Designing a Trustworthy Voice User Interface for Payments and Transactions

A study in user experience design

Jessica Lundqvist
Abstract

Talking to computers originated as a science fiction element but is now becoming a reality with the developments of Intelligent Personal Assistants from big companies such as Google, Apple, Microsoft and Amazon. This thesis aims to investigate the design aspects to get users to feel trust in voice user interfaces with the use case of performing an invoice payment. To be able to design for trust in a Voice User Interface for banks it is of utmost importance to first understand the fundamentals of a conversation between humans. Further, it is necessary to follow already existing guidelines.

A Service Design methodology was implemented with a research phase consisting of a literature study and a workshop to form an understanding from the users point of view of the problem. A mockup was made to compare two types of conversations to interact with the voice assistant. The literature study, workshop and mockup tests later resulted in a prototype created in Dialogflow, Google’s development tool for voice user interfaces. The prototype test was evaluated with interviews and a System Usability Score questionnaire. The overall opinion and grade the prototype received was that it was easy and fun to use and could add a value for when users have their hands or eyes occupied with something else at the same time. Based on the results from the workshop, mockup interviews and feedback from the users testing the prototype, design guidelines for increasing the feeling of trust were compiled. The guidelines point out the importance of a personalized application, to provide a welcoming phrase, describing instructions so the user knows what is expected from them, and for the user to be able to use catchphrases or short sentences when talking to the system. The system should mimic the way the users provides their inputs and by adding a visual interface as complement to the application the user might feel more in control. These guidelines could provide a bigger sense of control and trust during the usage of the system.

Most users participating in the tests described in this thesis did not really feel entirely ready to use voice, especially not in public, yet. To increase the interest and usage it is necessary to find ways to integrate voice into the users everyday life.
Acknowledgements

This author would first of all want to send a big thank to the CX team, innovation lab and emerging tech at SEB for providing insights, ideas, support, fun discussions, and many fika moments with lots of laughter. Another special thank goes to the great supervisors at SEB, Daniel Regestam and Caroline Malmberg, for their support, guidance through this process, help to structure and discuss this thesis by reading it multiple times. Additional thanks go to the fantastic people participating in the workshop, mockup tests and prototype tests.

Further, a big thank you to the author’s supervisor at Umeå University, Ola Ringdahl, for all constructive feedback and times he has read this rapport helping to improve this thesis.

Last but not least a warm thank to family and friends for their support, and to Rebecca for her support and company at the office during these months.
# Contents

1 Introduction 1
   1.1 Problem statements and objective 2
   1.2 Contribution 2
   1.3 Delimitation 3

2 Background 5
   2.1 SEB 5
      2.1.1 SEB and voice assistants 5

3 Theory 7
   3.1 User experience design 7
   3.2 User Interface Design 8
   3.3 Voice User Interface 8
      3.3.1 Conversational User Interface 9
      3.3.2 Intelligent Personal Assistants 10
      3.3.3 Discoverability 11
      3.3.4 Pros and Cons with VUIs 11
   3.4 Technology Acceptance Model 12
   3.5 What is trust? 13
   3.6 Cognitive load 14
   3.7 Design concepts of Voice User Interfaces 15
      3.7.1 Context 18
      3.7.2 Providing Context 19
      3.7.3 Disambiguation 19
      3.7.4 Confirmations 20
      3.7.5 Error Handling 21
   3.8 Transactions with voice 22
   3.9 Summary of the Theory Chapter 22
4 Method

4.1 Workshop

4.1.1 Presentation of scenarios

4.2 Mockup

4.3 Prototype

4.3.1 Dialogflow

4.3.2 System Usability Score questionnaire

4.3.3 Interviews

4.4 Evaluation

5 Results

5.1 Workshop

5.2 Mockup

5.3 Prototype

5.3.1 How the prototype works

5.3.2 SUS questionnaire

5.3.3 Interviews

5.3.4 Observations and errors in the interaction

6 Discussion

6.1 Result

6.2 Connections to theory

6.3 Method

7 Conclusion

7.1 Future work

Bibliography
1 Introduction

In the movie Iron Man (2008) the lead character Tony Stark took intelligent personal assistants (IPAs) to another level with his operating system Jarvis which assists him in his everyday life and heroic adventures [1]. We also have the movie Her (2013) which takes place in a near future and is about a lonely writer who falls in love with his operating system called Samantha, she is designed to meet his every need [2]. If we go back 40 years from now we find C-3PO in the famous movie series Star Wars which started out in 1977 [3]. He was a golden protocol designed to communicate in a conversational manner with other organics. Backing up another 10 years to the movie 2001: A Space Odyssey (1968), where astronauts go on a quest in space to find a mysterious artefact with help from an intelligent supercomputer named HAL 9000 [4]. These are all intelligent conversational interfaces communicating with humans and they have set our imaginary bar and expectations very high for how we want an Intelligent Personal Assistant to behave. We have Apple’s Siri, Google’s Google assistant, Microsoft’s Cortana and Amazon’s Alexa, they are in the process of making it into peoples everyday lives. Even though they still have a long way left to be able to behave like Jarvis, Samantha, C-3PO or HAL 9000 our IPAs have quickly risen in numbers of people using them all around the world [5].

Users would like to see themselves as instant experts, this results in a very low tolerance for errors and high expectations in applications. If something goes wrong it is not the user’s fault - it is the system that is the problem. The reason for users performing in an incorrect manner is often because the system has poor usability. There should not be anything that the users have to learn before to be able to use a product [6].

Voice-enabled technology can be found in our smartphones, televisions, smart homes and many other smart products. Along with the advancements which have been made within voice recognition technology and smart home technology, voice interactions are only expected to continue to grow in the costumers’ interest. There are times voice user interfaces are suited as a feature and work together with a graphical user interface, for example when searching for a show on your AppleTV. However, there are times when voice is the only option to interact with a product, examples of those kinds of products are smart speaker devices such as Google Home or Amazon’s Echo. By adding voice user interfaces as a feature to a graphical user-centred design, voice can enable much more convenience for the user or broaden the product to work as a voice-first application functioning on smart speaker devices opens up possibilities for creating smart homes. Although, it is important to keep in mind that designing a voice user interface is nothing like designing a graphical user interface. The design guidelines for graphical user interfaces cannot be applied to voice user interface because of the lack of a visual interface removing the possibilities to create visual affordances. This means that users will receive no indirect visual indication of what the user can perform with the interface.

Speech is a more natural way for humans to communicate than interacting with touch screens or typing. The users are excited about how quickly they can complete routine tasks
thanks to voice-enabled technology and daily tasks such as checking traffic on the way to work without even lifting a finger [7]. This adds value for anyone and especially for people with disabilities. People are also very excited about the technology and what future possibilities it can create for interactions between a device and a human. People who use voice assistants use them with purposes of completing various number of tasks to help them run their everyday routines. Using devices with voice assistants is a quicker and more user-friendly way than typing. Nearly one in four adults in the USA have a smart speaker device and about 58% of smartphone users talk to voice assistants through their smartphones. 60% of the users like voice assistants because it’s hands-free. 48% like it because they can do other things at the same time as driving or cooking. 39% because it is more convenient that touch or typing. 19% because it is fun and 6% think that the answers are better than with text-based searches [8].

To be able to create great user experiences with voice interactions, it is important to have an understanding of how humans naturally communicate using their voices. It is also important to understand the fundamentals of voice interaction. But first, it is a good idea to begin with examine the attributes of voice communication between humans. From there move on to examining design guidelines, especially for voice user interfaces interacting with humans [9].

1.1 Problem statements and objective

There are existing general guidelines for how to design a voice user interface and voice assistant but it is interesting to look at how to handle these interactions when handling sensitive information such as a person's bank account information. It is important that all industries follow the trends and this includes the banks as well, especially since voice interaction usage are increasing [8]. Providing private and personal information is important for users that it stays private, and that it feels safe and secure to provide it, especially when a technology is new to the market. This is why it is important to investigate how to design for trust when there is nothing visual for the user to read and re-read, so how do we design for trust for voice-first interactions? Trust is important because it is the foundation around which all human relationships revolve [10].

The questions this thesis will try to answer is: Which guidelines can be found and developed for additional design guidelines focusing on perceived sense of trust specific to VUIs and bank interactions? Is the market is ready for such a voice application?

1.2 Contribution

The contribution lies in finding additional guidelines for designing voice assistants when it comes to banking matters and sensitive information that is private, where trust is required for the user to dare to use such a system.
1.3 Delimitation

In this thesis, the focus has mainly been on the actual design of the conversation between the user and the system. This means that no log-in to the system will be included, the prototypes which have been developed will simulate a system where the user has already logged in and activated the system. The final step to approve a payment via an additional system such as mobile bank ID, voice recognition or something similar is also not a part of the prototype.
2 Background

This chapter describes shortly about SEB, why they are interested in this subject.

2.1 SEB

This thesis was done in cooperation with SEB, Skandinaviska Enskilda Banken AB, a Northern European financial group, in their headquarter in Stockholm, Spring of 2019. SEB’s interest in this project is to get more insights into Voice interactions, conversational user interfaces and Intelligent Personal Assistants and how they could be integrated into their services. SEB is a universal bank founded in 1856 as Stockholm’s first private bank and one of the first business banks in Sweden. At the time this thesis was written (2019), they have approximately 16000 employees. SEB’s vision is to be a relationship bank which delivers world-class service. They are and want to be more than just a bank for their clients [11, 12]. SEB strives to be at the forefront of meeting customers from all directions and with every opportunity. They have a great drive and curiosity in technology and always want to find new ways to develop their services for the customers need.

2.1.1 SEB and voice assistants

SEB is interested in voice and what possibilities it may bring is because it is just as important for banks to keep track of the technological developments and improvements happening every year as for any industry. It is important to follow trends and where the customers’ interests lie to improve the customers’ impression of the company. Since SEB is a bank they have a very broad target group. Everyone has a bank or involved somehow in the banking business. The bank needs to be able to reach everyone and meet their customers from their point in life. This also includes making bank errands easy for people with visual impairments or anyone with reading and writing difficulties. Too many companies are unfortunately neglecting the quick technological development happening right now and they have a lot of catching up to do [13].
3 Theory

Voice is a relatively new up-and-comer, not because of when it was invented but because it has quite recently become realistic and getting smart enough for us to want to continue to use in everyday living and to move into our homes. To be able to continue this development and journey of voice there are some things that are good to know. For example:

- How do developers and designers design for a user?
- What are the previous takes on designing Voice Interfaces?
- How do developers and designers design Voice Interfaces?
- How do developers and designers design a conversation?
- What is a conversation?
- How do developers and designers make users want to continue using it?
- Why is trust important?

These questions are treated in this Theory chapter.

3.1 User experience design

User experience, often shorted as UX, is a term used to describe how a user experiences a product, service, website etc. It can be applied to many different applications, for example, a website, to enhance the user’s experience during the visitation time no matter how long or short it may be [14].

Usability and UX can be about making everyday living easier and more fun to use both at work and leisure time. Good user experience leads to good use and benefits for both the user and the business no matter what the solution is about. In every case it is about defining the businesses goal and understanding the users’ goal and behaviour, then to shape solutions to help achieve both goals [15].

The goal with UX design is to make the user’s interaction as simple and effective as possible so the user can achieve the goal with the product. In other words, the focus with UX design is on how we use a system. At this stage, we concentrate on foreseeing what the users’ needs are and making sure that the interface has elements that are easy to access, understand and use [16].

Usability is as much about user-friendliness, i.e how simple, smooth and logical a system is to understand and use, as user benefit, i.e how the system can support the user to achieve
their goals by providing valuable information and functionality. Usability is not unique to web development, but is, or should be, a natural part of all product and service development [17, 18].

3.2 User Interface Design

User interface (UI) design, is the interface which a user interacts with. Often it is referred to as a Graphical User Interface (GUI) when a visual interaction is involved. It shows how a product will look like and perceive. This is where colours, brightness, contrast, formatting of graphical elements, pictures, fonts and font formatting to capture the target group’s interest is applied. The focus is on how a system will look like. What is important is to have a well-functioning interface, informative elements and colours that are comfortable to the eye. It is the graphical necessities which sets the tone on the appearance of the products UI [16].

3.3 Voice User Interface

Voice User Interfaces allow people to use their voice as input instead of fingertaps for smartphones and mouse clicks and keyboard for computers to control various devices [19]. It is what a user interacts with through communicating with a spoken language application [20]. The ability for a voice user interface (VUI) to understand what the user says and to be able to act on it requires a combination of two technologies that are important for it to work: Automated Speech Recognition (ASR) and Natural Language Understanding (NLU). An example of ASR is, if someone would speak to you in a language you do not understand, you might be able to write down the said words but you would have no idea what the person really means to say to you [21]. With NLU, computers can understand the intent of what the person means, not just the words that are spoken. It is the NLU that makes it possible for voice assistants to understand that if asked "what is it like outside?" the intent is probably to ask for a weather forecast near the current location [22].

For humans, communication is a constant exercise in decoding different meanings. Humans have been thinking, learning, evolving and defining language and norms for several thousands of years. Talking and body language as a tool to interact each other is the humans’ natural habitat way of communicating. The machines that we interact with have had a much shorter time to learn how to talk then humans have had [20]. Sometimes humans use the wrong words and often the words humans say are not actually the words they actually meant. The purpose of NLU is about providing the computers with the required context behind what we say and the adaptability to understand the countless different variations in how we say identical things but in different ways [22].

To succeed with a Voice User Interface it is necessary to understand the philosophy of conversation. Speaking is the most basic and ubiquitous way for humans to communicate. It is the easiest way to interpret peoples intent with what they say and answers can vary but still mean the same thing [21, 23]. When a computer understands the user intents and what user means to say without having to ask for it, in a very specific way, using only voice it starts to feel like you are actually having a real conversation with someone and not just a
To have a conversation is a lot more than just an exchange of information. The participants of a conversation usually share natural assumptions of how the subject of a conversation will further develop. The participants have expectations about the quality and quantity of the contributions which every participant make. Furthermore, there are sets of natural conversation requirements which includes consistency, politeness etc. On top of that, people instinctively feel to ignore from superficial sayings if they somehow are unclear or abstract and search for a deeper and non-literal interpretation. All humans do it naturally, but a conversation is still a rather complicated process.

A theory called Cooperative Principle, created by a philosopher named H. Paul Grice is about assumptions that people attending in conversations generally attempts to be truthful, clear, relevant and informative, and shows interest in wanting a conversation to proceed.

### 3.3.1 Conversational User Interface

A definition of the word 'conversation' is an exchange of information which can be both private or formal. It is often used to describe a person’s thoughts, feelings, ideas, questions or news. In order to have a successful conversation, the listeners and speakers must act cooperatively. The Cooperative Principle by Paul Grice has divided this idea into four maxims:

- **Quality** Say what you believe to be true.
- **Quantity** Say as much information as is needed and no more, while being as informative as possible.
- **Relevance** Talk about what is relevant to the conversation at hand.
- **Manner** Try to be clear and explain in a way that makes sense to others.

To be understood, it is necessary that people speak cooperatively. In conversations but also in VUIs where these rules are not followed and confusions and frustration can easily arise. Here follow a few examples provided by Pearl where VUI interactions might go against these maxims by Grice and can affect the user’s experience in a negative way.
Quality Advertising things you can not live up to, such as saying, "How can I help you?" when really all the VUI can do is take hotel reservations.

Quantity Extra verbiage, such as "Please listen carefully, as our options may have changed." (Whoever thought, "Oh, good! Thanks for letting me know??"). Limit the amount of information.

Relevance Giving instructions for things that are not currently useful, such as explaining a return policy before someone has even placed an order.

Manner Using technical jargon that confuses the user.

A conversational system is, according to Radlinski and Craswell [27], a system for retrieving information that allows a mixed-initiative back and forth between users and agents, where the actions of the agents are performed in response to current user needs based on the conversation, using both short and long-term knowledge of the user. Radlinski and Craswell discuss that the system should at least contain these following features:

User Revealment The system helps the user express, and potentially discover their true needs.

System Revealement The system reveals its capabilities, building the user’s expectations of what it can and can not do.

Mixed Initiative Both the user and the system can take initiative for the conversation.

Memory The user can refer to part conversations and statements which remain true unless told otherwise.

Set Retrieval The system can reason about the utility of sets of complementary items.

3.3.2 Intelligent Personal Assistants

IPAs are taking us closer to making science fiction real, opening up the possibility to interact and talk with our devices and system, it is now only a question of how to utilize them. These IPAs are systems designed for the purpose to help people with basic everyday tasks by providing information through human (natural) language. These assistants often use online resources to answer the user’s questions, they are therefore very reliant on internet connection. More features are timers and alerts, calendar and meeting reminders and health monitoring [28]. Apple’s Siri was first released as an application in 2010 but was first integrated to the iPhone in 2011 [29]. Amazon’s Alexa was first released in 2014 [30], Microsoft’s Cortana in 2015 [31] and Google Assistant in 2016 [32]. These are all Intelligent Personal Assistants (IPAs) with systems that run on different devices e.g purpose built speakers with embedded assistants and applications for different smartphones. When
activated, through talking activation-phrase or physical interaction, the system records the user’s voice and sends it to a receiving server which processes the input as a command. Depending on the input and context, the assistant will receive back from the server the appropriate information to read back to the user [33].

IPAs are defined by their ability to handle the natural language as inputs and outputs, such as the ability to process the input to answer questions and execute actions given to the system [34]. The systems are built with several linked subsystems. Two of these subsystems are speech recognition and natural language processing (NLP), these translate the users’ output, into requests which the computer can understand. The requests are then processed and executed [35]. In recent years IPAs has beginning to set their mark on the market, much thanks to the progression within deep learning, machine learning and artificial intelligence. Neural networks can now handle a large amount of data making the computers much better at NLP, speech recognition and answering questions [35]. It has been made possible to use deep learning to train the artificial intelligence that is decent enough to be useful in consumer products thanks to combining these subsystems with faster, cheaper and more accessible computer powers and high amounts of available data. The opportunity to build IPA products has to lead to a rising trend with leading tech companies to produce IPAs [35].

3.3.3 Discoverability

Don Norman [18] describes the term discoverability of a product as: It is possible to determine what actions are possible, as well as the current state of the device. If a product would have high discoverability it would be easy for the user to just know which actions that are possible and what the current state of the product is. The Norman Door is one of many famous, examples provided by Norman, to explain an object which lacks discoverability. It was first introduced in 1988 and was presented as a door where people struggle to figure out how to open it, if they should pull or push. His explanation was that the door lacked affordances, i.e. affordances exist to help people figure out what actions are possible and what the desired action specific to the problem is. To help discoverability be more clear one can use a signifier to give out clues on what the possible actions are, an example of a signifier can be the sign 'PUSH' on the door [18].

This is where it can get tricky with a voice-only interaction, the users are often unaware of the features that are available [36]. Corbett and Weber [37] discussed the challenges to design a voice user interface with both good and high discoverability. They concluded back in 2016 that when interacting with a voice interface there has always been issues with discoverability and learnability. Discoverability of voice commands is cumbersome since they are hard to give out clues for, compared to graphical interfaces. Corbett and Weber [37] also concluded that users will often experience inconsistency and interpret an incorrect mental model of how the system works along with and what it can and can not do.

3.3.4 Pros and Cons with VUIs

Interacting with mainly voice-activated systems can come with issues. Because language is humans main way to communicate and can have consequences with too high expectations which Voice User Interfaces still have a bit to work on to be able to reach these expectations [38]. Speech is asymmetric, it is faster than typing but listening is slower than reading, so VUIs are thereby harder to scan compared to GUIs [38]. This is because speaking is lin-
ear. There is no rewind button on listening, as there is with the possibility to go back and re-read just one specific sentence, the Voice Assistant will probably read back the entire paragraph. This results in the conclusion that reading is better for ones’ memory retention and comprehension. To listen also challenge humans skill for real-time comprehension, because it requires that the individual listening need to interpret and understand instantaneously in order to understand what the other person is saying [39].

Another issue is that the user can not see which options and actions the interface has available in its system when the system is no longer speaking to us. This applies pressure to the user to remember what the user is supposed to say and where the user is in the ongoing conversation. This can be very cumbersome for the user because of the limited capacity in humans short term memory. This concludes it to be a poor idea to present too much information to the user at the same time leaving the user overwhelmed and making it hard to recall all the presented information. Voice can also be limited due to noise, public places or just someone else talking next to the person using the assistant, either the assistant or the user can not hear the responses or interprets incorrectly leaving the user confused and frustrated. If and when errors occur in communication between a voice assistant and a user can be quite cumbersome to get back on track [40]. These limitations should be well understood before designing a VUI.

However, conversations are very intuitive and they let the user say what they want to say. It is the ultimate shortcut since it can provide speed, simplicity and ubiquity because with the voice we can get our answers and actions pretty much instantaneously. It can save time and effort for the user comparing to a screen-based user interface. A conversation can be a lot quicker compared to the number of taps a task would take on a graphical user interface. It reduces frictions with these shortcuts to get the user want they want. Conversations can also provide the opportunity for the user to multitask. Help the user in busy situations when their hands and eyes are occupied [41, 42].

3.4 Technology Acceptance Model

Technology Acceptance Model, TAM, [43] is an established model for measuring the users’ acceptance when using an information system. This model is said to be able to be applied to any form of technology. So, can this model also be applied to Voice Interface? To be able to answer that question, first step is to understand this model. The TAM is an extension on Ajzen and Fishbein’s theoretical model TRA - Theory of Reasoned Action developed during 1975 [44]. The intention with TRA is to explain and predict social behaviour in specific situations. TAM has been developed for two main reasons, first to increase understanding with acceptance in users’ by describing the motivating process that occurs between the system and the user’s behaviour. Second, to create a practical way to empirically test acceptance in systems. By allowing users to take part in the prototyping it is possible to test how the users would use the system. This is important since the design of systems and functions often lies in the hands of different developers. No matter if a system is developed for internal usage or for selling, the people developing the system have a great impact on the functionality and abilities. These functions have in their turn a major effect on how the user will use the system [43]. The goal with TAM was to try to understand the potential users’ overall attitude towards a system [43] and it was supposed to explain why people accept or reject new technologies. The attitude is said to be the main factor for whether the user
will use the system. The path to attitude towards using a system is divided into two tracks, perceived usefulness and perceived ease-of-use.

Perceived usefulness is defined as the degree of how well a person believes that using a system will increase their job performance. People tend to use an application if they believe it can help them perform their jobs better. However, the benefits of an application are outweighed by the effort involved in using the application if potential users believe that the system if too complex to use. Perceived usefulness can be influenced by external variables, for example, comparing two different forecast systems, but still equally easy to use. If one of them would produce an objectively more accurate forecast that application would be seen as the more useful system despite the ease of use equality. Likewise, if one graphics application produces higher quality graphs than another equally ease of use counterpart, that one would also be considered the more useful application. As the importance placed on perceived usefulness increases over time, the difference in importance between usefulness and ease of use increases [43].

Perceived ease-of-use is defined as the degree of how free from effort it is to use a system. Perceived ease-of-use has, in turn, a casual connection to perceived usefulness. If something is considered to be easy to use, it affects how easy it is to fulfil a function and thereby perform something which in turn is linked to achieving utility. The easier a system is to interact with the greater should the users’ feeling of efficiency and personal control concerning their ability to execute the sequences of behaviour needed to operate a system. Efficiency is one of the most important underlying theorized factors for internal motivation. Influences on ease-of-use and usefulness are different design properties since design and various functions influence ease-of-use and utility. The design has no direct influence on the attitude or use of the system but has an indirect impact on the usefulness and utility [43].

Users adopt an application mainly because of the functions it holds and the benefits it generates, while the severity in using the application is regarded as a secondary factor in the process of adoption [45]. For instance, if the system provides critically necessary functions, users are often willing to look pass some level of severity in using it. A system that is too complex to use can lead to a weak adoption, even though the system is otherwise useful. Although, no amount of ease of use can compensate for a system that does not perform a useful function. It is therefore important for system designers, who have a tendency to look pass usefulness in favour of ease of use, to realize that usefulness has proven to have a significantly more important role than ease of use. If consumers can predict the clear benefits of using complex technology, they may be more motivated to learn about and try to use the technology [45, 46]. When it comes to Voice banking adoption, it is important that consumers understand that voice banking is useful to them and can make their everyday lives easier, they may decide to choose traditional banking if voice banking is interpreted as too difficult to use [45].

3.5 What is trust?

To be able to design for trust it is first important to establish what trust is for humans and also why it is so important for humans. Trust is a brain process that binds representations of self, other, situation, and emotion into a special pattern of neural firing called a semantic pointer [47]. It is a multitude of several psychological expectations, means that a trust-
worthy party will not act in an unpredictable manner. Trust means a feeling of not being vulnerable to another party. A relationship based on trust builds on both parties will behave benevolently. The correlation between both parties is that one can not control the other. However, both parties are dependent on each other for trust to emerge. This means that both risk and dependency are necessary factors for trust [48]. Trust is a willingness to be vulnerable. A belief that a person will behave and act in a specific way that depends on another. An abstract mental attitude regarding an argument that someone is reliable. A feeling of confidence and security that another one cares. And a complex neural process that binds different beliefs and portrayals into a semantic indicator that involves feelings. It is something a person chooses to do, to take a risk and be vulnerable to the other party. Meaning that if someone says "I trust you" it means "I’m willing to be vulnerable to you" [49]. Trust is necessary for relationships because with no trust for the other part there can be no continuous relationship.

Kim, Shin and Lee [48] formulate that things that happen on the internet are more anonymous. This is probably not the case when something happens in reality between two people. It results in an even greater risk, uncertainty and loss of control in the digital world. This means that to reach success, it requires that customers can feel confident. Creating this trust, however, is something that is both costly and time consuming [48]. Kim, Shin and Lee believe that trust is something that has been shaped for a longer period, which means that when new innovative services are launched, there may be no previous experiences to look back on.

Trust can be divided into two different phases: initial trust and cumulative trust. As users gain more and more experience in something, initial trust evolves into cumulative trust. This means that it is important to try to strengthen users’ initial trust in a product or service because cumulative trust is not yet established. Trust is a feeling characterized by insecurity, vulnerability and dependence. These aspects are recurring in terms of online transactions. In the environment in which these online transactions are carried out, it is not always possible to see who the seller is or to take the purchased product and leave immediately after the payment has been done. These are examples of how trading in the digital world entails greater risk than ordinary trading. In order to feel trust in the digital world, this means that the customers must feel trust in the seller, the seller’s technical expertise and previous experience the seller [50]. From a designer’s perspective, the goal is to get their users engaged and quickly establish trust. To establish trust is for many users something that is difficult, trust can quickly and easily be affected and if once it has been damaged it is difficult to rebuild again [51].

3.6 Cognitive load

The term cognitive load mainly refers to the level of mental effort a user needs to successfully complete a task or to respond to a request [52]. In the context of User Experience, the prime concern is to not overwhelm the user with too much information or options all at once and also to find the most efficient and best path to reach a specific goal. Simply wanting it to be a logical flow for the user to recognize instead of the need to recall from memory.

When a user interacts with a graphical user interface (GUI), even if not actually performing
a task, the user reads or navigates the interface mentally, meaning that the user interacts with
the GUI in a constant way. This tells us that the level of attention that is required to operate
the interface continues to be more or less consistent throughout the entire experience [52].
However, the pattern of attention when a user interacts with a voice user interface (VUI)
is slightly different. To activate the interface the users needs to use an activation phrase
and from this the user will need to be fully alert to receive and understand the system’s
responses. There are no options to skip information like there is in a GUI or take time to
understand the response, e.g. re-read a text. The user is forced to pay full attention during
the interaction, except when no direct interaction is taking place when there is no need to
pay attention at all. These different cognitive load patterns of attention can be displayed
with a graph as seen in Figure 1.

![Lead of Attention: Graphical vs Voice User Interfaces](image)

**Figure 1**: Peak of Attention: Graphical vs Voice UI [52].

The figure shows when the attention and thereby cognitive load of the user is high and low.
For the voice and graphical UI’s respectively the main cause for the peaks is that the user
can not control the speed of the information that comes in from the VUI, which can lead to
a negative experience and increased numbers of errors that the user makes. To avoid this
high cognitive load on the user and keeping them from getting overwhelmed, the length and
information density of responses should be limited [52].

### 3.7 Design concepts of Voice User Interfaces

To avoid issues and friction between a user and a voice assistant there are some guidelines
and concepts to follow. This section will present some of the concepts that are relevant to
this study. A voice user interface uses a variety of grammars, prompts and dialogue logics in
combinations with each other [20]. A prompt is a system written message scripted through
the VUI’s design and sent the user. The logic of the dialogue determines which prompt and
grammars are the predicted and defined responses the user can say to the system [53].

Four tips to design natural voice-first experiences [22]:

1. **Keep prompts simple and clear.**
2. **Provide clear and concise responses.**
3. **Anticipate user input and prepare for common responses.**
4. **Monitor and adjust the conversation based on user feedback.**
1. **Identity Intent**
   What are the different goals someone wants to accomplish? Be specific.

2. **Identify Utterances**
   What are the different words or phrases people might say to signal their goal and intent? The more examples the better.

3. **Cover Corrections**
   Natural conversations is not perfect. Give users the opportunity to correct errors or change their answers.

4. **Build Exceptions**
   It is always better to say "I do not know the answer to that", than to pretend and give a wrong answer.

Beneath follow some further recommended guidelines to have in mind for designing VUIs [54]:

1. Rethink your information hierarchy
2. Design for how people talk, not how they type
3. Recognize a variety of questions and commands
4. Keep the assistant’s responses short
5. Try to stick to a maximum of 3 choices
6. Add variety to your answers
7. Personalisation is important

Simply looking at one prompt at a time separately will not lead to a natural conversation, but instead, result in graceless and stiff conversations. Sample dialogues is a series of prompts and responses that be used for conversational design, emphasizing on the conversation. Cathy Pearl [21] recommends that the VUI designer pick five of the most common use cases for VUI as well as a number of error dialogues for when errors occur. After or during the process of dialogues are read out loud to determine if they are too awkward or too formal, these dialogues can then be used in mockup testings.

When the best dialogues have been tested and iterated, the VUIs dialogues can be sketched out. Dialogues in VUIs are diagrams that work as a map for all available paths that can be taken through the system. In cases with a very large number of possible paths in an open-ended conversation, the flow can be sorted into groups by type of interaction, but all possible interactions do not need to be put into the diagram [21].

**Command-and-control or conversational**

When designing the system for VUI’s it’s the dialogue-interaction between the interface and the user which has to be designed and structured [55]. There are a variety of ways that the user can interact with an IPA. One way is by a **command-and-control** interaction, and this is also the most common way of interacting with a VUI. In these types of interaction, the users have to explicitly indicate each time they want to talk with an awakening word or phrase (e.g "Hey, Google" or "Alexa") to the voice assistant. The assistant will thereby
show the user that the assistant is listening by indicating it with a nonverbal-audio and/or some kind of visual feedback of lights or animated dots in motion. This is a suitable course of action when the assistant does not know when the user is expected to talk to it, to avoid disturbing a conversation that is not intended for the conversation and allowing the user to think before answering. The assistant estimates when the user is ready to talk and then responds. The conversation is not expected to continue after the first exchange [21]. The other way to design a conversation is the conversational design technique, this approach is becoming more and more common and it uses a more natural turn-taking interaction technique between the user and the voice assistant. It can be considered awkward and unnatural for humans to always keep indicating the assistant that they wished to speak to it as in command-and-control. The conversational approach instead lets the assistant use, for example, pauses, explicit directions or questions and waits for the user to respond without closing the conversation. In a system using the conversational technique the users is more involved in a conversation without the users having to activate the assistant every time they want to talk. For example: some systems which often uses a conversational technique are chatbots [21].

It can at times be possible to switch between command-and-control and conversational modes if the situation is clear and the user understands when the mode has been changed. An example of a suitable situation for switching the mode is the game Jeopardy! on Amazon Echo where the user only needs the activation word to start playing the game but not again after it has begun [21].

Keep it simple

Since all audio outputs are volatile, linear and always moving forward in an ephemeral manner and constantly fading, the VUI has some unique challenges. Important guidelines for designing a voice user interface are provided by the Interaction Design Foundation [9] are described below:

- Provide users with information about what they can do.
- Tell the user where they are in the system.
- Express intentions in examples.
- Limit the amount of information.
- Use visual feedback (to show that the system is listening).

A goal when designing a visual user interface is to try to limit the number of clicks a user has to perform to complete an action, this approach is to create a better experience for the users. More clicks, calls and answers have a negative effect on the experience because it often results in a feeling that the task is bulky, boring and ineffective. The designer should seek to limit the number of back and forth in the conversation between the user and the VUI. The rule of thumb to keep them limited is to closely related to the principle of keeping the interaction short. While a user of a visual interface may skim it, a VUI user only has the option to linearly listen for the assistant’s outputs [21], this makes a VUI less
much less scan-able than a GUI [51]. An additional factor to consider is humans’ short-term memory, which makes long instructions difficult to remember and follow [38]. It is therefore important to keep it short and just present the most critical information. This critically relevant information itself should be hierarchically ordered depended on the relevance. In visual experience, users can quickly move their attention to the information that is most relevant to them, the users have the option to skip certain uninteresting parts on the interface. But when it comes to a voice user interface it is not possible to scan the interface for the most relevant information for the user, nor is there a way to skip given information from the voice assistant. The user is forced to listen to everything that the application has to say and because of this, it is important to be short and consist, in other words, keep it short and simple. Only present the most important information and with the options for the interaction. An example to compare a VUI with a GUI is of a shopping app, the user is to search results for a given product. Things to present from the VUI could then be the title of the product, price and rating. Everything else can be presented later and letting the user know that they can ask more specifically if they want to know more. This list of products in a GUI can provide much more information in the interface for the users to scan by them self [21].

In GUI’s the user can by them self see where they are but with VUI’s the user must be told what functionality they are currently using. Otherwise, the user can quickly get confused or might activate an unwanted functionality by mistake. The users a basically running blind in the dark [9].

When people talk it is not common that they fully express their full intention since people are used to talking in slang, shortcuts or hinting about what they really mean. This usually works fine in a conversation with other humans since they usually get what the other person means. But in VUI’s it is necessary to express intentions so the system understands the users’ true intention. An example with Alexa is that a user can say “Alexa, ask Astrology Daily for the horoscope for Leo.” and get the right information which the user is searching for, instead of first saying “Alexa, ask Astrology Daily.” and then ask for the horoscope the user wants information on [56]. It is therefore important to guide the user to let them know what is possible to do and for what they can say [9].

As mentioned before, with visual interfaces users can go back to overlook any information they might have forgotten. This is not the same case with verbal interfaces. This is why it is important to keep sentences and information as short and brief as possible to not confuse the user and make them feel comfortable and help them stay in that space they are currently in. We do not want the users to think too hard so we have to respect the humans’ short-term memory [9].

3.7.1 Context

For a user to experience an intelligent conversation with a VUI it requires that the assistant can handle more than one exchange of information from the user, called a turn. For this to be possible the system behind the assistant needs to keep memorizing and keep track on the user’s recent turn history.

Here follows an example by Pearl [21] of how the Google Assistant handles conversations and remembers the previous context:

USER
OK, Google. Who was the 16th president of the United States?

**GOOGLE**
Abraham Lincoln was the 16th president of the United States.

**USER**
How old was he when he died?

**GOOGLE**
Abraham Lincoln died at the age of 56.

**USER**
Where was he born?

**GOOGLE**
Hogenville, KY.

**USER**
What is the best restaurant there?

**GOOGLE**
Here is Paula’s Hot Biscuit

In this example, by Pearl [21] the assistant remembers the context and manages to keep the conversation moving forward making it feel more like an actual and realistic human-to-human conversation. A system that would not be able to keep track of the context and recent information would only be able to handle on-turn actions. This is something that plays a vital part in communication, to be able to use two different terms but still refers to the same thing and this is called coreference. Without coreference, a conversation would quickly crash. This example above shows why coreference is important. The assistant understands that Abraham Lincoln is the same as the pronoun "he" and that the pronoun "there" refers to Hogenville, KY [21].

### 3.7.2 Providing Context

One of VUI’s biggest challenges is to teach the user what is possible for them to do. The difference VUIs have from GUIs is that there are no visual objects available, no buttons or menus to browse through. This is why it is so important that VUI’s are designed in a way that informs the user on how to interact, what actions are available to perform and what the correct responses are when interacting with the system [21].

Pearl [21] suggests that a system to initiate and help the user by informing how the user can respond or possible actions to ask for. A good VUI design has to be capable to handle when and if a user gets lost by always allowing the user to ask for assistance. A possible way to help the user out is by providing their current context (if there is one) and then give hints to guide the user in the correct direction [21].

### 3.7.3 Disambiguation

By using contextual information the system can make assumptions when a user is, for example, asking about the weather in London [21]. The user is in his/her head referring to London in the United Kingdom and not London in Canada but not providing all the information. In these situations when the user is not providing all the necessary information the system is could use contextual information based on if the user has previously looked up information about London (UK) the system assumes that the user is referring to the city in
the United Kingdom and not the city in Canada. If there is no previous contextual information available it is necessary for the system to ask the user "Do you mean the one in the UK or the one in Canada?" [21].

3.7.4 Confirmations

When designing a VUI it is important to make sure that the user feels understood or to let the user know when and if they are not being understood. Even if confirmations are a good thing there is a way to overdo it, this can result in a very unnatural interaction creating a negative user experience. It might provide a closer accuracy but is can also irritate the user greatly. When designing confirmation methods the situation is an important aspect to keep in mind. Think about the consequences it could be if something was to go wrong? An example where confirmations play an important role is with transferring money, this could have very big consequences, compared to asking a music service to play music which has much lower consequences.

Pearl [21] explains that there are different ways to confirm information to the user:

**Three-tiered confidence:**
The system will explicitly confirm the information that has a certainty threshold between 45-80% accuracy, but will reject if it has less than 45% and will implicitly confirm anything over 80%.

**Implicit confirmation:**
This method implicitly confirms things without rejecting the user to perform any actions. An example is when you ask "Which is the world's highest mountain range" so the system’s response can result in "The world’s highest mountain range is the Himalayas in Asia, highest of all these peaks is Mount Everest which with its 8,848 meters is the world’s highest mountain".

**Nonspeech confirmation:**
In this case, there is no spoken answer that is performed when an action is completed. For example, an application that turns on and off the lights then becomes its confirmation. A spoken confirmation is then not necessary in this situation.

**Generic confirmation:**
In a more conversational system, other types of confirmations may be useful. for example in situations where the user uses the system more openly without a real goal. For example, if a system asks how a user slept the night before. The user can then respond poorly. The answer from the system can be "I’m sorry to hear that".

**Visual Confirmation:**
In multimodal interfaces (a system that has both a screen and a voice to interact with, such as a mobile phone), it is common to use both voice and visual confirmation when for example asking the device a question.

Google Home shows confirmations with visual feedback, a nonspeech confirmation, for example when a user activates the assistant with the awakening phrase "Hey, Google" or
"OK, Google" the device lightens up at the top with animated dots in motion to show that the assistant is listening. These dots also start to blink as the user talks to indicate that the assistant is listening and processing the input. As the user stops talking the dots start to move in a quicker motion which also quickly fades away to show that the assistant has stopped listening and is now trying to formulate a response to the user. These light-shows in different motions are to indicate that something is happening and that the device is paying attention to the user.

3.7.5 Error Handling

A system that uses an Interactive Voice Response (IVR) has the system to prompt a response. Example: if a user said something that was not understood or heard, the system is to prompt the user to repeat themselves. An example of an IVR system can be a system for booking appointments. It is important to prompt the user to repeat what they said so the user knows that they were simply just not understood, not that the system has been cut off or is not working. Because of this, the system uses a fixed script it is required to make sure that the user provides the right input so that the conversation can keep moving forward. Therefore when the time is out it asks the user to speak again (Pearl [21] recommends a time-gap of 10 seconds for the user to answer even when they are not responding). In case these timeouts are not carefully planned the user and the system can create an awkward and frustrating interaction by cutting each other off.

It is not always mandatory to re-prompt the user when an error occurs, for instance, Amazon's Echo does nothing if the user says nothing after awakening the assistant. When a user uses a device, they answer with silence in the same way as if they would be talking to a human by repeating itself. An IPA does not use fixed scripts and does not wait for the next input from the user as an IVR does, simply because it is most often a one off-command and there is no need since it is only a transaction which fails and not the entire conversation, according to Pearl [21].

What is important to remember is that even though voice recognition has improved tremendously in the last 10 years to more than 90% accuracy with the right conditions. This however not guarantee a smooth interaction for users when interacting with VUIs. A great design should always be able to handle if errors occur or things go wrong [21].

- No speech detected
- Speech detected, but nothing recognized
- Something was recognized correctly, but the system uses it in the wrong way
- Something was recognized incorrectly

Escalating errors is a strategy with cases when speech is expected. These types of cases can be if the user does not provide all the necessary information or simply is indistinct in the response. Then the system can prompt the user to repeat and even add what information is missing to make it clearer for the user to know what to provide in the next response [21].
3.8 Transactions with voice

There are a few companies particularly in the USA who have already hopped on the trend with bank transactions through Voice Assistants. You can now transfer money to friends with Google Pay through Google Assistant but not with Google Home. Although, it only requires that the receiver exists in the sender’s contact list and that they also have Google Pay. This only exists in the USA and other English speaking countries. But then there are some banks as well, for example, Ally Bank and CapitalOne although they use Alexa and not Google. With Ally Bank [57] a user can make one-time transfers between one’s internal accounts, or to and from the user’s registered accounts at different banks. For extra added security, Alexa will ask the user to say a previously registered voice passcode as six digits to verify it’s the correct user and then continue with the transfer. A user can among a few more things check one’s balance, transfer money, hear recent transactions and get today’s rates.

Say things like:
"Alexa, open Ally and transfer money."
"Alexa, ask Ally to make a transfer."
"Alexa, open Ally and transfer $100 to my savings account."
"Alexa, ask Ally to transfer $100 from my savings account to my checking account." [57]

CapitalOne is also a bank in the USA which has enabled this feature and also uses Amazon’s Alexa and a user can check one’s balance, track one’s spending and even pay a bill [58].

3.9 Summary of the Theory Chapter

It is important to understand how the human brain works, how we remember and handle information when it comes to processing only through listening. To respect humans short-term memory, the cognitive load and that users are easy to annoy. Understand what a conversation consists of and how humans communicate with each other to how to design a VUI.

User Experience is important when designing a product or service to increase the understanding for the user, their needs and the problem to solve. To understand how to design a voice user interface it is beneficial to understand the differences in designing a graphical user interface. We have for a long time now interacted with visual interfaces and this has resulted in many guidelines for designing them. But since voice user interfaces are not that old on the market, recent general guidelines from Amazon and Google has been developed. Voice user interfaces do not have visual affordances so the user can not get a hint of what features are available. It is important to remember to not overwhelm the user with too much information at the same time and try to only present what is the most important for the user to know in each step. Voice can also be cumbersome in surroundings with lots of noise. If errors occur the user can quickly get frustrated. That is why it is important to understand the limitations that exist when designing a voice user interface. Although, looking at how natural it is for us humans to speak our minds instead of typing this type of interaction can help users perform tasks without lifting a finger or open up the possibility to multitask. The Technology Acceptance Model (TAM) is about the user accepting a system if it is useful even if there are lots of possible improvements to make the system will still be used if it
helps someone in their everyday life. TAM can be applied to any form of technology which includes voice assistants, with the purpose of finding if there is a value in using them. If a system looks good but gives us no value, then why would we use it? If we find a purpose within a system (value) and it looks good as well, why would we not want to use it?

When designing these voice-first interactions some recommended guidelines are to design for how people talk and not how they write. If a voice assistant would start talking as scientifically as the way people write it would sound very unsuited and with a lack of comprehending from the person listening. Try to keep the assistant’s responses short and to a maximum of three choices to not overwhelm the user with information. To set the assistants tone with some personalizing characteristics would increase the level of convenience with the user. Try to start off by understanding conversations, sketch dialogues and test them out as mockup to users and listen for their opinions, there are lots of ways to design a conversation. Should it be by commands or more of a conversation? Then it is appropriate to start developing. When developing the system, the designers should design the VUI so it hints to the user what the user can do and what to say. It is also important for the user to understand where they are in a conversation so let the user know what they are doing. When errors occur it is important how the assistant handles those errors and not blame the user. Confirmations can be handled in different ways depending on the situation and conversation. Example: if the user wants to start interacting with a Voice Assistant, a visual confirmation pops up which indicates that the assistant is listening, otherwise a short spoken or non-speech audio confirmation. Another example of a confirmation: when the user has asked the assistant a question the assistant can reply with a short repetition before answering, e.g "Which mountain is the tallest in the world?" and the assistant replies with "The tallest mountain in the world is Mount Everest with its 8,848 meters". All recommended guidelines are valid points for designing conversations which can improve the feeling of comfort and trust in the usage of voice assistants. But when handling sensitive information like money with one’s bank and sending them it is important that it also works. But how do we design for improving trust? This is hopefully something that will be found in this thesis, but we must first understand why trust is so important for us and how does it matter in using technology. What is trust and how does the technology acceptance model applies to voice technology.

Voice has already started to grow in interest within the banking industry and banks in the USA has already released some voice-only functionalities with Amazon’s Alexa. This shows that the industry is trying to catch up on the trend and technological development.
4 Method

In this chapter the process of work for the thesis is presented. The purpose and how each phase was structured. How the implementation and evaluation went will be presented.

This chapter consists of five main sections: Literature study, Workshop, Mockup, Prototype, Evaluation and Compile guidelines. This has been performed with a Service Design methodology as shown in Figure 2. The first stage of this method was to understand the challenge with voice, understand the user, find an idea to move on with, create a prototype and test the idea and then to iterate the entire process.

![Figure 2: Way of working in Service Design.](image)

SEB uses Service Design as their design process in projects. The employees working as UX designers (or CX designers as SEB calls it, meaning Customer Experience) at SEB holds a Service Design course for all SEB employees in Sweden and other countries for them all to understand what the methodology is all about and why it is useful in the process of new innovative ideas.

The aim of the literature study presented in Chapter 3 was to dig deeper into the subject, finding facts about current (spring 2019) market situation, understand the research area and the challenges about voice assistants, facts about human conversations and the existing design aspects of it. This was followed by a workshop with UX designers and others to understand the user and get a hands-on discussion on the value of voice interface and peoples experience of voice assistants to understand the users’ usage, needs and challenges using assistants currently available on the market. Using the knowledge and research found during previous steps ideating a use-case to mockup on pieces of paper digging deeper into the conversational aspect between a voice assistant and it’s user. Testing the mockup with new test participants and having them answer a few interview questions, then it was time to iterate the process.

The results from literature study, workshop and mockup were evolved into a prototype by creating a test application to simulate an invoice-payment voice application. Finding test users to test the prototype, and having them perform the tests by having an authentic invoice in-front of them and talking to the prototype assistant to make the payment go through. Each participant testing the prototype also answered a System Usability Score questionnaire and interviewing the participants in conjunction with the test. The SUS questionnaire part of the test generated quantitative data whereas the interviews are considered to be qualitative. The results are presented in Chapter 5 and discussed in Chapter 6.
4.1 Workshop

A workshop with the aim of finding where the interest of voice in everyday payments may lie was organized with UX designers and others from SEB to figure out how a conversation with a voice assistant could behave for different contexts. The purpose of the workshop was to brainstorm about the subject voice assistants and how they could come to use in financial errands, and from this find a use case to move on with and create a voice application prototype.

The workshop was conducted in Swedish and had 12 participants, six women and six men in ages between 26 and 53. The entire workshop session took two hours and it started with a short introduction to the subject of voice assistants. Next step was to randomly divide the participants into three groups of four people each. Each group was to decide two different scenarios and present two ways of conversing and interacting with a voice assistant in forms of role-playing, one pretending to be the voice assistant and the other person to perform as the user. These scenarios were to be chosen by mixing between the options of two different kinds of errands and four different kinds of environments where these errands were to be taken place.

1. Transfer money between internal accounts or to someone else’s account.

2. Pay an invoice received through email and it has to be paid right away. In this errand, there were two following options to chose by.
   
   (a) invoice reminder and the due date is tomorrow. The first invoice amount was 500 SEK but now the amount has been added with 60 SEK extra because of the reminder. So the total amount is now 560 SEK.
   
   (b) You have ordered a new suitcase for 500 SEK from a website which sends the invoice to your email. The invoice has been added with 49 SEK for the shipping costs, so the total amount is on 549 SEK. You have never purchased from this website before and the final due date is tomorrow.

Environments to choose by:

- In the cashier queue at the grocery store
- Sitting or standing in the subway train
- Walking in the city rush
- Calm day at home

4.1.1 Presentation of scenarios

The groups had 45 minutes to come up with suitable conversations depending on their scenarios. Each group had 10 minutes to presented their scenarios of conversation and then discuss, talk about their reflection and take on what they just had presented. The role play was filmed in purpose for the leader of the workshop to be able to go back and watch the interactions and analyze them afterwards. After each groups presentation, there was time for an open reflection before moving on to the next group’s presentation. When all the
group’s had presented a last and final open discussion and reflection session took part for the remaining 20 minutes.

From the workshop all of the ideas, thoughts and reflections that were considered were collected for thoughts to the future prototype for this project.

4.2 Mockup

After the workshop, it was time to decide a use case and to start designing a conversation flow for it. The use case that was chosen was invoice payments with the argument, if the invoice had been received through e-mail it might be complicated to pay the bill on the phone. The reasons behind performing the tests with a mockup was because it is a quick and low-cost method to quickly get results.

A mockup was tested with seven participants, four women and three men in ages between 26 and 46. They were asked questions with the aim to find out the wanted ways to pay an invoice. The questions asked were:

- How do you pay your invoices today?
- How would you like it to work today?
- How would you wish it to work today - what would the dream scenario be?
- How do you think voice could contribute to reaching the dream scenario?

After answering these questions the users got a bill in front of them that they were supposed to pay and they were asked to openly explain out loud how they would like the conversation to go when paying an invoice through voice. What the users wished to say and what things they want the assistant to say or perform.

The following step was to choose between two conversational flows between a user and a voice assistant. The test leader had written these two flows on post its and read it out to the participants for them to decide which one that felt better than the other.

4.3 Prototype

The results from the workshop and the mockup tests were used to develop a prototype. This prototype was created in Swedish, both because all potential users had Swedish as a native language and it would also be interesting to see how the conversation and the system could behave in Swedish. There were 9 participants in this user test of this prototype, five women and four men in ages between 24 and 45. Four of the users worked at SEB, one was a SEB customer and four remaining users had other banks. The participants were first asked to read a few short instructions on what they were to perform for the test.

_During this user test you will be recorded (only audio and the file will be deleted when the final evaluation is completed) and you may cancel at any time during the test. The test will take approximately 30 - 45 minutes in total. The test includes first testing the prototype, then answering a questionnaire with 10 questions and finally answer some interview questions._
The scenario the test users were presented with was:
You have ordered lenses from the internet and received the invoice via email. It was the first time you tried to buy lenses online and therefore the company is not included with your recipients in your internet bank. You are now going to pay this invoice via SEB’s new feature through Google Assistant.

From this, the users got to see the invoice they were to pay from an iPhone X and they got a moment to look through it before starting the prototype. When the users felt ready the prototype was started by the test leader on a MacBook Pro through Dialogflow and a Google Assistant simulator.

### 4.3.1 Dialogflow

The prototype was built in Dialogflow. It is a Google-owned development tool for human-computer interaction specialized on text-based conversations and voice interfaces [59, 60]. The tool works on natural language understanding and uses machine learning to train the application. The logic of the application built from Dialogflow can be described as decision trees and paths trained with machine learning used for the agent to better understand what the user input actually mean. This means that there are examples of what users inputs would be and those examples are written in and the agent learns from that. So if a user would say something similar the assistant would try and generate based on the training to an already given example and that is where the machine learning comes of use. This is what makes the idea of this as an intelligent agent defined as having the ability to act rationally which means that it can take the appropriate actions for given its goals and circumstances. It is flexible, able to learn and able to make appropriate choices given the resources it has. An agent has three main components, these are intents, context and entities [61]. Dialogflow matches users input to a specific intent, see Figure 3.
An intent [61] is how the agent determines what to do with an input. It also handles how it responds, what parameters to save, it establishes context to guide the conversation, handles fallback intents for how to handle unmatched inputs and how it should respond if the user is not understood. There are multiple different types of fallback intents to call for within specific contexts [59]. In each intent, the developer provides training sentences as examples of what a user might say as user inputs. The more examples provided in the training the better as it uses machine learning to build up its model. Each intent has contexts associated with it, both an input context and an output context. The contexts are used to guide the conversation from different intents and pass on the saved parameters, see Figure 4.

Figure 3: Inside Dialogflow example of intents and how the phrases that was trained. The marked words are entities the system picks up.
**Figure 4:** The picked up entities from the intents and which ones that are required and prompts out responses to the user.

**Figure 5:** Training the entities with different synonyms and formats which can be provided by the user.

Entities [61, 60], showed in Figure 5, are the parameters that the agent is looking for in different intents. The system uses these to generating responses further by making these entities required (as shown in figure 4). An example is an intent trained to find the users name, if the name is required in the intent and the user only says "my name is..." and never provides the actual name the program will ask the user for the required information. In this example, to get the user to provide the required information it simply should ask for the
user’s name, like “sorry I didn’t get your name”. Once the assistant has received the name it will now be stored inside as name-parameter value [61, 60].

4.3.2 System Usability Score questionnaire

After performing the use case with the prototype the participants was to answer a questionnaire with questions concerning what feelings arose during the test. The purpose of using a System Usability Score (SUS) questionnaire was to get quantitative data on how good usability the prototype has [62, 63]. The questionnaire was originally in Swedish and contained 10 standard item questions where the user answers each question with a five response option, from strongly disagree (1) to strongly agree (5) [63]. The questions sometimes concern the same thing but asked in reverse, first question asked in a negative and then second in a positive way, and so on. This questionnaire was originally created in 1986 by John Brooke [62, 63, 64, 65]. The purpose of it is that it can evaluate a wide variety and different kinds of services or products for example hard- and software, any kind of device, application and websites. The benefits of using this questionnaire are that it is a “quick and dirty” but reliable tool for measuring the usability in a service [63].

The questions were:

1. I think I would like to use this product often.
2. I found this product unnecessarily complex.
3. I thought this product was easy to use.
4. I think that I would need the support of a technical person to be able to use this product.
5. I found the various functions in this product were well integrated.
6. I thought there were too much inconsistency in this product.
7. I would imagine that most people would learn to use this product very quickly.
8. I found the product very cumbersome to use.
9. I felt very confident using the product.
10. I needed to learn a lot of things before I could get going with this product.
Figure 6: The SUS questionnaire that was handed out to all participants.

For odd questions, in the questionnaire shown in Figure 6, the user’s response was subtracted with 1 from and for even-numbered questions the user’s response was subtracted from 5. This scales all values of users’ responses from 0 to 4 (with four being the most positive response). Adding up the converted responses for each user and multiply the total grade with 2.5. This converts the range of possible values from 0 to 100 instead of from 0 to 40 [66].

Figure 7: SUS Scale which determines the grade of usability in system [67]

The final score of the SUS square is analyzed according to Figure 7 which shows that values below 60% are indications of significant usability problems. Values between 60% and 80% are interpreted as borderline to good. Values above 80% indicate good to excellent usability. 100% corresponds to a perfect system without usability problems [66, 68].
4.3.3 Interviews

After the participants answered the SUS questionnaire they got interviewed and asked about their answers in the questionnaire, why they had answered the way that they did, what their reasons and motivations were and what feelings that arose. The aim of the interview questions was to get some qualitative data about the concept of using a voice assistant for payments. After evaluating their answers in the questionnaire the participants still had a few open-ended interview questions to answer before finished. The questions concerned both the test of the prototype, the perceived sense of security, the lack thereof, possible requirements before using a voice service to handle money and also overall concerning voice in everyday use.

What previous experiences do the users have in handling/talking to voice assistants? What kind of emotions arose during the usage of the prototype? General experiences of the prototype, concerning the conversation, the language, suggestions for improvements on the prototype etc.

**General questions concerning trust in voice assistants**

1. What would be required for this prototype to feel safe?
2. What is your experience of voice assistants and the trust for them?
3. What would be required for you to feel trust for voice assistants?

**General questions concerning trust in voice assistants in situations of handling money**

4. What would have to be required for you to use voice assistants when handling money or banking services?
5. Any demands on the design?

**General question about trust in voice assistants among handling sensitive content**

6. What is required to enable you to use voice assistants when handling something with private and sensitive content?

**General questions on the usage of voice assistants in different contexts**

7. What types of situations/contexts can you see yourself use this type of services?
8. (If the answer on the previous question is “only in private”) What would be required for you to use this service in public places? Think openly and it does not have to be like this prototype. Dream scenario - how would you like it to be, how would you like to use voice assistant services in public.
4.4 Evaluation

After the tests were completed they were all analyzed and transcribed. The goal with the tests were to find out how the participants experienced the prototype, how they perceived the usability of the prototype, how they experienced performing a payment through voice and if there is a value in using voice instead of current payment options. Another desired outcome from all tests is to combine and compile guidelines. The data from the test was compiled for each participant.
5 Results

This chapter contains all the answers and results from the workshop, mockup, prototype tests and interviews.

5.1 Workshop

The purpose with this workshop is to get a deeper knowledge and insights about understanding the user in the context of using an IPA and voice assistant doing bank errands and how users would wish the interaction to behave and how they wished the conversation would flow. The workshop had 12 participants that were evenly divided into 3 groups.

**Group 1** presented two scenarios. Both played out at the cashier queue inside a grocery store but the differences were that the first scenario was without earphones and the other was with earphones. Standing at the cashier queue as the next person to load the treadmill for payment with another person standing after the main character. The role play showed how an embarrassing conversation between a voice assistant and a user could go and how they would not wish it to be. The assistant reached out to the user telling him that he has received an invoice reminder for 560 SEK to Transportstyrelsen and that he needs to pay it right away. The user approved the payment by telling the assistant confirmation, a phrase like "okay, pay now". The assistant tries to make the payment but noticed that the user did not have enough money in the account and asked the user if he would like the assistant to first transfer money from the savings account and then make the payment. The user confirmed once again by saying "yes, do that". The payment goes through, the assistant thanks for the user's attention and confirms that the payment is completed. The second scenario was the exact same conversation flow but with earphones to keep the conversation private but still the user only answered with confirmations and relying on the assistant to guide the conversation forward.

**Group 2** also had two scenarios to present. The first scenario was played out at the cashier in the grocery store but the twist in this scenario is that the user was trying to pay for her groceries but gets declined with the reason being that there was not enough money in their bank account on the bank account. The user taps on the earphones to activate Siri and tell it that the user wished to transfer money from saving to the main account by saying "Top off my account". The assistant can see that the user has been declined in the grocery store and also for what amount so the assistant transfers that exact amount to the user’s account and replies to the user with a confirmation that the transaction has now been made, and the user can now make the payment. In the second presented scenario, the user was at home sitting on the couch. This time the user only has a visual impairment with only 10% vision making the user almost blind. Here the assistant reaches out to the user telling him that he has received a new invoice from Ridklubben on 2000 SEK and asks if the user would like to pay that invoice. The user confirms by answering "Yes", and the assistant asks from which
bank the user would like to pay from. The user responds with which bank they would like to pay from and the assistant confirms the payment by responding to the user the user that "the invoice has now been paid".

**Group 3** had also thought out two example scenarios. In the first scenario, the user was standing in the subway talking to the assistant with earphones. The user takes the initiative telling the assistant that he wants to talk to SEB and send money to a friend named Bengt. The assistant has access to the user’s phone book and notices that the user has two contacts with the name Bengt and therefore asks if the Bengt that the user has since before had the most contact with is the right Bengt, "Do you wish to send to Bengt Nilsson?". The user confirms that it was the right Bengt that the assistant has found and the assistant replies with a swishing sound of money getting sent. In the second scenario, the user stood by the cashier in the grocery store and wanted to transfer money from savings account to the main account. The user started with activating the assistant and said that they wanted to talk to SEB. The assistant asked what the user wanted to do and replied that they wanted to top up the account with 500 SEK and then the assistant replied with a sound of coins scrambling.

**Discussion and reflection after presentation of scenarios**
The participants of the workshop thought that since it was important that the interaction was smooth and quick it is important to use small catch phrases when telling the assistant what the user wants to do or to confirm. It can be fine that the assistant reaches out to the user first if using earphones. The preferred way to communicate with the assistant was when using earphones and the user only needs to confirm and does not need to tell the assistant any form of informative input. If the user was at home and alone or with family it can be fine to use the assistant out loud and to give the assistant informative inputs. If using the assistant in public with earphones the preferred way was if it sounds for people not attending to the entire conversation like the user is having a normal phone call with another human at the other end of the phone call. Telling the assistant information was fine but there was a need for controlling the entered information, so a really long OCR number was maybe not preferred.

According to the participants in this workshop the most desired way to interact with a voice assistant especially when in public surroundings, was when the user was in control but there is a need in personalizing and making settings individually. They continue by demanding that the voice assistant needs to be calm and have understanding for the user, was welcome to contact the user but in that case the assistant needs to know when the user was wearing earphones.

### 5.2 Mockup

The purpose with this section is to get a deeper knowledge about how the design of the concept and its mockup was developed.

The majority of the users’ dream scenario when the users pay bills that are not e-invoices is that they want them to be able to work as e-invoices that are received through the bank, at least those that are via the e-mail.

The participants were to decide between two optional flows in the conversation and the majority thought that option 1 from Figure 8 was better for a first time user but after having
used the application a few times before one would be considered an experienced user and option 2 in Figure 8 was then considered better. Option 1 has less cognitive load but more activation peaks. Here are the options the participants were presented with:

![Conversational Flow](image)

**Figure 8:** The participants could choose one of these conversational flows. Option 1 to the left and Option 2 on the right.

Both options from Figure 8 had this following control check, Figure 9, from the assistant to wrap up the payment conversation.

![Control Flow](image)

**Figure 9:** Control flow which followed both previous options.

### 5.3 Prototype

The purpose with this section is to get a deeper knowledge about how to further develop the design of the interaction and conversation and how the prototype was developed. The participants tested the prototype, answered a SUS questionnaire and further open interview questions concerning voice assistants and the trust for voice assistants.
5.3.1 How the prototype works

The prototype is activated as a test application through Google Assistant and invokes by users either saying "Hey Google, Talk to SEB Invoice", or type "Talk to SEB Invoice". The tests start after the invocation step meaning that the test started from the step that the assistant says "Hello what can I help you with today?" (except in Swedish). From this point the user, which in the test scenario is to pay an invoice, can either just start off by telling that they wish to pay a bill, but they can also continue their sentence "I want to pay a bill..." with more information like the company name or/and the plusgiro/bankgiro number, the amount, the date and Optical Character Recognition (OCR) number. It is optional how much information the user wants to provide, the application is trained to be able to take in all the information, but company information, amount, date and OCR is required. This means that if the user says "I want to pay a bill to Lensway" the system has registered in the company name but has to ask for the amount, date and OCR afterwards since they all are required in the same intent. When all the invoice information has been collected the users can either decide if they want to check the information that the assistant has collected or trust that the assistant has understood it all correctly and confirm that the assistant can go through with the payment. If the user selects the first option, to check the information, the assistant will give out the company information, amount to pay, which date for the transaction and the OCR number all in one message, followed by asking if the user wants to change anything or continue with the payment. The users can now decide if they want to change anything (or everything) or if they want to confirm that the information is correct and approve the payment. To change the information the users say what information they want to change and the assistant will ask for that information again. After telling the information again the assistant will tell the user what they have changed and what the assistant has changed it to. The users can now still decide if they want to control everything again or confirm the payment. When confirming the payment the assistant will let the users know that the payment has now been put into the system.

The prototype activates through a simulator via Google Dialogflow but can also be accessed through the creators Google account as a test application. The prototype works both as a chatbot with a visual conversation but is only tested as a voice application. The result of how the conversation between the user and the application went is shown in Figure 10 and 11. The grey chat-bubbles represent what the user says and the white bubbles represent what the system replies to the user.
5.3.2 SUS questionnaire

This section covers the result from the SUS questionnaire combined and calculated, questions shown in Section 4.3.2 System Usability Score questionnaire. Figure 12 shows the scores from the 9 test users combined resulted in a score of 83.3. Since this score is above average, according to Figure 7, this prototype would be considered to have a grade B or a corresponding Good Usability but maybe not Great.
Figure 12: Each participant’s answer on each question has been calculated and together they are combined into a final grade of 83.3.

### 5.3.3 Interviews

This section covers the interviewees’ answers, opinions and suggestions for the prototype. The questions asked can be found in Section 4.3.3 Interviews.

Figure 13: Responses to what the participants’ previous experiences of voice assistants was.

From Figure 13, the 33.3% which had a Google Home at home, had a little more experience than the others. 22.2% had never used a voice assistant before this test even if they had Siri in their iPhone but had never tried the service. 44.4% remaining had used voice assistants a few times before. Mostly they have played around a few times with Siri or another voice application. Examples of tasks they have tried were calling a friend through Siri or tried...
talking to their GoPro.

However, all test users have previous experiences from a voice user interface. From e.g. talking or pressing options on the phone when for example calling medical centre, customer service, booking travel tickets and so on. Those type of voice interface were not so popular according to the participants, but they can still be easy to handle, as long as the interface provides a confirmation.

Opinions about the prototype

This section covers the interviewees’ overall opinion regarding the prototype.

The prototype got the grade "easy to use" from all participants, most of them thought that this would be easier to use than paying an invoice by typing in the information to the smartphone, and it would be a nice feature to use if the user was doing something else at the same time e.g cooking, cleaning or driving etc, something that has the users’ hands occupied. However, if one is to pay lots of bills at the same time, two out of nine interviewees thought that it might go faster to use the computer, but that it depends on how the user handles the conversation, meaning how many inputs the user provides at the same time, one or multiple inputs creates different paths in the system. It at least requires that the user has used this voice assistant payment before.

The control reading created the most insecurity with the test users since they felt that the system read out the information a bit too fast. The participants felt that the OCR number on eighteen digits was too hard to collect in a fast speech pace even when having the numbers in front of the user. Seven out of nine thought that the reading of the OCR was done too fast, the remaining two users could recall the given OCR och read at the same time as the assistant read it out making it no problem to understand. Two participants said that the control information could be divided into two messages. Besides the control of information, the users felt that since there was no visual interface the system was still good and that the flow was simple and that is was easy to follow the questions and given instructions since it was performed step-by-step.

Beneath follows some quoted comments from the participants concerning the prototype:

“It might have gone easier if sitting next to the computer but the advantage with voice is that I would be able to do something else at the same time.” - User 2.

"Knew exactly what to do because I just did what the assistant asked of me" - User 3.

“She sounds a bit robot-like, but it’s good that the assistant does not sound completely human either because then you might think that the assistant is better than you think” - User 4.

"Do not know how seniors would handle this - I think they could experience a problem” - User 2.

"Very long delay for the assistant to process what the user have said and to move on to the next step/question." - User 8.

"Very nice if one does not want to go and get the phone just to pay an invoice but instead to pay with the voice to the Google assistant, but the scanning-functionality is very good, so that is a very big competitor for this scenario with doing invoice payments.” - User 4.
“Does not feel like a problem to want to pay several invoices at the same time because at the end the assistant asked if there is anything more it can help me with so I could add the next invoice then. It probably could go faster like this too than typing, at least if I enter multiple inputs at once that just one information after each question.” - User 9.

Suggestions for improvements on the prototype

This section contains the interviewees’ opinions and comments summarized on how they would like the prototype to be further developed.

User 9 said that as a first time user it might be necessary with onboarding, a welcoming phrase telling the user what is required for the assistant to make a payment, and by that the user will also know what to expect and get a mental picture of how long the interaction might take and what is asked of them. This type of onboarding could also influence the conversation to go in another direction than how the users have interacted with this prototype during the tests. It could give the users an idea for telling more information at once instead of only one after each given question.

To prevent the user from feeling insecure when the assistant does not understand the user an idea, which several interviewees suggested, was to use clearer descriptions on how the user should answer the questions/ and say things. At least a clearer description would be necessary the second time the assistant asks the same question. Another tip for preventing the user sensing a lack of control or feeling insecure could be to get feedback about the receiver of the invoice right away. Example: “You have created an invoice payment to —, how much do you want to send?” just so the user will know that the assistant has perceived the right (or wrong) company information. Another opinion was that it should also be possible for the user to be able to control the entered information at any time. The assistant also needs to be well trained and to be able to handle any type of input and any type of intent at any time, according to the majority of the interviewees. If asked to change one specific information the system should only change that specific thing, and then be able to change something else if the user wants, otherwise the conversation should move on. The system should not change everything if the user requests to change only one thing. More training, fewer bugs and easy navigation to move around in the application and ask the assistant control questions is required according to participants number eight. There should also be limited delays, at least faster between users’ answers and the next question, according to the same participant, number eight.

Since seven out of nine interviewees felt that the control reading was read out too fast. The general tip the users said might help solve this problem was that the assistant would read back the control message in the same manner and pace as the user had told the assistant. A few of the participants experienced a problem with the assistant’s pronunciation. They thought that the assistant had a too robotic pronunciation and that it dragged down the whole experience when pronouncing a word incorrectly. One participant said that it sometimes felt like the assistant did not register the full stops after each sentence, that it did not notice when a new sentence started. However, when the assistant only asked one question or said only one sentence like “how much money do you want to send?” the talking speed appeared fine. Participant number four found he missed something in this prototype and that was to be able to add a message to the sender and/or the receiver for the specific payment.

If performing one invoice payment through voice as the prototype is built right now there
would be no problem to do many at the same time as long as the user can say all the invoice information at the same time, that would go even faster than typing it in, according to participant number nine. Two out of the nine participants felt that a larger amount would not feel any different than doing small ones, while other would maybe want to listen to the control message one extra time or that the control message has to be slowed down or divided into two control messages.

An extra feature that was recommended by one of the participants was something to be able to access in the settings in a hypothetical application, to be able to choose own settings for the assistant. Like politeness, jargon/slang, dialect, gender and more. Also, be able to choose how robotic the assistant should sound.

Opinions about trust and security in voice assistants

This section contains the interviewees’ overall opinion about trust and security in voice assistants and not just about the tested prototype.

*It’s good that the assistant does not sound completely human either because then you might think that the assistant is better than you think.* - User 4.

Two of the participants felt insecurities from not understanding or withholding information about the technology behind the system. *“Can someone try to disguise as me? Would the assistant be able to hear the difference between me and someone that sounds similar to me? If I would know if there is a voice recognition I would feel safer. I want to know more about the technology behind, what kind of safety measurements the systems would use.”* - User 2.

While other participants felt that if this would be a voice application that runs live right now, there would be a lot of safety measures behind to prevent those type of things. Those users did not think that knowing the technology behind the system would be an issue when using these type of applications.

“For the actual payment, one must trust that the service and the bank are doing right. You can trust the application/service and Google that if the service is live then it should be safe and otherwise it is Google Home/the bank’s fault. The first time you might want to double check the transaction at the bank application so that it actually became a payment and once you see that is has been placed you will continue to trust it. I feel like it works like this: Google fills in the invoice and sends it to the bank, it’s Google home’s fault if the bank does not receive the invoice and the bank’s fault if the invoice has entered but has not been paid.” - User 1.

Bank errands through voice interaction

This section contains the interviewees’ thoughts about using a voice assistant with overall bank errands, not just invoice payments as the prototype’s use case concerns.

The overall opinion was that the users did not want the assistant to reach out to them and blurt out information. Participant number seven claimed that one way for the assistant to let the user know that something new has happened is that the assistant can introduce that it has something to tell and then let the users decide when or if they want to know more, through a notification. An example of this kind of notification could sound like this: “this
What would be required for users to use voice banking business in public

This section contains the interviewees’ thoughts and comments concerning using bank business errands through voice assistants but performed in public situations.

Using a voice assistant in public close to other people is however something that the participants felt reluctant to do. More specifically in situations where it can disturb other people, e.g. on the tube, in the office etc. In these situations, they would also be reluctant to talk on the phone since it might disturb the people around them. What the interviewees thought would have to be required for them to use this in public would be if the assistant is the one who initiates the interaction. The assistant has to be the one to lead the conversation forward by asking the user questions which allows the user to only confirm as a short reply. Most participants also desired that the conversation between the assistant and the user has to sound human-like so it would appear for other people as the user is talking to another human. This would improve the feeling of comfort and they would not feel as awkward to talk to the assistant in public while still wearing earphones. Most users do simply not want to give the assistant the invoice information, they want the assistant to tell them when it has found an invoice and ask if the user wants to pay that invoice. It is also necessary that the assistant can comprehend and understand what the user says in every kind of words or sentences.

When the users were asked how they would like the conversation to behave if they would be forced to use a voice assistant in public situations they wanted only a small notification that the assistant requested the users attention and then as soon the users had earphones and confirmed that it was okay for the assistant to continue, so no one else could access the information, then it would be fine.

Normally none of the participants would pay bills in public, but when discussing other bank errands such as transferring money from one account to another, internal accounts or to someone else’s account, that would be something they could maybe be open to. That would, in that case, require that the user could use only small catch phrases like “top of my account” and the assistant is to know from which account to move money from and to which account to sent money to.

Added value

This section covers discussions about where voice assistants add value to the user.

When discussions about added value came up and why the participants felt an interest in assistants the overall response was that they are easy to use, and in situations where both hands are occupied or it is easier than finding and using their smartphone. Also in situations when being lazy and sitting on the couch at home not wanting to go to the phone or computer but instead just tell the assistant that the user wants to make a payment. One participant said that it reduces friction for commonly done tasks and therefore adds value. Voice can also make a boring task as paying an invoice less dull and also much easier. There are also many
benefits if this would work as the requirements in Section 5.4.2 especially for people with disabilities to be able to pay by voice. The disabilities mentioned was if one was to break both arms and if a person is visually impaired or paralyzed. The users also feel that using one’s voice should probably go faster than typing something in, otherwise, there would be no profit from using voice compared to existing way to perform payments.

Another option, by participant number two, was to be able to Swish money to someone through voice, like "send 200 SEK to Bengt" and the assistant then knows who Bengt is or ask "Which Bengt do you mean, Nilsson or Andersson?" and then confirm through voice or face recognition or mobile bankID. If it would be through voice recognition a message or notification would be required that 200 SEK has been sent to Bengt Andersson.

![Figure 14: Responses to the question if the participants would want to use this type of prototype.](image)

Participants answering "Maybe" from Figure 14 motivates their answer with that the assistant needs to be smarter, more trained and a lot quicker otherwise they would rather use mobile phone or computer for invoice payments. Following participants answering "No" does not feel ready for voice interactions overall.

5.3.4 Observations and errors in the interaction

Sometimes the prototype was unable to fully comprehend what the user said when this occurred the test leader was forced to calmly tell the user to repeat itself to the prototype. This caused an interference with the interaction between the user and the system but was necessary to keep the test going. Observing the users when answering the system’s questions they often sat hoping that the system would have understood what they had said. Often the system understood correctly but sometimes it did not and the user sighed and repeated with a clearer articulation, and sometimes they tried to say something in a completely different way holding their breath hoping that the system would now understand. For participants of the tests, the system understood the user completely with no errors making the user feel happy about him/herself and positive about the system.
6 Discussion

The purpose with this chapter is to discuss and analyze the concept and the developed prototype and relate it to the problem statement.

6.1 Result

It was very divided between opinions on voice control from all participants from the workshop, mockup and prototype tests. Some had completely fallen for the technology while some were very sceptical about talking to a gadget. The voice assistant is most useful in scenarios where it can simplify and reduce friction in everyday living and make multitasking easier. Examples from the workshop are quick invoice payments and transfers between accounts/people, control and get information about purchases and reminders of payments. Voice can make tasks possible to perform even though the users’ eyes or hands are not available, this also makes voice a feature that would create opportunities for people with visual impairment.

The results of the workshop show a clear indication that the users do not want to be the one to lead the conversation but still being the one with the control, only telling the assistant yes/no answers confirming the actions the assistant is asking what the user wants the assistant to do. The assistant is to behave like a true lifelike human assistant would do. The assistant would reach out to the user and present that an invoice had been received along with presenting where from. The user would then be able to choose to ask more about the invoice or confirm that the assistant would go on and pay the invoice. A confirmation would, on the other hand, be nice to have after the payment has been done, and most would like to have that confirmation visually shown or to be able to find either in the bank or as a confirmation email. There was a strong agreement that the users would not want to speak out loud a 20 digits long OCR number and would probably not want it read back to them as well. About 50% of the participants felt quite sceptical in the usage of voice assistants, the thought of talking to a “thing” could only provide an awkward feeling. The participants in the user workshop all mentioned, some more passionately than others, that it was important that the interaction would be quick and simple. Some also perceived voice control as something innovative and intriguing, that could encourage exploring.

Even though the users from the workshop said that they did not want to tell the assistant the invoice’s information it felt important to investigate why they did not want to do that and if they would be forced, how would they want the interaction and conversation to behave? Wanted to find pain points, find out why they felt that way and test to see if there is a way to solve it.

The result from the mockup tests followed in the same direction as the workshop had. But in this case, the users were forced to decide between two conversational flows in which
the user decides when to talk to the assistant but then the assistant leads the conversations by telling the user what the system requires to be able to complete the task, in this case, payment of an invoice. The two different options were to test the recommended guidelines to not overwhelm the user with too much information all at once. The first option only asked for one required piece of information at a time at the time out of all required pieces of information that the system needed to complete. The second option told the user what all the required information was for the system to be able to complete, meaning the company information, amount to pay, due date and OCR number. This forced the user to remember all the required pieces of information. The users wanted the conversation to move along in a slow manner to not overwhelm them but still be able to choose what information to say, if they wanted to mix the order or tell more than one input to the assistant, the system would still have to be able to handle that. The two flows that the users chose by is shown in Figure 8 and there was an agreement between the users that the first option felt good as a first time user but the second option could be better fitted to experienced users making the process easier to speed up with less need of guidance on the information required to say. Another option which only was discussed was that as an experienced user the second option could feel over-explained and unnecessary, that the user would rather want to start the conversation by telling all the invoices information instead of the assistant to start onboarding with the required information for using the application. This shows that if the user knows, as in this use case, how to pay an invoice and what information usually is required this recommended guideline does not always need not be followed precisely. Although a recommendation is still to not provide with too much information, a rule of thumb is 7 plus/minus 2.

The results from the prototype tests together with the SUS questionnaire with followed interview questions was interesting. Generally, there were many things the participants agreed on but it was also some things that stood out. There was a strong agreement in that the user wanted the assistant to be able to find the invoice’s information just as a bank app would with an e-invoice where the system collects the receiver, amount and OCR number and which date the final due date is. This would be very convenient. The overall emotion the users experienced when interacting with the prototype was that it was fun to use. But when it came to discussing how reliable and trustworthy they thought it was, even when no errors had occurred, the answers were often that they felt a small lack of security of how much they thought they could trust the system. This had a lot to do with how the conversation is designed and how much training the prototype was put through, but it was also a need for understanding the technology behind it. The overall usage of the service was thought to be simple and easy but it was the reading back from the assistant about the invoice information that made the user feel insecure. Something very clear in the test of the prototype which many participants commented on was the speed but also fallback intents. A good fallback response from the system is important for the user to not feel insecure or experience the sense of losing control. The assistant also needs to be a bit smarter in the sense of always be able to control previously given input, to be able to always handle control questions from the user about what information it has collected.

The fact that all participants had previous experience in paying a bill and knew what information was important could have contributed a lot because the users knew to expect what the assistant would ask for, making it feel a bit more secure.
6.2 Connections to theory

A general recommendation was to stick to a maximum of three things that the user needs to handle and remember [54]. This is how the prototype was designed, by offering the users to only provide one input at a time in the prototype. Except with the control check from the system, the assistant put out all the provided information like company information, amount, due date and OCR number. This was made on purpose to get a reaction on how the users would prefer it. Most liked the way it was but that the assistant just needs to present the information at a slower pace. Some participants would have liked that information to be divided into two or to present at least the company name before. What the users felt more concerned about was if the OCR number was collected correctly.

The dialogue was considered polite but not too formal. The language and words seemed to suit the use case and the fact that the users were supposed to talk to a bank application. The prototype used a conversational manner and was designed to imitate how a conversation could behave if a user could phone a banker to send the money to the invoice’s receiver. It only used confirmations in terms of when the user answered the assistant’s questions it moved on to ask for followed information indicating that the previous information had been received. Sometimes that was not always good, some users wanted to know right away what information the assistant had registered. However, some users said that they liked not hearing what the assistant had registered right away but rather afterwards when the user wanted to check the information it had registered.

Technology Acceptance Model (TAM) felt appropriate to use for evaluation of what value using voice assistants could bring to the users and if there is a value with voice for this use case. The found value is in those situations where the users have their hands and/or eyes occupied and also might even have an urgent invoice to pay, or with users with disabilities.

6.3 Method

The implementation of this thesis used a Service Design methodology to foremost investigate the problem and formulate the fundamentals of the essential facts which was done in the literature study presented in Chapter 3 the Theory-section but also in the workshop.

The workshop was interesting and fun to perform and the participants did a good job with their role-plays. What was searched for in this workshop was to understand the participants from a user’s point of view, how the users would wish the conversation to behave without being provided with an example of a conversation based on existing guidelines.

The fact that many of the participants did not immediately believe in the idea of using voice when paying invoices since many did not receive invoices through e-mail, when they bought something online they preferred paying right away. If listening to the participants in the Mockup interviews for how they would like to use a voice assistant in bank errands the use case could have been developed into something else. The participants could have been even more interested and curious about what value voice interactions could give. The decision to go on with the chosen use case of paying an invoice was because of the feeling of higher complexity but it still did not require many steps to perform comparing to "only" transferring money to a different account.
The approach to, as a test leader, manually activate when the prototype is supposed to listen and to when in need (when the user was not heard by the system) ask the users to repeat themselves interfered in the sense of finding how the users and the prototype would handle and act for these type of errors. Since it was an early stage prototype and the test was performed through a chat-based prototype a better way would have been to test without the test leader to activate the assistant to listen to get a full evaluation. Instead, the focus lied with the flow of the conversation between the user and the assistant. If the user sometimes was not heard could have to do with the limitations of the hardware. Since the prototype was played out from a MacBook Pro but the computer did not face the test person but instead the test leader meaning that the microphone was not directed towards the user.

The prototype's final grade from the SUS questionnaire was an 83,3 on a scale from zero to 100. The reason for the SUS questionnaire was not only to get quantitative data but to also discuss the questions with why the users answered the way they did. An interesting example was when the participants answered the first question "I think I would like to use this product often" some marked down a four which is equal to a pretty strong agreement. But when asked afterwards discussing their answers and the question their answer could sometimes be "No, I do not think I would use this product right now". Some users in the other hand answered in the reversed way, marking down a three as an answer to the question but during oral discussion answered "Yes I would definitely use this". During the evaluation of the questions some also answered that they would maybe want to changed the way they had marked before in the questionnaire, but the marked answer stayed as the had answered the first time. When performing a SUS questionnaire it was a good idea to evaluate the questions to get the users to fully understand the questions and think about their answers. This could have affected the resulting grade and the fact that it was the person who developed this prototype which also led the tests could have affected the test users to answer more positive than what they otherwise would have, just to be kind to the test leader.

The interview phase included good discussions but could have had an even deeper focus on the feeling of trust and not specifically for the prototype but to also talk about the users relation to trust overall, what their definition of it would be, especially concerning in technology.
7 Conclusion

What was to be investigated in this thesis was if it is possible to find any further guidelines specifically for designing trust with VUIs for bank interactions and to get an understanding if the market is ready for a voice application. The problem statement of this thesis was how to design a voice user interface to make the user feel safe while using it and how to design trust in a voice user interface. The already existing guidelines from the voice assistant developers such as Google and Amazon provide a very good and thorough foundation for the design of a conversational user interface. When it comes down to designing a conversational user interface application for banks and in this case invoice payments, I have found some points which the interviewees thought would decrease the users feeling to trust the application.

In the banking industry, there are already banks preparing for the area of voice with making it possible to interact with one’s bank account through Amazon’s Alexa or send money to a friend through Google Pay. As the voice functionality is growing it will also become more socially acceptable, and since it is growing it is also important to follow the trends to stay alert and listen to the customers’ needs.

These guidelines that follow are found from the result of the tests and discussions revolving designing to increase the feeling of trust when using a voice assistant for bank errands, especially transactions.

- **Welcoming/onboarding phrase:** Present the application so the user knows what to expect from the application and conversation, what is expected from the user and how long time the conversation might take. This could provide the user with a mental map on how many things they have said and how many things they had left to say.

- **Informative and describing responses:** Be very informative in fallback responses if the assistant does not understand something the user says, e.g. the OCR number of an invoice. The second time the assistant might need to tell the user more precisely how to pronounce the date for the assistant to understand. This would lead the users back in control since they would then know exactly what is required from the user. This prototype wanted to test and see how the users provide their input and to keep it open, that the voice assistant can understand no matter how the user provides it, just as in a conversation between two humans. But many wanted it to be described to them, giving them the control to know what to say for the system to understand them. A user can provide the system with the OCR number in numerous ways. But sometimes to be sure that the system collects it correctly the assistant can guide the user, as an example, to provide the OCR number in a slow pace, and still keep it conversational and turn-taking.

- **Personalize the voice application to improve user experience:** To allow the user to create a personal touch to the assistant as dialects, slang, gender, talking speed, how
formal/informal it should sound, how robotic etc.

- **Copy the users’ way of speech:** The system should try and copy the speed of how the users’ speak and use the same words and format. E.g. the user input for the due date is June 10th, then the assistant should answer with the same format: June 10th, instead of 2019-06-10. Same with the OCR number, if the user wants to provide each number individually such as one, two, three, four, five, six etc. or if the user wants to say it like twelve, thirty-four, fifty-six etc.

- **Decide own catchphrases and synonyms:** In settings to the voice application, be able to decide which account is the savings and which is the primary account so it is easy for the user to talk with short sentences and for the assistant to still understand what the user wants to do. Make it possible for slang expressions to work as synonyms and train the intents well. There are many kinds of words to use instead of, for example, *cash* in Swedish. Make sure to provide many synonyms or else provide the opportunity for the user to provide synonyms they like. It is still a good idea to design the application and tell the user how to talk to the assistant but just as a personal touch and if the user has a better idea on how to say something, let the user be able to decide settings.

- **A graphical user interface as a complement:** for users interacting with the bank application as they currently do. Allow voice through the existing application as an alternate way to insert payment information, search for financial advice or other bank services. Speaking is still our natural way of communicating and it does not require the user to know exactly what they are looking for but to still understand their true intentions.

Whether the market is ready for this type of applications the answer would be "Yes and no, maybe in a few years". For people to start using and to socially accept voice assistants it is necessary to provide use cases for the customers. This requires that the industries start developing voice assistants. It has however already started, companies such as H&M Home and SJ are just two of them. These are good starting points which I think will increase the interest and usage of them overall. With more applications developed with a voice interface as a complement to a previous graphical one can also reach out to more users. As these applications start to integrate more into our everyday lives it will get more socially acceptable to use these types of interactions. When introducing these applications to potential users it might be important to talk about how errors are handled, how the technology works and how accurate the system is.

### 7.1 Future work

A suggestion for future work would be to do another round of iteration. To make changes from the result into the prototype and do another round of iterations. What is important besides the designing for trust is of course that the system behind is also worthy of the trust. Therefore the development of the system, that the system actually does what it says it does, is almost more important for the usability than how the conversation between the user and the assistant works.
Another suggestion would be to test with similar methods as the one in this thesis but with other types of payments or bank errands through voice and see how users react, behave and what their opinion is. It would also be interesting to try to develop a prototype through voice where the assistant leads the conversation and has the information about the invoice just as if the system had received an e-invoice through the bank application, just as most users from this thesis preferred it to behave. Try and find if this is more useful and preferred by users than the prototype that has been tested in this thesis.
Bibliography


