Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5B
D-70569 Stuttgart

Conception and implementation of a vocal assistant for the use in vehicle diagnostics

Mousa Yacoub
Master thesis

Prüfer:	Prof. Dr. Ngoc Thang Vu
Prüferin:	Dr. Antje Schweitzer
Betreuer:	Dipl. Wirt-Inf. Benjamin Leinenbach
Betreuerin:	Dr. Antje Schweitzer

Beginn der Arbeit:	03.12.2018
Ende der Arbeit:	31.05.2019

Erklärung (Statement of Authorship)

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet. Die eingereichte Arbeit ist weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen. Sie ist weder vollständig noch in Teilen bereits veröffentlicht. Die beigefügte elektronische Version stimmt mit dem Druckexemplar überein.

[1]

(Mousa Yacoub)

[1]Non-binding translation for convenience: This text is the result of my own work, and any material from published or unpublished work of others which is used either verbatim or indirectly in the text is credited to the author including details about the exact source in the text. This work has not been part of any other previous examination, neither completely nor in parts. It has neither completeley nor partially been published before. The submitted electronic version is identical to this print version.

Abstract

One of the most important fields have been researched and developed actively lately is the virtual/vocal assistants field. Developers and companies are extending its ability in a rapid rate. Such assistants could nowadays perform any daily task the user tends to do like booking a flight, ordering products, managing appointments, etc. Since this technology is always extended and developed further, we want in this work to survey how far we can adopt it in the vehicle diagnosis field. Concretely we want to develop an Alexa skill that will be used by technicians at workshops to run the vehicle diagnosis process. Main goal there is to avoid any physical interaction with the diagnosis system and at the same time to increase the usability. That's the biggest advantage of vocal assistants, as we can operate the diagnosis system vocally through a conversation with the developed Alexa skill without the need to interact with any keyboard/touch-screen enabled systems. The thesis is organized as follows: First we start with an introduction to the topic, where we discuss the motivation and the problem we want to solve. In the second section we will review some fundamentals for a better understanding. In section 3 we review a related research and in section 4 we get to know the used technology in this thesis, particularly how to build an Alexa skill and from which components it consists. Section 5 reviews my solution approach and which steps I took to reach the goal of the thesis. The implementation and technical details are reflected in section 6. To evaluate the thesis and the resulting product, a study has been conducted. The concept and results of this study are discussed in section 7. Last section includes the conclusion and promising approaches that could be developed on top of the result delivered from this thesis to address further limitations.

Kurzfassung

Eins der wichtigsten Gebiete, das in der letzten Zeit viel erforscht und entwickelt wurde, ist das Gebiet von Sprachassistenten. Entwickler erweitern ständig die Fähigkeiten solcher Sprachassistenten. Solche Assistenten können heutzutage alle tägliche Aufgaben erledigen, die der Benutzer normalerweise selbst erledigen kann, z. B. Flüge buchen, Produkte bestellen, Termine verwalten usw. Da diese Technologie immer erweitert und weiterentwickelt wird, möchten wir in dieser Arbeit untersuchen, inwieweit wir sie im Bereich der Fahrzeugdiagnose einsetzen können. Konkret wollen wir eine Alexa-Skill entwickeln, mit der die Techniker in den Werkstätten den Fahrzeugdiagnoseprozess durchführen. Das Hauptziel dabei ist es, jegliche physische Interaktion mit dem Diagnosesystem zu vermeiden und gleichzeitig die Benutzerfreundlichkeit zu erhöhen. Das ist der größte Vorteil von Sprachassistenten, da wir das Diagnosesystem mit der entwickelten Alexa-Skill sprachlich durch ein Gespräch führen können, ohne mit Tastatur-/Touchscreen-fähigen Systemen interagieren zu müssen. Die Masterarbeit ist wie folgt aufgebaut: Zuerst beginnen wir mit einer Einführung in das Thema, in der wir die Motivation und das Problem, das wir lösen wollen, erläutern. Im zweiten Abschnitt werden wir einige Grundlagen zum besseren Verständnis darstellen. Im dritten Abschnitt dsikutieren wir eine verwandte Arbeit und im vierten Abschnitt lernen wir die verwendeten Technologien kennen, inbesondere aus welchen Komponenten die Skill besteht und wie eine Alexa-Skill entwickelt werden kann. Abschnitt fünf befasst sich mit meinem Lösungsansatz und den Schritten, die ich unternommen habe, um das Ziel der Arbeit zu erreichen. Die Implementierung und die technischen Details sind im Abschnitt sechs wiedergegeben. Zur Bewertung der Masterarbeit und des daraus resultierenden Produkts wurde eine Studie durchgeführt. Das Konzept und die Ergebnisse dieser Studie werden im Abschnitt sieben diskutiert. Der letzte Abschnitt enthält Schlussfolgerungen und vielversprechende Ansätze, die basierend auf der Ergebnisse dieser Arbeit entwickelt werden können, um weitere Probleme zu beheben.

Contents

1 Introduction

In this section I will give an entrance to the topic, some motivation about the environment of the work and I will introduce the problem we are facing and we want to fix by having this work done.

1.1 Motivation

Since the 80's it has been investigated how natural language in the form of text or speech data can be algorithmically processed with the help of the computer. This area of data processing is known as "computational linguistics". Speech recognition is a branch of this area, which is further a branch of Artificial Intelligence (AI), and it is used for Natural Language Processing (NLP), which generally involves text processing and linguistic categorization. Nowadays people relay on technology a lot. Virtual Personal Assistants (VPAs) are one of the most used applications in technology every day. VPAs like Siri, Cortana, and Alexa are online assistant applications that help people and make their life easier. These virtual assistants are not like search engines, yet they are able to set reminders, send messages, tell directions, and make calls. Such VPAs are connected to the Internet and can access the services and information used by voice control. Some of them, like Alexa from Amazon, provide so-called Skills (small apps) which can add more features and extend their capabilities. A skill consists of speech recognition (what did the user say) and speech processing (how should the skill react to what has been said). Most of the applications have so far been only used for private purposes and tasks, such as setting alarms, asking for weather, booking flights and so on. The main focus of my research is to survey how beneficial it is to use such a virtual personal assistant in the industry and not only in the private domain. Specifically we want to examine how efficient it is to employ vocal assistants in the field of vehicle diagnostics at workshops and which current problems this usage addresses.

1.2 Problem description

The title of the thesis already expresses the main problem we are considering through this work. The environment we are working with is vehicle diagnosis. Vehicle diagnosis describes, based on the medical term diagnosis, the task to detect errors in electrical and electronic components on automobiles. The term "vehicle diagnosis" encompasses a number of technical procedures and applications which are used, for example, in error analysis, in quality assurance for statistical evaluations and in vehicle development. In addition, vehicle diagnosis is used to inform or warn the driver about errors that have occurred and to signal limitations of some vehicle properties.
There are multiple stages for running the vehicle diagnosis depending on by whom, how and where it takes place. One possible approach is to do it directly by the customer (the driver), as there are multiple techniques and software already installed on vehicles to allow customers diagnose their vehicle by themselves. Additionally there are some techniques called Over-The-Air (OTA) troubleshooting, i.e., the customer gets support from the customer service and runs the diagnosis process remotely without the need for transporting the vehicle to the workshop. By that they survey if any solutions already exist which could be applied or if the customer should head to one of the nearest workshops to get the problem fixed. This leads us to the next possibility for running the vehicle diagnosis process, namely at the automobile workshop by the car Mechanic.
Normally this procedure is done by using a monitor or special PC located somewhere in the workshop garage as the technician has to give input to it, either via a touch screen or a keyboard. Some problems of this scenario are that the technician would rather give the input without having any physical interaction with the system especially that their hands in the workshop are mostly unclean. Furthermore when he needs to check something in the car and then give the input, he will need to move multiple times between the car and the PC, whereas in best case it's more efficient to give any input directly. So practically, he needs a hands-free system to support him in these tasks.

2 Fundamentals

Here I will provide the basics and principals of the fields we are working with. Those fields form the background, on which the whole work is built.

2.1 Natural Language Processing

2.1.1 Definition

One of the first motivational questions were asked before introducing the Natural Language Processing (NLP) is, if computer programs will be able to convert a piece of written text in some language into a programmer friendly data structure that have the same semantics as the meaning of the natural language text? The existence of such data structure was not clear, until fundamental Artificial Intelligence problems were resolved and opened the insight for promising approaches. These representations are often motivated by specific applications or by our belief that they capture something more general about natural language. NLP is the computerized approach to analyzing text that is based on both a set theories and a set of technologies. It is one of the areas that are being researched and developed very actively and for which there is still no standard or agreed-on definition. NLP includes two main processing areas complementing each other, language processing and language generation [3]. The first area deals with the analysis of language with the goal of providing a meaningful representation, whereas the second area deals with the generation of language from a representation. The authors in [3] gave a definition of NLP as "Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications". There are several ideas in this definition to point out, 'range of computational techniques' interprets that there are variety of approaches can be used for the analysis of a particular type of languages.

'Naturally occurring texts' reflects that there are no boundaries or limitations with respect to the text, it could be of any language, mode or genre also that it can be either written or oral. The only needed property is that it should be in a language used by humans to communicate with each other. 'levels of linguistic analysis' illustrates the idea that there are variety of language processing techniques applied when humans produce or comprehend language. For this reason various NLP systems make use of different levels (or a combination of them) of linguistic analysis. 'Human-like language processing' refers to the fact that NLP is considered a discipline within Artificial Intelligence (AI). Lastly, 'For a range of tasks or applications' explains that NLP is used to achieving a specific goal and is applied in multiple domains and applications, we are focusing on those applications in the following. **Information Retrieval (IR):** Is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections. It is very wide domain applying multiple technologies like NLP to achieve its goal. Statistical approaches for accomplishing NLP also have seen a utilization of NLP. **Information Extraction (IE):** Is built on top of IR and focuses on the recognition, tagging, and extraction into a structured representation. Those extractions can be beneficial and used for a variety of applications such as question-answering and data mining. **Question-Answering:** Can be compared to IR, which returns for a user and a query a list of relevant documents found that most possible fulfill the query. **Dialogue Systems:** Comes as one of the most important use-cases applying NLP techniques. Since such dialogue systems make use of the phonetic and lexical levels of languages. Furthermore this application represents the core of the domain we are dealing with in this thesis.

2.1.2 Processing Pipeline

We will focus in this subsection on the steps in which the NLP pipeline goes through to achieve the NLP objective. An overview of the processing pipeline can bee seen in Figure 1. Analyzing and processing natural language is in general a difficult task, since meaning depends highly on the current context of the speech. To achieve this goal multiple steps are introduced and implemented either by rule-based or machine learning systems [4]. First two steps are called, **Tokenization and Sentence Splitting:** They refer to the orthographic processing part of the pipeline. The main objective here is to break the text apart into separate sentences and for each sentence break it further into separate words, i.e., split the text into lexical unites based on white spaces and punctuation. Tokenization and sentence segmentation are very important since writing a program to understand a single sentence is much easier than to understand a whole paragraph. This task could be easy for splitting sentences apart whenever we reach a punctuation mark or further splitting words based on white spaces between them. But it is not that simple for some cases, such as proper nouns like "United Kingdom of great Britain" should be considered as one unit, or also grammatical abbreviations like "can't, don't and wouldn't.. etc.". Furthermore there are languages without white spaces like Chinese and Korean. For such cases modern NLP pipelines often use more complex techniques that work even when a document isn't formatted cleanly. Third step is **Part-of-Speech Tagging:** This processing step represents the morphologic part of the pipeline and -as its name self explains- it assigns each token to its part-of-speech category. Such categories are for example nouns, verbs, adjective, preposition and so on. Knowing the role of each word in the sentence will help us start to figure out what the sentence is talking about. One possible technique to achieve that is by feeding each word and some extra words around it for context into a pre-trained part-of-speech classification model. Models in this sense are completely based on statistics, i.e., it doesn't actually understand what the words mean in the same way as humans do (it knows how to guess the corresponding part-of-speech category). Next step is **Named-Entity Recognition:** It implements information extraction methods with main goal on finding spans of the text that constitute proper names and tagging them according to their type. In another words the goal of this step is to detect

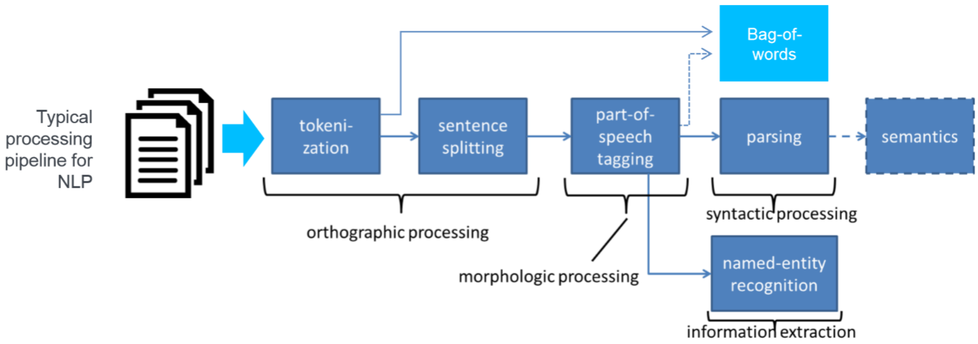

Figure 1: Typical NLP processing pipeline. [4]

and label found nouns with the real-world concepts that they represent. Examples of different types are Persons, organizations, countries and so on. For a better understanding on which role this step has, let us consider the following sentence which represents the input into a named-entity recognition system: "Jim bought 300 shares of Acme Corp. in 2006. " then the output would look something like "$[Jim]_{Person}$ bought $[300]_{Number}$ shares of $[AcmeCorp.]_{Organization}$ in $[2006]_{Time}$". The before last step is **Parsing:** It is the step responsible for syntactic processing of the pipeline. Concretely it assigns for each sentence the corresponding syntactic structure similarly to parsing of programming languages and compilers. Biggest challenge in this step is that natural language sentences are highly ambiguous and could be interpreted in multiple different ways. This is what last step of the pipeline tries to address. **Semantics:** Aims at giving a semantic role to a syntactic constituent of a sentence. One technique to do that is defining arguments (actors) in the sentence and the relation between those arguments (namely, verbs). So for a given sentence like "John ate the apple" it will produce the output "$[John]_{first-argument}[ate]_{Relation}[\text{the } apple]_{second-argument}$". The precise arguments depend on a verb's frame and if there are multiple verbs in a sentence some words might have multiple tags.

2.2 Vocal Assistants

2.2.1 Voice Computing

Voice or Vocal Computing describes the ability to interact (vocally) by language with computers, mobiles and stationary devices, as well as with their software-based applications. As an intuitive and natural Interface Voice Computing represents an innovative development progress in the context of human-machine-interaction [7]. Currently, voice computing is primarily integrated into virtual intelligent assistants, which diffuse into two main application contexts: The technology either extends the functionality of smart phones and other mobile devices as an app-based solution or it plays the role of a smart home application which helps to network and connect the private living environment. While applications such as "Siri" from Apple and "Now" from Google are already widely integrated into the practice of using smart phones through their distribution over IOS and Android, other artifacts such as "Alexa" from Amazon or "Home Assistant" from Google form a new category of products for home use, which further promotes digital penetration of everyday life. The technological novelty of mentioned applications is that they provide a natural language-driven interface that is not restricted by individual commands but can capture entire sentences and the context of their utterance. The resulting intelligent assistance functions rely in their role on self-learning algorithms (machine learning) and artificial intelligence. One much-discussed aspect concerns the ambivalence between latent availability on the one hand and a possible loss of privacy on the other hand; by activating the assistance systems via voice, the microphones are latently ready to receive the signals and thus they start to monitor all sounds in the area. In addition, as virtual agents are deployed on digital platforms of large technology providers, their assistance functions, such as searching for information on the Internet or making bookings and other commercial transactions, at least call into question the decision-making sovereignty of their users. Similar to the development of visual user interfaces or touch-sensitive touchscreens of smart phones, tablets etc. Voice-computing interfaces also form a technological innovation in the context of human-machine interaction. The critical point for this development is the desire to make operating interfaces as intuitive as possible. Since the usage of natural language as input and output signal enables a virtually unbroken and therefore natural communication with computer programs, voice computing in this context represents a significant progress in development. The current main relevance of this technology is revealed by its usage in virtual assistants that link mobile or stationary devices with AI and deep learning algorithms. At the product level, voice computing and virtual assistants are embedded in two application contexts: While programs such as Siri or Now have already been largely integrated into smart phone usage practices, devices such as Alexa or Home Assistant are still relatively new to the market. As a combination of loudspeakers, microphones, and interfaces that digitally network them, they create a new product category at the intersection of smart home components and the Internet of Things (IoT), which promotes digital penetration of private homes and the home environment networked with global data streams. In order to translate the potential of voice computing and digital assistants into practical applications, interaction processes need to be made even more responsive and systems must be able to communicate with people in a coherent and consistent manner.

2.2.2 Automatic Speech Recognition

The main purpose of introducing the Automatic Speech Recognition (ASR) is to give systems the ability for mapping an acoustic signal to a string of words (sentences). On top of it there is also an approach called Automatic Speech Understanding (ASU) that deals with extracting the semantics and meanings out of the recognized words [1] [14]. There are still open challenges and issues which need to be addressed in this field, however ASR evolved a lot over the passed years. Most known domain using this technology is the Human-Computer Interaction (HCI). Some studies proved that variety of tasks can be solved better with visual or pointing interfaces, but on the other side for tasks that require full natural language communication or for tasks where no keyboard could be used, speech provides the best interface for the usage in such cases. Some of the scenarios for applying the ASR are for example so called hands-busy or eyes-busy applications, i.e., scenarios where the user has objects to manipulate or equipment to control. Another famous application for this technology is telephony (spoken dialogue systems), where the user should pronounce digits or confirm some terms orally by saying "yes". Recently ASR was also plugged for the usage in the dictation field. This domain is mostly used in law and for the interaction between computers and humans with disabilities (for example who lost the ability to type or to speak).

In this context some parameters play a role for making the goal of ASR easily achieved or make it harder to reach. Such a parameter is the number of distinct words one would need to recognize. The impact is that tasks requiring the recognition of two word vocabulary only (for example only yes or no) are considered to be easy. Whereas tasks for large vocabularies like detecting human conversations which require the ability to recognize around 64,000 words are much harder to solve. A further parameter is how fluent, natural or conversational the speech is. So called isolated words, in which words are distinguishable by separating them from each other are much easier to recognize in contrast to continuous speech, where words in this case need to be segmented first. Furthermore in continuous speech there are different difficulty levels for solving tasks, for example human-to-machine speech are much easier to recognize than human-to-human speech. Last parameter I want to point out is the noise. The laboratory research in speech recognition is done with high quality microphones in very calm environment where no other voices than the one to be dictated exist. Any kind of noise makes recognition much harder, for example recognizing a speaker in a calm room is much easier than recognizing it where also other noises are present. This is a very important parameter, which plays a direct role in our case. As mentioned in the motivation, the Alexa skill is developed to be used by technicians at

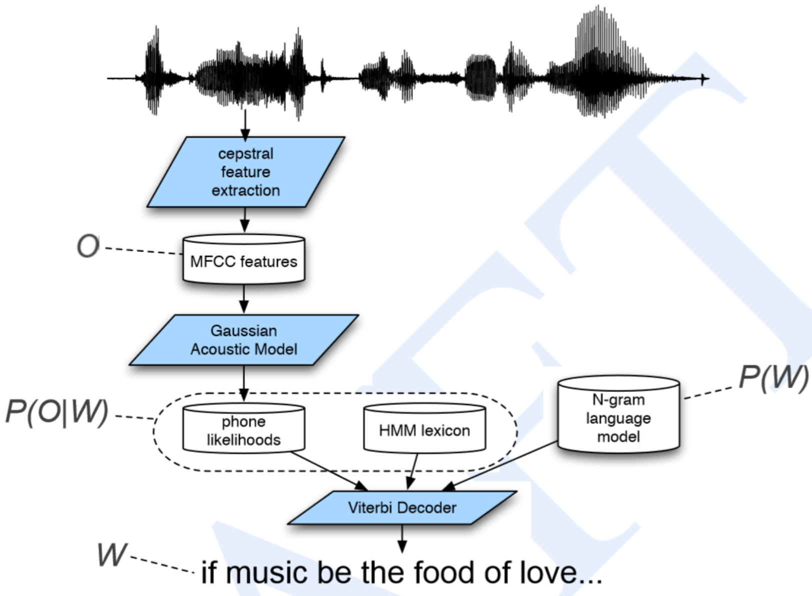

Figure 2: Schematic architecture for a speech recognizer [1] (P. 293).

car workshops and such an environment is very noisy. We will discuss this parameter also later in this thesis.
The architecture of a simplified speech recognizer can be seen in Figure 2. which it shows the needed component for processing a single utterance. The recognition process goes through three main steps: The feature extraction (a.k.a signal processing) step, it takes as input the acoustic waveform and samples it into frames, which are transformed into so called spectral features. Such spectral information and other information about energy and spectral changes are represented by a vector of around 39 features which they represent a time window. Second step is called the acoustic modeling (a.k.a. phone recognition), it computes the likelihood of the observed spectral feature vectors of the previous step given words, phones and subparts of phones (called linguistic units). The output of this step is that for each time frame exists a sequence of probability vectors. Each of these vectors at the corresponding time frame contains the likelihood that each phone or subphone unit generated the acoustic feature vector observation at that time. The last

step is called decoding, in which the sequence of acoustic likelihoods (i.e., the acoustic model AM) is taken with an Hidden Markov Model (HMM) dictionary of word pronunciations and combining them with the language model (LM) then producing as an output the most likely sequence of words which it represent the result for the original acoustic waveform from the beginning. To get an idea of what the HMM dictionary includes, it contains a list of word pronunciations, which are represented by a string of phones. So each word is represented as an HMM, where phones are the states of the model and the HMM output likelihood function for each state is estimated by the Gaussian likelihood estimator. For this step there are multiple known algorithms such as Vitrebi for decoding and speeding up the decoding process by using sophisticated augmentations like pruning, fast-match and tree-structured lexicons.

2.2.3 Dialogue and Conversational Agents

Currently there is a wide range of applications for the usage of computer linguistics. I will introduce in the following the most interesting domain in which it can be applied and which is relevant for the purpose of this thesis. The application is called dialogue and conversational agents [1]. Conversation is the most fundamental and specially privileged area of the language. It is the first way to communicate between each other as humans and first kind of language we learn as children. Also the way of communication between conversational agents (a.k.a. spoken dialogue systems) is speech rather than text. These systems communicate with users in spoken natural language to execute some tasks, for example to answer some questions about weather and sports or to route telephone calls. Let us in the following focus on the basics of dialogue systems and their architecture (see Figure 3). It consists of six components, firstly speech recognition followed by natural language understanding to extract the meaning from the input as a first step. As a further step the components natural language generation and TTS synthesis run a mapping between the meaning and speech. For monitoring and con-

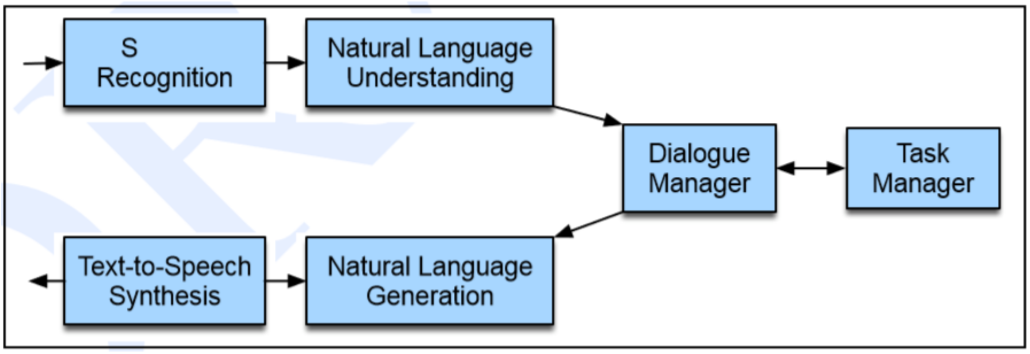

Figure 3: Conversational agent architecture. [1] (P. 835)

trol, the component dialogue manager controls the whole process as well as the component task manager which has knowledge about the task domain. The first component represents the ASR part of the architecture (see previous subsection), which transforms an audio input into a transcribed string of words. For the usage in conversational agents there exist multiple optimization possibilities for the ASR system. The speech recognizer needs to be able to recognize only sentences that can be understood by the natural language understanding component. This is why non-probabilistic language models based on finite-state grammars are mainly used in commercial dialogue systems. These grammars define all possible responses that the system could understand and are generally handwritten. In most of the cases what the user says relates to what the system has just said, this is why language models in conversational agents are developed to be dialogue-state dependent, i.e., they keep the current state of the conversation for optimization possibilities. For example if the system asks the user "What city are you requesting the weather for?", the ASR language model could only expect city names or sentences of the form "what is the weather like in -City Name-?". Secondly, a semantic representation which is appropriate for the dialogue task is produced by the natural language understanding component (NLU). Thirdly, the natural language generation component (NLG) decides on the concepts to express to the user, decides on how to describe those concepts in words, and also in some cases maps prosody to the words. The process of this component can be split into two main tasks, "what to say" and "how to say it". As a fourth step the TTS component takes these words and their prosodic annotations and synthesizes a waveform out of it (see subsection 2.1.1). As a last step comes the role of the dialogue manager component, which controls the architecture and structure of the dialogue. It gets the input from the ASR and NLU components, then obtains a global state of the process by exchanging needed information with the task manager, and lastly delivers the output to the NLG and TTS components.

3 Related Work

One of the core goals mentioned for this work is to survey the usage of vocal assistants in the industry and to make benefit out of its advantages in this field. None of the retrieved papers and approaches tried so far to consider this aspect. However, the following paper surveyed if we could make some everyday life activities easier with the usage of vocal assistants.

3.1 Voice Assistant for Outpatient Care - Concept

Outpatient care services make it possible for (old) people who need regular help in everyday life to stay at home. This care can include advanced medical technology and procedures even when provided outside of hospitals. To learn more about outpatient care, you can find sufficient information in the book of Doris Schäffer [6].
In the following we reflect the goal and results delivered by the authors in [5]. Their main goal is to evaluate whether available voice user interfaces from Amazon Echo and Google Home are suitable to be operated by seniors or not, and to find out which factors must be considered when developing such an interface for seniors and integrating them into the home environment. Their purpose through the paper is to provide a guide, which support for the integration of commercial vocal assistants into households. They additionally

deliver development instructions for voice-controlled applications (such as Alexa skills from Amazon) for the use in outpatient care with seniors as the target group. Interviews and an online survey were conducted to collect the functional and non-functional requirements on vocal assistants for seniors in need of care and for caregivers. The interviews and the online survey revealed the following usage categories for the requirements analysis, which could support the seniors in need of care: Smart home control, memorial work, infotainment, communication and documentation.

3.2 Voice Assistant for Outpatient Care - Results

Seventeen seniors participated in the survey (nine women and eight men) aged between 60 and 95. Those seniors reflected their needs against the five defined categories, like how they would operate the lighting system at home or how they would receive information and news etc. Authors focused then on how the seniors react on the functionality of vocal assistants and if they would like to use them to get their tasks done.

In the results they found out that the needs of seniors in need of care could be met partly under the usage of vocal assistants, which contribute preserve the quality of life. By that, even not professional caregivers could be supported by such vocal assistants in their word. By allowing seniors doing everyday life tasks through vocal assistants by themselves, they give the chance for caregivers to concentrate on other more important tasks related seniors' health. They further concluded that in the case of heavy cognitive restrictions, it can be decided after a test phase whether the use of a vocal assistant system is beneficial. Authors believe that Amazon Echo and Google Home are constantly evolving and offer potential for outpatient care. For that more development and research must be done in this area.

4 Used technologies

Technologies and the example environment we are using in this work will be introduced under this section. I will provide a small comparison between the current technologies and then give some background on the development with them.

4.1 Alexa vs. Siri vs. Home

Voice assistants are software agents that can interpret human speech and respond via synthesized voices. Apple's Siri [8], Amazon's Alexa [9], and Google's Home [10] are the most popular voice assistants and are embedded in smart-phones or dedicated home speakers. Users can ask their assistants questions, control home automation devices and media playback via voice, and manage other basic tasks such as email, to-do lists, and calendars with verbal commands. The software constantly listens for a key word to wake it up. Once it hears that key word, it records the user's voice and sends it to specialized server, which processes and interprets it as a command. Depending on the command, the server will supply the voice assistant with appropriate information to be read back to the user, play the media requested by the user, or complete tasks with various connected services and devices. The number of services that support voice commands is growing rapidly, and Internet-of-Things device manufacturers are also building voice control into their products. Apple's Siri assistant has been around the longest, released as a standalone app in 2010 and bundled into IOS in 2011. Amazon launched Alexa with its Echo-connected home speaker in 2014, and Google's Assistant was announced in 2016 along with its Home speaker and is also embedded in the Google Application for smart-phones based on Android. Each assistant has its own unique features, but the core functions are the same. Siri started as a virtual personal assistant (VPA) application on the iPhone 4S, iPod touch 5, iPad touch 3, iPad Air and all of the iPad minis. It was introduced to the world in 2011. It experienced many developments. In IOS 6 Siri was able to provide information about sports, movies, and restaurants, and was able to open applications. In IOS 7 it was able to have better voices and languages. In IOS 8 it could listen to the wake word "hey Siri". It is available on Apple devices and Apple TV. Siri is good on the go because it can be in your pocket.

Google Home speakers enable users to speak voice commands to interact with services through Google's personal assistant software called Google Assistant. A large number of services, both in-house and third-party, are integrated, allowing users to listen to music, control playback of videos or photos, or receive news updates entirely by voice. Google Home devices also have integrated support for home automation, letting users control smart home appliances with their voice. Multiple Google Home devices can be placed in different rooms in a home for synchronized playback of music. An update in April 2017 brought multi-user support, allowing the device to distinguish between up to six people by voice.

On the other side Alexa is a VPA created and developed by Amazon. It was named after the ancient library Alexandria. It was created in 2014. Alexa is a touch free device that has a speaker. It can play music, read the news, and set alarms. Alexa is able to connect to a phone as well as show directions and estimate traffic times. Also, it can be connected to a smart home system and control for example the lights by voice. Since 2017, over 5000 workers work on Alexa and its programs. To understand how it works, let us consider the Figure 4: It starts when the user says the wake up word "Alexa" then the Echo device starts to listen to what the user says afterwards. Subsequently the Alexa enabled device sends the voice stream as a request to the Amazon Alexa Service which is located in the Amazon Cloud, for that such devices should be connected to the internet (i.e., WIFI). The Alexa Service is the brain behind Alexa devices and represents the voice service that provides capabilities and so called skills to allow customers to interact with devices in a more intuitive way using voice commands. At this step the speech-to-text process is done, then the information requested by user will be gathered from Services and Data Sources (such as Wikipedia, weather services, etc.). This information takes the same path but the other way around. Alexa converts it to a voice stream and sends it as a response back to the Alexa Echo, which in its role speak out the requested information to the user. And this scenario is the typical pipeline for processing user's requests and orders.

Figure 4: Alexa Echo processing steps [11].

4.2 Alexa Skills

Alexa provides a set of built-in capabilities, referred to as skills. For example, Alexa's abilities include playing music from multiple providers, answering questions, providing weather forecasts, and querying Wikipedia. Skills are functionalities or capabilities built by third party to extend core skills of Alexa and build new customer experiences. We could think of Skill in Alexa as an application in mobile parlance. Skill logic could be developed in any language and web stack (Java, Node.js, Python, C#, Go, PHP) and it could be hosted in the cloud (AWS, Azure, Google Cloud Platform, Heroku) or also in data centers. For illustrating at which part of the processing architecture custom skills are plugged, we consider Figure 5 in contrast to

Figure 4. Main parts for building Alexa custom skills are Alexa Voice Service (AVS), which is hosted in the Amazon cloud and the Alexa Skill Kit (ASK). AVS is a scalable cloud service that adds voice-enabled experiences to any connected product (like Echo). Furthermore it performs speech to text operations (and vice versa), Natural language processing (NLP, see section 2.1) and other functions based on Artificial Intelligence (AI). The Alexa Voice Service enables us to access cloud-based Alexa capabilities with the support of AVS APIs, hardware kits, software tools, and documentation. It simplifies building voice-forward devices with Alexa built-in by handling complex speech recognition and natural language understanding in the cloud, reducing development costs and time. Whereas the ASK lets one teach Alexa new skills, customers can access these new abilities by asking Alexa questions or making requests. We can build skills that provide users with many different types of abilities. In other words ASK is a collection of self-service APIs, tools, documentation and code samples that make building new Alexa Skills easy and fast. So for building a custom skill we create a cloud-based service that handles the requests for the skill type and host it in the cloud. The Alexa service routes incoming requests to the appropriate service. Different types of skills require different types of services; for a custom skill either we set an AWS Lambda Function or a web service as an endpoint. AWS Lambda (offered by Amazon Web Services) is a service that allows us to run code in the cloud without managing servers. Alexa sends the code user requests and the code can inspect the request, take any necessary actions (such as looking up information online) and then send a response back. It also allows scaling without thinking of servers or nodes. Alternatively, we can write a web service and host it with any cloud hosting provider (AWS, Google Cloud, Microsoft Azure, etc.) and this web service must accept requests over HTTPS. In this case, Alexa sends requests to the web service and the service takes any necessary actions then sends back a response.

In practice when designing and building a custom skill, one need to create a custom interaction model for the skill. This defines the requests the skill

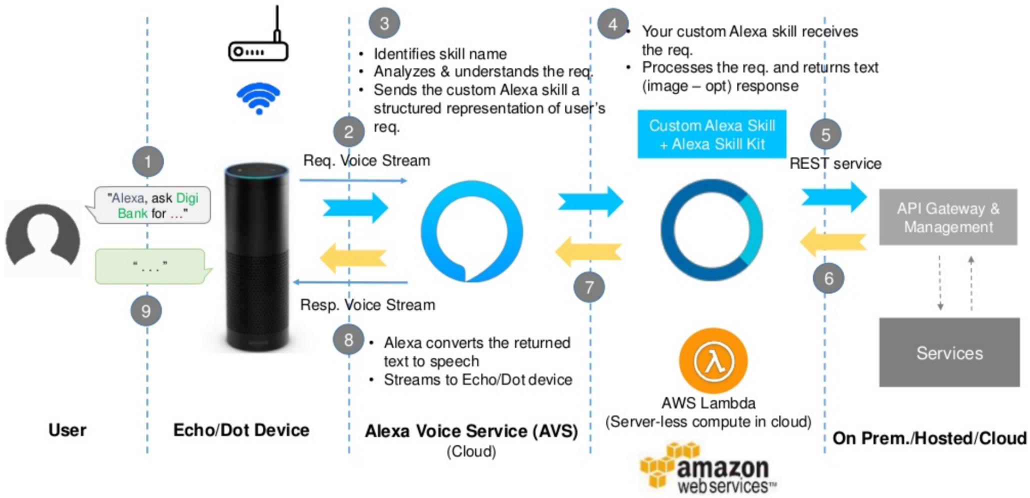

Figure 5: Alexa Echo processing steps of custom skills [12].

can handle and the words users can say to invoke those requests. It consists of the following components:

- A set of intents that represent actions that users can do with the skill. Intents are specific requests, questions or commands. These intents represent the core functionality for the skill. In this context there exist so-called "Built-In Intents", which are defined by Amazon and allow users to engage with different skills in a consistent way (E.g. StopIntent).

- Intents can optionally have arguments called slots. Those slots represent a place-holder for parameters given by users, such as the time to set alarms or a city for weather requests. Slots are defined with different types. Also for slots exist Amazon's Built-In slot types, such as "AMAZON.DATE" type to convert words that indicate dates (example, "today" and "next Friday") into a date format. Additionally one can define custom slot types for possible values. Custom slot types are used for lists of items that are not covered by one of Amazon's built-in slot types.

- A set of sample utterances that specify the words and phrases users can say to invoke those intents. A mapping between these utterances and the defined intents should be done. This mapping forms the interaction model for the

skill.
- An invocation name that identifies the skill. The user includes this name when initiating a conversation with the custom skill.
- A cloud-based service that accepts these intents as structured requests and then acts upon them. This service must be accessible over the Internet as an endpoint for the skill.
- A configuration that brings all of the above together so that Alexa can route requests to the service for the skill.
To put everything together into one scenario, we consider Figure 6 which represents the parsing process for Invocation Name, Intent and Slots. Each statement should start with the wake world "alexa", so the device expect the request. Then we can use words such as "ask" or "tell" followed by the skill specified invocation name to make it clear, to which skill we want to send the request. The rest contains the sample utterance which is mapped to the corresponding Intent followed by the value of the required slot.

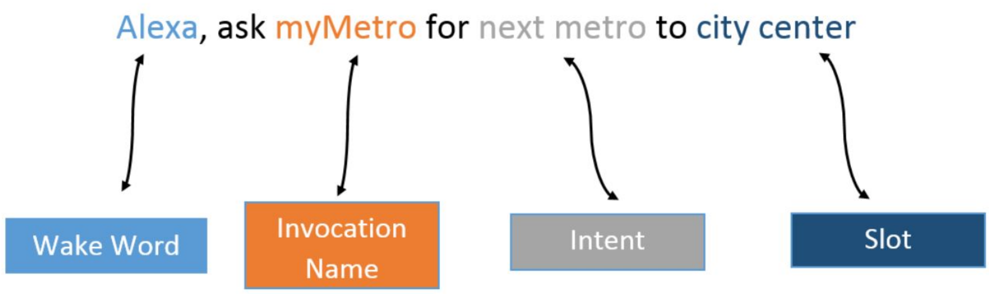

Figure 6: Parsing for Invocation Name, Intent and Slots.

5 Solution approach

For solving the problem we want to survey how far we can make use of vocal assistants in the usage of vehicle diagnostics. In the following we take a look at my solution road-map to the problem introduced in the first section. We want to find alternative ways to support technicians at car workshops to make their work easier in the vehicle diagnostics field. The solution is based on the idea of establishing an environment hosting a vocal assistant which guides the technician (car Mechanic) to diagnose the failure in vehicles, as well as to determine which steps and procedures need to be taken for addressing the detected failure. We consider in the following my method and approach which I went through to reach the objective of this work (see Figure 7). At the beginning I focused on the processes of vehicle diagnostics and how it is operated, in particular I collected information about the steps and procedures needed to run a failure diagnose of a vehicle. As a second step I considered the possible methods which already exist for carrying out the diagnose process. It is mainly based on the practical check up of the vehicle to detect the failure, which is a time and effort consuming solution. Secondly there are currently more efficient alternatives, such as web-based applications and software that can be used to deliver the failure information. Such systems obtain information as input, for example the car model and the error code, then decide on possible failures which could led to the current behavior. Each of these failures is realized as a decision tree with multiple different decisions that need to be made so the system can determine the problem causing this case. Those decisions are made by the user also as input to the system in order to process the decision tree (see next section). One main practical limitation of this approach is that the technician needs to interact with the system through a touch-screen or by using a keyboard to insert the input to the system. In such cases the technician would like to keep his hand free (from pressing and typing) and use them for checking the car. Furthermore in car workshops and while working on cars, hands of the worker get dirty and in this case he would not use them to interact physically with any system. This drives us to the third step for figuring a solution that address this limitation and deliver better usability and interaction model. As a possible improving solution comes at this point the employment of voice assistant for solving the introduced problem. Such technology has multiple advantages for increasing the usability, especially for the concrete defined use-case. After I defined the possible solution, as a next step I realized it practically by implementing it with the objective of providing a proof-of-concept prototype that reflects the main principal and functionality of the final product. This prototype should also examine the usage of such technologies in the industry and not only for private needs. As already mentioned the solution is based on the environment of Amazon Alexa and so called Alexa skills. As a last step an overall evaluation is conducted to assess the performance of the resulted product, focused on some parameters like usability, complexity, accuracy and error rate.

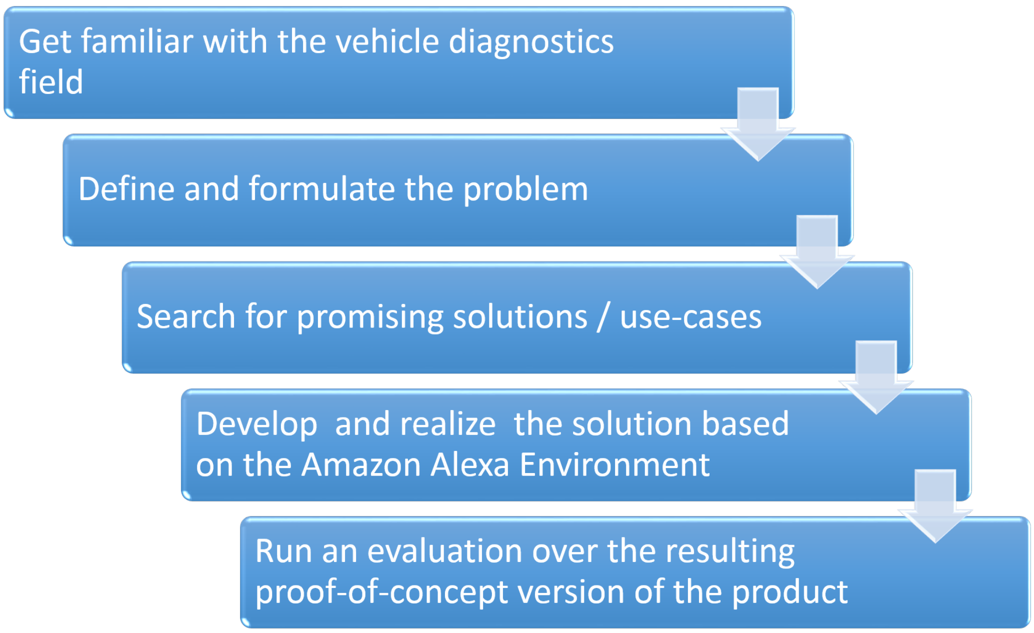

Figure 7: A diagram showing my approach in steps for my research.

6 Implementation

In this section we focus on the implementation done to achieve the goal of this work, i.e., we consider the Alexa Skill developed as a practical solution to the problem described at the beginning. As part of this Thesis an Alexa Skill called "Porsche Diagnostics" has been developed and implemented.

6.1 Decision Trees

Before diving into details of the skill components, it is worth at this step to mention some details about the process of error detection and diagnose of vehicles as a completion of the previous section. The main process is realized through a decision tree. Such a tree can be seen in Figure 8 which consists of nodes and edges. Nodes are lettered with their ID as a number. We distinguish between binary decisions, where the node has exactly two child nodes and multiple decisions, where nodes could have more than two child nodes and a value is expected to be given by user. As an example for this in the given Tree, nodes with ID's 19, 20 and 21 represent binary decisions, whereas node 16 expect an answer from the user for a specific color value and depending on the value we iterate to the corresponding child node. The same is the case for nodes 17 and 18, as a number value is expected and depending on in which interval the value lies we move on to the corresponding child node. Another aspect shown in the given decision tree is illustrated in node 20 as we simulate the case in which Alexa requires information from some other sources (like measuring oil pressure in the vehicle) and depending on the delivered value from the source the skill navigates to the corresponding child node. After taking the needed decisions we land on one of the page nodes to complete the diagnose process and specify the occurred error.
For processing the decision tree, replies from the user are needed constantly at each node to process further to the matching child-node. This requires some kind of an interaction model between the system and the user, for

which realizing the solution as a dialogue system in the environment of Alexa, concludes to the importance of the approach introduced through this thesis. It represents the best case for the usage of vocal assistants in such interactive environments.

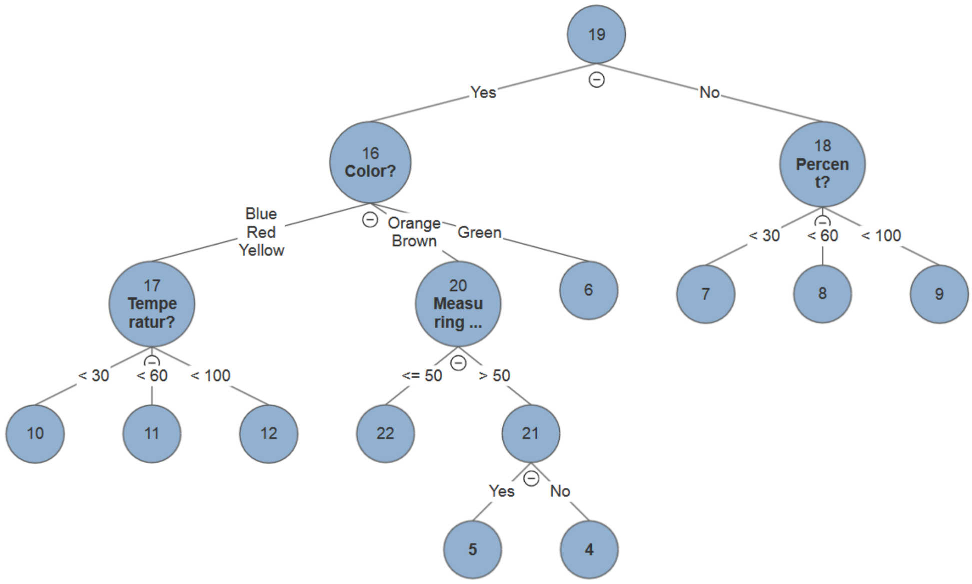

Figure 8: Example decision tree for error detecting.

6.2 Sample dialogue with the skill

In the following we get first impressions on how the skill works and how to interact with it. We start the skill by speaking the skill invocation name to initiate the dialogue (the given letters will be referenced to in the next section):

-User: start porsche diagnostics. (A)

-Alexa: welcome to the Porsche Assistant! I will guide you to get the problem fixed. When ready, say start

-User: start. (B)

At this step the skill is started and Alexa will ask for some information from the user. The model is designed in such a way that the user needs to define

the Porsche Model and the Error Code. Each combination of them is mapped to a subset of Decision Trees (errors) that could be processed to diagnose the error in the vehicle. So we continue the scenario as the following possible way:
-Alexa: What kind of Porsche is it? (C)
-User: Macan.
-Alexa: what is the error code? (C)
-User: 789.
-Alexa: OK, so Porsche Model is macan and Error Code is 789. Is that correct?
-User: correct
Up to this point the skill collected first needed information to decide on possible decision trees. Each decision tree is nothing else but a possible error which produces the failure in the vehicle. Alexa at this step tells the user about found -possible- errors which correspond to the give data.
-Alexa: Okay processing to the error overview I have found 2 errors: first one: break failure second one: turbo error. please choose one to proceed. (D)
-User: Second one. (E)
-Alexa: Is Oil Temprature above forty degree? (F)
-User: yes it is.
-Alexa: Which color is the shown error sign? (F)
-User: blue.
-Alexa: What is the oil Temperature? (F)
-User: ......
-Alexa: ...
In this state the decision tree is now specified and ready to be processed by making decisions on each node as shown in the example scenario. Alexa asks for values and the user answers with the actual state until all decisions are made and depending on which page node is reached in the tree, the possible failure is defined with respect to it. In the upcoming step after specifying the failure, Alexa can tell the user what procedures need to be taken in order to

solve the problem:
-Alexa: Okay decision tree is processed and I have detected the error information. You could recall error overview or I can tell you in the following which actions need to be taken. OK?
-User: OK.
-Alexa: As a first step, Move the spark plugs from the old cap to the new cap. (G)
-User: I did that.
-Alexa: Then, you will need to set the distributor caps down on a flat surface. (G)
-User: finished.
-Alexa: ...
With that the skill delivers all steps needed to be taken for solving the problem, depending on the specified error resulted from the decision tree. By that we reach the last state in the skill after all needed information is answered back to the user:
-Alexa: The session for the macan with the error code 789 was successfully done. Any further requests?
-User: no thanks.
-Alexa: Okay, session is closed. To start new session say restart. Or you may close the skill now.
-User: close the skill. (H)
-Alexa: Closing the skill ... See you next time! Good Bye!
What we have seen is a sample scenario for the interaction model between the skill and the user. We saw how information is exchanged between both sides and how the goal is achieved by using the developed skill.

6.3 Skill components

In the following we consider some technical details behind this functionality. As described in Section 4, the development of the skill is based on one side on the Alexa Skill Kit (ASK) and on the other side on building the end-point for handling the incoming requests and replying with a corresponding response. We discuss those two sides in the following and there are markers added which map each of the following technical information with the corresponding step in the scenario described above. Please consider Figure 9 which shows the whole image of the defined intents and the interaction between them. We recall that for activating a skill we need to set the invocation name, which is in our case "Porsche Diagnostics" (referenced with 'A' in the previous scenario)(in most of the cases it is set to be the same as the skill name). For reaching the intended behavior we need as a next step to define required intents and corresponding slots.

6.3.1 Defined Intents

The first defined Intent is called "InteractionIntents", which is responsible to catch the request after starting a new session with the skill. Some of its Sample Utterances are "Start", "Restart" and "Start a session" (referenced with 'B'). So if the user speak one of those samples, incoming requests will be handled by the InteractionIntents. For this intent two slots have been defined for storing user inputs, namely the Porsche Model and Error Code. For the functionality of those slots, we recall the information from section 2.2.3, which concludes that conversational agents are developed to be dialogue-state dependent. That is, the application expects at this point the user to fulfill the values of those slots (referenced with 'C'). For the first slot I have defined the list "porscheCars" as custom Slot Type, whereas for the later one I used the Amazon built-in Slot Type "Amazon.Number". The skill in this case compares the user input against those two types to notify any wrong unacceptable values (for example giving car model of other brands). Second

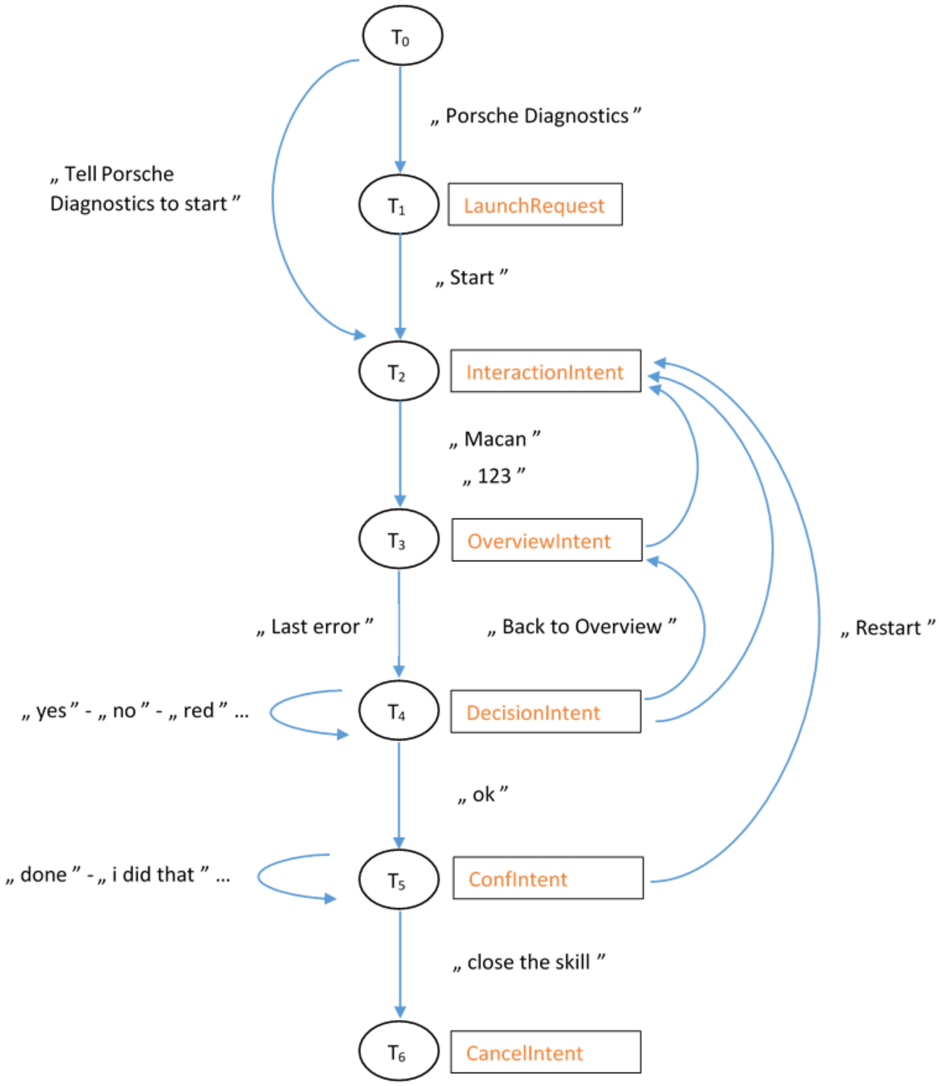

Figure 9: Flow Diagram of the interaction between the components.

intent defined is called "OverviewIntent", which is responsible for representing the found possible errors (referenced with 'D'). It also get triggered when the user answers with "back to error overview". A slot called errorID from the custom slot type "errorIDType" is defined for this intent, which represents the user input on choosing the intended error. Please distinguish between the two slots Error Code and errorID; The first one refers to the error code combined with the Porsche model, whereas the later one refers to the multiple possible errors, which could occur for the defined Porsche model and error code. Third defined intent is "DecisionIntent", which receives requests after the user chooses one of the found errors (referenced with 'E') and it's responsible for processing the specified decision tree and moderating the dialogue between the skill and the user for making decisions. Multiple types of answers are expected from the user, e.g. there are only binary subtrees with only yes-no answers, there are questions which need to be answered by users like giving a color, as well as giving values on temperature or percentage number (referenced with 'F'). Those possible answers are mapped to corresponding slots in the Alexa Skill Kit, e.g. when colors are expected from user then the value is mapped to the slot from type "Amazon.Color". As soon as the decision tree is processed, we reach the intent "ConfIntent", in which Alexa delivers the steps required for solving the diagnosed failure (referenced with 'G'). Lastly we land in the "CancelIntent" after calling the command for closing the skill (referenced with 'H'). By that terminates the flow through the components of the skill.

6.3.2 The Endpoint

Moving to the other side of the skill, namely the endpoint which is implemented based on Amazon Lambda Functions. I used for implementing those functions the programming language "Node.js.8.10", where js. stands for JavaScript. Each of the intents defined in the ASK, triggers a function in the end-point by sending a request to it and getting responses to deliver them back to the user. For that, functions must have same name as the intent name triggered, e.g. we have in the end-point 'InteractionIntents': function () {...}. Those Lambda Functions are also called handlers and are stored in a handlers-collection. There are some extensions to JavaScript which support Alexa orders, such as: this.emit(':ask', speechOutput). This construct interprets that Alexa should ask the user what the string variable speechOutput includes and expect the user to answer on that. Depending on the answer and by comparing it with defined Sample Utterances, the corresponding intent will further trigger the suitable lambda function in the end-point and so on. Furthermore for reading user input and map it to the slot value we use: model = this.event.request.intent.slots.porscheModel.value, where porscheModel the slot name is.

6.3.3 Technical Details

Considering decision trees in the end-point, we can see an illustration in Figure 9 which codes the same tree from Figure 10. It is coded into some kind of Hash-Map Format, where Node ID represent the key in the collection and the object represent the corresponding child nodes as array of ID's. Therefore we can also indicate page nodes by having empty arrays of child nodes. In the example database each node is mapped to a specific decision, e.g. the node with ID 7 contains the decision "do you see any Security Lights on?" which will be asked to the user by the skill when iterating through the tree.

As already mentioned, each combination of Porsche Model with an Error Code is mapped to multiple different errors. This is realized in the end-point as an array including found errors. This has the advantage of providing more flexibility, as the number of found errors equals the length of this array and allows accessing the chosen error directly by the index. Furthermore it improves usability to answer requests as "last error" by simply taking last element in the array.

For the special case, in which what the user says doesn't correspond to any of the defined intents, we can include to the handlers-collection a function with the keyword "Unhandled". This handler is some kind of a generic han-

```
let dicisionTree0 = {
     0: [19],
    19: [16, 18],
    16: [17, 20, 6],
    18: [7, 8, 9],
    17: [10,11,12],
    20: [22, 21],
    6: [2, 3],
    7: [],
    8: [],
    9: [],
    10: [],
    11: [],
    12: [],
    22: [],
    21: [5, 4],
    2: [],
    3: [],
    5: [],
    4: []
},
```

Figure 10: Decision tree coded in the end-point.

dler which receives all undefined cases to produce a proper feedback to the user instead of resulting in an error and the skill could crash at some point without informing the user about the case.

7 Evaluation

After the skill has been developed, we want to run an evaluation over its performance to examine it under different criteria. For this purpose I conducted a study, in order to test the skill and its functionality.

7.1 Concept

The focus group for the study is in the first place my work colleagues at MHP - A Porsche Company. They represent best testers of the skill since they work in the same environment and are to some limit close to the vehicle diagnosis field. The main focus of the study is on one side to evaluate the performance of the developed skill and on the other side to examine how far we can employ Alexa in the environment of vehicle diagnostics. Main keywords I have considered are usability, accuracy, reaction time, error rate and efficiency. The study took averagely around 15 minutes for each participant. At the beginning I introduced the main idea of my thesis and the task of the developed skill for the vehicle diagnosis. Then I gave some general information about Porsche models and binary trees used for the diagnosis process. I prepared a Google form [13] for collecting participants' answers and feedback. At the beginning they needed to answer two introduction questions (can be seen in Figure 11). First question indicates if the participant has any Alexa product or if he has already used it. Second question reflects how deep is the participant's knowledge in the vehicle diagnosis field, where he can define his knowledge according to a linear scalar (between strong knowledge and no idea at all).

For running the interaction with Alexa I gave each user some tasks to accomplish using the skill. The main scenario is that we imagined the participant to be the technician at the Porsche Workshop and he needs to run a vehicle diagnosis using the developed skill. Samples of the tasks can be seen in Figure 12. For this scenario we simulate the case of a Porsche Macan with error

Figure 11: Google form of the first two questions.

code 123. The expected interaction looks as follows:
- Alexa: Welcome to the Porsche Assistant! I will guide you to get the problem fixed. When ready, say start.
- User: start.
- Alexa: What is the model of the Porsche?
- User: Macan.
- Alexa: Tell me the Error code.
- User: 123.
- Alexa: Okay processing to the error overview I have found 4 errors: first one: fog lights damages second one: oil filter failure third one: break failure fourth one: sensor damages. please choose one to proceed.
- User: sensor damages.
- Alexa: OK processing the fourth error, Is Oil Temperature above forty degree?
- User: yes it is.
- Alexa: Which color is the shown sign?
- User: Red.
- Alexa: What is the Motor Temperature?
- User: 25 degrees.
- Alexa: Is the Fog Lamp Indicator turned on?
- User: No it's not.
- Alexa: Okay decision tree is processed and I have detected the error information. You could recall error overview or I can tell you in the following which actions need to be taken. okay?
- User: ...
- Alexa: ...
- User: Restart session.

At the end new session will be started to process another Porsche model and error code to go through different decisions and test other functionalities of the skill.

After running both use-cases the participant gets a good impression about

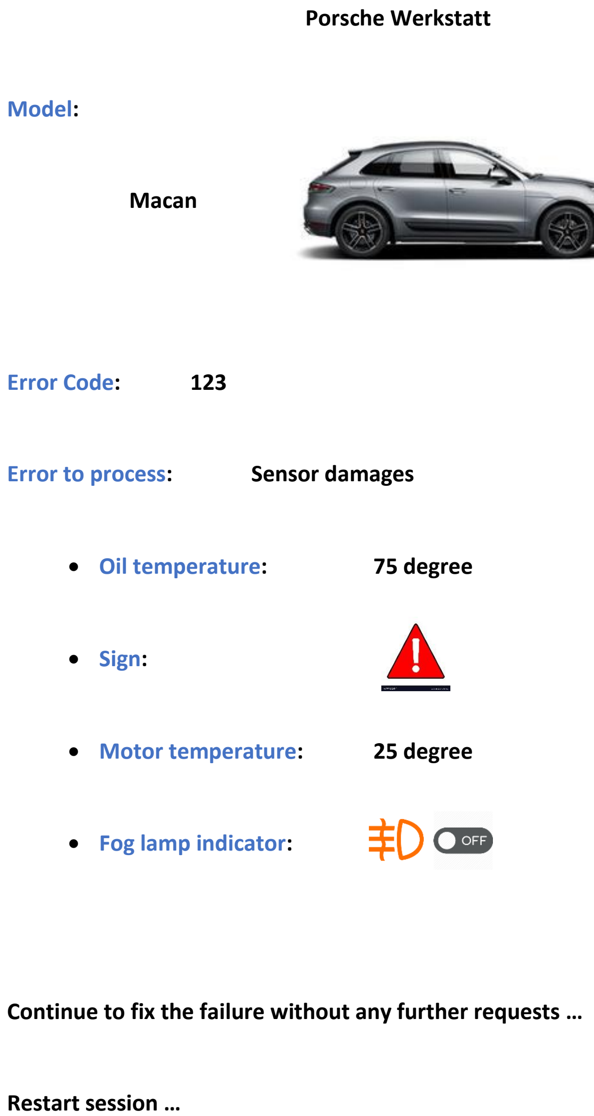

Figure 12: Sample tasks for running the vehicle diagnosis using the developed Alexa skill.

Figure 13: The remaining questions after using the skill.

the skill and its functionality. At this point he is ready to fill out the second section of the Google form. A part of the questions can be seen in Figure 13. First question evaluates how natural was the interaction with Alexa, i.e., how much the participants had the impression that they are communicating and having a conversation with a real human being and not with a device controlled by AI. The next question indicates the reaction time after the participant asks the skill or answers it back with some values. Concretely it means the time needed for Alexa to collect the response, process it and replies back to the user. The next question allows the participant to define how good was Alexa in recognizing his orders and requests. For example, did Alexa interpret exactly what he said and did Alexa receive the right value of the different parameters. The questions so far handle the behavior of Alexa

and its performance. The next questions consider the usage of such a technology in the vehicle diagnosis field. So the forth question is: "How beneficial do you find it for the vehicle diagnosis that Alexa is operated hands-free?". This question points out the importance that Alexa is used hands-free so that the technician can use his hands to interact with the vehicle as he needs to check out some data or adjust some parts. The following question asks the participants about their opinion, if they would as technician use a hands-free virtual assistant to perform the vehicle diagnosis process. This question also indicates how promising is the usage of virtual assistants in such a field. As we can see in Figure 13, the answers to these questions are given as a linear scalar (from 1 to 4) 1 stands for positive, 2 rather positive, 3 rather negative and 4 negative.
After that follows a multiple choice question, namely, "Would you rather use the typical diagnostic system with touch screen/keyboard instead?". Here we compare to some limit the existed approach by using the physical interaction with the vocal interaction approach provided by the developed Alexa skill. There are three possible questions to choose from, "yes" to indicate that the typical approach is better, "no" to indicate the other way around case, or "Both systems are equivalent" to express that both systems have its own advantages and we can't choose one of them over the other.
Subsequently in the form there are two supplementary questions. First one asks for the personal opinion about the usage of virtual assistants in vehicle diagnostics, i.e., if they find it very promising (or the opposite), what they liked most about the skill, or if they have any additional suggestions. The last question could be answered by the participant is a comparison between the typical approach with physical interaction and the suggested approach with vocal interaction. When the participant is finished with filling out the form, the end of the study is reached.

7.2 Results

For the concept described in the previous section, 15 participants took a part in the study. So the results introduced in this section depend on their evaluation of the skill.
First question indicated how many of the participants already used Alexa. Results show that 12 out of the 15 participants (80%) already used it and 3 of them did not have any interaction with Alexa before. So most of the participants already had an interaction with Alexa and they are familiar with its services. Furthermore about 73% of the participants already have good knowledge in the vehicle diagnosis and have already worked for projects in this field. The rest are divided as follows: 1 has sufficient knowledge in the topic, 2 have heard about it and 1 participant had no idea about this field. These are the results for the first two introduction questions that reflect the knowledge of the participants in Alexa and vehicle diagnosis.

The following answers are for the questions answered after using the developed skill (a statistical visualization of the participants' feedback can be seen in Figure 14). The participants were split between finding the interaction with Alexa very natural and finding it relatively natural (approximately 50% each), i.e, through the study they had the feeling that they are communicating with a real person. This result was expected since the skill provides different tones and expressions on replying the user, which gives them the expression that a real person is answering their requests not an artificial virtual assistant. The results show further that Alexa's main strength was the reaction time. Concretely 14 out of the 15 participants found the response time to their requests very fast, whereas the remaining one chose it to be relatively fast. One of the strength points of the cloud services provided by Amazon is the strong processability. This is the reason behind the fast request processing by Alexa. Third question evaluates how good did Alexa recognize the orders. To some limits this point seems to be the weakest tested ability of Alexa through the conducted study. Participants feedback varies for this question the most and are divided as follows: 5 participants (about 33%) found that

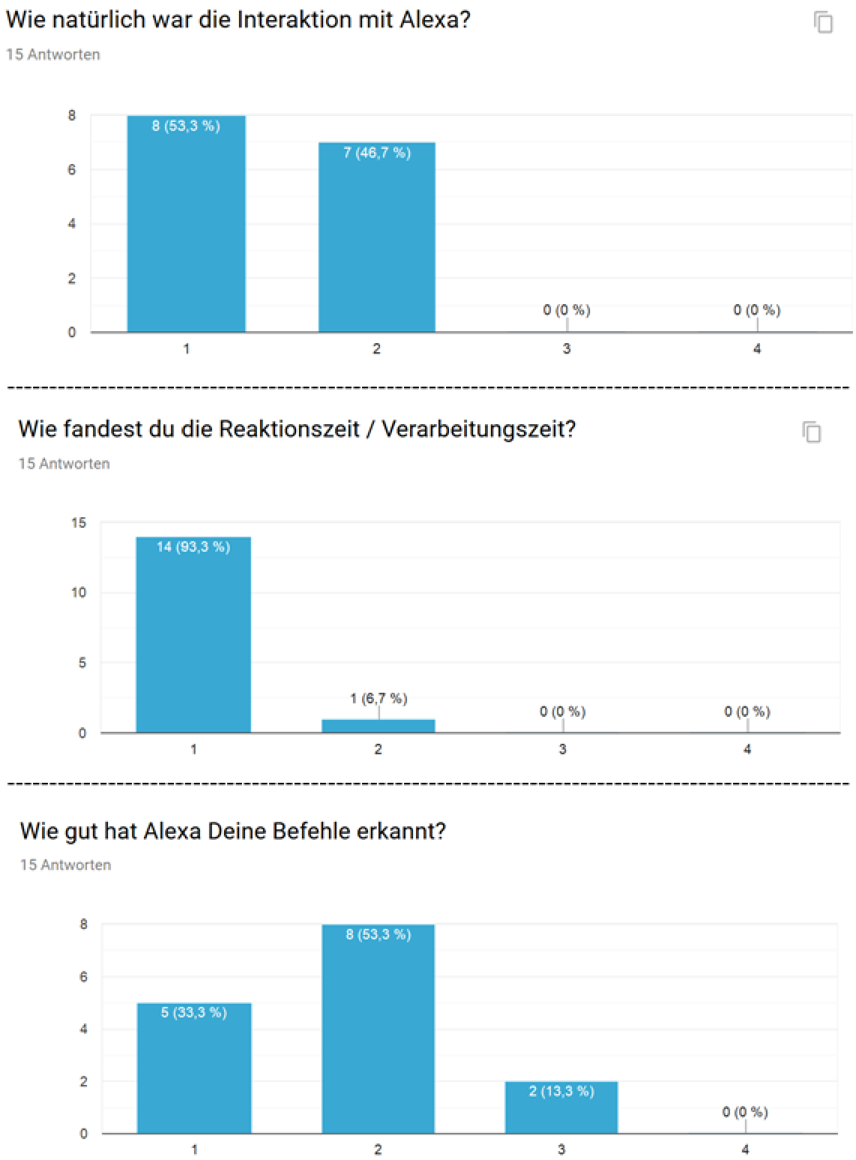

Figure 14: Participant feedback considering Alexa performance.

Alexa understood their requests very good, 8 participants (over 50%) found it rather good and the remaining 2 found it rather bad. Those numbers reflect that there were some cases where Alexa misunderstood what the user said or where Alexa did not recognize the user's request. One possible explanation for why most of the participants in this case did not decide for the best performance in recognizing their requests, is the usage of a development test environment, in which participants had to use a headset to interact with the test environment over the web browser. That is, the study was not conducted on a real Alexa device (for example, Alexa Echo) and this produced some limitations on the command recognizability. In another words, the usage of a real Alexa device would have delivered better results for this question. Most important criteria for that is the quality of the regular Microphone used in the user study in contrast to the high quality built-in Microphones on the Amazon Echo products. I preferred using the test environment over a real Alexa device, because with the test environment all user's requests and Alexa's responses get documented in a written format. With that, I can track back the whole conversation to find out if anything went wrong through the study. By the evaluation I found out that some words were misunderstood, e.g. the user said "finished" and Alexa received it as "furnished". An additional aspect for this point is that all of the participants in the user study are not English native speakers. To some limits, not pronouncing the words correctly also led to a bad evaluation for this question. That is, if the skill was developed in German, the study would have delivered better evaluation for this question.

This represented the part for evaluating the Alexa performance. In the following we take a look at the feedback considering the usage of vocal assistants in the field of vehicle diagnosis (see Figure 15). Next question indicates from the participant, how positive is the hands-free usage of Alexa in the vehicle diagnosis at workshops. As we can see on the bar chart, 80% considered this usage very beneficial, whereas the rest 20% found it relatively positive. The feedback of the following question delivers barely same results. All 15 partic-

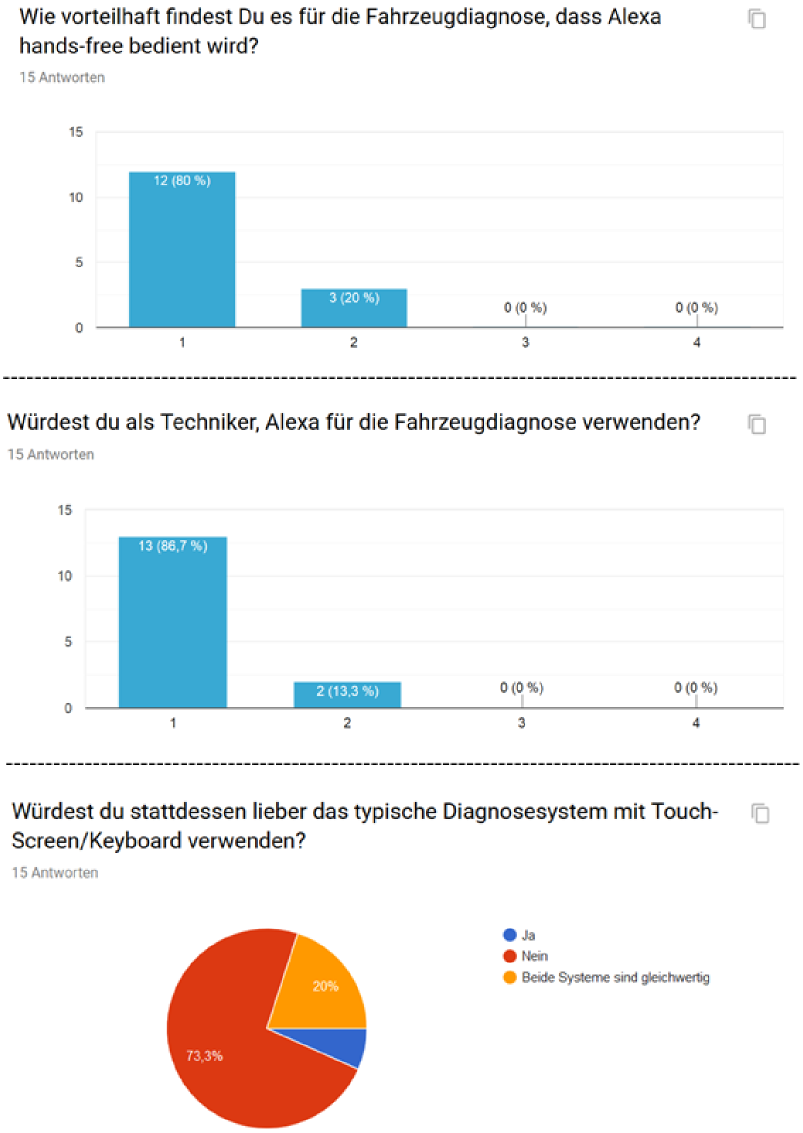

Figure 15: Participant feedback considering the usage of Alexa in the field of vehicle diagnosis.

ipants would -as technicians at workshops- use Alexa for running the vehicle diagnosis process. The last question handles a comparison between the typical diagnosis system with touch-screen and the Alexa skill. The pie chart illustrates that around 73% of the participants would not use the typical diagnosis system with the physical interaction over Alexa, whereas exactly 20% found both systems equally relevant for the vehicle diagnosis process. We can conclude that all participants of the study find the usage of vocal assistants in the field of vehicle diagnosis is very beneficial, especially that it is operated hands-free.

Last part of the feedback sheet consists of two questions requesting ideas and opinions of the participants. Firstly considering the usage o such an Alexa skill in the vehicle diagnosis process and secondly considering a comparison against the typical diagnosis system with the touch-screen. The participants could answer those questions with short sentences. In the following I will review most important ideas mentioned for this section. Participants found in general that utilizing vocal assistants in such a field is very beneficial mostly because it is operated hands-free. This feature is very relevant since technician at workshops need their hands for checking cars through the diagnosis process and because technician would avoid any physical contacts due to their unclean hands. By adopting this technology those problems are addressed, since all the technician needs to do is talking through his headset. Further positive aspect was mentioned by most of the participant is the improved usability. They argued that vocal interaction with the system is very easy and operating the diagnosis process through a conversation with the developed Alexa skill clearly improves the usability. Other advantages mentioned are the fast reaction time, naturality of the conversation, step for step queries and feedback impressions of Alexa. However on the other side they also mentioned some challenges that could occur. For example, to adopt such a new technology in the field of vehicle diagnosis, technicians should be trained well to use the developed skill and to know how this process works vocally. The participants argued further that they are not sure if technicians in this case would really want to learn the new system instead of what they are already used to and been using so far. In addition a very important limitation is also mentioned, namely that by using vocal assistants we give up on images and the visual feedback, i.e., there is no screen anymore to display any content that could help more in some situations. Additional to this aspect, faster operation and less error rate are mentioned as advantages for the typical diagnosis system in contrast to the developed skill. Furthermore we recall the noise parameter introduced in section 2.2.2. Noises at workshops could also produce a limitation on the Alexa skill functionality, particularly they could affect the voice recognition negatively. Unfortunately we didn't have the chance to test the skill at workshops to measure the size of this effect on the performance of the skill, but by using the headset it shouldn't reduce the quality of the voice recognition process, especially that Alexa is very powerful when it comes to voice recognition in noisy environments. In this context, Maryam et al. [14] surveyed the performance of virtual personal assistants in noisy environments by testing a developed Alexa skill with a noise simulator for the purpose to learn robust dialogue policies and to improve speech recognition in noisy fields.

8 Conclusion and future work

As we have seen the previous section, the study showed us that working with vocal assistants in the vehicle diagnosis field is very promising and this approach solves multiple problems considering usability, efficiency and performance. The main purpose of the thesis is to provide the technician at workshops new means for running the diagnosis process without any physical interaction with any device. At the same time to improve the usability and make the technician's work easier by using the hands-free headset for the vocal interaction with the system. Concretely and by having the headset, the technician doesn't need to swing a lot between the car and the system for delivering the feedback, rather he can simply check the car and deliver the input vocally without the need to go back to the touch-screen system. Additionally we have seen that there is also a dark side for this approach. Most important criticism is the visual limitation, as under the usage of vocal assistants we don't have the ability to display images or any additional information that could help the technician through the diagnosis process. This point is actually the advantage of the typical diagnosis system with the touch-screen over the approach introduced through this paper. This bring us to the discussion on how to address those limitations and which possible approaches could be developed in the future to improve our introduced work. There exist two actual approaches to further build on top of what I have developed through this thesis.

First approach is to use the so called Echo Show [15] [16]. Echo Show is an Alexa-enabled device with a screen developed by Amazon which also supports displaying content and images. Those devices such as Echo Show, Echo Spot, Fire TV Cube, Fire HD 8, and Fire HD10 allow skill developers to create skills that use both screen and voice interaction. So as we can notice by developing the skill with Echo Show instead of the Echo Dot, we can additionally to the voice commands also display useful content on the screen of the device. To mention here is that the interaction with the Echo Show is moreover vocally without the need to have any physical interaction with the device. A representation of the Echo Show device can be seen in Figure 16. With this device one could develop the skill to also represent helpful content and images on its screen for each step in the vehicle diagnosis process and by that addresses the limitation of not being able to display any visual content. If we evaluate the usage of the Echo Show against the Echo Dot and the typical diagnosis system we find out that it addresses the physical interaction problem of the typical system and the limitation on displaying visual content of the Echo Dot. However on the other side the placement of the device could be challenging. Since we want to avoid any physical interaction, the technician shouldn't carry it in his hands and by that it should be placed somewhere around him. This could require the technician to frequently go

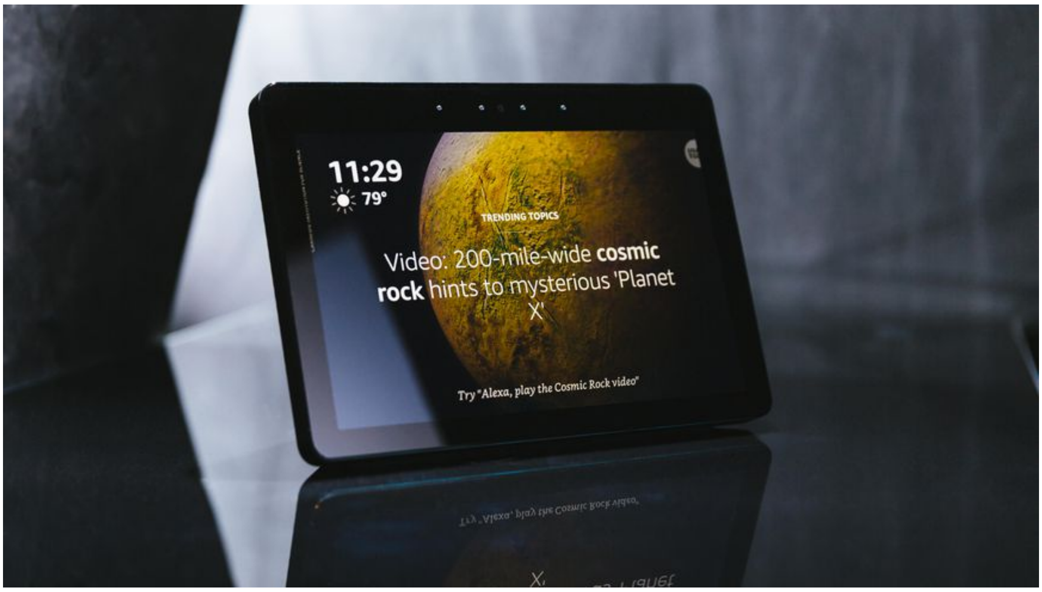

Figure 16: Image showing the Echo Show device from Amazon [16].

back and take a look at the screen of the device to receive the visual content. This drawback of the usage of Echo Show brings us to the most promising approach that solves all problems mentioned in this work. This approach is also mentioned by some participants of the study in the previous section. Instead of using the screen-enabled Echo Show, we can combine the Alexa skill with the Augmented Reality (AR) technology [17]. By that we get rid of the physical device with the screen and replace it with AR glasses that the technician can wear [18]. Such Augmented Reality glasses can be seen in Figure 17. This technology addresses clearly the challenge of having a screen-enabled device located somewhere by wearing glasses instead. The expected behavior is then that the technician delivers input vocally to the Alexa skill and by the reply he additionally gets visual feedback through the AR glasses, such as highlighting car parts with different color which need to be checked or other similar visual effects. With this combination of Alexa and AR we address the visual limitation of relying only on the Alexa skill and at the same time we get rid of any screen-enabled devices. But this approach also have its challenges. The biggest challenge is the development complexity. To

Figure 17: Image showing the augmented reality glasses from Google [18].

achieve and develop this approach we need to integrate two different environments to work together, the amazon Alexa skill on one side and Augmented Reality on the other side, which is a big challenge.
Those are two possible approaches which could bring more advantages of the usage of vocal assistants to run the vehicle diagnosis process. We mentioned at the beginning that the development of technology is running relatively fast, especially in the last decade. That's why I believe that in the nearest future, the usage of vocal assistants in the vehicle diagnosis field will be more researched and more promising solutions will be introduced.

References

[1] Jurafsky, Dan, and James H. Martin. "Speech and language processing." Vol. 3. London: Pearson, 2014.

[2] Ladd, D. Robert. "Intonational phrasing: the case for recursive prosodic structure." Phonology 3 (1986): 311-340.

[3] Liddy, Elizabeth D. "Natural language processing." In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc. (2001).

[4] Dennis, Tom. "Information Visualization and Visual Analytics." University of Stuttgart (2018).

[5] Hellwig, Andre, et al. "Sprachassistenten in der ambulanten Pflege." Zukunft der Pflege Tagungsband der 1. Clusterkonferenz 2018: 72.

[6] Schaeffer, Doris. "Grenzen ambulanter Pflege" Veröffentlichungsreihe der Forschungsgruppe Gesundheitsrisiken ISSN-0935-8137. 1992.

[7] Ferdinand, Jan-Peter, and Tobias Jetzke. "Voice Computing–allgegenwärtige Spracherkennung." Retrieved 16 (2017): 2017.

[8] Apple Inc. "SiriKit". Link: "https://developer.apple.com/documentation/sirikit" (visited on 15.05.2019).

[9] Amazon.com, Inc. "Build Skills with the Alexa Skills Kit". Link: "https://developer.amazon.com/docs/ask-overviews/build-skills-with-the-alexa-skills-kit.html" (visited on 10.03.2019).

[10] Google. "Actions on Google". Link: "https://developers.google.com/actions/" (visited on 15.05.2019).

[11] Smith, Sam. "How does Amazon Echo work?". Link "https://www.quora.com/How-does-Amazon-Echo-work" (visited on 15.05.2019), 2018.

[12] Rajaraman, Anand. "Building Custom Alexa Skill". link: "https://erpinnews.com/building-custom-alexa-skill" (visited on 15.05.2019), 2019.

[13] Google LLC. "Google Docs, Sheets, Slides und Forms". link: "https://www.google.com/intl/de_de/docs/about/" (visited on 15.05.2019).

[14] Fazel-Zarandi, Maryam, et al. "Learning robust dialog policies in noisy environments." arXiv preprint arXiv:1712.04034 (2017).

[15] Amazon.com, Inc. "Amazon Echo Show". link: "https://www.amazon.de/Amazon-Premiumlautsprecher-brillantem-10-Zoll-HD-Display-Schwarz-Alexa/dp/B0793C38JY" (visited on 15.05.2019).

[16] CNet. "Amazon Echo Show (2nd Gen) review". link: "https://www.cnet.com/reviews/amazon-echo-show-second-generation-review/" (visited on 15.05.2019).

[17] Azuma, Ronald T. "A survey of augmented reality." Presence: Teleoperators & Virtual Environments 6.4 (1997): 355-385.

[18] Muensterer, Oliver J., et al. "Google Glass in pediatric surgery: an exploratory study." International journal of surgery 12.4 (2014): 281-289.