrastruktur, von welcher das Ergebnis danach wieder zurück an den Nutzer geschickt wird. Da die geografische Lage dieser Rechenzentren dem Nutzer nicht bekannt ist, kann auch nicht eindeutig festgestellt werden, auf welcher Gesetzesgrundlage die Datenverarbeitung durchgeführt wird. Des Weiteren stellt Pfeifle [Pfe18] heraus, dass durch Alexa erzeugte Aufzeichnungen mindestens teilweise gespeichert werden und für weiteres Training verwendet werden. Auch wird in verschiedenen Medien immer wieder über neue Probleme berichtet. So wurde im April 2019 aufgedeckt, dass sich Amazonmitarbeiter die Aufzeichnungen der Geräte anhören, um diese dann für die Verbesserung der Geräte zu verwenden. Darüber wurden die Nutzer zu keinem Zeitpunkt informiert, so dass sensible Informationen an Dritte gelangen konnten. [1]. Andere Recherchen haben auch ergeben, dass andere Hersteller, zumin-

[1] https://www.spiegel.de/netzwelt/gadgets/amazon-mitarbeiter-hoeren-sich-tausende-privatgespraeche-mit-alexa-an-a-1262315 [Abgerufen am 20.04.2019]

dest bislang, ähnlich verfahren sind. Berichten aus dem Sommer 2019 zu Folge ist Apple auf ähnliche Art und Weise vorgegangen [2].
Die Vermutung, dass solche Vorgänge für Verunsicherung bei Nutzern sorgen kann, wird durch Lau et al. [Lau+18] bestätigt. So stellen sie fest, dass besonders Personen, die keinen Sprachassistenten besitzen, große Bedenken bezüglich des Schutzes der Privatsphäre haben. Zeitgleich sind die Nutzer solcher Systeme stark darauf angewiesen, den Herstellern zu vertrauen.
Da sich insbesondere sprachgesteuerte Assistenzroboter stark für die Unterstützung älterer Menschen eignen, ist diesem Punkt große Bedeutung zu zumessen. Denn gerade ältere Menschen nutzen das Internet im Allgemeinen nur selten, wie die Initiative D21 in ihrem jährlichen Lagebild zur digitalen Gesellschaft 2018/19 erläutert [Ini19]. Entsprechend selten verwendet diese Nutzergruppe auch Sprachassistenzsysteme. Es ist daher notwendig, Bedenken bezüglich des Datenschutzes aus dem Weg zu räumen, um eine hohe Akzeptanz des Systems Sprachassistent-Assistenzroboter zu erzielen.
Darum sind digitale Assistenten, die ihren Fokus auf den Datenschutz legen, zu favorisieren. Einen solchen Schwerpunkt legene beispielsweise das OpenSource Projekt *Mycroft.ai* oder das teilweise quelloffene System *Snips.ai*. Im Rahmen dieser Arbeit sollen dabei mögliche Einsatzszenarien eines Sprachassistenten in Zusammenarbeit mit einem Assistenzroboters betrachtet werden. Zunächst soll die Architektur und Funktionsweise von Sprachassistenten untersucht werden. Auf dieser Grundlage wird dann ein Konzept erstellt und prototypisch umgesetzt. Abschließend wird mithilfe einer Pilotstudie die Praxistauglichkeit des Prototyps untersucht.

1.1 Ziele

Hauptziel dieser Arbeit ist es, ein Konzept zu entwickeln, das die Interaktion von Mensch und Roboter mittels natürlicher, gesprochener Sprache ermöglicht. Die Interaktion soll auch für Menschen mit beschränktem technischen Verständnis oder motorischen Einschränkungen möglich sein, so dass dieses Konzept universell verwendbar ist.
Dafür ist es nötig, eine Sprachassistenzsoftware zu identifizieren, die für diese Interaktion gut geeignet ist und zeitgleich den Schutz der Privatsphäre berücksichtigt. Außerdem bedarf es einer prototypischen Umsetzung des Konzepts, um dessen Funktionsfähigkeit nachzuweisen. Dieser Nachweis soll im Rahmen einer Nutzerstudie geschehen.

[2] https://www.golem.de/news/datenschutz-apple-hoert-durch-siri-drogengeschaefte-und-sex-mit-1907-142817.html [Abgerufen am 28.07.2019]

1.2 Aufbau der Arbeit

Grundlagen

Durch die Betrachtung der allgemeinen Architektur von Sprachassistenzsystem und Assistenzrobotern wird eine Grundlage für die weitere Arbeit gelegt. Außerdem wird die aktuelle Datenschutzsituation anhand der Datenschutzgrundverordnung (DSGVO) betrachtet und welche Angriffe auch die Privatsphäre mit Sprachassistenten möglich sind. Außerdem werden Einsatzszenarien sowie relevante Aspekte der Mensch-Roboter Interaktion betrachtet.

Ausgewählte Sprachassistenten

Im Anschluss an die Grundlagen ist es möglich, einzelne Sprachassistenten (Mycroft AI, Snips AI, Amazon Alexa) im Hinblick auf ihre Umsetzung der allgemeinen Architektur sowie Abwehrmaßnahmen von Angriffen auf die Privatsphäre zu betrachten. Außerdem werden solche Einsatzszenarien gewählt, bei denen ein gemeinsamer Einsatz eines Sprachassistenten und Assistenzroboters besonders lohnenswert erscheint. Diese Szenarien werden anschließend auf gemeinsame Grundfunktionen untersucht.

Einsatzkonzept

Für die Erstellung eines Einsatzkonzeptes ist es zunächst nötig, die sich aus den Einsatzszenarien sowie den weiteren zuvor angestellten Betrachtungen zu analysieren. Anhand dieser Kriterien ist es möglich, einen bestimmten Sprachassistenten auszuwählen, mit dem die Anforderungen bestmöglich erfüllt werden können. Auf der gleichen Basis können die spezifischen Bestandteile für die Verarbeitungspipeline gewählt werden. Auf Grundlage dieser Auswahl und den Anforderungen kann anschließend ein Konzept für die Zusammenarbeit eines Assistenzroboters mit einem Sprachassistenten im Rahmen der Mensch-Roboter-Interaktion erstellt werden.

Prototyp

Um das Konzept auf seine Funktionsfähigkeit in der Realität überprüfen zu können, ist die Umsetzung in Form eines Prototyps erforderlich. Daher wird in diesem Kapitel betrachtet, welche Besonderheiten in der Entwicklung von Anwendungen mit dem verwendeten Roboter und Sprachassistenten zu beachten sind. Außerdem wird die Funktionsweise des Prototyps und der darin umgesetzten Funktionen genauer betrachtet, was auch einen Blick auf die Architektur erfordert.

Evaluation

Um eine Bewertung des Konzepts mittels des Prototypen vorzunehmen, muss dieser von Versuchspersonen getestet werden. Die Tests werden in diesem Kapitel im

Rahmen einer Pilotstudie vorgenommen. Dafür wird beschrieben, wie die Fragen der Studie aufgebaut sind und wie sie im Speziellen abläuft. Auf Basis der Analyse der Antworten der Probanden ist es abschließend möglich, das zuvor erstellte Konzept zu bewerten.

Zusammenfassung

Zum Abschluss der Arbeit werden alle zuvor erlangten Erkenntnisse zusammengefasst und eingeordnet. Außerdem wird ein Ausblick auf weitere Entwicklungsmöglichkeiten gegeben.

2 Sprachassistenzsysteme und Assistenzroboter

Um eine Arbeitsbasis für den Aufbau Mensch-Roboter-Interaktion auf Basis natürlicher Sprache zu schaffe, ist es nötig, die einzelnen Bestandteile und eng damit verknüpfte Themen grundlegend zu betrachten. Dieses Kapitel gibt dafür zuerst einen Überblick über die prinzipielle Funktionsweise von Sprachassistenten. Da mit diesen Systemen häufig datenschutzrechtliche Bedenken verknüpft sind, wird die aktuelle rechtliche Lage genauer betrachtet und mögliche Risken, die Sprachassistenten in sich bergen.
Außerdem ist es nötig, zu betrachten inwiefern sich Assistenzroboter von Industrieroboter unterscheiden. Unterscheidungsmerkmale sind wirken sich dabei sowohl auf deren Aufbau sowie Einsatzmöglichkeiten aus. Um eine Interaktion von Mensch und Assistenzroboter zu ermöglichen, bedarf es auch einer Untersuchung relevanter Aspekte der Mensch-Roboter-Interaktion.

2.1 Allgemeine Architektur von Sprachassistenzsystemen

Die allgemeine Architektur von Sprachassistenten folgt gewissen Grundprinzipien, die von Edu et al. [Edu+19] sowie Sridhar et al. [ST17] beschrieben werden und denen die Erläuterungen in diesem Kapitel folgen. Dabei wird der Fokus der Analyse auf die Software gelegt. Zwar wird auch gewisse Hardware (Mikrofon, Lautsprecher, Analog-Digital-Wandler, Verarbeitungseinheit) benötigt, die Auswahl dieser Bestandteile kann jedoch durch den Anwender getroffen werden und hat nur bedingt Einfluss auf die Unterscheidung der spezifischen Systeme, da dieselbe Hardware mit unterschiedlicher Software genutzt werden kann. Die folgenden Erläuterungen orientieren sich an de Darstellung einer allgemeinen Verarbeitungspipeline von Sprachassistenten in Abbildung 2.1. Die in Bereich I der Abbildung 2.1 Aktionen stellen die Interaktion zwischen Nutzer und Assistenzsystem dar. Dafür wird ein `Client` mit entsprechender Hardware für Audioein- und Ausgabe benötigt. Damit der Nutzer mit dem System interagieren kann, muss zunächst ein sogenanntes `Aktivierungswort` (auch als `Signalwort`, `Weckwort` bezeichnet), gesagt werden. Der der Client mittels eines Mikrofons konstant den Umgebungsgeräuschen lauscht,

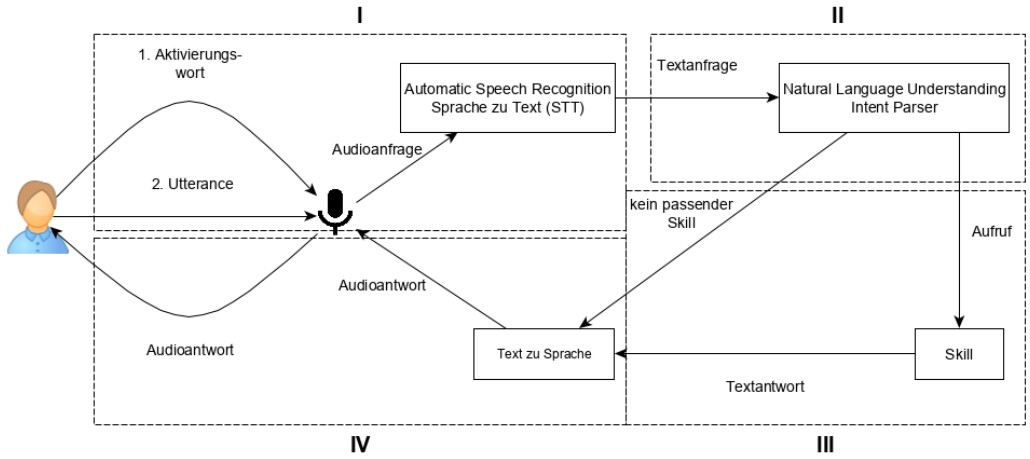

Abb. 2.1: Übersicht über die allgemeine Verarbeitungspipeline eines Sprachassistenten nach [Edu+19]

wird er dadurch aktiviert. Daraufhin beginnt die Aufzeichnung der Audiosignale der durch den Nutzer getätigten Aussage, die als `Phrase` bezeichnet wird. Daran schließt sich die Analyse dieser Signale an. [ST17]
Dafür ist eine Umwandlung der Signale in maschinenverständliche Sprache nötig, die mit einer Umwandlung der analogen in digitale Signale einhergeht, welche wiederum zur Weiterverarbeitung in Textform vorliegen müssten. Für diese Transformation kommt ein `Sprache-zu-Text (STT)` Umwandler zum Einsatz. Dieser nutzt die natürliche Sprachverarbeitung (NLP), wodurch zugleich auch Störgeräusche entfernt werden. Auch erlaubt NLP es, dass das System akzentbehaftete Sprache verstehen kann [Edu+19]. Die Filterung der störenden Geräusche wird durchgeführt, bevor es zu einer analog-digitalen Umwandlung kommt. Aus dieser digitalisierten Wellenform können durch Analyse von Frequenz und Tonhöhe die einzelnen Bestandteile (Features) herausgefiltert werden. Diese werden danach auf ein zuvor trainiertes Modell angewendet, um so die entsprechende Textrepräsentation zu ermitteln. Dazu wird immer der aktuelle Signalausschnitt mit dem vorherigen und folgenden betrachtet und in einem Lexikon der Wellenformen wird analysiert, was gesagt wurde. [Edu+19]
Dieser Prozess wird vereinfacht in Abbildung 2.2 dargestellt.

Im Abschnitt II der Abbildung 2.1 erfolgt die Verarbeitung der Anfrage auf Textbasis. Diese wird zuerst an den `Intent Parser` weitergeleitet. Der Parser untersucht den erhaltenen Text auf die gewünschte Aktion, also die Intention des Nutzers, die ausgeführt werden soll. In der einfachsten Umsetzung wird dann eine JSON Ausgabe generiert, die angibt, welcher `Intent` erkannt wurde. Für diese Erkennung definiert der Entwickler zuvor verschiedene Aussagen, die der Intention zugeordnet werden können, wodurch die gewünschte Aktion hervorgerufen wird. Außerdem beinhaltet die Ausgabe die Wahrscheinlichkeit (`confidence`) dieser Intention und

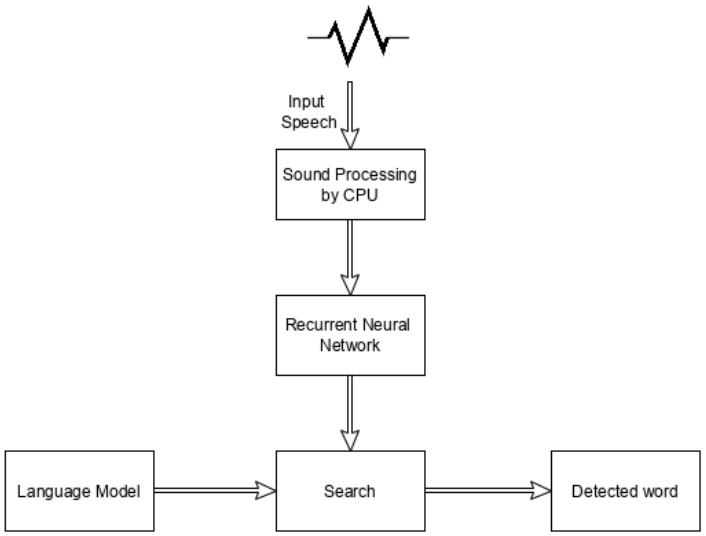

Abb. 2.2: Vereinfachte Darstellung der Umwandlung gesprochener Sprache in Text nach [Amb+18]

```
{
    "intent": "wetter",
    "confidence": 0.9577,
    "entity": "Dresden"
}
```

Abb. 2.3: Beispiel JSON für Anfrage „Wie ist das Wetter in Dresden?“

zugehörige Parameter. Für die Beispielanfrage „Wie ist das Wetter in Dresden?“ ist dies in Abbildung 2.3 dargestellt. Bei einer `entity` handelt es sich um eine Variable, die für die Durchführung der Handlung zwingend notwendig ist. Würde diese Information im dem Beispiel nicht mit geliefert, könnte dem Nutzer nicht die gewünschte Information geliefert werden. [1]

Der Bereich III der Abbildung 2.1 zeigt dabei die Abläufe, wenn die Zuordnung der Anfrage zu einer passenden Anwendung geschehen ist. Diese Anwendungen werden als `Skills` bezeichnet. Hierbei handelt es sich um Fähigkeiten, die das System besitzt, um Anfragen zu bewältigen. Ein Skill ist dabei ähnlich einer App auf einem Handy, das heißt, sie stellt eine Schnittstelle zu dem entsprechenden Dienst dar. Außerdem gibt es in der Regel einen sogenannten „Marktplatz“, auf dem diese Anwendungen angeboten werden. Daraus kann sich der Nutzer dann die gewünschten Anwendungen auswählen und seinem System hinzufügen. Es besteht auch die Möglichkeit, sich seine eigenen Skills zu definieren. [Edu+19]
Außerdem werden die Fähigkeiten in zwei Arten unterschieden. Zum einen sind dies die nativen Skills, welche direkt durch den Hersteller des Assistenzsystems

[1] https://mycroft-ai.gitbook.io/docs/mycroft-technologies/adapt

angeboten werden und mit dem System ausgeliefert werden. Außerdem gibt es Drittanbieterskills, die durch eigenständige Entwickler bereitgestellt werden und die unabhängig vom Systemhersteller agieren. [Edu+19]
Basiert die Analyse auf der zuvor erläuterten Zuweisung von Wahrscheinlichkeiten zu den möglichen Intents, dann wird nach der abgeschlossenen Analyse derjenige Skill ausgeführt, der die höchste Wahrscheinlichkeit aufweist. In dem in Abbildung 2.3 gezeigten Beispiel wäre dies der Wetter-Skill. Dabei versucht das System zuerst, einen nativen Skill auszuführen. Wenn es keinen solchen gibt, wird versucht, den eines Drittanbieters zu nutzen. Sollte es auch weiterhin keinen passenden Skill geben, wird der Nutzer darüber informiert [Edu+19].
Nach der Ausführung des Skills kommt es zunächst zu einer Antwort an den Aufrufer, die zu diesem Zeitpunkt in Textform vorliegt. Es kann jedoch sein, dass in der Antwort nach weiteren Informationen gefragt wird, die zur korrekten Durchführung des Befehls benötigt werden. Wenn mit dem Befehle Geräte kontrolliert werden sollen, werden die entsprechenden Anweisungen durch den Skill an diese weitergeleitet. Bei solchen Geräten kann es sich beispielsweise um verschiedene Teile einer Smart-Home-Umgebung handeln. [Edu+19]
Sofern es eine Antwort an den Nutzer geben soll, muss diese aus der Textform in gesprochene Sprache übersetzt werden. Dargestellt wird dies schematisch im Bereich IV der Abbildung 2.1. Dabei kommt ein `Text-zu-Sprache (TTS)` Umwandler zum Einsatz. Der Umwandler besteht neben einem Sprachmodell auch aus Lautdefinitionen inklusive einer Verslehre, damit es zu einer akustischen Sprachausgabe kommt. Dafür wird der Text in Tokens unterteilt und mittels Textanalyse zu einer Audioausgabe synthetisiert [ST17]. In der Regel hat der Nutzer die Möglichkeit, zwischen verschiedenen Stimmen zu wählen.

2.2 Betrachtungen zum Datenschutz

Im Zusammenhang mit Sprachassistenten stellen sich auch immer datenschutzrechtliche Fragen, da theoretisch ein ununterbrochenes Belauschen der Nutzer möglich ist. Damit diese Fragen beantwortet werden können, ist es nötig, zuerst die aktuelle Rechtslage zu betrachten. Über diese kann im Rahmen dieser Arbeit aufgrund ihrer Komplexität nur ein Überblick gegeben werden. Abschließend kann sich potentiellen Gefährdungen des rechtlichen Rahmens sowie möglichen Angriffspunkten auf die Privatsphäre zugewendet werden.

2.2.1 Relevante aktuelle Rechtslage

Die aktuell bedeutendste rechtliche Grundlage für den Datenschutz innerhalb der Europäischen Union (EU) ist die seit dem 25. Mai 2018 geltende Datenschutz-

grundverordnung (DSGVO). Deren wesentliches Ziel ist die Anpassung der Datenschutzgesetze an die Gegebenheiten des Internetzeitalters, wobei auch Wert auf eine technologieneutrale Formulierung der Bestimmungen gelegt wurde. [Par16]
Aus der Perspektive von Sprachassistenten sind die folgenden Artikel die relevantesten:

- Artikel 3 „: Räumlicher Anwendungsbereich“
- Artikel 5 „: Grundsätze für die Verarbeitung personenbezogener Daten“
- Artikel 17 „: Recht auf Löschung “
- Artikel 25 „: Datenschutz durch Technikgestaltung und durch datenschutzfreundliche Voreinstellungen“

In Artikel 3 wird das Prinzip des Marktortes eingeführt. Das bedeutet, dass alle Regeln der DSGVO für alle Unternehmen gelten, die in der EU geschäftlich aktiv sind. Dabei ist es irrelevant, an welchem Ort sich der Unternehmenssitz befindet. [Dat17b]
Artikel 5 legt fest, dass Daten nur sparsam erhoben werden dürfen („Datensparsamkeit“) und auch nur für zuvor festgelegte Zwecke („Zweckbindung“) [Par16]. Diese Daten sind nach Artikel 17 nach der Erfüllung der Aufgabe zu löschen, außerdem kann jederzeit eine Löschung durch die betroffene Person verlangt werden [Dat17a].
Durch Artikel 25 wird festgelegt, dass der Datenschutz bereits bei der Entwicklung („Privacy-by-Design“) berücksichtigt werden muss. Außerdem sollen standardmäßig solche Einstellungen aktiviert sein, die die Privatsphäre schützen („Privacy-by-Default“). [Par16]

2.2.2 Potentielle rechtliche Konflikte beim Einsatz von Sprachassistenten

Die folgenden möglichen rechtlichen Konflikte beziehen sich auf die im Abschnitt 2.2.1 erläuterten Aspekte der DSGVO. Da alle drei betrachteten Systeme offiziell in Deutschland angeboten werden, gelten die in Kapitel 2.2.1 betrachteten Gesetze aufgrund des Artikels 3 der DSGVO auch für diese Sprachassistenten.
Konflikte entstehen vor allem bei einer Speicherung der Daten nach der Abarbeitung der Anfrage, da der vom Nutzer benötigte Zweck erfüllt ist. Auch stellt sich die Frage, ob die standardmäßige Speicherung der Daten dem Prinzip „Privacy-by-Default“ entspricht. Jedoch haben Furey und Blue [FB18a] festgestellt, dass es unter anderem die weit gefasste Regelung der „Verbesserung der Nutzererfahrung“ gibt. Unter anderem ließe sich die Nutzung von Daten für das Training der Sprachassistenten sowie auch darauf basierende personalisierte Werbung auf dieser Art klassifizieren. [FB18a]
Es kann auch aktuell kein klares Urteil gefällt werden, ob die Datenverarbeitun-

gen durch Sprachassistenten rechtmäßig ist. So ist der wissenschaftliche Dienst in einer Betrachtung zu dem Thema, ob die Transkribierung und Auswertung von Mitschnitten durch Alexa rechtmäßig ist, zu keinem klaren Urteil gekommen. Es wird hervorgehoben, dass man davon ausgehen kann, dass Amazon seinen Pflichten zur Informationsvermittlung nachgekommen ist. Gleichzeitig wird aber die Frage gestellt, auf welche Weise ausgeschlossen werden kann, dass Daten von unbeteiligten Dritten sowie Minderjährigen gesammelt werden. [DB19]
Von Storr und Storr wurde außerdem herausgefunden, dass die in der Cloud befindlichen Daten durch die Anbieter für Analysen mittels künstlicher Intelligenz benutzt werden und zum Teil auch an andere Anbieter weiterverkauft werden [SS17].

2.2.3 Mögliche Bedrohungen durch Dritte

Die Bedrohungen der Privatsphäre durch Dritte sollen im Folgenden anhand eines Szenarios betrachtet werden. Dieses gestaltet sich folgendermaßen: eine Nutzerin, Alice, fragt einen Sprachassistenten nach dem Wetter. Dafür nutzt der Assistent den Skill eines Drittanbieters, namens Bob, welcher auch andere Anwendungen anbietet. Damit die Anfrage verarbeitet werden kann, ist eine Verbindung zur Cloud nötig. Außerdem besteht eine Verbindung des Assistenzsystems mit SmartHome-Geräten in der Wohnung von Alice.
Für ein solches Szenario werden durch Chung et al. [Chu+17] Angriffsmöglichkeiten geschildert, die von Jacksong et al. [JO18] erweitert werden. Dargestellt sind diese in Abbildung 2.4 und entsprechend der folgenden Beschreibung mit einer Kennzahl versehen.

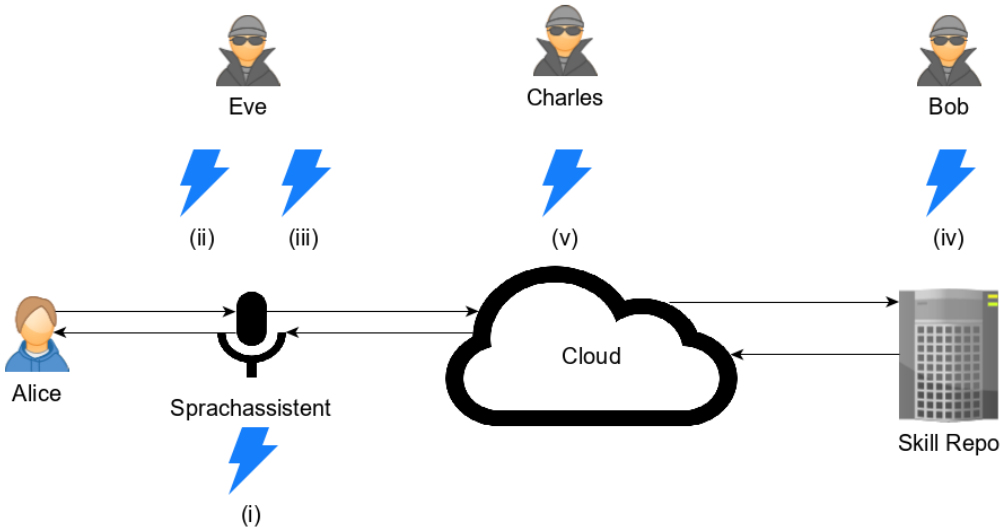

Abb. 2.4: mögliche Angriffspunkte beim Datenaustausch mit der Cloud

(i) ungewollte Aufnahme von Geräuschen Zu einer ungewollten Aufnahme von Geräuschen kann es beispielsweise dann kommen, wenn Alice mit einem Freund namens Alexander redet und ihr Sprachassistent, wie in dem von der Verbraucherzentrale NRW untersuchten Fall, eine Alexa ist. Äußert sich Alice mit der Aussage: „Ich denke, dass Alexander das weiß“, kann dies durch das Assistenzsystem zur fehlerhaften Annahme kommen, dass das Aktivierungswort „Alexa“ geäußert wurde. Die Verbraucherzentrale Nordrhein-Westfalen hat die Reaktion von Alexa auf die vier standardmäßig verfügbaren Aktivierungsworte („Alexa“, „Amazon“, „Echo“, „Computer“) sowohl in abgewandelter Form, als auch innerhalb eines Satzes betrachtet. Mit Verwendung des Signalwortes innerhalb eines Satzes kam es in der Hälfte der Fälle zu einer Aktivierung des Assistenten. Dieser startet im Anschluss eine, durch den Nutzer ungewollte, Aufzeichnung und Auswertung der anschließend getroffenen Aussagen. [e.V17]
Außerdem kommt es hier möglicherweise zu einem Konflikt mit der DSGVO da die Einwilligung zur Datenverarbeitung nach Artikel 3 in diesem Moment diskutabel ist.

(ii) Manipulation des Engerätes Während sich Alice nicht in ihrer Wohnung befindet, versucht die Angreiferin Eve Zugang zur Wohnung zu erhalten. Da Alice unter anderem ihre Wohnungstür über den Sprachassistenten steuert, will Eve diesen dafür manipulieren [2]. Dazu hat sie zwei Möglichkeiten. Zum einen kann sie einen entsprechenden Befehl durch ein nicht vollständig geschlossenes Fenster rufen [Haa+17]. Zum anderen kann sie einen sogenannten „Delfinangriff“ durchführen, wie er von Roy et al. geschildert wird. Dabei werden hochfrequente Töne verschickt, die für Menschen nicht hörbar sind, aber trotzdem von den Assistenzsystemen verarbeitet werden. Getestet wurde das mit gängigen Systemen wie Alexa, Cortana, Siri und führte zumindest im Umkreis von 1,5m zum Erfolg. [Roy+18]

(iii) Abhören der Daten Eve kann auch versuchen, Rückschlüsse aus den Daten zu ziehen, die an die Cloud verschickt werden. Dabei kann sie Nachrichten mitlesen, die unverschlüsselt versendet werden. Aber auch aus verschlüsselten Nachrichten ist es möglich, Rückschlüsse über das Nutzungsverhalten zu ziehen. Beispielsweise kann sie so herausfinden, wann normalerweise Personen im Haushalt anwesend sind, um einen Angriff wie (ii) durchzuführen. [Apt+17]

(iv) Manipulation der Skills Auch der Anwedungsanbieter Bob kann Angriffe auf das System von Alice tätigen. Diese Möglichkeit hat er durch die Manipulation von Skills. Er kann seine Anwendung so gestalten, dass sie zwar nach außen hin vertrauenswürdig wirkt, aber für die Ausführung des Skills ein weiterer, manipulierter Skill aufgerufen wird, ohne das Alice dies merkt. Eine andere Option besteht darin,

[2] https://opensource.com/article/19/2/mycroft-voice-assistant [Abgerufen am 07.06.2019]

mehr Daten anzufordern, als für die erfolgreiche Durchführung der Anfrage nötig sind. Möglicherweise gibt Alice die Einwilligung dazu, da sie im Installationsprozess unaufmerksam war. [Edu+19]
Da Bob mehrere Skills anbietet, kann er auch die Daten der verschiedenen Skills aggregieren und somit ein umfassendes Bild von Alice erhalten. Ein sehr ähnliches Szenario wird von Memon und Anwar am Beispiel von Smartphone Apps beschrieben. [MA15]

(v) Zugriff auf Daten in der Cloud Für den Hacker Charles ergibt sich auch eine Angriffsmöglichkeit, in dem er auf die in der Cloud abgelegten Daten zugreift. Dies ist ein lohnenswertes Ziel, da er nur Zugriff auf dieses eine Element benötigt, um an viele sensible Daten gelangen zu können. Für einen solchen Angriff hat er dabei verschiedene Alternativen zur Durchführung, da ein Zugriff auf die Cloud auf unterschiedlichen Wegen möglich ist (z.B. per App oder Web-Access) [Mod+13].
Ein solcher Angriff ist besonders lohnenswert, da sich die Daten durch ihre Vielfältigkeit auszeichnen. Es ist durchaus möglich, dass ein Nutzer weitere Geräte (z.B. Fitnesstracker) mit dem Sprachassistenten verbindet, welcher alle Daten dann zentral ablegt. Außerdem entsteht durch Verknüpfung dieser Daten, sogenanntes „daisy chaining“, ein umfassendes Bild des Nutzers. Daran sind auch die Hersteller verschiedener Systeme interessiert, um ihre Produkte und Werbung besser an den Nutzer anzupassen. [FB18b]

2.3 Grundlagen des Einsatzes von Assistenzrobotern

> *A service robot is a robot which operates semi or fully autonomously to perform services useful to the wellbeing of humans and equipment, excluding manufacturing operations.*
>
> **— International Federation of Robotics**
> [Kar00]

Für eine genauere Untersuchung von Assistenzrobotern ist es zunächst nötig zu betrachten, welche Roboter als solche klassifiziert werden können und wie sie sich von anderen Robotern unterscheiden. Für die Eingrenzung bedient sich diese Arbeit an der Definition von Assistenzrobotern durch die „International Federation of Robotics“. Diese besagt, dass Assistenzroboter (teil-)autonom Aufgaben erfüllen, die Menschen helfen. Allerdings handelt es sich dabei nicht um Produktionsaufgaben [Kar00]. Dem gegenüber stehen die Industrieroboter, deren Hauptaufgabe in der

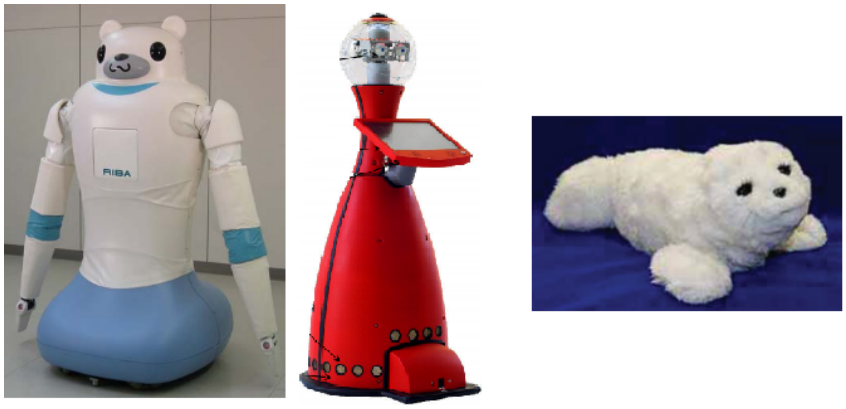

Abb. 2.5: Roboter RIBA [Muk+10], TOOMAS [Gro+09], PARO [Cal+11]

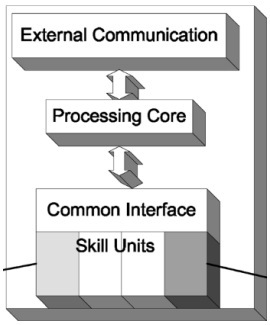

Abb. 2.6: Basisarchitekur eines Assistenzroboters aus [Gal+06]

Herstellung von Produkten oder der Unterstützung des Herstellungsprozesses liegt [Kum+05].
In diesem Zusammenhang ist es sinnvoll zu untersuchen, welche Architekturmerkmale Assistenzroboter gemeinsam haben, sowohl unter Hard- als Softwareaspakten. Außerdem kann basierend auf der Eingangs erwähnten Definition untersucht werden, welche Einsatzszenarien es für Assistenzroboter gibt.

2.3.1 Aufbau von Assistenzrobotern

In Abbildung 2.5 werden verschiedene physischen Erscheinungen der Assistenzroboter anhand von Beispielen veranschaulicht. Dabei ist klar ersichtlich, dass das Aussehen der Roboter an ihre jeweiligen Aufgaben angepasst wurde.
Beispielsweise zeichnet sich RIBA [Muk+10] durch ein sehr menschenähnliches Erscheinungsbild aus. Für den Einsatz zum Heben von Patienten sind mit den Armen wichtige Voraussetzungen getroffen. Für den Navigationsroboter TOOMAS [Gro+09] hingegen reicht eine abstrakte Menschenähnlichkeit, da er auf die Bewegung im Raum spezialisiert ist und somit vollständig auf Arme verzichten kann.
Dem gegenüber steht mit PARO [Wad+04] ein Roboter, dessen Äußeres sehr stark dem einer Robbe ähnelt. Dadurch wird er von den Patienten als Lebewesen angesehen und erleichtert beziehungsweise fördert die Interaktionen.

Diese Beispiele veranschaulichen, dass es nur schwer möglich ist, eine allgemeine Aussage über die Architektur von Assistenzrobotern zu treffen. Wie in Abbildung 2.6 ersichtlich, besitzen Assistenzroboter drei verschiedene Basisverarbeitungseinheiten. Mit einer kann der Roboter mit der Außenwelt kommunizieren. Dies umfasst sowohl Ein- als auch Ausgabe auf physischer, schriftlicher oder sprachlicher Basis. Außerdem gibt es eine Verarbeitungseinheit, die sowohl auf die Eingaben reagiert als auch die Ausgaben des dritten Bestandteils. Dieser dient dem Aufruf der verschiedenen Fähigkeiten des Systems sowie der Rückgabe von Resultaten. Die Architektur wurde aus dem von Galindo et al. [Gal+06] im Zusammenhang mit einem Roboterrollstuhl beschriebenen Konzept entnommen.

2.3.2 Einsatzszenarien für Assistenzroboter

Für die Analyse von möglichen Einsatzgebieten für Assistenzroboter spielt es keine Rolle, ob bestimmte Modelle aufgrund ihrer physischen Erscheinung besser oder schlechter für bestimmte Aufgaben geeignet sind. Diese Analyse basiert alleine auf allgemein möglichen Einsatzfeldern, die Architektur der einzelnen Modelle orientiert sich dann wiederum an ihren angedachten Aufgaben. [Böh02]
Eine häufige Gemeinsamkeit verschiedener Einsatzszenarien ist der Bereich der Pflege. Dort können Roboter zur Unterstützung von pflegebdürftigen Personen eingesetzt werden. Einige solcher Aufgaben werden von Hans et al. [Han+02] beschrieben. So kann ein solcher Roboter im Haushalt zum Einsatz kommen, um einer Person Gegenstände - beispielsweise Getränke - zu bringen. Gerade bei der Unterstützung von bettgebundenen Personen wird dies als sinnvoll erachtet. Ebenfalls sind dabei Steuerungstätigkeiten von Systemen wie Heizung, Klimaanlage, Belüftung vorstellbar.
Auch gehbehinderte Menschen können von Assistenzrobotern unterstützt werden. In diesem Zusammenhang sind die Funktionen als Lauf- sowie Orientierungshilfe von großer Relevanz. [Han+02]
Außerdem wird die Unterstützung auf Kommunikationsebene als sinnvolles Einsatzgebiet gesehen. So kann der Roboter bei der Kommunikation eines Patienten mit, z.B. Ärzten, Pflegern oder Angehörigen, mittels Telepräsenz helfen. Dies können Verwandte sein, aber auch Ärzte oder Pfleger. Auch kann mit Hilfe des Roboters eine Verwaltung der Medieninfrastrutur (z.B. Fernseher, Telefon) vorgenommen werden. [Han+02]
Es ist auch möglich, die Roboter zur Unterstützung von Menschen einzusetzen, die an Demenz erkrankt sind. Für einen solchen Einsatz wurde beispielsweise der robbenähnliche Roboter `PARO` entwickelt [Cha+13]. Dieser hilft den Patienten, soziale Interaktionen zu trainieren, indem er auf Berührungen und Geräusche reagiert und damit physische Nähe erzeugt. Dadurch animiert er die Patienten, wieder vermehrt mit anderen Menschen zu kommunizieren.

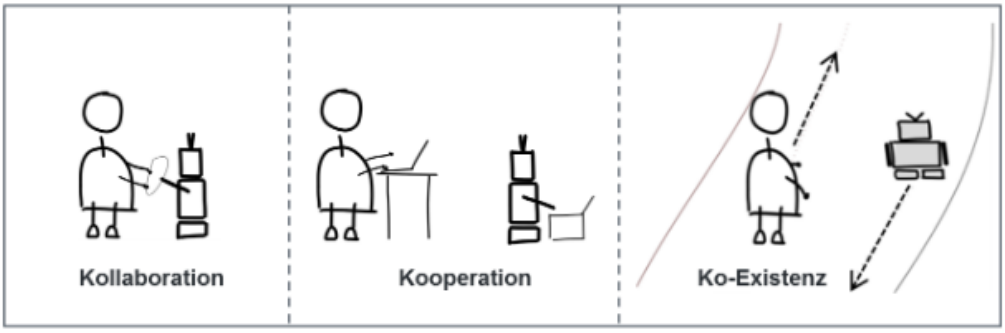

Abb. 2.7: Arten der Interaktion zwischen Mensch und Roboter aus [Onn+16]

Auch eine direkte Entlastung des Pflegepersonals kann der Roboter umsetzen, in dem er Patienten zum Beispiel an Routineaufgaben, wie Medikamenteneinahme, Flüssigkeitszufuhr oder Nahrungsausfnahme erinnert [Pol+02]. Des Weiteren kann der Roboter auch die pflegebedürftige Person auf Lebensfunktionen überwachen und bei Bedarf einen Notruf tätigen [Han+02].
Eine weitere Unterstützungsmöglichkeit in diesem Bereich umfasst Bewegunsabläufe. So wird von Hayashi et al. [Hay+05] ein tragbares System vorgestellt, dass sie als „Robot Suit“ bezeichnen. Es wird am Körper getragen und unterstützt die Beinbewegungen des Trägers. Ein anderer Einsatz wird von Mukai et al. [Muk+10] präsentiert. Dabei übernimmt ein selbständiger Roboter Hubbewegungen. Damit kann ein Patient beispielsweise aus einem Rollstuhl in ein Bett gehoben werden.
Neben der Pflege gibt es auch noch Einsatzmöglichkeiten auf dem Gebiet der Information und Navigation in öffentlichen Umgebungen (z.B. Museen, Supermärkte). Beispielsweise wurde von Böhme mit SCITOS ein Roboter entwickelt, der Kunden in einem Baumarkt den Weg zu den gesuchten Produkten zeigt [Böh02]. Der Roboter TOOMAS ist dabei die Spezialiserung dieses Robotermodells für den Einsatz im Baumarkt der Kette Toom [Gro+09].
Von Buhman et al. [Buh+95] wurde ein Roboter mit ähnlichem Zweck eingeführt. Dieser agiert als Museumsführer und erkennt für diesen Zweck automatisch Personen, die er dann auf einer Tour durch das Museum führen kann.
Andere Einsatzszenarien spielen sich wiederum im Haushalt ab. Dabei wird ein Roboter als Reinigungshelfer eingesetzt. Die Tätigkeiten umfassen dabei beispielsweise Mäh-, Saug- oder Poolreinigungsfunktionen [Kum+05]. Ähnliche Reinigungsfunktionen finden sich aber auch in Robotern wieder, die als allgemeine Haushaltshilfe eingesetzt werden können [Yam+12].

2.4 Betrachtung relevanter Aspekte der Mensch-Roboter-Interaktion

Um relevante Aspekte der Mensch-Roboter-Interaktion betrachten zu können, ist es zunächst nötig, diese allgemein zu betrachten. Anschließend daran ist es möglich, auf Basis der in Kapitel 2.3.2 erläuterten Einsatzszenarien von Assistenzrobotern relevante Aspekte zu betrachten.
Als Veranschaulichung der folgenden Betrachtungen dient Abbildung 2.7, welche verschiedenen Möglichkeiten es für die Interaktion von Mensch und Roboter darstellt. Damit beide gemeinsam ein Ziel erreichen, können sie sowohl kollaborieren als auch kooperieren. Im ersten Fall sind ihre Tätigkeiten direkt voneinander abhängig, das heißt, dass mindestens ein Partner auf den anderen angewiesen ist, um die eigenen Aufgabe erfolgreich abschließen zu können. Im Falle der Kooperation besteht keine solche direkte Abhängigkeit zwischen den beiden Teilaufgaben. Beide Partner können ihre Teilaufgaben unabhängig von dem anderen durchführen, jedoch kann die Gesamtaufgabe nur gemeinsam erfüllt werden. Wenn beide Partner eigene Aufgaben haben, die in keiner Beziehung zueinander stehen, aber sich in der gleichen Umgebung abspielen, handelt es sich um eine Ko-Existenz. Dabei muss bei den eigenen Aktionen auf die des anderen Rücksicht genommen werden. [Onn+16]
Für die in Kapitel 2.3.2 erarbeiteten Szenarien sind alle diese Arten der Interaktion von Relevanz. Eine Kollaboration liegt beispielsweise dann vor, wenn der Roboter die Aufgabe hat, einen Gegenstand zu besorgen. Wird dieser beispielsweise von einer Pflegeperson für die erfolgreiche Versorgung eines Pflegebedürftigen benötigt, so kann er die Aufgabe erst erfolgreich abschließen, nachdem er von dem Roboter diesen Gegenstand erhalten hat. Eine Kooperation zwischen Roboter und Mensch ist zum Beispiel dann der Fall, wenn dem Roboter die Aufgabe der Erinnerung an Routinetätigkeiten zukommt. In diesem Fall ist die Gesamtaufgabe die optimale Unterstützung des Pflegebedürftigen, allerdings hängen die Tätigkeiten des Pflegers nicht direkt von denen des Roboters ab. Aber es kommt auch zur Ko-Existenz. Es kann angenommen werden, dass in einer Umgebung, in der Personen gepflegt werden, nicht nur ein Pfleger und ein Patient existieren, sondern jeweils mehrere. Unterstützt der Roboter mit seinen Tätigkeiten Pfleger A, indem er einen Gegenstand besorgt, dann steht seine Aufgabe nicht in Beziehung zu denen von Pfleger B. Trotzdem muss dieser in den Handlungen des Roboters berücksichtigt werden.
Damit die Interaktion zwischen Roboter und Mensch problemlos ablaufen kann, sind zwei Faktoren von großer Bedeutung. Zum einen ist dies die Art der Kommunikation. Zum anderen ist auch das Verhalten des Roboters wichtig. Gerade dadurch kann Nutzerakzeptanz erzeugt werden, die wiederum nötig ist, damit auf die Unterstützung des Roboters zurückgegriffen wird.

Arten der Mensch-Roboter Kommunikation Für die Kommunikation zwischen Roboter und Mensch bieten sich prinzipiell verschiedene Wege an. So kann diese komplett auf grafischer Ebene stattfinden, beispielsweise mit einem berührungsempfindlichen Display. Über dieses können Informationen ein- sowie ausgegeben werden. Allerdings birgt eine solche Kommunikationsart auch Probleme in sich. So haben Personen mit motorischen Einschränkungen [LS14] und Menschen mit begrenzten Sehfähigkeiten [Zen+18] Probleme mit der Nutzung von Displays für die Ein- und Ausgabe.
Alternativ kann die Kommunikation auch auf Basis von Gesten stattfinden. Dazu muss der Roboter in der Lage sein, Menschen visuell zu erkennen und aus ihren Gesten die richtigen Schlüsse zu ziehen. Als Geste sind dabei die Bewegungen der Hände und des Kopfes zu betrachten [Böh02]. Damit die Gestenerkennung korrekt abläuft, werden von Waldherr et al. [Wal+00] zwei verschiedene Ansätze vorgeschlagen. Zum Einen ist es möglich, Gestentemplates zu definieren. Diese werden mit der aktuellen Armhaltung verglichen und auf Basis von Wahrscheinlichkeiten zugeordnet. Zum Anderen wird der Einsatz eines neuronalen Netzes ins Spiel gebracht. Damit können beispielsweise die Winkel zwischen Ober- und Unterarm bestimmt werden. Gleichzeitig stellen die Autoren aber fest, dass eine rein gestenbasierte Kommunikation Schwierigkeiten mit sich bringt und besser als Unterstützung von Sprache dienen kann. [Wal+00]
Eines dieser Probleme ist, dass zwangsläufig eine Sichtverbindung zwischen der interagierenden Person und dem Roboter bestehen muss. Die Gesten müssen dabei eindeutig sein, damit sie klar voneinander unterscheidbar sind. Dafür ist es notwendig, dass der Roboter nah genug bei der entsprechenden Person ist. Da Gesten bestimmte Bewegungen verschiedener Gliedmaßen voraussetzen, ist diese Art der Steuerung für Menschen mit motorischen Einschränkungen ungeeignet. [Wal+00]
Vorstellbar ist auch, dass der Roboter anstelle von Gesten die Verhaltensweisen des Menschen analysiert und daraus Rückschlüsse auf das Wohlbefinden des Pflegebedürftigen zieht. Dafür werden aber zusätzliche Sensordaten benötigt, um beispielsweise das Schlafverhalten analysieren zu können [Cor+13].
Von Prodanov et al. [Pro+02] wird die Kommunikation auf Basis von natürlicher, gesprochener Sprache vorgestellt. Diese wird von den Autoren als die nutzerfreundlichste Art der Kommunikation bezeichnet. In ihrem Fall wird sie für einen Roboter eingesetzt, der Besucher durch ein Museum führt. Sowohl Prodanov et al. als auch Böhme heben dabei hervor, dass sich die Interaktion mit gesprochener Sprache besonders dafür eignet, Personen mit geringen technischen Vorkenntnissen zur Nutzung eines solchen Systems zu befähigen [Pro+02; Böh02]. Allerdings muss der Roboter auf ein sehr umfangreiches Vokabular zurückgreifen, um mit den unterschiedlichsten Nutzern interagieren zu können. Erst dieses Vokabular erlaubt es dem Roboter, entsprechende Befehle aus den Wörtern abzuleiten, ohne dass der Nutzer diese wortgetreu ausspricht. [Böh02]
Da Kommunikation nicht nur in eine Richtung funktioniert, ist es außerdem wichtig,

dass der Roboter dem Nutzer Rückmeldungen gibt. Diese kann er, in Abhängigkeit der verfügbaren Hardware, durch akustische Signale, Ausgaben auf dem Display oder aber Bewegungen geben [Loi+15]. Gerade durch eine Antwort des Roboters durch gesprochene Sprache bleibt für den Nutzer ein Gefühl der Natürlichkeit erhalten [Böh02].
Wichtig ist hierbei auch, dass die Spracherkennung mit einer ausreichenden Genauigkeit durchgeführt wird [Wan+16]. Damit der Nutzer die Möglichkeit hat, bei möglicherweise falsch verstandenen Befehlen einzugreifen, bevor diese ausgeführt werden, wird durch Green et al. [Gre+00] ein Antwortverhalten vorgestellt. Dieses kombiniert Gesten und Sprache. Dargestellt ist es in Abbildung 2.8. Dabei fragt der Roboter bei dem Nutzer nach, ob die Anweisung korrekt verstanden wurde und unterstreicht dies mit einer Geste. Erst nach Antwort des Nutzers wird die Aktion ausgeführt und dabei der Befehl noch einmal wiederholt.
Eine Nutzerbefragung von Green et al. [Gre+00] zu diesem Thema hat unterstreicht die Beudeutung der Interaktion auf Basis von gesprochener, natürlicher Sprache. So haben 82% der Befragten geantwortet, dass sie für die Interaktion mit einem Roboter natürliche Sprache bevorzugen. Lediglich 52% empfinden demnach die Gestensteuerung als praktikablen Weg [Gre+00].

```
Nutzer:  Roboter, hole Kaffee aus der Küche!
Roboter: Kaffee aus der Küche holen?
# Roboter performt Geste
Nutzer:  Ja, bitte
Roboter: Hole Kaffee aus der Küche
```

Abb. 2.8: Dialog zwischen Mensch und Roboter mit natürlicher Sprache aus [Gre+00]

Verhalten des Roboters Ein nicht zu vernachlässigender Faktor der Interaktion ist das Verhalten des Roboters. Damit sind vor allem die Bewegungen des Roboters gemeint. So hat Lohse [Loh07] herausgefunden, dass die Geschwindigkeit der Bewegungen einen maßgeblichen Einfluss auf die Akzeptanz des Roboters haben. So wird eine zu große Geschwindigkeit häufiger als aggressiv empfunden. Besser ist eine Anpassung an die Bedürfnisse des Nutzers. [Loh07]
Auch das Sozialverhalten ist ein wichtiger Aspekt. Hierbei handelt es sich um die Einhaltung von allgemeinen Normen, zum Beispiel die Respektierung des persönlichen Raumes. Auch sollte sich der Nutzer bewusst sein, dass der Roboter seine Gegenwart erkennt und seine Reaktionen automatisch entsprechend anpasst. Andernfalls kann dies gerade bei Menschen, die sich um Umgang mit derartigen Maschinen unsicher fühlen, diese Unsicherheit noch maßgeblich verstärken. [Pac+05].

2.5 Zusammenfassung

Ein Schwerpunkt in diesem Kapitel stellt die allgemeine Architektur von Sprachassistenten dar, wobei die softwareseitige Umsetzung der Verarbeitung von besonderem Interesse ist. Um ein Interaktion zu starten, benötigt der Nutzer ein Signalwort, mit dem er den Assistenten aktiviert. Danach lässt er einen Befehl folgen, der durch das System verarbeitet wird und eine entsprechende, zuvor installierte, Fähigkeit aufruft. Daraufhin erhält der Nutzer im Normalfall eine akustische Antwort. Mit der Nutzung von Sprachassistenten werden auch zusätzliche Angriffspunkte auf die Privatsphäre des Nutzers geschaffen. Dabei kann der Nutzer sowohl ausspioniert werden, als auch mittels der Systeme physischer Zugang zur Wohnung des Nutzers geschaffen werden.
Des weiteren wurde eine Definition von Assistenzrobotern eingeführt, die sich dadurch von Industrierobotern unterscheiden, in dem sie keine Unterstützung in Herstellungsprozessen bieten. Vielmehr unterstützen sie im Bereich der Pflege von Menschen, wobei sich ihr Aussehen stark an den zu erfüllenden Aufgaben orientiert. Außerdem sollte der Roboter in seinem Verhalten gewissen Regeln folgen, damit sich Nutzer bei der Bedienung des Geräts sicher fühlen.

3 Einsatz von Assistenzrobotern mit Sprachassistenzsytemen

Für den gemeinsamen Einsatz von Assistenzrobotern und Sprachassistenten ist es nötig, die im vorherigen Kapitel beschriebenen Einsatzszenarien im Hinblick auf ihre Gemeinsamkeiten in den Grundfunktionen zu betrachten. Außerdem bedarf es einer Untersuchung verschiedener Sprachassistenzsysteme im Hinblick auf ihre Umsetzung der einzelnen Verarbeitungsschritte. Zusätzlich können die verschiedenen Systeme auf ihre Möglichkeiten zur Abwehr von Angriffen auf die Privatsphäre untersucht werden.

3.1 Mögliche Sprachassistenzsysteme

Da es das Ziel dieser Arbeit ist, ein Konzept für zur unabhängigen natürlich-sprachlichen Mensch-Roboter-Interaktion zu erstellen, kommen nur solche Sprachassistenzsysteme in Frage, welche die Datenverarbeitung mittels selbstgewählter Infrastruktur erlauben. Dadurch behält der Nutzer auch die Hoheit über die eigenen Daten, was wiederum unter Datenschutzaspekten ein sehr wichtiger Faktor ist.
In diesem Zusammenhang sind zwei verschiedene Projekte von Interesse. Zum Einen *Mycroft AI*, ein Sprachassistent dessen Quellcode komplett offen gelegt ist. Ein anderes Projekt ist *Snips AI*, dessen Kernfunktionen frei verfügbar sind und das den Fokus vorrangig auf Datenschutz legt. Gerade der Open-Source Charakter ist wichtig, damit zur Verbesserung der Basisfunktionen jederzeit manuelle Anpassungen vorgenommen werden können. Um die beiden System einordnen zu können, wird ein Vergleich mit dem aktuellen Marktführer im Bereich Sprachassistenten, Amazon Alexa (siehe Abbildung 3.1), vorgenommen.

Mycroft AI Bei Mycroft AI (kurz Mycroft) handelt es sich um einen Open-Source Sprachassistenten, der 2015 erstmals veröffentlicht wurde. Er wurde zunächst als komplettes Gerät inklusive Lautsprecher und Mikrofon über Crowdfunding finanziert. Drei Jahre später wurde ein Nachfolgemodell erneut auf dem gleichen Weg

[1] https://www.handelsblatt.com/unternehmen/it-medien/elektronikmesse-ifa-siri-alexa-und-google-home-wie-sprachassistenten-die-technikwelt-veraendern/22971046.html [Abgerufen am 11.05.2019]

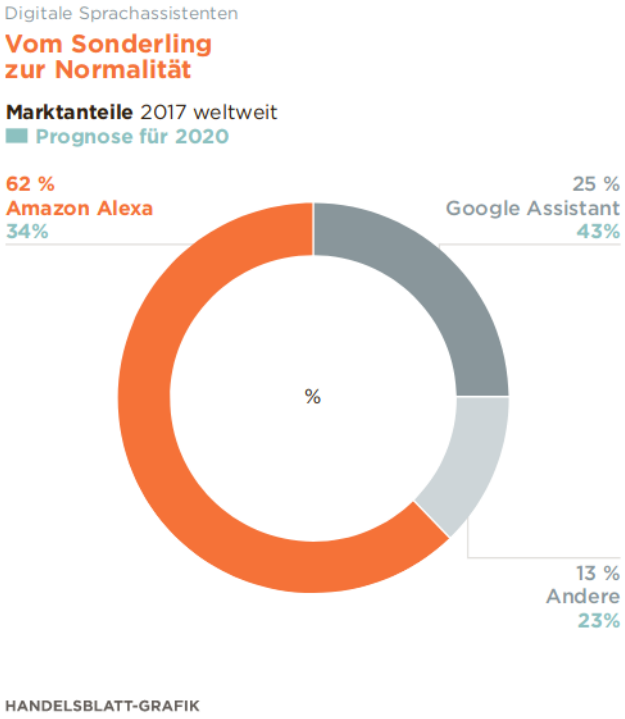

Abb. 3.1: Marktanteile der Sprachassistenten 2017 [1]

finanziert[2]. Zeitgleich mit der Markteinführung des ersten Geräts wurde auch die Software des Systems unter der Apache 2.0 Lizenz frei zur Verfügung gestellt[3]. Diese wird direkt in Form eines eigenen Betriebssystems, Picroft, zur Verfügung gestellt. Hierbei handelt es sich um einen Ableger von Raspbian. Es ist allerdings auch möglich, die Software mit anderen Linux Derivaten zu betreiben, wozu allerdings die manuelle Installation weiterer Pakete benötigt wird. Außerdem wird durch den Hersteller eine Android Anwendung zur Verfügung gestellt. Die Betriebssysteme Windows und MacOS werden mit Stand September 2019 noch nicht unterstützt. [4] Mycroft zeichnet sich insgesamt besonders durch seine Modularität aus. Somit kann sich der Nutzer für jeden Schritt der Verarbeitung zwischen verschiedener Software entscheiden. Diese umfasst sowohl durch Mycroft entwickelte Software, als auch die von anderen Anbietern. Teilweise werden für den selben Schritt durch Mycroft verschiedene Produkte angeboten, deren Schwerpunkt jeweils unterschiedlich ist [5]. Details zu den verschiedenen Möglichkeiten werden in Kapitel 3.2.1 dargestellt.

Snips AI Snips AI (kurz Snips) ist eine Sprachassistenzsoftware, die sich dem Schutz der Privatsphäre verschrieben hat. Dabei steht das Architekurprinzip „Privacy-by-Design“ im Fokus, welches den Datenschutz in den Mittelpunkt der Entwicklung

[2]https://www.panbachi.de/mycroft-ai-opensource-alternative-zu-alexa-und-co-350/ [Abgerufen am 11.05.2019]
[3]https://github.com/MycroftAI/mycroft-core [Abgerufen am 11.05.2019]
[4]https://mycroft.ai/get-started/ [Abgerufen am 19.09.2019]
[5]https://mycroft.ai/documentation/mycroft-software-hardware/ [Abgerufen am 11.05.2019]

stellt. Außerdem wirbt das Unternehmen damit, dass die Software die DSGVO-Regularien (siehe Kapitel 2.2.1) erfüllt und alle Verarbeitungsschritte direkt auf dem Endgerät durchgeführt werden können, wodurch keine Internetverbindung nötig ist [6]. Das System kann mit allen gängigen Betriebssystemen, außer Windows, benutzt werden. Es werden alle Linux Distributionen sowie MacOS direkt unterstützt, außerdem werden für Android und iOS entsprechende SDKs zur Verfügung gestellt, um eine einfache Integration in die Anwendungen zu ermöglichen [7].
Allerdings werden durch den Hersteller nicht alle Teile der Software frei zur Verfügung gestellt. Lediglich der Teil zur Verarbeitung der gesprochenen Sprache ist offen gelegt. Zu diesem wurden auch die theoretischen Betrachtungen veröffentlicht veröffentlicht [Cou+18].

Amazon Alexa Bei Amazon Alexa (kurz: Alexa) handelt es sich um die Sprachassistenzsoftware von Amazon, welche für die unternehmenseigenen intelligenten Lautsprecher der Echo-Reihe entwickelt wurde. Dieser Service wird über die Cloud zur Verfügung gestellt und erlaubt auch Drittanbietern, eigene Funktionen anzubieten[8]. Durch den Alexa Voice Service (AVS) bietet Amazon die Möglichkeit, Alexa direkt in eigene Geräte zu integrieren. Dabei werden mit Android, iOS, macOS, Ubuntu sowie Windows alle gängigen Betriebssysteme offiziell unterstützt. Außerdem kann das System auch direkt mit Raspbian eingesetzt werden. [9]

3.2 Architekturdetails der vorgestellten Systeme

In diesem Kapitel wird genauer betrachtet, wir die zuvor vorgestellten System die einzelnen Bestandteile der Sprachverarbeitung umsetzen. Dabei hat sich herausgestellt, dass die Unternehmen nicht immer kommunizieren, wie die einzelnen Verarbeitungsschritte funktionieren. Teilweise werden auch Schritte, die in Kapitel 2.1 als Einzelschritte betrachtet werden, zu einem zusammengefasst.

3.2.1 Mycroft AI

Mycroft zeichnet sich neben seiner Modularität auch dadurch aus, dass die Software komplett frei zu Verfügung gestellt wird. Dies umfasst eigene Implementierungen für jeden Schritt der Verarbeitung der natürlichen Sprache und auch ein Kooperation mit Mozilla zur Verbesserung des Sprachverständnisses. Außerdem verspricht das

[6] https://snips.ai/ [Abgerufen am 11.05.2019]
[7] https://snips.ai/developers/ [Abgerufen am 11.05.2019]
[8] https://developer.amazon.com/alexa-skills-kit [Abgerufen am 11.05.2019]
[9] https://github.com/alexa/avs-device-sdk/wiki [Abgerufen am 11.05.2019]

Unternehmen durch diese Offenlegung, dass der Nutzer die Hoheit über seine Daten behält und auch keine Daten verkauft oder für Werbezwecke benutzt werden [10].

Aktivierungsworterkennung Die Aktivierungsworterkennung kann mit verschiedenen Engines durchgeführt werden. Zum einen ist dies PocketSphinx [HD+06], welches unterschiedliche Worte auf Basis von Phonemen unterscheidet. Das ist die kleinste Einheit von Lauten der gesprochenen Sprache, die eine Unterscheidung zwischen Wörtern zulässt. Jedoch führt die Tatsache, dass diese in verschiedenen Sprachen unterschiedlich ausgesprochen werden, zu einem Problem. Wenn die interagierende Person nicht in ihrer Muttersprache kommuniziert, kann es zu fehlerhafter Aussprache von Phonemen kommen, wodurch die Worterkennung nur schlecht oder gar nicht funktioniert. Eine Internetverbindung für die Aktivierungsworterkennung wird nicht benötigt. [Amb+18]
Zum anderen kann Snowboy [11] von KITT.ai eingesetzt werden, welches auf einem neuronalen Netz basiert. Dieses muss entsprechend vor der Nutzung trainiert werden, damit das gewünschte Aktivierungswort erkannt wird. Die Erkennung kann offline durchgeführt werden, allerdings kann laut Aberkar et al. [Amb+18] nur ein Signalwort trainiert werden. Im Gegensatz zu PocketSphinx ist diese Software nicht Open-Source.
Außerdem kann die durch Mycroft entwickelte Engine Precise [12] eingesetzt werden. Als Basis dient hierfür ein rekurrentes neuronales Netz, das auf Geräuschmuster trainiert wird. Dadurch ist die Erkennung unabhängiger von einer Sprache oder einem Akzent. Genauso wie die beiden anderen Lösungen für diesen Verarbeitungsschritt mit Mycroft, funktioniert Precise ohne Internetverbindung.

Sprache-zu-Text Umwandlung Für die Sprache-zu-Text Umwandlung sind zwei verschiedene Umsetzungen von Interesse, auch wenn Mycroft prinzipiell die Nutzung vieler auf dem Markt verfügbarer Engines für diese Aufgabe ermöglicht (u.a. Watson STT, Google STT, Wit.ai) [13].
Eine der Umsetzungen ist Kaldi, eine Open-Source Umsetzung von STT. Diese zielt darauf ab, in möglichst vielen Umgebungen einsetzbar zu sein und mit Daten des Linguistic Data Consortium zu funktionieren. Dadurch wird eine große Zahl unterschiedlicher Sprachen unterstützt. Außerdem ist den Entwicklern wichtig, dass die Software während der Entwicklung ausgiebig getestet wird. Eine Besonderheit dieser Implementierung ist, dass sie direkt auf dem Endgerät funktioniert und somit keine Internetverbindung benötigt. [Pov+11]
Durch Mycroft wird wiederum eine Alternative beworben, die gemeinsam mit Mozilla entwickelt wird. Diese ist `OpenSTT`, wobei von Mycroft die Trainingsdaten für die

[10] https://mycroft.ai/ [Abgerufen am 14.05.2019]
[11] https://snowboy.kitt.ai/ [Abgerufen am 14.05.2019]
[12] https://mycroft.ai/documentation/precise/ [Abgerufen am 14.05.2019]
[13] https://mycroft.ai/documentation/mycroft-software-hardware/ [Abgerufen am 14.05.2019]

Künstliche Intelligenz geliefert werden, während durch Mozilla die Implementierung zur Verfügung gestellt wird. [14]

Die frei verfügbare Umsetzung der Sprachumwandlung wird von Mozilla als `DeepSpeech` bezeichnet und beruht auf den theoretischen Überlegungen von Hannun et al. [Han+14]. Dieses System nutzt Deep Learning und ermöglicht dadurch kontinuierliche Verbesserungen bei vergleichsweise einfachem Aufbau. Beispielsweise müssen durch den Entwickler keine Modelle zur Geräuschfilterung erstellt werden, da solche System selbst lernen. Zudem kommt es auch ohne Konzepte wie Phoneme aus, vielmehr findet es selbständig Unterscheidungskriterien. In Tests hat dieses System dabei Fehlerquoten von 6,5% erreicht, während die menschliche Fehlerrate für die gleichen Daten mit circa 10% beziffert wird. Die Implementierung durch Mozilla basiert auf dem Machine Learning Framework `TensorFlow` von Google. Für das Training werden Daten genutzt, die im Rahmen des „Common Voice“ Projekts gesammelt wurden. [15]

Da die Entwicklung von DeepSpeech maßgeblich durch die Community vorangetrieben wird, werden auch deren Interessen widergespiegelt. So standen im Juni 2019 circa 50 Sprachen zu Verfügung. Dies sind neben bekannten Sprachen wie Englisch oder Deutsch unter anderem auch Katalanisch oder Irisch. [16]

Um die Menge an Daten des Common Voice Projekts zu vergrößern, wurde durch Mycroft das „Open Dataset“ ins Leben gerufen. Dieses verwendet die Daten, die bei der Interaktion zwischen Mensch und Sprachassistent anfallen, wenn durch den Nutzer dieser Übermittlung explizit zugestimmt wurde. [17]

Der Einsatz von DeepSpeech kann nur mittels eines Server geschehen. Hierbei kann entweder ein eigener, lokaler Server oder eine Cloudumgebung genutzt werden. Aktuell gibt es Bestrebungen durch Mycroft, die Komplexität der Berechnungen so anzupassen, dass sie lokal auf dem Endgerät durchgeführt werden können.[17]

Intent Parser Die Erkennung der Intention des Nutzers kann entweder mit Adapt oder Padatious durchgeführt werden. Ersteres nutzt dabei einen Schlüsselwortabgleich und berechnet daraus einen Wahrscheinlichkeitswert. Anhand dessen wird danach der gewünschte Skill ausgewählt. Dieser Ablauf ist dabei identisch mit dem in Kapitel 2.1 beschriebenen. [18]

Ein anderer Ansatz wird mit Padatious verfolgt. Er basiert auf einem neuronalen Netz, wofür im Gegenteil zu Adapt keine einzelnen Wörter sondern ganze Sätze genutzt werden. Dadurch ist dieses System flexibler, da der Aufbau der Aussagen keiner starren Struktur folgen muss. [19]

Bei beiden Ansätzen kann es aber dazu kommen, dass mehr als nur ein Skill den

[14] https://mycroft.ai/blog/deepspeech-update/ [Abgerufen am 14.05.2019]
[15] https://hacks.mozilla.org/2017/11/a-journey-to-10-word-error-rate/ [Abgerufen am 15.05.2019]
[16] https://voice.mozilla.org/en [Abgerufen am 03.06.2019]
[17] https://mycroft.ai/voice-mycroft-ai/ [Abgerufen am 15.05.2019]
[18] https://mycroft.ai/documentation/adapt/ [Abgerufen am 20.05.2019]
[19] https://mycroft.ai/documentation/padatious/ [Abgerufen am 20.05.2019]

entsprechenden Intent ausführen kann. Aus diesem Grund kommt zusätzlich das Common Play Framework zum Einsatz. Dieses weist den verschiedenen Entitäten unterschiedliche Gewichte zu, wodurch dann die auszuführenden Aktion genauer bestimmt werden kann. [20]

Text-zu-Sprache Umwandlung Für die Umwandlung von Text in eine akustische Ausgabe, stellt Mycroft zwei verschiedene, selbst entwickelte Engines zur Verfügung. Dies sind Mimic TTS und Mimic 2 TTS. Dabei zeichnet sich Mimic durch einen geringen Ressourcenbedarf aus und kann somit direkt auf dem Endgerät eingesetzt werden. Für die Ausgabe stehen zwei verschiedene Stimmen für Englisch zu Verfügung, wobei in der Zukunft noch weitere Sprachen hinzugefügt werden sollen [21]. Entwickelt wird diese Umwandlung in Zusammenarbeit mit VocaliD, einem Unternehmen, dass Text so synthetisiert, dass auch Menschen mit Sprachbehinderung eine individuelle Stimme nutzen können. Ziel dieser Zusammenarbeit ist es, klare und möglichst natürlich klingende Sprachausgaben für Mycroft zu erzeugen [22]. Dafür werden durch Mimic verschiedene kurze Audioaufnahmen zu den gewünschten Wörtern kombiniert. [Jre+09]
Der Nachfolger, Mimic 2 TTS, setzt auf Deep Learning für die Generierung der Sprachausgabe. Dadurch kann das System Eigenschaften der Sprache, wie Betonung oder Rhythmus, besser umsetzen und somit eine menschlichere Ausgabe erzeugen. Allerdings bedarf diese Lösung mehr Rechenleistung, weshalb sie in der Cloud ausgeführt werden muss und somit eine Internetverbindung erfordert. [23]

3.2.2 Snips AI

Aktivierungsworterkennung Die Aktivierungsworterkennung erfolgt mittels des Service „snips-hotword“. Dieser wird direkt durch Snips zur Verfügung gestellt, jedoch gibt das Unternehmen keine Informationen über die Funktionsweise dieses Dienstes. Für den Nutzer besteht aber die Möglichkeit, eigene Aktivierungswörter zu trainieren. [24]

Sprache-zu-Text Umwandlung und Intent Parser Die Umwandlung von gesprochener Sprache in Text sowie die Analyse der dabei entstandenen Ausgabe findet bei Snips mittels eines einzigen Services statt. Die Extraktion des Textes aus der Audioeingabe geschieht, wie in Abbildung 3.2 dargestellt, mittels `Automatic Speech Recognition (ASR)` statt. Dieser folgt dem prinzipiellen Ablauf, wie er für STT in Kapitel 2.1 beschrieben wird. Dabei wird die Eingabephrase zuerst in eine phonetische

[20] https://mycroft.ai/documentation/skills/common-play-framework/ [Abgerufen am 20.05.2019]
[21] https://mycroft.ai/documentation/mimic/ [Abgerufen am 20.05.2019]
[22] https://mycroft.ai/blog/vocalidmimic/ [Abgerufen am 20.05.2019]
[23] https://mycroft.ai/blog/mimic-2-is-live/ [Abgerufen am 20.05.2019]
[24] https://docs.snips.ai/articles/platform/wakeword/personal[Abgerufen am 20.05.2019]

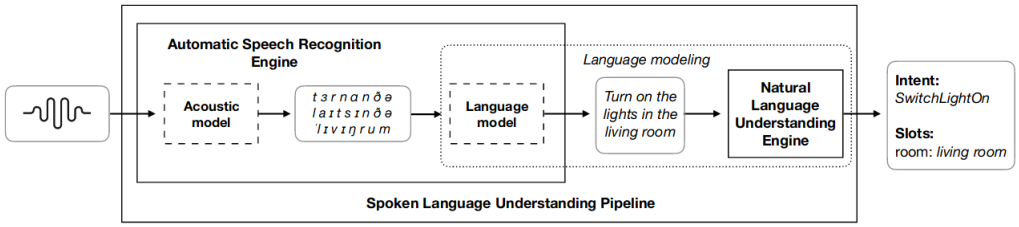

Abb. 3.2: Verarbeitungspipeline für gesprochene Sprache aus [Cou+18]

Repräsentation umgewandelt, aus der dann wiederum der Text generiert wird. Der Unterschied zu den vorherigen Beschreibungen liegt im Teil der `Natural Language Understanding (NLU)`. Diese ordnet die in Textform vorliegende Aussage den zuvor definierten Intents zu. Dabei werden die Phrasen und deren Entitäten nicht nur deterministische sonder auch probabilistisch voneinander unterschieden. Es können also sowohl solche Sätze verwendet werden, die den Trainingsdaten entsprechen (deterministische Unterscheidung) als auch solche, die diesen ähneln (probabilistische Unterscheidung), für die Bestimmung der Nutzerintention verwendet werden. [Cou+18]

Text-zu-Sprache Umwandlung Standardmäßig erfolgt die Text-zu-Sprache Umwandlung bei Snips offline. Hierbei kann zwischen verschiedenen Diensten gewählt werden. Diese können auch ohne Snips auf Linux Distributionen eingesetzt werden. Dabei handelt es sich unter anderem um „picotts“, „makerstts“, „pico2wave“. [25]
Es können aber auch andere Services für die Audioausgabe eingebunden werden, unter anderem sind dies Amazon Polly TTS und Google WaveNet TTS. Diese System wiederum benötigen eine Internetverbindung, da die Umwandlung cloudbasiert stattfindet. Außerdem ist es möglich, Mimic von Mycroft einzusetzen. [26]

3.2.3 Amazon Alexa

Im Mittelpunkt von Amazon Alexa (kurz Alexa) steht der Alexa Voice Service (AVS). Dieser gibt Zugang zu den cloudbasierten Komponenten von Alexa und ermöglicht eine Integration von Alexa auf unterschiedlichen Endgeräten. Da ein Großteil des Services auf Infrastruktur von Amazon funktioniert, benötigen Geräte, auf denen dieser Sprachassistent zum Einsatz kommt, nur eine geringe Menge an lokalen Ressourcen. [27]

[25] https://medium.com/snips-ai/is-it-google-aiy-nah-its-snips-f67c9dc2139a [Abgerufen am 21.05.2019]
[26] https://github.com/snipsco/awesome-snips#customisations [Abgerufen am 21.05.2019]
[27] https://developer.amazon.com/alexa-voice-service [Abgerufen am 23.05.2019]

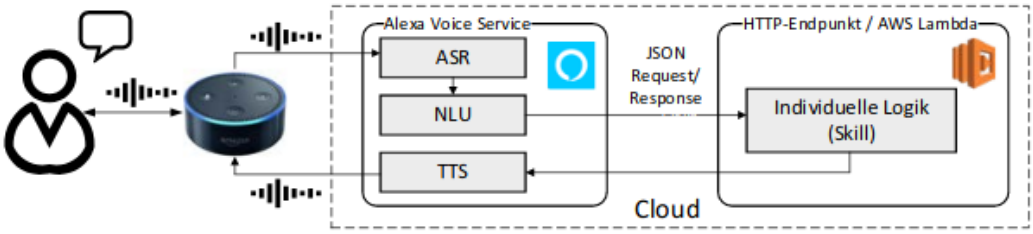

Abb. 3.3: Ablauf des Aufrufs eines Skills mit Alexa aus [Ank+19]

Aktivierungsworterkennung Es ist von Amazon vorgesehen, dass sowohl Sensory TrulyHandsfree als auch Snowboy für die Erkennung des Aktivierungswortes eingesetzt werden können [28]. Snowboy wird Von KITT.ai, einem Unternehmen, das inzwischen zu der chinesichen Firma Baidu gehört, zur Verfügung gestellt. Mit diesem Service ist es möglich eigene Aktivierungswörter für Alexa zu nutzen. Die Erkennung selbst wird direkt auf dem Endgerät durchgeführt [29].
Bei TrulyHandsfree von Sensory handelt es sich um eine Software, die nach Unternehmensangabe die am meisten genutzte Engine für Spracherkennung ist. Diese zeichnet sich besonders durch hohe Genauigkeit, geringen Ressourcenverbrauch sowie Unterstützung vieler Plattformen aus. Dabei ist es möglich, sowohl vordefinierte Aktvierungswörter zu nutzen als auch eigene zu definieren [30].

Sprache-zu-Text Umwandlung Wie in Abbildung 3.3 dargestellt, findet dieser Umwandlungsschritt vollständig auf Infrastruktur von Amazon statt. Die nach dem Aktivierungswort getätigten Aussagen werden mittels Automatic Speech Recognition (ASR) automatisch in Text umgewandelt. Dafür kommt der Dienst „Amazon Lex“ zum Einsatz. Die dabei verwendeten Funktionen wurden zu diesem Zweck zuvor mittels Deep Learning trainiert. [31]

Intent Parser Die Erkennung des Intents des Nutzers findet mittels Alexa Skill Kit statt. Dieser greift dabei auch auf Natural Language Understanding (NLU) zurück, welches ebenfalls Bestandteil von AVS ist und auch auf mittels Deep Learning trainierte Funktionen zurückgreift. In dem Skill Kit werden verschiedene Fähigkeiten definiert, die mit Alexa ausgeführt werden können. Außerdem sind die Aussagen festgelegt, die zu der Ausführung des gewünschten Skills führen. Genauere technische Funktionalitäten werden durch Amazon nicht aufgeführt. [32]

[28] https://github.com/alexa/avs-device-sdk/wiki/cmake-parameters#Wake-word-detector [Abgerufen am 23.05.2019]
[29] https://snowboy.kitt.ai/ [Abgerufen am 23.05.2019]
[30] https://www.xda-developers.com/sensory-trulyhandsfree-low-power-hotword/ [Abgerufen am 23.05.2019]
[31] https://developer.amazon.com/alexa-skills-kit/asr [Abgerufen am 23.05.2019]
[32] https://developer.amazon.com/docs/custom-skills/create-intents-utterances-and-slots.html [Abgerufen am 23.05.2019]

Text-zu-Sprache Umwandlung Die Umwandlung des Textes in gesprochene Wörter geschieht mittels Amazon Polly, welcher wie in Abbildung 3.3 ersichtlich, auch Teil von AVS ist. Dieser Service nutzt Deep Learning, um eine möglichst menschenähnliche Sprachausgabe zu erzeugen. Diese kann in acht verschiedenen Sprachen mit 28 unterschiedlichen Stimmen erfolgen. [33]

3.3 Möglichkeiten zur Verhinderung von Angriffen auf die Privatsphäre mit den einzelnen Sprachassistenten

Die Basis der in diesem Kapitel betrachteten Angriffe stellt das Kapitel 2.2.3 dar. Dabei leiten sich die Maßnahmen zur Verhinderung von Angriffen sinnvollerweise von den verschiedenen Angriffsszenarien ab. Manche Maßnahmen können durch den Nutzer ergriffen werden, andere wiederum müssen durch den Hersteller implementiert werden.
Beispielsweise kann der Nutzer einem Szenario (i), wie im vorigen Kapitel beschrieben, einfach vorbeugen. Dafür muss der Ort des Sprachassistenten mit Bedacht gewählt werden (z.B. nicht in der Nähe von Fenstern) oder das System bei Verlassen der Wohnung von der Stromquelle getrennt werden [JO18].
Für einen Vergleich der verschiedenen Assistenzsysteme ist es jedoch von größerem Interesse, die herstellerspezifischen Abwehrmaßnahmen zu betrachten.
Damit ungewollten Aufnahmen von Geräuschen verhindert werden können, bietet es sich einerseits an, das Aktivierungswort so zu wählen, dass es nicht versehentlich in Konversationen auftaucht. Dies kann entweder ein entsprechendes, durch den Hersteller voreingestelltes Aktivierungswort sein oder ein passendes selbstgewähltes [e.V17].
Sinnvoll ist auch, den Nutzer mittels einer Benachrichtigung darüber zu informieren, dass eine Aufnahme begonnen wurde. Dafür ist ein Signalton am besten geeignet, der möglicherweise noch durch ein Lichtsignal unterstützt wird. [JO18; Edu+19].
Die Manipulation des Endgerätes durch hochfrequente Töne lässt sich von Seiten des Nutzers nur dadurch vermeiden, in dem eine Sprachauthentifizierung genutzt wird. Diese kann einen Nutzer anhand seiner Sprachmuster erkennen und weiß somit, ob dieser die Berechtigung zur Verwendung des Systems hat [Edu+19]. Eine solche Funktion wird aktuell von Alexa angeboten, wobei dieses Feature erst manuell durch den Nutzer aktiviert werden muss [FB18b]. Eine praktische Umsetzung wurde von Amazon im September 2018 durch eine entsprechende Patentanmeldung vorgestellt. Dabei erstellt der Nutzer bei der Konfiguration für jeden Befehl eine als Wasserzeichen bezeichnete Aufnahme. Anschließend wird bei jeder Verwendung der jeweilige

[33] https://aws.amazon.com/de/polly/ [Abgerufen am 23.05.2019]

Befehl mit seinem Wasserzeichen abgeglichen. [AT18]
Allerdings wird von von Chen et al. [Che+17] beschrieben, das auch eine solche Verifizierung keine absolute Sicherheit garantiert. Zur Verbesserung wird von Feng et al. [Fen+17] vorgeschlagen, einen tragbaren Sicherheitstoken einzusetzen. Dieser vergleicht die Körpervibrationen mit den gesprochenen Phrasen, wodurch der Nutzer mit einer Genauigkeit von 97 % korrekt authentifiziert werden konnte. Eine offizielle Integration mit den verschiedenen Sprachassistenten gibt es bislang nicht, allerdings wurde das Projekt mit Google Now prototypisch implementiert. Dieser Prototyp basiert darauf, dass der Assistent über Bluetooth mit dem Token kommuniziert. Eine Auswertung der Ergebnisse wiederum findet in der Cloud statt. [Fen+17]
Attacken auf Daten, die sich in der Cloud befinden, können am effektivsten dadurch unterbunden werden, in dem keine Daten in dieser abgelegt werden. Das entsprechende Prinzip ist das der Datensparsamkeit, das besagt, dass nur so viele Daten wie unbedingt nötig erhoben werden sollten. In diesem Falle würde dies bedeuten, dass die Verarbeitung von Daten lokal stattfindet. Ermöglicht wird dies von Mycroft [34] und Snips [35], bei hingegen Alexa bedarf es einer Verbindung zu der Cloud. [Ank+19]
Wenn es nicht möglich ist, eine Kommunikation mit der Cloud zu umgehen, sollte diese mindestens verschlüsselt stattfinden. Da es aber beispielsweise Apthorpe et al. [Apt+17] gelungen ist, auch aus den verschlüsselten Daten Informationen über den Nutzer zu gewinnen, ist dies keine hinreichende Schutzmaßnahme.
Sobald eine Auswertung der Daten in der Cloud erfolgt, werden diese auch dort abgelegt, selbst wenn entsprechend dem Prinzip der Datensparsamkeit nur so wenig Daten wie möglich übermittelt werden. Aus diesem Grund wird von Jackson und Orebaugh [JO18] vorgeschlagen, diese Daten regelmäßig auf unerlaubten Zugriff zu kontrollieren und sie generell so schnell wie möglich zu löschen. Im Falle von Alexa ist die mittels einer entsprechenden App oder per Webzugriff auf das eigene Nutzerkonto problemlos möglich. Da sowohl Snips als auch Mycroft eine Datenverarbeitung ohne den Einsatz einer Cloud ermöglichen, bedarf es keiner derartigen Datenverwaltung.
Einen effektiven Schutz gegen manipulierte Skills schlagen sowohl Jackson als auch Edu et el. nicht vor [Edu+19; JO18]. Auch wird von den Anbietern der Assistenzsoftware kein Schutz vor solchen manipulierten Skills angeboten. Allerdings hat der Nutzer aufgrund der Quelloffenheit von Mycroft bei diesem Sprachassistenten die Möglichkeit, die einzelnen Skills und ihre verbundenen Funktionen selbst zu untersuchen sowie zu bewerten. Für eine solche Untersuchung wird aber ausreichendes technisches Verständnis benötigt, welches nicht allgemein vorausgesetzt werden kann.
Eine Übersicht über die Umsetzung der verschiedenen Abwehrmaßnahmen mit den jeweiligen Assistenzsystemen gibt die Tabelle 3.1.

[34] https://mycroft.ai/blog/deepspeech-update/ [Abgerufen am 20.05.2019]
[35] https://snips.ai/products/flow/ [Abgerufen am 20.05.2019]

	Mycroft	Snips	Alexa
Individuelles Aktivierungswort	✓	✓	✓
Benachrichtigung	✓ akustisches Signal	✓ akustisches Signal	✓ hardwareabhängig
Sprach-authentifizierung	X	X	✓ Aktivierung durch Nutzer [FB18b]
Sicherheitstoken	manuelle Implementierung	manuelle Implementierung	manuelle Implementierung
Verzicht auf Cloud	✓ möglich, wenn STT auf eigenem Server	✓	X
Verschlüsselte Kommunikation	✓	nur lokale Verarbeitung	✓
Datenspeicherung in Cloud	X	X	✓
Skillverifizierung	eigenständig durch Nutzer	?	?

Tab. 3.1: Möglichkeiten zur Umsetzung der Abwehrmaßnahmen

3.4 Geeignete Einsatzszenarien für die gemeinsame Nutzung von Sprachassistenten und Assistenzroboter

Damit es möglich ist, ein allgemeingültiges Konzept für die Mensch-Roboter-Interaktion mithilfe eines Sprachassistenten zu formulieren, müssen zunächst solche Szenarien gewählt werden, deren Grundanforderungen sich ähneln und häufig auftreten. Aufgrund der Orientierung an den grundlegenden Anforderungen ist es außerdem möglich, das später in diesem Kapitel erstellte Konzept auf weitere Szenarien anzuwenden. Die Grundlage für die Auswahl geeigneter Szenarien stellen die Betrachtungen in Kapitel 2.3 dar. Um folgende Szenarien handelt es sich:

i. Unterstützung bei der Pflege von Patienten
ii. Einsatz als Gehhilfe
iii. Navigationshilfe
iv. Haushaltshilfe (z.B. Reinigung von Böden)

Einige dieser Szenarien lassen direkt erkennen, dass sie sehr ähnliche Anforderungen an den Roboter stellen. So muss dieser in den Szenarien (iii) und (iv) dazu in der Lage sein, sich eigenständig in einer Umgebung zurecht zu finden und ein Ziel zu erreichen. Während dieses im Fall der Navigationshilfe ein klar definierter Punkt in einer Umgebung ist, ist das Ziel bei der Haushaltshilfe abstrakter. Beispielsweise muss der Roboter bei der Bodenreinigung jeden Punkt im Raum abfahren. Das heißt, er muss wissen, wo er noch nicht war und wie er diese Position erreichen kann.
Auch Falle des Einsatzes als Gehhilfe, Szenario (ii), kann es sein, dass der Roboter dabei hilft, den Weg zu einem zuvor festgelegten Ziel zu finden. Dies wird zum Beispiel bei dem Care-o-Bot II so gehandhabt [Gra+04].
Im Einsatz für die Unterstützung bei der Pflege von Patienten sind die Anwendungsfälle, siehe Kapitel 2.3.2, vielfältiger. Aber auch hierbei muss sich der Roboter selbständig bewegen. Beispielsweise ist es vorstellbar, dass ein Pfleger einen Roboter eine Aufgabe der Art „Gehe zu Patient X“ oder „Gehe in Raum Y“ gibt. In diesem Fall muss der Roboter auch wissen, wo er sich befindet und wie er zu seinem Ziel gelangen kann.
Ein weiterer Anwendungsfall im Zusammenhang mit der Pflege ist, dass einer Person (Pfleger oder zu Pflegender) ein Gegenstand gebracht werden soll. Auch dafür muss der Roboter in der Lage sein, den Ort des Gegenstandes zu erreichen und wieder den Weg zu seinem Ausgangspunkt zu finden.

3.5 Zusammenfassung

Um sinnvolle Anforderungen an ein Konzept für die Mensch-Roboter-Interaktion mittels natürlicher Sprache zu formulieren wurden zunächst die Einsatzszenarien von Assistenzrobotern auf ihre wesentliche Grundfunktionen zerlegt. Außerdem wurden drei verschieden Sprachassistenten genauer darauf untersucht, wie sie einzelnen Schritte der Verarbeitungspipeline umsetzen und wie sie Angriffen auf die Privatsphäre der Nutzer vorbeugen.

4 Konzept für den Einsatz von Sprachassistenten mit einem Assistenzroboter

Basierend auf der Analyse der Funktionsweise von Sprachassistenten (siehe Kapitel 3.2) und der Einsatzszenarien für Assistenzroboter im vorigen Kapitel, ist es möglich Anforderungen an den gemeinsamen Einsatz der beiden Komponenten zu formulieren.
Anhand dieser Anforderungen ist es anschließend möglich, einen passenden Sprachassistenten aus den im vorigen Kapitel vorgestellten zu wählen. Außerdem können mittels dieser Kriterien passende Komponenten für die einzelnen Verarbeitungsschritte gewählt werden. Eine Auswahl eines Roboters geschieht aber in diesem Zusammenhang nicht, da sich das Konzept durch Allgemeingültigkeit für Assistenzroboter auszeichnen soll.
Abschließend ist es möglich, ein Konzept für die Zusammenarbeit zu erstellen, das die Abläufe zwischen Sprachassistent und Assistenzroboter so festlegt, dass die Anforderungen bestmöglich erfüllt werden.

4.1 Anforderungen auf Basis der vorherigen Betrachtungen

Im Folgenden werden die Anforderungen an den Einsatz für einen Assistenzroboter mit einem Sprachassistenten ausgehend von den im vorigen Abschnitt beschriebenen Szenarien analysiert. Als grundlegende Szenarien dienen dabei die in Kapitel 3.4 herausgearbeiteten. Deren Gemeinsamkeit ist, dass die Grundfunktion des selbständigen Erreichens eines zuvor kommunizierten Zieles. Dafür wird auch explizit definiert, welche Anforderungen nicht Teil des Konzepts sind, da sie den Umfang dieser Arbeit übersteigen. Eine grundlegende Anforderungen an das Gesamtsystem ist, dass es mit verschiedensten Assistenzrobotern eingesetzt werden kann.

Bewegung Damit die eingangs erwähnte Grundfunktion des Erreichens eines Punktes im Raum möglich ist, benötigt der Roboter die Fähigkeit, sich selbständig fortzubewegen. Dabei ist irrelevant auf welche Art die Bewegung stattfindet. Vorstellbar

sind Räder oder auch Beine, die Umsetzung obliegt dem Entwickler des Roboters. Entscheidend ist aber, dass die Fortbewegung aus eigener Kraft geschieht und nicht wie bei Hayashi et al. [Hay+05] dadurch, dass der Roboter mit dem Menschen fest verbunden ist.
Bei dieser Anforderung handelt es sich um eine Grundvoraussetzung für den sinnvollen Einsatz des Konzepts, allerdings ist es nicht Teil dieses. Eine spezifischere Festlegung würde außerdem die Einsatzmöglichkeiten des Konzeptes zu stark beeinträchtigen, da auf diesem Weg eine zu hohe Zahl von Geräten von dem Einsatz mit diesem Konzept ausgeschlossen werden würden. Es sollte für das Konzept davon ausgegangen werden, dass passende Steuerungsbefehle über passende Schnittstellen versendet werden können, die entsprechende Handlungen hervorrufen.

Orientierung Damit die eigenständige Bewegung des Roboters sinnvoll funktioniert, benötigt diese die Fähigkeit, Hindernisse sowie Gegenstände und Menschen zu erkennen. Dafür werden entsprechende Sensoren benötigt, die vermutlich einfachste Umsetzung ist dabei die mittels eines Bildsensors. Dieser erlaubt die Untersuchung der erhaltenen Daten auf verschiedene Parameter und ermöglicht somit auch eine Vielzahl weiterer Anwendungen. Jedoch ist auch diese Umsetzung dem Entwickler des Roboters überlassen, wobei insbesondere eine Erkennung von Menschen nötig ist, damit dem Nutzer signalisiert werden kann, das seine Präsenz erkannt wurde.
Ein weiterer wichtiger Punkt ist die Fähigkeit, selbständig den Weg zu einem Ziel zu erkennen. Dafür ist es nötig, dass der Roboter die möglichen Wege in seiner Umgebung kennt. Dies beutetet, dass er über entsprechendes Kartenmaterial verfügt und sich auf diesem auch selbst lokalisieren kann. Für diese Arbeit ist allerdings davon auszugehen, dass es die Möglichkeit gibt, entsprechende Funktionen mittels entsprechender Schnittstellen anzusprechen und die Antworten gut auswertbar sind. Die Funktionen der Orientierung sind allerdings auch nicht durch dieses Konzept definiert. Sie werden vielmehr durch das Konzept benutzt, um einen Einsatz des Roboters entsprechend der in Abschnitt 3.4 erörterten Szenarien zu ermöglichen.

Architektur des Roboters Das Erscheinungsbild des Roboters ist kein Teil des Konzepts und stellt auch wird auch nicht durch das Konzept bedingt. Anhand der in Kapitel 3.4 beschriebenen Szenarien ist erkennbar, dass sich die spezifischen Anwendungen der Assistenzroboter teilweise stark unterscheiden. Aus diesem Grund kann nicht gesagt werden, über welche spezifischen Extremitäten ein einzelner Roboter verfügen sollte, da jeder Asssistenzroboter auf seine Einsatzgebiete spezialisiert ist. Die Fähigkeit, Dinge zu greifen und zu transportieren beispielsweise ist sehr hilfreich, aber neben einer Nutzung von Armen sind auch andere Umsetzungen vorstellbar.
Zwangsläufig benötigt werden mindestens ein Mikrofon und ein Lautsprecher, die Teil des Roboters sind. Wünschenswert sind mehrere Mikrofone, die gemeinsam als Arraymikrofon eingesetzt werden können. So ist es möglich, dass die Richtung er-

kannt wird, aus der mit dem Roboter kommuniziert wird. Anhand dieser Erkennung kann sich der Roboter anschließend in die Richtung des Kommunikaktionspartners bewegen, so dass das Gefühl entsteht, dass sich der Roboter dem Nutzer zugewendet hat. Für ein Gefühl der Natürlichkeit der Interaktion ist die Möglichkeit eines solchen Verhaltens förderlich.
Eine Anforderung an das Konzept ist, dass es sich unabhängig des Betriebssystems, das auf dem Roboter läuft, benutzt werden kann.

Interaktion Die Interaktion zwischen dem Nutzer und dem Roboter stellt einen wesentlichen Teil der Anforderungen dar. Insbesondere aufgrund der Tatsache, dass die Nutzergruppe sehr heterogen ist, wie in Kapitel 2.4 erarbeitet, muss die Interaktion möglichst selbsterklärend aufgebaut sein. Personen, die bereits Erfahrungen im Umgang mit vergleichbaren Assistenzsystemen haben, sollten genauso in der Lage sein das System zu benutzen, wie jene ohne jegliche Vorkenntnisse. Entscheidend ist außerdem, dass sich die Interaktion nicht wie eine zusätzliche Last anfühlt, damit die Bereitschaft zur Nutzung des Systems maximiert wird [Böh02].
Aus diesem Grund ist es wichtig, dass die Nutzer nicht auf feste Befehle zurückgreifen müssen, sondern möglichst frei in der Wortwahl sind. Dies setzt voraus, dass die Umwandlung von Sprache zu Text mit einer ausreichenden Genauigkeit vollzogen wird. Außerdem ist es wichtig, dass die passenden Befehle auch aus einer längeren Wortgruppe korrekt entnommen werden. Das bedeutet, dass das System in der Lage sein muss, entsprechende Filterungen vorzunehmen, so dass nur die für die Ausführung eines Befehls benötigten Informationen weiterverwendet werden. Dies setzt voraus, dass der Intent Parser auch eine hinreichend große Menge an Vokabeln einem bestimmtem Befehl zuordnen kann.
Dabei darf es für das Konzept keine Rolle spielen, auf welcher Sprache die Interaktion stattfindet, jedoch ist es wünschenswert, ohne großen Konfigurationsaufwand die benutzte Sprache zu wechseln. Dadurch ist der Einsatz des Systems nicht auf einen bestimmten Sprachraum beschränkt und zum anderen kann individuell auf die Bedürfnisse der Nutzer eingegangen werden. Beispielsweise ziehen aktuell immer häufiger Senioren aus Deutschland ins Ausland. [1] Nicht immer sprechen sie die Landessprache, was besonders die Pflege dieser Personen maßgeblich erschweren kann. Können sie aber in ihrer Sprache mit dem Roboter kommunizieren, werden solche Schwierigkeiten minimiert und im Idealfall die Interaktion mit dem Pflegepersonal verbessert.
In Kapitel 2.4 wurden zusätzlich noch Anforderungen an das Verhalten des Roboters im Rahmen der Interaktion erarbeitet. So ist es von Vorteil, wenn der Roboter in der Lage ist, einer Person zu signalisieren, dass er diese wahrgenommen hat. Darauf folgenden Aktionen können, unter Berücksichtigung des Kontexts, beispielsweise

[1] https://www.zeit.de/wirtschaft/2019-07/deutsche-rentenversicherung-zahlungen-deutsche-ausland [Abgerufen am 05.10.2019]

Hilfsangebote oder Ausweichmanöver sein. Beschrieben wird ein ähnliches Verhalten von Prodanov et al. [Pro+02] für den Fall der Nutzung eines Assistenzroboters als Guide in einem Museum. Dabei ist die Nutzergruppe sehr heterogen. Personen in dieser haben möglicherweise Vorbehalte gegenüber der Nutzung eines Roboters, welche ihnen dadurch genommen werden, in dem sie aktiv von diesem angesprochen werden und somit die Interaktion durch den Roboter gestartet wird.
Des weiteren ist es wünschenswert, wenn der Roboter die in Kapitel 2.4 erarbeiteten Faktoren auf die Akzeptanz des Roboters berücksichtigt. Das heißt, er sollte sich in einer an die Nutzer angepassten Geschwindigkeit bewegen und auch einen Abstand zu ihnen einhalten, der den persönlichen Raum respektiert. Damit einher geht auch die Anpassung der eigenen Aktionen an eine erkannte Person.
Eine weitere Anforderung ist, dass die Sprachausgabe des Roboters gut verständlich ist. Dabei ist wünschenswert, dass diese Stimme zusätzlich möglichst natürlich klingt, so dass für den Nutzer anhand der Sprache eher der Eindruck entsteht, mit einem Menschen als mit einer Maschine zu interagieren.
Außerdem ist wichtig, dass sich der Roboter höflich und in vollständigen, korrekten Sätzen ausdrückt. Wünschenswert ist zusätzlich, dass er über ein gewisses Repertoire an Aussagen verfügt, so dass diese variiert werden können, auch wenn sie die selbe Nachricht überbringen sollen.

Datenschutz Da die Einsatzszenarien aus Kapitel 3.4 sich alle im persönlichen Umfeld der Nutzer abspielen, ist der Datenschutz eine wichtige Anforderung an das Konzept. Auch gerade im Fall der Pflege werden teilweise hochsensible Daten ausgetauscht, die nicht nur Zahlungsdaten umfassen, sondern Einblicke in den aktuellen Gesundheitszustand der einzelnen Personen gibt. Gerade dem Schutz solcher Daten ist große Bedeutung zu geben.
Des weiteren spielt der Datenschutz eine wichtige Rolle im Zusammenhang mit der Nutzeraktzeptanz und somit der Bereitschaft, einen Assistenzroboter mit Sprachassistenten zu nutzen. Beispielsweise wurde im Rahmen einer bitkom Studie herausgefunden, dass circa drei Viertel der Bundesbürger keinen Sprachassistenten nutzen möchten, da sie ihre Daten nicht an Unternehmen wollen [2].
Aber auch die Ergebnisse von Fruchter et al. [FL18] bestätigen, den Einfluss des Datenschutzes auf die Nutzererfahrung. In ihrer Arbeit haben sie herausgefunden, dass auch Personen, die bereits einen Sprachassistenten verwenden, Bedenken haben, von diesem ausspioniert zu werden.

Funktionen Prinzipiell soll es möglich sein, dass mit dem Konzept unterschiedlichste Funktionen aufgerufen werden können. Allerdings soll der Fokus darauf gelegt werden, dass mit den Funktionen eine Steuerung des Roboters möglich ist.

[2] https://www.bitkom.org/Presse/Presseinformation/Digitale-Sprachassistenten-als-intelligente-Haushaltshelfer.html#item-911-close [Abgerufen am 07.07.2019]

Das bedeutet, dass durch den Sprachassistenten die gesprochenen Aussagen so in maschinenverständliche Befehle übersetzt werden, dass der Roboter diese ohne Probleme ausführen kann.
Es sollte jedoch auch keine Probleme bereiten, weitere Funktionen manuell zu installieren oder bestehende Funktionen anzupassen. Nur auf diese Art kann das System an veränderte Bedürfnisse angepasst werden und viele Anwendungsfälle abdecken.
Wünschenswert ist außerdem, dass der Roboter ohne Probleme einige grundlegende Informationen geben kann, so wie dies aktuell mit vergleichbaren Systemen möglich ist. Solche Informationen sind beispielsweise das aktuelle Wetter oder lexikalische Daten von Begriffen.

Infrastruktur am Einsatzort Ziel ist es, dass der Einsatz des Assistenzroboters mit minimaler Infrastruktur stattfinden kann. Sollte der Einsatz mit einem erheblichen Installationsaufwand und damit verbundenen Zusatzkosten einhergehen, könnte dies potentielle Nutzer abschrecken. Um eine maximale Funktionsbreite des Gesamtsystems zu ermöglichen, sollte der Aufbau einer Internetverbindung mittels WLAN durch den Roboter möglich sein. Dieses ist beispielsweise nötig, damit Informationen über das aktuelle Wetter geliefert werden können. Aber auch für Einsatzszenarien, die in Kapitel 2.3 beschrieben werden, wird zwangsläufig eine Internetverbindung benötigt. So kann eine Telepräsenz nur dann durchgeführt werden, wenn der Roboter eine Verbindung mit der Außenwelt hat.
Für andere Anwendungen, wie der Dokumentation der Pflege reicht auch eine lokale Netzwerkverbindung aus, die mit dem entsprechenden Server für die Dokumentation verbunden ist. Dafür muss aber ebenfalls mit der selben Infrastruktur (z.B. Router) die Drahtlosverbindung eingerichtet werden.

Zusammenfassung der Anforderungen Aus den vorherigen Betrachtungen ergeben sich somit die folgenden Anforderungen:

i. Voraussetzungen:
   a) Roboter:
      i. API für Bewegungssteuerung
      ii. API für Navigation und Orientierung
      iii. WLAN Modul
      iv. Sensor, der Erkennung von Objekten/Personen ermöglicht
      v. Lautsprecher und Mikrofon in ausreichender Qualität
   b) Infrastruktur:
      i. Drahtlosnetzwerkverbindung (vorzugsweise mit Internetanbindung)

ii. Muss-Ziele:
   a) Interaktion:
      i. Steuerung des Roboters mittels natürlicher Sprache
      ii. Personenerkennung und Signalisierung der Erkennung
      iii. möglichst natürliche Antworten (im Bezug auf Klang der Stimme und Ausdrucksweise)
      iv. zuverlässige Sprache-zu-Text Umwandlung
      v. Erzeugen eines natürlichen Gefühls der Interaktion
      vi. möglichst freie Wortwahl für Erteilung von Befehlen
   b) Einhaltung Regelungen des Datenschutzes
   c) Anpassung der Sprache problemlos möglich
   d) Einsatz des Sprachassistenten unabhängig vom Betriebssystem des Roboters
   e) von bestimmter Sprache unabhängiges Konzept
   f) Problemlose Installation neuer Funktionen
   g) Anpassung installierter Funktionen mit geringem Aufwand durchführbar
   h) Bedienung ohne Vorwissen über die Funktionsweise
   i) Anpassungen der Kernfunktionen nach eigenem Bedarf möglich

iii. Kann-Ziele:
   a) Nutzung weiterer Funktionen (z.B. Wetterinformationen)
   b) Fähigkeit, ein Arraymikrofon zu nutzen
   c) Antworten mit unterschiedlichen Sätzen, deren Nachricht die Gleiche ist
   d) Rücksicht des Roboters auf Sozialnormen (Abstand, Geschwindigkeit, Reaktion)
   e) Verwendung eines individuellen Aktvierungswortes

f) schnelle Reaktionen des Roboters

iv. Anforderungen, die nicht Teil dieser Arbeit sind:
   a) Kartierung der Umgebung
   b) Ausweichen
   c) Objekerkennung
   d) Funktionsweise der Bewegung, Navigation, Orientierung

4.2 Auswahl eines geeigneten Sprachassistenten auf Basis der Anforderungen

Da das Konzept unabhängig von dem eingesetzten Roboter funktionieren soll, kommt der Wahl eines geeigneten Sprachassistenzsystems für die Interaktion eine entscheidende Rolle zu.
Auf Basis der in Kapitel 3.1 untersuchten Sprachassistenzsyteme wird im Folgenden die Auswahl eines geeigneten Systems getroffen. Bei den Sprachassistenten handelt es sich um Mycroft, Snips und Alexa. Entscheidend für die Auswahl sind die Anforderungen, die sich auf die Sprachverarbeitung beziehen. Anforderungen an die Fähigkeiten des Roboters können in diesem Zusammenhang nicht betrachtet werden, da es sich um zwei separate Systeme handelt.
Dafür wird der Fokus auf die Erfüllung der Muss-Ziele gelegt, wobei die Sprachassistenten so gewählt werden sollten, dass auch die Kann-Ziele erfüllbar sind.
Dabei erlauben alle Assistenten die Steuerung des Roboters mittels natürlicher Sprache. Zum einen können sie Eingaben dieser Art alle verarbeiten (vgl. Kapitel 2.1), zum anderen erlauben sie auch die Installation von Skills, die durch Drittanbieter zur Verfügung gestellt werden. Ein solcher Skill würde die unterschiedlichen Befehle an den Roboter versenden. Damit einher geht die Möglichkeit, die Wörter für die Eingabe frei zu wählen.
Für eine Interaktion muss der Nutzer auch bei allen Systemen einzig das Aktivierungswort kennen, welches auch bei Bedarf individualisiert werden kann.
Da alle drei Assistenzsysteme mindestens ein Mikrofon sowie Lautsprecher bedürfen, ist dies auch kein Unterscheidungskriterium.
Die Differenzierung der System kann jedoch anhand der Datenschutzanforderungen vorgenommen werden. Da mit Alexa die Sprachdaten zwangsläufig auf Servern von Amazon verarbeitet werden müssen, tritt der Nutzer seine Datenhoheit ab. Da somit die Anforderungen an den Datenschutz nicht erfüllt werden können, eignet sich Alexa nicht für das zu erstellende Konzept.
Mycroft und Snips hingegen legen großen Wert auf Datenschutz und können diesen auch vollständig garantieren. Dabei bedarf es bei Mycroft der Konfiguration eines eigenen Servers zur Umwandlung von Sprache-zu-Text. Diese Konfiguration

würde im produktiven Einsatz durch den Anbieter durchgeführt. Der Server würde in diesem Fall neben dem Sprachassistenten und Assistenzroboter auch eine zugehörige Verarbeitungseinheit für die lokale Sprache-zu-Text Umwandlung erhalten. Sollte dies nicht möglich sein, kann die Verarbeitung auch auf den Mycroft-Servern durchgeführt werden, deren Nutzung sich durch sehr datenschutzfreundliche Nutzungsbedingungen auszeichnet. [3]
Dem gegenüber steht Snips mit der Möglichkeit, jegliche Berechnungen direkt auf dem Endgerät durchzuführen. Allerdings gibt es hier das Problem, dass nicht alle Teile des Programmcodes offen gelegt sind. Somit kann die Funktionsweise nicht lückenlos nachvollzogen werden, was dazu führt, dass man auf die Angaben des Herstellers vertrauen muss. Außerdem kann es zu Problemen bei der genauen Anpassung des Systems an spezifische Anforderungen führen.
Aufgrund der Tatsache, dass es sich bei Mycroft um ein komplettes Open-Source Projekt handelt und somit alle Verarbeitungsschritte lückenlos nachvollzogen werden sowie jederzeit eigene Anpassungen vorgenommen werden können, ist Mycroft System das der Wahl.

4.3 Auswahl der Bestandteile für die einzelnen Verarbeitungsschritte des Sprachassistenzsystems

Da sich Mycroft durch Modularität auszeichnet, kann für jeden einzelnen der Verarbeitungsschritte eine Vielzahl verschiedener Lösungen eingesetzt werden. Aus diesem Grund werden im Folgenden die am besten passenden Umsetzungen für die einzelnen Schritte gewählt, indem sie daraufhin untersucht werden, wie gut sie die Anforderungen jeweils erfüllen.

Aktivierungsworterkennung Für die Aktivierungsworterkennung können mit PocketSphinx, Precise sowie Snowboy drei verschiedene Systeme eingesetzt werden (siehe Kapitel 3.2.1). Jede dieser Erkennungen kann dabei offline eingesetzt werden. Somit erfüllen sie gleichermaßen die Anforderungen an den Datenschutz.
Snowboy hat allerdings das bereits im Zusammenhang mit Snips erwähnte Problem, das der Code nicht frei zugänglich ist. Somit begibt man sich mit diesem System in eine Abhängigkeit vom Hersteller, die bei den beiden anderen, durch Mycroft entwickelten Systemen, nicht besteht. Aus diesem Grund sollten sie gegenüber Snowboy bevorzugt werden.
Precise hebt sich durch seine Architektur hervor, die es ermöglicht, verschiedene Sprachen und Aussprachen zu verstehen und dadurch eine höhere Nutzeraktzeptanz

[3] https://mycroft.ai/embed-privacy-policy/ [Abgerufen am 01.10.2019]

verspricht. Da für die Nutzung von PocketSphinx Phoneme definiert werden müssen, ist die Unterstützung mehrerer Sprachen mit diesem System durchaus mit Schwierigkeiten behaftet.
Mit Precise bestehen zwei verschiedene Möglichkeiten, ein Aktivierungswort festzulegen. Einerseits werden vier verschiedene Aktivierungswörter standardmäßig angeboten. Dabei handelt es sich um „Hey Mycroft“, „Hey Ezra“, „Christopher“ sowie „Hey Jarvis“. Diese Modelle werden dabei ständig verbessert, in dem Nutzer explizit ihre Zustimmung zur Weiterverwendung dieser Daten geben. Andererseits ist es möglich, eigene Wörter zu trainieren. Dabei wird durch den Hersteller darauf hingewiesen, dass diese mindestens drei Silben umfassen sollen, weshalb die Standardsignalwörter mehrheitlich aus zwei Wörtern bestehen. [4]
Nach einer erfolgreichen Erkennung des Aktivierungswortes wird ein Signalton abgespielt. Dieser signalisiert dem Nutzer, dass das Gerät bereit ist und Befehle entgegen nimmt.

Sprache-zu-Text Umwandlung Wie in Kapitel 3.2.1 beschrieben, können mit Mycroft mehrere Systeme für STT eingesetzt werden. Von größerem Interesse sind dabei das Open-Source Sytem Kaldi sowie das von Mozilla entwickelte System DeepSpeech. Beide legen in ihrer Funktionalität Wert auf Datenschutz.
Aufgrund der konstanten Weiterentwicklung und Unterstützung vieler Sprachen, bietet sich Mozillas DeepSpeech am meisten an. Durch die Vielsprachigkeit kann das System jederzeit bei Bedarf an eine andere Sprache angepasst werden, ohne dass etwas in der Programmlogik verändert werden muss. Außerdem verspricht die konstante Weiterentwicklung auch immer genauere Ergebnisse, ohne sich in Abhängigkeit eines bestimmten Herstellers zu begeben.
Standardmäßig wird durch Mycroft im September 2019 zwar Google STT benutzt. Allerdings weißt die Nutzung eine Besonderheit gegenüber anderen Herstellern auf: alle Anfragen werden über einen Server von Mycroft versendet. Damit kann durch Google eine Anfrage nicht mehr einem spezifischen Nutzer zugeordnet werden, sondern nur noch der Menge an Mycroft Nutzern. Dies erhöht den Datenschutz maßgeblich. [5]
Mycroft behält sich vor, sämtliche Anfragen auf eigener DeepSpeech Infrastruktur durchzuführen. Aktuell wird durch Mycroft die Präzision von Google STT als bedeutend höher eingeschätzt, als die von Deepspeech. Aus diesem Grund wird im Folgenden Google STT via der durch Mycroft zur Verfügung gestellten API verwendet. Da die Weiterleitung der Anfragen an einen Service für Sprache-zu-Text Umwandlung in Händen von Mycroft liegt und mittels einer API angesprochen wird, hat eine mögliche Umstellung keinen Einfluss auf die Funktionsfähigkeit des

[4] https://github.com/MycroftAI/mycroft-precise/wiki/Training-your-own-wake-word [Abgerufen am 19.06.2019]
[5] https://cutt.ly/CeyEYx7 (Mycroft Roadmap: Speech-To-Text, Google Doc) [Abgerufen am 01.10.2019]

Systems. Es ist davon auszugehen, dass die Umstellung erst dann erfolgt, wenn DeepSpeech eine ausreichende Genauigkeit aufweist, weshalb dies zu keiner messbaren Beeinträchtigung der Nutzererfahrung führen sollte.

Intent Paser Für die Erkennung der Nutzerintention kann sowohl Adapt als auch Padatious eingesetzt werden. Beide sind offizielle Entwicklungen von Mycroft. Da Padatious keine starre Struktur der Aussagen benötigt, sondern flexibel auf diese reagieren kann, verspricht es prinzipiell bessere Ergebnisse. Allerdings ist dieses System mit Stand September 2019 noch in der Entwicklung [6], so dass noch keine zuverlässige Funktionsweise garantiert werden kann. Deshalb wird im Folgenden Adapt eingesetzt. Dieses System zeichnet sich besonders durch seine Leichtgewichtigkeit, Zuverlässigkeit und Geschwindigkeit aus. Es bedarf zwar eines starren Vokabulars, welches aber aufgrund der relativ geringen Komplexität der Aufgaben in den Einsatzszenarien (siehe Kapitel 3.4) kein Problem darstellt. Jedoch ist die Wahrscheinlichkeit, dass ein Nutzer möglichst natürlich mit dem System kommunizieren kann, mit Padatious größer. Dies liegt daran, dass dieser auf Basis eines neuronalen Netzes funktioniert und somit keine feste Struktur der Aussagen nötig ist.

Text-zu-Sprache Umwandlung Durch Mycroft werden verschiedene Systeme für die Text-zu-Sprache Umwandlung unterstützt, aufgrund der datenschutzrechtlichen Bedenken sind jedoch nur die beiden Umsetzungen des Unternehmens von Interesse. Dabei handelt es sich um Mimic und seinen Nachfolger Mimic II.
Mit Mimic klingt die Sprachausgabe sehr stark nach Maschine, allerdings wird sie auch direkt auf dem Gerät ausgeführt. Dem Gegenüber steht Mimic II mit einer natürlicheren Sprachfarbe und der Unterstützung verschiedener Sprachen. Damit sind Anpassungen an das System im Falle ohne größere Problem den Einsatz einer anderen Sprache als Englisch. Aufgrund dieses Umstandes und dem natürlicheren Klang eignet sich Mimic II besser für die Erfüllung der zuvor formulierten Anforderungen.

Übersicht der Bestandteile Eine Übersicht über die einzelnen Bestandteile von Mycroft die am besten geeignet sind, um die Anforderungen an das Konzept erfüllen, sind in Tabelle 4.1 aufgeführt.

[6] https://mycroft-ai.gitbook.io/docs/mycroft-technologies/overview [Abgerufen am 01.09.2019]

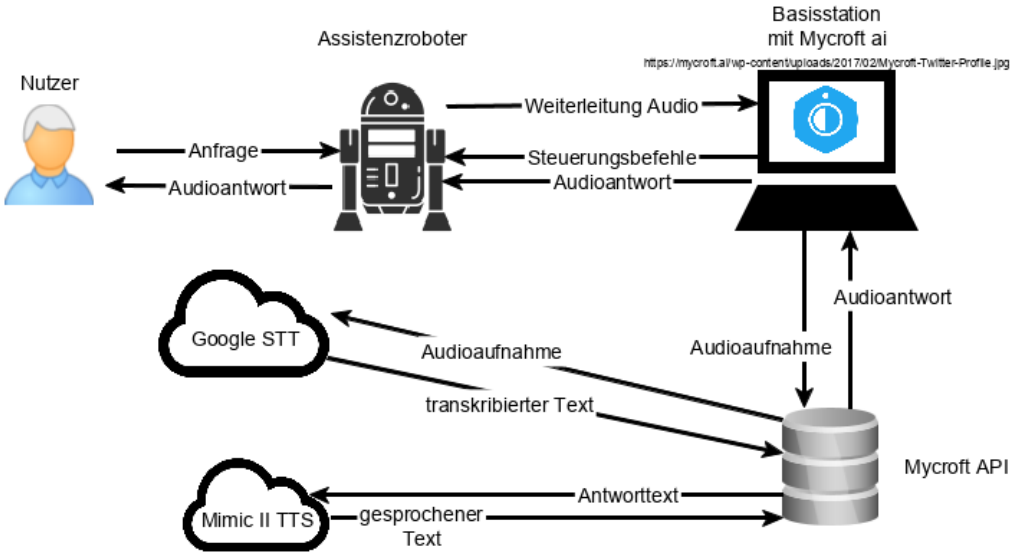

Abb. 4.1: Schematischer Ablauf der Interaktion

Sprachassistenzsystem	Mycroft Ai 19.02
Aktivierungsworterkennung	Precise
Sprache-zu-Text Umwandlung	Google STT via Mycroft Proxy
Intent Parser	Adapt
Text-zu-Sprache Umwandlung	Mimic II TTS

Tab. 4.1: gewählte Systembestandteile für Sprachassistenz

4.4 Erstellung des Konzepts

Ziel des Konzeptes ist es, die Anforderungen aus Kapitel 4.1 möglichst gut umzusetzen. Dabei wurden die datenschutzrechtlichen Bedingungen bereits durch die Auswahl der einzelnen Komponenten im vorherigen Kapitel berücksichtigt und erfüllt. Das Konzept umfasst die Abläufe der Kommunikation, hat aber keinen Einfluss auf die die Infrastruktur (z.B. WLAN) am Einsatzort oder die Fähigkeit des Roboters zur Fortbewegung.
Ziel des Konzeptes ist es, auch für andere Einsatzszenarien als die in 3.4 beschriebenen nutzbar zu sein. Der schematische Ablauf der Interaktion wird in Abbildung 4.1 dargestellt. Daraus ist ersichtlich, dass der Roboter als eine Art bewegliches Mikrofon mit Lautsprecher eingesetzt wird. Der Nutzer interagiert zwar mit dem Roboter, die Verarbeitung dieser Aktion findet jedoch auf einem separaten Gerät statt, welches im Folgenden als Basisstation bezeichnet wird. Diese beinhaltet das Sprachassistenz-

system und schickt passende Befehle mittels einer drahtlosen Netzwerkverbindung an den Roboter. Außerdem ist die Basisstation mit dem Internet verbunden, damit die in Abschnitt 4.3 gewählten Bestandteile des Sprachassistenten nutzbar sind.
Diese Unterteilung ist notwendig, da Mycroft aktuell nur auf Linux Distributionen problemlos eingesetzt werden kann. Da aber nicht davon ausgegangen werden kann, dass dieses Betriebssystem auf allen Assistenzrobotern benutzt wird, können beliebige Roboter eingesetzt werden. Trotzdem soll der Nutzer das Gefühl haben, nur mit dem Roboter zu interagieren. Ein festes Mikrofon an der Basisstation hätte zur Folge, dass der Nutzer möglicherweise akustisch nicht verstanden wird. Alternativ gäbe es die Möglichkeit, weitere Mikrofone am Einsatzort aufzustellen, was aber einen nicht vertretbaren zusätzlichen Aufwand darstellen würde. Gleiches gilt für den Einsatz stationärer Lautsprecher.
Wie in Kapitel 2.1 erläutert, folgt die Interaktion mit einem Sprachassistenten immer gewissen grundlegenden Abläufen. Dafür muss der Nutzer zuerst das Aktivierungswort benutzen und erst daran anschließend die Phrase, die den eigentlichen Befehl beinhaltet. Bei kommerziellen Sprachassistenten, wie Alexa, erfolgt die Verarbeitung des Aktivierungswortes in der lokalen Einheit und erst die daran angeschlossene Interaktion wird in der Cloud analysiert (vgl. Abbildung 3.3). Im Rahmen dieses Konzepts werden die Audiodaten konstant durch den Roboter an die Basisstation versendet. Erst diese analysiert die Eingabe auf ein Aktivierungswort und leitet bei Bedarf die anschließenden Daten an die API von Mycroft weiter.
Bei der Umwandlung von Sprache zu Text kann es zu Ungenauigkeiten kommen, die zur Ausführung eines falschen Befehls führen können. Deshalb soll dem Nutzer eine Möglichkeit gegeben werden, Handlungen abzubrechen, bevor sie inkorrekt ausgeführt werden. Eine mögliche Abfolge wurde dafür in Kapitel 2.4 erwähnt. Dieses Konzept wurde von Green et al. [Gre+00] vorgestellt und das hier Erstellte ist daran angelehnt. Dargestellt wird es in Abbildung 4.2. Der größte Unterschied zu dem Konzept von Green ist, dass der Roboter keine Geste ausführt, um einen Befehl zu bestätigen, da der Roboter nicht zwangsläufig über entsprechende Extremitäten verfügt. Außerdem ist keine explizite Bestätigung der Befehle durch den Nutzer nötig.
Damit der Nutzer darüber in Kenntnis gesetzt wird, welcher Befehl verstanden wurde, wird dieser vom Roboter vor der Ausführung wiederholt. Sollte der Nutzer mit diesem einverstanden sein, reicht eine implizite Bestätigung (Abb. 4.2a). Es wird dabei davon ausgegangen, dass der Nutzer der Durchführung zustimmt, wenn er nicht explizit widerspricht.
Im Falle einer Ablehnung hat der Nutzer die Möglichkeit, direkt nach der Wiederholung des Befehls durch ein einfaches „Stop“ die Durchführung abzubrechen. Daraufhin bittet der Roboter um eine erneute Eingabe des Befehls. Dabei handelt es sich um eine durchgängige Interaktion, wodurch es nicht der erneuten Benutzung des Aktivierungswortes bedarf. Beispielhaft ist ein solcher Ablauf in Abbildung 4.2b dargestellt, dabei versteht der Roboter statt „Kaffee“ das Wort „Tee“.

```
Nutzer:  Roboter, hole Kaffee aus der Küche!
Roboter: Hole Kaffee aus der Küche.
# Roboter wartet kurz
# Roboter führt Handlung durch
```

(a) Roboter hat Anweisung korrekt verstanden und führt sie aus

```
Nutzer:  Roboter, hole Kaffee aus der Küche!
Roboter: Hole Tee aus der Küche.
Nutzer:  Stop!
Roboter: Leider habe ich etwas falsch verstanden.
         Bitte wiederholen Sie den Befehl.
```

(b) Roboter versteht Anweisung falsch und Nutzer bricht sie vor Ausführungsbeginn ab

Abb. 4.2: implizite Bestätigung und explizite Ablehnung des Befehls

Wurde der Befehl durch den Nutzer (implizit) bestätigt, wird durch Mycroft eine entsprechende Steuerungsnachricht an den Roboter versendet. Dafür wird, wie auch für die bisherige Kommunikation, ein Socket benutzt. Damit die Analyse der Befehle universell und ohne Schwierigkeiten möglich ist, werden sie in Form von JSON Nachrichten verschickt.
Außerdem soll der Nutzer die Aktion des Roboters jederzeit abbrechen können, zum Beispiel wenn diese fehlhaft ausgeführt wird oder nicht mehr notwendig ist. Um das Beispiel aus Abbildung 4.2 aufzugreifen, könnte fehlerhafter Weise einen Teepackung als Kaffepackung erkannt wird.
Außerdem ist es wichtig, dass der Roboter auf erkannte Personen reagieren kann. Dabei bietet es sich an, aktuelle Umgebungsbilder auf die Existenz von Menschen zu untersuchen. Wenn eine Person erkannt wurde, soll der Roboter dieser seine Hilfe anbieten. Dabei kann die Person direkt eine Interaktion starten, ohne das ein Aktivierungswort benötigt wird. Dargestellt ist der Ablauf für diesen Fall in Abbildung 4.3.
Nicht zu vernachlässigen ist auch die Ausdrucksweise des Roboters. Diese sollte immer höflich sein, auch wenn sich Dialoge dadurch verlängern. Dieses führt zu einem positiven Nutzergefühl. Im Gegensatz dazu wird bei sehr kurzen Antworten des Roboters der Nutzer ständig daran erinniert, dass er mit einer Maschine kommuniziert.
In Abbildung 4.4 werden die einzelnen Schritte der Kommunikation mittels eines Sequenzdiagramms dargestellt. Darin werden die Abhängigkeiten der einzelnen Aktionen von den jeweiligen Verarbeitungen durch einen Teilnehmer der Interaktion veranschaulicht.

```
# Roboter erkennt Person
Roboter: Wie kann ich Ihnen behilflich sein?
Nutzer:  Bringe mir einen Kaffee
Roboter: Bringe Kaffee
# Roboter wartet kurz
# Roboter führt Handlung durch
```

Abb. 4.3: Dialog zwischen Mensch und Roboter auf Initiative des Roboters

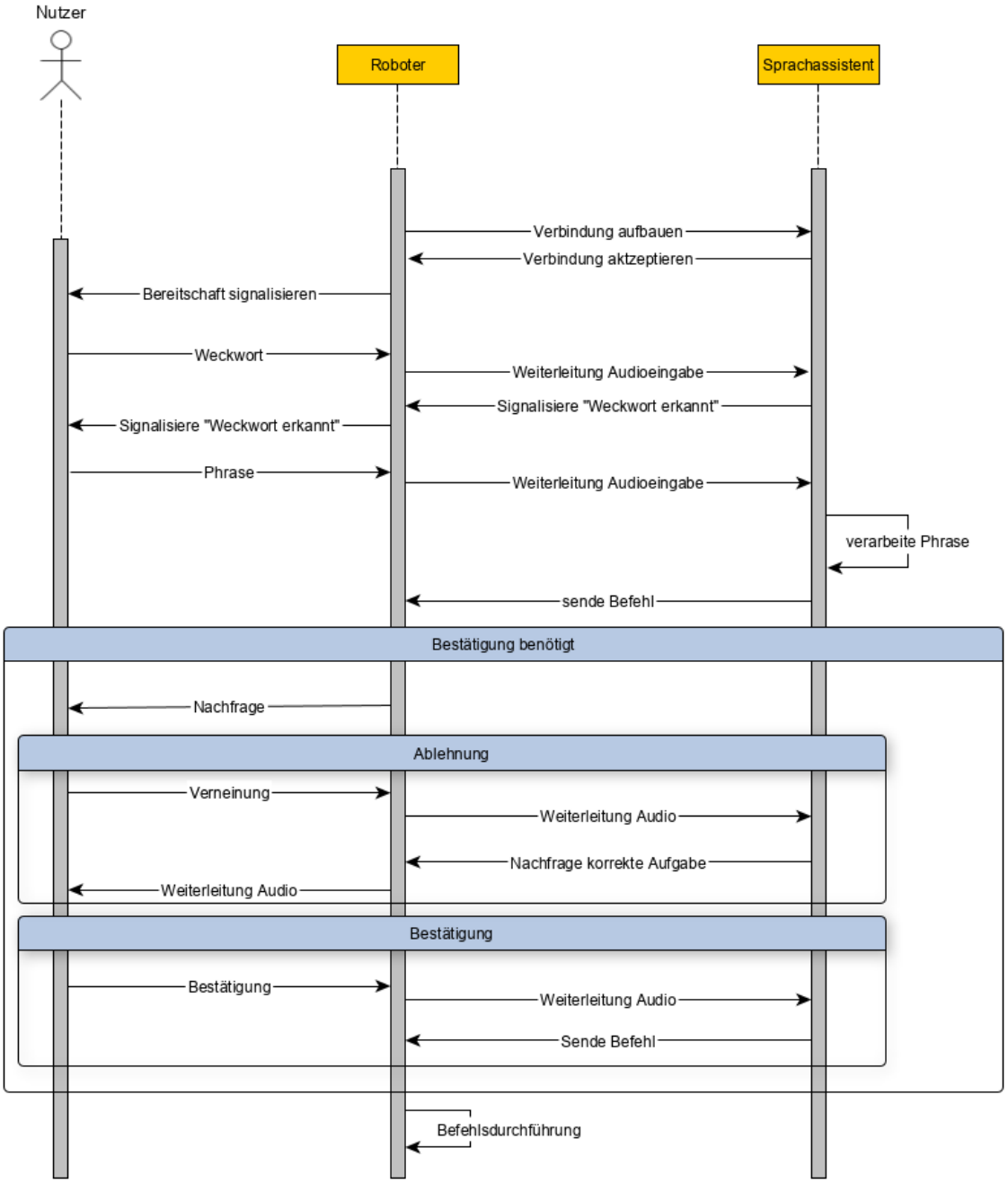

Abb. 4.4: Sequenzdiagramm der Interaktion ANPASSUNG SEQUENZDIAGRAMM??

4.5 Zusammenfassung

In diesem Kapitel wurde zuerst auf Basis der Anforderungen ein geeignetes Sprachassistenzsystem gewählt. Die Wahl ist dabei auf Mycroft gefallen, da dieses System vollständig quelloffen ist und sich durch eine starke Modularität auszeichnet. Anschließend wurde die Umsetzungen der einzelnen Verarbeitungsschritte so gewählt, dass sie den Anforderungen bestmöglich gerecht werden. Darauf aufbauend konnte ein Konzept erstellt werden, das den Einsatz eines Sprachassistenten gemeinsam mit einem Assistenzroboter erlaubt.

5 Prototypische Umsetzung des Konzepts

Ziel dieses Kapitels ist es, das im vorherigen Teil dieser Arbeit erstellte Konzept umzusetzen. Dafür wird ein zweirädriger, selbst-balancierender Roboter von Segway Robotics eingesetzt. Um die Entwicklung einer Anwendung mit diesem Roboter zusammen mit Mycroft zu ermöglichen, wird zuerst betrachtet, wie diese beiden Systeme im Speziellen aufgebaut sind und wie sie miteinander kombiniert werden können. Abschließend wird die Funktionsweise des entstandenen Prototyps genauer erläutert.

5.1 Eingesetzte Hardware

Für die Erstellung der prototypischen Anwendung wird der Roboter Loomo von Segway Robotics eingesetzt. Dieser ist in Abbildung 5.1 dargestellt. Er verfügt über zwei Räder sowie eine zentrale Einheit, die sowohl Sensoren als auch die Recheneinheit umfasst. Dabei werden die Berechnungen mittels eines Intel Atom Z8750, mit 4 Kernen die jeweils eine Taktrate von 2,4 GHz erreichen, durchgeführt. Der Kopf des Roboters lässt sich dabei unabhängig vom restlichen Roboter bewegen. Die Rotation ist horizontal und vertikal möglich. In diesem Kopf befinden sich ein LCD Touchbildschirm sowie eine HD Kamera. Der Bildschirm kann dabei sowohl für Ein- als auch Ausgaben verwendet werden. Unterhalb des Kopfes befindet sich außerdem ein Array Mikrofon, mit dem prinzipiell die Richtung der Geräusche festgestellt werden kann. Zudem ist es möglich, mittels eines Lautsprechers Soundausgaben zu erzeugen. [1] Es ist außerdem möglich, zusätzliche Hardware, wie beispielsweise Arme, mittels des „Hardware Extension Bays“ an den Roboter anzubringen.
Wie in Kapitel 4 erläutert, wird der Roboter mit der Sprachassistenzsoftware Mycroft interagieren. Diese wird dabei auf einem Laptop mit folgender Konfiguration zur Verfügung gestellt:

- CPU: Intel Core i7-5600 @ 2,60 GHz
- Arbeitsspeicher: 8 GB DDR3
- Betriebssystem: Manjaro Linux KDE 18.04

[1] https://developer.segwayrobotics.com/developer/documents/segway-robot-overview.html [Abgerufen am 10.07.2019]

Abb. 5.1: Loomo Assistenzroboter [2]

Der Laptop fungiert als Basisstation für die Interaktion. Das heißt, dass auf ihm Mycroft läuft, welches entsprechend Kapitel 4.3 konfiguriert ist. Für die Interaktion sind beide Geräte mit einem WLAN-Netz verbunden, welches auch eine Verbindung mit dem Internet erlaubt und es ermöglicht, nach Bedarf Ports freizugeben.

5.2 Besonderheiten in der Entwicklung

In diesem Kapitel werden einige Besonderheiten der beiden Teilsysteme betrachtet, die bei der Entwicklung mit dem jeweiligen System zu beachten sind. Zu berücksichtigen ist dabei, dass beide auf Basis verschiedener Programmiersprachen agieren. Während Loomo mit Android 5.1 (API 22, ohne Play Services) betrieben wird, nutzt Mycroft Python 3 mit einigen extra Packages.

5.2.1 Entwicklung von Anwendungen mit Loomo

Der Roboter ist so konfiguriert, dass einzelne Bestandteile mittels eines, vom Hersteller zu Verfügung gestellten, Software Development Kit (SDK) genutzt werden können. Die benötigten Teile des SDK werden dafür mittels Dependencies in der Gradle Datei der Android Anwendung importiert.
Folgende Teile können eingesetzt werden: [3]

[2] https://www.indiegogo.com/projects/loomo-mini-transporter-meets-robot-sidekick#/ [Abgerufen am 02.10.2019]
[3] https://developer.segwayrobotics.com/developer/documents/segway-robots-sdk.html [Abgerufen am 02.10.2019]

- **Vision** erlaubt Zugriff auf die Kameras (HD, Intel Real Sense, Fischauge), beinhaltet außerdem ein detection-tracking system (DTS) , mit dem Personen erkannt und verfolgt werden könnn
- **Speech** ermöglicht Zugriff auf das Array Mikrofon und ein Speaker Modul inklusive Loomo eigener TTS
- **Locomotion** dient Steuerung der Fortbewegung (Base) und Kopfbewegung (Head)
- **Sensor** ermöglicht Zugriff auf die verschiedenen Sensordaten (z.B. aktuelle Ausrichtung, Bewegungsgeschwindigkeit)

5.2.2 Entwicklung mit Mycroft

Da sich Mycroft noch in aktiver Entwicklung befindet, ändern sich auch verschiedene Bestandteile der Implementierung bei Verbesserungen. Allerdings hat das System eine spezifische Grundstruktur, an der sich alle Weiterentwicklungen orientieren. Die verschiedenen Skills folgen auch einem festen Aufbau, der vorschreibt, auf welche Art die einzelnen Elemente zu strukturieren und abzulegen sind.

Android Da der Roboter Loomo auf Android basiert, gibt es prinzipiell auch die Überlegung, Mycroft auf dem Roboter zu installieren und somit verbindungsbedingten Problemen bei der Kommunikation zwischen Roboter und Basisstation vorzubeugen. Dies bringt aber zwei Probleme mit sich. Zum einen kann nicht davon ausgegangen werden, dass Assistenzrobotern prinzipiell mit Android betrieben werden. Damit wäre somit keine valide Überprüfung des Konzepts mittels des Prototyps möglich.
Außerdem wird aktuell Android nicht offiziell von Mycroft unterstützt. Es existieren zwar zwei Implementierungen, jedoch sind beide durch die Community entstanden. [4]

Die sogenannte Companion App greift für die STT und TTS Umwandlung jeweils direkt auf die Google APIs zurück und interagiert mit der Basisstation in Form von JavaScript Object Notation (JSON) Nachrichten.
Die andere Implementierung versucht, die Core Implementierungen komplett auf Androidbasis umzustellen [5]. Jedoch sind eigene Tests mit dieser Anwendungen immer an verschiedenen Fehlern unbekannter Ursache gescheitert.

[4] https://mycroft-ai.gitbook.io/docs/using-mycroft-ai/get-mycroft/android [Abgerufen am 05.10.2019]
[5] https://github.com/MycroftAI/MycroftCore-Android [Abgerufen am 20.05.2019]

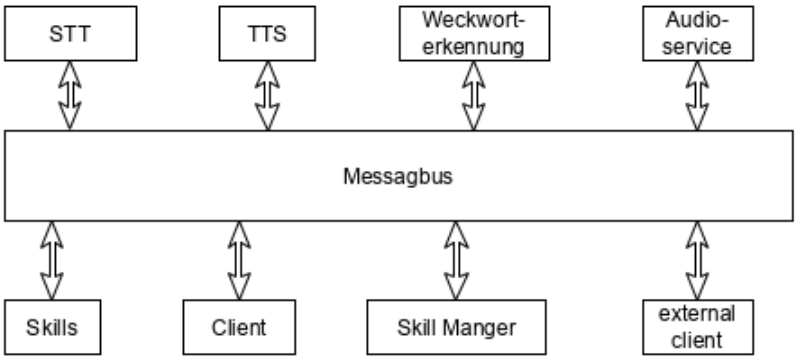

Abb. 5.2: schematische Darstellung der Funktionsweise des Messagebus

Übersicht der Architekturdetails Im Mittelpunkt der Architektur von Mycroft steht der sogenannte Messagebus [6]. Dieser erlaubt die Kommunikation zwischen den einzelnen Bestandteilen der Anwendung. Die Nachrichten werden in Form von JSON Nachrichten ausgetauscht. Außerdem ist es möglich, dass sich weitere Anwendungen mit diesem Bus verbinden, indem sie eine Verbindung zu dem entsprechenden Websocket aufbauen. Ab dem Moment der erfolgreichen Verbindung können sowohl Nachrichten gelesen als auch geschrieben werden. Schematisch ist der Messagebus in Abbildung 5.2 dargestellt.

Entwicklung von Skills Einen entscheidenden Teil der Entwicklung nimmt die Erstellung von eigenen Skills für Mycroft ein. Damit erhält der Sprachassistent neue Fähigkeiten, die entsprechend selbst festgelegter Aussagen ausgelöst werden. Jeder einzelne Skill hat eine definierte Ordnerstruktur, in der die entsprechenden Bestandteile untergebracht werden. Diese sind:

- **dialog** umfasst jede mögliche Antwort, die Mycroft nach der Ausführung des Skills geben kann
- **vocab** definiert die Intents, die Reaktion von Mycroft hervorrufen. Dies setzt die Festlegung von Vokabeln vorraus, die in der Phrase vorkommen müssen damit der Skill aktiviert wird.
- **regex** optional erlaubt die Definition von regulären Ausdrücken, die in Intents vorkommen können
- **init.py** Python Datei, in der Aufrufe der Intents behandelt und verarbeitet werden

Systemkonfiguration Für die Nutzung von Mycroft wird ein Nutzerkonto benötigt. Nach der erfolgreichen Installation der Software wird diese mittels eines Codes

[6] https://mycroft-ai.gitbook.io/docs/mycroft-technologies/mycroft-core/message-bus [Abgerufen am 04.10.2019]

mit dem Nutzerkonto verknüpft. Das erlaubt, über die Webseite Einstellungen vorzunehmen. Diese umfassen beispielsweise die Angabe eines Standortes, des Aktivierungswortes oder der Stimme. Alternativ kann das System auch mittels einer Konfigurationsdatei angepasst werden. Die einzelnen Systembestandteile können auch individuell festgelegt werden. Dabei wird lokal die Remotekonfiguration überschrieben, insofern diese voneinander abweichen. [7]

5.3 Prototyp

Der Prototyp legt seinen Fokus auf die Umsetzung des Konzepts zur natürlichsprachlichen Interaktion mit einem Roboter (vgl. Kapitel 4.4). Aus diesem Grund werden Funktionen, denen tiefergehende Konzepte anderer Art zugrunde liegen, nur als „Dummy“ Funktionen implementiert. Diese simulieren ein entsprechendes Verhalten nur in den Grundzügen. Ein Beispiel ist der Befehl „Hole mir Kaffee“. Dafür müsste der Roboter seine Umgebung kartiert haben, sich auf der Karte orientieren können und in der Lage sein, Hindernisse zu erkennen. Außerdem bräuchte er eine Objekterkennung, die Gegenstände, in diesem Fall Kaffee, eindeutig erkennt. Zusätzlich wären noch Extremitäten nötig, mit denen er diesen greifen kann. Damit erkennbar ist, wie das entsprechende Verhalten aussehen könnte, fährt der Roboter in diesem Fall ein Stück nach vorn, dreht sich um 180 Grad und kommt wieder zurück.
Die Implementierung des Sprachassistenten wurde auf Englisch durchgeführt, die Interaktion also nur in dieser Sprache stattfinden. Der Grund ist, dass Mycroft von einem Unternehmen aus den Vereinigten Staaten entwickelt wird und auch die Community auf Englisch kommuniziert. Dies verspricht die besten Ergebnisse für die prototypische Umsetzung, insbesondere im Bezug auf die Umwandlung von Sprache zu Text.
Als Aktivierungswort wurde die Standardphrase „Hey Mycroft“ belassen. Das Training eines eigenen Aktivierungswortes hatte geringere Aussichten auf Erfolg, da für gute Trainingsergebnisse sowohl korrekte als auch falsche Aussprachebeispiel benötigt werden. Damit soll die fehlerhafte Erkennung des Wortes minimiert werden. Die anderen, durch Microft angebotenene Signalwörter, haben sich nach Tests als weniger zuverlässig erwiesen.

5.3.1 Implementierte Funktionen

Damit aussagekräftige Tests mit dem Roboter durchgeführt werden können, muss dieser verschiedene Befehle ausführen können. Da die verschiedenen Bewegungen

[7] https://mycroft-ai.gitbook.io/docs/using-mycroft-ai/customizations/mycroft-conf [Abgerufen am 02.10.2019]

aus einer Kombination von einfacheren Befehlen besteht, wurden folgende Grundfunktionen umgesetzt. Die Drehung um 90 Grad nach links beziehungsweise rechts sowie um 180 Grad um die eigene Achse. Eine weitere Basisfunktion ist die lineare Fortbewegung. Bei den Drehungen dreht sich der Roboterkopf mit dem Körper mit, so dass er immer nach vorne schaut.
Außerdem kann der Roboter beauftragt werden, dem Nutzer einen Gegenstand zu bringen. Aktuell wird auf „coffee“, „tea“, „milk“ und „water“ reagiert. Die daraus resultierende Handlung des Roboters ist aber immer identisch, da es sich hierbei um eine Dummy Funktion handelt. Der Roboter fährt lediglich ein Stück nach vorne, dreht sich um seine eigene Achse und fährt die selbe Strecke wieder zurück. Bevor die Aktion ausgeführt wird, bekommt der Nutzer die Frage gestellt, ob der Roboter den gewünschten Gegenstand korrekt verstanden hat.
Es ist auch möglich, den Roboter in die Küche zu schicken. Dies ist ebenfalls nur eine Dummy Implementierung. Dafür fährt der Robter ein Stück geradeaus und sagt sein Ziel an. Auf diese Art besteht für den Nutzer noch die Möglichkeit, die Ausführung abzubrechen.
Weitere Funktionen sind: einem Hindernis ausweichen und zum Nutzer zurückkommen. Bei letzterem wird davon ausgegangen, dass sich der Roboter vor Ausführung des letzten Befehls vor dem Nutzer befunden hat. Deshalb wird dieser Befehl einfach in entgegengesetzter Richtung ausgeführt.
Auf manuelle Aktivierung ist es auch möglich, dass mittels der HD Kamera eine Personenerkennung durchgeführt wird. Dafür wird auf die Personenerkennung des DTS der Loomo SDK zurückgegriffen. Wenn eine Person erkannt wurde, wird dieser Hilfe angeboten. Daraufhin kann ein entsprechender Befehl geäußert werden, ohne dass das Aktivierungswort benötigt wird. Zu Testzwecken wird eine erkannte Person, unabhängig von der vorherigen Antwort, nach 30 Sekunden erneut gefragt, ob ihr geholfen werden kann.
Um im Rahmen der Studie überprüfen zu können, ob die in Kapitel 4.4 erarbeitete Art der Bestätigung einer Handlung von Nutzern als sinnvoll eingeschätzt wird, wurde zusätzlich eine Nachfrage implementiert. Dies bedeutet, dass der Roboter bei einigen Aktionen (z.B. zu einem definierten Ort fahren) dem Nutzer mitteilt, an welchen Ort er fahren wird. Bei Befehlen, der Inhalt ist, dass der Roboter einen bestimmten Gegenstand zu dem Nutzer bringt, wird der Nutzer gefragt, ob die verstandene Handlung korrekt ist. Diese Abweichung von dem in Kapitel 4.4 erstellten Konzept wurde bewusst vollzogen. Dadurch ist es möglich, die Meinung von Nutzern im Rahmen der Studie einzuholen und zu überprüfen, ob die bei der Konzepterstellung angestellten Überlegungen korrekt sind.
Es ist außerdem möglich, alle Aktionen zu jeden Zeitpunkt abzubrechen. Dafür bedarf es zuerst der Nutzung des Signalwortes, damit der Sprachassistent in einen aktiven Zustand versetzt wird. Anschließend reicht ein „Stop“ für den Abbruch der Handlung aus. Dieser Ablauf ist sinnvoll, da eine dauerhaft aktivierte Verarbeitung der Audioaufnahmen zu häufig zu Missverständnissen und dadurch zu unerwarteten

Handlungen des Roboters führt.
In Tabelle 5.1 werden die verschiedenen Aktionen und jeweils eine mögliche Phrase aufgeführt. Für jeden dieser Befehle musste das entsprechende Vokabular festgelegt werden, mit dem dieser ausgelöst wird. Dabei ist anzumerken, dass möglicherweise korrekte Aussagen von dem Sprachassistenten nicht verstanden werden, da sie sich nicht in dem definierten Vokabular befinden.

Aktion	**mögliche Phrase**
Drehung	turn left/right/around
zurück kommen	comeback
Gegenstand holen	get coffee/tea/milk/water
zu Ort fahren	go to kitchen/door
aus dem Weg fahren	out of my way
Aktion abbrechen	stop

Tab. 5.1: Funktionen und mögliche auslösende Phrasen

5.3.2 Kommunikation zwischen Roboter und Sprachassistent

Für die Kommunikation zwischen dem Roboter und dem Sprachassistenten wird auf Sockets zurückgegriffen. Zum einen wird der, in Kapitel 5.2 vorgestellte, Messagebus eingesetzt,über den Steuerungsbefehle in Loomo versendet werden. Diese sind JSON Nachrichten, die sowohl den Typ des Befehls („loomoInstruction“) als auch die die auszuführende Handlung beinhalten. Außerdem versendet Loomo über diesen Bus die Nachricht, falls er eine Person entdeckt hat.
Weiterhin befinden sich noch zwei weitere Sockets im Einsatz, welche beide in separaten Threads benutzt werden, damit sie die Ausführung weiterer Aktivitäten nicht behindern. Die Entscheidung, mehrere Sockets zu benutzen, wurde bewusst getroffen. Auf diese Weise kann mit absouluter Sicherheit ausgeschlossen werden, dass sich die Datenströme gegenseitig behindern.
Einer der Sockets wird durch Mycroft initiiert und dient als Eingang des Mikrofonsignals. Einen weiteren initiiert Loomo, um Audio Dateien zu erhalten, die als Ergebnis der Sprachausgabe abgespielt werden sollen.

5.3.3 Architekturdetails

Für die Architektur des Prototypen ist es notwendig, den Sprachassistenten und den Roboter getrennt zu betrachten, da es sich um jeweils eigenständige Anwendungen handelt.

Mycroft In Abbildung 5.3 wird der Aufbau von Mycroft mittels eines Klassendiagrammes dargestellt. Aus Gründen der Übersichtlichkeit werden nur die wichtigsten Klassen abgebildet. Die dabei vorgenommenen Einfärbungen dienen dem besseren Verständnis. In manchen Fällen, beispielsweise TTS, ist es möglich, zwischen dem Einsatz verschiedener Klassen auszuwählen. Die dabei gewählte Klasse ist jeweils blau umrahmt. Eigens für diese Anwendung erstellte Klassen wiederum sind in orange eingefärbt.
Eine dieser beiden Klassen ist `socketServer` und läuft in Form eines Threads ununterbrochen im Hintergrund. Diese erhält die Mikrofoneingabe des Roboters in Form eines Bytestreams. Dafür wird das Pythonpackages `PyAudio` [8] verwendet. Dieses ermöglicht den Einsatz von `PortAudio` [9]. Dabei handelt es sich um API, die zur Aufnahme und Abspielen von Audiodateien benutzt werden kann. Da es mit Mycroft nicht möglich ist, einen eigenen Audiostream direkt als Mikrofonsignal zu verwenden, muss für die Verwendung der Aufnahmen ein Umweg gegangen werden. Dazu ist es nötig, das Signal auf dem Host mittels PyAudio auszugeben. Gleichzeitig kann Linux so konfiguriert werden, dass eine Audioausgabe auf das Mikrofon umgeleitet wird. Somit ist es möglich, die Mikrofondaten zu verwenden, ohne die tiefere Programmlogik von Mycroft anpassen zu müssen.
Die andere hinzugefügte Klasse nennt sich `fileOpener`. Diese verbindet sich dann mit dem Roboter, wenn es durch die TTS Umwandlung zu einer Audioausgabe kommt. Sie wird in Form einer Datei im Waveformat lokal abgespeichert und normalerweise über die Lautsprecher der Verarbeitungseinheit ausgegeben. Durch eine Anpassung wird dieser Schritt durch die fileOpener Klasse durchgeführt. Anstelle einer Audioausgabe wird die Datei dafür über den Socket zum Roboter versendet und entsprechend ausgegeben.

[8] https://people.csail.mit.edu/hubert/pyaudio/docs/ [Abgerufen am 05.10.2019]
[9] http://www.portaudio.com/ [Abgerufen am 05.10.2019]

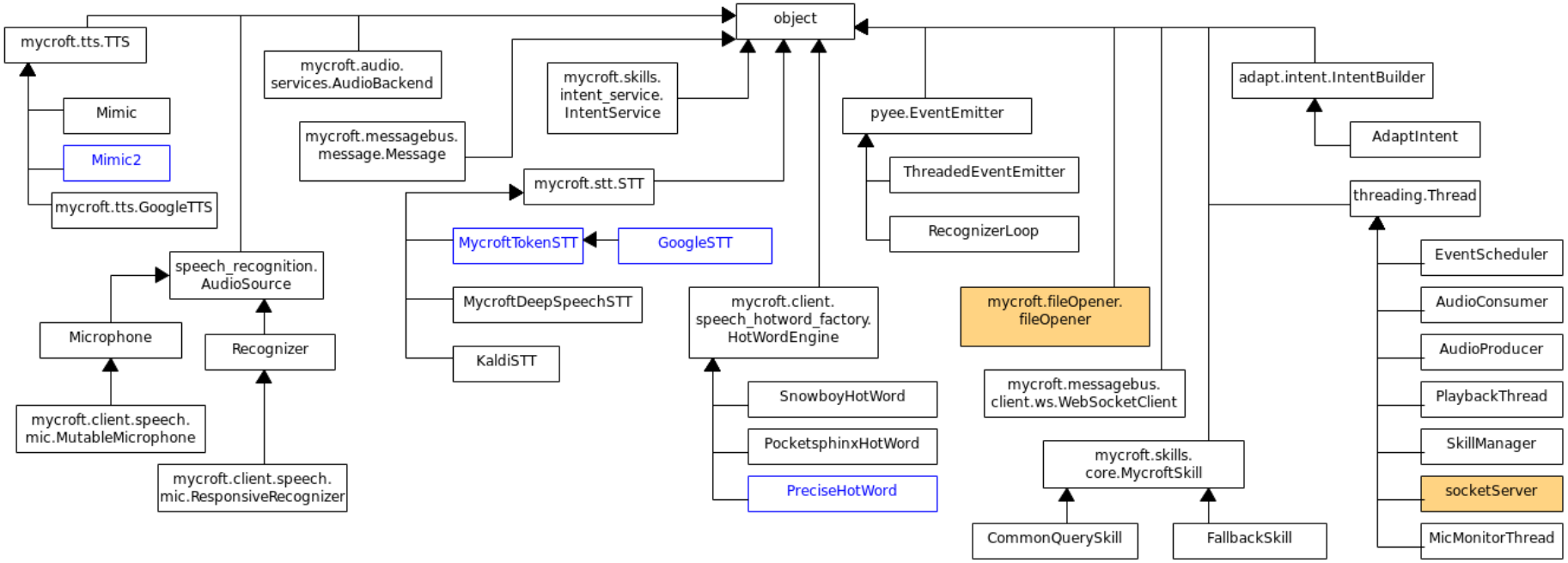

Abb. 5.3: Klassendiagramm von Mycroft, beschränkt auf die relevantesten Teile

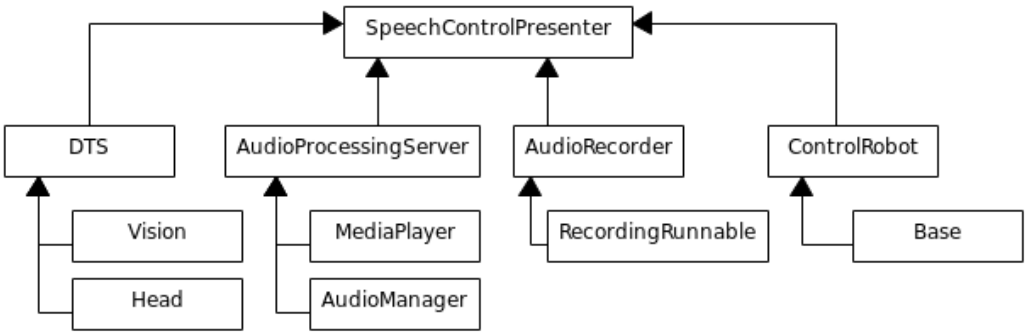

Abb. 5.4: schematische Darstellung der Androidanwendung

Loomo Der grundlegende Aufbau der Android Anwendung orientiert sich an einer Beispielanwendung, die durch Segway zur Verfügung gestellt wird [10]. In dieser ist bereits die Funktion zur Erkennung und Verfolgung von Menschen integriert, wobei im vorliegenden Fall Personen lediglich durch Kopfbwegungen verfolgt werden. Standardmäßig ist die Personenerkennung mittels des DTS deaktiviert, sie kann aber durch den Nutzer über den Bildschirm manuell aktiviert werden.
Die Basisapplikation wurde noch um Bestandteile für das Versenden und Empfangen der Audiodateien sowie eine Verbindung zu dem Websocket erweitert. Außerdem wurden die Funktionen für die Steuerung des Roboters mittels der Loomo SDK integriert. Eine schematische Darstellung des Aufbaus ist in Abbildung 5.4 zu sehen.
Für die Ausgabe der erhaltenen Audio Dateien wird der in Android integrierte MediaPlayer verwendet, der automatisch auf die Lautsprecher des Roboters zugreift. Während einer Audioausgabe wird mittels des AudioManagers das Mikrofon deaktiviert, um zu verhindern, dass die Ausgabe zeitgleich als Eingabe verwendet wird und somit zu falschen Ergebnisse führt.
Im Hintergrund bleibt die Verbindung zu Mycroft konstant erhalten, sodass der Roboter eine eingehende Nachricht im Wave Format umgehend ausgeben kann.
In der Entwicklung hat sich herausgestellt, dass es nicht möglich ist, die Array Mikrofone von Loomo als solche einzusetzen. Somit kann nicht festgestellt werden, aus welcher Richtung der Nutzer mit dem Roboter interagiert, da die Mikrofone nur als ein einziges verwendet werden können. Die Aufnahme der Geräusche geschieht in einem Thread, der RecordingRunnable, kontinuierlich im Hintergrund. Auf diesem Weg werden die Interaktionen mit dem Roboter nicht beeinträchtigt.
Für die Steuerung des Roboters kommt das von Segway zur Verfügung gestellte SDK zum Einsatz. Mittels des Base Paketes ist es möglich, den Roboter nach vorn oder zurück fahren zu lassen. Außerdem können damit Drehrichtungen festgelegt werden. Damit ein Befehl Steuerungsbefehl ausgeführt werden kann, benötigt die JSON Nachricht zusätzlich ein Datenfeld. Mit dieser Information wird anschließend die zu der Aktion passende Funktion aufgerufen.

[10] https://github.com/SegwayRoboticsSamples/FollowMeSample [Abgerufen am 17.07.2019]

5.4 Zusammenfassung

In diesem Kapitel wurden die Eigenschaften der prototypischen Umsetzung des Konzeptes genauer beschrieben. Als Roboter wurde dafür ein selbst-balancierender Roboter von Segway Robotics eingesetzt. Dieser basiert auf Android und kommuniziert über Sockets mit Mycroft. Dabei versendet Mycroft die Nachrichten an den Roboter über einen Messagebus, der auch für den Nachrichtenaustausch der anderen Systembestandteile dient. Damit die Nachrichten effektiv zwischen den beiden Systemteilen ausgetauscht werden können, werden Sockets verwendet, über die auch die Audiodateien versendet werden. Teilweise wurden die Funktionen des Prototypes so implementiert, dass sie lediglich das reale Verhalten simulieren.

6 Evaluation des Konzepts anhand des Prototyps

Schwerpunkt dieses Kapitels ist die Bewertung des Konzepts im Rahmen einer Pilotstudie. Dafür wurden Probanden mehrere Aufgaben gestellt, die sie mithilfe des Prototypen lösen sollten. Im Anschluss daran haben sie sich jeweils zu einigen quantitativen Fragen geäußert und ihre Eindrücke zusätzlich in einem kurzen Gespräch genauer beschrieben. Mithilfe der gesammelten Antworten ist es möglich, die Einsatzfähigkeit und Qualität des Konzepts zu bewerten. Zusätzlich können so auch Verbesserungspotentiale aufgezeigt werden.

6.1 Studiendesign

Die folgenden beiden Aussagen stehen im Fokus der Studie:

(a) Die Nutzung des Systems ist für den Nutzer selbsterklärend und vermittelt ein Gefühl von natürlicher Interaktion.

(b) Ein stärkerer Fokus auf den Datenschutz erhöht die Bereitschaft zur Nutzung von Sprachassistenten, auch wenn bislang kein solcher Assistent verwendet wird.

Dabei liegt der Schwerpunkt auf einer Bewertung der Aussage (a), welcher sich aus den Anforderungen in Kapitel 2.3 ergibt. Durch die Analyse von Aussage (b) soll es möglich sein, Rückschlüsse darauf zu ziehen, ob die Einhaltung des Datenschutzes ein System attraktiver macht.
Für die Durchführung der Aufgaben wurde den Probanden das System zuerst kurz vorgestellt. Dabei wurde explizit darauf hingewiesen, dass einige Funktionen das gewünschte Verhalten nur simulieren. Aus diesem Grund konnte durch die Studienteilnehmer nicht die Qualität der Aufgabendurchführung bewertet werden, sondern nur die Erkennung der Aufgaben. Ob diese jeweils korrekt durchgeführt wurde, soll für die Teilnehmer anhand der Reaktionen des Roboters eindeutig nachvollziehbar sein.
Ein Unterweisung in die Funktionsweise des Roboters wurde nicht durchgeführt, die Nutzer kannten lediglich das Signalwort zur Aktivierung des Sprachassistenten. Außerdem wussten sie, dass die Interaktion ausschließlich auf Englisch durchgeführt

werden konnte.
Damit die Antworten der Probanden vergleichbar sind, wurden jedem die selben Aufgaben vorgelegt, die in Kapitel 6.1.1 genauer vorgestellt werden. Damit die beiden eingangs aufgestellten Hypothesen bewertet werden können, wurden die Teilnehmer gebeten, einen Fragebogen auszufüllen sowie in einem Gespräch ihre Eindrücke genauer zu schildern. Damit eine verbesserte Grundlage für das Gespräch geschaffen werden konnte, wurden die Probanden außerdem darum gebeten, ihre Gedanken während der Durchführung der Aufgaben laut zu äußern.

6.1.1 Aufgabenstellung

Ziel der Aufgabenstellung ist es, jede der implementierten Funktionen zu abzudecken. Auf diese Weise bekommen die Teilnehmer den breitesten Eindruck der Reaktionen des Systems und es ist auch möglich, das Verhalten der Nutzer in verschiedenen Situationen zu bewerten. Um die Hypothese (a) möglichst gut bewerten zu können, wurden die Aufgaben so allgemein wie möglich formuliert. Dadurch kann beurteilt werden, ob das System so selbsterklärend ist, wie dies angedacht wurde. Allerdings konnten Aufgaben mit geringer Komplexität nur eingeschränkt allgemein formuliert werden. Ein Beispiel dafür ist eine Aufgabe mit der folgenden Struktur: „Lassen Sie den Roboter in eine bestimmte Richtung drehen“. Eine allgemeinere Formulierung dieser Aufgabe ist nicht möglich.
Die vorgelegten Aufgaben waren:

i. Sorgen Sie dafür, dass sich der Roboter nach links beziehungsweise rechts dreht.
ii. Lassen Sie den Roboter in die Küche fahren.
iii. Teilen Sie dem Roboter mit, dass er wieder zu Ihnen zurückkehren soll.
iv. Beauftragen Sie den Roboter damit, Ihnen Tee zu holen.
    a) Teilen Sie dem Roboter mit, dass Sie für Ihren Kaffee noch Milch benötigen.
    b) Während der Roboter die Aufgabe ausführt, ändert sich Ihre Meinung und sie möchten doch keine Milch.
    c) Sorgen Sie dafür, dass der Roboter die aktuelle Handlung abbricht.
v. Gehen Sie davon aus, dass Ihnen der Roboter im Weg steht. Teilen Sie ihm nun also mit, dass er Ihnen Platz machen soll.
vi. Nach manueller Aktivierung der Personenerkennung durch den Studienleiter: Lassen Sie sich von dem Roboter ein Wasser bringen.

Damit die auf Englisch stattfindende Interaktion nicht durch Probleme in der Übersetzung verfälscht wurde, bekamen die Probanden sowohl eine deutsche als auch eine englische Fassung der Aufgaben.

6.1.2 Quantitative Fragen

Die quantitativen Fragen lassen sich in zwei Teile unterscheiden. Im ersten wurden die Fragen des System Usability Score (SUS) von Brooke et al. [Bro+96] gestellt. In einem zweiten Teil haben sich die Fragen stärker an systemspezifischen Eigenschaften orientiert.
Der SUS wurde für die Bewertung des System genutzt, da mit diesem eine einfache und schnelle Einschätzung der Benutzerfreundlichkeit des Systems möglich ist [Bro+96]. Diese Einordnung hilft für die Beurteilung, ob das System selbsterklärend ist, da ein benutzerfreundliches System nur äußerst wenige Instruktionen für die Nutzung bedarf.
Der Ablauf der Befragung für den SUS wird dabei von Brooke et al. vorgegeben. Dem Nutzer werden zehn verschiedene Aussagen vorgelegt, deren Inhalt zwischen positiv und negativ alterniert. Für die Bewertung wird eine 5-Punkt-Likert Skala eingesetzt, wobei durch den Wert „Eins“ eine starke Ablehnung ausgedrückt wird, während „Fünf“ starke Zustimmung ausdrückt. Damit eine Vergleichbarkeit der Bewertung möglich ist, werden die einzelnen Antworten gemäß einer Berechnungsvorschrift addiert. Bei geradzahligen Fragen wird der erhaltene Wert von fünf subtrahiert, während er bei geraden Fragen um den Wert „Eins“ verringert wird. Die erhaltene Summe wird anschließend mit dem Faktor 2,5 multipliziert, wodurch ein Ergebnis zwischen 0 und 100 erzielt wird. Dabei gilt pauschal: höhere Werte stehen für eine bessere Benutzerfreundlichkeit. [Bro+96]
Eine Einordnung des erhaltenen Ergebnisses kann anschließend auf der Basis einer Analyse von Bangor et al. [Ban+09] durchgeführt werden. Dafür wurden fast 1.000 Fragebögen ausgewertet und auf einer Skala eingeordnet.
Die folgenden zehn Aussagen sind für die Ermittlung des SUS durch die Probanden zu bewerten:

i. I think that I would like to use this system frequently.
ii. I found the system unnecessarily complex.
iii. I thought the system was easy to use.
iv. I think that I would need the support of a technical person to be able to use this system.
v. I found the various functions in this system were well integrated.
vi. I thought there was too much inconsistency in this system.
vii. I would imagine that most people would learn to use this system very quickly.

viii. I found the system very cumbersome to use.

ix. I felt very confident using the system.

x. I needed to learn a lot of things before I could get going with this system.

Mit dem zweiten Teil der Fragen soll es ermöglicht werden, die Teile der Hypothesen aus 6.1 zu bewerten, die durch den SUS nicht abgedeckt werden. Außerdem können mithilfe dieser Fragen Designentscheidungen aus den vorherigen Kapiteln bewertet werden. Dafür wurden den Probanden elf weitere Aussagen vorgelegt, die mittels der selben 5-Punkt-Likert Skala wie im vorherigen Teil zu bewerten waren. Diese Aussagen handelt es sich um die Folgenden:

i. The spoken answers by the system were clear and natural.

ii. The system responded quickly to my requests.

iii. The system had no problems to correctly understand my requests.

iv. The interaction with the system felt natural.

v. I had no problems using the wakedword.

vi. I found the delay until the system answered disturbing in the interaction.

vii. When interacting with a voice assistant, answers should be given as quickly as possible even when the answer sounds more like a machine.

viii. It is useful, that the robot is offering help, when he sees me.

ix. Privacy is very important for me.

x. I would us a voice assistant more often, if my privacy is guranteed.

xi. When interacting with a voice assistant, it's very important, that answers sound natural even when they take a bit longer.

Für eine bessere Einordnung der Probanden wurden sie gefragt, ob sie bereits einen Sprachassistenten verwenden. Außerdem wurden einige persönliche Angaben abgefragt. Dazu zählen Geschlecht, Altersgruppe, höchste Qualifikation, aktuelle Arbeitssituation sowie eine Selbsteinschätzung über die Technikversiertheit.
Um konsistente Gedankengänge zwischen der Erledigung der Aufgaben und Beantwortung der Fragen weiterhin zu erhalten, wurden die Fragen auf Englisch vorgelegt.

6.1.3 Qualitative Fragen

Die qualitativen Fragen wiederum ließen sich nicht so genau formulieren wie die quantitativen, da sie auch immer im Bezug zu der vorherigen Aufgabendurchführung sowie der Entwicklung des Gesprächs zu sehen sind. Der Verlauf des Gesprächs orientiert sich dabei jedoch an den Hypothesen aus Kapitel 6.1. Mittels dieser Befragung sollen die Einschätzungen der Probanden besser verstanden und die

Ursachen für einzelne Einschätzungen nachvollzogen werden.
Besonderes Augenmerk wurde auf die Gründe für die Nutzung beziehungsweise die Ablehnung von Sprachassistenten gelegt und damit war es auch möglich, genauer auf den Zusammenhang zwischen Datenschutz und Nutzung einzugehen. Es wurde außerdem genauer auf Unterschiede im Hinblick auf existierende, kommerzielle System eingegangen und mögliche Konsequenzen für die Funktionsfähigkeit der Systeme eingeschätzt.
Einen weiteren Schwerpunkt stellt die Art der Handlungsbestätigung dar. Ziel war es herauszufinden, welche Art der Bestätigung in bestimmten Situationen sinnvoll erscheint. Für die Beantwortung dieser Problemstellung konnte ein direkter Bezug zu den vorherigen Erfahrungen aus dem Nutzertest hergestellt werden.
Zusätzlich wurden auch die Meinungen über verschiedene Stimmen thematisiert. Besonders von Interesse waren dabei die Faktoren Verzögerung zwischen Anfrage und Antwort, Natürlichkeit der Sprachausgabe sowie Deutlichkeit der gesprochenen Worte.

6.2 Auswertung der Studie

Für die Auswertung der Studie bietet es sich an, die empirisch erhobenen Daten zu analysieren und mit den qualitativ erhaltenen Informationen zu verknüpfen. So lassen sich die Hypothesen aus Kapitel 6.1 einfacher bewerten und einordnen. Außerdem können Verbesserungspotentiale aufgezeigt werden
Insgesamt wurde die Studie von 13 Teilnehmern durchgeführt, deren Demografie in Kapitel 6.2.3 beschrieben wird.

6.2.1 System Usability Score

Der System Usability Score (SUS) wurde unter den Berechnungsvorschriften von Brooke et al. [Bro+96], wie auch in Kapitel 6.1.2 beschrieben, ermittelt. Damit er aussagekräftig ist, wurde er für die einzelnen Antworten, unter Berücksichtigung möglicher Enthaltungen, normalisiert. Die Werte für die einzelnen Aussagen werden in Abbildung 6.1 dargestellt.
Insgesamt hat das System ein Ergebnis von 70 Punkten erreicht, dies entspricht nach Bangor et al. [Ban+09] dem Adjektiv „gut“ beziehungsweise der (im englischsprachigen Raum) verwendeten Schulnote C. Gemäß den Analysen von Sauro [Sau11] ist es somit besser bewertet, als alle mit dem SUS bewerteten Systeme im Durchschnitt. In Abbildung 6.1 ist ersichtlich, dass es bei der Bewertung der einzelnen Aussagen einige Ausreißer von dem durchschnittlich erzielten Wert von 2,7 gibt.
Eine genauere Betrachtung dieser Ausreißer ermöglicht eine bessere Aussage über mögliche Schwachpunkte und Stärken des Systems. Die Bewertung der Aussage 1

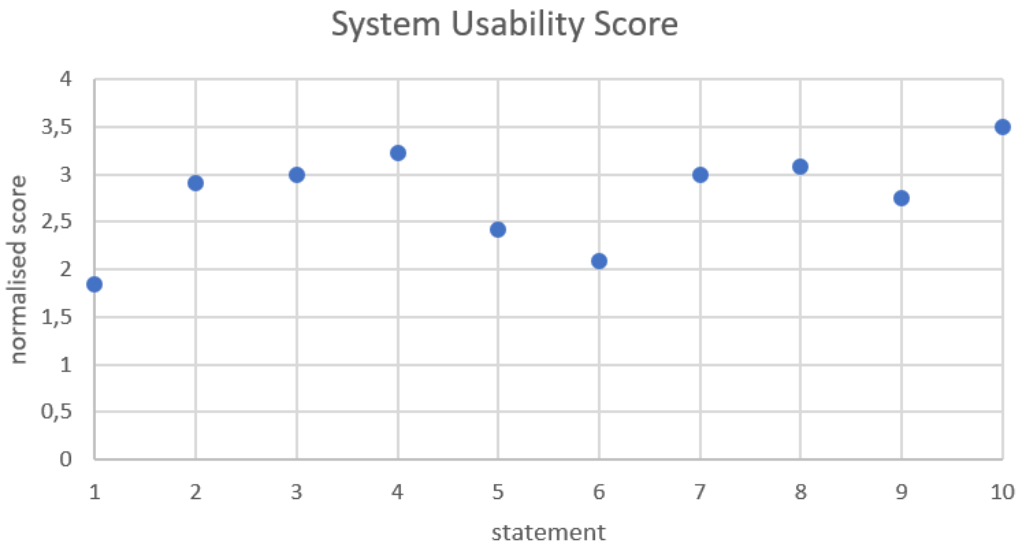

Abb. 6.1: Ergebnisse System Usability Score (SUS)

(siehe Abbildung 6.2) bedeutet, dass sich die Probanden unsicher sind, ob sie das System öfter nutzen würden. Dabei bewerten sie allerdings nicht nur die Sprachsteuerung an sich, sondern auch die Funktionen, die ein solcher Roboter in ihren Augen übernehmen könnte. Neben Problemen bei der Interaktion, zum Beispiel korrektes Verständnis von Aussagen, beider Interaktionsteilnehmer, liegt diese Wertung auch an der Einschätzung einiger Teilnehmer, dass ihnen Sprachassistenen im Allgemeinen keinen Mehrwert bieten.
Einen weiteren Ausreißer stellen die Antworten zu Aussage 6 dar. Dabei ist erkennbar, dass sich die Studienteilnehmer sehr uneinig sind, ob das System zu viele Inkonsistenzen aufweist. Eine Ursache für negative Bewertungen liegt darin, dass einigen Teilnehmern nicht klar war, wann der Kontext einer Interaktion sich verändert hat. Somit waren ihnen die erwünschten konsistenten Interaktionen nicht möglich. Andererseits haben ebenfalls fünf Probanden nicht zu viele Inkonsistenzen in dem System empfunden.
Eine besonders hohe Ablehnung hat die Aussage 10 erfahren, deren Inhalt ist, dass für die Nutzung des Systems viel erlernt werden muss. Die Testpersonen waren sich also nahezu einig, dass die Bedienung des Systems leicht zu erlernen ist. Die Verteilung der Antworten ist in Abbildung 6.3 dargestellt, wobei auch im Gespräch der einzige Ausreißer nicht geklärt werden konnte.

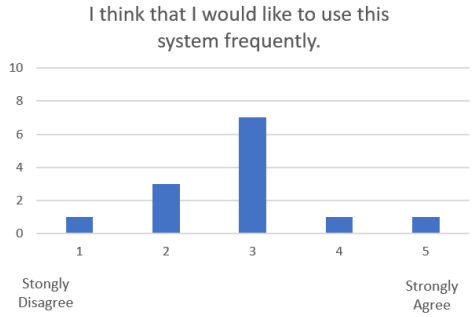

Abb. 6.2: Bewertung Aussage 1

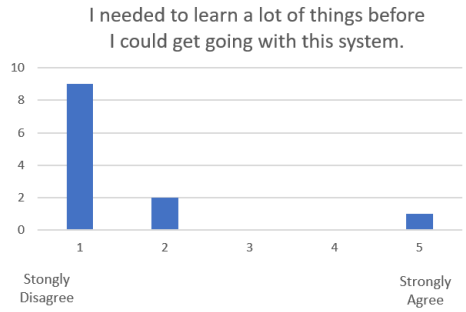

Abb. 6.3: Bewertung Aussage 10

6.2.2 Systemspezifische Fragen

Der zweite Teil des Fragebogens besteht nicht wie der vorherige Teil aus standardisierten Aussagen, sondern soll es ermöglichen, die vorherigen Bewertungen nachvollziehbarer zu gestalten. Außerdem sollen diese Aussagen die Bewertung der Hypothesen und zuvor getroffenen Designentscheidungen ermöglichen.
Als erstes war die Aussage zu bewerten, ob die Antworten des Systems als klar und natürlich empfunden wurden. Dies wurde von 61 % der Befragten abgelehnt. Mögliche Ursache ist der, teilweise sehr starke, Akzent der Stimme von Mimic II. So wurde beispielsweise das Wort „tea“ nur von den beiden Muttersprachlern direkt verstanden, die restlichen Teilnehmer konnten dies auch nach einer mehrfachen Wiederholung nicht verstehen. Eine ähnliche Situation gab es bei dem Angebot der Hilfe nach erfolgter Personenerkennung. Das Hilfsangebot wurde erneut von einem Großteil der Personen nicht im ersten Versuch verstanden. Prinzipiell empfanden aber über 60 % der Probanden die Initiierung der Interaktion durch den Roboter als hilfreich. Neben der Aussprache durch den Sprachassistenten haben auch die Lautsprecher des Roboters einen negativen Einfluss auf das Verständnis. Da diese als eher unterdurchschnittlich zu betrachten sind, tragen sie durch eine unklare Ausgabe zu Verständnisschwierigkeiten bei. Aufgrund dieser Tatsachen und dem maschinellen Grundton wurde die Aussage auch als weniger natürlich bewertet.
Die Interaktion wurde jedoch trotzdem von mehr als der Hälfte der Probanden als natürlich empfunden. Dabei wurde durch Probanden geäußert, dass sich die Interaktion mit einem Sprachassistenten nie natürlich anfühlen wird. Unter anderem liegt dies daran, dass vor jedem Start einer Interaktion das Aktivierungswort benötigt wird. Im Gespräch von zwei Personen wird dies meist nur für die Initiierung genutzt, auch wenn es danach mehrere Einzelinteraktionen gibt. Zusätzlich wurde angemerkt, dass nicht immer klar ist, ob ein voriger Kontext noch erhalten ist oder nicht. Gerade im Umgang mit Personen kann davon ausgegangen werden, dass ein solcher über einen längeren Zeitraum erhalten bleibt.
Mit der Nutzung des gewählten Aktivierungswortes hatten mehr als 75 % der Probanden keine Probleme. Wobei dies nicht bewertet, ob das Aktivierungswort an den

einzelnen Stellen als sinnvoll erachtet wurde, sondern ob die Nutzung von „Hey Mycroft“ zu Problemen geführt hat. Ein Proband hat das System wiederholt versucht mit „OK Mycroft“ zu aktivieren. Dieses Signalwort rührt dabei von der Nutzung des Google Assistant, wobei Mycroft dies auch als Aktivierung angesehen hat.
Eine wichtige Rolle hat außerdem die Geschwindigkeit der Reaktionen des Systems gespielt. Dies umfasst die Dauer zwischen Nutzung des Signalworts und Aktivierung des Systems sowie zwischen Beenden einer Phrase und Aktion des Systems. Kein Proband war der absoluten Überzeugung, dass das System schnell auf die Anfragen reagiert hat. Die Verteilung der Antworten ist in Abbildung 6.4 zu sehen. Jedoch wurde durch eine größere Zahl der Teilnehmer die Dauer bis zu einer Reaktion als störend in der Interaktion empfunden. Dies ist in Abbildung 6.5 dargestellt. Teilweise resultierte diese Einschätzung daraus, dass es von dem System keinerlei Reaktion gab, wenn eine Aussage nicht verstanden wurde. Diese Verzögerung wurde auch als Dauer zwischen Ende der Phrase und Aktion des Systems empfunden. Wobei aus den Beobachtungen der Aufgabendurchführung hervorgegangen ist, dass die Verzögerung zwischen Aktivierungswort und Aktivierung störender war. Teilweise wurde der Befehl direkt nach dem Signalwort angehängt und dann durch den Assistenten nicht oder nur teilweise verstanden. In anderen Situationen haben sich Probanden wiederholt, nachdem der Signalton abgespielt wurde.
Entsprechend haben mehr Personen die Aussage abgelehnt, dass das System Anfragen korrekt verstanden hat. Teilweise wurden auch Wörter verstanden, die nur entfernt verwandt mit den eigentlich Gesagten waren. Durch das System wurde dann geäußert, dass nicht klar ist, welche Aktion ausgeführt werden soll. Für die Nutzer wurde aber nicht ersichtlich, welche Wörter falsch verstanden wurden, was gerade beim wiederholten Eintreten dieser Situation zu Frustration geführt hat. Teilweise wurden Befehle auch in Sätzen formuliert, die zwar inhaltlich eindeutig waren, aber nicht in der Definition des Vokabulars von Intents aufgeführt wurden. Somit wurden die gewünschten Aktionen nicht durchgeführt, was bei Probanden den Eindruck erweckte, nicht korrekt verstand worden zu sein.
Damit diese Fehler nicht mehr auftreten, haben einige Probanden versucht, langsamer zu sprechen. Dies wiederum hat dazu geführt, dass das System keine der geäußerten Wörter verstanden hat.
Wie in Abbildung 6.6 zu sehen, ist der Mehrheit Probanden dieser Studie der Datenschutz sehr wichtig. Unter anderem aus diesem Grund verzichten aktuell über zwei Drittel der Befragten auf die Nutzung eines Sprachassistenten. Jedoch können sich 45 % vorstellen, einen Sprachassistenten generell oder häufiger zu nutzen, wenn der Datenschutz garantiert werden kann (siehe Abbildung 6.7). Dabei sind aber einige Probanden der Meinung, dass sie die Nutzung schon jetzt maximiert haben, während andere der Meinung sind, dass ihnen ein Sprachassistent kein Mehrwert bringt und sie somit auch im Falle des verbesserten Datenschutzes kein Interesse an der Nutzung haben.

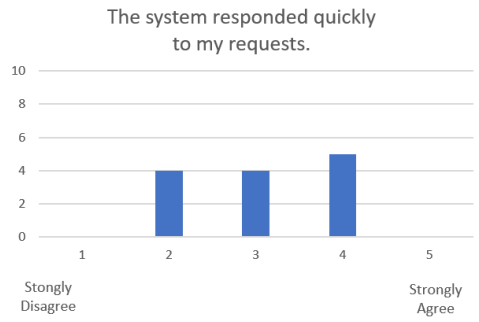

Abb. 6.4: Beurteilung der Antwortgeschwindigkeit

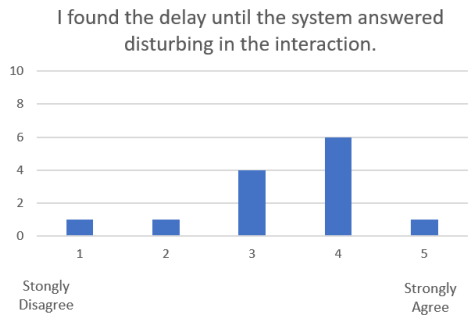

Abb. 6.5: Einfluss der Antwortgeschwindigkeit auf die Interaktion

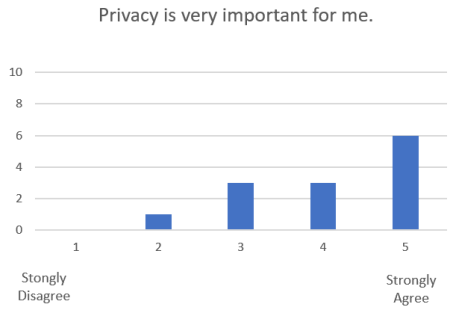

Abb. 6.6: Relevanz des Datenschutzes

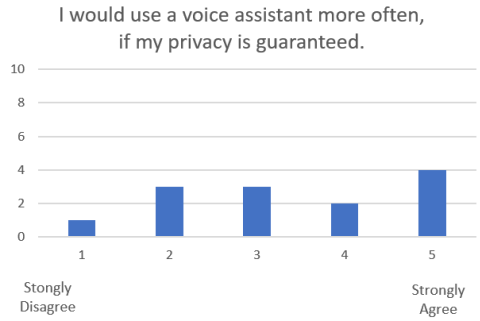

Abb. 6.7: Einfluss des Datenschutzes auf die Nutzungshäufigkeit

6.2.3 Probandendemografie

Da es sich bei dieser Studie um eine Pilotstudie handelt, können die Probanden nicht als repräsentativ betrachtet werden. Es wurde versucht, die Gruppe möglichst heterogen zu wählen. In den Abbildungen 6.8, 6.9 und 6.10 sind die Probanden näher beschrieben.
Dabei ist erkennbar, dass die Mehrheit der Probanden maximal 30 Jahre alt ist und somit durch die Studie keine Aussagen getroffen werden können, inwiefern Senioren mit dem System interagieren. Auch durch die beiden Probanden in der Altersgruppe 51-60 erlauben keine solche Aussagen, da auch bei ihnen davon ausgegangen werden kann, regelmäßig mit moderner Informationstechnik zu interagieren. Dies ist insofern wichtig, als dass insbesondere Senioren weniger an den Umgang mit Computern und Smartphones gewöhnt sind und somit möglicherweise anders interagieren würden. Auch die Aussagen zu der Technikaffinität (siehe Abbildung 6.11) sind mit diesem Hintergrundwissen zu betrachten. So kann davon ausgegangen werden, dass die Technikaffinität der Gruppe unter 30-jähriger grundsätzlich größer ist, als die von Senioren.

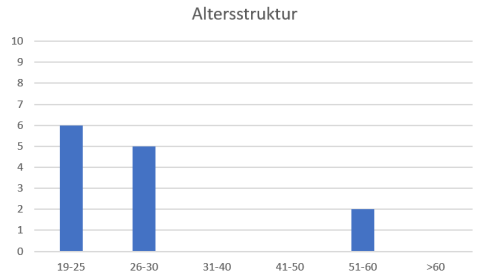

Abb. 6.8: Altersstruktur der Probanden

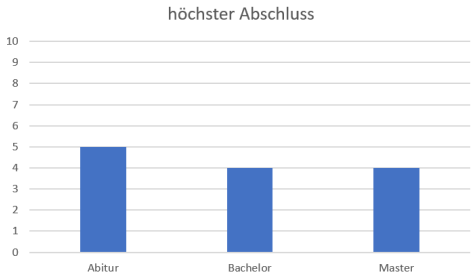

Abb. 6.9: höchster Abschluss der Probanden

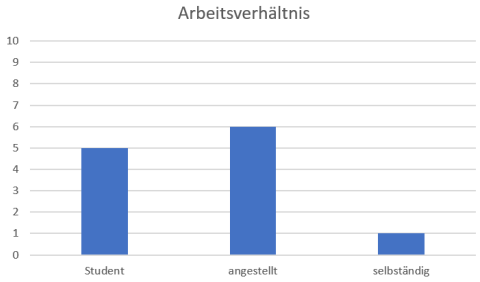

Abb. 6.10: aktuelles Arbeitsverhältnis der Probanden

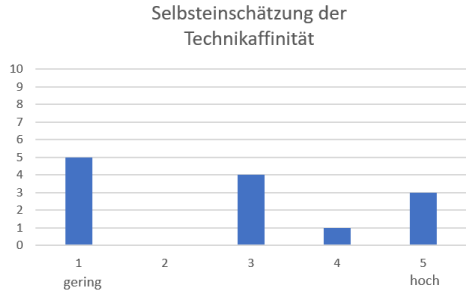

Abb. 6.11: Selbsteinschätzung der Technikaffinität der Probanden

6.2.4 Weitergehende Meinungen über das System

Im Rahmen des im Anschluss an die Befragung mit jedem Probanden geführte Gespräch wurden einige Themen behandelt, die nicht durch die vorherigen Fragen abgedeckt wurden oder mit diesen im direkten Zusammenhang standen.
Im Bezug auf die möglichen Stimmen des Roboters hat sich herausgestellt, dass die Stimme von Mimic I gegenüber der von Mimic II bevorzugt wird, insofern sie denn schneller abgespielt wird. Denn aus Sicht einiger Testpersonen sind beide von stark mechanischem Klang, so dass die schnellere Antwort allgemein bevorzugt wird.
Auf die Frage, welche Arten der Bestätigung bevorzugt werden, ähneln sich die Antworten stark. So wurde es gemeinhin als gut empfunden, wenn einfache Aktionen nicht akustisch bestätigt werden, da dies die Ausführung unnötig verzögern könnte. Für komplexere Tätigkeiten wurde die Bestätigung des Befehls durch den Roboter als sehr hilfreich empfunden, da immer noch korrigierend eingegriffen werden kann, bevor es zur fehlerhaftem Ausführung kommt. Nachfragen hingegen wurden nur dann als sinnvoll angesehen, wenn sich der Roboter unsicher über eine Handlung ist.
Als kritisch wurde auch die fehlende Rückmeldung durch den Sprachassistenten angemerkt, wenn Befehle trotz Aktivierung nicht verstanden wurden. In diesem Fall wurde sich eine Meldung der Art „Leider habe ich Sie nicht verstanden. “ gewünscht.

6.2.5 Bewertung des Untersuchungsziels

Anhand der Antworten der Nutzer sowie des beobachteten Verhaltens ist es möglich, die in Kapitel 6.1 formulierten Hypothese zu bewerten. Dabei handelt es sich um:

(a) Die Nutzung des Systems ist für den Nutzer selbsterklärend und vermittelt ein Gefühl von natürlicher Interaktion.

(b) Ein stärkerer Fokus auf den Datenschutz erhöht die Bereitschaft zur Nutzung von Sprachassistenten, auch wenn bislang kein solcher Assistent verwendet wird.

Für eine Bewertung ist es sinnvoll, Hypothese (a) in zwei Bestandteile zu zerlegen. Zunächst sollte die Nutzung selbsterklärend sein. Dieser Teil lässt sich gut mit dem System Usability Score (SUS) bewerten, da -wie eingangs erwähnt-, eine hohe Benutzerfreundlichkeit mit einem hohen Maß an selbsterklärender Bedienung einhergeht. Aufgrund des erreichten Wertes von 70 Punkten und der vielfach vertretenen Meinung, dass bei der Nutzung des Systems keine Unterstützung Dritter nötig ist, kann dieser Teil der Hypothese als korrekt betrachtet werden.
Im zweiten Teil der Hypothese wurde behauptet, dass die Systemnutzung ein Gefühl von natürlicher Interaktion vermittelt. Diese Aussage hat sich mit der Einschränkung bestätigt, dass sich die Nutzer immer noch der Interaktion mit einer Maschine bewusst waren. Trotzdem wurde offenbar ein gewisses Gefühl von natürlicher Interaktion vermittelt.
Der Hypothese (b) kann uneingeschränkt zugestimmt werden. Es besteht eine eindeutige Bereitschaft, den Sprachassistenten bei garantiertem Datenschutz häufiger einzusetzen. Erkennbar wird dies an den Abbildungen 6.6 und 6.7.
Insgesamt ist das Konzept also funktionsfähig, in einigen Punkten gibt es aber Verbesserungspotential. Dabei handelt es sich zumeist um das Zusammenspiel der einzelnen Komponenten handelt und nicht um die Abläufe.

6.3 Zusammenfassung

Mittels der Pilotnutzerstudie war es möglich, das zuvor erstellte Konzept auf seine Funktionsfähigkeit zu untersuchen. Dabei wurden die Untersuchungsschwerpunkte darauf gelegt, ob die Systemnutzung selbsterklärend ist und sich natürlich anfühlt. Außerdem wurde untersucht, ob eine Garantie des Datenschutzes mehr Personen von Sprachassistenten überzeugen kann. Dafür wurde eine Bewertung des Systems mittels des System Usability Score (SUS) und einige systemspezifischen Fragen vorgenommen. Vertieft wurden diese Antworten im Rahmen von persönlichen Gesprächen. Es hat sich herausgestellt, dass das Konzept prinzipiell einsatzfähig ist. Gerade die Fragen danach, ob das System selbsterklärend ist und ob Datenschutz

mehr Menschen überzeugen kann, konnten bestätigt wurden. Auch bescheinigt werden, dass ein gewisser Eindruck von Natürlichkeit in der Interaktion entsteht, es aber in dieser Hinsicht noch Entwicklungspotential gibt.

7 Fazit

Zum Abschluss dieser Arbeit werden die vorherigen Ergebnisse zusammengefasst. Außerdem werden diese diskutiert und Verbesserungspotentiale auf Basis der vorherigen Erkenntnisse aufgezeigt. Zudem wird ein Ausblick auf die Entwicklungsmöglichkeiten dieses und ähnlicher System gegeben.

7.1 Zusammenfassung der Arbeitsergebnisse

Im Rahmen dieser Arbeit wurde ein Konzept erarbeitet, mit dem es möglich ist, auf Basis natürlicher Sprache mit einem Assistenzroboter zu interagieren. Dafür kommt ein Sprachassistent zum Einsatz, welcher allgemein eine Umsetzung der Umwandlung von gesprochener Sprache in maschinenverständlichen Text durchführt und in der Lage ist, eine Antwort in Form von gesprochener Sprache zu geben. Für diese Aufgabe hat sich *Mycroft AI* als am besten geeignet erwiesen. Dieses System ordnet dem Datenschutz eine große Bedeutung und ist zudem modular aufgebaut. Somit können für die einzelnen Schritte der Sprachverarbeitung unterschiedliche Implementierungen eingesetzt werden.
Damit auch Personen, die keine Vorkenntnisse über die Funktionsweise des Systems besitzen, in der Lage sind, dieses für die Interaktion mit einem Assistenzroboter zu nutzen, wurde ein Interaktionskonzept erstellt. Mittels diesem ist es möglich, dass die den Nutzer getätigte Aussagen automatisch den entsprechenden Befehlen zugeordnet, die durch den Roboter ausgeführt werden. Dabei entsteht für den Anwender das Gefühl, direkt mit dem Roboter zu interagieren. In Wirklichkeit leitet dieser die Audiodaten an den Sprachassistenten weiterleitet, der die passenden Handlungen einleitet. Auf diese Art werden die Anforderungen an den Roboter minimal gehalten, wodurch eine Vielzahl unterschiedlicher Roboter mit diesem Konzept eingesetzt werden können.
Dieses Konzept konnte erfolgreich mit einem selbst-balancierenden Roboter von Segway Robotics umgesetzt werden, wobei dabei einzelne Funktionen aus Komplexitätsgründen lediglich simuliert werden. Mittels dieser Umsetzung war es möglich, das Konzept durch eine Nutzerstudie auf seine Funktionsfähigkeit zu untersuchen. Dabei waren die Untersuchungsschwerpunkte neben der Natürlichkeit der Kommunikation und ob diese selbsterklärend ist auch der Einfluss des Datenschutzes auf das Nutzungsverhalten von Sprachassistenten. Im Ergebnis hat sich das Konzept als

einsatzfähig erwiesen, wobei sich die Interaktion ein gewisses Gefühl von Natürlichkeit erzeugt und dem System eine gute Bedienbarkeit bescheinigt wurde. Auch hat sich als korrekt erwiesen, dass Nutzer einen Sprachassistenten häufiger verwenden würden, wenn dieser den Datenschutz garantiert.

7.2 Diskussion

Um das erstellte Konzept sinnvoll bewerten zu können, ist es nötig einen Bezug zu den Anforderungen herzustellen. Diese können in Kapitel 4.1 im Detail nachgelesen werden.
Insgesamt hat sich das Konzept als funktionsfähig erwiesen. Jedoch sind einige Designentscheidungen kritisch zu hinterfragen, da sie sich im Rückblick nicht immer als ideal erwiesen haben, auch wenn die Gründe nachvollziehbar sind.
Zuerst sollen die Entscheidungen für die einzelnen Bestandteile von Mycroft analysiert werden, da diese zum Teil die schnellsten Verbesserungen versprechen. Festzuhalten ist, dass mit Mycroft die meisten Muss Kriterien erfüllt werden konnten, einige aufgrund der gewählten Systembausteine aber nur als begrenzt erfüllt gesehen werden können.
So hat sich die Wahl von Mimic II für die Sprachausgabe nicht als ideal erwiesen. Die Hoffnung war, damit innerhalb vertretbarer Zeit Antworten zu erzeugen, deren Klang nicht zu maschinell ist. Da jedoch die Stimme durch die Probanden als nur gering natürlich eingeschätzt wurde, während die Dauer bis zu einer Antwort als zu lang empfunden wurde. Möglicherweise ist es aktuell nicht möglich, beide dieser Kriterien mit zeitgleicher Rücksicht auf den Datenschutz zu erfüllen.
Auch hat sich die Sprache-zu-Text Umwandlung mittels der Google STT als weniger zuverlässig erwiesen als erwartet. Da für dieses System aber Abstriche im Bezug auf den Datenschutz gemacht werden mussten, wirkt ein Überprüfung der Funktionsfähigkeit des DeepSpeech Servers sinnvoll. Da dieser mit in dem aktuellen Entwicklungsstand eigenständig eingerichtete werden muss, sollten die einzelnen Konfigurationsmöglichkeiten genauer betrachtet werden. So könnte eine Kopplung der Wahrscheinlichkeiten einzelner Wörter an das für einzelnen Skills definiert Vokabular die Erkennungsgenauigkeit erhöhen. In diesem Fall würde bei der Umwandlung dem Vokabular eine höhere Wahrscheinlichkeit zugeordnet, das auch die einzelnen Skills hervorruft. Zeitgleich könnte eine lokale Installation eines STT Umwandlers die Antwortzeiten maßgeblich reduzieren. Möglicherweise geringere Rechenkapazitäten könnten in dem Fall durchaus durch die geringeren Umlaufzeiten der Anfrage aufgewogen werden.
In diesem Zusammenhang sollte auch das Feedback an den Nutzer verbessert werden, wenn einer Aussage kein Skill zugeordnet werden kann. Konnten überhaupt keine Wörter verstanden werden, sollte dies dem Nutzer mitgeteilt werden. Ein

andere Meldung sollte erfolgen, wenn die verstandenen Wörter nicht klar einem Skill zuzuordnen sind.
Auch die Wahl von Adapt hat in den Nutzertests wiederholt gezeigt, dass es nur schwierig möglich ist, alle für einen Befehl möglichen Formulierungen zu berücksichtigen. Aus diesem Grund sollten weitergehende Tests mit Padatious durchgeführt werden, auch wenn dieses System bislang noch nicht vollends ausgereift ist. Da dieses jedoch aufgrund seiner Basis auf maschinellem Lernen flexibler bei der Zuordnung von Wörtern zu Skills ist, könnten die Nutzerfahrung maßgeblich verbessert werden.
Einige Kann Kriterien konnten umgesetzt werden, wobei sich nach der Auswertung herausgestellt hat, dass einige dieser vermutlich keinen messbaren Einfluss auf das Nutzererlebnis haben werden. Beispielsweise hatte nur wenige Nutzer ein Problem mit dem aktuell verwendeten Aktivierungswort, während zugleich keine fehlerhaften Aktivierungen festgestellt werden konnten. Ob ein individualisiertes Wort zu Verbesserungen führen würde, kann hinterfragt werden.
Die Verwendung des Arraymikrofons konnte nicht umgesetzt werden, da Loomo zwar über ein solches verfügt, aber in der Dokumentation nicht erläutert wird, wie dieses zu verwenden ist. Auch war eine manuelle Nutzung nicht möglich.
Jedoch ist im Rückblick auch die Entscheidung für Mycroft kritisch zu betrachten. Zwar hat es die Eigenschaft als Open-Source Produkt Anpassungen an allen relevanten Stellen erlaubt. Allerdings haben sich an einigen Stellen immer wieder systembedingte Schwierigkeiten aufgetan. Besonders auffällig waren wiederholte Stabilitätsprobleme, die auch während des Nutzertests aufgetreten sind und somit das Studienergebnis beeinflusst haben. Das diese und weitere Probleme durch den Hersteller erkannt wurden und auch an ihnen gearbeitet wird, hat sich zum Projektende gezeigt. So wurde Anfang Oktober 2019 eine neue Version der Kernfunktionen veröffentlicht, die einige dieser Probleme beheben sollen. [1]
Möglicherweise könnten die Ergebnisse mit der Verwendung von Snips weiter verbessert werden, besonders im Hinblick auf Antwortzeiten.
Die Aussagen der Studie sind aber auch immer unter dem Aspekt zu betrachten, dass diese als Pilotstudie durchgeführt wurde und somit keine repräsentativen Aussagen zu lässt. Vielmehr ermöglicht sie grundlegende Aussagen, von denen sich möglicherweise Tendenzen ablesen lassen.
Einen, neben dem Datenschutz weiteren, wesentlichen Vorteil gegenüber herkömmlichen Sprachassistenten weißt die in dieser Arbeit beschriebene Lösung auf. Durch die Personenerkennung ist es besser möglich, verschiedene Personen zur Nutzung des Systems zu animieren, auch wenn sie sich zuerst noch unsicher fühlen oder nicht um die Funktionen des Systems wissen.
Insgesamt kann gesagt werden, dass trotz der Verbesserungsmöglichkeiten dieses Konzept für den Einsatz prinzipiell geeignet ist. Besonders durch die größere Flexibi-

[1] https://mycroft.ai/blog/mycroft-core-19-08-release/ [Abgerufen am 09.10.2019]

lität, vor allem im Bezug auf die Umsetzungen der Verarbeitungsschritte sowie die Befehle, zeichnet es sich gegenüber bereits existierenden, starren Konzepten aus.

7.3 Ausblick

Generell ist festzuhalten, dass sich die unabhängigen Sprachassistenten aktuell noch in einer Anfangsphase der Entwicklung befinden und dadurch nur schwierig mit den Anbietern kommerzieller Lösungen konkurrieren können. Jedoch können sie bereits jetzt für die Interaktion mit einem Assistenzroboter eingesetzt werden. Allerdings sind dafür noch weitere Tests nötigt. So sollte zum einen das aktuelle System mit anderen Umsetzungen der einzelnen Bestandteile der Verarbeitungspipeline getestet werden. Dies verspricht in bestimmten Teilen bereits ein hohes Verbesserungspotential.
Außerdem sollte das Gesamtsystem mit alternativer Software getestet werden. In diesem Zusammenhang würde sich eine tiefergehende Betrachtung von Snips anbieten, dass bereits in seiner Grundkonfiguration einige der Anforderungen erfüllt.
Außerdem wirkt es erstrebenswert, die Datenverarbeitung zu noch größeren Teilen lokal durchzuführen. Zum einen ist dies die Sprache-zu-Text Umwandlung, mit der neben Verbesserungen im Bezug auf den Datenschutz auch noch eine Erhöhung der Verarbeitungsgeschwindigkeit möglich ist.

Abkürzungsverzeichnis

ASR	Automatic Speech Recognition
AVS	Alexa Voice Service
AWS	Amazon Web Services
DSGVO	Datenschutzgrundverordnung
DTS	detection-tracking system
EU	Europäischen Union
JSON	JavaScript Object Notation
NLP	natürliche Sprachverarbeitung
NLU	Natural Language Understanding
STT	Sprache-zu-Text
SDK	Software Development Kit
SUS	System Usability Score
TTS	Text-zu-Sprache

Literatur

[Amb+18] Aditya Amberkar, Parikshit Awasarmol, Gaurav Deshmukh und Piyush Dave. „Speech Recognition using Recurrent Neural Networks“. In: *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*. IEEE. 2018, S. 1–4 (zitiert auf den Seiten 7, 24).

[Ank+19] Jürgen Anke, Uwe Fischer und René Lemke. „Integration digitaler Sprachassistenten in den Kundenservice am Beispiel der Stadtwerke Leipzig“. In: *Digitalisierung von Staat und Verwaltung* (2019) (zitiert auf den Seiten 28, 30).

[Apt+17] Noah Apthorpe, Dillon Reisman und Nick Feamster. „A smart home is no castle: Privacy vulnerabilities of encrypted iot traffic“. In: *arXiv preprint arXiv:1705.06805* (2017) (zitiert auf den Seiten 11, 30).

[AT18] Inc. Amazon Technologies. „Detecting replay attacks in voice-based authentication“. US-Pat. 16/129,081. 12. Sep. 2018 (zitiert auf Seite 30).

[Ban+09] Aaron Bangor, Philip Kortum und James Miller. „Determining what individual SUS scores mean: Adding an adjective rating scale“. In: *Journal of usability studies* 4.3 (2009), S. 114–123 (zitiert auf den Seiten 65, 67).

[Bro+96] John Brooke et al. „SUS-A quick and dirty usability scale“. In: *Usability evaluation in industry* 189.194 (1996), S. 4–7 (zitiert auf den Seiten 65, 67).

[Buh+95] Joachim Buhmann, Wolfram Burgard, Armin B Cremers et al. „The mobile robot Rhino“. In: *Ai Magazine* 16.2 (1995), S. 31–31 (zitiert auf Seite 15).

[Böh02] Hans-Joachim Böhme. *Serviceroboter und intuitive Mensch-Roboter-Interaktion*. Techn. Univ., Fachgebiet Neuroinformatik, 2002 (zitiert auf den Seiten 14, 15, 17, 18, 37).

[Cal+11] Christopher James Calo, Nicholas Hunt-Bull, Lundy Lewis und Ted Metzler. „Ethical implications of using the paro robot, with a focus on dementia patient care“. In: *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*. 2011 (zitiert auf Seite 13).

[Cha+13] Wan-Ling Chang, Selma Šabanovic und Lesa Huber. „Use of seal-like robot PARO in sensory group therapy for older adults with dementia“. In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2013, S. 101–102 (zitiert auf Seite 14).

[Che+17] Si Chen, Kui Ren, Sixu Piao et al. „You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones“. In: *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE. 2017, S. 183–195 (zitiert auf Seite 30).

[Chu+17] Hyunji Chung, Michaela Iorga, Jeffrey Voas und Sangjin Lee. „Alexa, can I trust you?“ In: *Computer* 50.9 (2017), S. 100–104 (zitiert auf Seite 10).

[Cor+13] Silvia Coradeschi, Amedeo Cesta, Gabriella Cortellessa et al. „Giraffplus: Combining social interaction and long term monitoring for promoting independent living“. In: *2013 6th International Conference on Human System Interactions (HSI)*. IEEE. 2013, S. 578–585 (zitiert auf Seite 17).

[Cou+18] Alice Coucke, Alaa Saade, Adrien Ball et al. „Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces“. In: *arXiv preprint arXiv:1805.10190* (2018) (zitiert auf den Seiten 23, 27).

[Dat17a] Datenschutzkoferenz. „Kurzpapier Nr. 11: Recht auf Löschung / „Recht auf Vergessenwerden““. In: (2017) (zitiert auf Seite 9).

[Dat17b] Datenschutzkoferenz. „Kurzpapier Nr. 7: Marktortprinzip – Regelungen für außereuropäische Unternehmen“. In: (2017) (zitiert auf Seite 9).

[DB19] Wissenschaftlichen Dienst des Deutschen Bundestags. „Zulässigkeit der Transkribierung und Auswertung von Mitschnitten der Sprachsoftware Älexa"durch Amazon“. In: 2019 (zitiert auf Seite 10).

[e.V17] Verbraucherzentrale NRW e.V. *Amazon Alexa: Wann ist der Sprachassistent ganz Ohr? Ein Reaktions-Check. Kurzuntersuchung der Verbrau-cherzentralen*. 2017 (zitiert auf den Seiten 11, 29).

[Edu+19] Jide S Edu, Jose M Such und Guillermo Suarez-Tangil. „Smart Home Personal Assistants: A Security and Privacy Review“. In: *arXiv preprint arXiv:1903.05593* (2019) (zitiert auf den Seiten 5–8, 12, 29, 30).

[FB18a] Eoghan Furey und Juanita Blue. „Alexa, emotions, privacy and GDPR“. In: *Proceedings of the 32nd International BCS Human Computer Interaction Conference*. BCS Learning & Development Ltd. 2018, S. 212 (zitiert auf Seite 9).

[FB18b] Eoghan Furey und Juanita Blue. „She Knows Too Much–Voice Command Devices and Privacy“. In: *2018 29th Irish Signals and Systems Conference (ISSC)*. IEEE. 2018, S. 1–6 (zitiert auf den Seiten 12, 29, 31).

[Fen+17] Huan Feng, Kassem Fawaz und Kang G Shin. „Continuous authentication for voice assistants“. In: *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. ACM. 2017, S. 343–355 (zitiert auf Seite 30).

[FL18] Nathaniel Fruchter und Ilaria Liccardi. „Consumer attitudes towards privacy and security in home assistants“. In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, LBW050 (zitiert auf Seite 38).

[Gal+06] Cipriano Galindo, Javier Gonzalez und J-A Fernandez-Madrigal. „Control architecture for human–robot integration: application to a robotic wheelchair“. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 36.5 (2006), S. 1053–1067 (zitiert auf den Seiten 13, 14).

[Gra+04] Birgit Graf, Matthias Hans und Rolf D Schraft. „Care-O-bot II—Development of a next generation robotic home assistant“. In: *Autonomous robots* 16.2 (2004), S. 193–205 (zitiert auf Seite 32).

[Gre+00] Anders Green, Helge Huttenrauch, Mikael Norman, Lars Oestreicher und K Severinson Eklundh. „User centered design for intelligent service robots“. In: *Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000 (Cat. No. 00TH8499)*. IEEE. 2000, S. 161–166 (zitiert auf den Seiten 18, 46).

[Gro+09] H-M Gross, H Boehme, Ch Schroeter et al. „TOOMAS: interactive shopping guide robots in everyday use-final implementation and experiences from long-term field trials“. In: *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2009, S. 2005–2012 (zitiert auf den Seiten 13, 15).

[Haa+17] William Haack, Madeleine Severance, Michael Wallace und Jeremy Wohlwend. „Security analysis of the Amazon Echo“. In: *Allen Institute for Artificial Intelligence* (2017) (zitiert auf Seite 11).

[Han+02] M Hans, B Graf und RD Schraft. „Robotic home assistant care-o-bot: Past-present-future“. In: *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication*. IEEE. 2002, S. 380–385 (zitiert auf den Seiten 14, 15).

[Han+14] Awni Hannun, Carl Case, Jared Casper et al. „Deep speech: Scaling up end-to-end speech recognition“. In: *arXiv preprint arXiv:1412.5567* (2014) (zitiert auf Seite 25).

[Hay+05] Tomohiro Hayashi, Hiroaki Kawamoto und Yoshiyuki Sankai. „Control method of robot suit HAL working as operator's muscle using biological and dynamical information“. In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2005, S. 3063–3068 (zitiert auf den Seiten 15, 36).

[HD+06] David Huggins-Daines, Mohit Kumar, Arthur Chan et al. „Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices“. In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Bd. 1. IEEE. 2006, S. I–I (zitiert auf Seite 24).

[JO18] Catherine Jackson und Angela Orebaugh. „A study of security and privacy issues associated with the Amazon Echo“. In: *International Journal of Internet of Things and Cyber-Assurance* 1.1 (2018), S. 91–100 (zitiert auf den Seiten 10, 29, 30).

[Jre+09] Camil Jreige, Rupal Patel und H Timothy Bunnell. „VocaliD: Personalizing text-to-speech synthesis for individuals with severe speech impairment“. In: *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. ACM. 2009, S. 259–260 (zitiert auf Seite 26).

[Kar00] J Karlsson. „UN world robotics statistics 1999“. In: *Ind. Robot* 27.1 (2000), S. 14–18 (zitiert auf Seite 12).

[Kum+05] Vijay Kumar, George Bekey, Yuan Zheng et al. „Industrial, personal, and service robots“. In: *DRAFT REPORT* (2005), S. 41 (zitiert auf den Seiten 13, 15).

[Lau+18] Josephine Lau, Benjamin Zimmerman und Florian Schaub. „Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers“. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), S. 102 (zitiert auf Seite 2).

[Loh07] Manja Lohse. *Nutzerfreundliche Mensch-Roboter-Interaktion. Kriterien für die Gestaltung von Personal Service Robots*. 2007 (zitiert auf Seite 18).

[Loi+15] Claudia Loitsch, Michael Schmidt und Gerhard Weber. „Position paper: accessible human-robot interaction (AHRI)“. In: *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. ACM. 2015, S. 16 (zitiert auf Seite 18).

[LS14] Daniel A Lazewatsky und William D Smart. „Accessible interfaces for robot assistants“. In: *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2014, S. 106–111 (zitiert auf Seite 17).

[MA15] Atif M Memon und Ali Anwar. „Colluding apps: Tomorrow's mobile malware threat“. In: *IEEE Security & Privacy* 13.6 (2015), S. 77–81 (zitiert auf Seite 12).

[Mod+13] Chirag Modi, Dhiren Patel, Bhavesh Borisaniya, Avi Patel und Muttukrishnan Rajarajan. „A survey on security issues and solutions at different layers of Cloud computing“. In: *The journal of supercomputing* 63.2 (2013), S. 561–592 (zitiert auf Seite 12).

[Muk+10] Toshiharu Mukai, Shinya Hirano, Hiromichi Nakashima et al. „Development of a nursing-care assistant robot RIBA that can lift a human in its arms“. In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2010, S. 5996–6001 (zitiert auf den Seiten 13, 15).

[Onn+16] Linda Onnasch, Xenia Maier und Thomas Jürgensohn. *Mensch-Roboter-Interaktion-Eine Taxonomie für alle Anwendungsfälle*. Bundesanstalt für Arbeitsschutz und Arbeitsmedizin Dortmund, 2016 (zitiert auf den Seiten 15, 16).

[Pac+05] Elena Pacchierotti, Henrik I Christensen und Patric Jensfelt. „Human-robot embodied interaction in hallway settings: a pilot user study“. In: *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE. 2005, S. 164–171 (zitiert auf Seite 18).

[Par16] Europäisches Parlament. *Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung)*. Techn. Ber. Tech. Rep. Amtsblatt der Europäischen Union, 2016 (zitiert auf Seite 9).

[Pfe18] Anne Pfeifle. „Alexa, What Should We Do about Privacy: Protecting Privacy for Users of Voice-Activated Devices“. In: *Wash. L. Rev.* 93 (2018), S. 421 (zitiert auf Seite 1).

[Pol+02] Martha E Pollack, Laura Brown, Dirk Colbry et al. „Pearl: A mobile robotic assistant for the elderly“. In: *AAAI workshop on automation as eldercare*. Bd. 2002. 2002, S. 85–91 (zitiert auf Seite 15).

[Pov+11] Daniel Povey, Arnab Ghoshal, Gilles Boulianne et al. *The Kaldi speech recognition toolkit*. Techn. Ber. IEEE Signal Processing Society, 2011 (zitiert auf Seite 24).

[Pro+02] Plamen J Prodanov, Andrzej Drygajlo, Guy Ramel, Mathieu Meisser und Roland Siegwart. „Voice enabled interface for interactive tour-guide robots“. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Bd. 2. IEEE. 2002, S. 1332–1337 (zitiert auf den Seiten 17, 38).

[Roy+18] Nirupam Roy, Sheng Shen, Haitham Hassanieh und Romit Roy Choudhury. „Inaudible voice commands: The long-range attack and defense“. In: *15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18)*. 2018, S. 547–560 (zitiert auf Seite 11).

[Sau11] Jeff Sauro. *A practical guide to the system usability scale: Background, benchmarks & best practices*. Measuring Usability LLC, 2011 (zitiert auf Seite 67).

[SS17] Christine Storr und Pam Storr. „Internet of Things: Right to Data from a European Perspective“. In: *New Technology, Big Data and the Law*. Springer, 2017, S. 65–96 (zitiert auf Seite 10).

[ST17] Smruthi Sridhar und Matthew E Tolentino. „Evaluating Voice Interaction Pipelines at the Edge“. In: *2017 IEEE International Conference on Edge Computing (EDGE)*. IEEE. 2017, S. 248–251 (zitiert auf den Seiten 5, 6, 8).

[Wad+04] Kazuyoshi Wada, Takanori Shibata, Tomoko Saito und Kazuo Tanie. „Psychological and social effects in long-term experiment of robot assisted activity to elderly people at a health service facility for the aged“. In: *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*. Bd. 3. IEEE. 2004, S. 3068–3073 (zitiert auf Seite 13).

[Wal+00] Stefan Waldherr, Roseli Romero und Sebastian Thrun. „A gesture based interface for human-robot interaction“. In: *Autonomous Robots* 9.2 (2000), S. 151–173 (zitiert auf Seite 17).

[Wan+16] Ning Wang, Frank Broz, Alessandro Di Nuovo, Tony Belpaeme und Angelo Cangelosi. „A user-centric design of service robots speech interface for the elderly“. In: *Recent Advances in Nonlinear Speech Processing*. Springer, 2016, S. 275–283 (zitiert auf Seite 18).

[Yam+12] Kimitoshi Yamazaki, Ryohei Ueda, Shunichi Nozawa et al. „Home-assistant robot for an aging society“. In: *Proceedings of the IEEE* 100.8 (2012), S. 2429–2441 (zitiert auf Seite 15).

[Zen+18] Limin Zeng, Björn Einert, Alexander Pitkin und Gerhard Weber. „HapticRein: Design and Development of an Interactive Haptic Rein for a Guidance Robot“. In: *International Conference on Computers Helping People with Special Needs*. Springer. 2018, S. 94–101 (zitiert auf Seite 17).

[Ini19] Initiative D21 e.V. *D21 Digitial Index 2018/2019 Jährliches Lagebild zur Digitalen Gesellschaft*. 2019 (zitiert auf Seite 2).

Abbildungsverzeichnis

Tabellenverzeichnis

Dokumentation Prototyp

Studienergebnisse

	Strongly Disagree 1	2	3	4	Strongly Agree 5	abstention
I think that I would like to use this system frequently.	1	3	7	1	1	0
I found the system unnecessarily complex.	3	6	0	2	0	2
I thought the system was easy to use.	0	0	3	7	3	0
I think that I would need the support of a technical person to be able to use this system.	9	1	1	1	1	0
I found the various functions in this system were well integrated.	1	0	5	5	1	1
I thought there was too much inconsistency in this system.	2	3	2	4	1	1
I would imagine that most people would learn to use this system very quickly.	0	1	3	4	5	0
I found the system very cumbersome to use.	5	5	2	1	0	0
I felt very confident using the system.	1	1	2	4	4	1
I needed to learn a lot of things before I could get going with this system.	9	2	0	0	1	1

Ergebnisse der Studie, SUS Fragen

	Strongly Disagree 1	2	3	4	Strongly Agree 5	abstention
The spoken answers by the system were very clear and natural.	2	6	3	2	0	0
The system responded quickly to my requests.	0	4	4	5	0	0
The system had no problems to correctly understand my requests.	0	6	3	3	1	0
The interaction with the system felt natural.	3	2	1	5	2	0
I had no problems using the wakeword.	1	2	0	2	8	0
I found the delay until the system answered disturbing in the interaction.	1	1	4	6	1	0
When interacting with a voice assistant, answers should be given as quickly as possible even when the answer sounds more like a machine.	2	3	1	4	2	1
It is useful, that the robot is offering help, when he sees me.	2	1	2	3	5	0
Privacy is very important for me.	0	1	3	3	6	0
I would use a voice assistant more often, if my privacy is guaranteed.	3	3	1	2	4	0
When interacting with a voice assistant, it's very important, that answers sound natural even when they take a bit longer.	1	3	3	2	4	0
I am already using a voice assistant	yes: 4 no: 9 abstention: 0					

Ergebnisse der Studie: systemspezifische Fragen