Improving user trust towards conversational chatbot interfaces with voice output

RAMÓN BURRI
Improving user trust towards conversational chatbot interfaces with voice output

Author: Ramón Burri  ramonb@kth.se
Supervisor: Gabriel Skantze  gabriel@speech.kth.se
Examiner: Olov Engwall  engwall@kth.se

Degree Project in Information and Communication Technology, Human-Computer Interaction and Design

October 17, 2018
Improving user trust towards conversational chatbot interfaces with voice output  Ramón Burri

Abstract

This thesis investigates the impact of the voice modality on user trust in conversational chatbot interfaces. The assumption is that trust can be increased by adding voice output to a chatbot and by a higher quality of a used text-to-speech synthesis. The thesis first introduces chatbots and the concept of conversational interfaces then defines trust in an online context. Based on this, a model for trust and perceiving factors for credibility, ease of use and risk is defined. An online experiment is conducted where participants run through conversational scenarios with a chatbot while varying the voice output. Followed by a survey to collect data about the perception of the trust factors for a scenario with no voice and two scenarios with different speech synthesis qualities. To analyse the ordinal data from the survey the "Wilcoxon signed-rank test", a nonparametric statistical test, is conducted to compare trust for the voice output types. Results show that adding the voice output modality to a conversational chatbot interface increases the user trust towards the service. Furthermore, the assumption that synthesis quality has an effect on trust could not hold true because the results are not statistically significant. On this basis, the limitations of the used methods are discussed and suggestions for further research are proposed.

Keywords

Chatbot · Conversational interface · Human-computer trust · Speech synthesis · Trust · Voice
Contents

1 Introduction .............................................. 1
   1.1 Problem Statement ..................................... 1
   1.2 Voice Modality ........................................ 1
   1.3 Hypotheses ............................................ 3

2 Related Work ......................................... 3
   2.1 Conversation .......................................... 3
   2.2 Trust .................................................. 5

3 Trust model ........................................... 7
   3.1 External Factors ....................................... 8
   3.2 Perceiving Factor: Credibility ......................... 8
   3.3 Perceiving Factor: Ease of Use ........................ 9
   3.4 Perceiving Factor: Risk ................................ 9
   3.5 Trust .................................................. 9

4 Evaluation ............................................. 9
   4.1 Experiment ............................................ 9
   4.2 Experiment Design .................................... 11
   4.3 Chatbot implementation ............................... 11
   4.4 Survey ............................................... 12

5 Results .................................................. 13
   5.1 Descriptive Statistics ................................. 13
   5.2 Nonparametric Statistical Test ....................... 14

6 Discussion and Conclusion ......................... 17
   6.1 Ethical Implications ................................... 19
   6.2 Research Limitations and Suggestions ............... 19
   6.3 Conclusion ............................................ 20
1 Introduction

This thesis investigates the effect of the voice output modality on the user trust in conversational interfaces. The goal is to add voice as a modality to a text-based chatbot interface and evaluate if the user trust towards the interface and the service changes. The added chatbot voice is a synthetic computer generated speech. A speech synthesis can consist of various qualities, ranging from human-like voice to the one of a synthetic robot. Therefore, this thesis is further evaluating the impact on trust for different synthesis qualities.

1.1 Problem Statement

The usage of conversational interfaces and in particular chatbots are increasing in online services [1]. The conversational interface is providing the primary means of interaction with chatbots, messaging apps, and virtual personal assistants are therefore necessary. McTear defines a conversational interface as “a conversational interface, also known as conversational user interface (CUI), provides the front-end to a chatbot or virtual personal assistant, allowing the user to interact with the app using speech, text, touch, and various other input and output modes.” [1]. A chatbot is a software application which performs automatic tasks. The idea is that a chatbot mimics a conversation with a real human and is therefore a conversational interface. Since trust is an important factor in human conversations and conversational interfaces [2], the assumption is that the importance applies as well in an online context. The global chatbot market is expected to reach USD 1.25 billion by the year 2025 [3]. In Nguyen’s article, it is shown that 80% representatives of a study in the countries United Kingdom, Netherlands, France and South Africa are already using chatbots or are willing to adopt them by 2020 [4]. Moreover, the article says that Within the global chatbot market, approximately 45% of end users prefer chatbots as the primary mode of communication for customer service inquiries. Thus, conversational chatbot interfaces are on the rise.

Trust is an important factor in service design since an increase in trust can result in higher usability. More usage leads to more benefit in transactional activities and increased user engagement [5]. The user interface should convey the message of a trustful service [6]. Being able to detect if voice output has an impact in a conversational interface could change the perspective of future service design, enabling chatbot providers to enhance or adapt their interfaces for a more trustworthy interaction.

1.2 Voice Modality

This thesis focuses on the addition of the spoken voice modality in a conversational chatbot interface and neglects the impact of voice input. The reason is that the technical implementation or addition of a voice output service is simpler to achieve than a voice input processing [7]. Moreover, the user is an actor in a human-computer conversation and embodying the chatbot with an own voice is a first step to investigate the impact of voice output on trust. Santen and Pear describe the two speech output techniques: Text-to-Speech (TTS) and pre-recorded speech [7]. Playback of pre-recorded speech provides a high-quality speech output since a human speaking can be recorded. Pre-recording every sentence of a system can be problematic due to memory limitations for sentence sizes or the high effort of applying changes to pre-recorded audio files. A TTS system creates an artificially generated voice of human speech. This process is also called speech synthesising. TTS is synthetic and perceived as less natural and the resulting speech is, therefore, less acceptable. A synthesis which is similar to a
human voice is considered of higher quality and depends on the TTS technology used and how natural the language is perceived. This thesis focuses on a TTS approach with different synthesis technologies for generating voice output of a chatbot. Pre-recorded speech is limited for chatbots since not every answer can be recorded beforehand. TTS speech instead allows generating speech from any text-based user input. TTS can be separated into two parts: text analysis and waveform synthesis [8][9]. The text analysis normalizes the text in a first step to detect boundaries of sentences for a natural speech flow. Numbers, dates and abbreviations need to be converted to be later processed. In the second step, the text is analysed for phonetic and prosodic features to find natural pronunciations and phrasing. This information is needed to generate the waveform for the speech output. Converting the symbolic linguistic representation form the text analysis into sound is called synthesizing.

There are various speech synthesis technologies on the market or in research. Most of them can be divided into the following categories: concatenative synthesis, formant synthesis, articulatory synthesis and deep learning based synthesis. The technologies differ mainly in how speech is parameterized for storage and synthesizing [10]. Concatenative synthesis technology is a very common technique since it can produce natural-sounding synthesis speech. The synthesized speech is created by concatenating pieces of recorded speech units stored in a database. Speech synthesis systems are using different storage sizes for speech units. Smaller units consist of distinct speech sound or gesture which is called phone or pairs of them which are called diphones. If the system stores phones or diphones, it can deliver a large speech output range but may lack clarity. Also entire words, phrases or sentences can be stored in unit selection synthesis for higher output quality but increased storage size. This is mostly used for specific domains as in weather or schedule announcements. Formant synthesis, on the other hand, does not use human speech samples at all. Rules are used to describe the resonant frequencies of the vocal tract. Formant synthesis uses additive synthesis and an acoustic model to create a speech output. Adjustable parameters such as fundamental frequency, the degree of voicing and the intensity of the source signals are varied over time to create the artificial speech waveform. Formant synthesis quality can sound unnatural because it is difficult to estimate the vocal tract model and source parameters. But compared to concatenative technologies, the formant synthesizers are usually smaller programs since they do not need to store speech samples. Furthermore, the speech can be reliably intelligible which is an important factor for users with visual impairments. Articulatory synthesis is based on the physical model of the human speech organs. Simulating acoustic functions of the vocal tract to mimic lips, jaw, tongue and velum is used to create the speech output. With simulation of the air flow through the vocal tract based on pressure, vocal cord tension and the relative position of the organs an articulatory model can be reproduced. This technology is mostly used in research because of the limitations of parameters for modelling the system and the need for an accurate three-dimensional vocal tract representation. Deep learning based synthesis uses deep neural networks which are trained on recorded speech data. With the trained set the technique models directly the waveforms to achieve a more realistic-sounding, human-like voice. A popular deep learning technology is called WaveNet [11][12]. WaveNet’s neural network is fed with real waveforms. As they pass through the network, it learns how to describe the evolving audio waveform over time. The trained network is capable of creating new waveforms at more than 16’000 samples per second. The ability to generate raw waveforms means that it can model any kind of audio, including languages with accents, was one of the reasons that Google chose WaveNet for it’s Google Assistant [13]. The quality of the synthesized speech output is judged by its similarity to the human voice and its ability to be understood clearly. This thesis compares the quality of two different TTS technologies. The WaveNet technology from Google [12]
Improving user trust towards conversational chatbot interfaces with voice output  Ramón Burri

is used and compared with the standard unit selection concatenative TTS synthesis Festival Text-to-Speech from the centre of speech technology research at the University of Edinburgh [14]. These two technologies are chosen because they are representatives of a common TTS approach (concatenative) and a modern, popular TTS system used in Google’s commercial products (WaveNet). Validating if voice output and synthesis quality have an impact on user trust will provide a solid base for further research on voice input or the combination of both.

1.3 Hypotheses
The thesis consists of two alternate hypotheses and one null hypothesis. The first hypothesis is that adding voice output to a conversational chatbot interface increases the user’s trust towards the service. The second hypothesis is that the synthesis of the voice output has an impact on the trust towards the service. The null hypothesis states that there are no differences when measuring trust between a text-based chatbot without voice and one with voice output.

- $H_1$: Adding voice output to a chatbot interface increases the user’s trust towards the service.
- $H_2$: The user’s trust towards the service can be increased by increasing the voice synthesis quality towards a more human like voice.
- $H_0$: There is no difference in trust towards the service between a chatbot with no voice and a chatbot using speech synthesis as voice output.

It is expected to see differences in trust for the ease of use, credibility and risk. Adding voice should increase the credibility, the ease of use and decrease the risk of using the service for the user. This assumption should hold true to validate the hypothesis.

2 Related Work
To answer the question, if voice output increases the user’s trust towards conversational interfaces, research about related work was conducted. We can distinguish between the two categories of conversation and trust. With the context of the conversation, we take a closer look at the characteristic of chatbot interfaces, how they are perceived and which role they are playing in a human-computer conversation. Along with the TTS system, the synthesis quality is analysed. The second part of the background research is dedicated to trust. We define trust and see its implications and the role of trust in conversational interfaces.

2.1 Conversation
Chatbots are user interfaces which support conversational interactions between humans and machines. Thus, the base of a conversational interface is the concept of a human to human conversation. Conversations can be split into spoken interactions and written interactions. Chatbots make mostly use of written interactions due to their text-based nature but conversational interfaces are capable of spoken interactions as well. McTear describes the term conversation as “informal spoken interaction in which news and views are exchanged and in which one of the main purposes is the development and maintenance of social relationships.” [8]. Thus, an interface mimicking this behaviour should act as similar to a human being as possible. Humans engage in conversations to perform actions. Actions like asking questions, making promises
Improving user trust towards conversational chatbot interfaces with voice output Ramón Burri

or others are based on a request which produces a dialogue [15]. For a question, humans are requesting answers which can be seen as a conversational act [16], leading, in the end, to a conversation. The acts can be broken down to understand what is requested from the other party. Conversations follow in general a turn-taking approach where each participant is taking a turn to talk. A simple view of turn-taking in conversations is that one participant talks and the other participants wait until the first one finishes to take over. Turn-taking, in reality, is more complex and is based on rules which are mutually accepted [17]. But even with such rules, humans overlap or interrupt other participants in spoken conversations. In case when a machine takes the role of a participant it is important to identify when it is their turn to talk. Human-machine conversations take a different form of turn-taking because of speech recognition technology requirements. A machine needs to determine when the person began to speak and when the utterance is complete with only sound cues [8]. In all the cases a machine should be prepared for the user’s expectations and interaction patterns. Some systems allow experience users to interrupt a talking machine when they are familiar with the prompt. This technique is called barge-in and helps a user to take over the lead in spoken interactions. For text-based interactions as used in chatbots, the start and end of a turn could be determined by the message submitted by the user. Since text-based messages usually get send in chunks of phrases and sentences the machine can wait until receiving a message before taking the turn. Furthermore, a machine could check if the user is typing which could indicate that they are not finished with their turn. Besides turn-taking, there are many factors which have to be handled in a conversation by a machine. Contextual awareness [7], including previous turns and context of the conversation, grounding [18], conversational repair for miscommunication [19] and linguistics are only a few of the factors. Additionally, we have to take into account that humans can be aware that they are not interacting with another human and therefore change their behaviour as shown by Hill’s research [20]. Thus, there are many challenges for interfaces to provide experiences as engaging and as realistic as a human to human conversation.

A common interface for conversations is the chatbot which takes a user’s input to create a response. A chatbot is a text-based interface which can have an additional voice user interface on top of it to make use of speech input and output. In chatbots, the conversations are mainly triggered by the user’s input and consists of smaller task-oriented dialogues [8]. Exceptions are when the chatbot asks a question to continue a conversation. Each user input is matched against a set of patterns to create an output. Chatbots are popular for service-oriented applications such as in customer support, process-oriented tasks like ordering goods or information retrieving [1]. These elements are generating a lot of interest in the field of ubiquitous computing because it allows having chatbots and chatbot services on mobile and wearable devices. Products like IBM Watson [21], Google Actions [22] or the Microsoft Bot framework [23], to name only a few, provide cloud-based chatbot systems based on machine learning and Artificial Intelligence concepts [24] [25]. Being able to train a chatbot’s responses and reactions will generate a personality which can be used for artificial agents or avatars.

Therefore, conversation based interfaces are on the rise and with the rapid development of technologies, the interfaces will become more usable and effective. In the right context and application, conversational interfaces are popular tools to support human-computer interactions. A voice supported conversational interface consists of different speech and language technologies which form a flow of interaction for a user. This flow or chain of processes can be separated into components as seen in figure [1] such as automatic speech recognition (ASR), natural language understanding (NLU), dialogue management (DM), natural language generation (NLG) and text-to-speech synthesis (TTS) which operate together for a system to understand the user voice input and generate a voice output [19]. In this chain, the ASR and NLU are responsible for
Improving user trust towards conversational chatbot interfaces with voice output  Ramón Burri

handling the user’s speech input and the NLG and TTS for generating natural speech output. The DM, on the other hand, takes care of the dialogue flow logic based on context and discourse. This chain of components and technologies is a common architecture for speech systems as well as audio supported conversational interfaces. This thesis focuses mainly on the second part of the process chain which is responsible for the voice output. One of the factors which has an impact on the usability is the synthesis quality of TTS systems. A good quality in TTS systems is perceived when the generated voice sounds natural and is intelligible. One of the natural language elements is the prosody. It is a concept that describes the rhythm of speech, stress patterns and intonation which have an impact on the naturalness of the perceived speech [10] [26]. In the end, a high-quality speech synthesis can optimize for naturalness and intelligibility or focus on only one element depending on the use case.

![Components flow of voice supported conversational interfaces](image)

Figure 1: Components flow of voice supported conversational interfaces, adapted from Skantze [19]

Pauletto’s paper describes that the technology is serving as a proxy for human emotions which it does not and cannot feel [27]. Pauletto’s statement “When humans complain that a synthesizer sounds ‘robotic’ or ‘alien’, the problem is not that there is no emotion contained in the signal. The problem is that the signal is expressing the wrong emotion, leading to confusion and miscommunication.” shows that the intonation of the voice synthesis is important for the perceived quality. Thus, for chatbot services in which trust is an important factor, a low-profile prosody scheme could be suitable instead of an excited speech with varying tempo.

2.2 Trust

A major part of this thesis is to elaborate a way of measuring user trust. To measure trust we first need to define what trust is. Corritore et al. say, for example, that online trust, which is the trust in the context of the digital world [6]. There is only little research about online trust in particular but extensive research about offline trust, trust in the “real world”. Trust is a research topic from the fields of philosophy, sociology, human-computer interaction, marketing, psychology and many more. This variety of use of trust makes it difficult to find a definition across fields. And even within a specific field, there is a lack of agreement and focused effort [28]. The research about trust is, therefore, a collection of various concepts and models. Thus, Corritore et al. propose that the research on online trust can be build of the body of work, of
Improving user trust towards conversational chatbot interfaces with voice output  Ramón Burri

the offline trust researches. They argue, that both, the online and offline worlds have much in common when it comes to social interaction concept of users. Especially in the field of human-computer interaction, the literature on trust uses the base research from the offline world. Since the understanding of online business builds on the definitions of offline trust, we can assume that the concepts are valid in both scenarios.

Blois research shows that there is not always a clear distinction between trust and trustworthiness [29]. In spite of the distinction, a logical link is present [30]. Trust is seen as an act of a trustor which places his or her trust in some object. This is valid despite the fact, that the user’s trust can be proven to be well placed or not. Trustworthiness, on the other hand, is the object of trust and defined by its characteristic. Furthermore, trust is not the same as reliance. A person can rely on another person without trusting him or her [29]. Trust itself is split into many cognitive cues which are formed by a person on the perception of an object which is to be trusted [31]. Thus, the key concepts that are used in the literature to define trust are risk, expectation, confidence, vulnerability and exploration [32] [33] [28] [34] [35]. The research about the concepts shows that offline trust is multi-dimensional and varies in many forms. For the online trust, we focus on the concepts which are relevant to the chosen trust model which is used to evaluate trust.

In the field of human-computer interaction, trust can be seen as an intervening variable that mediates user behaviour with computers [36]. The user’s perception of the expertise of system, including its functionality, forms the trust towards such a system. It is shown that errors in the system’s functionality play a factor in the loss of trust [37]. Users are able to recover and rebuild trust when errors do not repeat but the recovery of trust takes a long time to prior gains. This leads to the question of when people do trust computers? In the first place, we can define that trust is only needed in situations of vulnerability and risk [38]. Thus, the user’s expertise and confidence towards a system impact their risk assessment. Experienced users, which are aware that a system provides full error recovery will be more likely to trust it than others. Thus, we have not a single answer to when users trust a system but we know that various concepts as expertise and risk impact user trust. This is used throughout the thesis to measure trust not as a single variable, but rather as a collection of concepts.

Trust is also researched a lot in the context of electronic commerce (e-commerce) where a buyer-seller relationship is important. Various pieces of research show that trust plays an important role for business to customer services [39] [40] [41] [42] [43] [44]. They focus on different criteria on what increases or impacts user trust in an e-commerce relationship such as the age of the relationship, the expertise of the company selling goods or brand recognition. Each study states the importance of trust in a business to customer relationship to increase the business’s revenues. Chatbots are getting more commonly used on e-commerce platforms for product searches, travel bookings and promotions companies for companies to gain a competitive advantage and profile their customers [1]. Since e-commerce businesses start using more and more chatbots and want to increase trust, the result of this thesis could provide guidance. Trust is also a crucial factor in the context of financial services. Hoffman et al. show how introducing new products and services rely heavily on trust [5]. In the same paper they make an example where introducing new computer technology can lead to loss of trust when the provided service is poorly implemented. Lee mentions that the information quality is more important than the interface quality of a services [45]. A reason for this could be the factor of security in those service interactions. Users perceive a higher risk in financial services due to the fact that they need to provide or retrieve sensitive data over a system which they have to trust [46]. To improve the quality of the system Zhou recommends to increase the structural assurance of services which was the main impacting factor for trust in his research. This proves
that context and risk generated by it influence user trust.

In the context of conversational interfaces, we can find studies for relational agents, artificial-intelligent agents and other humanoid representations. Qiu and Benbasat state that the presence of a 3-dimensional avatar does not increase user trust towards a conversational interface [47]. The representation of a human does not affect trust but the conversational strategies do. In Bickmore and Cassell’s paper they discuss the effect of conversational strategies like small talk on trust and the fact that interface agents could use those techniques to establish and increase user trust [48]. Research about computer-computer or computer-human interaction for agents was conducted by Taddeo to show when and how computer systems can trust users [49]. For this thesis, we focus on human-computer trust and user trust towards a chatbot interface.

Jensen et al. prove that the voice modality has a big impact on social cooperation and that voice has a positive effect on trust between the participants [50]. This outcome is aligned with Cockburn’s philosophical discussion that trust is part of every conversation [2]. Thus, we can conclude that conversations and therefore conversational interfaces like chatbots affect trust in one or more ways. As well we can say that lack of trust is a barrier for people to engage in e-commerce services which results in reduced service usage. This shows that increasing user trust in services can benefit customer to business relationships and in the end revenues. Increasing trust is also important for the user itself to lower perceived risk towards a service and interaction with a chatbot to provide a better user experience.

3 Trust model

Evaluating trust or measuring the impact of voice output on trust in a conversational interface can only be done when the factors of importance are known. To find and classify the factors for this thesis evaluation, a theoretical model for trust is needed. Several researchers have created theoretical models for trust and trust relationships [51] [52] [53] [6]. Lee and Turbans’s trust model for consumer internet shopping and Kim’s trust formation model for B-to-C e-commerce are both related to an e-commerce research but are identifying the technology and its familiarity as one of the key factors for trust. The technology factor is also present in a conversational chatbot where a voice modality is present. The model of Herbig et al. focuses more on credibility and reputation instead. Credibility is also one of the key factors in Corritore et al.’s trust model beside ease of use and risk. All the models take various external factors into account which influence the previously listed internal factors to define trust or a trust relationship.

The theoretical model which shares factors with all other models is Corritore et al.’s trust model [6]. The model itself is generic enough to be used in the context of this thesis and provides a structure which can be evaluated. The trust model is divided into external factors, and perceiving factors “credibility, ease of use and risk” which then result in trust. An adapted trust model can be seen in figure[2]. The following sections map the factors of the trust model to the context of conversational interfaces and voice output.
3.1 External Factors

Corritore et al. also define external factors as aspects of the environment, both physical and psychological, surrounding a specific online trust situation [6]. In the case of conversational interfaces, we could list the user’s familiarity with chatbots, messaging application and the technology behind it [54] [55] [51] [42]. The design of the interface, the interaction flow and consistency of the design, as well as branding and reputation of the service provider, can be seen as external factors [56] [43] [57] [58] [55].

The list of external factors is endless which can be accounted for the use case of the chatbot interface. The focus of this thesis lies on the addition of a modality and therefore the external factors are less relevant for the evaluation of the hypothesis. Since external factors can be perceived differently by users, the main goal is to reduce and control them during the evaluation.

3.2 Perceiving Factor: Credibility

Credibility is divided in the model in 4 parts: honesty, expertise, predictability and reputation which is based on the offline trust models [59] [60] [28] [61] [55] [42].

Interesting for the modality voice is the focus on predictability because the user can not be sure if for example entered sensitive data will be said out loud by the chatbot. The offline trust research suggests that predictability is a trustor’s expectation that an object of trust will act consistently based on past experience [54] [62] [55]. When an object of trust acts in a predictable manner, credibility is lent to that object. Thus, it can be interesting to see if voice or synthesis quality impacts the predictability of a service. Within credibility, we also have the expertise. Expertise is the perceived level of professionality of a system or service. Adding voice output to a chatbot can have an impact on how the user perceives the expertise level of the service [31]. The thesis’s assumption is that adding a good synthesis quality will increase and a bad robotic synthesis decreases the user trust.
3.3 Perceiving Factor: Ease of Use

The factor ease of use focuses on the technology acceptance by the user [63] [64]. It assesses how easy it is to use the service and chatbot to achieve the user’s task. It is interesting to measure if the voice modality increases the ease of use. Ease of use can be split into accessibility and user experience. Accessibility focuses on the aspect of how a user can access and solve his task in terms of contextual support or clear instructions, whereas the user experience is more related to the overall experience of using a chatbot service. Therefore, in the case of conversational interfaces, those two factors need to be evaluated in the context of ease of use.

3.4 Perceiving Factor: Risk

Risk is an important factor in trust. If an interaction or use of service shows no sign of risk then there is no need for the user to trust it [38] [65]. Risk is the likelihood of an undesired outcome and is therefore closely related to trust [66] [67] [68]. In the evaluation of this thesis hypotheses, the factor risk is measured by conducting an experiment in different contexts. It is assumed that based on Luhmann’s theory [38] the risk is higher in chatbot services where sensitive data is requested from the user and no trust issues if risk is absent.

3.5 Trust

According to Corritore et al., trust is related to perceiving factors credibility, ease of use and risk. Increasing trust means, therefore, increasing credibility and the ease of use and reducing the risk towards the service. In this thesis, we are going to break down the factors for trust in predictability and expertise for credibility, accessibility and user experience for ease of use and risk.

4 Evaluation

To test the hypothesis an experiment is conducted with several chatbot conversation scenarios. The participant’s experience and trust, evaluated by the perceiving factors, are measured after the experiment in form of a survey. The experiment and survey try to measure trust by the identified factors: predictability, expertise, accessibility, experience and risk. Trust can then be compared among the different voice types.

4.1 Experiment

The experiment is an online task where participants work through three different interactions with a chatbot. Each interaction consists of a scenario and a voice type which is randomly chosen for each participant. The voice types for a scenario are either no voice, text-based only, or one of two TTS synthesis technologies along with the text. The scenarios represent parts of three common chatbot service interactions with different conversation choices and questions about the participant’s data. The experiment consist of three scenarios:

- Restaurant Giraffe: A conversation in which the user has to do a table reservation in a restaurant. The user has to choose a preferred day and time for the reservation as well as how many guests the reservation should be done. Most of the choices are offered by the chatbot and no personal data is requested from the user to complete the scenario (See figure 3a).
Improving user trust towards conversational chatbot interfaces with voice output  Ramón Burri

- **TomorrowBank**: A conversation in which the user applies for a mortgage loan for an apartment. The conversation mimics a web form which requests the user to select the amount of the loan, his or her yearly income and financing strategy. Sensitive data such as the yearly income and a telephone number need to be entered to complete the scenario (See figure 3b).

- **RealTech**: A conversation in which the user goes through a checkout process of an online shop. The user has the choice to either pay by credit card or by invoice. Both choices require the user to enter valid credit card data for the payment or an address to which the invoice will be sent (See figure 3c).

The scenarios differ in context and severity of personal data requested by the chatbot. With three different scenarios, the bias due to environmental factors or participants behaviour is reduced. Additionally, this setup helps to validate the hypothesis since they have to hold true for all scenarios. Example conversation flows for each scenario can be seen in figure 3.

![Example conversation flows of the three experiment scenarios](Figure 3: Example conversation flows of the three experiment scenarios)

Besides the scenarios three different voice types are used during the experiment:

- **No Voice**: The conversation is only text-based and no audio is used for the chatbot.

- **WaveNet Synthesis**: The chatbot’s questions and answers are spoken out loud with the Google TTS service using WaveNet [12]. Google’s modern deep learning technology is used to generate voice which is similar to the one of a human.

- **Concatenative Synthesis**: The chatbot’s questions and answers are spoken out loud with the Festival TTS service of the centre of speech technology research at the University of Edinburgh [14]. Their unit selection method for synthesis is used to generate voice which should be of high quality but distinguishable as synthetic.
The voice types can be split into two groups for further analysis. First we have the no voice and voice group and second, we can compare the two different synthesis qualities.

4.2 Experiment Design

The experiment’s goal is to find out if the voice type causes a change in the user’s trust. Thus, this experiment consists of an independent variable voice type ($VT$) which is controlled through the experiment and a dependent variable trust ($TR$) which is the average of the 5 measured factors predictability ($PR$), experience ($EX$), accessibility ($AC$), user experience ($UX$) and risk ($RI$).

The experiment is structured as a basic design because only one independent variable is investigated. A within-group design is chosen due to the fact that every participant is exposed to each condition of the independent variable and the isolation of individual differences. The randomisation of the scenarios and the associated voice types are introduced with the within-group design due to the smaller sample pool required. Two problems which come with the within-group design are the learning effect and fatigue which could bias the result as stated by Lazar in “Research methods in human-computer interaction” [69]. The learning effect can be controlled by the experiment’s randomized order and conditions. Fatigue, on the other hand, is controlled through the fact that the experiment itself only takes 5 to 10 minutes to run and therefore no fatigue effects are expected.

The participants start on an informational website with detailed instructions. It prepares the participant for audio playback and that a survey will follow the experiment. Due to the fact that the conversations ask for sensitive user data, a note is displayed to ensure the participant that none of the entered data is stored in any way or used outside of the conversation window. This is important since the experiment does not force a participant to enter their correct data but also not prevents it. Allowing the participant to run the experiment on their desktop computer or mobile device increases the external validity. All measurements and collected data are coming from the survey and not the chatbot interactions.

4.3 Chatbot implementation

The chatbot used in the experiment was created as an own Javascript, CSS and HTML web application. The participants are first lead to an introduction page which described the purpose, conditions and process of the experiment. From there on the participant starts with the first random chatbot interaction. At the end of each interaction, the participant is redirected to either the next random scenario or after completion of all scenarios to the survey.

The Festival TTS and the Google WaveNet TTS services require different technologies for the implementation. Since the chatbot conversations allow only basic interactions, we decided to record the chatbot conversation parts as audio files. All possible chatbot interactions are therefore played over the respective TTS service and recorded. Having a bunch of audio files has the benefit to pre-load them in JavaScript. Pre-loading all the files for each scenario provides an interaction which is not influenced by network quality. Controlling the network factor is important since the participants can run the experiment also on a mobile device. The chatbot then played the already loaded audio file whenever the chatbot’s turn is. When the participant decided to interrupt the chatbot, then the audio will be stopped. This implementation needs only to keep track of the position inside the conversation but does not need to store any user data.
Table 1: Survey questions

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Factor</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PR</td>
<td>I expected that the chatbot will say out loud what I entered as input.</td>
</tr>
<tr>
<td>2</td>
<td>EX</td>
<td>The chatbot service was professional.</td>
</tr>
<tr>
<td>3</td>
<td>AC</td>
<td>The voice output helped to understand the context of the conversation.</td>
</tr>
<tr>
<td>4</td>
<td>UX</td>
<td>It was clear at which point the chatbot was speaking.</td>
</tr>
<tr>
<td>5</td>
<td>RI</td>
<td>I was confident entering sensitive data.</td>
</tr>
<tr>
<td>6</td>
<td>PR</td>
<td>It was clear what happens with the entered data.</td>
</tr>
<tr>
<td>7</td>
<td>EX</td>
<td>I was confident that “TomorrowBank” provides a reliable chatbot.</td>
</tr>
<tr>
<td>8</td>
<td>AC</td>
<td>It was clear what answers I can provide.</td>
</tr>
<tr>
<td>9</td>
<td>UX</td>
<td>The conversation was structured and easy to follow.</td>
</tr>
<tr>
<td>10</td>
<td>RI</td>
<td>I felt that TomorrowBank’s service was trustful.</td>
</tr>
</tbody>
</table>

Survey questions for the banking scenario. Two questions to measure each perceiving factor.

4.4 Survey

The survey participation group is based on random sampling. Since the target user group consists of English speaking persons only which are older than 18 years old and have access to the online experiment. Therefore, reaching a census would be out of the scope of this research. The sampling followed a self-selected non-probabilistic-based survey. This means that selected participants receive the link to the experiment and are able to invite others to participate. The nature of this sampling allows inviting various groups of people among various demographics.

The survey itself is structured in three parts. The first part consists of general questions about which devices are used to do the experiment, the age of the participant and how experienced they are with chatbots. This part helps to get a better picture of what the sampling group consists of since it is not a closed and controlled experiment.

The second part questions the experiment scenarios. There are three equal questions blocks for each scenario and voice type combination the participant has played through. To ensure that we have the right scenario and voice type combination multiple surveys are created and depending on the randomized run the participants get the correct responding survey. The questions are all in a Likert scale format with 5 possible answers from 1 (strongly disagree) to 5 (strongly agree). The participant can select option 3 (neutral) if they do not have an opinion. All questions are formed positively, short and written in clear English to reduce any biases. Each scenario has 10 questions where always two are used to measure one of the perceiving factors as seen in table[1]. Having two questions for each factor reduces the error rate of misunderstood questions or validate them. Since the survey follows after the participant played through all three scenarios, an image of a sample chatbot conversation is shown above the questions. This helps the participant to remember the conversation context.

The last part of the survey is more general and asks the users to sort the three scenarios by trust and usability. Furthermore, questions like which voice type was preferred and if they could identify the difference between human and a synthesized voice are asked. All questions in this part help to get more detailed feedback to the experiment.
Improving user trust towards conversational chatbot interfaces with voice output  Ramón Burri

5 Results

The experiment was conducted during a period of two weeks and collected responses from 76 participants. Each participant played through all scenarios and voice types combinations. The data set got cleaned up and the data normalized. 10 questions ($Q_n$) are used to measure the different factors of trust $TR$. The factors are predictability ($PR$), experience ($EX$), accessibility ($AC$), user experience ($UX$) and risk ($RI$). They are aggregated and presented as seen in the formula[1]. We assume that all five factors have the same weight to calculate $TR$.

$$TR = \frac{PR + EX + AC + UX + RI}{5}$$

The data values of each of the factors are between 1 and 5. Values below 3 show a negative and values above 3 show a positive correlation to the factors. This means that we have an ordinal scale of values to measure trust and it factors. The question $Q_5$ for risk is positively formulated and will, therefore, have a high value when high risk is perceived. Aligning this value with the others is done by inverting it, that 1 is a perceived as a high risk and 5 as a low risk.

5.1 Descriptive Statistics

A statistical analysis is conducted to understand the nature of the data set and its variation. Running tests on the first part of the survey gives a better picture of the sampling group. We know that the majority, 63% of the participants, are between 25 and 34 years old and only 6% are older than 45. Two third of the participants run the test on a desktop computer and one third on a mobile device. The result of the question about the experience with chatbots is interesting. 13% of the participants use chatbots on a weekly basis or more and 65% on a monthly basis or less. That means that every fifth participant never used a chatbot before. We can assume that the sampling mostly consists of people in their 30s and 40s and have used chatbots rarely or never before.

![Figure 4: Trust value for each voice type compared by trust factors and scenarios. Each bar shows shows the standard deviation of the values.](image-url)
Comparing the mean of the TR values of different scenarios shows that the trust is always above 3 and therefore in a positive correlation. Even split by each individual factor as shown in figure 4a makes it clear that all values are above 3. Furthermore, the conversations which use voice show a higher value than the text-only conversations. The participants showed the biggest trust towards the Google TTS synthesis with WaveNet. This means that the general trust is increased by 7% when adding voice to a chatbot and increased by 9.5% if using the Google TTS synthesis.

Comparing the trust by voice type and scenario as seen in figure 4b we get a clearer picture. In the scenario of the restaurant table reservation, TR among the voice types is almost identical. There is only a slight increase by adding voice to the conversation. Looking at the other scenarios, the differences by voice type are significantly higher. The trust level of the text only option stays at 3 the VT with WaveNet is at least half a point higher. Another interesting observation is that in the banking and online shop scenario, the trust increases even more if the synthesis is closer to the human voice. The perceived factors also determine how big the differences are between the two synthesis types. For the factors, accessibility and user experience, the synthesis quality only varies within 0.1 on the trust scale. The greatest difference is detected in experience and predictability, followed by risk. We can say that credibility has an impact on perceived synthesis quality, since both factors EX and PR are from the same trust model category.

5.2 Nonparametric Statistical Test

The descriptive statistical analysis clearly showed a difference in trust when adding voice to a chatbot conversation. But comparing the means of two or more conditions does not validate the hypotheses. A statistical significance test is needed to evaluate the effect of the independent variables. Since the difference in means can occur by chance the probability of it should be calculated and controlled. To control type 1 errors during our experiment which can occur when the null hypothesis assumption is wrong but it isn’t, we strive for a very low significance level also known as p-value in our test. A widely adopted p-value threshold in human-computer interaction is 0.05. Thus, if our tests show that the significance level is lower than 0.05 means that the chance of mistakenly rejecting the null hypothesis is below 0.05% [69].

Testing the significance of our hypothesis is done with a nonparametric statistical test named "Wilcoxon signed-rank test" which is an alternative for the dependent-samples t-test [70] [71]. Because the data collected from the survey is ordinal are not scaled by intervals one of the main assumptions for a parametric test is not met. The "Wilcoxon signed-rank test" requirements are that we have a null hypothesis where the median difference between pairs of observation is equal to zero. The participants had to run a chatbot conversation for each voice type which allows us to observe the difference. Furthermore, the "Wilcoxon signed-rank test" can be applied in a within-group experiment design where only one dependent variable, in our case TR, is analysed which has an ordinal nature like a Likert scale has.

The "Wilcoxon signed-rank test" does not use a singular formula but instead a test procedure with several steps. First, we define N as the sample size of our measurements which are the number of pairs. A sample consists of a pair of no voice and a voice measurement. Therefore, we can say $i = 1, ..., N$ where $x_{1,i}$ and $x_{2,i}$ are the data points. The null hypothesis $H_0$ would show that the difference between the pairs $x_{1,i}$ and $x_{2,i}$ does not follow a symmetric distribution around zero. Then, we calculate the differences between the repeated measurements and the absolute differences. This is done for $i = 1, ..., N$ with $|x_{2,i} - x_{1,i}|$ and the sign function $\text{sgn}(x_{2,i} - x_{1,i})$. From the result we exclude all pairs with $|x_{2,i} - x_{1,i}| = 0$ and get a new sample size $N_r$. 

14
Improving user trust towards conversational chatbot interfaces with voice output Ramón Burri

"Wilcoxon signed-rank test" excludes ties because they do not support the evidence in favour of the $H_0$ and only contain the lack of evidence of them. The next step is to order the cases by increasing absolute differences. For all remaining cases, we assign their relative rank $R_i$. In case of tied ranks, the average rank is calculated. With the sum of the signed ranks, we are going to calculate the test statistics $W$ (See formula $2$).

$$W = \sum_{i=1}^{N_r} [\text{sgn}(x_{2,i} - x_{1,i}) \cdot R_i]$$  \hspace{1cm} (2)

Since the $H_0$ of the "Wilcoxon signed-rank test" says that there is no difference between the pairs, we can assume that $W$ follows a specific distribution with an expected value of 0 and a variance of $\frac{N_r(N_r+1)(2N_r+1)}{6}$. Thus, for a two-sided test, we can reject $H_0$ if $|W| > W_{\text{critical},N_r}$ for sample sizes smaller than 10. For bigger samples a $z$-score can be calculated with $z = \frac{W}{\sigma_W}$ where as $\sigma_W$ is defined as in formula $3$.

$$\sigma_W = \sqrt{\frac{N_r(N_r+1)(2N_r+1)}{6}}$$  \hspace{1cm} (3)

With calculated $z$ we can reject $H_0$ if $|z| > z_{\text{critical}}$. The critical value can be found in Wilcoxon’s paper “Probability tables for individual comparisons by ranking methods” [72] by a required significance level. A limitation of the "Wilcoxon signed-rank test" is that observations are discarded where the differences between the pairs are zero as stated by Pratt [73]. Pratt provides, therefore, an alternative which incorporates the zero differences. For data on an ordinal scale, the modification of the "Wilcoxon signed-rank test" seems more robust.

Figure 5: Pretest for the “Wilcoxon signed-rank test”. The boxplot shows the medians compared by voice type ($VT$)

For this case we can run two different tests, each comparing the no voice with a voice value of the independent variable $VT$, once with the "Wilcoxon signed-rank test” and once with the modified version of Pratt. A pretest of the median comparison can be seen in figure 5. From the boxplot, each box has a midline representing the median of your data, with the upper and lower limits of the box being the third and first quartile (75th and 25th percentile) respectively. The boxplot shows that the median does differ from each voice type which supports our alternative hypothesis $H_1$ and $H_2$. 

15
To calculate the actual p-values with the "Wilcoxon signed-rank test" we used the 
\texttt{wilcox.test} function of the R statistical computing software with a confidence interval of 
0.95 [24]. Comparing the mean of no voice to WaveNet shows that the calculated p-value is 0.0001142. Whereas the no voice with concatenative comparison results in a p-value of 0.03107. For Pratt’s alternative, we used the \texttt{wsrTest} function with a confidence interval of 
0.95 [25]. The resulting p-values are 0.00006495 and 0.03216 and are aligned with the original 
version of the "Wilcoxon signed-rank test" as shown in table [3]. For hypothesis $H_2$, comparing 
the two synthesis types, results in a p-value of 0.07388 and a Pratt’s p-value of 0.09038.

Both values, from the no voice to synthesis voice type test, are lower than 0.05 which is an 
indicator for rejecting $H_0$. But we learned that for a bigger sample size the z-scores must be 
bigger than $z_{\text{critical}}$ to reject $H_0$. We want a confidence level of 95%, therefore looking up the 
$z_{\text{critical}}$ value results in a value of 1.960 [22]. The z-scores we get from the Wilcoxon test are 
$-3.858286$ for the first and $-2.156134$ for the second test.

The z values are higher than 1.960 and therefore the null hypothesis $H_0$ can be rejected. Note 
that the WaveNet synthesis with the Google TTS has a high confidence in increasing trust than 
the Festival’s concatenative TTS. On the other hand, $H_2$ does not show a significant p-value or 
z-score. Thus, we see an increase in trust in our experiment but the change from concatenative 
to WaveNet is not significant and the hypothesis can not be approved.

With the resulting z values, we can further calculate the effect size for the signed-rank test. 
The effect size is a quantitative measure of the significance of an effect. Cohen classifies 
the following correlation coefficient levels $r$ for social sciences to show if the effect of the 
correlation is low, moderate or high [76] (See table [2]).

Table 2: Cohen’s effect size

<table>
<thead>
<tr>
<th>Effect size</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.10</td>
</tr>
<tr>
<td>Moderate</td>
<td>0.30</td>
</tr>
<tr>
<td>High</td>
<td>0.50</td>
</tr>
</tbody>
</table>

Correlation coefficient effect size by Cohen [76].

The correlation coefficient $r$ can be calculated by dividing the absolute (positive) standardised 
test statistic $z$ by the square root of the number of pairs as seen in formula [4].

\[ r = \frac{\text{abs}(z)}{\sqrt{N_r}} \] (4)

The two correlation coefficients are 0.4425758 for the first z-score and 0.2473255 for 
the second. Comparing the values we can state that adding the voice with WaveNet to 
the conversation has a moderate to a high effect on trust whereas the effect of adding the 
concatenative voice is only low to moderate.

If we break down TR into the individual factors we can see in table [3] that all factors except 
PR are significant for $H_1$. UX among all of the others shows a high effect and significance 
on trust. For $H_2$ the factors show an opposite behaviour. When comparing the two synthesis 
qualities, only the factor PR is significant, similar to the comparisons of the means as elaborated 
earlier in figure [4a].
Table 3: Wilcoxon test result

<table>
<thead>
<tr>
<th>Test</th>
<th>H</th>
<th>Var</th>
<th>VT A</th>
<th>VT B</th>
<th>p-value</th>
<th>p-value (Pratt)</th>
<th>z-score</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H₁</td>
<td>TR</td>
<td>No Voice</td>
<td>Concatenative</td>
<td>0.00011</td>
<td>0.00007</td>
<td>-3.858</td>
<td>0.443</td>
</tr>
<tr>
<td>2</td>
<td>H₁</td>
<td>TR</td>
<td>No Voice</td>
<td>WaveNet</td>
<td>0.03107</td>
<td>0.03216</td>
<td>-2.156</td>
<td>0.247</td>
</tr>
<tr>
<td>3</td>
<td>H₁</td>
<td>PR</td>
<td>No Voice</td>
<td>Concatenative</td>
<td>0.6787</td>
<td>0.8823</td>
<td>-0.414</td>
<td>0.048</td>
</tr>
<tr>
<td>4</td>
<td>H₁</td>
<td>EX</td>
<td>No Voice</td>
<td>Concatenative</td>
<td>0.00323</td>
<td>0.00331</td>
<td>-2.945</td>
<td>0.338</td>
</tr>
<tr>
<td>5</td>
<td>H₁</td>
<td>AC</td>
<td>No Voice</td>
<td>Concatenative</td>
<td>0.01084</td>
<td>0.01071</td>
<td>-2.548</td>
<td>0.292</td>
</tr>
<tr>
<td>6</td>
<td>H₁</td>
<td>UX</td>
<td>No Voice</td>
<td>Concatenative</td>
<td>8.663e-10</td>
<td>8.844e-12</td>
<td>-6.132</td>
<td>0.703</td>
</tr>
<tr>
<td>7</td>
<td>H₁</td>
<td>RI</td>
<td>No Voice</td>
<td>Concatenative</td>
<td>0.04042</td>
<td>0.04625</td>
<td>-2.050</td>
<td>0.235</td>
</tr>
<tr>
<td>8</td>
<td>H₂</td>
<td>TR</td>
<td>Concatenative</td>
<td>WaveNet</td>
<td>0.07388</td>
<td>0.09038</td>
<td>-1.787</td>
<td>0.205</td>
</tr>
<tr>
<td>9</td>
<td>H₂</td>
<td>PR</td>
<td>Concatenative</td>
<td>WaveNet</td>
<td>0.00236</td>
<td>0.00293</td>
<td>-3.041</td>
<td>0.349</td>
</tr>
<tr>
<td>10</td>
<td>H₂</td>
<td>EX</td>
<td>Concatenative</td>
<td>WaveNet</td>
<td>0.05023</td>
<td>0.1265</td>
<td>-1.958</td>
<td>0.225</td>
</tr>
<tr>
<td>11</td>
<td>H₂</td>
<td>AC</td>
<td>Concatenative</td>
<td>WaveNet</td>
<td>0.6654</td>
<td>0.6224</td>
<td>-0.432</td>
<td>0.050</td>
</tr>
<tr>
<td>12</td>
<td>H₂</td>
<td>UX</td>
<td>Concatenative</td>
<td>WaveNet</td>
<td>0.545</td>
<td>0.4825</td>
<td>-0.605</td>
<td>0.069</td>
</tr>
<tr>
<td>13</td>
<td>H₂</td>
<td>RI</td>
<td>Concatenative</td>
<td>WaveNet</td>
<td>0.1257</td>
<td>0.1631</td>
<td>-1.531</td>
<td>0.176</td>
</tr>
</tbody>
</table>

Results of the “Wilcoxon signed-rank test” for trust (TR) and the trust factors.

6 Discussion and Conclusion

At the beginning of the thesis, we saw that trust is closely related to any kind of conversation between humans [2]. The chatbot played the role of a participant in the scenarios and got evaluated by the same trust factors as humans are and showed therefore that trust is as an important factor in conversations between humans but as well between humans and computers. The result of this thesis supports the alternative hypothesis $H₁$ that adding a voice output modality to a chatbot increases the user’s trust towards such a service. As seen in the result table 3 a significant effect is found for the factors expertise, accessibility, user experience and risk when voice is added to a text-based chatbot scenario. Four out of five perceived trust factors can be influenced by the voice modality.

Analysing the factor $RI$, we could say that the addition of voice to the chatbot increases the similarity to a conversation with a human. This does not mean that a user is fooled to interact with a human [20] but rather create an interaction which is more familiar. Furthermore, a reason that the two factors $EX$ and $RI$ are increasing together could be, that experience is influenced by the risk assessment of the user as stated by Luhmann and therefore low risk correlates to high expertise [6] [38]. The added voice can help with accessibility if reading text is difficult or not possible which could be the reason for the factor $AC$ to increase. The factor $UX$, on the other hand, could have been influenced by the survey question Q4 “It was clear at which point the chatbot was speaking.”. The chatbot always started speaking after the text message of the chatbot is shown which could have lead to a negative impact due to confusion on the scenario without voice output.

An interesting fact is that $PR$ has a lower value when adding voice to the chatbot. This could be for several reasons. In the sample group has over 65% participants which use a chatbot once a month or less. This means that probably most of them never used a chatbot with voice before. Following this assumption, the predictability of a chatbot with voice would go down similar to the thesis’s result. Participants which do not know what a chatbot with voice is capable of, have a harder time to predict it.

The result of the mean comparison in figure 4b further shows that the differences in the
restaurant table reservation scenario are almost zero. A possible conclusion could be that
the restaurant scenario did not ask for any sensitive data from the user which would confirm
Luhmann’s theory that without a sign of risk there is no need to generate trust [38]. The only
scenario where participants do not have to enter any personal data has an overall high trust level
of almost 4 which is higher than any other value in the other scenarios. Another reason for this
outcome could be that an added voice to a chatbot had a direct impact on the user experience.
It would explain why the factor UX has the biggest difference when adding a voice output.

The second set of tests for the alternative hypothesis $H_2$ showed almost a mirrored result
when comparing the two speech synthesis. As clearly seen for all factors and scenarios in figure
4a and 4b the Google WaveNet synthesis dominates the concatenative synthesis from Festival.
As discussed earlier the quality of the speech synthesis can have an impact on perceiving factors
and based on the thesis’s results we can see the effect on PR. Besides predictability, we can see
in table 5 that the p-value of 0.05023 for EX is very close to 0.05 and a significant factor.
The fact that both factors are significant or close to it tells that the synthesis quality mainly
affects credibility in the trust model. Corritore et al. state that credibility affects the user’s
perception of risk in an inverted way. If the user has a perception of high credibility, he will
perceive the risk as lower [6]. Therefore, in our experiment, we would say that a more realistic
synthesised voice is perceived as less risky. Or the other way around, that a synthetic, robotic
voice increased the risk. Assuming this is true it would support what Hoffman et al. mentioned
in his research. Poorly implemented systems or in this case synthesis quality can lead to loss
of trust [5]. Nevertheless, the results for TR shows that hypothesis $H_2$ cannot be significantly
supported.

Throughout the experiment, a clear difference could be detected in trust or its factors when
comparing all scenarios. As described in the experiment design section, the idea of testing
different scenarios was to reduce and control biases due to environmental factors or participants
behaviour. Biases could still remain and affect users behaviour as well as the flow of the
conversation. Analysing the individual feedback from the survey showed that participants
struggle in the online shop scenario when they had to enter a credit card number. The entered
credit card number in the experiment is checked against a valid credit card number format.
Participants tried sometimes to enter a fake credit card number which leads to the chatbot
refusing it. The RealTech online shop scenario was the only conversation where input was
checked for validity. Similar to the online shop scenario the mortgage application conversation
did not check against valid input but ask for sensitive data of a participant. Providing income
statements or a phone number was for some participants reason enough to fake them as they
stated in the feedback. This is a valid response in the context of the experiment but it shows that
there was a risk perceived by the participants. This could explain the differences in overall trust
per scenario. Other factors which could have biased the result, could have been the company
names and association a user builds with it or previous experiences for similar cases. It was
interesting to see from the feedback that the participants felt immersed into the conversation
even though they knew that they are in an experimental task with set-up conversations. The
participants have associated trust towards the chatbot and not the service or company behind it.
This could show that chatbots are not perceived as a technical interface but furthermore
as a conversation partner in which human conversation strategies, including trust building, are
applied.
6.1 Ethical Implications

Investigating trust is a highly complex topic and is perceived differently from person to person. This thesis touches a psychological part of the human-computer interaction when quantifying trust in different variables. Thus, interpreting that the results can be applied to every person in the targeted population size can be dangerous. Individual experiences, conditions and context can affect trust or its factors in unexpected ways. Therefore, treating trust as a scientific variable needs to be done carefully. All humans participate in conversations but not everybody is fond of small talk or feeling comfortable with it. Since the addition of voice can increase trust towards an interface, an option to use a text-only chatbot should be available to respect different preferences. Increasing modalities on interfaces as the voice in chatbots are not only affecting the user trust but makes the interface accessible to a bigger user group. People with visual impairments or other disabilities could benefit from an additional modality to interact with. Adding voice can open doors for many users but should be considered carefully in which situation it is suitable.

6.2 Research Limitations and Suggestions

One of the limitations of this thesis is the selected sample of the target population. Most of the participants are coming from friends, colleagues and their connections which biases the sample by the seen factors age and experience. The target population probably consists of a lot of elderly persons with less or none experience with chatbots. On the other side, young people under the age of twenty have another relation to technology and mobile devices and could perceive a chatbot service differently. Besides the participant selection, controlling the experiment execution is limited as well. The introduction, experiment and survey are online and could have been conducted at any time and place since mobile devices are supported. Thus, having external environmental factors or different time constraints may have impacted the result. These biasing conditions could reflect a real scenario of a chatbot service but are hard to control. A laboratory experiment approach with a random sample of the population could support the results. Moreover, having a bigger sample size would allow conducting an experiment in which each participant would run the same scenario three times with random order of the voice types. This could strengthen the result of a "Wilcoxon signed-rank test" since the compared pairs refer to the exact same scenario. However, as mentioned, the sample size needs to be big enough so that the learning effect can be neglected.

Another limitation is the selection of TTS synthesis during this thesis. Only two different synthesis qualities, each of a different technology, are compared. Both, the Google WaveNet and Festivals unit selection method are of better synthesis qualities. To understand the impact of synthesis quality on trust a test with multiple TTS systems of different technologies needs to be conducted. Having a more linear synthesis quality detection could show what the minimum quality is before it has a negative trust impact. It can as well show that TTS technologies are more suitable for chatbot voice output and which detailed properties of a TTS system have the biggest effect on trust.

The thesis focuses on voice output for the chatbot but, as we could see, the chain of processes for spoken dialogue systems also consists of other components. Jensen et al. state that voice as a modality is crucial in building trust in conversations and therefore an extended research on ASR, NLU and DM could show if the investigated results hold true when users can speak to a chatbot [50]. The fact that two out of the three tested scenarios are non-symmetric in terms of interaction should not be neglected. The participant is always typing the answers whereas the chatbot’s response is written and spoken for the two scenarios with voice. This imbalance of
modality, where the participant is typing but the chatbot is talking aloud, could affect the result since the dialogue is not natural.

Based on the result of the thesis, the recommendation would be for application or service designers to plan for a voice modality when designing a chatbot conversation. As the results show, the addition of voice by a TTS engine improves the user trust towards the service as long as the synthesis is of higher quality as the WaveNet synthesis from Google. For use cases where no sensitive data is content of the interaction, like the table reservation scenario, user trust does not change with voice addition. The result supports as well the theory of Corritore et al. [6] that online trust can be based on the concepts of offline trust which allow further research on the topic of trust in an online context.

6.3 Conclusion

Trust is a part of human-human conversations [2] and also a factor in human-computer conversations. Since conversational interfaces like chatbots are on the rise, building trust for users is important to ease the interaction with computer systems and increase the usability and usage of interfaces [5]. This thesis provides evidence that adding the voice output modality to a conversational chatbot interface increases the user trust towards the service. Results show that a high-quality TTS synthesis has a positive impact on the perceived experience, accessibility, user experience and risk by a user. Thus, the hypothesis that trust can be improved by adding voice output to a conversational interface holds true. The results also support Luhmann’s thesis that trust is only generated when there is a perceived risk [38]. The thesis shows that the TTS synthesis quality can have an impact on trust in chatbot interactions but fails to make statistically significant findings. The fact that trust in services can be improved by adding voice output is important for application and service designers when creating chatbot interfaces. Increasing trust can strengthen the relationship with users for businesses and provide a better user experience for increased service usage.

References


Improving user trust towards conversational chatbot interfaces with voice output  Ramón Burri


Improving user trust towards conversational chatbot interfaces with voice output Ramón Burri


Improving user trust towards conversational chatbot interfaces with voice output Ramón Burri


Improving user trust towards conversational chatbot interfaces with voice output  Ramón Burri


Improving user trust towards conversational chatbot interfaces with voice output Ramón Burri


Improving user trust towards conversational chatbot interfaces with voice output  Ramón Burri


