UNIVERSITÄT KOBLENZ · LANDAU

Fachbereich 4: Informatik

COMPUTERVISUALISTIK

Evaluation einer sprachgesteuerten Lösung in der Produktion mit multimodaler Ausgabe

Masterarbeit

Zur Erlangung des Grades Master of Science (M. Sc.)
im Studiengang Wirtschaftsinformatik

vorgelegt von

Sikandar Khan
Matrikelnr. 207110092
skhan@uni-koblenz.de

Erstgutachterin:	Prof. Dr. Karin Harbusch Institut für Computervisualistik
Zweitgutachter:	Denis Memmesheimer Institut für Computervisualistik

Koblenz, im Januar 2018

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Mit der Einstellung dieser Arbeit in die Bibliothek bin ich einverstanden. Ja ☒ Nein ☐

Koblenz, 21.01.2018 .....

(Ort, Datum) (Unterschrift)

Zusammenfassung

Die vorliegende Masterarbeit thematisiert die Evaluation einer sprachgesteuerten Lösung in der Produktion mit multimodaler Eingabe. Dabei wurden die Usability und die Benutzerfreundlichkeit eines gewählten Sprachdialogsystems bewertet. Die Bewertung wurde mit Hilfe von Benutzertests und eines modifizierten SASSI-Fragebogens durchgeführt. Weiterhin wurden auch technische Kriterien, wie die Wortfehlerrate und die Out-of-grammar Rate zur Hilfe gezogen.
Für den Versuch wurden zwei verschiedene Szenarien aus einer realen Produktionsumgebung definiert. Dabei sollten die Teilnehmer verschiedene Aufgaben mit Hilfe des Testsystems erledigen. Die Interaktion mit dem Sprachdialogsystem fand anhand von gesprochenen Befehlen statt, welche durch eine Grammatik definiert wurden. Die Sprachkommandos wurden durch die Zuhilfenahme eines WLAN-Headsets an das Sprachsystem übertragen. Während des Versuchs wurden Aussagen der Teilnehmer protokolliert und die technischen Kriterien notiert.

Das Ergebnis der Evaluation verdeutlicht, dass das Sprachdialogsystem eine hohe Qualität bezüglich Usability und Benutzerfreundlichkeit aufweist. Dabei sind die Wortfehlerrate und die Out-of-grammar Rate sehr niedrig ausgefallen und das System wurde von den Benutzern deutlich positiv bewertet. Nichtsdestotrotz wurden einige Kritikpunkte genannt, die zu einer Verbesserung des Systems beitragen können.

Abstract

The present master thesis deals with the evaluation of a voice-controlled solution in the production environment with multimodal input. In the process, the usability and user-friendliness of a selected speech dialogue system were evaluated. The evaluation was carried out with the help of user tests and a modified SASSI questionnaire. Furthermore, technical criteria such as the word error rate and the out-of-grammar rate were also involved.

Two different scenarios from a real production environment were defined for the experiment. The participants had to do different tasks with the help of the test system. The interaction with the speech dialogue system took place on the basis of spoken commands, which were defined by a grammar. The voice commands were transmitted to the voice system with the aid of a WLAN headset. Statements of the participants were recorded and the technical criteria were noted during the experiment.

The result of the evaluation shows that the speech dialogue system has a high quality in terms of usability and user-friendliness. The word error rate and the out-of-grammar rate have been very low and the system was rated very positively by the users. Nonetheless, some criticisms have been mentioned that can help to improve the system.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einführung

Die menschliche Sprache gehört so selbstverständlich zu unserem Alltag, dass wir uns kaum Gedanken über die Komplexität dieser Art von Kommunikation machen. Erst wenn man versucht die verbale Kommunikation mit einem Computer zu verwirklichen, zeigt wie anspruchsvoll und schwierig dieses Vorhaben sein kann (vgl. Pfister & Kaufmann 2017).

Die gesprochene Sprache ist die natürlichste Form der Kommunikation und hat zum großen Vorteil, dass viele Informationen in kurzer Zeit ausgetauscht werden können. Diese Gründe führen dazu, dass für die Entwicklung vieler Anwendungen der Einsatz von Sprache in der Mensch-Maschine Interaktion als Ziel angesehen wird (vgl. Schukat-Talamazzini 1995). Die Mensch-Maschine-Kommunikation hat in den vergangenen Jahrzehnten immer mehr an Bedeutung gewonnen. Dafür sind nicht nur die komplex zu bedienenden elektronischen Endgeräte, sondern auch die rasche Leistungssteigerung von Computern und Embedded-Systems[1] (vgl. Bender 2005), wie Mobiltelefone, verantwortlich (vgl. Schenk, Rigoll 2010).

Im Industriesektor wird es im Zeitalter von Industrie 4.0[2] (vgl. Schwab 2016) für Produktionsunternehmen immer wichtiger ihre Produktionsprozesse durch IT-Systeme zu unterstützen, um ihre Wettbewerbsfähigkeit zu erhalten und zu steigern. Durch das Internet und die weitere Entwicklung der Automatisierung wird durch den Einsatz von Sensoren und eingebetteten Systemen der Grad an Beobachtbarkeit der Fertigungsprozesse weiter erhöht. Somit entsteht ein Abbild der realen Welt. Die entstehenden Daten müssen zur Nutzung gefiltert, aggregiert und entsprechend dargestellt werden. Dabei sollen die Systeme nur die relevanten Informationen zur Verfügung stellen und das möglichst einfach, schnell und kontextbezogen. Diese in Echtzeit entstehenden Daten können mit Hilfe von

[1] Ist der integrierte Teil einer Maschine oder eines Geräts und ist nach außen nicht als Rechner, sondern nur als Träger intelligenter Systemfunktionen erkennbar.

[2] Wird als die vierte industrielle Revolution bezeichnet.

internetbasierten Diensten verarbeitet und für Regelprozesse genutzt werden. Dadurch wird eine Systematisierung und Selbststeuerung des Produktionsprozesses unterstützt (vgl. Scheer 2013).

Die Erfassung der Informationen aus der physikalischen Welt führt zu einem impliziten Expertenwissen über Maschinen und Fertigungsprozesse, das wiederrum zu einem besseren Verständnis über die Wirkungszusammenhänge in der Produktion führt. Trotz der Automatisierung und der autonomen funktionsweise der Systeme ist es notwendig, dass der Mensch die Prozesse plant und Entscheidungen trifft, wenn Probleme auftreten und Änderungen notwendig werden (Scheer 2013). IT-Systeme, welche die Steuerung und Kontrolle der Produktion, sowie die Bereitstellung, Verwendung und Nachverfolgbarkeit von Real-Time-Daten über den gesamten Produktionsprozess unterstützen, werden als Manufacturing Execution Systems (MES) bezeichnet (vgl. MESA 2000).
Ein Beispiel für solch ein MES ist die iTAC.MES.Suite des Unternehmens iTAC Software AG. Neben der Traceability mit einer durchgehend dokumentierten Chargenerfassung, bietet das System auch Funktionen um Falschbestückungen zu erkennen und entsprechend zu melden (vgl. iTAC 2017).

Die ME-Systeme (MES) und Maschinen in der Produktion arbeiten zwar automatisiert, allerdings müssen sie vom Menschen überwacht und gewartet werden. Wie kann nun der Mensch bei seiner Tätigkeit in der Produktion, neben den Werkzeugen, welche vom ME-System selbst bereitgestellt werden, weiter unterstützt werden? Dieser Leitgedanke führt zu der Überlegung, eine sprachgesteuerte Lösung in Form eines Dialogsystems bereitzustellen, um den Menschen effizienter bei seiner Arbeit zu unterstützen.

Solch eine sprachgesteuerte Lösung findet sich zum Beispiel in der Logistikbranche zur Unterstützung des Menschen bei der Kommissionierung wieder, welche als Pick-by-Voice[3] (vgl. Theel 2015, S.20) System bezeichnet wird (vgl. Theel 2015). Es ist denkbar,

[3] Der Kommissionierer erhält die zu bearbeitenden Aufträge als akustische Anweisungen über eine Sprachgarnitur und kann diese per Spracheingabe bestätigen.

dass eine ähnliche sprachgesteuerte Lösung auch in der Produktion eingesetzt werden kann, um den Menschen bei seiner Tätigkeit zu unterstützen.

Das Ziel dieser Arbeit ist es, drei verfügbare Spracherkennungssysteme auf dem Markt entsprechend ihrer Tauglichkeit für die Anbindung an ein MES anhand der „Vergleichenden Evaluation“ auszuwerten und das gewählte Spracherkennungssystem anhand der empirischen Evaluation hinsichtlich Benutzerfreundlichkeit und Bedienbarkeit zu evaluieren.

Im ersten Teil der Arbeit wird die theoretische Grundlage anhand der Literatur gelegt. Hierbei werden grundlegende Begrifflichkeiten erklärt. Darüber hinaus wird das Konzept der Sprachsynthese dargestellt und ein Überblick über die Funktionsweise von Dialogsystemen gegeben. Im Anschluss werden die verschiedenen Dialogstrategien erläutert.

Im zweiten Teil der Masterarbeit werden drei Spracherkennungssysteme ausgewählt und im Hinblick auf die Integration an ein MES evaluiert. Dabei werden verschiedene Anforderungskriterien definiert, anhand dessen die Spracherkennungssysteme ausgewertet werden. Im Anschluss wird mithilfe der vergleichenden Bewertung ein Spracherkennungssystem für die bewertende Evaluation ausgewählt.

Im letzten Teil der Arbeit werden einige Anwendungsszenarien aus der Praxis erläutert. Mithilfe dieser Anwendungsszenarien wird das gewählte Spracherkennungssystem anhand von Task-Szenarien und Fragebögen evaluiert. Die Ergebnisse werden im Fazit zusammengefasst und es wird eine Empfehlung zur Verbesserung des Systems ausgesprochen.

2 Theoretische Grundlagen

Die Ausführungen dieser Masterarbeit werden mit einigen theoretischen Grundlagen in diesem Kapitel eingeleitet. Dazu werden einige Begrifflichkeiten erklärt um ein grundlegendes Verständnis für die Thematik zu schaffen und die dazugehörigen Zusammenhänge zu verdeutlichen.

2.1 Industrie 4.0

Die Kurzform „Industrie 4.0“ wird als die vierte industrielle Revolution bezeichnet. Als vorangegangene industrielle Revolutionen können folgende Innovationen aufgelistet werden (vgl. Schwab 2016):

- Ab etwa dem Jahr 1760 gilt die Erfindung der Dampfmaschine als erste industrielle Revolution.
- Bei der zweiten industriellen Revolution handelt es sich um das Konzept der arbeitsteiligen Massenfertigung mit Hilfe elektrischer Energie ab 1870.
- Ab ungefähr 1960 spricht man von der dritten industriellen Revolution, welche durch IT und Elektronik getrieben eine variantenreiche Serienproduktion ermöglichte.
- Die vierte industrielle Revolution ist geprägt durch die Integration sogenannter cyber-physischer Systeme in industrielle Prozesse (vgl. Kagermann et. al. 2013).

In Abbildung 1 werden die vier industriellen Revolutionen in Abhängigkeit zu dem Grad der Komplexität verdeutlicht.

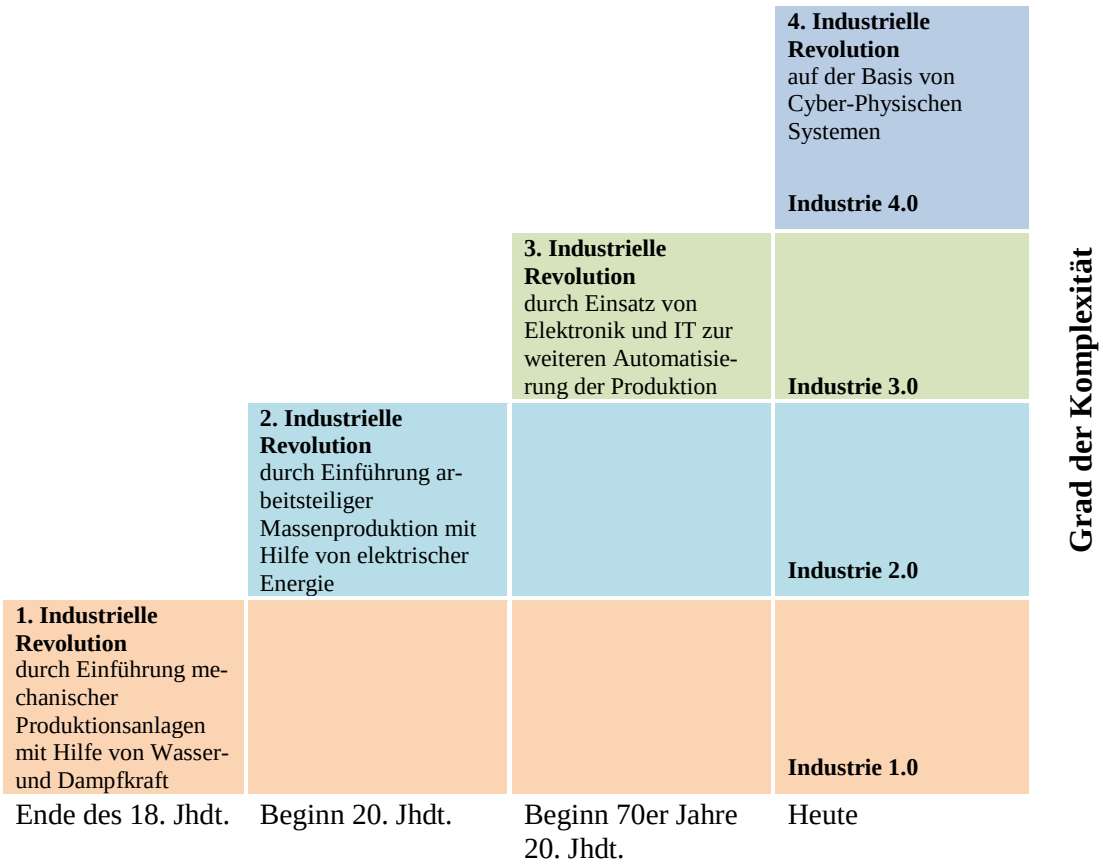

Abbildung 1: Die vier Stufen industrieller Revolutionen (Quelle: Schlick, Stephan, Zühlke 2012)

Der Kerngedanke von Industrie 4.0 beinhaltet die Aufhebung der Trennung zwischen physischer und virtueller Welt und somit die Verschmelzung der dinglichen Welt und deren digitaler Modelle (vgl. Bauernhansl 2017).

2.2 Cyber-physisches System

Der Begriff Cyber-physisches System oder im englischen auch „Cyber-Physical System“ (CPS) wurde bedeutend von Edward Lee geprägt und wird definiert als:

„Cyber-Physical Systems are integrations of computation with physical processes. Embedded computers and networks monitor and control the physical processes, usually with feedback loops where physical processes affect computations and vice versa“ (vgl. Pinnow und Schäfer 2017, S. 135).

Diese Definition ist allerdings zu allgemein gehalten (vgl. Uhlmann et al. 2013), denn demnach wären alle digitalen Regler, Steuerungen und Prozessleitsysteme der letzten 50 Jahre Cyber-physische Systeme. CPS sind mehr als klassische Automatisierungssysteme. Sie sind datentechnisch vernetzte Produktionsanlagen, Produkte und Materialien sowie Transporttechnologien, welche ihre Nutzung und den Ablauf der Bearbeitungsprozesse autonom organisieren, steuern und an externe Anforderungen anpassen. Ein weiteres Merkmal ist die Vernetzung dieser intelligenten Komponenten und Teilsysteme durch die Verfügbarkeit einer informationstechnischen Infrastruktur in Form von industriell einsetzbaren Internetverbindungen, die sich konzeptionell an dem Begriff „Internet der Dinge“ orientieren (vgl. Uhlmann et al. 2013).

Dadurch soll die Integration von realer und virtueller Welt ermöglicht werden. Dies führt zum Zusammenwachsen von Produkten, Geräten und Objekten mit eingebetteter Software zu verteilten und integrierten Systemen, wie in Abbildung 2 verdeutlicht.
Die Thematik dieser Arbeit bewegt sich im rot-markierten Bereich und beschäftigt sich mit der Mensch-Maschine-Kommunikation.

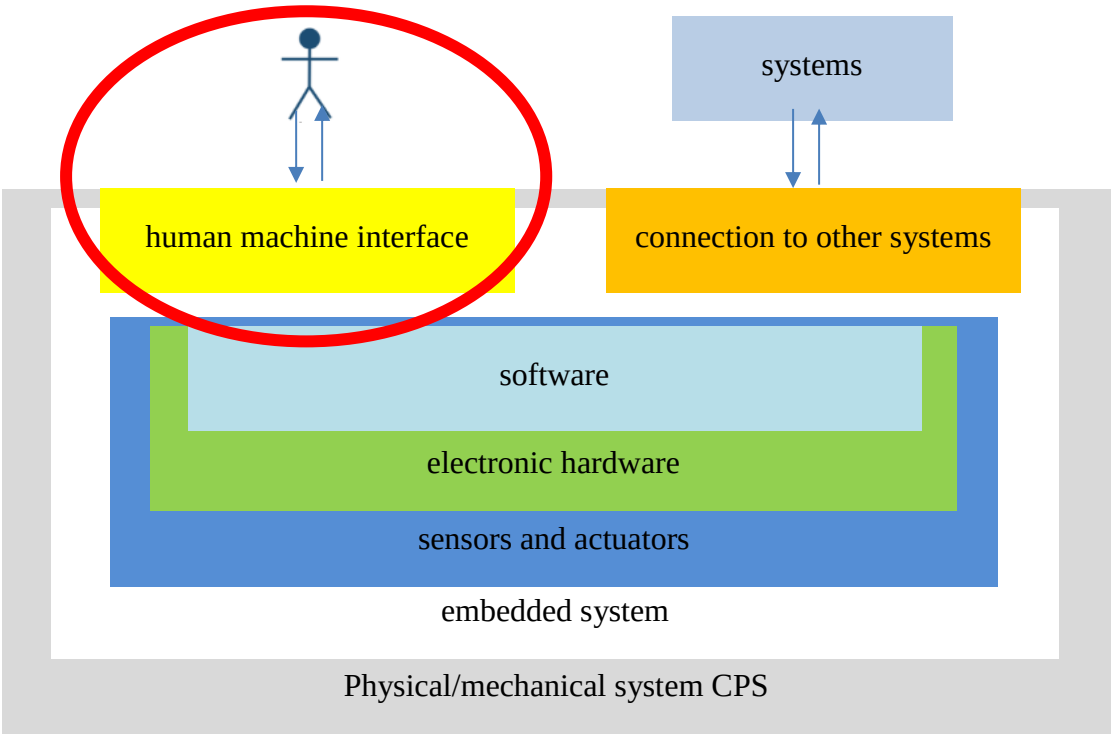

Abbildung 2: Modell eines Cyber-physischen Systems (Quelle: Broy 2010)

2.3 Manufacturing Execution System

Die Manufacturing Enterprise Solutions Association (MESA) besteht aus mehreren Vertretern von Industrie, Softwareherstellern und Unternehmensberatern und ist eine Non-Profit-Organisation. Sie wurde ins Leben gerufen, um einen Informationsaustausch untereinander über die Möglichkeiten und die Integration von produktionsnaher Informationstechnologie sowie die Verbreitung von MES zu führen (vgl. MESA 2000).

Die MESA definiert ein MES als ein System, das die Bereitstellung von Real-Time-Daten über den gesamten Produktionsprozess als Aufgabe übernimmt. Weiterhin zählen die Steuerung und Kontrolle der Produktion, eine schnelle Reaktionsfähigkeit bei Abwei-

chungen, sowie die Integration der Produktionsprozesse in die ERP-Systeme durch Anbindung der Automatisierungssysteme und die Unterstützung der manuellen Produktionsprozesse zu den wichtigen Merkmalen eines MES (vgl. MESA 2000).

2.4 Sprachsynthese und Sprachsynthese Systeme

Der Einsatz von Sprachsynthese findet überall dort statt, wo es erforderlich ist die Ausgabe von Information auf akustischem sprachlichem Weg zu leiten. Heutzutage wird die Sprachsynthese zunehmend in Auskunftssystemen, wie zum Beispiel Navigationssysteme, Reiseauskünfte, Verkehrsmeldungen, eingesetzt. Dabei kann die Sprachsynthese als ein zweistufiger Prozess beschrieben werden. Zuerst wird der Eingabetext linguistisch analysiert und danach wird die aus der Analyse resultierende linguistische Repräsentation in ein synthetisches Sprachsignal umgesetzt (vgl. Carstensen et. al. 2004, S. 517).

Das Sprachsynthesesystem, auch TTS-System (Text-to-Speech-System) genannt, ist ein komplexes System. Die Leistungsfähigkeit ist stark von der Qualität der einzelnen Komponenten des TTS-Systems abhängig.

3 Dialogsysteme und Dialogstrategien

Im praktischen Teil dieser Masterarbeit wird ein Sprachdialogsystem mit multimodaler Eingabe evaluiert. Daher wird in diesem Kapitel auf Dialogsysteme und die damit verbundenen Grammatiken eingegangen. Dazu wird in 3.1 die Funktionsweise eines Sprachdialogsystems erläutert. Im Abschnitt 3.2 wird das multimodale Dialogsystem veranschaulicht und in 3.3 werden die verschiedenen Darstellungsformen von Grammatiken erklärt.

3.1 Sprachdialogsysteme

Sprachdialogsysteme erlauben es einem Benutzer, mit einer Maschine mittels sprachlicher Ein- und Ausgabe zu kommunizieren und finden überall dort Anwendung, wo es dem Benutzer ermöglicht werden soll auf intuitive Art und Weise auf Informationen zuzugreifen (vgl. Carstensen et al. 2004, S. 532).

Als Urvater aller Dialogsysteme gilt das berühmte Eliza System, das eine Interaktion über Tastatur und Bildschirm, also in geschriebener Sprache, ermöglichte. Heutige Dialogsysteme sind aufgrund von leistungsfähigen Spracherkennungs- und Sprachsynthesekomponenten in der Lage gesprochene Benutzereingaben zu verarbeiten und ihre Antworten ebenfalls in Form von gesprochenen Äußerungen zu realisieren (vgl. Carstensen et al. 2004, S. 532).
Es folgt eine kurze Erläuterung der Anforderungen an die einzelnen Komponenten eines Sprachdialogsystems, welche in Abbildung 3 ersichtlich sind.

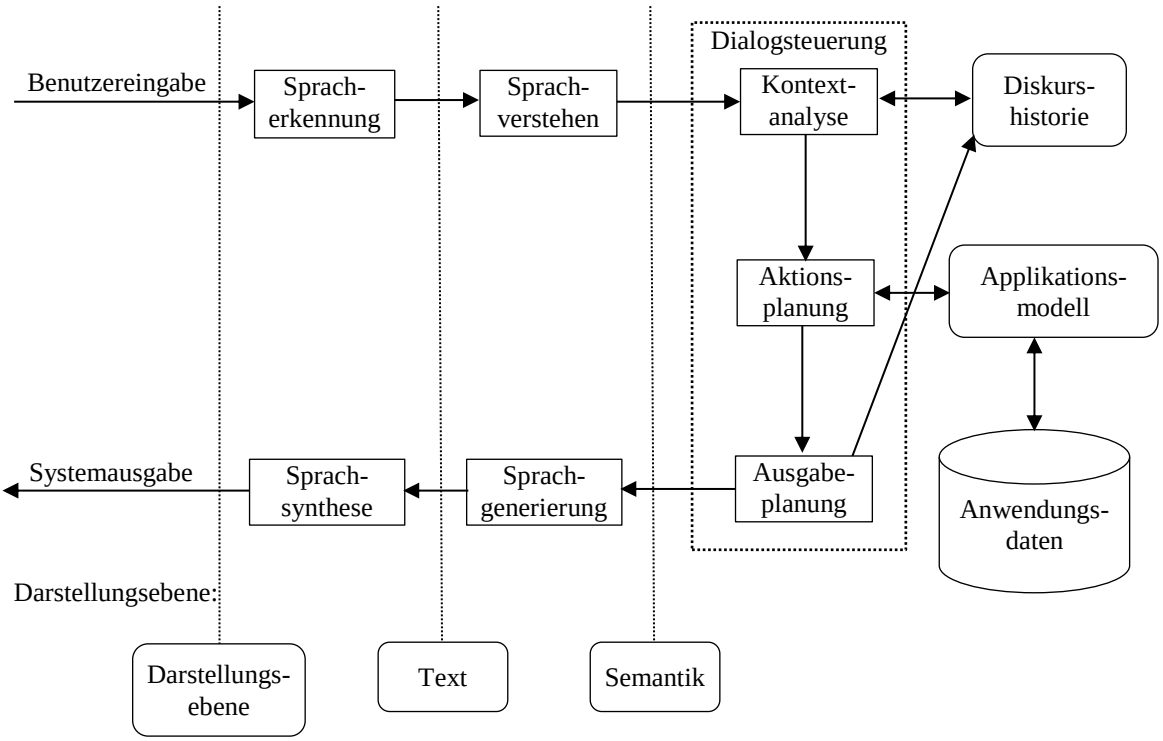

Abbildung 3: Architektur eines natürlichsprachlichen Dialogsystems
(Quelle: Carstensen et. al. 2004, S. 535)

Die sprachliche Eingabe des Nutzers liegt als Schallsignal vor und muss gefiltert, analysiert und einem definierten Satz an Symbolen zugeordnet werden. Dabei generiert die **Spracherkennungskomponente** eine Folge von Worthypothesen. Diese können im einfachsten Fall eine Sequenz an Wörter bilden oder aber komplexere Strukturen mit mehreren Hypothesen aufweisen. Die Sprachabdeckung eines Spracherkenners kann auf zwei Arten erfolgen, erstens eine regelbasierte Grammatik oder zweitens ein statistisches Sprachmodell.

Die regelbasierte Grammatik beinhaltet alle als Eingabe möglichen Konstituenten und Regeln, wie diese Konstituenten miteinander kombiniert werden können. Dies ist bei Sprachdialogsystemen mit geringer Funktionalität vergleichsweise einfach zu bewerkstelligen. Komplexe Sprachdialogsysteme erfordern hingegen viel Expertenwissen (vgl.

McTear 2002), denn Menschen können identische Anfragen auf unterschiedliche Weise ausdrücken (vgl. Fromkin et al. 2002).

Beim statistischen Sprachmodell werden Übergangswahrscheinlichkeiten zwischen Wörtern berechnet. Hierfür müssen verschiedenartige Trainingsdaten erfasst werden, um eine gute Qualität der Spracherkennung zu gewährleisten. Dieses Verfahren wird beispielsweise oft in Diktiererkennern genutzt, findet aber auch zunehmend als Analyseverfahren in Sprachdialogsystemen Verwendung (vgl. Hofmann, Ehrlich, Reichel et al. 2013). Eines der wichtigsten Performancekriterien für das Spracherkennungsmodul in einem Dialogsystem sind die Echtzeitanforderungen an die Verarbeitungsgeschwindigkeit. Denn das System sollte ohne große Latenzzeiten und unmittelbar nach Ende der Benutzeräußerung oder -eingabe eine entsprechende Reaktion aufrufen. Verzögerungen werden von dem Benutzer als störend und ineffizient empfunden. Weiterhin muss ein Dialogsystem eine Robustheit gegenüber den Charakteristika verschiedener Sprecher aufweisen und auch gegenüber Hintergrundgeräuschen und gestörten Sprachsignalen robust sein. Dies wird anhand von entsprechenden Signalverarbeitungsalgorithmen und durch die Verwendung von möglichst großen Trainingsdaten erreicht (vgl. Carstensen et al. 2004, S. 535-536).

Die nächste Komponente ist die **Sprachverstehenskomponente**, welche auf Grundlage der vom Erkenner gelieferten Satzhypothesen die Intention des Benutzers erkennt und die entsprechenden anwendungsrelevanten Informationen extrahiert. Dabei spielen auch Echtzeitfähigkeiten und Robustheit eine große Rolle, denn Benutzereingaben sind spontansprachliche Äußerungen, die unvollständig oder grammatikalisch inkorrekt sein können. Der Parser muss also in der Lage sein diese Faktoren zu berücksichtigen und die Äußerungen entsprechend korrekt zu verarbeiten. Des Weiteren erschweren zusätzliche Erkennungsfehler die syntaktische Analyse der Benutzereingabe. Diesbezüglich parsen viele Systeme den Satz nur partiell, d.h. es werden nur einzelne bedeutsame Abschnitte der Benutzeräußerung extrahiert, falls eine vollständige Interpretation nicht möglich ist. Die bedeutungstragenden Phrasen werden anhand von anwendungsspezifischen Grammatiken modelliert (vgl. Carstensen et al. 2004, S. 536).

Die Komponente **Dialogsteuerung** kann als Schaltzentrale eines Dialogsystems betrachtet werden. Dabei ist sie für die Interpretation der aktuellen Benutzeräußerung unter Berücksichtigung des Dialogkontextes, die Planung und Durchführung von Aktionen, wie z.B. die Interaktion mit externen Anwendungen und zuletzt für die Auswahl einer angemessenen Ausgabe an den Benutzer zuständig. Ein weiteres Merkmal der Dialogsteuerung ist die Diskurshistorie, welche alle bisherigen Äußerungen von System und Benutzer repräsentiert. Dies hat den Vorteil, dass nicht nur sprachliche Konstrukte wie Ellipsen oder Anaphern aufgelöst werden, sondern auch Fehleingaben erkannt und die entsprechenden Systemaktionen ausgelöst werden. Anhand dieser Historie ist es möglich, die von der Sprachverstehenskomponente extrahierte Semantik im Kontext des bisherigen Dialogverlaufs zu interpretieren. Der Abgleich des Systemwissens mit den Benutzereingaben erfolgt durch explizite oder implizite Verifikationsfragen. Dadurch erhält der Benutzer die Gelegenheit Fehleingaben zu korrigieren oder Informationen hinzuzufügen.
Die **Aktionsplanung** entscheidet welche Aktionen im aktuellen Dialogzustand durchzuführen sind und initiiert die Disambiguierungs- und Verifikationsfragen. Der Dialogablauf kann durch endliche Automaten oder durch planbasierte Ansätze beschrieben werden. Bei einfachen Anwendungen finden endliche Automaten ihre Anwendung. Sie beschreiben für jeden möglichen Dialogzustand die möglichen Benutzereingaben und die erforderlichen Systemreaktionen. Die planbasierten Ansätze eigenen sich gut für komplexe mixed-initiative Dialogsysteme (vgl. Carstensen et al. 2004, S. 537).
Durch die **Ausgabeplanung** werden die entsprechenden Systemäußerungen bestimmt und in Form von Sprachakten repräsentiert. Diese werden von den nachfolgenden Komponenten in textuelle und akustische Repräsentation umgesetzt. Dabei liefert die Aktionsplanung vorgefertigte Ausgabetemplates, sodass nur die entsprechenden Variablenwerte eingefügt werden müssen. Aufgrund dessen wird die Sprachgenerierung als separates Modul übergangen (vgl. Carstensen et al. 2004, S. 537).
Werden die Systemäußerungen nur in abstrakter semantischer Form vorgegeben, so ist es die Aufgabe des **Sprachgenerierungsmoduls**, diese in eine textuelle Darstellung umzusetzen. Anschließend wird diese textuelle Darstellung mit Hilfe eines Sprachsynthesesystems ausgegeben (vgl. Carstensen et al. 2004, S. 537).

3.2 Multimodale Dialogsysteme

Bei multimodalen Dialogsystemen erfolgt die Kommunikation zwischen Mensch und Maschine über mehrere Ein- und Ausgabemodalitäten. Die Kombination von mehreren Ein- und Ausgabemodalitäten führt zu einer deutlichen Effizienzsteigerung und erhöhter Benutzerfreundlichkeit, denn je nach Präferenz und Aufgabenstellung kann der Benutzer für seine Eingabe die passende Eingabemodalität wählen oder mehrere Modalitäten kombinieren. Weiterhin ist für die Systemausgaben die Kombination verschiedener Modalitäten von Vorteil, denn es lassen sich einige Informationen, wie z.B. lange Listen oder komplexe Zusammenhänge, textuell oder grafisch effizienter beschreiben als durch die Ausgabe von Sprache (vgl. Carstensen et al. 2004, S. 538).
Neben den spezifischen Analysemodulen ist für die Realisierung von multimodaler Eingabe eine weitere Verarbeitungskomponente, die Medienfusion, notwendig. Diese Komponente kombiniert die Informationen aus den einzelnen Modalitäten und leitet sie an die Dialogsteuerungseinheit weiter.

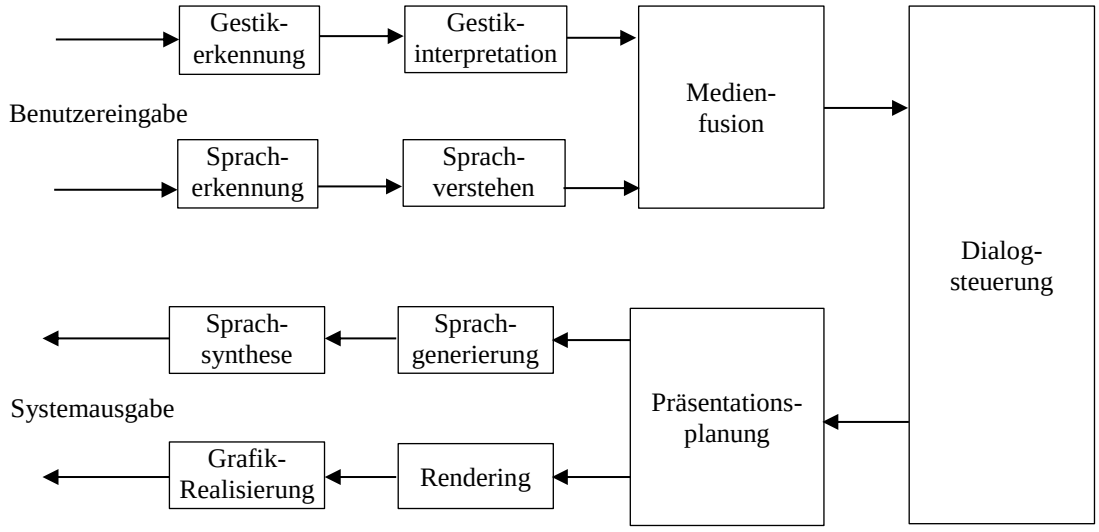

Abbildung 4: Architektur eines multimodalen Dialogsystems
(Quelle: Carstensen et. al 2004, S. 538)

3.3 Grammatiken

Für spracherkennende Systeme stellt die Grammatik einen der essentiellen Entwurfsparameter dar. Ein befehlsbasiertes Sprachsystem, beispielsweise das Navigationssystem, akzeptiert nur wenige Befehle. Anders ist es bei einem Diktiergerät. Hierbei möchte man dem Anwender keine Beschränkungen auferlegen (vgl. Euler 2006).
Für Grammatiken ergeben sich vielfältige Anwendungsbeispiele, die von der linguistischen Beschreibung von Sprachen bis hin zur Definition von Programmiersprachen reichen. Bei der Mensch-Maschine-Kommunikation treten Grammatiken bei natürlichsprachlichen Systemen und bei der Modellierung von Dialogen auf (vgl. Schenk und Rigoll 2010, S. 93).

Kontextfreie Grammatiken, im englischen auch „context free grammars“ bezeichnet (CFGs), sind für Einsatzgebiete mit einfacher Grammatik geeignet. Durch CFGs kann eine endliche Menge an syntaktischen Regeln definiert werden. Beim Konzept der kontextfreien Grammatik erfolgt die Analyse (engl. „parsing“) und die semantische Interpretation einer Äußerung automatisch. Allerdings ist das Konzept bei stark variierenden Benutzeräußerungen, aufgrund der großen Anzahl an möglichen Formulierungen, die nur schwer durch eine kontextfreie Grammatik darstellbar ist, eher weniger geeignet (vgl. Levasseur und Doerr 2017). CFGs können durch das Set G = {V, T, P, S} beschrieben werden mit (vgl. Schenk und Rigoll 2010, S. 94 – 95):
V ≡ Variable z. B. <Ausdruck> (normalerweise in Großbuchstaben angegeben).
T ≡ Terminale z. B. „x“ (normalerweise in Kleinbuchstaben angegeben).
P ≡ Produktionsregel z. B. <Ausdruck> → „x“
S ≡ Startsymbol.
Die Produktionsregeln sind dabei von der Form $A \rightarrow \alpha$, mit $A \in \{V\}$ und $\alpha \in \{V \cup T\}$. Wenn für dieselbe Variable A mehrere Produktionsregeln bestehen, z. B. $A \rightarrow \alpha$; $A \rightarrow \beta$; $A \rightarrow \gamma$, dann kann dafür auch $A \rightarrow \alpha \mid \beta \mid \gamma$ geschrieben werden. Somit wird in einer CFG jede Variable unabhängig von ihrem Kontext durch eine Regel ersetzt (vgl. Schenk und Rigoll 2010, S. 95).

Für die Darstellung von Grammatiken existieren zwei Normalformen. Eine Form zur Darstellung einer Grammatik ist die **Chomsky-Normalform** (CNF). Eine CNF enthält Produktionsregeln bei denen auf der rechten Seite entweder nur zwei Variablen oder ein terminaler Ausdruck stehen. Ein Beispiel dafür zeigt die folgende Darstellung (vgl. Schenk und Rigoll 2010):
A → BC oder A → a.
Hierbei mit A, B, C ∈ {T}.

Die zweite Form zur Darstellung einer Grammatik für spracherkennende Systeme stellt die **Backus-Naur Form** (BNF) dar. Bei der BNF werden in den Ableitungsregeln Nichtterminalsymbole, die auch syntaktische Variablen genannt und durch spitze Klammern (‚<' und ‚>') gekennzeichnet werden, definiert. Alternativen werden durch einen senkrechten Strich (‚|') dargestellt. Weiterhin können auch Terminalfolgen (Sequenzen) definiert werden, die sowohl Terminal- als auch Nichtterminalsymbole enthalten und Wiederholungen werden durch Rekursion dargestellt. Ein Beispiel für eine Rekursion wäre die folgende Darstellung (vgl. Schenk und Rigoll 2010, S. 96):
<A> → a | <A>.

Beispielsweise kann man die Regeln A → B1; A → B2; A → Bn, welche Teil einer Grammatik sind, folgendermaßen in BNF definieren:
<A> ::= <B1> | <B2> | <Bn> (vgl. Levasseur und Doerr 2017).

Ein konkretes Beispiel für eine Sprachverarbeitungsregel wird anhand des Wortes „Bestätigen" mit seinen Alternativen verdeutlicht:
„bestätigen" → „akzeptieren"; „bestätigen" → „okay"; „bestätigen" → „annehmen."
In der Backus-Naur Form wird diese Regel folgendermaßen deklariert:
<bestätigen> ::= <akzeptieren> | <okay> | <annehmen>.

3.4 Dialogstrategien für Sprachdialogsysteme

In der Mensch-Maschine Interaktion kann die Dialoginitiative klassifiziert werden in: Vom Nutzer geführter, vom System geführter und gemischt-geführter Dialog (vgl. Karat et al. 2012). Dabei weist jede dieser Dialogvarianten verschiedene Ausprägungen und unterschiedliche Einsatzfelder auf, welche im Folgenden erläutert werden.

3.4.1 System-geführter Dialog

Bei dieser Dialogvariante steht das System im Vordergrund und dem Nutzer wird eine passive Rolle zugeteilt (vgl. Karat et al 2012). Der Dialogablauf wird anhand des Stellens von direkten Fragen durch das System bestimmt. Nutzer müssen auf diese Fragen antworten, um den Dialog fortzusetzen. Dabei wird durch die gezielte Formulierung der Frage die Antwortvarianz eingeschränkt (vgl. Kamm 1995).

Ein Beispiel für den system-geführten Dialog ist das Navigationssystem:

System: „Wie lautet der Zielort?“
Nutzer: „Berlin“
System: „OK Berlin, wie lautet die Straße?“
Nutzer: „Goethe Straße“
System: „Navigation nach Berlin, Goethe Straße starten? Sagen Sie nein für Korrekturen.“
Nutzer: „Ja“

Durch die vorgegebenen Antwortmöglichkeiten anhand der restriktiven Spracherkennungsgrammatik können dem Nutzer nur die erfragten Informationen zur Verfügung gestellt werden. Weiterhin ist die Einführung neuer Themen mit relativ viel Aufwand verbunden, da die Grammatik entsprechend angepasst werden muss. Allerdings ist bei Laien ohne Systemerfahrung ein solch systemgeführter Dialog von Vorteil (vgl. Karat et al. 2012).

3.4.2 Nutzer-geführter Dialog

Beim nutzer-geführten Dialog werden die Dialogschritte allein durch den Nutzer bestimmt. Dabei können offene Fragen an das System gestellt werden und das System übernimmt die passive Rolle des Antwortgebers. Das folgende Beispiel zeigt einen entsprechenden Dialog einer Navigationsanwendung:

Nutzer: „Navigation nach Köln zum Marktplatz."
System: „OK, Navigation zum Marktplatz in Köln wird gestartet."
Nutzer: „Aber bitte die Autobahn vermeiden."
System: „Routenoptionen geändert, keine Autobahn zugelassen."
Bei dieser Variante erfordert der Nutzer genaue Kenntnis darüber, was das System versteht und was nicht. Diese Art der Dialogführung ist daher eher für Experten als für Systemneulinge gedacht (vgl. Karat et al. 2012).

3.4.3 Gemischt-geführter Dialog

Beim gemischt-geführten Dialog werden die Vorteile beider zuvor genannten Dialogvarianten kombiniert. Diese Form der Bedienung eines Sprachdialogsystems stellt die einfachste und natürlichste Art der Dialogsteuerung für den Nutzer dar. Der Dialogablauf wird abhängig vom Kontext entweder vom System oder vom Nutzer geführt. Der Nutzer kann dem System dadurch Informationen in beliebiger Reihenfolge zur Verfügung stellen und dem System wird, bei unpräzisen Aussagen, die Möglichkeit eingeräumt explizit nachzufragen. Im Folgenden liegt die Dialoginitiative beim Nutzer und wechselt aufgrund unvollständiger Informationen zum System:

Nutzer: „Navigation nach Düsseldorf."
System: „Zu welcher Straße in Düsseldorf möchten Sie navigieren?"
Nutzer: „Zum Friedrich-Ebert-Ring bitte."
System: „Soll die Navigation zum Friedrich-Ebert-Ring gestartet werden?"
Nutzer: „Ja, aber ohne Autobahn."
System: „Routenoptionen geändert, Autobahn wird vermieden."

Ein gemischt-geführter Dialog ist, aufgrund des Überbeantwortens (engl. over-answering) ein guter Kompromiss für Experten und Laien (vgl. Jokinen und McTear 2010).

3.4.4 Adaptiver Dialog

Adaptive Dialogstrategien passen sich an den Nutzer und deren Situation an und stehen dabei in der aktuellen Forschung im Fokus. Bertrand (2014) entwickelte eine adaptive Dialogsteuerung, die den Dialogfluss in Abhängigkeit von kontextuellen Informationen anpasst.

Aufgrund der bedingt eingrenzbaren Nutzereingaben sind, aus technischer Sicht, nutzergeführte, gemischt-geführte und adaptive Dialogalternativen hoch komplex und stellen hohe Anforderungen an die automatische Spracherkennung und das Verständnis von gesprochener Sprache dar.

4 Evaluationstechniken für Sprachdialogsysteme

Die Evaluation wird nach Sonntag (1999, S. 7) definiert als „*die Bestimmung eines Wertes.“* Im Konkreten wird dabei einem Gegenstand ein bestimmter Wert zugeordnet. Bei dieser Zuordnung muss ein bestimmter Maßstab bestehen, nachdem die Zuordnung erfolgt, anderenfalls ist die Zuordnung nicht sinnvoll. Dabei muss die Messmethode genau das messen, was zu messen ist, das als Validität bezeichnet wird. Die Reliabilität bezeichnet die Tatsache, dass die Messergebnisse reproduzierbar sein müssen (vgl. Carstensen et al. 2010).

Ziel dieser Arbeit ist eine Evaluation durchzuführen, um die Usability des angebundenen Spracherkennungssystems an ein MES auszuwerten. Anhand der Resultate wird ein Fazit erstellt und es werden mögliche Verbesserungsvorschläge getätigt.

4.1 Evaluationstechniken

Im Allgemeinen existieren zwei wesentliche Gründe für die Evaluierung von Softwaresystemen: Erstens eine Evaluation, um als Ergebnis eine Systemverbesserung hervorzubringen und zweitens eine Bewertung, um Systeme miteinander zu vergleichen (vgl. Zühlke 2012).

Die aufgeführten Evaluationstechniken können, einerseits anhand der eingesetzten Methoden unterteilt werden. Dabei wird zwischen Testing (Testmethoden), Inspection (Kontrollmethoden) und Inquiry (Erkundungsmethoden) unterschieden. Andererseits können Evaluationsverfahren auch nach dem Kriterium der Vorgehensweise und der dabei eingesetzten Methoden in den drei Gruppen, formale, heuristische und empirische Evaluation unterteilt werden (vgl. Zühlke 2012).

4.1.1 Unterteilung anhand der eingesetzten Methoden

Bei der Unterteilung anhand der eingesetzten Methoden unterscheidet man zwischen den Testmethoden (Testing), Kontrollmethoden (Inspection) und Erkundungsmethoden (Inquiry).

Testing

Beim Testing werden durch repräsentative Nutzer typische Aufgaben mit dem Prototyp erledigt. Die Ergebnisse werden benutzt um festzustellen, ob die Nutzer bestimmte Aufgaben mit dem System erfüllen können (vgl. Zühlke 2012).

Inspection

Die Inspection liefern Ergebnisse über die Wünsche, Probleme, Anforderungen und das Verständnis der Nutzer. Diese Ergebnisse werden durch direktes Reden mit den Nutzern, durch das Beobachten der Nutzer bei der Arbeit oder durch mündliche oder schriftliche Fragen an den Nutzer erzielt (vgl. Zühlke 2012).

Inquiry

Bei den Erkundungsmethoden werden bestimmte Benutzbarkeitsprobleme der Schnittstelle durch Ergonomie-Experten selbst untersucht (vgl. Zühlke 2012).

4.1.2 Unterteilung nach der Vorgehensweise und den eingesetzten Methoden

Die Unterteilung nach der Vorgehensweise und den dabei eingesetzten Methoden umfasst die drei Gruppen der formalen, heuristischen und empirischen Evaluation.

Formale Evaluation

Die formale Evaluation fokussiert die Bedienung interaktiver Systeme. Dabei wird sie unterteilt in Ziele, Operationen, Methoden und Selektionsregeln. Sie basiert auf dem GOMS-Modell nach Card, Moran und Newell (1983).

Heuristische Evaluation

Die heuristische Evaluation ist ein methodisches Vorgehen, um Probleme bei der Nutzung des Systems zu identifizieren. Dabei werden Usability-Probleme mit Hilfe von Richtlinien erkannt. Als Grundlage für die heuristische Evaluation dient ein bestehender Prototyp des Systems. Hierbei ist, je nach Entwicklungsstand des Prototyps, von einer kalten, warmen und heißen Einschätzung die Rede. Bei der kalten Einschätzung existiert das System nur auf dem Papier, bei der warmen Einschätzung existiert nur eine Teilfunktion des Systems und bei der heißen Einschätzung handelt es sich bei dem System um ein nahezu vollendetes Produkt (vgl. Zühlke 2012).

Die heuristische Evaluationsmethode unterteilt sich erstens in die iterative heuristische Expertenevaluation und zweitens in die Focus-Group-Research Methode. Bei der iterativ-heuristischen Expertenevaluation wird auf Basis von Heuristiken ein Untersuchungsgegenstand auf mögliche Probleme für die Endnutzer untersucht. Dabei wird die Evaluation von mindestens einem Tester durchgeführt und die Anzahl der gefundenen Bedienprobleme steigt degressiv mit zunehmender Anzahl der Testpersonen. Dieser Zusammenhang wird in Abbildung 5**Fehler! Verweisquelle konnte nicht gefunden werden.** verdeutlicht (vgl. Zühlke 2012).

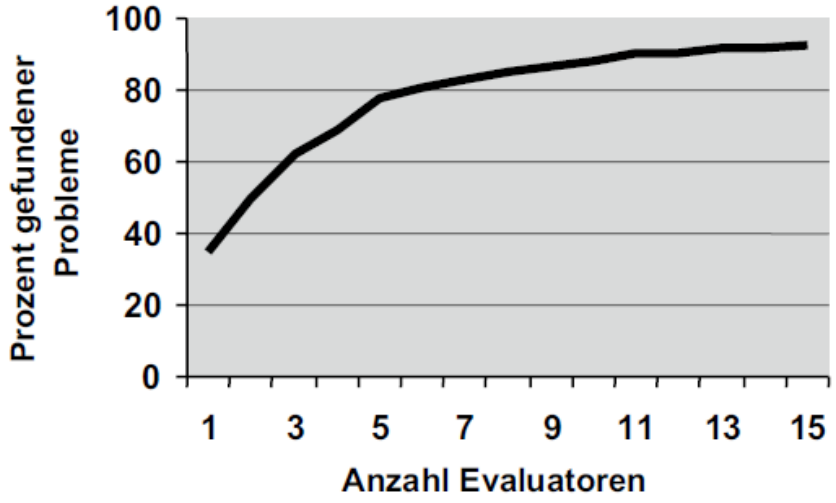

Abbildung 5: Verhältnis zwischen Anzahl der Testpersonen und dem Anteil gefundener Probleme (Quelle: Nielsen 1993)

Somit können fünf Evaluationsteilnehmer bereits ca. 80% der Usability-Probleme identifizieren. Dabei ist festzustellen, dass bei mehr als zehn Testpersonen nur noch marginale Verbesserungen zu erwarten sind. Bei dieser Methode werden als erstes die Heuristiken, nach denen der Untersuchungsgegenstand untersucht werden soll, festgelegt. Diese Art der heuristischen Evaluation stellt den Einstieg für die weiteren Evaluationsschritte dar (vgl. Zühlke 2012).

Bei der Focus-Group-Research Methode handelt es sich um moderierte Diskussionsrunden, die zu einem entsprechenden Thema geführt werden. Die Teilnehmer setzen sich mit den verschiedenen Sichtweisen, Ideen und Wahrnehmungen der anderen Teilnehmer auseinander und tauschen sich dabei untereinander aus. Ziel dieser Methode ist es, erste Konzepte durch repräsentative Nutzer zu evaluieren. Dabei dienen die Fokusgruppen der qualitativen Datenerhebung bezüglich wichtigen Ansprüchen und Nutzererwartungen (vgl. Zühlke 2012).

Empirische Evaluation

Als empirische Evaluation werden die Tests von Endnutzern an einem Prototyp bezeichnet. Diese Evaluationsmethode unterscheidet sich in den folgenden aufgeführten Methoden (vgl. Zühlke 2012, S. 122).

Thinking-Aloud-Methode

Hier arbeitet die Testperson in Gegenwart eines Prüfers am System und wird dazu animiert auszusprechen, was ihr bei der Arbeit mit dem System in den Sinn kommt. Dabei sollen die Vorstellungen, Gedanken und unbewusste Haltungen der Nutzer zum Vorschein gebracht werden und Hinweise über die auftretenden Probleme bei der Mensch-Maschine-Interaktion aufgezeigt werden (vgl. Thimbleby 1990).

Concurrent-Thinking-Aloud-Methode

Bei der Concurrent-Thinking-Aloud-Methode werden die Nutzer gebeten ihre Gedanken zu verbalisieren (vgl. Nielsen 1993).

Retrospective-Thinking-Aloud-Methode

Bei der Retrospective-Thinking-Aloud-Methode dürfen die Nutzer erst nach Beantwortung einer Frage ihre Äußerungen tätigen. Der Test wird mitgeschnitten und wird durch den Prüfer mit dem Nutzer besprochen, um die auftretenden Probleme zu klären. Der große Vorteil ist dabei, dass der Tester die Aufnahme jederzeit anhalten und gezielt Fragen stellen kann. Der große Nachteil dieser Methode ist die doppelt benötigte Zeit im Gegensatz zur Concurrent-Thinking-Aloud-Methode (vgl. Zühlke 2012).

Inteviews

Die Durchführung von Interviews stellt eine weitere Art der empirischen Evaluation dar. Hier werden den Nutzern, die mit dem System vertraut sind oder Testaufgaben erledigt haben, Fragen zum System gestellt und die Erfahrungen der Befragten werden dokumentiert (vgl. Zühlke 2012).

Fragebögen

Die Fragebögen stellen eine Alternative zu den Interviews dar und haben den großen Vorteil, dass diese eine schnellere Auswertung ermöglichen, welche oft automatisch durchgeführt werden kann. Ein weiterer Vorteil ist, dass die Hemmschwelle der Testpersonen geringer ausfällt. Bei den Fragebögen werden ein Background-Fragebogen und ein Post-Test-Fragebogen angefertigt. Der Background-Fragebogen liefert Hintergrundinformationen über die Testpersonen um auftretende Probleme und Verhaltensweisen der Nutzer besser zu verstehen. Dieser Fragebogen wird von den Nutzern vor dem eigentlichen Test ausgefüllt. Ziel des Post-Test-Fragebogens ist es die Stärken und Schwächen des Systems zu liefern. Dabei setzt sich der Fragebogen aus zwei Teilen zusammen. Der erste Teil beinhaltet die Aspekte des „Joy-of-use“ und der zweite Teil die Aspekte des „Ease-of-use.“ Im Teil „Joy-of-use“ handelt es sich um Skalen, mit welchen die Testpersonen ihre Bewertung abgeben. Die Skala besteht dabei aus fünf Feldern, die sich durch die Assoziation mit Schulnoten als einfach benutzbar erweist. Somit ist eine Erfassung von Tendenzen problemlos möglich. Der Ease-of-use Teil beinhaltet verschiedene Aussagen mit denen die Testpersonen konfrontiert werden. Dabei geben die Nutzer den Grad ihrer Zustimmung oder Ablehnung zu jeder Aussage in einer Likert-Skala an (vgl. Rubin 1994).

Ein sehr bekannter Fragebogen, der speziell für Sprachdialogsysteme entwickelt wurde, ist der „Subjective Assessment of Speech System Interfaces (SASSI)“-Fragebogen (vgl. Hone und Graham 2000). Dieser enthält 34 Elemente, die auf einer fünfstufigen Likert-Skala bewertet werden. Dabei werden die einzelnen Elemente den folgenden sechs Faktoren zugeordnet:

- Genauigkeit der Systemantwort („system response accuracy“) beschreibt, wie genau die Nutzereingaben vom System richtig verstanden wurden und ob die Intention und die Erwartungen des Nutzers erfüllt wurden.
- Der Faktor Beliebtheit („likeabililty“) enthält die Meinung und das Empfinden des Nutzers über das System.

- Die kognitive Beanspruchung („cognitive demand“) beschreibt die vom Nutzer empfundene kognitive Anstrengung während der Interaktion.
- Ärgernis („annoyance“) erfasst die negativen Gefühle des Nutzers während der Interaktion mit dem System.
- Die Bewohnbarkeit („habitability“) beschreibt die Fähigkeit des Nutzers den Umgang mit dem System schnell und problemlos zu erlernen.
- Schnelligkeit („speed“) erfasst die vom Nutzer wahrgenommene Geschwindigkeit der Interaktion und des Systems.

Task-Szenarien

Nach Preece (1994) beschreiben die Task-Szenarien eine Folge von Aktionen, die bei der Arbeit mit dem Produkt auftreten können. Dabei handelt es sich nicht nur um eine reine Aneinanderreihung von Aufgaben. Sie beinhalten auch ein Endresultat, welche die Nutzer erreichen sollen. Darüber hinaus werden auch die Motive beschrieben, die hinter den Aufgaben stehen (vgl. Rubin 1994).

4.2 Technische Kriterien

Bei sprachverarbeitenden Systemen handelt es sich um Softwaresysteme. Somit bilden die Standards für die Evaluation von Softwaresystemen die Basis für die Evaluation von sprachverarbeitenden Systemen. Bei der Evaluation von Softwaresystemen liegt der Fokus auf die Qualität. Nach ISO/IEC-Norm 9126 ISO/IEC (2001, S. 20) wird die Qualität beschrieben als „*Die Gesamtheit der Eigenschaften eines Software-Produkts, die sich auf dessen Eignung beziehen, festgelegte oder vorausgesetzte Erfordernisse zu erfüllen.“* Diese Norm wurde im Jahr 2011 durch die ISO/IEC 25010 ersetzt. Dabei wird die Bedienbarkeit als eine Komponente der Softwarequalität angesehen.

4.2.1 Wortfehlerrate

Ausschlaggebend für die Evaluation von Spracherkennern ist die Wortfehlerrate (WER) oder die Wortakkuratheit (WA). Dabei gilt die Beziehung WER = 100% - WA. Bei der Evaluation wird ein Testset verwendet, indem für jede Äußerung eine Referenztranskription vorliegt, mit der die entsprechend erkannte Wortfolge verglichen wird. Diese Art der Evaluation ist unter dem Begriff referenz-basierte Evaluation bekannt, wobei die erkannte Wortfolge auf die Referenzwortfolge abgebildet wird. Dabei ist es entscheidend, dass die Summe der notwendigen Korrekturen minimal ist. Jede Korrektur wird sodann als Fehler vermerkt. Bei den Fehlern unterscheidet man zwischen Ersetzungen (substitutions), Auslassungen (deletions) und Einfügungen (insertions). Die Wortfehlerrate errechnet sich somit aus der Anzahl der Ersetzungen Sn, Auslassungen Dn und Einfügungen In. Dadurch ergibt sich die Formel: $100\% \cdot \frac{Sn+In+Dn}{N}$, wobei N die Anzahl der Wörter im Referenzsatz darstellt (vgl. Carstensen et al. 2010).

4.2.2 Konzeptfehlerrate

Die Konzeptfehlerrate (CER) wird von der Wortfehlerrate (WER) abgeleitet und betrachtet nicht die einzelnen Wörter als grundlegendes Maß, sondern semantische Einheiten. Die CER misst somit den Prozentsatz korrekt verstandener semantischer Einheiten pro Dialog und wird genau wie die WER berechnet (vgl. Carstensen et al. 2010).

4.2.3 Out-of-Grammar Rate

Ein weiteres wichtiges Kriterium, besonders bei grammatikbasierter Spracherkennung, ist die Rate der Äußerungen, die nicht von der verwendeten Grammatik abgedeckt werden. Im Englischen wird dies auch als out-of-grammar rate (OOG) bezeichnet. Für die Berechnung gilt: $OOG = \frac{N_{OOG}}{N_{total}}$. Wobei N die Anzahl aller Äußerungen und N_{OOG} die Anzahl der Äußerungen, die nicht von der verwendeten Grammatik abgebildet werden können, darstellen (vgl. Carstensen et al. 2010).

5 Auswahl eines Sprachdialogsystems

Der Markt bietet verschiedene Alternativen von Spracherkennungssystemen, die jeweils eine führende Position im jeweiligen Marktsegment aufweisen. Im Folgenden wird eine Übersicht über die verfügbaren Spracherkennungssysteme gegeben und anschließend werden Anforderungen, die für die Entscheidungsfindung essentiell sind, definiert. Im Anschluss werden mit Hilfe der vergleichenden Evaluationsmethode alle genannten Spracherkennungssysteme aufgelistet und den Kriterien zugeordnet. Anhand dieser Evaluation wird ein Spracherkennungssystem für die Evaluation selektiert.

5.1 Marktübersicht

Es existieren marktabhängig verschiedene Spracherkennungssysteme, die sich auf den ersten Blick zur Anbindung eignen. Im Folgenden werden aktuelle Spracherkennungssysteme erläutert, welche von verschiedenen Herstellern angeboten werden.

Im Bereich Gesundheitswesen ist die Spracherkennungssoftware SpeechMagic der Firma Nuance Communications Inc., ansässig in 1 Wayside Road, Burlington, MA 01803, United States, weit vertreten.

Die Firma topsystem Systemhaus GmbH, ansässig in der Monnetstraße 24, in 52146 Würselen, vertreibt die Produktlinie Lydia® Voice Solutions für die Logistik- und die Luftfahrtindustrie. Die Produkte umfassen neben Hardware, wie Lydia Headsets, auch Software für die Spracherkennung und –wiedergabe. Nach Rücksprache mit dem Unternehmen topsystem Systemhaus GmbH fiel die Wahl auf ein Blutooth-Headset mit externem Spracherkennungsmodul und dem Lydia® Telegram Client als Core (vgl. topsystem 2017).

Das dritte System ist von der Firma Google Inc., ansässig in 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA. Google vertreibt den Dienst Google Cloud Speech API, welches die Spracherkennung ermöglicht und das Gesprochene in digitalen Text

umwandelt. Dabei werden mittels einer benutzerfreundlichen API leistungsstarke neuronale Netzmodelle angewendet. Die API unterstützt über 80 Sprachen und Varianten. Mithilfe des Dienstes ist es möglich den Text von Nutzern, die in das Mikrofon einer Anwendung sprechen, zu transkribieren, Sprachsteuerung zu aktivieren oder Audiodateien zu transkribieren. Die Textergebnisse können auch direkt gestreamt werden, sodass der Text unmittelbar während des Sprechens erscheint. Weiterhin wird auch die kontextbasierte Spracherkennung unterstützt, sodass die Spracherkennung auf den Kontext zugeschnitten werden kann, indem mit jedem API-Aufruf eine separate Liste mit Worthinweisen bereitgestellt wird. Dies ist insbesondere bei der Steuerung von Geräten oder Apps wichtig. Allerdings ist der Dienst cloudbasiert, d.h. jedes gesprochene Wort wird an den Google-Server gesendet und dort anhand von Algorithmen und neuralen Netzwerken in digitalen Text konvertiert (vgl. Google 2017).

Der Dienst Speech-to-Text und Text-to-Speech von der Firma IBM Deutschland GmbH, ansässig in IBM-Allee 1, 71139 Ehningen, wird als Cloud-Dienst IBM Watson angeboten. Dabei unterstützt der Dienst das automatische Übersetzen von gesprochener Sprache aus sieben Sprachen in Echtzeit. Ein weiteres Merkmal des Dienstes ist das schnelle Identifizieren und Transkribieren vom Gesprochenen und es werden über eine Vielzahl von Audio-Formaten und Programmierschnittstellen unterstützt. Anhand von Schlüsselwörtern ist es möglich die Genauigkeit für die Spracherkennung zu verbessern und somit eine bessere Erkennung in Echtzeit zu erzielen (vgl. IBM 2017).

5.2 Anforderungen

Die Anforderungsdefinition aus der Systemanalyse, welche den Kern der Systemanalyse darstellt (vgl. Rupp & die SOPHISTen 2013) wird herangezogen, um die Kriterien zur Auswahl der Sprachsoftware zu eruieren.

Nach Rupp & die SOPHISTen (2013, S. 13) ist eine Anforderung definiert als:

1. *„Eine Bedingung oder Fähigkeit, die von einem Benutzer (Person oder System) zur Lösung eines Problems oder zur Erreichung eines Ziels benötigt wird.“*

2. *„Eine Bedingung oder Fähigkeit, die ein System oder Teilsystem erfüllen oder besitzen muss, um einen Vertrag, eine Norm, eine Spezifikation oder andere, formell vorgegebene Dokumente zu erfüllen.“*
3. *„Eine dokumentierte Repräsentation einer Bedingung oder Eigenschaft gemäß (1) oder (2).“*

Dabei werden zwei Arten von Anforderungen unterschieden, die funktionalen und die nicht-funktionalen Anforderungen. Die funktionalen Anforderungen beschreiben die Funktionalität des zu erstellenden Systems. Die funktionalen Anforderungen liefern ausschließlich Antworten auf die Frage „Was soll das System machen?“ Zu den nicht-funktionalen Anforderungen zählen alle anderen Anforderungen, die nicht den funktionalen Anforderungen zugeordnet werden können. Weiterhin lassen sich die nicht-funktionalen Anforderungen in zwei Gruppen einteilen, in Qualitätsanforderungen und Randbedingungen. Die Qualitätsanforderungen beziehen sich in der Regel auf andere funktionale Anforderungen und beschreiben, in welcher Qualität das System seine Aufgaben erfüllen soll. Durch die Randbedingungen wird der Handlungsspielraum bei der Systementwicklung zusätzlich eingeschränkt. Diese zwei Gruppen lassen sich noch weiter wie folgt untergliedern:

Qualitätsanforderungen (vgl. Rupp & die SOPHISTen 2013)

- Anforderungen an die Funktionalität (z.B. Genauigkeit einer Berechnung)
- Anforderungen an die Effizienz (z.B. Antwortzeitverhalten)
- Anforderungen an die Zuverlässigkeit (z.B. Fehlertoleranz)
- Anforderungen an die Benutzbarkeit (z.B. Bedienbarkeit, Erlernbarkeit)
- Anforderungen an die Änderbarkeit (z.B. Erweiterbarkeit des Systems)
- Anforderungen an die Übertragbarkeit (u.a. auch Konformität zu Standards)

Randbedingungen (vgl. Rupp & die SOPHISTen 2013)

- Technische Anforderungen (z.B. Vorgaben an die Hardware, die Schnittstellen oder die Systemarchitektur)

- Anforderungen an die Benutzerschnittstelle (z.B. Vorgaben an die Benutzerschnittstelle, Ergonomie)
- Anforderungen an sonstige Lieferbestandteile (z.B. an ein Betriebshandbuch)
- Anforderungen an die Durchführung der Entwicklung (z.B. Vorgehensmodell)
- Rechtlich-vertragliche Anforderungen (z.B. Zahlungsmeilensteine)

Die Abbildung 6 fasst die verschiedenen Arten der Anforderungen zusammen und ermöglicht einen Überblick.

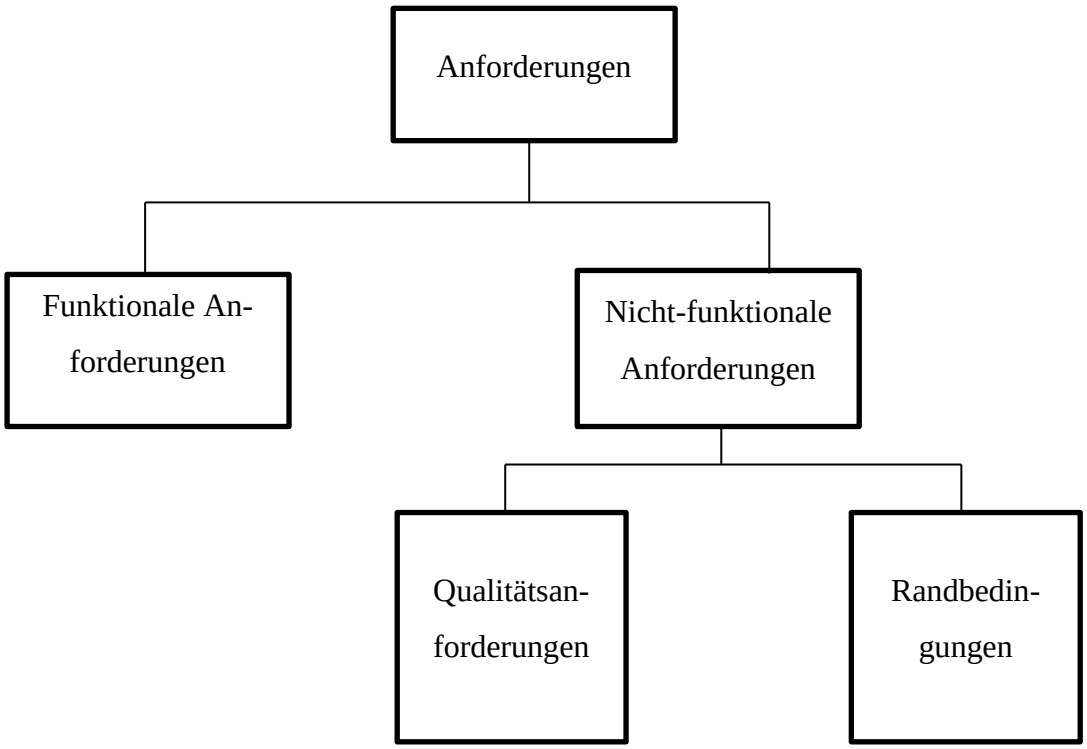

Abbildung 6: Arten von Anforderungen (Quelle: Rupp & die SOPHISTen 2013, S. 22)

5.3 Vergleichende Evaluation

Für die Auswahl des Sprachsystems sind vorwiegend nicht-funktionale Anforderungen von Bedeutung. Dabei spielen die technischen Anforderungen als Randbedingung eine große Rolle. Anhand von Recherche und Nutzerbefragung werden folgende technische Anforderungen definiert:

- Datenschutz nach dem Bundesdatenschutzgesetz: Das Sprachsystem darf die Benutzereingaben nur intern verarbeiten, sodass Spracheingaben nicht an externe Server gesendet werden müssen.
- Das Sprachsystem soll eine API bieten, wodurch es an andere Systeme angebunden werden kann.
- Das Sprachsystem soll in der Lage sein, Umgebungslärm - wie Maschinenlärm - zu filtern, um die wesentlichen Spracheingaben der Benutzer zu erkennen.
- Das Sprachsystem soll Mehrsprachlichkeit unterstützen, sodass Spracheingaben in mehreren Sprachen möglich sind. Dabei sollen mindestens die Sprachen Deutsch und Englisch unterstützt werden.
- Das Sprachsystem soll die Möglichkeit bieten, dass Benutzersprache antrainiert werden kann, um eine höchstmögliche Erkennung von Sprachbefehlen zu gewährleisten.

Die Bewertung der einzelnen Kriterien erfolgt anhand eines Punktesystems. Für jedes erfüllte Kriterium gibt es einen Punkt, sodass insgesamt fünf Punkte erreicht werden können. Die Kriterien sind dabei nach ihrer Priorität aufgelistet, wobei das erste Kriterium die höchste Priorität besitzt. Somit stellt das Kriterium des Datenschutzes ein K.o.-Kriterium dar. Sprachsysteme, die dieses Kriterium nicht erfüllen, fallen somit aus der Entscheidungsfindung heraus.

Anforderungen \ Sprachsysteme		Lydia	Nuance	Google Cloud Speech API	IBM Watson
Qualitätsanforderungen (Randbedingungen)	Datenschutz (Daten dürfen das Unternehmen nicht verlassen)	1	0	0	0
	API Anbindung	1	1	1	1
	Umgebungslärm filtern	1	1	1	0
	Mehrsprachlichkeit (Deutsch & Englisch)	1	1	1	1
	Benutzersprache antrainieren	1	1	0	0
Gesamtergebnis (maximal 5 Punkte möglich)**:**		**5**	**4**	**3**	**2**

Tabelle 1: Vergleichende Evaluation der Systeme (eigene Darstellung)

Das erste und damit wichtigste Kriterium ist – wie bereits zuvor erwähnt - der Datenschutz. Die digital gesprochenen Daten dürfen das Unternehmensnetzwerk nicht verlassen. Im spezifischen Fall bedeutet dies, dass die Konvertierung der gesprochenen Daten im Unternehmensnetzwerk stattfinden muss. Es darf keine Cloud-Lösung zum Einsatz kommen, welche darauf ausgerichtet ist, dass die Daten zur Konvertierung an einen Server außerhalb des Unternehmensnetzwerks gesendet werden und die konvertierten Daten zurück vom Server an das Unternehmensnetzwerk gesendet werden. Die Konvertierung der gesprochenen Daten muss entweder im Unternehmensnetzwerk geschehen oder falls möglich, direkt auf der Hardware mit Hilfe von einem Embedded System durchgeführt werden.

Das zweite Kriterium ist die Verfügbarkeit einer API (application programming interface). Die API wird dazu verwendet, dass andere Systeme mit dem Sprachsystem kommunizieren und Daten austauschen können. Dadurch wird gewährleistet, dass die konvertieren Daten auch durch andere Systeme für ihre Zwecke genutzt werden können, um gewisse Aktionen auszuführen.

In der Produktion herrscht normalerweise ein gewisser Lärmpegel, welcher zum Beispiel durch Maschinen verursacht wird. Daher ist ein weiteres Kriterium die Möglichkeit diesen Umgebungslärm zu filtern, damit die auditiven Benutzereingaben auch fehlerfrei erkannt werden.

Das System soll die Möglichkeit zur Mehrsprachlichkeit bieten, sodass die auditiven Benutzereingaben in zumindest zwei Sprachen erfolgen können, Englisch und Deutsch. Somit ist die vierte Anforderung an das System die Mehrsprachlichkeit.

Jeder Nutzer hat seine eigene Aussprache und bei internationalen Nutzern spielt oft auch der Akzent eine große Rolle. Dadurch ist das letzte und damit optionales Kriterium das Antrainieren der Benutzersprache, damit das System die Aussprache der einzelnen Benutzer lernen kann, und dadurch die Qualität der Spracherkennung erhöht wird.

5.4 Auswertung der vergleichenden Evaluation

Das Sprachsystem Lydia® bietet als einziges System die Möglichkeit, dass die auditiven Benutzereingaben zur Erkennung und Konvertierung nicht an einen Server oder eine Cloud-Lösung gesendet werden müssen. Somit ist der Datenschutz dahingehend garantiert, dass die Daten das Unternehmensnetzwerk nicht verlassen. Die Erkennung und Konvertierung kann direkt über eine Serverinstallation im eigenen Netzwerk erfolgen. Das führt auch zu schnelleren Ladezeiten, sodass eine Live-Konvertierung der auditiven Daten möglich ist. Weiterhin ist das System Multiuser fähig, sodass mehrere Nutzer gleichzeitig am System angemeldet werden können.
Die Firma topsystem Systemhaus GmbH bietet zum Spracherkennungssystem auch die passende Hardware, wie Wireless-Headsets mit angebundenem, direktem Spracherkennungsmodul. Durch die Kompatibilität von Hardware und Software wird erreicht, dass die Hardware fehlerfrei mit der Spracherkennungssoftware kommunizieren kann.

Bei den Sprachsystemen Nuance, Google Cloud Speech und IBM Watson müssen die auditiven Benutzereingaben, für die Erkennung und Konvertierung, an einen externen Server gesendet werden. Sobald diese erfolgreich erkannt und entsprechend konvertiert wurden, werden diese Daten zurück an das jeweilige Sprachsystem gesendet. Damit wird das Kriterium des Datenschutzes nicht erfüllt und diese Systeme scheiden explizit aus. Aufgrund dessen wird für die Implementierung und die Anbindung an die iTAC.MES.Suite das Sprachsystem Lydia® von der Firma topsystem Systemhaus GmbH ausgewählt.

6 Evaluationsdesign und Durchführung

Die Evaluation des Prototyps wird anhand von drei Evaluationstechniken durchgeführt. Zuerst werden empirische Evaluationsmethoden angewandt. Dabei werden Task-Szenarien anhand von Benutzertests durchgeführt und danach wird ein Fragebogen mit spezifischen Fragen zur Bedienbarkeit des Systems von den Probanden ausgefüllt. Während der Erledigung der Tasks werden auch die technischen Kriterien anhand der referenzbasierten Evaluation dokumentiert. Dazu zählt die Wortfehlerrate, welche mit einer Referenzwortfolge (oder vorgegebenen Grammatik – Anhang 5) verglichen wird. Dabei werden Einfügungen, Ersetzungen und Auslassungen von Wörtern dokumentiert. Weiterhin wird ein Beobachtungsprotokoll erstellt, um die Reaktionen der Probanden während des Versuchs zu erfassen.

Ziel dieser Evaluation ist es, das Testsystem anhand von Benutzertests auf Usability zu testen, um mögliches Fehlverhalten oder Fehler des Systems aufzudecken.

6.1 Systemaufbau des mobilen Testsystems

Das Testsystem umfasst mehrere Hardwaregeräte, eine Instanz der iTAC.MES.Suite, die Lydia@ Voice Control Software mit entsprechender API, welche an die iTAC.MES.Suite angebunden wird und weitere Komponenten, die für die Durchführung der Tasks benötigt werden. Zur Hardware gehören ein PC mit dem Betriebssystem Windows 8.1, ein Lydia® WLAN-Headset, ein Lydia@ Infrarot-Handscanner und ein Router von der Firma Linksys. Zu den weiteren Komponenten zählen ein Dummy-Mainboard mit einer Seriennummer und drei Behälter mit Seriennummern mit jeweils einer Gruppe von Montagekomponenten für das Mainboard. Der Router wird genutzt, um ein eigenes, geschlossenes Netzwerk zu schaffen und somit die fehlerfreie Konnektivität des Lydia® WLAN-Headsets zu gewährleisten.
Weitere technische, sowie Implementationsdetails sind für die Evaluation nicht relevant und werden nicht weiter beschrieben.

Abbildung 7 gibt einen Überblick über die einzelnen Komponenten des Testsystems und Abbildung 8 verdeutlich den Versuchsaufbau des mobilen Testsystems.

Linksys WLAN Router

Lydia Bluetooth-Headset

Dummy Mainboard

Lydia Handscanner

Monitor

Rechner

Abbildung 7: Übersicht der Testgeräte (eigene Darstellung)

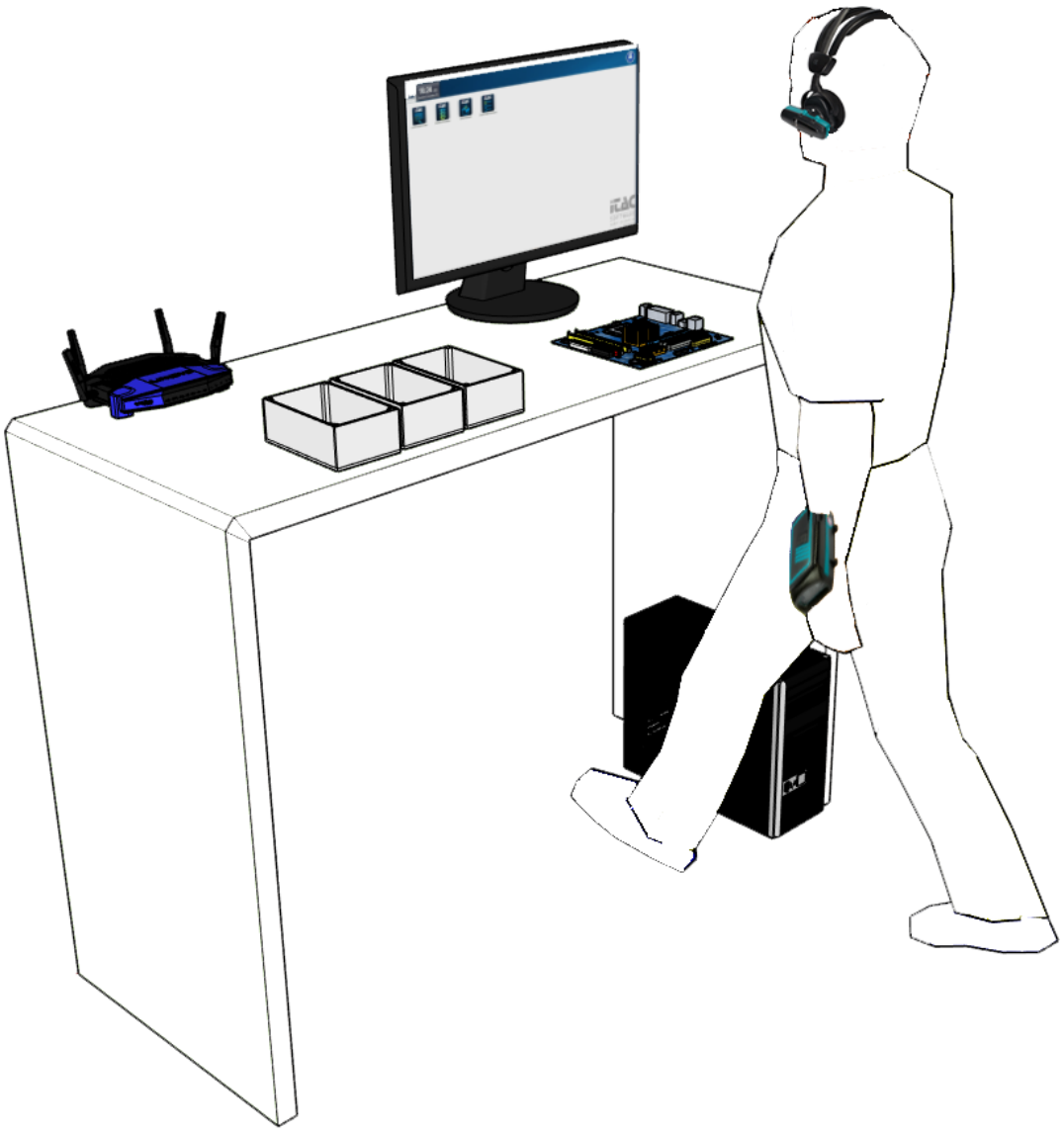

Abbildung 8: skizzierter Versuchsaufbau (eigene Darstellung)

6.2 Versuchsmaterial

Dieser Abschnitt umfasst die Versuchsmaterialien, die während des Experiments herangezogen werden. Es folgt eine Auflistung der Materialien mit entsprechender Erläuterung.

6.2.1 Schriftliche Anweisungen (Anhang 3)

Vor Beginn des eigentlichen Versuchs wird den Versuchsteilnehmern ein Blatt mit Informationen zum Versuchsablauf ausgehändigt. Darin wird auch kurz das zu testende System und die durchzuführenden Aufgaben erläutert. Die Instruktionen sind im Anhang 3 zu finden.

6.2.2 Personenbezogener Fragebogen (Anhang 1)

Der personenbezogene Fragebogen dient zur Erfassung demographischer Daten wie Alter und Geschlecht der Probanden. Weiterhin wird abgefragt ob die Probanden bereits über Erfahrungen mit Sprachdialogsystemen verfügen und in wie weit sie ihre Englischkenntnisse auf einer 4-stufigen Skala, von „keine“ bis „verhandlungssicher“ einschätzen, denn das Sprachdialogsystem wird englischsprachig bedient.
Dieser Fragebogen wird vor dem Versuch an die Teilnehmer verteilt und soll entsprechend vor dem durchzuführenden Versuch von den Probanden ausgefüllt werden.

6.2.3 SASSI Fragebogen (Anhang 2)

Der modifizierte SASSI-Fragebogen dient zur Erfassung der Gebrauchstauglichkeit des Systems durch 15 Items, wobei die ersten 12 Items anhand einer 5-stufigen Likert-Skala bewertet werden und die letzten 3 Items offene Fragen darstellten.
Für den Versuchsablauf sind einige SASSI-Items nicht von Relevanz, daher wird der Fragebogen gekürzt und die irrelevanten Items entsprechend gestrichen.
Die Beurteilung der Gebrauchstauglichkeit erfolgt über eine modifizierte Version des

SASSI-Fragebogens nach Hone und Graham (2000) mit deutscher Übersetzung nach Strauss (2010) (siehe Anhang A.3.2). Der Fragebogen basiert auf einer 5-stufigen Likert-Skala. Die Bewertung erfolgt von „lehne stark ab“ (-2) bis „stimme stark zu“ (+2) bzw. von „sehr schlecht“ bis „sehr gut.“ Die Aufgabe ist erfolgreich absolviert bei richtiger sprachlicher Selektion des gesuchten Elements.

Nutzer bewerten die Varianten somit hinsichtlich der SASSI-Dimensionen Genauigkeit der Systemantwort, Beliebtheit und Ärgernis. Zusätzlich geben drei Fragen des ITU-T Rec. P.851 (International Telecommunication Union (ITU), 2003) Aufschluss über nicht ausreichende Hilfestellung (7.3 Q4), notwendige Konzentration (7.2 Q6) und Gesamteindruck. Zur Bestimmung des Aufgabenerfolgs erfolgt eine manuelle Annotation, ob der Nutzer die Aufgabe korrekt abschließen konnte oder nicht.

6.2.4 Beobachtungsprotokoll und technisches Protokoll (Anhang 6)

Während des Experiments werden die Probanden bei der Erledigung der Tasks beobachtet und ihre Reaktionen und Emotionen protokolliert. Wichtig dabei ist es zu erkennen, ob die Probanden bei der Interaktion mit dem System Frustration empfinden oder Freude. Weiterhin werden auch die eigenen Gedanken der Versuchsteilnehmer, die während des Versuchs laut geäußert werden, protokolliert.
Die technischen Kriterien, dazu zählt die Wortfehlerrate, die Konzeptfehlerrate und die Out-of-grammar Rate, werden während des Versuchs vom Versuchsleiter vermerkt.

6.3 Benutzertests

Die Benutzertests finden bei der Firma Limtronik GmbH in Limburg statt. Das Unternehmen agiert als Dienstleister für Electronic Manufactoring Services[4] (vgl. Zhai, Shi, Gregory 2007) und ist spezialisiert auf die Fertigung von elektronischen Baugruppen, sowie maßgeschneiderten Systemen. Weiterhin betreibt das Unternehmen am Standort Limburg

[4] EMS - Unternehmen, die fertigungsbezogene Dienstleistungen für Hersteller anbieten.

an der Lahn eine der bundesweit modernsten Fertigungsstätten im Sinne der Industrie 4.0 (vgl. Limtronik 2018).
Im Folgenden werden allgemeine Informationen zu den Probanden, den gestellten Aufgaben, sowie dem Versuchsablauf gegeben.

6.3.1 Probanden

Die Anzahl der Teilnehmer wird aufgrund des im Kapitel 6.1.2 (heuristische Evaluation) erläuterten Zusammenhangs auf sechs Teilnehmer festgelegt.
Insgesamt nehmen drei weibliche und drei männliche Teilnehmer an dem Experiment teil. Bei den Teilnehmern handelt es sich um Auszubildende aus dem Bereich Mechatronik und Elektrotechnik im Alter von 16 bis 21 Jahren. Alle Teilnehmer sind als unerfahrene Probanden anzusehen, da sie das Testsystem oder ein ähnliches System vor dem Versuch nicht bedient haben.

6.3.2 Aufgaben (Anhang 4)

Den Teilnehmern werden zwei verschiedene Aufgaben gestellt, die mittels des Testsystems durchgeführt werden sollen.
Dabei handelt es sich bei der ersten Aufgabe um eine Wartungsarbeit aus der Produktion. Eine Maschine soll auf einen gewissen Zustand gesetzt werden. Sobald der Zustand erfolgreich gesetzt ist, meldet das System automatisch einen neuen Task. Dieser Task wird einer Aufgabenliste (Tasklist) zugeordnet und soll vom Benutzer per Sprachbefehl (confirm oder okay) angenommen werden. Sobald die Aufgabe angenommen wurde, soll der Benutzer die Priorität dieser Aufgabe auf hoch setzen. Danach soll der Benutzer die Aufgabe von der Aufgabenliste löschen, was die erfolgreiche Erledigung der entsprechenden Aufgabe simuliert.
Bei der zweiten Aufgabe handelt es sich um eine Montagearbeit aus der Produktion. Dabei soll der Benutzer an einer Arbeitsstation sitzen und eine Platine mit drei verschiedenen Komponenten bestücken. Die Aufgabe beginnt mit dem Einscannen der Seriennummer der Platine. Wird die Seriennummer erfolgreich vom System erkannt, so zeigt das

System die Gesamtanzahl der zu verbauenden Komponenten mit den genauen Bezeichnungen auf dem Bildschirm an. Parallel dazu folgt die Aufforderung zum Einscannen der Seriennummer der ersten zu verbauenden Komponente. Nach erfolgreicher Erkennung der Komponente, soll diese vom Benutzer an der entsprechenden Stelle, welche auf dem Monitor genau angezeigt wird, verbaut werden. Sobald der Benutzer die Komponente an der richtigen Stelle angebracht hat, soll der Proband es mit einem Befehl (confirm oder installed) bestätigen. Danach folgt die Aufforderung zum Einscannen der zweiten Komponente. Wird diese erfolgreich erkannt und entsprechend vom Benutzer verbaut und bestätigt, folgt die nächste Aufforderung zum Einscannen der letzten Komponente. Sobald alle Komponenten erfolgreich verbaut sind, fordert das System den Benutzer auf, die Erledigung der Aufgabe für die eingescannte Platine per Befehl (confirm) zu bestätigen. Sobald der Benutzer dies bestätigt hat, kann die Seriennummer der nächsten Platine eingescannt werden. Für das Experiment wird nur eine Platine mit insgesamt drei zu verbauenden Komponenten verwendet.
Im Abschnitt 7.4 werden die Task-Szenarien Schritt für Schritt mit den entsprechend zu verwendenden Befehlen beschrieben.

6.3.3 Versuchsablauf

Vor Beginn des Versuchs erhalten die Teilnehmer genaue, schriftliche Anweisungen zum Versuch und füllen einen kurzen Fragebogen mit persönlichen Daten aus (siehe 6.2.2). Weiterhin wird den Probanden das Testsystem anhand einer kurzen Demonstration der zu erledigenden Aufgaben erklärt. Die Teilnehmer werden dazu angehalten, während der Durchführung der Task-Szenarien, bei auftretenden Störungen der Interaktion, ihre Gedanken dem Beobachter laut mitzuteilen. Nach erfolgreicher Erledigung der zwei Aufgaben, füllen die Probanden den SASSI-Fragebogen zum Testsystem aus und damit wird der Versuch abgeschlossen.
Fehler! Verweisquelle konnte nicht gefunden werden. verdeutlich den Versuchsablauf n sequenzieller Reihenfolge.

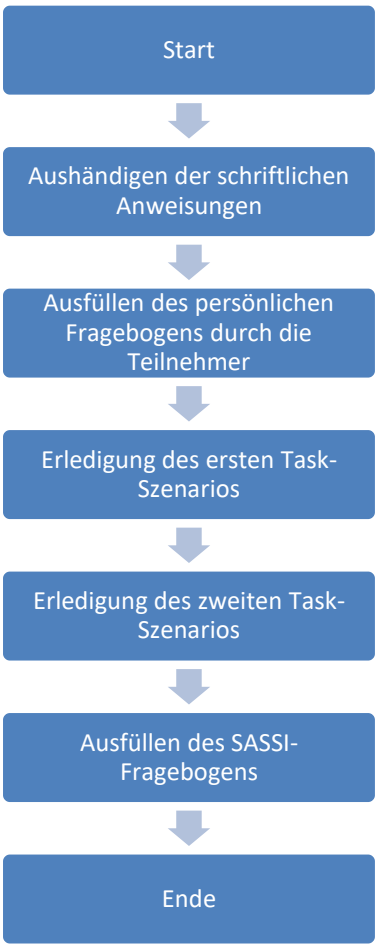

Abbildung 9: Versuchsablauf (eigene Darstellung)

6.4 Task-Szenarien

Aufgrund der Tatsache, dass die Sprachsoftware an die ITAC.MES.Suite angebunden ist, werden für die Evaluation verschiedene Szenarien aus der praktischen Produktion ermittelt. Bei der Erstellung des ersten Szenarios steht der Gedanke im Vordergrund, dass die Produktionslinien in einer Produktion auch gewartet werden müssen. Speziell bedeutet es, dass während einer Produktion die Maschinen vom Menschen gewartet und überwacht werden. Beispielsweise müssen Bestückungsrollen mit Widerständen in den Maschinen eingelegt oder Flüssigkeitsbehälter ausgetauscht werden. Das Zielsystem, als Pilotsystem, ist ein Dialogsystem, mit dessen Hilfe ein Mitarbeiter in der Produktion die Wartungsarbeiten erledigen kann. Um den Dialog mit dem System zu vereinfachen wird festgelegt, dass gewisse Schlüsselwörter bzw. Befehle genannt werden, die eine entsprechende Aktion des Systems auslösen. Das erste Szenario umfasst dabei folgende Tasks mit den dazugehörigen Befehlssätzen.

6.4.1 Szenario 1 (maintenance):

1. Aufgabenliste abrufen und bearbeiten
    a. Es existiert eine Aufgabenliste. Der Benutzer kann diese Aufgabenliste mit dem Befehl „task list“ aufrufen.
    b. Mit dem Befehl „next“ wird die nächste Aufgabe in der Aufgabenliste abgerufen.
    c. Die Aufgaben können drei verschiedene Prioritäten annehmen, „high“, „medium“ und „low.“
    Die Priorität einer Aufgabe wird mit den Befehlen „high priority“, „low priority“ und „medium priority“ vergeben.
    d. Wenn eine Aufgabe erledigt ist, kann diese mit dem Befehl „remove“ von der Aufgabenliste entfernt werden.

   e. Um die Aufgabenliste zu schließen, soll der Befehl „finish“ benutzt werden. Ist die Aufgabenliste leer so wird dieser Befehl vom System ignoriert.

2. Maschinenzustand definieren
   a. Um den Maschinenzustand zu definieren wird der Befehl „set condition“ benutzt.
   b. Mit dem Befehl „cancel“ wird das Festlegen des Maschinenzustands storniert
   c. Mit Hilfe des Befehls „select“ wird ein definierter Maschinenzustand selektiert.
   d. Zum Setzen des Maschinenzustands wird der Befehl „confirm“ verwendet.
3. Allgemein
   a. Um das Mikrofon stummzuschalten wird der Befehl „close microphone“ definiert.
   b. Um das Mikrofon wieder aktiv zu schalten wird der Befehl „open microphone“ definiert.
   c. Um die letzte Aussage wiederholen zu lassen wird der Befehl „repeat“ definiert.

Beim zweiten Szenario steht eine Bestückungsstation im Fokus. Eine Platine soll durch einen Mitarbeiter bestückt werden. Dabei soll der Mitarbeiter genaue Anweisungen erhalten, welche Komponente an welchem Platz auf der Platine angebracht werden soll. Zuerst wird die Platine anhand ihrer Seriennummer eingescannt, dies geschieht über einen Handscanner. Danach wird dem Mitarbeiter durch das System gemeldet, welche Komponente als erstes aus welchem Behälter entnommen werden soll. Dazu wird der Barcode des Behälters über den Handscanner eingescannt, wodurch die entsprechende Komponente vom System erkannt wird. Danach wird dem Mitarbeiter visuell dargestellt, wo die Komponente auf der Platine angebracht werden soll. Dies geschieht über einen Windows Client mit entsprechender grafischer Oberfläche. Sobald der Mitarbeiter die Komponente

an dem richtigen Platz angebracht hat, wird der Arbeitsschritt durch den Mitarbeiter auditiv an das System bestätigt. Danach wird die nächste Komponente eingescannt und die vorherigen Schritte werden wiederholt, solange nicht alle Komponenten korrekt auf der Platine angebracht wurden. Somit umfasst das zweite Szenario folgende Tasks mit den dazugehörigen Befehlssätzen.

6.4.2 Szenario 2 (assembly):

1. Seriennummer der Platine mit dem Handscanner einscannen
    a. Ist die Seriennummer falsch oder nicht im System gespeichert, gibt das System entsprechend eine Meldung zurück.
    b. Ist die Seriennummer richtig, so wird eine Liste mit den zu verbauenden Komponenten angezeigt und das System fordert den Mitarbeiter auf, die erste Komponente einzuscannen, um diese zu verifizieren.
2. Die Seriennummer des Behälters mit der entsprechenden Komponente wird per Handscanner eingescannt, um die vom System geforderte Komponente zu verifizieren.
    a. Ist die eingescannte Seriennummer nicht bekannt, so gibt das System eine entsprechende Rückmeldung.
    b. Ist die eingescannte Seriennummer richtig, so weist das System den Mitarbeiter die Komponente an die richtige Stelle auf der Platine zu verbauen.
    Die richtige Stelle wird auf dem Monitor visuell per Kreuzlinien markiert.
    c. Wurde die Komponente vom Mitarbeiter an der richtigen Stelle installiert, so kann der Mitarbeiter dies mit den Befehlen „confirm“ oder „installed“ dem System zurückmelden.
    d. Sind weitere Komponenten vorhanden, so fordert das System den Mitarbeiter dazu auf, die nächste Komponente einzuscannen und es werden wieder die Schritte a bis c durchgeführt.

e. Wurden alle vorhandenen Komponenten erfolgreich verbaut, so meldet das System, dass der Prozess abgeschlossen ist und fordert den Mitarbeiter dazu auf, den Abschluss mit dem Befehl „confirm“ zu bestätigen.

3. Allgemein
    a. Der Prozess kann jederzeit durch den Mitarbeiter mit dem Befehl „cancel“ abgebrochen werden. Bei Abbruch setzt sich das System wieder in den Ausgangszustand zurück.

7 Evaluationsergebnisse und Diskussion

In diesem Kapitel werden die Evaluationsergebnisse präsentiert. Dabei werden zuerst die Ergebnisse der indirekten Daten veranschaulicht. Danach erfolgt die Auswertung der direkten Daten, welche durch den modifizierten SASSI-Fragebogen ermittelt wurden.

7.1 Indirekte Daten

Zu den indirekten Daten zählen die Wortfehlerraten, welche anhand der Unterteilung in Einfügungen, Auslassungen und Ersetzungen vermerkt wurden. Abschließend wird das Ergebnis zu der Anzahl der Äußerungen, die nicht von der Grammatik abgedeckt wurden (OOG), präsentiert. Die indirekten Daten werden für beide Szenarien separat veranschaulicht.

WER

Die Wortfehlerrate wird in Tabelle 2 verdeutlicht. Bei der Berechnung wurden auch Äußerungen der Probanden berücksichtigt, die beim Sprachsystem nicht zu einem Fehler geführt haben. Bei dem Versuch hat das Sprachsystem keine Wörter hinzugefügt bzw. die Sprachbefehle nicht falsch interpretiert. Vom System wurden keine Einfügungen vorgenommen. Jedoch wurden insgesamt zwei Sprachbefehle ersetzt und drei Sprachbefehle gelöscht bzw. nicht interpretiert. Ganzheitlich betrachtet ist die Wortfehlerrate mit 35,71% niedrig ausgefallen.

	N = 14
WER in %	35,71
Einfügungen	-
Ersetzungen	2
Auslassungen	3

Tabelle 2: WER in % und Anzahl an Spracherkennungsfehlern (eigene Darstellung)

CER

Die Konzeptfehlerrate wird in Tabelle 3 verdeutlicht. Bei der Berechnung der Konzeptfehlerrate wurden Äußerungen der Probanden, die nicht zu einem Fehler bei der Aufgabenerledigung geführt haben, nicht berücksichtigt. Somit wurden nur die semantischen Einheiten in die Berechnung mit aufgenommen. Insgesamt wurden keine Einfügungen und keine Auslassungen vom System vorgenommen. Allerdings fanden zwei Ersetzungen statt. Somit ist die Konzeptfehlerrate mit 14,3% sehr niedrig ausgefallen.

	N = 14
CER in %	14,3
Einfügungen	-
Ersetzungen	2
Auslassungen	-

Tabelle 3: CER in % und Anzahl an Konzepterkennungsfehlern (eigene Darstellung)

OOG

Tabelle 4 fasst die Ergebnisse der OOG-Rate zusammen. Bei dem Versuch wurden den Probanden explizit die Befehlssätze zu den einzelnen Szenarien schriftlich mitgeteilt (siehe Anhang 4). Dadurch wussten die Teilnehmer genau, welche Befehle sie benutzen konnten, wobei das Sprachsystem auch Kommandos verstanden hat, welche an der passenden Stelle ein Wort mehr als der Befehlssatz enthalten haben.
Ungeachtet dessen haben im ersten Szenario drei Teilnehmer versucht, anstatt des Befehls „remove“ intuitiv das englische Wort „delete“ zu benutzen. Insgesamt fällt die OOG-Rate mit 3,85% sehr niedrig aus. Allerdings ist zu beachten, dass die Benutzung des Wortes „delete“ intuitiv erfolgt ist. Somit ist es empfehlenswert, dieses Wort mit in die Grammatik aufzunehmen.

	N = 26
OOG-Rate in %	3,85

Tabelle 4: OOG-Rate (eigene Darstellung)

7.2 Direkte Daten des modifizierten SASSI-Fragebogens

Der modifizierte SASSI-Fragebogen enthielt insgesamt 15 Fragen. Die ersten 12 Fragen wurden anhand einer 5-stufigen Likert-Skala bewertet und die letzten drei Fragen waren offene Fragen. Der Fragebogen enthielt sowohl positiv formulierte, als auch negativ formulierte Fragen. Für die Antworten wurde die Likert-Skala von „lehne stark ab“ (-2) bis hin zu „stimme stark zu“ (2) zur Hilfe herangezogen. Bei der Frage nach dem Gesamteindruck des Systems (siehe Abbildung 21) wurden die Skalenstufen von „sehr schlecht“ (-2) bis „sehr gut“ (2) angepasst.

Im Folgenden werden zunächst die Antworten von allen Probanden zu den einzelnen Fragen jeweils durch eine Abbildung veranschaulicht. Dabei werden die Ergebnisse erörtert und gegebenenfalls vorhandene Zusammenhänge aufgezeigt.

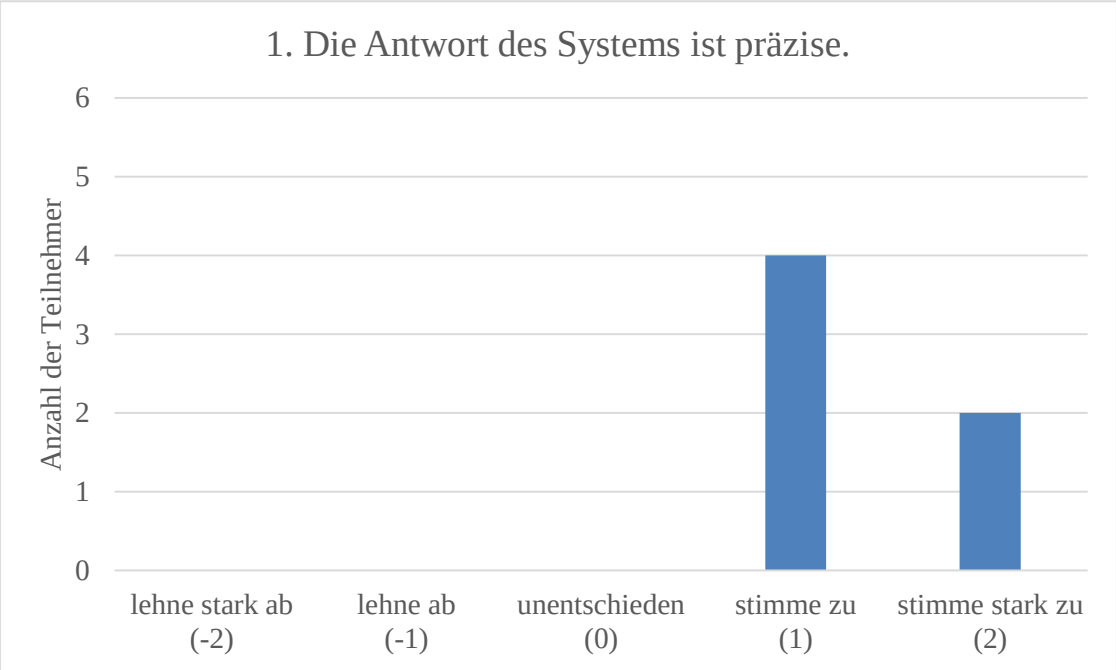

Abbildung 10: Gesamtbewertung der Probanden zu Frage 1 (eigene Darstellung)

Abbildung 10 veranschaulicht, dass alle Teilnehmer der Aussage *„Die Antwort des Systems ist präzise“* zustimmen. Insgesamt stimmen zwei Teilnehmer stark zu und drei Teilnehmer stimmen einfach zu. Somit kann angenommen werden, dass die Sprachrückmeldungen des Systems für alle Teilnehmer verständlich waren.

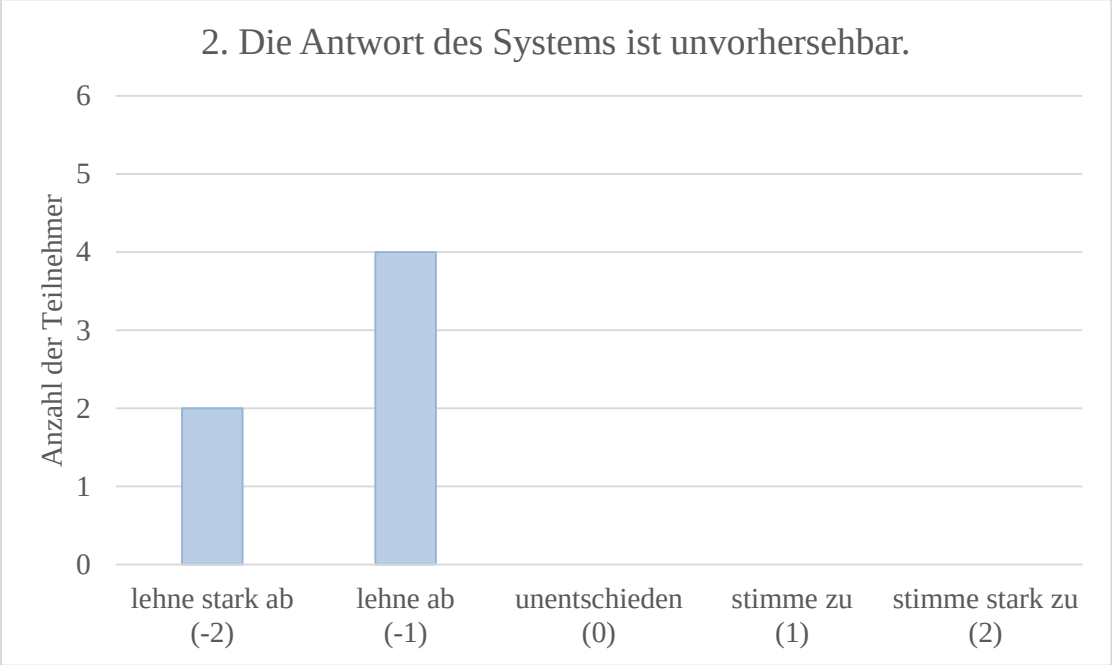

Abbildung 11: Gesamtbewertung der Probanden zu Frage 2 (eigene Darstellung)

Aus der vorangegangenen Abbildung 11 lässt sich ableiten, dass alle Teilnehmer der Aussage *„Die Antwort des Systems ist unvorhersehbar*“ widersprechen. Dabei lehnen vier Teilnehmer diese Aussage ab und zwei Teilnehmer lehnen diese Aussage stark ab. Dies führt zu der Schlussfolgerung, dass die Rückmeldungen des Sprachsystems demnach korrekt sind und das System kein Fehlverhalten aufweist.

Diese Folgerung kann durch die Beobachtung bestärkt werden, dass die Teilnehmer während des Versuchs keine Fragen bezüglich der gegebenen Systemantworten gestellt haben.

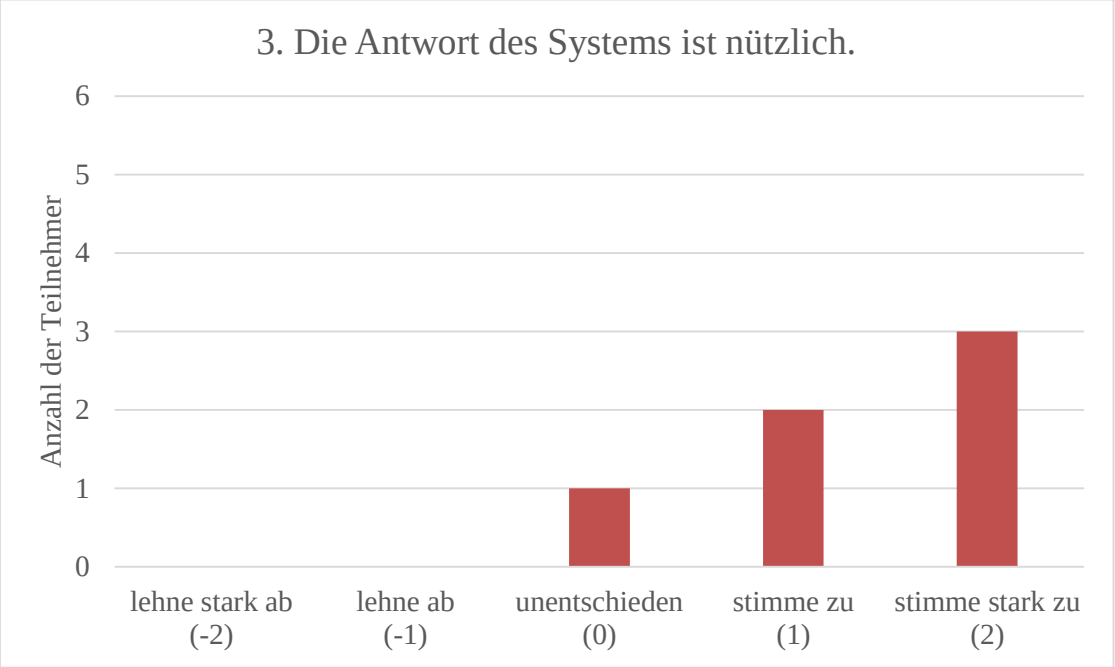

Abbildung 12: Gesamtbewertung der Probanden zu Frage 3 (eigene Darstellung)

Die Bewertung der Aussage *„Die Antwort des Systems ist nützlich“* fällt insgesamt positiv aus. Es ist anzumerken, dass kein Teilnehmer diese Aussage negativ bewertet hat. Ein Proband war unentschieden und hat diese Frage neutral bewertet, fünf Teilnehmer fanden die Antworten des Systems nützlich.

Zusammengefasst ist festzustellen, dass die Antworten des Systems bei beiden Szenarien für die Erledigung der Aufgaben nützlich waren. Während des Versuchs haben alle Teilnehmer die Rückmeldungen des Sprachsystems verstanden. Diese Annahme wird durch die Tatsache bestärkt, dass alle Probanden die Interaktion mit dem passenden Befehl fortgeführt haben. Daher kann die einzige neutrale Bewertung als Ausreißer gedeutet werden.

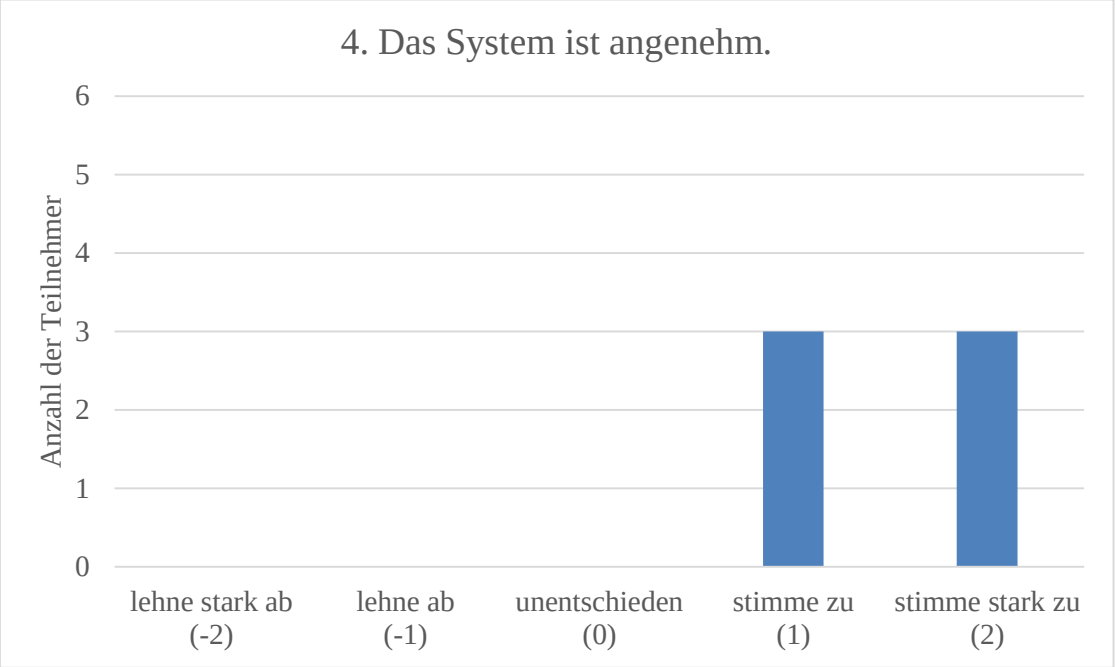

Abbildung 13: Gesamtbewertung der Probanden zu Frage 4 (eigene Darstellung)

Abbildung 13 veranschaulicht die Bewertung der Probanden bezüglich der Bedienbarkeit des Sprachsystems. Alle Teilnehmer fanden, dass das System angenehm zu bedienen war. Drei Teilnehmer fanden das System ansprechend und drei Teilnehmer fanden das System sehr angenehm.

Ganzheitlich betrachtet kann von der Bewertung abgeleitet werden, dass die Usability des Systems als hoch eingestuft werden kann.

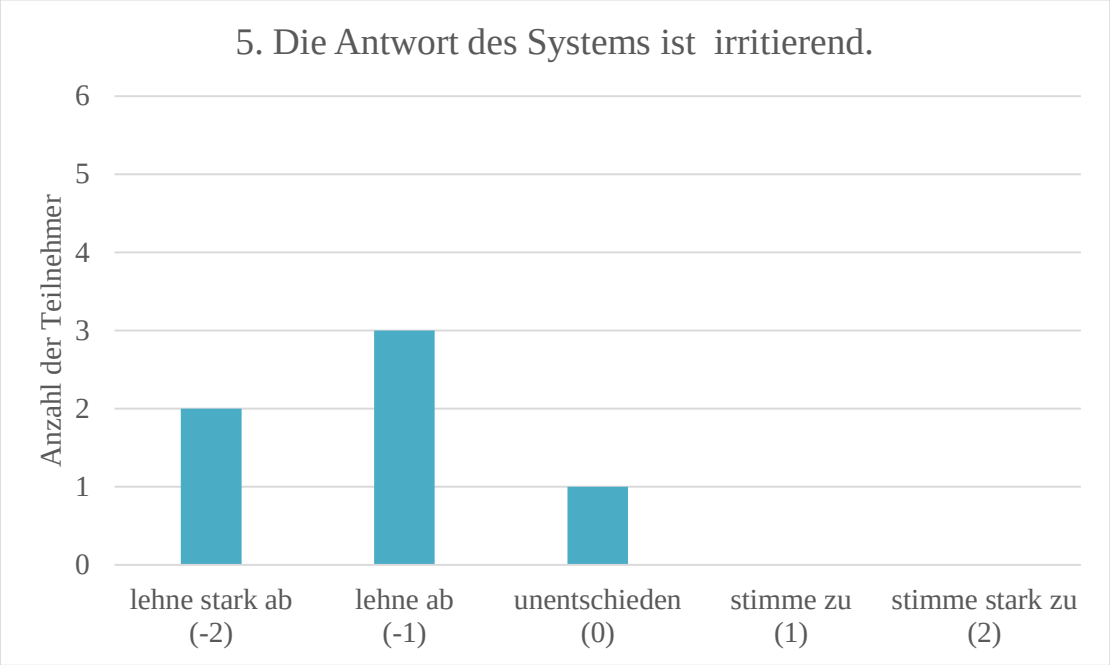

Abbildung 14: Gesamtbewertung der Teilnehmer zu Aussage 5 (eigene Darstellung)

Die Frage, ob die Antwort des Systems irritierend sei, wurde insgesamt von den Teilnehmern abgelehnt. Dabei haben zwei Teilnehmer diese Aussage stark abgelehnt, drei Teilnehmer haben diese Aussage abgelehnt und ein Teilnehmer hat diese Aussage neutral bewertet.

Die einzelnen Bewertungen verdeutlichen, dass die Antworten des Sprachsystems die Teilnehmer in beiden Szenarien nicht irritiert haben. Vergleicht man diese Bewertung mit der Bewertung der Aussage *„Die Antwort des Systems ist nützlich“*, ist festzustellen, dass beide Bewertungen zusammenhängen und ähnlich sind. Der Aussage hinsichtlich der Nützlichkeit der Systemantwort wurde insgesamt zugestimmt, wobei hier auch ein Teilnehmer diese Aussage neutral bewertet hat.

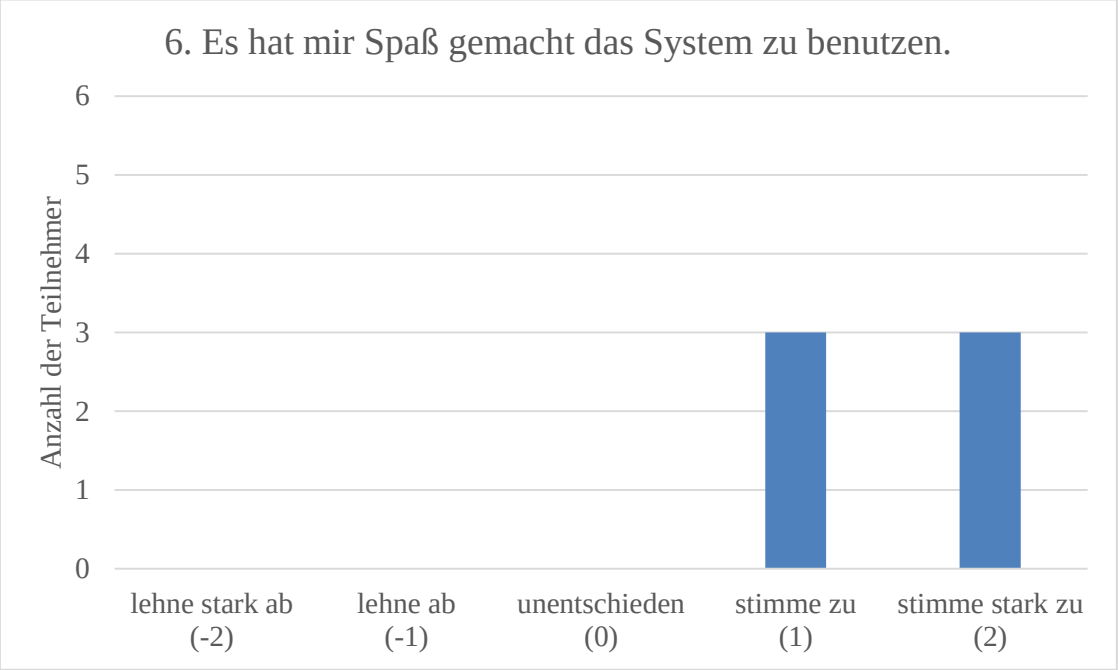

Abbildung 15: Gesamtbewertung der Aussage 6 (eigene Darstellung)

Die Aussage bezüglich des Spaßfaktors wurde von allen Teilnehmern insgesamt gut bewertet. Drei Teilnehmer haben zugestimmt und drei Teilnehmer haben stark zugestimmt.

Dies führt zu der Folgerung, dass die Bedienung des Systems und die Interaktion mit dem System den Teilnehmern Spaß bereitet hat. Daher kann angenommen werden, dass die Benutzung des Systems für die Probanden nicht langweilig war und mit einem gewissen Spaßfaktor einherging.

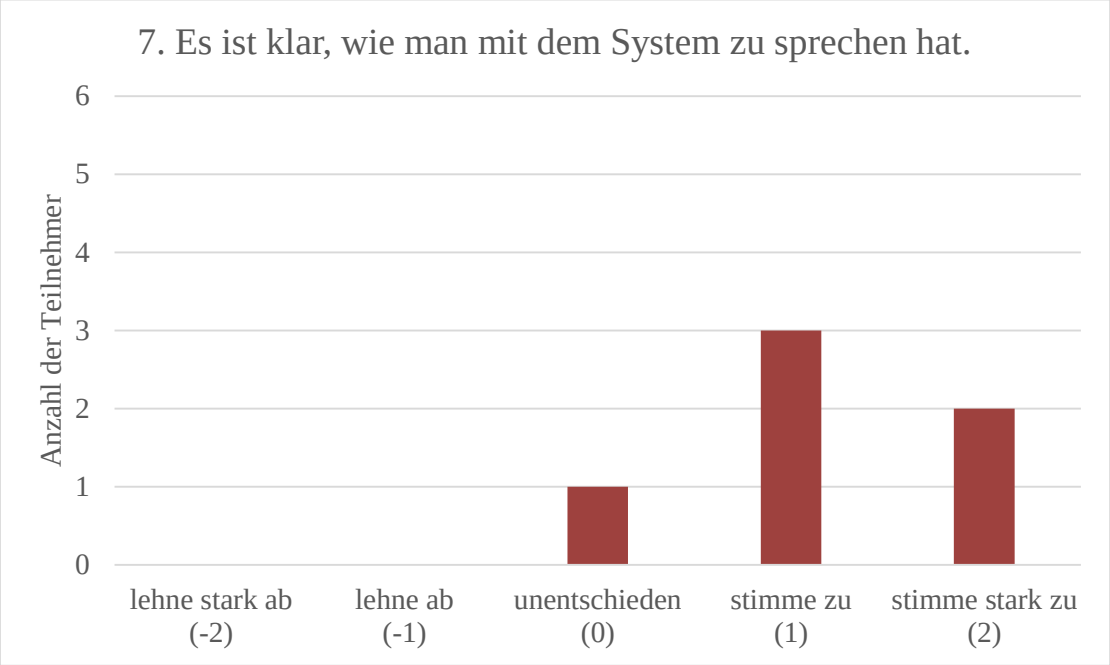

Abbildung 16: Gesamtbewertung zur Aussage 7 (eigene Darstellung)

Die Abbildung 16 veranschaulicht, dass fast alle Teilnehmer im Bilde darüber waren, wie man das Sprachsystem zu bedienen hat. Fünf Teilnehmer haben der Aussage *„Es ist klar, wie man mit dem System zu sprechen hat“* zugestimmt und ein Teilnehmer war unentschlossen.

Bezugnehmend auf die technische Evaluation (Kapitel 8.1, Abschnitt OOG) ist anzumerken, dass von drei Teilnehmern ein Befehl, welcher nicht durch die Grammatik abgedeckt war und dementsprechend vom Sprachsystem nicht erkannt wurde, benutzt wurde. Die neutrale Bewertung (unentschieden) eines Teilnehmers kann darauf zurückgeführt werden, dass dieser Befehl nicht genutzt werden konnte.

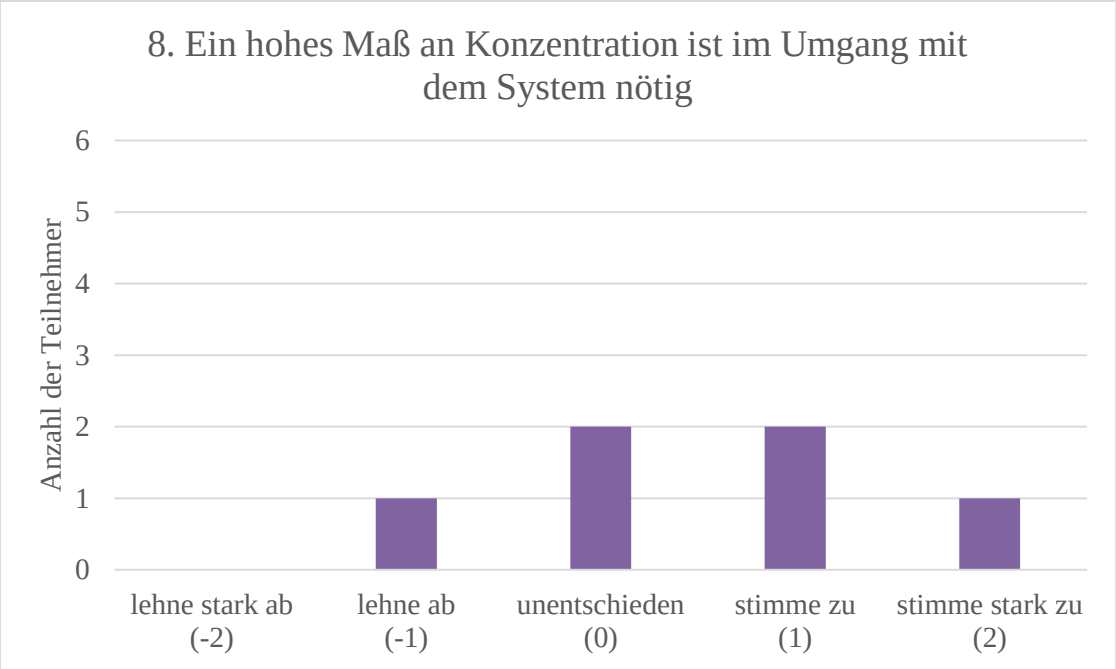

Abbildung 17: Gesamtbewertung zu Aussage 8 (eigene Darstellung)

Hinsichtlich der Bewertung des Maßes an der für den Umgang mit dem System erforderlichen Konzentration sind die Bewertungen sehr unterschiedlich ausgefallen. Ein Teilnehmer fand, dass der Umgang mit dem Sprachsystem keine große Konzentration erforderte, während drei Probanden ein hohes Maß an Konzentration für den Umgang mit dem Sprachsystem voraussetzten. Zwei Teilnehmer waren unentschlossen und haben diese Aussage neutral bewertet.

Es ist schwierig hier eine Schlussfolgerung zu ziehen, aber es kann abgeleitet werden, dass die Bedienung des Systems durchaus ein gewisses Maß an Konzentration erfordert.

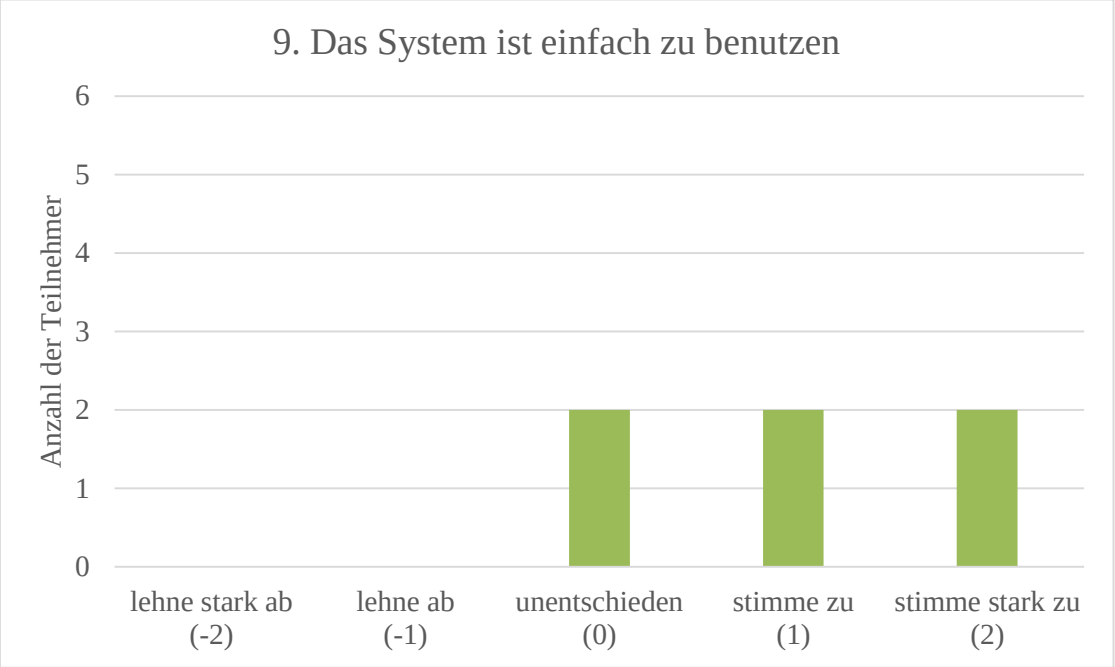

Abbildung 18: Gesamtbewertung zu Aussage 9 (eigene Darstellung)

Kumuliert betrachtet wurde der Aussage *„Das System ist einfach zu benutzen“* zugestimmt. Zwei Teilnehmer stimmten stark zu und zwei stimmten einfach zu. Die restlichen zwei Teilnehmer waren gegenüber dieser Aussage unentschlossen. Vergleicht man diese Bewertung mit der Bewertung zur Aussage 6, so ist anzumerken, dass diese beiden Bewertungen zusammenhängen. Wenn etwas einfach zu benutzen ist, so ist die Bedienung auch mit einem gewissen Spaßfaktor verbunden.

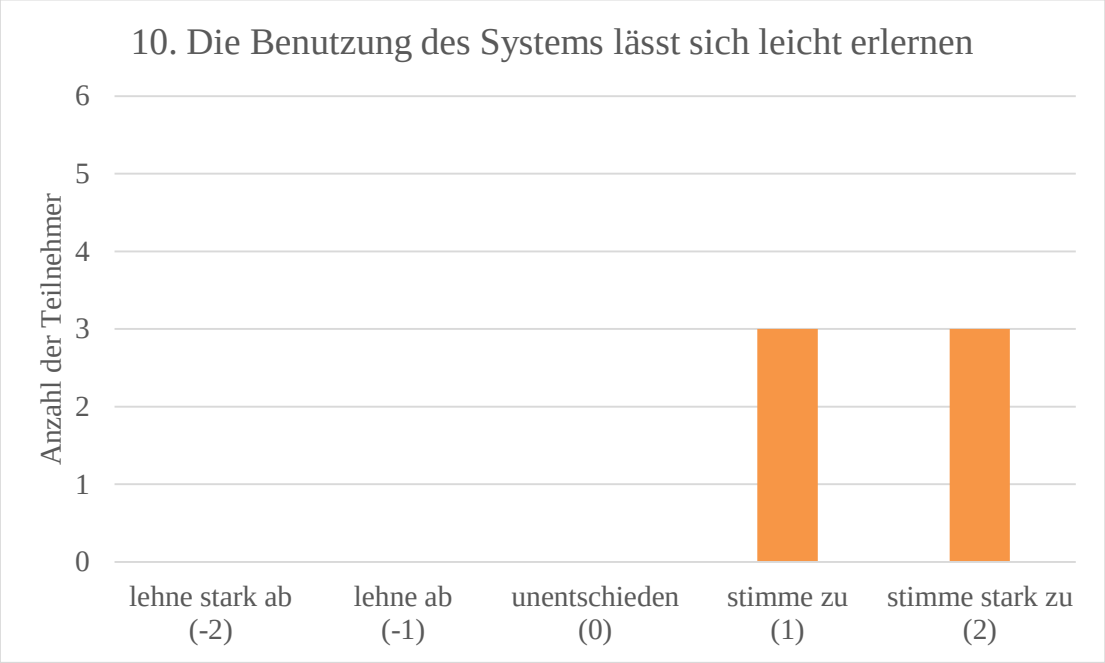

Abbildung 19: Aggregierte Ergebnisse zur Aussage 10 (eigene Darstellung)

Die aggregierten Ergebnisse zu dieser Aussage zeigen auf, dass es den Teilnehmern leichtgefallen ist, die Bedienung des Systems zu erlernen. Dabei ist zu bemerken, dass den Teilnehmern die Bedienung einmal vorgeführt wurde. Weiterhin haben die Teilnehmer auch die Befehlssätze mitgeteilt bekommen, welche sie benutzen konnten.
Diese positive Bewertung ist mit der sechsten Aussage verknüpft, denn ist ein System leicht erlernbar, so bereitet der Umgang mit dem System auch Freude.

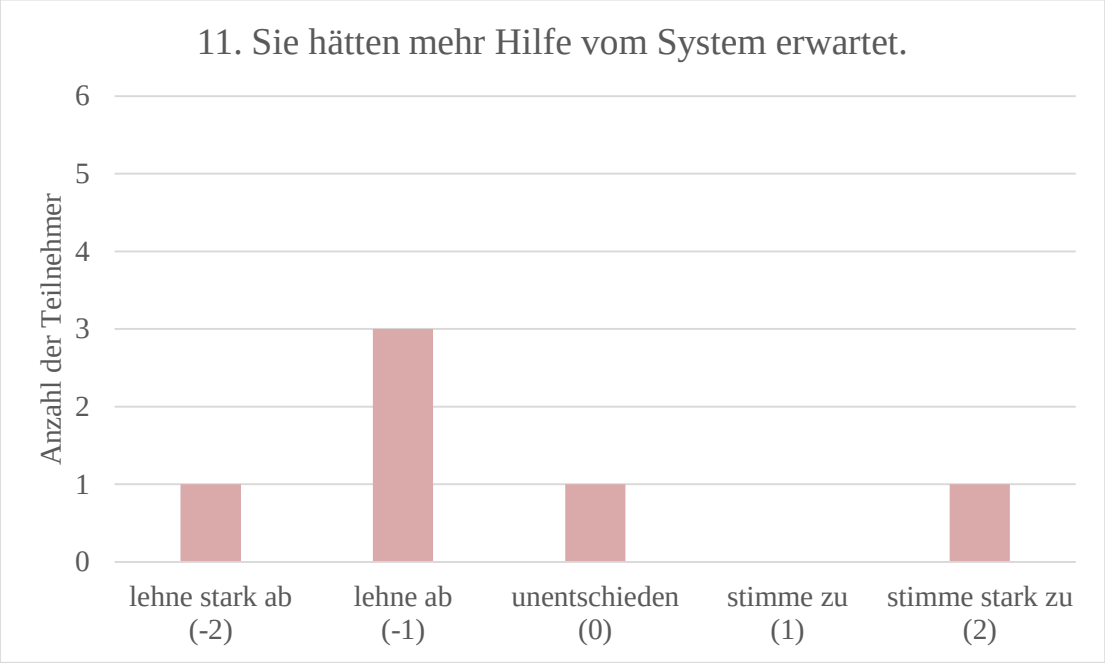

Abbildung 20: Kumulierte Ergebnisse zur Aussage 11 (eigene Darstellung)

Die kumulierte Bewertung zu Aussage 11 zeigt auf, dass die Mehrheit der Teilnehmer mit der vorhandenen Hilfestellung des Systems zufrieden war. Vier Teilnehmer haben nicht mehr Hilfe vom System erwartet. Ein Teilnehmer war unentschlossen und ein Teilnehmer hat entsprechend mehr Hilfestellung vom System erwartet.
Diese Streuung ist nicht nachvollziehbar, denn alle Teilnehmer haben angegeben, dass die Bedienung des Systems leicht erlernbar sei (siehe Abbildung 19) und dass es ihnen Spaß bereitet hat, das System zu benutzen (siehe Abbildung 15).

Aufgrund dieser Gegebenheit kann der Wunsch eines Teilnehmers nach der starken Hilfestellung des Systems als Ausreißer interpretiert werden.

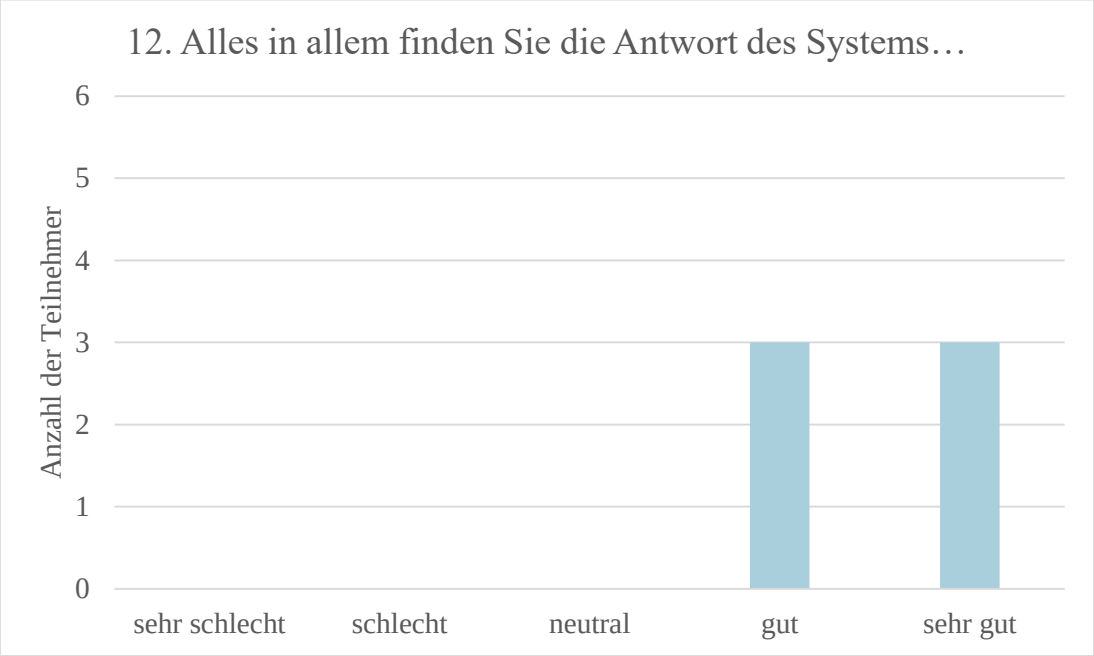

Abbildung 21: Aggregierte Ergebnisse zur Aussage 12 (eigene Darstellung)

Die abschließende Aussage nach dem Gesamteindruck bezüglich der Rückmeldungen des Systems wurde von allen Teilnehmern positiv bewertet. 50% der Teilnehmer haben diese Aussage mit „gut“ und die restlichen 50% haben diese Aussage mit „sehr gut“ bewertet.

Das führt zu der Schlussfolgerung, dass die Rückmeldungen des Sprachsystems insgesamt sehr zufriedenstellend waren.

Die nachfolgende Tabelle 5 fasst die Ergebnisse zu den drei offenen Fragen zusammen, welche auch den Schlussteil der direkten Daten repräsentieren. Es wurden drei offene Fragen gestellt, wodurch, - in erster Linie, ermittelt werden sollte, ob die Probanden haptische Feedbacks des Handscanners (Vibrieren des Handscanners bei erfolgreichem Scanvorgang der Seriennummern) gegenüber den akustischen bevorzugen würden. Abschließend sollten die Teilnehmer anmerken, ob sie sich einen personalisierten Arbeitsplatz mit multimodaler Bedienung (Touchscreen, Spracheingabe, Gestensteuerung) vorstellen können.

Offene Frage / **Teilnehmer**	Würden Sie in Szenario 2 nach dem Einscannen der Seriennummer eine haptische (z.B. vibrieren des Handscanners) gegenüber einer akustischen Rückmeldung des Systems vorziehen?
01	-
02	Ja
03	Ja
04	Nur wenn die Umgebung zu laut wäre, ansonsten beide Varianten gut
05	-
06	Nein

Offene Frage / **Teilnehmer**	Was hat Ihnen nicht gefallen?
01	Anfangs schwer durchzublicken.
02	Nichts
03	Es wäre schön, wenn man alles nochmal lesen könnte, was gesagt wird.
04	Szenario 1 wusste ich in manchen Situationen nicht, ob ich mich in der Tasklist befinde oder nicht.
05	Das beim Bestücken die Position angezeigt wird, bevor man das Bauteil gescannt hat.

06	Alles gut.
Offene Frage / **Teilnehmer**	Wäre eine Personalisierung ihres Arbeitsplatzes für Sie denkbar (Multitouch-Screen, Spracheingabe, Gestensteuerung)?
01	Ja
02	Ja, wenn man die Tätigkeit ausübt.
03	Ja
04	Ja
05	Ja, bei vollständiger Funktion.
06	Ja

Tabelle 5: Ergebnisse zu den offenen Fragen (eigene Darstellung)

Aus den einzelnen Antworten der Teilnehmer zu der offenen Frage nach dem haptischen Feedback des Handscanners kann man entnehmen, dass die Hälfte der Gesamtteilnehmer ein haptisches Feedback vorziehen würde, beziehungsweise sich auch ein haptisches Feedback vorstellen könnte. Ein Drittel der Teilnehmer hat zu dieser Frage keine Aussage getätigt und ein Teilnehmer würde eine akustische Rückmeldung gegenüber einer haptischen Rückmeldung vorziehen. Somit kann abgeleitet werden, dass eine haptische Rückmeldung, die man ein- und ausstellen kann, durchaus denkbar wäre.

Die offene Frage, ob den Teilnehmern etwas missfallen hat, wurde von zwei Teilnehmern verneint. Ein Teilnehmer gab an, dass es anfangs etwas schwierig war, mit der Bedienung des Sprachsystems zurecht zu kommen *(„Anfangs schwer durchzublicken“)*. Mit dieser Aussage kann die Bewertung des Ausreißers zu Aussage 11 *(sie hätten mehr Hilfe vom System erwartet)* begründet werden. Demnach hat ein Teilnehmer mehr Hilfestellung

vom System erwartet. Ein Teilnehmer hat angegeben, dass eine textuelle Darstellung der eingegebenen Sprachbefehle und akustischen Rückmeldungen des Systems hilfreich gewesen wäre. Diese Aussage ist nicht nachvollziehbar, denn den Teilnehmern wurde vor dem Test auch die visuelle Benutzeroberfläche des Windows-Clients am Monitor erläutert. Zu den Funktionalitäten des Clients zählte sowohl die textuelle Darstellung der Sprachbefehle, als auch der akustischen Rückmeldungen des Systems. Somit ist diese Aussage des Teilnehmers nicht plausibel.

Ein weiterer Kritikpunkt eines Teilnehmers war, dass bei Szenario 2 nicht immer eindeutig war, ob man sich in der Aufgabenliste (tasklist) befindet oder nicht. Demnach wurde der Befehl „repeat“ verwendet um die letzte akustische Rückmeldung des Systems wiederholen zu lassen und somit sich zurecht zu finden. Diese Aussage des Teilnehmers ist sehr hilfreich und veranschaulicht eine wichtige Schwachstelle des Sprachsystems. Daher wäre es denkbar, das System entsprechend - um eine Art Navigation - zu erweitern. Dies wäre beispielsweise durch einen zusätzlichen Befehl wie „current state“, in dem das Sprachsystem die aktuelle Position in der Navigation rückmeldet, umsetzbar.

Ein zusätzlicher Kritikpunkt eines Teilnehmers war es, dass bei Szenario 2 beim Bestücken der Platine bereits die Position der zu verbauenden Komponente auf dem Bildschirm angezeigt wurde, bevor man das entsprechende Bauteil eingescannt hatte. Also direkt nachdem man die richtige Anbringung der vorangegangenen Komponente dem System mit dem entsprechenden Sprachbefehl bestätigt hatte, wurde sofort die Position der nächsten Komponente auf dem Bildschirm angezeigt. Dieser Zustand, nach Angabe eines Teilnehmers, würde dazu animieren, die Komponente auf der Platine zu verbauen, ohne vorher die Komponente durch das Einscannen zu verifizieren. Aufgrund der einfachen Architektur des Systems wäre es möglich, diesen Punkt zu verbessern und dadurch die Usability des Systems zu erhöhen.

Zusammenfassend kann festgestellt werden, dass das Testsystem durchaus Probleme aufweist, welche sich allerdings im Rahmen halten und durch geringen Aufwand gelöst werden können.

8 Fazit

Gegenstand dieser Masterarbeit war die Evaluation eines Sprachdialogsystems in der Produktionsumgebung mit multimodaler Eingabe. Dazu wurden verschiedene, marktunabhängige Sprachsysteme ermittelt und anhand der vergleichenden Evaluation - mit Hilfe von spezifischen Kriterien - miteinander verglichen. Aus dieser Bewertung wurde das Sprachsystem Lydia® Voice Control, der Firma topsystems Systemhaus GmbH, ausgewählt, um an das spezifische Manufacturing Execution System der Firma iTAC Software AG integriert zu werden.
Das Testsystem umfasste eine Koppelung des MES mit dem ausgewählten Sprachsystem, sowie weitere Hardwaregeräte und Komponenten. Um die Integration des Sprachsystems auf Benutzerfreundlichkeit und Bedienbarkeit zu testen, wurden sowohl Benutzertests mit Hilfe von Task-Szenarien durchgeführt, als auch ein SASSI-Fragebogen mit spezifischen Fragen zur Usability des Systems erstellt. Dieser Fragebogen wurde von den Versuchsteilnehmern nach den Benutzertests entsprechend ausgefüllt. Weiterhin wurden auch technische Kriterien anhand der referenzbasierten Evaluation für die Bewertung des Systems herangezogen. Zu den technischen Kriterien zählen die Wortfehlerrate, die Konzeptfehlerrate und die Out-of-grammar-Rate. Die vorliegende Arbeit dokumentiert und diskutiert dabei die gewonnenen Ergebnisse aus den Versuchen.

Eingangs wurden allgemeine theoretische Grundlagen gelegt, um einen Einblick in die Thematik zu gewähren. Danach wurden Dialogsysteme und die dazugehörigen Dialogstrategien erläutert. Darauf folgte die Veranschaulichung der verschiedenen Evaluationstechniken und der technischen Kriterien für Sprachdialogsysteme. Weiterhin wurden auch Grammatiken - speziell kontextfreie Grammatik - beleuchtet, da im Rahmen dieser Masterarbeit eine solche Grammatik für das Testsystem erstellt wurde.

Die technische Implementation des Prototyps ist hingegen nicht Teil dieser Arbeit und wird aufgrund dessen auch nicht weiter thematisiert.

Im Anschluss wurde die Herangehensweise bei der realisierten Usability-Evaluation dargelegt. Dabei wurden anhand von Benutzertests zwei Szenarien arrangiert, in denen die Probanden unterschiedliche Aufgaben bewältigen sollten. Das erste Szenario stellte die Arbeitsumgebung eines Wartungsmitarbeiters in einer oder mehreren Produktionslinien dar. Hierbei sollten Maschinenzustände verändert werden und auftretende Aufgaben, welche durch die verschiedenen Maschinenzustände anfielen, erledigt werden. Konkret waren zwei Maschinenzustände zu setzen und zwei Aufgaben von der virtuellen Aufgabenliste abzuarbeiten. Für die Steuerung standen den Versuchsteilnehmern verschiedene Sprachbefehle zur Verfügung, die ihnen anfangs kommuniziert wurden.
Das zweite Szenario stellte die Arbeitsumgebung eines Mitarbeiters an einer Bestückungsstation dar. Dabei sollte eine Platine mit drei verschiedenen Komponenten bestückt werden. Die Interaktion mit dem Dialogsystem erfolgte einerseits anhand von Sprachbefehlen und andererseits anhand eines Handscanners, wodurch Seriennummern eingescannt wurden. Zuerst sollte die Seriennummer der Platine eingescannt werden. Bei erfolgreicher Verifizierung der Platine durch das System, wurde der Benutzer dazu aufgefordert, die Seriennummer der ersten zu verbauenden Komponente einzuscannen. Wurde die Komponente vom System erfolgreich verifiziert, erfolgte die Aufforderung zum Einscannen der zweiten Komponente. Wurden alle Komponenten vom Benutzer erfolgreich an die richtige Position verbaut, musste der Benutzer die Fertigstellung der Platine dem System mit dem passenden Sprachbefehl bestätigen. In beiden Szenarien wurde den Probanden die zu erledigenden Aufgaben einmal vom Versuchsleiter vorgeführt.

Mit Blick auf die Ergebnisse in Kapitel 7 kann zusammengefasst werden, dass die Probandinnen/Probanden die Benutzerfreundlichkeit des Testsystems als gut empfunden haben. Die Mehrheit der Versuchsteilnehmer hatte beim Umgang mit dem Testsystem keine großen Probleme und konnte die Sprachbefehle gut anwenden. Es ist dazu anzumerken, dass einige Benutzer auch intuitiv Kommandos um passende Wörter ergänzt haben. Dabei erkannte das Testsystem den entsprechenden Befehl anhand von Schlüsselwörtern und führte die entsprechende Aktion aus. Diese Funktionalität hat die Versuchsteilnehmer sehr erfreut wodurch sie auch Spaß an der Bedienung des Testsystems hatten. Obwohl

die Benutzer beim Versuch die entsprechenden Befehle schriftlich vorliegen hatten, haben einige Benutzer beim ersten Szenario eins bis zwei Befehlswörter nochmal nachgefragt. Dies führte jedoch nicht zu irgendwelchen Schwierigkeiten. Aufgrund dieser Vorkommnisse ist es empfehlenswert, das System um eine Hilfsfunktion zu erweitern, wodurch dem Benutzer alle Sprachbefehle durch das System nochmals auditiv mitgeteilt werden können.
Abschließend wurde von allen Versuchsteilnehmern kommuniziert, dass sie sich einen personalisierten Arbeitsplatz mit sprachgesteuerter und/oder multimodaler Bedienung gut vorstellen können, denn es würde zu einer effizienteren Arbeitsumgebung beitragen. Konkret haben auch alle Teilnehmer angegeben, dass das System leicht zu erlernen und einfach zu benutzen sei. Daraus kann abgeleitet werden, dass die Benutzerfreundlichkeit des Testsystems als qualitativ hoch eingestuft werden kann. Dies wird durch die Testergebnisse der indirekten Daten bestätigt, denn die Wortfehlerrate (35,71%) und die Out-of-grammar Rate (3,85%) des Systems sind sehr niedrig ausgefallen. Dabei ist jedoch zu betonen, dass einige Benutzer ein spezifisches Sprachkommando („delete“) benutzt haben, welches nicht Teil der Grammatik war und dadurch vom Testsystem nicht erkannt wurde. Daher ist es empfehlenswert, dass der Befehl „delete“ als Ergänzung mit in die Grammatik aufgenommen wird.

Zweifellos ist ein Verbesserungspotenzial im Hinblick auf die Usability des Testsystems vorhanden. Gleichwohl bietet das getestete System ein Hohes Maß an Usability und Benutzerfreundlichkeit, die durch die genannten Ergänzungen gesteigert werden können.

Anhang 1: Personalisierter Fragebogen

Geschlecht: **Datum:**

- ○ weiblich
- ○ männlich

Alter:

Haben Sie Erfahrungen mit Systemen die über Sprachsteuerung bedient werden?

○ ja ○ nein

Wie schätzen Sie Ihre Englischkenntnisse ein?

keine	geringe	gute	verhandlungssicher
○	○	○	○

Anhang 2: SASSI Fragebogen

Die Antwort des Systems ist präzise.

lehne stark ab	lehne ab	unentschieden	stimme zu	stimme stark zu
○	○	○	○	○

Die Antwort des Systems ist unvorhersehbar.

lehne stark ab	lehne ab	unentschieden	stimme zu	stimme stark zu
○	○	○	○	○

Die Antwort des Systems ist nützlich.

lehne stark ab	lehne ab	unentschieden	stimme zu	stimme stark zu
○	○	○	○	○

Das System ist angenehm.

lehne stark ab	lehne ab	unentschieden	stimme zu	stimme stark zu
○	○	○	○	○

Die Antwort des Systems ist irritierend.

lehne stark ab	lehne ab	unentschieden	stimme zu	stimme stark zu
○	○	○	○	○

Es hat mir Spaß gemacht das System zu benutzen.

lehne stark ab lehne ab unentschieden stimme zu stimme stark zu

○ ○ ○ ○ ○

Es ist klar, wie man mit dem System zu sprechen hat.

lehne stark ab lehne ab unentschieden stimme zu stimme stark zu

○ ○ ○ ○ ○

Ein hohes Maß an Konzentration ist im Umgang mit dem System nötig.

lehne stark ab lehne ab unentschieden stimme zu stimme stark zu

○ ○ ○ ○ ○

Das System ist einfach zu benutzen.

lehne stark ab lehne ab unentschieden stimme zu stimme stark zu

○ ○ ○ ○ ○

Die Benutzung des Systems lässt sich leicht erlernen.

lehne stark ab lehne ab unentschieden stimme zu stimme stark zu

○ ○ ○ ○ ○

Sie hätten mehr Hilfe vom System erwartet.

lehne stark ab lehne ab unentschieden stimme zu stimme stark zu

○ ○ ○ ○ ○

Alles in Allem finden Sie die Antwort des Systems…

sehr schlecht	schlecht	neutral	gut	sehr gut
○	○	○	○	○

Würden Sie in Szenario 2 nach dem Einscannen der Seriennummer eine haptische (z.B. vibrieren des Handscanners) gegenüber einer akustischen Rückmeldung des Systems vorziehen?

Was hat Ihnen nicht gefallen?

Wäre eine Personalisierung ihres Arbeitsplatzes für Sie denkbar (Multitouch-Screen, Spracheingabe, Gestensteuerung)?

Anhang 3: Instruktionen

Liebe/r Probandin/Proband,

herzlichen Dank, für die Teilnahme an diesem Test. Im Folgenden werden Sie gebeten ein Sprachdialogsystem zu bewerten. Mit Hilfe des Systems ist es möglich die Wartung und Überwachung von Maschinenzuständen und die Bestückung einer Platine vorzunehmen.

Für den Test bitten wir Sie 2 Szenarien mit verschiedenen Aufgaben zu erledigen. Während der Durchführung werden Ihnen evtl. Fragen gestellt, die Sie bitte direkt dem Versuchsmitarbeiter beantworten. Sie müssen sich nichts notieren.
Vor dem Versuch wird Ihnen ein Fragebogen mit persönlichen Informationen ausgehändigt. Dieser Fragebogen ist anonym und dient zur Erfassung von demografischen Daten. Wir möchten Sie bitten diesen persönlichen Fragebogen vor dem Versuch auszufüllen. Danach werden Ihnen die zwei Szenarien vorgestellt und die Erledigung der Aufgaben einmal vorgeführt.

Ihre Meinung und ihr spontanes Verhalten über das System sind uns wichtig. Wir möchten Sie nicht testen, sondern unser System. Sollte es Schwierigkeiten mit der Erledigung der Aufgaben geben, so ist das nicht Ihr Fehler, sondern ein Problem unseres Systems. Sie helfen uns mit diesem Test ungemein bei der Weiterentwicklung. Alle Daten, die während des Versuchs aufgenommen werden, werden von uns vertraulich behandelt und anonym ausgewertet. Bei Fragen können Sie sich jederzeit an den Versuchsleiter wenden.

Bearbeiten Sie bitte beide Szenarien der Reihe nach.
Nach der Bearbeitung beider Szenarien bitten wir Sie einen systembezogenen Fragebogen auszufüllen. Dabei sollen Sie das System und die Interaktion mit dem System bewerten. Sollten Sie Hinweise oder Kommentare haben, die durch die Fragen nicht abgedeckt werden, so bitten wir Sie diese am Ende des Fragebogens festzuhalten.

Für die Sprachsteuerung stehen Ihnen die vorgestellten Befehle in englischer Sprache zur Verfügung. Geben Sie dem System klare Befehle, die entweder aus einem oder mehreren Wörtern bestehen. Die passenden Befehlssätze finden Sie auf der nächsten Seite.

Am Anfang des ersten Szenarios können Sie die Interaktion mit dem System anhand des Befehls „set condition“ oder „set machine condition“ aktivieren. Die weitere Interaktion erfolgt aufgabenorientiert mit den dazugehörigen Kommandos.
Beim zweiten Szenario können Sie die Interaktion mit dem System durch das Einscannen der Seriennummer der Dummy-Platine aktivieren (das System weist Sie am Anfang darauf hin).

Und nun wünschen wir Ihnen viel Spaß beim Versuch!

Anhang 4: Szenarien

Aufgaben zu Szenario 1:

1. Setzen Sie in Linie S01 den Status der Station „Pick&Place“ auf „line maintenance.“
2. Nehmen Sie den neuen Task an.
3. Rufen Sie die Tasklist auf.
4. Setzen Sie die Priorität des ersten Tasks auf „high“.
5. Löschen Sie Task 1 aus der Tasklist.
6. Wechseln Sie zum Setzen des Maschinenstatus.
7. Veranlassen Sie das System den Status der Station „Pick&Place“ in der Linie S01 auf „change log setup“ zu setzen.
8. Das System meldet einen neuen anfallenden Task. Löschen Sie diesen Task direkt und nehmen Sie ihn nicht an.

Verfügbare Befehle in Szenario 1:

- *„Set condition“* oder *„set … condition“*
- *„Select“* oder *„select …“*
- *„Next“* oder *„next ….“* oder *„… next …“*
- *„Previous“* oder *„previous …“* oder *„… previous …“*
- *„Confirm“* oder *„… confirm“*
- *„Tasklist“* oder „*… tasklist*“
- *„Low/medium/high priority“* oder *„… low/medium/high priority“*
- *„Remove“* oder *„remove …“*

Aufgaben zu Szenario 2:

1. Scannen Sie die Seriennummer der Platine ein.
2. Erledigen Sie die Bestückung der Platine indem Sie den Anweisungen des Systems folgen.
3. Bestätigen Sie die Bestückung der Platine an das System.

Verfügbare Befehle in Szenario 2:

- *„confirm“* oder *„… confirm“*
- *„installed“* oder *„… installed“*

Anhang 5: GRAMMATIK

= <Noise> | <Commands> | <Number>

<Noise> = $Replace(a, Noise1) | $Replace(i, Noise2) | \
$Replace(s, Noise3)

<Commands> = <cmdRepeat> | <cmdRestart> | \
<cmd1> | <cmd2> | <cmd3> | <cmd4> | <cmd5> | <cmd6> | <cmd7> |
<cmd8> | <cmd9> | <cmd10>

<cmdRepeat> = $Dynamic(CmdRepeat, repeat)

<cmdContinue> = $Dynamic(CmdContinue, continue)

<cmdCancel> = $Dynamic(CmdCancel, cancel)

<cmdFinished> = $Dynamic(CmdFinished, location empty)

<cmdRestart> = $Dynamic(CmdRestart, restart)

<cmd1> = $Dynamic(Cmd1, task list)

<cmd2> = $Dynamic(Cmd2, next)

<cmd3> = $Dynamic(Cmd3, confirm)

<cmd4> = $Dynamic(Cmd4, remove)

<cmd5> = $Dynamic(Cmd5, high priority)

<cmd6> = $Dynamic(Cmd6, medium priority)

<cmd7> = $Dynamic(Cmd7, low priority)

<cmd8> = $Dynamic(Cmd8, <finish>)

<cmd9> = $Dynamic(Cmd9, set condition)

<cmd10> = $Dynamic(Cmd10, select)

<finish> = $Replace(finish | cancel, finish)

<Number> = <n1> | <n2> | <n3> | <n4> | <n5> | <n6> | \
<n1NE> | <n2NE> | <n3NE> | <n4NE> | <n5NE> | <n6NE>

<n1> = $Dynamic(Num1, $Digits(0N, 1) <okay>)

<n2> = $Dynamic(Num2, $Digits(0N, 2) <okay>)

<n3> = $Dynamic(Num3, $Digits(0N, 3) <okay>)

<n4> = $Dynamic(Num4, $Digits(0N, 4) <okay>)

<n5> = $Dynamic(Num5, $Digits(0N, 5) <okay>)

<n6> = $Dynamic(Num6, $Digits(0N, 6) <okay>)

<n1NE> = $Dynamic(Num1NE, $Digits(0N, 1))

<n2NE> = $Dynamic(Num2NE, $Digits(0N, 2))

<n3NE> = $Dynamic(Num3NE, $Digits(0N, 3))

<n4NE> = $Dynamic(Num4NE, $Digits(0N, 4))

<n5NE> = $Dynamic(Num5NE, $Digits(0N, 5))

<n6NE> = $Dynamic(Num6NE, $Digits(0N, 6))

<okay> = $Replace(okay | okee | okeh, okay)

Anhang 6: Technische Auswertung

Szenario 1		
Insertions	**Substitutions**	**Deletions**
Szenario 2		
Insertions	**Substitutions**	**Deletions**

Literaturverzeichnis

Bauernhansl, T. 2017: Die Vierte Industrielle Revolution – Der Weg in ein wertschaffendes Produktionsparadigma. In B. Vogel-Heuser, T. Bauernhansl & M. ten Hompel (Hrsg.), Handbuch Industrie 4.0. Band 4: Allgemeine Grundlagen (S. 1–31). Springer Berlin Heidelberg.

Bender, Klaus 2005: Embedded Systems – qualitätsorientierte Entwicklung. Springer Berlin.

Bertrand, G. 2014: Situation- and User-Adaptive Dialogue Management. Dissertation, Universität Ulm.

Card S.-K.; Moran T.-P.; Newell A. 1983: The psychology of human-computer-interaction. Lawrence, Erlbaum Associates, New Jersey

Carstensen, K.-U.; Ebert, Ch.; Endriss, C.; Jekat, S.; Klabunde, R.; Langer, H. 2004: Computerlinguistik und Sprachtechnologie, Eine Einführung, 2. überarbeitete Auflage, Spektrum akademischer Verlag.

Carstensen, K.-U.; Ebert, Ch.; Endriss, C.; Jekat, S.; Klabunde, R.; Langer, H. 2010: Computerlinguistik und Sprachtechnologie, Eine Einführung, 3. Auflage, Spektrum akademischer Verlag.

Euler, Stephen 2006: Grundkurs Spracherkennung. Springer Viewig, Wiesbaden.

Google 2017: https://cloud.google.com/speech/?hl=de Stand 13.09.2017.

Hofmann, H.; Ehrlich, U.; Reichel, S.; Berton, S. 2013: „Development of a Conversational Speech Interface Using Linguistic Grammars.“ In: Adjunct Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Eindhoven, The Netherlands.

Hone, K.S. und Graham, R. 2000: Towards a tool for the subjective assessment of speech system interfaces (SASSI). Natural Language Engineering, 6 (3-4): S. 287 - 303.

IBM 2017: https://www.ibm.com/watson/services/speech-to-text/ Stand 13.09.2017.

International Telecommunication Union (ITU) 2003: Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. ITU-T Rec. S.851.

ISO / IEC 2001: International standard ISO/IEC 9126 -1:2001. Software engineering – Product Quality. Part 1: Quality Model. In International Organization for Standardization und International Electrical Commission (Hrsg.), ISO-IEC. Geneva, Switzerland: ISO.

iTAC 2017: http://www.itac.de/pages/company/about/index.html Stand 28.08.2017.

Jeschke, S.; Brecher, C.; Song, H.; Rawat, D. B. 2017: Industrial Internet of Things – Cybermanufacturing Systems, Springer, Cham Schweiz.

Jokinen, K. und M. McTear 2010: Spoken Dialogue Systems. Synthesis lectures on human language technologies. Morgan & Claypool Publishers.

Kagermann, H. (Hrsg.) 2013: Deutschlands Zukunft als Produktionsstandort sichern. Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0: Abschlussbericht des Arbeitskreises Industrie 4.0. Hrsg. von der Promotorengruppe Kommunikation der Forschungsunion Wirtschaft – Wissenschaft und acatech - Deutsche Akademie der Technikwissenschaften. Frankfurt/Main.

Kamm, C. 1995: "User Interfaces for Voice Applications." In: Voice Communication between humans and machines. Bd. 92. National Academy Press.

Karat, C.-M.; Lai, J.; Stewart, O.; Yankelovich, N. 2012: "Speech and Language Interfaces, Applications, and Technologies." In: The Human-Computer Interaction Handbook. Hrsg. von J. Jacko. CRC Press Taylor & Francis Group.

King, M.; Maegard, B.; Schutz, J. und des Tombes, L. 1996: EAGLES – evaluation of natural language processing systems. Technischer Bericht.

Levasseur, Ken und Doerr, Al 2017: Applied Discrete Structures – Part 2 – Algebraic Structures Version 3.3. faculty.uml.edu/klevasseur/ADS2.

Lee, E.A. 2006: Cyber Physical Systems –Are Computing Foundations Adequate? NSF Workshop on Cyber Physical Systems: Research Motivation, Techniques and Roadmap, Austin, USA.

Limtronik GmbH: http://www.limtronik.de/de/unternehmen/ Stand 11.01.2018.

McTear, M. 2002: "Spoken Dialogue Technology: Enabling the Conversational User Interface." In: Computing Surveys 34.1. ACM.

MESA 2000: Controls Definition & MES to Controls Data Flow Possibilities. White Paper Number 3. Pittsburgh, Manufacturing Enterprise Solutions Association.

Nielsen, J. 1993: Usability Engineering. 1. Aufl. AP Professional, Boston

Pinnow, C.; Schäfer, S. 2017: Industrie 4.0, Safety und Security – Mit Sicherheit gut vernetzt. Branchentreff der Berliner und Brandenburger Wissenschaft und Industrie. 1. Auflage. Beutel Verlag GmbH, Berlin.

Pfister, Beat, Kaufmann, Tobias 2017: Sprachverarbeitung. Grundlagen und Methoden der Sprachsynthese und Spracherkennung. 2. aktualisierte und erweiterte Auflage. Springer, Zürich.

Preece, J. 1994: Human-Computer Interaction. Addison Wesley, Harlow

Rubin, J. 1994: Handbook of usability testing: how to plan, design and conduct effective tests. Wiley, New York

Rupp, Chris; die SOPHISTen 2013: Systemanalyse kompakt, 3. Auflage. Springer Verlag, Berlin.

Scheer, August-Wilhelm 2013: Industrie 4.0 – Wie sehen Produktionsprozesse im Jahre 2020 aus https://www.researchgate.net/profile/August_Wilhelm_Scheer/publication/277717764_Industrie_40_-_Wie_sehen_Produktionsprozesse_im_Jahr_2020_aus/links/55ee9e5608ae0af8ee1a1d72/Industrie-40-Wie-sehen-Produktionsprozesse-im-Jahr-2020-aus.pdf

Schenk, J.; Rigoll, G. 2010: Mensch-Maschine-Kommunikation. Grundlagen von sprach- und bildbasierten Benutzerschnittstellen. Springer Verlag, Heidelberg.

Schlick, J.; Stephan, P.; Zühlke, D. 2012: Produktion 2020 – Auf dem Weg zur 4. industriellen Revolution. In: IM – Fachzeitschrift für Information Management und Consulting 27, Ausgabe 3, S. 26-33.

Schukat-Talamazzini, E.G. 1995: Automatische Spracherkennung: Grundlagen, statistische Modelle und effiziente Algorithmen. Vieweg, Braunschweig, Wiesbaden.

Schwab, K. 2016: Die Vierte Industrielle Revolution. Pantheon Verlag, München.

Sonntag, G. 1999: Evaluation der Prosodie. In *Berichte aus der Kommunikationstechnik.* Aachen.

Theel, Sven 2015: Kommissionierung im 21. Jahrhundert. Von Pick-by-Voice bis RFID. Diplomica Verlag, Hamburg.

Topsystem 2017: https://www.topsystem.de/de/lydia-copilot.html Stand 13.09.2017

Uhlmann, E./Hohwieler, E./Kraft, M. 2013: Selbstorganisierende Produktion mit verteilter Intelligenz. In: wt-online, Jg. 103 (2), S. 114-117

Zhai, E.; Shi, Y.; Gregory, M. 2007: The growth and capability development of electronics manufacturing service (EMS) companies. In: Int. J. Production Economics 107 (2007) 1–19. University of Cambridge, USA.

Zühlke, Detlef 2012: Nutzergerechte Entwicklung von Mensch-Maschine-Systemen. Useware-Engineering für technische Systeme. 2. Auflage. Springer Verlag, Heidelberg.