Building and evaluating a voice assistant with formality adaptation to its user

by

Martijn Faes

Snr: 2012824

Bachelor's Thesis

Liberal Arts and Sciences

Major Cognitive Neuroscience

University College Tilburg

School of Humanities and Digital Sciences

Tilburg University, Tilburg

Supervised by Dr. J.M.S. de Wit PDEng, Second Reader: Prof. Dr. E. J. Krahmer

August 2022

ABSTRACT

Even though being an influential factor on trust and other types of user experience in the interaction between the user and conversational software, formality remains underrepresented in both literature and measurement tools. Since it is beneficial to measure formality through audio, rather than through text, in this research a voice assistant with real-time formality adaptation functions is developed and evaluated. Additionally, the effects of real-time formality adaptation on the participant's trust and other attitudes towards the voice assistant are measured. Lastly, this study researches the impact of real-time data collection and processing on the participant's trust in conversational software with adaptive capabilities. Data is gathered through the newly developed voice assistant, as well as through a combination of surveys and an interview. The voice assistant was successfully created and real-time adaptivity was implemented. It was also found that the participants in the adaptive condition generally rated their feelings and trust with regards to the voice assistant as worse than the other two conditions, which can be caused by the inconsistencies in formality that affect the participant's experience negatively. Lastly it was found that gaining knowledge about the inner workings of the voice assistant raised the user's trust.

ACKNOWLEDGEMENTS

Thank you to Jan de Wit for supervising this thesis, and Emiel Krahmer for being second reader. I would like to thank them for their continued support and their passion for their students and their field, as well as for their advice and critiques.

I would like to thank my partner Chloe Jahne for supporting me all throughout this process, for her input and for being a loving partner.

Finally, I would like to thank all of the participants who volunteered their time in order to help me with my data collection.

Table of Contents

1. INTRODUCTION

The technological developments that have occurred in the past decades, or even the past years, have allowed people to engage in an expanding variety of interactions with technologies in their daily lives. At this point in time, users are able to interact with new technologies by using conversational human language. Through natural language processing (NLP) and artificial intelligence (AI), the user is able to have an interaction with the technology that closely resembles conversation between humans. One example of technology that makes use of natural language, and is being used with increasing frequency, is voice assistants. Commonly used voice assistants are Google Assistant, Siri by Apple or Microsoft Cortana. Through their convenient, natural language based nature, users utilize voice assistants for everyday tasks and interactions, like shopping, playing music, setting reminders and more. The number of active voice assistants is growing dramatically, with an expected number of 8 billion by 2023 (Malodia et al., 2021). Voice assistants are among the fastest growing products in consumer technology (Malodia et al., 2021) and are a way for consumers to have a "useful and meaningful" initial interaction with AI-based technologies, further enhanced by advanced natural language processing and machine learning structures (McLean, Osei-Frimpong, 2019, p1). Voice assistants are able to handle complex requests from users and are increasingly preferred over more traditional search engines (Malodia et al., 2021).

The broad implementation of new NLP and AI-based technologies and their rapid integration into our daily lives does not occur without causing frustrations (Brendel et al., 2020; Roshan-Ghias et al., 2020). When conversational AI aims to simulate the feeling of a person-to-person interaction, there are certain expectations that exist for the interaction with the technology (Chaves & Gerosa, 2021). Svikhnushina and Pu (2020) pointed out several of these expectations, relating to the AI giving intelligent emotional responses, adapting the conversational style to that of the user, and lastly, high task performance. As the user interacts with the conversational AI, with the notion that the expectations of the interaction will be met, Chaves et al. (2019) state that when these expectations are not met, frustrations will arise for the user.

These frustrations, following expectations not being met through improper performance of the software, can cause a decrease in feelings of trust (Chaves et al., 2021), and consequently a greater social distance to the conversational agent. Even though in the literature the relation between social distance and trust is often described as trust being a product of decreased social distance, there is support that this relation works the other way as well (Albanes, Scheepers & Sterkens, 2014), leading to the belief that increased trust can affect social distance in a beneficial way. In a study on human-robot interaction by Banks and Edwards (2019), it was noted that when the social distance is smaller between a human and a robot (embodied conversational agent), the human's willingness to integrate a certain technology into their lives and the number of social interactions with these technologies might improve. From this it can be theorised that a greater social distance would mean that people could be more reluctant to adopt the conversational AI. While increased social distance scores are thought to be related to negatively associated experiences, e.g. lack of trust, lower perceived anthropomorphism or decreased motivation to use the technology again, support for increased positive attitudes like anthropomorphism through decreased social distance was also found (Banks & Edwards, 2019).

With trust allegedly being an important factor for the acceptance of new technologies, many studies have pointed towards the role of perceived usefulness in building trust with conversational

AI. It is theorised that, if the technology is expected to perform the desired task well, raising user satisfaction in the process, a more trusted bond is established between the user and the AI conversational agent (Malodia et al., 2021). Increased trust is vital when the user interacts with conversational AI, especially when the application of it is involved in critical fields, like healthcare and education (MacArthur et al., 2017). Emphasizing the importance of gaining trust through improved user satisfaction, Chaves et al. (2021, p. 3-4) noted that in the field of chatbot research the user's satisfaction also directly correlates with the agent's use of "Appropriate degrees of formality".

Adopting appropriate degrees of formality in a conversation, and specifically in the context of interacting with a conversational agent such as a voice assistant, has been underrepresented in the literature. Based on the proposed relation between formality and satisfaction, and thereby trust in chatbots (Chaves et al., 2021), in this paper it is theorized that in the process of designing a voice assistant, it is of importance to allow the system to adapt to the user's level of formality. There are only a few existing implementations of automatic formality analysis and adaptation, and these are often only evaluated using metrics rather than studies involving participants that interact with the system (e.g. in Lai, Toral & Nissim, 2022; and in Niu, Martindale & Carpuat, 2017). Quick adaptation on sentence-level input not being sufficiently developed is an urgent AI problem. For this reason, a large focal point of this paper lies on the creation of a model that can both measure formality, as well as adapt in real-time to the level of formality that the user employs. Since there is no other model like this, this paper introduces a first attempt at implementing a way to automatically identify formality levels, and adapt a voice assistant's output accordingly. This study thus acts as a pilot for a model that is designed to answer the research question: How can we implement real-time formality adaptation for a voice assistant? (RQ1)

Considering the possible role of adaptation of formality levels in raising user satisfaction and building trust with the user, and in support of testing the model created for RQ1, this paper also poses the question: What is the impact of dynamic formality adaptation on people's attitudes and trust towards interactive conversational software? (RQ2)

In order to provide adaptive capabilities of software, a significant amount of user data needs to be obtained and processed. This collection of personal data is known to cause reluctancy in users to interact with conversational AI (Chaves & Gerosa, 2021; Svikhnushina, Pu, 2020), as well as non-adaptive systems. For example, due to security concerns, users might be held back from sharing their data, especially with regards to disclosing important personal details with certain technologies (Malodia et al., 2021). Terzopoulos and Satratzemi (2019) also point out that even when the concerns about data protection do not originate from distrust of the conversational software, there might be distrust with regards to the company providing the product or service to treat their data with respect and process or keep it in a manner that will warrant safety for the user. Nonetheless, for a consumer to interact with new technologies with adaptive features, it is impossible to provide the service without collecting a certain amount of data from that person. Since there are existing frictions between desire to use technologies and the collection of personal data (Lavado-Nalvaiz, Lucia-Palacios & Pérez-López, 2022), it is important to further investigate the impact of data collection on trust in conversational AI. To achieve this, this study aims to investigate the following: What is the impact of real-time data collection on trust towards interactive conversational software with adaptive capabilities? (RQ3)

2. THEORETICAL FRAMEWORK

Formality

Formality and informality are hard to define. A large part of the literature (e.g. Levin & Novak, 2009; Kaneyasu, 2022; Rüdiger & Mühleisen, 2022) uses Irvine's (1979) definition of formality and informality. As described in Kaneyasu (2022, p.5), formality can be understood as "conformity to social conventions of linguistic and nonlinguistic behaviors, positional identities, and a central situational and topic focus", while informality can be understood as "any practice that disrupts or undermines conformity to social norms or the systematic organization of social activity." In the Dutch language, one basic example of the contrast between formal and informal language, is the use of 'u' and 'je' (which both translate to 'you' in English). 'U' is used to address a person directly in a formal manner, while 'je' is used for the same purpose in an informal setting.

Understanding formality and informality according to Irvine's (1979) definitions, it can be seen that they both have definitions that are very reliant on the (social) context of the situation. This means that a person will continuously change their level of formality, as the situations in which they find themselves change. Just like interactions between humans, when utilizing a voice assistant for interactions with different people, the social context will differ per conversation the VA encounters, and may even vary within the same conversation. In order to suit these changes in social context in the most appropriate manner, it is important to explore the influence of formality adaptation by a VA on the experience of the user of the VA. Among other goals, this study aims to explore the appropriate manner of this formality adaptation by a voice assistant during a conversation with a person.

NLP, AI and formality measures

The use of NLP opens the door to many applications of conversational AI technology, e.g. in the form of commercial chatbots, social robots, conversational agents for mental health therapy (D'Alfonso et al., 2017), fitness motivation chatbots (Wiratunga et al., 2020), AI assistants used for education (Terzolpoulos & Satratzemi, 2019) and more. Through the use of NLP, it is even possible to perform text style transfers. This means that a new sentence can be generated in a different style from the original sentence, while still preserving its meaning. One application of this is the transfer between formal and informal text (Lai, Toral & Nissim, 2022).

Being capable of performing text style transfers between informal and formal text, NLP also facilitates the analysis of the level of formality in texts. Out of a small number of tools for formality analysis in Python3, the most dependable and frequently used code relies on the F-score (Heylighen, Dewaele, 1999). To calculate the F-score of a text, the frequencies of certain linguistic elements are used. Elements that contribute to formal style are nouns, adjectives, prepositions and articles, while elements that contribute to informal style are pronouns, verbs, adverbs and interjections. The greater the frequencies of these elements, the more they contribute to the style they represent. Since the level of formality that is most appropriate for the situation is dependent on the context of that situation, it can be seen as a disadvantage that the F-score merely looks at single linguistic elements instead of contextual elements, suggesting an incomplete judgement of formality. Another issue is that the F-score does not discriminate between different types of words within its elements, resulting in a system that is too rigid to accurately determine formality (Teddiman, 2009). One more disadvantage of the F-score is that it needs at least a small corpus of about 200 words to work accurately (Heylighen & Dewaele, 2002).

An alternative measure for the F-score is the CF-score proposed by Li, Cai and Graesser (2013). The CF-score relies on Coh-Metrix, as well as more sophisticated factors for formality classification, obtaining better results than the F-score. Elements that are used to calculate the CF-score are referential cohesion and deep cohesion for formal style, and narrativity, syntactic simplicity and word concreteness for informal style. Similarly to the F-score, the greater the value of the element, the greater its influence on the final CF-score. While obtaining superior scores to the F-score, limitations of the CF-score are that it is heavily reliant on linguistic differences, as oppose to differences in formality as perceived by humans (Li et al., 2015).

Voice

To the best of the researcher's knowledge, there are currently no formality analysis models available that measure formality from voice input. There are, however, large benefits to using voice-based interactions compared to text-based interactions. Rzepka, Berger and Hess found that using a VA, as opposed to a chatbot, results in "higher perceived efficiency, lower cognitive effort, higher enjoyment, and higher service satisfaction" (2021, p. 1). Following this, it becomes apparent that there is a great need to measure formality via VA, since it can expand its scope beyond the limits of a chatbot.

Trust

Having an appropriate degree of formality is shown to lead to more trust (Chaves et al., 2021). In the pursuit of defining trust, Przegalinska et al. (2019) found two different definitions that are both relevant in different ways to this study. The first is: "a firm belief in the reliability, truth, or ability of someone or something", with the second being "an arrangement whereby a person (a trustee) holds property as its nominal owner for the good of one or more beneficiaries". The first definition points at trust towards the system performing as intended and reliably to the user. In terms of the current study, this can be related to a factor that could influence the participant's experience of the VA in a significant way. The latter definition puts emphasis on one party holding onto one thing that the other does not hold control over. In the light of this study, this can be related to the companies that hold users' personal data. The user trusts the company to keep their data stored safely, and protect the user from being harmed in any way by entrusting their data to the company.

Trust has been shown to lead to more positive behavioural tendencies, reduced prejudice and reduced social distance in human-human interactions (Abanes, Scheepers & Sterkens, 2014), as well as human-chatbot interactions (Rhim et al., 2022) and human-voice assistant interactions, with an addition of increased intent to use (Choung, David & Ross, 2022; Lee, Ayyagari, Nasirian & Ahmadian, 2021).

Expectations towards conversational AI

Users are found to have expectations for intelligent responses when interacting with conversational software. On an emotional level, for example, in a previous study users expected that positive emotions were going to be mirrored by the software, but when the user was experiencing negative emotions, they preferred the software to react in a more intelligent and careful way than merely mirroring the emotion (Svikhnushina, Pu, 2020). Another example of an expectation found is that chatbots would adapt their conversational style and vocabulary to that of the user. Chaves and Gerosa (2021) found that when the conversational software would fail to properly adapt to the user, frustrations would arise. While not meeting the expectations with regards to adaptation can result in frustrations, positive experiences regarding adaptations hold the potential to significantly decrease social distance (Koppen, Ernestus & Mulken, 2016).

Social distance

Social distance can be defined as “people’s psychological boundary in a social relationship with a robot” (Kim, Kwak & Kim, 2013, p.1092) and plays a significant role in the success rate and satisfaction of interacting with conversational technologies. According to Koppen, Ernestus and Mulken (2016) the ability to adapt in terms of linguistic choices, leads to the reduction of social distance. A smaller social distance between user and technology has been found to lead to positive effects like increased trust, more positive attitudes towards robots in general and increased willingness to adopt a particular technology into one’s everyday social life (Banks & Edwards, 2019). Additionally, suggesting the importance of improving user experience through formality adaptation, greater similarity in character to the user by an AI-based conversational agent has been found to cause the user to comply more with the agent (El Hefny et al., 2020), while another study found that adaptability of a chatbot in healthcare situations is effective in aiding patients to achieve their healthcare related goals (Linder, 2020). Through the positive effects found, the importance of reducing social distance through adaptation becomes more apparent.

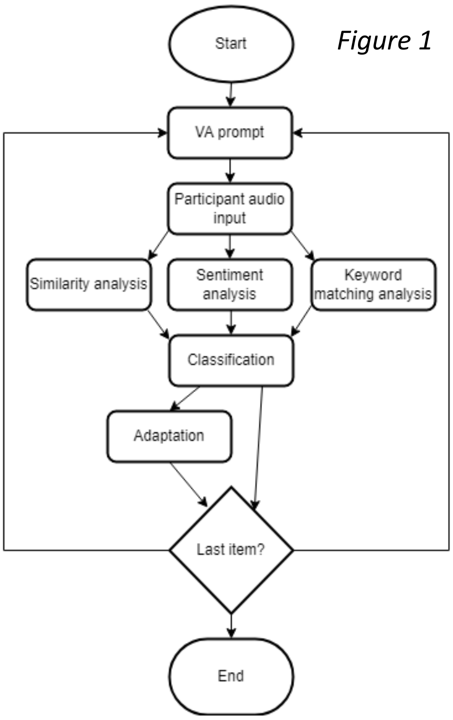
Anthropomorphism

Anthropomorphism can be defined as “the attribution of distinctively human-like feelings, mental states, and behavioral characteristics to inanimate objects, animals, and in general to natural phenomena and supernatural entities” (Salles, Evers & Farisco, 2020, p.89). Anthropomorphism has been widely researched with regard to automated systems and is found to be of significant importance in chatbots by Przegalinska et al. (2019). This research suggests that higher levels of anthropomorphism have a notable influence on the “increase of trust and overall performance of cooperation” (p.789). According to this information, as well as the knowledge that a smaller social distance can facilitate higher trust levels and increase feelings of anthropomorphism, it can be theorized that there is a positive feedback loop between increased levels of trust and decreased levels of social distance, leading to increased anthropomorphism, which then in turn can further increase trust levels. Following this reasoning, anthropomorphism is deemed as a strong factor in the development of a trusted voice assistant.

3. DEVELOPING AN ADAPTIVE VOICE ASSISTANT

In the software pipeline that is designed for this study (overview in *figure* 1), the data that is obtained through the audio input of the user is processed by means of three different analyses: a similarity analysis, a sentiment analysis and a keyword matching analysis. The results of these analyses are used to classify the level of formality of each user response to the VA. The system is then able to adapt its own level of formality to the detected formality level of the user. This adaptive system was compared in the current study to two non-adaptive versions of the same system, where the system’s formality level was either fixed to most formal or most informal. These analyses and classifications are applied to every user response, until the conversation ends. This section will provide deeper insight into the steps and

Figure 1



requirements for the development of the pipeline that was used in this study.

In order to make a pipeline that facilitates the measurement of and adaptation to levels of formality assumed by the user, several analyses need to be done. This paper focuses on the analysis of the transcripts of the user input, to which several analyses are applied that, given the previously presented literature, could be relevant for the measurement of formality. For the sake of clarity, the full pipeline will be described in three parts, chronologically following the order in which they are ran by the system. First, as soon as the program starts, necessary steps are taken to load the right dependencies and data, as well as to make sure other necessary parts are set up for everything to work (part A, visualized in *fig. 2-A*). Secondly, the appropriate set of data that the VA will use for its replies is selected (part B, visualized in *fig. 2*-B). This relies on the values for test conditions that are set in part A. Lastly, the data for the analyses is obtained and processed. The voice data is collected through the use of an audio stream, combined with voice transcription software. The audio stream starts after the VA has had their turn to speak and is closed when the participant finishes talking. The stream is then opened again after the VA's next prompt. Then, a similarity analysis, a sentiment analysis and a keyword matching analysis are performed. After this, the VA implements a Multi-class AdaBoosted Decision Trees algorithm to determine whether to increase or decrease the level of formality.

The NLP analyses (similarity, sentiment and keyword matching) and the boosting algorithm are not only applied to the adaptive condition, but also to the non-adaptive conditions, because of the value of the data and the short amount of time it takes to obtain. Contrastingly to the adaptive condition, the results of the boosting algorithm are not used to change the level of formality of what the VA will say next. In these non-adaptive conditions, the measurements into the level of formality are done to gain more insight into the workings and performance of the boosting algorithm. Alternatively, the data obtained is important for debugging purposes, since it will allow for more information on vital steps where something can possibly go wrong. All of the analyses on the user's transcript only take around a total of 300-350ms per input, this includes the decision of the boosting algorithm, which takes around 70ms to calculate its prediction.

Part A

In the following paragraphs, the actions of the VA in part A (*fig. 2*-A) of the pipeline will be explained in more detail.

Before running the code for the experiment, the test conditions are manually set. This is done by changing two values: whether it is the adaptive condition or the non-adaptive condition, and whether the test will use the data for the informal condition (level_1), the formal condition (level_5) or the adaptive condition (level_adaptable in the pipeline). Additionally, the number of the participant is also entered, since this allows separate experiment logs to be created for each participant.

Before the VA will utter its first prompt, preparations need to be made. The first step for this is importing all of the external code (dependencies) that is used to create this VA. All these dependencies were selected with the purpose of the VA being able to run locally (and therefore without an internet connection), allowing the experiment to be conducted in any quiet space, as well as improving the safety of the data collected, since the data is not sent to a third party in order for the pipeline to work. The following dependencies are imported:

Logging – The logging package is included in Python and is used to create an event logging system for this experiment. Later, in part B of the pipeline, the logging is set to show logging entries from the DEBUG level and up, which means that all logging entries will be shown. In the same section, the logging entries are set to show the time on which the log occurred (which can be used to measure the time between certain events/actions), what the name is of the level of the entry and what the logging message is. As one of the last steps in the logging configuration, the name of the file the logs will be written to is defined, creating a new log file for each participant through string formatting for the file name.

Playsound – The playsound module solely contains the playsound function, which can be used in combination with the path to an .mp3 or .wav format audio file, which it then plays. In the pipeline, it is used in part C to play the .mp3 sound files for the VA.

PyAudio – Pyaudio is a Python library that allows Python to access a microphone to record audio, which includes audio streaming. For this experiment, the built-in microphone of the researcher's laptop was used for the audio stream, as a similar microphone would likely be used if the user-VA interaction was held in practice. PyAudio is used in the VA to facilitate the audio stream through which a transcript can be obtained with Vosk. In part B of the pipeline, PyAudio is instantiated in the variable 'cap' (for 'capture') and used in part C to stream. Expanding on PyAudio's role in part C, for the non-adaptive and adaptive conditions respectively: After the VA output, the parameters for PyAudio are set. These parameters are optimized for the device that is used for the experiment. After the parameters have been set, the audio stream is opened up and finally closed after there is no more input from the participant.

Vosk – Vosk is a speech recognition toolkit that is used in the VA for the transcription of audio input that the user provides. Vosk is a large toolkit that supports over 20 languages and has many models available for these languages. The model used for this research is vosk-model-small-nl-0.22, which is trained on Dutch speech data. Out of three available models, this model yielded the most accurate results. In part B of the pipeline, the language model is loaded and applied to the transcription tool. The transcription tool is then used in part C of the pipeline right after the audio stream starts, through which it obtains the transcription in a JSON format.

Json – This is a built-in Python module that allows the program to understand and retrieve relevant information from data in JSON format. This module is used to obtain the transcription data from Vosk while the participant is speaking in part C of the pipeline.

AutoTokenizer, AutoModelForSequenceClassification – These are modules from Hugging Face's Hugging Face Transformers architecture. The modules are used to perform the sentiment analysis on the user's input. Another necessity for the sentiment analysis is a language model. Hugging Face provides a platform for developers to upload state-of-the-art, high

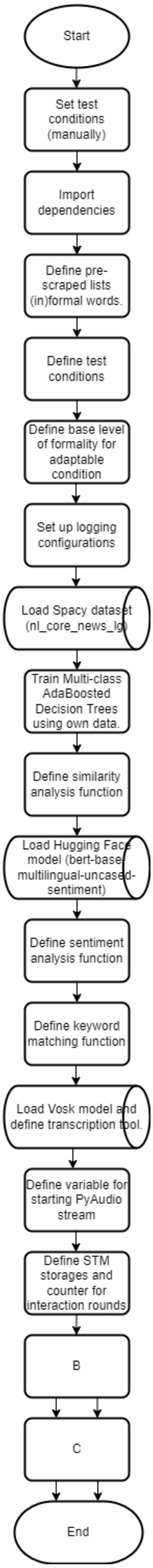

Start
Set test conditions (manually)
Import dependencies
Define pre-scraped lists (in)formal words.
Define test conditions
Define base level of formality for adaptable condition
Set up logging configurations
Load Spacy dataset (nl_core_news_lg)
Train Multi-class AdaBoosted Decision Trees using own data.
Define similarity analysis function
Load Hugging Face model (bert-base-multilingual-uncased-sentiment)
Define sentiment analysis function
Define keyword matching function
Load Vosk model and define transcription tool.
Define variable for starting PyAudio stream
Define STM storages and counter for interaction rounds
B
C
End

Figure 2-A

performance pretrained models, so for this study a model was chosen that is specifically trained for sentiment analysis: bert-base-multilingual-uncased-sentiment. The purpose of the AutoTokenizer is to convert the user input into a token format, which can be processed by the model. AutoModelForSequenceClassification instantiates the pretrained model, which can then be applied to the user's input.

The advantage of the state-of-the-art BERT-based model is the significantly higher performance than other types of language models. BERT uses bidirectional training of the language model, meaning that the context before and after the given excerpt is taken into account during training. This is different from unidirectional models where only preceding or subsequent words are looked at, and provides "a deeper sense of language context and flow" (Horev, 2018) than these models.

Torch – The torch package, which is part of the larger PyTorch library, "contains data structures for multi-dimensional tensors and defines mathematical operations over" them (PyTorch Contributors, n.d.). This package is a python implementation of Torch, a machine learning framework that is necessary for obtaining the results of the sentiment analysis.

spaCy – The spaCy package contains the '.similarity()' function that is necessary for the similarity analysis. Similarly to the sentiment analysis, a language model is needed to perform the similarity analysis. spaCy also has downloadable language models for the Dutch language. For this research the nl_core_news_lg model is used. The model is trained on a large set of Dutch news articles and is compatible with spaCy functions. It uses word vector representations that are compared to determine similarity.

AdaBoostClassifier, DecisionTreeClassifier – The class AdaBoostClassifier uses the AdaBoost-SAMME algorithm — which works like the original AdaBoost, however it performs better (Zhu, Zou, Rosset & Hastie, 2009) — to combine weak classifiers into a strong one. Weak classifiers are classifiers with high classification error, meaning that the model has trouble drawing the correct conclusion on how to classify data. The boosting algorithm is used to improve the performance of basic decision trees (Zhu et al., 2009). The DecisionTreeClassifier uses Decision Trees (DTs) to figure out and map the decisions made to reach a particular conclusion, according to the given data. In this research, the DTs create a model that predicts the level of formality, based on similarity score, sentiment score and keyword matching score. The data on which the Multi-class AdaBoosted Decision Trees are trained is generated by the researcher and was made to reflect different levels of formality, through which the algorithm can learn about the data that suit these levels.

NumPy – NumPy is a package that facilitates scientific computing with Python. The VA uses NumPy to structure training data as well as participant test data into the form of arrays for the Multi-class AdaBoosted Decision Trees. The training data will be used to train the model to recognize which level of formality is appropriate for the values given. The only value missing with the participant's data is the level of formality that they assume, which can be estimated by the Multi-class AdaBoosted Decision Trees.

Pandas – Pandas is used to further structure the arrays obtained with NumPy into dataframes that can be used by the Multi-class AdaBoosted Decision Trees to learn to make accurate predictions, as well as generate them.

Having discussed all of the dependencies and imports, the following steps that are described are all performed before there is any initial output of the VA, and serve as a means to greatly cut loading times over the course of the experiment. The next step of the pipeline will be to define lists of formal and informal words that have been scraped from an existing online overview (Genootschap Onze Taal, n.d.) before the start of the experiment. The scraping was done with the BeautifulSoup4 library, as well as the Requests library, and irrelevant data was filtered out by means of regular expressions. These lists of words will be used in the keyword matching analysis. Next, the test conditions which have previously been entered manually by the researcher will be defined. A base level for the starting level of formality for the adaptive condition is set at 3 (neutral). The dataset for spaCy is then loaded, followed by the training of the Multi-class AdaBoosted Decision Trees. The training is done by using the DT algorithm to fit scores of the similarity, sentiment and keyword analyses onto the correct levels of formality that fit the text that was used to obtain these scores. Both these scores and the matching levels of formality have been generated and determined by the researcher. Following this, a custom function for the similarity analysis is created, automating the process of selecting the VA prompt and the participant's response to it, and then applying spaCy's '.similarity()' function to them to obtain a similarity score. Defining this custom function allows the similarity analysis to be implemented efficiently into the code that facilitates the interaction with the VA in part C of the pipeline.

Then, the model from Hugging Face for the sentiment analysis is loaded. It contains a tokenizer which can be applied to the user input. A new custom function is created that can be used during the interaction with the VA in part C of the pipeline. First, the tokenizer of the Hugging Face model is used to tokenize the participant's most recent input. The sentiment model is then applied to the tokenized user input and through argmax function on the output of this model, the sentiment score is obtained.

Subsequently, two custom functions are defined that can be used during each iteration of the conversation with the AI in part C of the pipeline, in order to obtain the keyword matching score. The first function creates two counters that each keep track of the number of words that occurred in the participant's input and also occurred in the formal or informal list of words that were defined earlier. This way, the amount of formal and informal words can be derived from the participant's utterance. The second function uses the following formula to obtain a score for the keyword matching analysis:

$$\text{formality\_level\_score} = 0 + (n_f/n_u)*100 - (n_i/n_u)*100$$

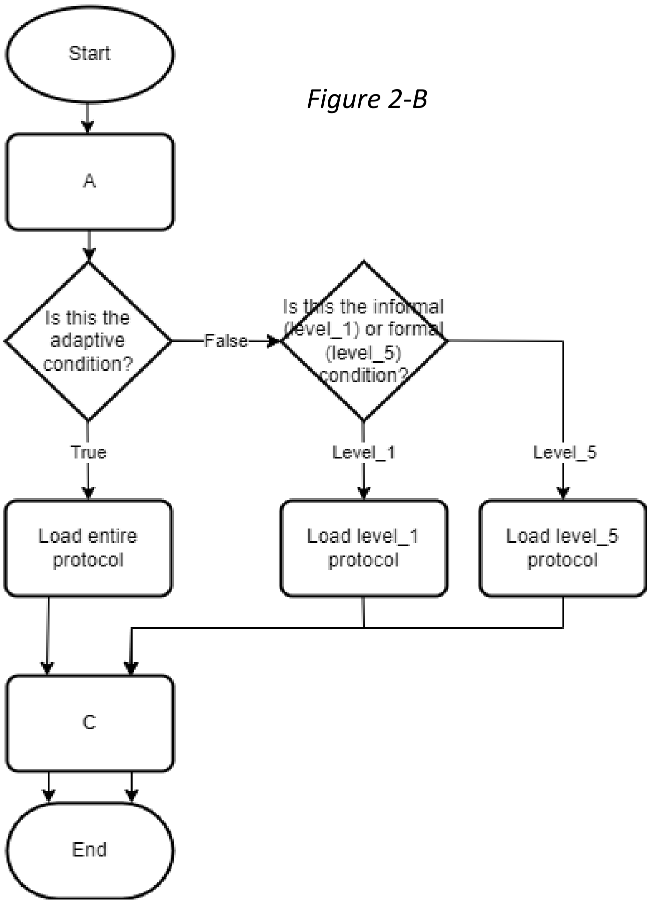


Figure 2-B

with n_f being the number of formal words that were counted by the first function, n_i being the number of informal words that were counted by the first function and n_u being the number of words in the utterance of the participant. Dividing by the length of the utterance allows the data to be measured on the same scale, across all inputs by the user. The outcome of this formula provides a score on a scale of -100 to 100, with -100 meaning that exclusively words that matched with the informal list of words were used and 100 meaning that exclusively words that matched with the formal list of words were used. The function calculates a new keyword matching score for each iteration over the protocol.

Then, the model for Vosk is loaded and used for the activation of the transcription software. Lastly, the variable to start the audio stream with is defined, a temporary storage for the transcript is created and a prompt counter is made, so that the right audio file can be selected during the experiment, and the log files can make separate logs for each prompt/input round of the experiment.

Part B

With all of the preparations done, the pipeline reaches the first forked path. This part of the pipeline is visualized in *figure 2-B*. The information these steps rely on has been defined manually by the researcher in part A of the pipeline. First, it is checked whether the test condition is adaptive or non-adaptive. If the test condition is adaptive, the full VA protocol is loaded, meaning the protocol with 5 different levels of formality for each prompt. It is necessary to load the whole protocol for this condition, since the program will be dynamically switching between levels of formality, depending on user input. If the test condition is non-adaptive, it is checked if the formality level of the VA in the experiment will be on Level_1 (very informal) or on Level_5 (very formal), after which the corresponding protocol is loaded. The VA will not switch between formality levels for these conditions, so a smaller amount of data than the whole protocol can be loaded, since these conditions only use a small part of it each. The right protocol is then sent to part C of the pipeline, where there is a slightly different procedure for the adaptive and non-adaptive conditions.

Part C

In this part (which is visualized in *figure 2-C*), the interaction between the VA and the user, as well as per-item data analysis on the user's input and a decision on the best matching level of formality are performed.

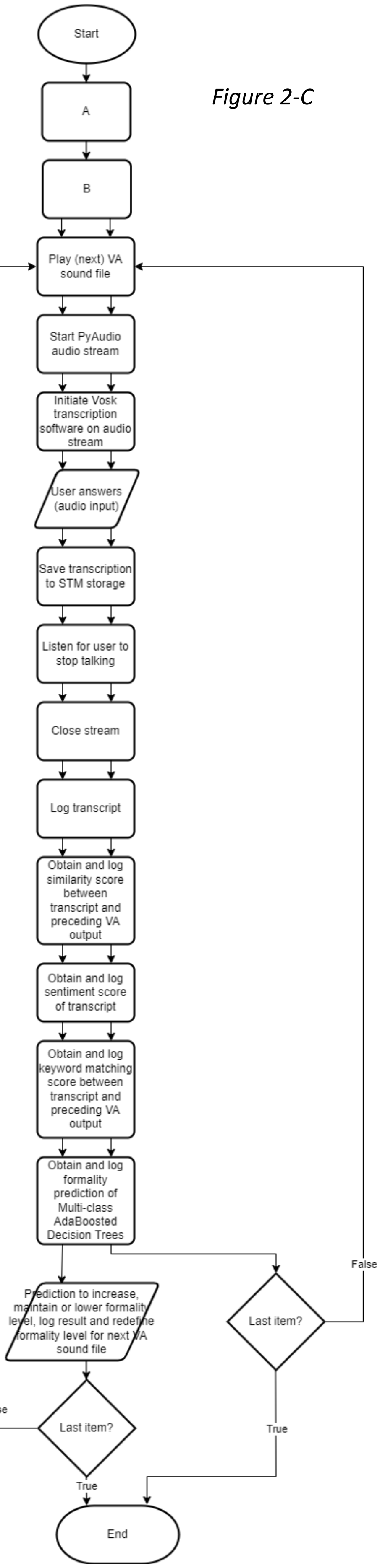

Start
A
B
Play (next) VA sound file
Start PyAudio audio stream
Initiate Vosk transcription software on audio stream
User answers (audio input)
Save transcription to STM storage
Listen for user to stop talking
Close stream
Log transcript
Obtain and log similarity score between transcript and preceding VA output
Obtain and log sentiment score of transcript
Obtain and log keyword matching score between transcript and preceding VA output
Obtain and log formality prediction of Multi-class AdaBoosted Decision Trees
Prediction to increase, maintain or lower formality level, log result and redefine formality level for next VA sound file
Last item?
False
True
Last item?
False
True
End

Figure 2-C

Part C of the pipeline starts with playing the sound file of the first VA prompt. In order to save processing time during the interaction between the VA and the participant, the sound files of each item in the protocol have been generated as .mp3 files before the experiment with the gTTS API. After the first sound file is played, in each of the following iterations of the loop, the sound file for the next prompt will play, which is selected from the directory in which the sound files are present. For the non-adaptive conditions, the next sound file for the same level of formality will always be played, which causes the output of the VA to be linear. A dynamic path is followed by the adaptive condition, through which the conclusion on the level of the user's formality is obtained with the Multi-class AdaBoosted Decision Trees. The level of formality that is found influences the sound file that is picked from the directory with sound files, by updating the variable for the level of formality that the VA should assume and selecting the right file using string formatting with this new level of formality.

After the VA has spoken, the PyAudio stream is opened up, in order to stream the audio input from the participant. Vosk is immediately activated after the stream starts, which will then in real time transcribe what the participant says. The participant starts speaking and while this happens, the transcription of Vosk is saved to a temporary storage. Since Vosk transcribes the streamed audio in short instances, the VA is programmed to listen for a period of more than one silent instance and then act out the next steps in the pipeline. This ensures that the user can speak relatively freely, as much as they deem necessary for the answer, and the VA should not interrupt them.

As soon as it is registered that the participant has stopped speaking, the audio stream is closed and the transcript is saved to a variable and logged. After this, the transcript is used for the similarity analysis, the sentiment analysis and the keyword matching analysis. The results of these analyses are saved to variables, as well as logged. The data obtained from the analyses are transformed into an array with NumPy and then put into a dataframe with the pandas library, for the data to be processed by the Multi-class AdaBoosted Decision Trees. The algorithm will then make a prediction of the level of formality that the participant uses.

This is the point where the adaptive condition and the non-adaptive conditions start to differ. While the non-adaptive conditions merely log the estimated level of formality as measured with the user's input and then go into the next iteration of the interactive loop, the adaptive condition will start a process to first measure what the level of formality was that the participant used, log it and then update the level of formality that the VA will use in its next prompt. In order for the VA not to jump dramatically between levels of formality, it is checked whether the newly measured score of the formality level is lower, higher or equal to the current level. If the level is equal, the formality level will stay the same in the next VA prompt. If the newly measured level of formality is higher than the preceding level, no matter how much, the level of formality will rise by 1. A similar mechanism is in place for lower levels of formality, meaning that if the measured level of formality is lower than the preceding level, the level of formality will decrease by 1. The single level increase was built into the pipeline to avoid dramatic changes in formality by the VA, with the assumption that it would cause confusion or frustration in the participant if they would speak with a very irregular conversational agent. The new level of formality will then be set in the variable that the sound file selection system uses to find and play the appropriate VA prompt in the next iteration of the interactive loop. This ensures that the VA will directly respond to changes in formality by the user. The new level of formality is logged. Having done this, the adaptive condition will also go to the next iteration of the interaction loop with the VA. Lastly, when every step of the protocol has been done, for each condition, the program will stop running.

4. EVALUATING THE ADAPTIVE VOICE ASSISTANT

In order to test the effect of formality adaptivity and real-time data collection on people's attitudes and trust towards interactive conversational software, as well as to provide insights into previously overlooked factors like trust, which F-score and other measures of formality do not take into account, a study was conducted in which participants were each presented with a scenario wherein they interacted with a custom-built VA. This scenario simulated a small conversation in which preferences for a holiday to the beach are discussed. Participants were randomly assigned to one of three groups, representing the three different test conditions: interaction with very informal VA, interaction with very formal VA and interaction with adaptive formality VA. The study follows a 1x3 between-subjects design.

Participants

In order to obtain data from a broad demographic with limited time and resources, convenience sampling was used. The participants were recruited through the researcher's personal network, and were tested in the participant's residence under quiet conditions. The recruitment resulted in n = 20 participants. Three participants were excluded from this research on the basis of incomplete data. This results in a total of n = 17 participants, consisting of n = 10 males and n = 7 females, with ages ranging between 19 and 78 (Mean = 49.76, SD = 21.64). Out of 17 participants, 11 had used voice assistants before, with four of them using VAs daily to weekly, and the remaining seven using them monthly to yearly. It is worth noting that because each experiment was performed in the participants' homes, each location different from the others. Though unavoidable, the difference between locations does simulate the different locations that the VA would be used in, if it was used as a tool at home. Another important point to note is that the nature of the relationship between the researcher and participant differs between samples.

After randomly assigning the first 10 participants, the decision was made to increase the number of participants in the adaptive condition, which would improve the ability to analyse the effects of real-time formality adaptation. While some participants were still assigned to the informal and formal conditions, the majority of the remaining participants was assigned to the adaptive condition from this point onward. The number of participants per condition were n = 4 for condition 1 (very informal), n = 4 for condition 2 (very formal) and n = 9 for condition 3 (adaptive formality).

Materials

In order to perform the experiment, the VA system as described in *section 3: Developing an adaptive voice assistant* was used. With the aim of improving the flow of the experiment, the system was optimized by loading all of the necessities for the interaction with the VA first and only then starting with the actual interaction. Furthermore, the time between the end of the user's utterance and the next VA output was also optimized by listening for the minimal amount of empty transcription instances by Vosk.

In the protocol that the participant is presented with, a conversation is held in which the VA asks the participants about their preferences regarding a vacation to the beach. The protocol was written in five different ways, through the use of words and phrases that share the same meaning, but each reflect a different degree of formality. These levels are on a scale of 1 to 5, with 1 being very informal and 5 being very formal. An example of the five different versions of one prompt is shown below (Figure 3). The full protocol with all levels of formality can be found in Appendix A.

2.1 Hoi daar! Het is superleuk om je te ontmoeten. Ben je er helemaal klaar voor een mooie vakantie naar het strand te plannen?

2.2 Wat leuk om je te ontmoeten! Heb je er zin in om een vakantie naar het strand te plannen?

2.3 Wat leuk om u te ontmoeten! Heeft u er zin in om een vakantie naar het strand te plannen?

2.4 Aangenaam om u te ontmoeten. Kijkt u ernaar uit om een vakantie naar het strand te plannen?

2.5 Aangenaam om met u kennis te maken. Bent u gereed om een vakantie naar het strand te boeken?

Figure 3

Measurements

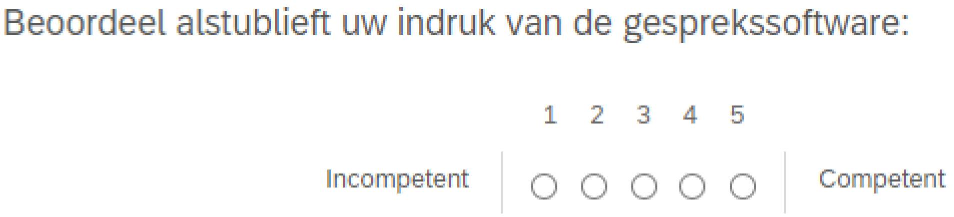
Online survey to measure trust and effect of real-time formality adaptation

This research uses a combination of direct and indirect measures to determine the trust that the user experienced with regards to machines, as well as the VA specifically. These measures are presented to the participant through Qualtrics. As a direct measure of trust, the Checklist for Trust between People and Automation (Jian, Bisantz, Drury & Llinas, 1998/2000) is used. In order to gain measurements of user experience with regard to real-time formality adaptation, as well as indirect measures of trust, the Godspeed questionnaire (Bartneck, Kulić, Croft & Zoghbi, 2009) and the Inclusion of the Other in the Self (IOS) task (Aron, Aron & Smollan, 1992) are made use of. For the Checklist for Trust between People and Automation (CTPA) and the individual subscales of the Godspeed questionnaire, the internal consistency was measured by calculating Cronbach's Alpha, which will be reported when each of these questionnaires and subscales are discussed in more detail. All levels of Cronbach's Alpha that were measured in this study were classified within the 'acceptable' (between 0.7 and 0.8) level and the 'good' (between 0.8 and 0.9) level of internal consistency, meaning that the surveys used in this study were found to be from acceptable or good reliability.

Checklist for Trust between People and Automation - The survey was conducted directly after the interaction with the VA, and includes a version of the Checklist for Trust between People and Automation that was translated into Dutch by the researcher. During the survey, the CTPA, a questionnaire that uses 7-point Likert scales to measure the participant's level of trust with regards to machines and with special focus on the interaction with the VA, is administered twice. The CTPA is used for the first time after the IOS task to measure the initial level of trust towards automated systems. After the participant has filled out the IOS task, the first CTPA (CTPA-1) and the Godspeed questionnaire, they are instructed to wait for the researcher to inform them more about the experiment. During this time, the researcher reveals information about the data that was collected and how it was processed during the interaction with the VA, that was previously left out. The participant is then asked to fill out the Checklist for Trust between People and Automation again (CTPA-2), which will provide information about the effect of real-time data collection and system adaptation on trust. This way, the CTPA questionnaire measures important data for both RQ2 and RQ3. An example of a question from the CTPA is "het systeem is misleidend" ("the system is deceptive"). Since the research of Jian et al. (2000) focused mainly on analysing the components of trust that should be measured in order to accurately determine trust, the questionnaire they put forward in their study serves as a means of measuring trust multi-dimensionally. The study lacks instructions or evidence for determining a single trust score, however, for the purposes of this study, the scores per participant were averaged after inversing the scores of the negative questions (question 1-5) in order to obtain a single score representing trust in the VA. With this in mind, the questionnaire (original can be found in appendix C) was still deemed relevant to the goals of the present study as a measure of trust. For the CTPA-1, Cronbach's α = 0.84, and for the CTPA-2, Cronbach's α = 0.71. Resulting in good and acceptable scores, respectively.

Godspeed questionnaire - The Godspeed questionnaire (original: appendix D) measures the user's experience with regards to the interaction with the VA through the subscales anthropomorphism, animacy, likeability, perceived intelligence and perceived safety. In this research, all subscales except for the animacy subscale are used, since there is considerable overlap between the anthropomorphism and animacy subscales (Bartneck et al., 2009). The anthropomorphism subscale is chosen to be administered instead of the animacy subscale, since the relevant literature for this study measures anthropomorphism more frequently than animacy. In addition to this, the last item of the anthropomorphism subscale was also excluded, as it measures the participant's experience on the movement of a robot. Because there is no robot or any other kind of visual information for the participant, this item was deemed irrelevant for the present study.

Each item in every subscale makes use of 5-point semantic differential scales, in which the participant indicates their attitudes towards the VA. The anthropomorphism subscale (Cronbach's α = 0.81) uses 4 different items, whereas the likeability (Cronbach's α = 0.86) and perceived intelligence (Cronbach's α = 0.83) subscales each have five items, and the perceived safety subscale (Cronbach's α = 0.77) has three. In order to maintain continuity in the use of the Dutch language during the experiment, the researcher translated the used items of the Godspeed questionnaire to Dutch. An example of a question from the Godspeed questionnaire, as used in the experiment, can be found in *figure 4*. The Godspeed questionnaire provides information about the user's experience of the VA, and measures data that is used to answer RQ2 of the current study.



Beoordeel alstublieft uw indruk van de gesprekssoftware:

1 2 3 4 5

Incompetent ○ ○ ○ ○ ○ Competent

Figure 4

Inclusion of the Other in the Self - The IOS task measures the user's perceived social distance to the other conversational agent – here the VA – by asking the user to indicate a degree of closeness visualized by a series of images depicting two circles representing both agents (*figure 5*). The data gathered through the IOS task is used to answer RQ2.

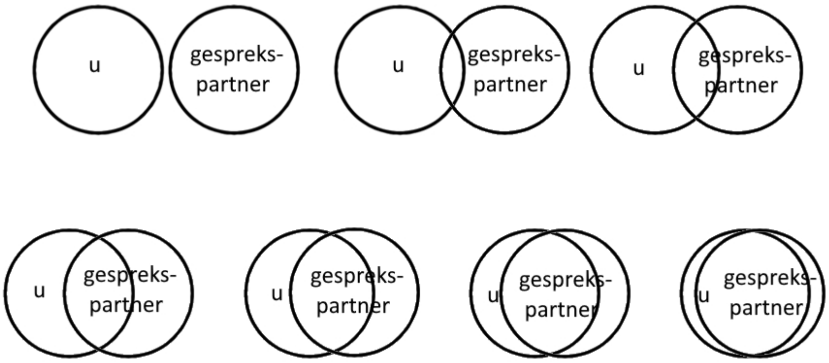


Figure 5

Interview to measure the user's experience of the system

In order to determine the performance of the VA system and to observe if the system functions in the way that a user would find desirable, the experiences of the participants are measured through various questions posed in a face-to-face structured interview. The questions (Appendix B asked are in part based on the Chatbot Usability Questionnaire (Holmes et al., 2019), with the rest of the questions being created by the researcher in order to measure system performance in other relevant ways. The results of the interview provide insight into the user's experience of the system, providing information necessary for RQ1. Since the system provides a new manner of measuring the user's level of formality, it is important to record the ability of the participant to understand and interact with the most important aspects of the system. The interview setting is chosen in order to measure the scores for the participant's experience with the VA in a structured and effective manner, while also being able to expand on this data by means of qualitative data.

Procedure

Once the participant sat down at the laptop with the researcher, a consent form was signed via Qualtrics. After this, the researcher asked the participant general questions about their age and gender. Then, the researcher ran the code for the experiment, after which the VA spoke its first item from the protocol. The participant responded by talking into the laptop's microphone. The VA subsequently went past each step of the protocol and notified the participant when the interaction was finished. After this, the participant was asked to fill out a questionnaire via Qualtrics. When the questionnaire was filled out to the point where participants were asked to wait for the researcher to inform them more about the experiment, the researcher informed them about the true amount of data that was gathered, as well as the nature of the data, which the program needs to measure in order for the system to fulfil its purpose. A last questionnaire (identical to the Checklist for Trust between People and Automation that the participant had already answered before) was filled out with the newly presented information in mind.

Finally, the researcher asked the participant six interview questions. These questions are 5-point semantic differential questions. After having asked what number on that scale would fit the best with the participant's experience of the conversation with the VA, they were asked what their motivation for this answer was. When the interview was concluded, the participant was thanked for their participation, which marked the end of the test. For each participant the experiment, from the consent form to the end of the interview, took around 25 minutes to complete.

5. RESULTS

After the development of the voice assistant, it was used in the pilot study in order to obtain data on the functioning of the VA, as well as to provide the interaction necessary to study the remaining research questions. In the following analyses, the data from VA prompt 1-11 is used, since these hold the relevant data.

Voice assistant performance

In addition to the development of the VA that was described in section *3: developing an adaptive voice assistant*, the functioning of the VA was also measured during the experiment. We found that during the interaction with the VA, the participants in the informal condition also responded to the prompts in the most informal way, with an average formality level of 3.20. Consistent with this, the participants in the formal condition were found to have the highest average level of formality, amounting to 3.50. Lastly, the participants in the adaptive condition were found to adopt formality levels averaging in the middle of the formerly mentioned conditions, with a score of 3.35.

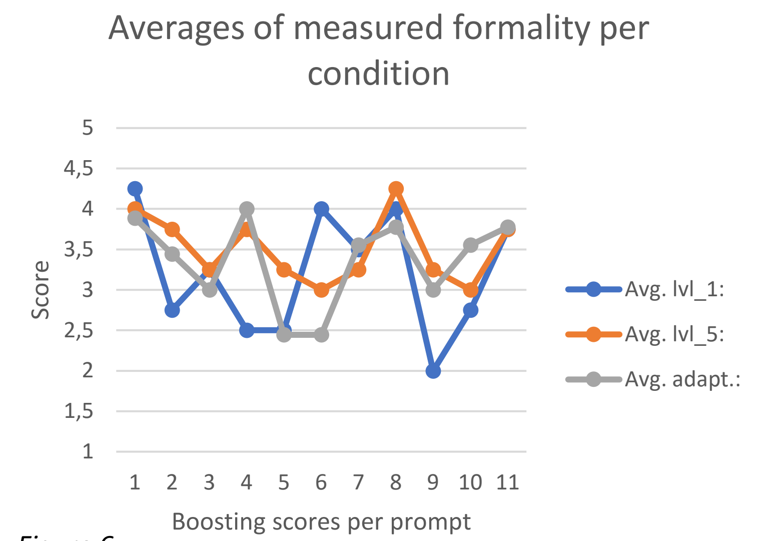


Figure 6

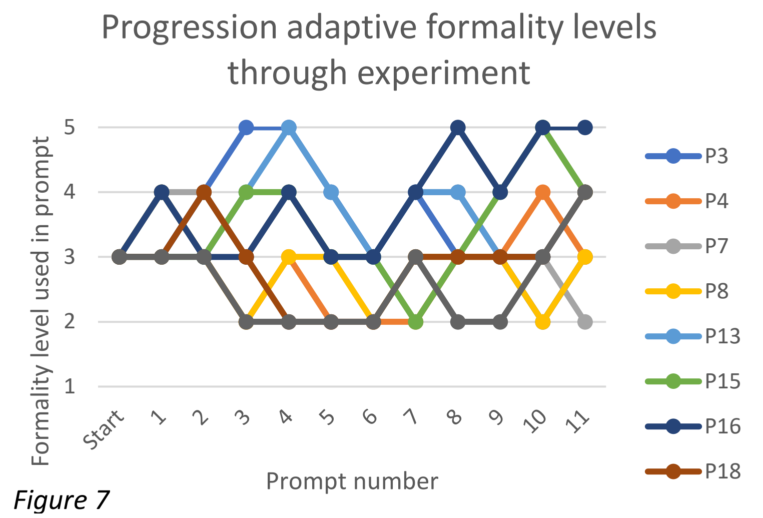


Figure 7

Figure 6 shows the measured level of formality over time, per condition. It is visible here that the participants' average responses in the formal condition (Avg. lvl_5) remain at a relatively high level of formality, never sinking below 3. The graph also shows that the average participants' responses in the informal condition reach the lowest formality score of all conditions with a score of 2, while also rising to more formal levels after prompt 1, 6, 7 and 8.

In *figure 7*, the formality adaptation by the model is visualized. All starting at formality level 3, the graph shows the progression of the new levels of formality adopted by the VA for each subsequent prompt, which are obtained through the Multi-class AdaBoosted Decision Trees algorithm.

The functioning and the usability of the VA were measured both quantitatively and qualitatively in an interview setting with the participant, partially adapted from the Chatbot Usability Questionnaire. The quantitative data obtained per question of the interview is shown in *figure 8*. In order to gain qualitative data, participants were asked to elaborate on their quantitative answer. The following are reoccurring comments with each of the questions asked:

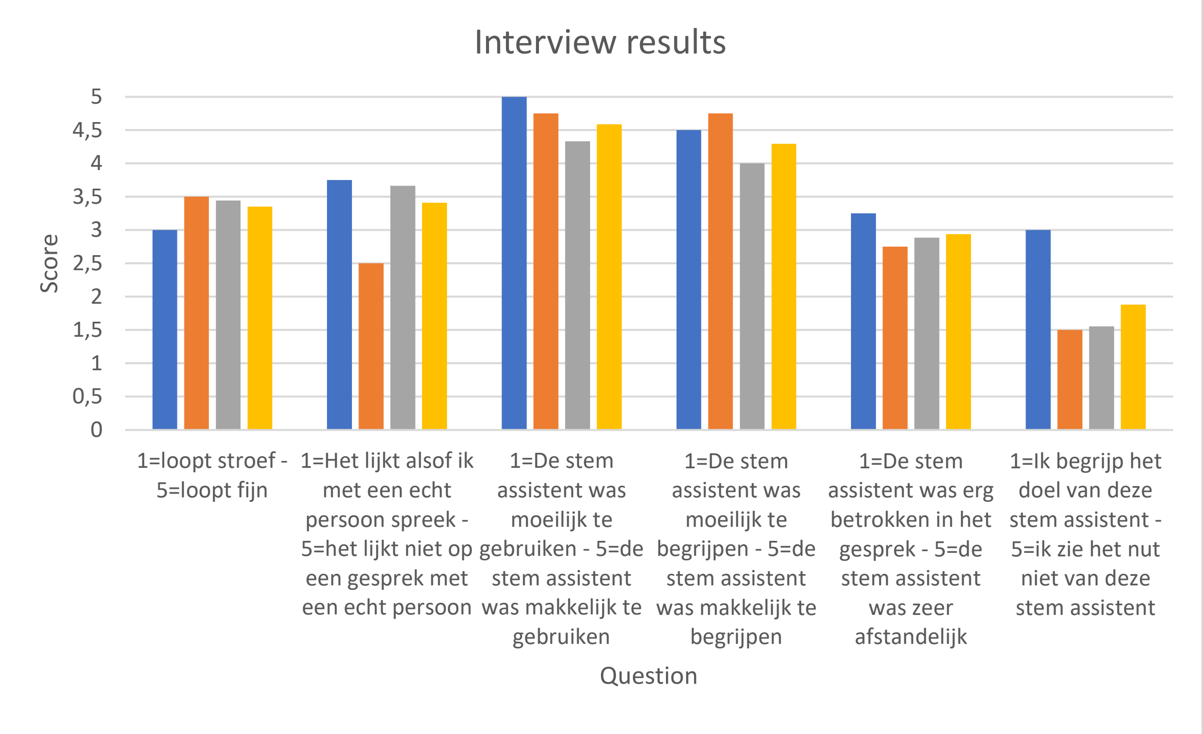


Figure 8

Question 1: While some participants (n=5) were purely positive about the flow of the conversation with the VA, the majority of the participants (n=11) commented on the delay experienced after each of their inputs. Within this last group, there was also a division between people that found it a problem and people that accepted this as part of the system.

Question 2: When asked whether the interaction with the VA felt like a conversation with a real person or not, a large part of the participants (n=8) commented that the text-to-speech voice did not sound like a real person. The voice of the VA was described as cold, too robotic, and not very alive. On the other hand, three people did describe the voice of the VA as sounding human.

Question 3: In the question about ease of use, the majority of the participants felt very positively about how easy it was (n=6) to go through the interaction with the VA, saying that 'you only need to talk and that is it' (n=6). Some participants said that the questions were sometimes vague.

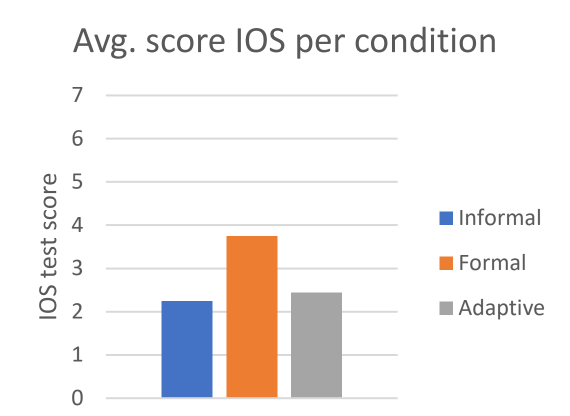


Figure 9

Question 4: In this question the participant was asked how difficult or easy it was to understand the VA. While the largest part of the participants had no trouble understanding the VA (n=10), there were also comments on the questions being hard to understand (n=3) or feeling out of place (n=5).

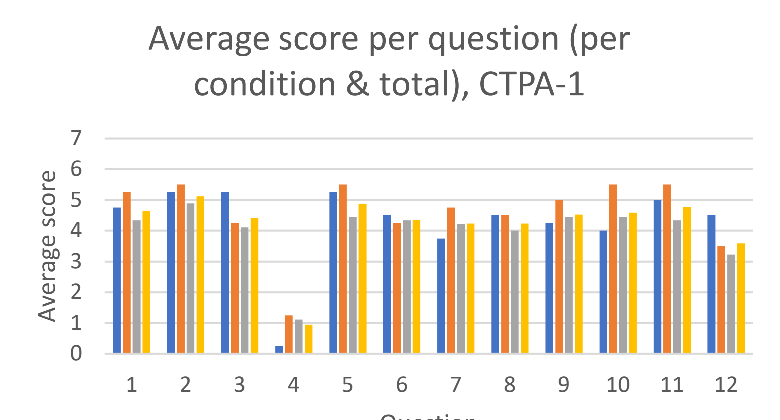


Figure 10-A

Question 5: This question dealt with how much involvement in the conversation the participant felt from the VA. Here, the comments that were primarily given pointed out that the participants felt a lack of connection with the VA. Across all conditions, participants pointed out a lack of reaction (n=7) or lack of back-and-forth interactions (n=5).

Question 6: In this last question, which focused on understanding the purpose and applications of the VA, nearly all participants indicated that they found it easy to understand the goal of the VA (n=13) and that they could understand the applications of the VA software well (n=9). However, some people specified that they had trouble understanding the possible further applications of it (n=7).

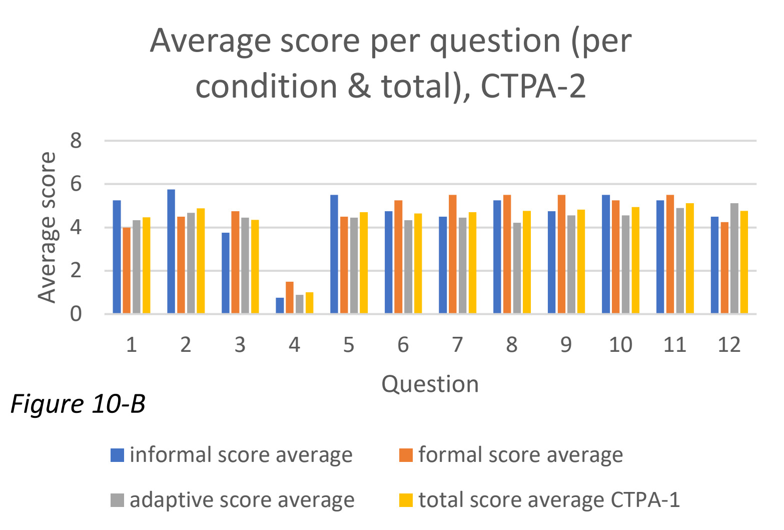


Figure 10-B

Attitudes and trust towards the voice assistant

In order to measure how people's attitudes and trust towards interactive conversational software are impacted by real-time formality adaptation, the differences between conditions are measured for the IOS scale (*figure 9*), Checklist for Trust between People and Automation (*figure 10-A and figure 10-B*) and the selected parts of the Godspeed questionnaire (*figure 11, A-D*). While the participants' indicated distance to the VA in the IOS scale is similar in the adaptive and informal conditions, with informal being the closest on average, the distance seems to be experienced as smaller in the formal condition. The formal condition also scores higher than the informal and adaptive conditions on every category of the Godspeed questionnaire, and on almost every question of the CTPA-1. The only point where the formal condition is noticeably overtaken by the informal condition is at the start of the CTPA-2. The adaptive condition scores below the average on every condition of the Godspeed questionnaire, as well as most items in both CTPA questionnaires.

Figure 11-A

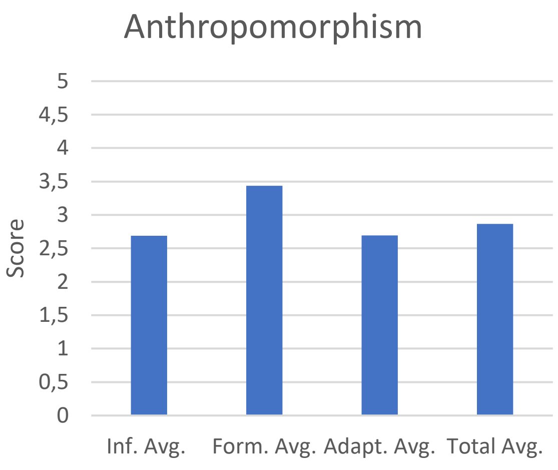


Figure 11-B

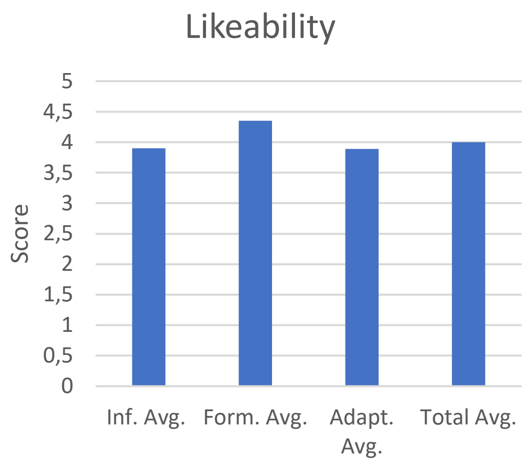


Figure 11-C

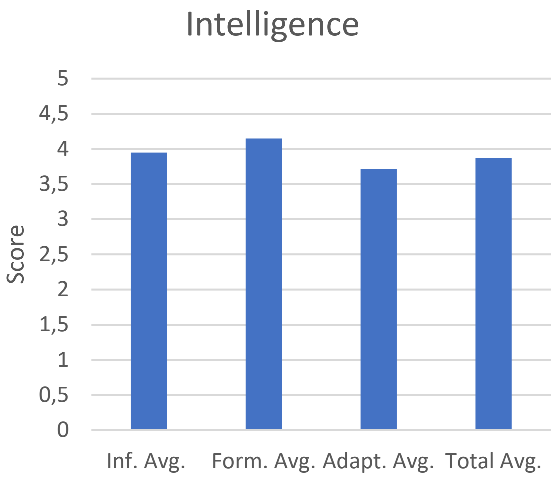
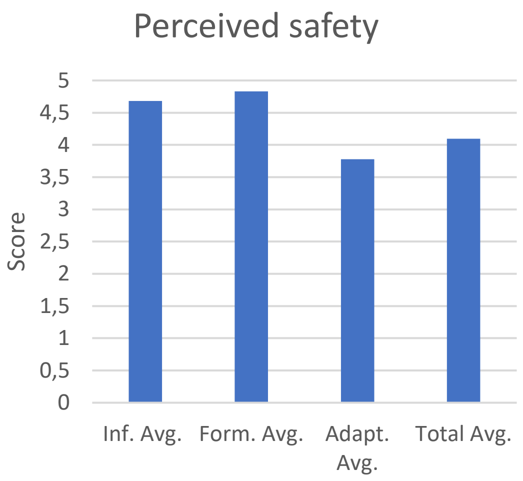


Figure 11-D



Trust and data collection

For the measurement of the effect of real-time data collection on the participant's feelings of trust towards the VA, the Checklist for Trust between People and Automation was administered. In *figure 10-A*, the data for the different test conditions, as well as the data for the sample as a whole are shown for the first use of the CTPA (CTPA-1). In *figure 10-B* the same type of data as in *figure 10-A*, only obtained the second time the CTPA was filled out (CTPA-2), is visualized in a similar way to CTPA-1. In order to show the influence of the knowledge of the data collection, the total sample scores for the CTPA-1 and CTPA-2 per prompt are visualized in *figure 12*, while the individual scores of the participants, and the change in their trust score between the CTPA-1 and CTPA-2 can be seen in *figure 13*. The thicker black line represents the average change in trust score over the entire sample, showing an increase in trust.

6. DISCUSSION AND CONCLUSION

Discussion

The present study aimed to develop the first voice assistant that adapts its own level of formality to that of the user in real-time, and to investigate the effects of this adaptation to the level of trust that the user experiences with regards to the VA. Additionally, this study aims to investigate the role of real-time data collection on the level of trust a user has towards the VA.

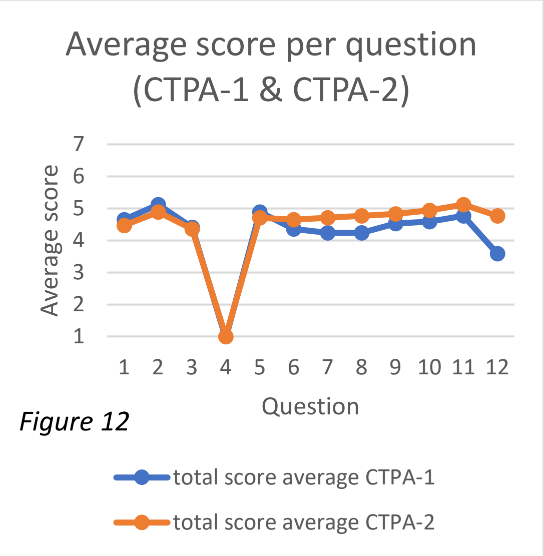


Figure 12

Main findings and interpretation

In this research the necessary components for real-time formality adaptation for a voice assistant have been identified and implemented. A voice assistant was built using a new way to measure formality from the participant's audio transcripts that were generated during a conversation with a VA. The VA is able to adapt the level of formality it assumes after each input it receives from the user, resulting in a model that is able to adapt and react quickly to the level of formality the participant uses, as well as changes to that level. What also becomes visible in *figure* 6 and *figure* 7, is that in certain points of the interaction, the scores across conditions converge. This means that for these questions, there is adaptation from the user to the VA. This also provides further support for the benefits of the adaptive condition, since there is a more human-to-human-like, two-way exchange of formality levels.

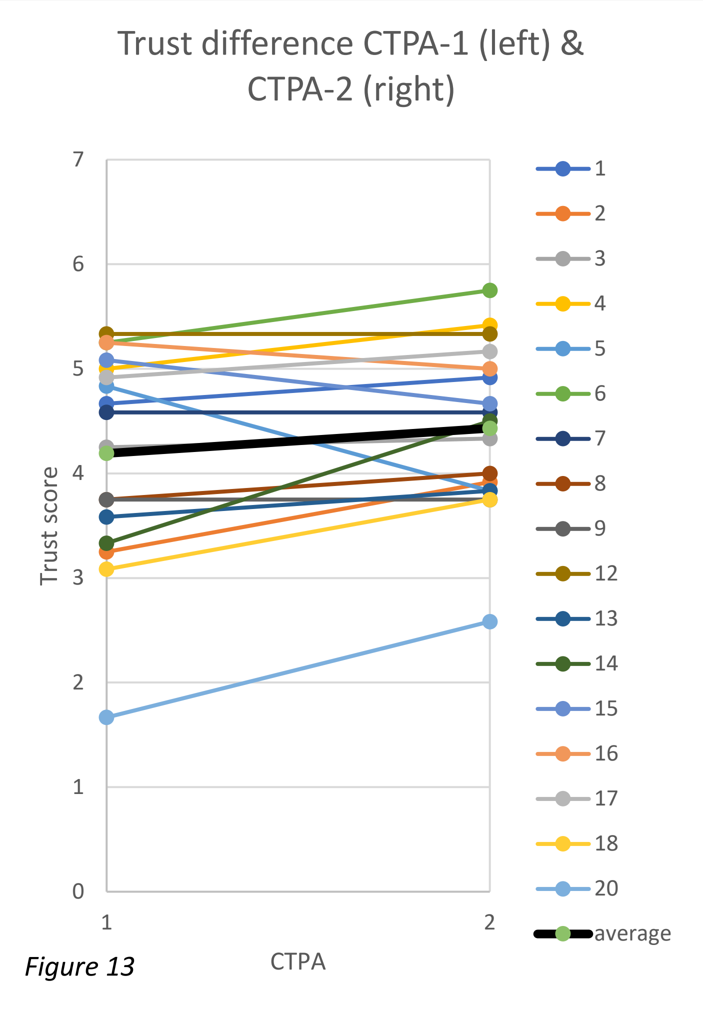


Figure 13

General feedback on the interaction with the VA was that the waiting times between the end of the user's input and the next VA output last too long, that the voice is too robotic to sound like a real person, that the VA was easy to use and easy to understand, that there is more desire for intelligent and specific responses from the VA to the user's input, and finally that the goal is mostly clearly understood, albeit that the applications of it are sometimes still hard to imagine.

As for the performance of the VA according to the surveys conducted through Qualtrics, the VA seems to perform well on trust in both the CTPA-1 and CTPA-2, averaging above the mean (4) on every question except for question four and question 12 (only CTPA-1), generally signifying above-average trust. The VA also scored well on Godspeed-likeability, Godspeed-intelligence and Godspeed-safety, scoring above the mean (3). Godspeed-anthropomorphism was found to score just below three on average, which means a slight negative score was achieved for this element and that the VA was seen as not very human-like.

Even though the adaptive condition was expected to perform better than the non-adaptive conditions, it was seen scoring lower on the majority of the measures that were taken. While obtaining higher scores in the interview on conversational flow (Q1), involvement with the user (Q5) and clarity of purpose (Q6), the adaptive condition scored the lowest in terms of ease of use (Q3) and understandability (Q4) in the interview, while scoring averagely on human-likeness (Q2). In the Godspeed questionnaire, the adaptive condition scored the lowest on all sections. The scores that were obtained through the CTPA-1 and CTPA-2 also showed the adaptive condition scoring below the sample average on nearly all items across both iterations of the questionnaire. Finally, through the IOS scale the adaptive condition, averaging at 2.44 was found to be experienced as noticeably more socially distant than the formal condition (3.75), but slightly less than the informal condition (2.25). Out of all conditions, the formal condition was found to score the highest in the majority of the items across the study.

Toader et al. (2019) offer a possible explanation for the low scores achieved by this condition. Because the manner that the participant gets spoken to during the experiment in the adaptive condition continuously changes in terms of level of formality, the VA might sound inconsistent and error-prone, which Toader et al. argued could impact the perceived competence, trust and consumer responses negatively, which is something that seems to be reflected in the data gained with this study. A more refined framework for adaptive formality could decrease these effects. Future research could build on the current pipeline, which could possibly start producing more positive results.

Looking at the effects of the real-time data collection and processing that occurs within the VA across every condition, and therefore looking at the effects of these on the global averages in the CTPA-1 and CTPA-2, we observe visible differences from item 6 onwards, where the CTPA-2 reaches higher levels of trust than the CTPA-1. These findings are consistent with most of the individual scores of the participants. While trust for some participants does go down after being informed about the data that is collected, in general, the trust of the participants increases. Since the CTPA-2 obtains scores higher than the CTPA-1, it becomes apparent that knowing more about the inner workings of the VA will increase trust, even if that knowledge includes information about machine learning algorithms being applied to user-generated data.

Study strengths, limitations & future directions

One strength that sets this study even further apart from existing research into formality and formality recognition models, is the use of voice as a means of communicating with the interactive conversational software. While benefits of using voice rather than text have been described for older, less technologically informed people (Kowalski et al., 2019), this study focuses on a broader demographic. In the interview, many participants were very eager to think about all of the processes that could be improved with the use of VAs with formality adaptation, as well as existing processes with VAs that could be improved with real-time formality adaptation. The use of voice is found to create a satisfactory experience by improving on the ease of use of interacting with the conversational software.

The use of voice leaves room for more analyses than just text-based analyses on the transcript of the user input. Possibilities for the inclusion of voice analyses, for example a pitch analysis and rate of speech analysis, were also explored by the researcher, but were not yet implemented. Using voice analyses would facilitate multimodal data collection, and is theorized to improve the performance in determining the formality level the participant uses. It would also provide the boosting algorithm with a broader set of data to learn from, possibly improving accuracy. For these reasons, it could be valuable for future research to look into the possible benefits of voice analysis.

Other limitations to this research are the sample size, which is relatively small and could therefore put more weight on outliers than a larger study would, and the size of the training dataset for the boosting algorithm. As becomes apparent in *figure 5*, the adapted formality levels that the VA uses after obtaining a score through the boosting algorithm, as well as the measured levels of formality across the entire sample, never reach 1. One reason this could be the case is the type of interaction that the participants dealt with. It could be that in another scenario, formality level 1 would be reached. Another possibility would be that level 1 is seen as too informal, with the participants refusing to interact with the VA on this level. One last reason could be the use of an inadequate training dataset. The training dataset is also limited to the researcher's view of levels of formality, which might not accurately represent degrees of formality in reality. In order to improve the accuracy of the labelled formality levels in the training dataset, the used inputs could be rated by a greater amount of people in order to develop a better supported score. It is important for future research to improve and contribute to the training dataset, as to improve model performance. Lastly, the voice of the VA is gendered female, which can have an effect on measured trust, as well as other social factors (Toader et al., 2019). Future research can also aim to address these limitations and find ways to reduce their effects.

Conclusion

The current study sought to answer three research questions regarding the development and evaluation of a voice assistant with real-time formality adaptation capabilities. A software pipeline for a voice assistant was created and a new manner of real-time formality adaptation was researched and implemented. The VA performed well on user trust, likeability, intelligence and understandability, but obtained below-average scores on anthropomorphism.

The study successfully used dynamic formality adaptation, but found that in an interaction with the VA, people's attitudes and trust were impacted in a way that was in many ways similar to the informal condition. The formal condition scored higher on most positive measures than the only two conditions, implying that this is the most appealing condition to interact with for participants. A likely explanation for lower scores obtained by the adaptive condition is the inconsistent nature in which the participant is addressed, being a reason to experience the system as inconsistent and error-prone.

With regards to the interaction between user trust and data collection, a positive effect of transparency in the collection and use of data was found. Even though it was theorised that the real-time data collection and processing would be of negative influence on people's trust towards the system, gaining more knowledge about the inner workings of the VA instead increased this trust. Because of the exploratory nature, small scale and other limitations of this study, it is important to gain more insights into real-time formality adaptation for VAs.

REFERENCES

Abanes, M. S., Scheepers, P. L. H., Sterkens, C. (2014). Ethno-religious groups, identification, trust and social distance in the ethno-religiously stratified Philippines. Research in Social Stratification and Mobility, 37, 61-75. doi: 10.1016/j.rssm.2014.02.001

Aron, A. Aron, E. N., Smollan, D. (1992). Inclusion of Other in the Self Scale and the Structure of Interpersonal Closeness. *Journal of Personality and Social Psychology, 63*(4), 596-612. https://doi.org/10.1037/0022-3514.63.4.596

Banks, J., Edwards, A. (2019). A Common Social Distance Scale for Robots and Humans*. Paper presented at the *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. Retrieved from: https://ieeexplore.ieee.org/abstract/document/8956316?casa_token=Zqro391Wp7EAAAAA:5QW6bJ5czve9aneQ2peZ3DAeuu6L9_aVr6pSWJX2424qLlO1bOKnb4znI22e1vYGRKVb3kp1YIuAnw

Bartneck, C., Kulić, D., Croft, E., Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. International Journal of Social Robotics, 1, 71-81. https://doi.org/10.1007/s12369-008-0001-3

Brendel, A. B., Greve, M., Diederich, S., Bührke, J., Kolbe, L. (2020). You are an Idiot! – How Conversational Agent Communication Patterns Influence Frustration and Harassment. *AMCIS 2020 Proceedings*, 13. Retrieved from: https://aisel.aisnet.org/amcis2020/sig_hci/sig_hci/13/

Brooke, J. (1996). SUS - A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, I. L. McClelland (Eds.), *Usability Evaluation In Industry* (pp. 189-194). London, United Kingdom: Taylor & Francis.

Carr, C. T. (2020). CMC Is Dead, Long Live CMC!: Situating Computer-Mediated Communication Scholarship Beyond the Digital Age. *Journal of Computer-Mediated Communication, 25*, 9-22. doi: 10.1093/jcmc/zmz018

Chaves, A. P., Doerry, E., Egbert, J., Gerosa, M. (2019). It's How You Say It: Identifying Appropriate Register for Chatbot Language Design. *HAI '19: Proceedings of the 7th International Conference on Human-Agent Interaction, 102-109.* doi: 10.1145/3349537.3351901

Chaves, A. P., Egbert, J., Hocking, T., Doerry, E., Gerosa, M. A. (2021). Chatbots language design: the influence of language variation on user experience. Retrieved from: https://arxiv.org/abs/2101.11089

Chaves, A. P., Gerosa, M. A. (2021). How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human-Computer Interaction, 27(8)*, 729-758. doi: 10.1080/10447318.2020.1841438

Choung, H., David, P., & Ross, A. (2022). Trust in AI and its role in the acceptance of AI technologies. International Journal of Human–Computer Interaction, 1–13. doi: 10.1080/10447318.2022.2050543

D'Alfonso, S., Santesteban-Echarri, O., Rice, S., Wadley, G., Lederman, R., Miles, C., Gleeson, J., Alvarez-Jimenez, M. (2017). Artificial Intelligence-Assisted Online Social Therapy for Youth Mental Health. *Frontiers in Psychology, 8:796*. doi: 10.3389/fpsyg.2017.00796

El Hefny, W., El Bolock, A., Herbert, C., Abdennadher, S. (2020). *Towards a Generic Framework for Character-Based Chatbots.* In F. De La Prieta et al. (Eds.), *Highlights in Practical Applications of Agents, Multi-Agent Systems, and Trust-worthiness. The PAAMS Collection. Communications in Computer and Information Science, vol 1233*. Cham, Switzerland: Springer. doi: 10.1007/978-3-030-51999-5_8

Genootschap Onze Taal. (n.d.). *Alternatieven voor ouderwetse en formele woorden*. Retrieved from https://onzetaal.nl/taalloket/modern-taalgebruik

Heylighen, F., DeWaele, J. (1999). *Formality of language: definition, measurement and behavioral determinants*. Retrieved from: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.30.6280&rep=rep1&type=pdf

Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., McTear, M. (2019). Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In M. Mulvenna, R. Bond (Eds.), *ECCE 2019: Proceedings of the 31st European Conference on Cognitive Ergonomics* (pp. 207-214). doi: 10.1145/3335082

Horev, R. [Rani Horev]. (2018, November 10). BERT Explained: State of the art language model for NLP [Blog post]. Retrieved from https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Irvine, J. T. (1979). Formality and Informality in Communicative Events. *American Anthropologist, 81*(4), 773-790. doi: 10.1525/aa.1979.81.4.02a00020

Jian, J., Bisantz, A. M., Drury, C. G., Llinas, J. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53–71. (Original work published 1998). doi: 10.1207/S15327566IJCE0401_04

Kaneyasu, M. (2022). Multimodal strategies for balancing formality and informality. *Internet Pragmatics, 5*(1), 143-164. doi: 10.1075/ip.00071.kan

Kim, Y., Kwak, S., Kim, M. (2013). Am I acceptable to you? Effect of a robot's verbal language forms on people's social distance from robots. *Computers in Human Behavior, 29*(3), 1091-1101. doi: 10.1016/j.chb.2012.10.001

Koppen, K., Ernestus, M., Van Mulken, M. (2016). The influence of social distance on speech behavior: Formality variation in casual speech. Corpus Linguistics and Linguistic Theory, 15(1). doi: 10.1515/cllt-2016-0056

Kowalski, J., Jaskulska, A., Skorupska, K., Abramczuk Cezary Biele, K., Kopeć, W., Marasek, K. (2019, May). Older Adults and Voice Interaction: A Pilot Study with Google Home. Paper presented at CHI EA '19: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. doi: 10.1145/3290607.3312973. Retrieved from: https://dl.acm.org/doi/10.1145/3290607.3312973.

Lai, H., Toral, A., Nissim, M. (2022). Multilingual Pre-training with Language and Task Adaptation for Multilingual Text Style Transfer. Retrieved from: https://arxiv.org/abs/2203.08552

Lavado-Nalvaiz, N., Lucia-Palacios, L., Pérez-López, R. (2022). The role of the humanisation of smart home speakers in the personalisation–privacy paradox. *Electronic Commerce Research and Applications, 53*. doi: 10.1016/j.elerap.2022.101146

Lee, O. D., Ayyagari, R., Nasirian, F., Ahmadian, M. (2021). Role of interaction quality and trust in use of AI-based voice-assistant systems. *Journal of Systems and Information Technology 23*(2), 154-170.

Levin, H., Novak, M. (2009). Frequencies of Latinate and Germanic words in English as determinants of formality. *Discourse Processes, 14*(3), 389-398. doi: 10.1080/01638539109544792

Li, H., Cai, Z., Graesser, A. C. (2013). Comparing Two Measures for Formality. *Proceedings of the Twenty-Sixth International FLAIRS Conference*. Palo Alto, CA: AAAI

Li, H., Graesser, A. C., Conley, M., Cai, Z., Pavlik, P. I., Pennebaker, J. W. (2015). A New Measure of Text Formality: An Analysis of Discourse of Mao Zedong. *Discource Processes, 53*(3), 205-232. doi: 10.1080/0163853X.2015.1010191

Linder, C. (2020). *The Effects of a Healthcare Chatbots' Language and Persona on User Trust, Satisfaction, and Chatbot Effectiveness* (Master's thesis, Clemson University). Retrieved from https://tigerprints.clemson.edu/all_theses/3299/?utm_source=tigerprints.clemson.edu%2Fall_theses%2F3299&utm_medium=PDF&utm_campaign=PDFCoverPages

Luger, E., Sellen, A. (2016). “Like Having a Really bad PA”: The Gulf between User Expectation and Experience of Conversational Agents. *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286-5297. doi: 10.1145/2858036.2858288

Malodia, S., Islam, N., Kaur, P., Dhir, A. (2021). Why Do People Use Artificial Intelligence (AI)-Enabled Voice Assistants? IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT. doi: 10.1109/TEM.2021.3117884

Mason, A. J., Carr, C. T. (2021). Toward a Theoretical Framework of Relational Maintenance in Computer-Mediated Communication. *Communication Theory.* doi: 10.1093/ct/qtaa035

McLean, G., Osei-Frimpong, K. (2019). Hey Alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior, 99*, 28-37. https://doi.org/10.1016/j.chb.2019.05.009

Niu, X., Martindale, M., Carpuat, M. (2017). A Study of Style in Machine Translation: Controlling the Formality of Machine Translation Output. In M. Palmer, R. Hwa & S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2814-2819). doi: 10.18653/v1/D17-1

Pelau, C., Dabija, D., Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior, 122*. doi: 10.1016/j.chb.2021.106855

Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P., Mazurek, G. (2019). In bot we trust: A new methodology of chatbot performance measures. *Business Horizons, 62*, 785-797. doi: 10.1016/j.bushor.2019.08.005

PyTorch Contributors (n.d.). *Torch* [code documentation]. Retrieved from https://pytorch.org/docs/stable/torch.html

Ray, C., Mondada, F., Siegwart, R. (2008). What do people expect from robots? Paper presented at the *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Retrieved from https://ieeexplore.ieee.org/abstract/document/4650714

Rhim, J., Kwak, M., Gong, Y., Gweon, G. (2022). Application of humanization to survey chatbots: Change in chatbot perception, interaction experience, and survey data quality. *Computers in Human Behavior, 124*. doi: 10.1016/j.chb.2021.107034

Roshan-Ghias, A., Mathialagan, C. S., Ponnusamy, P., Mathia, L., Guo, C. (2020). Personalized Query Rewriting in Conversational AI Agents. Retrieved from: https://arxiv.org/abs/2011.04748

Rüdiger, S., Mühleisen, S. (2022). Formality and informality in online performances. *Internet Pragmatics 5*(1), 1-11. doi: 10.1075/ip.00078.rud

Rzepka, C., Berger, B., Hess, T. (2021). Voice Assistant vs. Chatbot – Examining the Fit Between Conversational Agents' Interaction Modalities and Information Search Tasks. *Information Systems Frontiers*. doi: 10.1007/s10796-021-10226-5

Salles, A., Evers, K., Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neuroscience, 11*(2), 88-95. doi: 10.1080/21507740.2020.1740350

Svikhnushina, E., Pu, P. (2020). Social and Emotional Etiquette of Chatbots: A Qualitative Approach to Understanding User Needs and Expectations. Retrieved from: https://arxiv.org/abs/2006.13883

Terzopoulos, G., Satratzemi, M. (2019). Voice Assistants and Artificial Intelligence in Education. *Proceedings of 9th Balkan Conference on Informatics (BCI'19)*. doi: 10.1145/ 3351556.3351588

Toader, D., Boca, G., Toader, R., Măcelaru, M., Toader, C., Ighian, D., Rădulescu, A. T. (2019). The Effect of Social Presence and Chatbot Errors on Trust. *Sustainability, 12*(1), 256. doi: 10.3390/su12010256

Wiratunga, N., Cooper, K., Wijekoon, A., Palihawadana, C., Mendham, V., Reiter, E., Martin, K. (2020). FitChat: Conversational Artificial Intelligence Interventions for Encouraging Physical Activity in Older Adults. Retrieved from: https://arxiv.org/abs/2004.14067

APPENDIX A

Protocol (all formality levels)

Protocol introduction

In dit protocol gaat u een interactie aan met een virtuele stem-assistent. In deze sessie zult u een korte voorbeeld interactie met betrekking tot het plannen van een vakantie naar het strand volgen. Als de robot klaar is met praten, kunt u uw antwoord geven. Als u klaar bent, zal de robot verder gaan met de sessie. Bedankt voor uw deelname en succes!

Protocol 2:

1. **Hallo, mijn naam is Alex. Hoe heet u?**

2. **Wat leuk om u te ontmoeten! Heeft u er zin in om een vakantie naar het strand te plannen?**

2.1 Hoi daar! Het is superleuk om je te ontmoeten. Ben je er helemaal klaar voor een mooie vakantie naar het strand te plannen?

2.2 Wat leuk om je te ontmoeten! Heb je er zin in om een vakantie naar het strand te plannen?

2.3 Wat leuk om u te ontmoeten! Heeft u er zin in om een vakantie naar het strand te plannen?

2.4 Aangenaam om u te ontmoeten. Kijkt u ernaar uit om een vakantie naar het strand te plannen?

2.5 Aangenaam om met u kennis te maken. Bent u gereed om een vakantie naar het strand te boeken?

3. **Ik stel u graag een aantal vragen over uw ideale vakantie aan het strand. De eerste vraag is of uw voorkeur ligt bij een actieve vakantie of een relaxvakantie en waarom heeft u deze voorkeur?**

3.1 Ik ga je een paar vragen stellen over jouw perfecte vakantie aan het strand! Zou je liever op een doe-vakantie of een relaxvakantie gaan en waarom?

3.2 Ik stel je graag een aantal vragen over jouw ideale vakantie aan het strand. De eerste vraag is of jouw voorkeur ligt bij een actieve vakantie of een relaxvakantie, en waarom heb je deze voorkeur?

3.3 Ik stel u graag een aantal vragen over uw ideale vakantie aan het strand. De eerste vraag is of uw voorkeur ligt bij een actieve vakantie of een relaxvakantie en waarom heeft u deze voorkeur?

3.4 Ik zou u graag een paar vragen stellen over uw ideale vakantie aan het strand. Allereerst zou ik u willen vragen of uw voorkeur ligt bij een actieve vakantie of een vakantie waarbij u uit kunt rusten, en waarom u deze voorkeur heeft?

3.5 Ik wens u een paar vragen te stellen over de vakantie aan het strand die het meest passend is voor u. De eerste vraag luidt: Zou u mij kunnen vertellen of uw voorkeur ligt bij een meer actieve vakantie, of bij een vakantie waar u meer de rust opzoekt, en waarom uw voorkeur hier ligt?

4. **Inhakend op deze vraag, zou ik graag willen vragen naar wat voor activiteiten u uitkijkt bij een vakantie naar het strand?**

4.1 Dankjewel voor je antwoord! Ik zou je ook graag willen vragen welke activiteiten jij echt tof zou vinden bij een vakantie naar het strand?

4.2 Inhakend op de laatste vraag, zou ik graag willen vragen naar wat voor activiteiten je uitkijkt bij een vakantie naar het strand?

4.3 Inhakend op de laatste vraag, zou ik graag willen vragen naar wat voor activiteiten u uitkijkt bij een vakantie naar het strand?

4.4 Met het oog op de laatste vraag, zou ik u graag willen vragen naar wat activiteiten waarnaar u uitkijkt bij een vakantie naar het strand?

4.5 De laatste vraag volgend, kunt u enkele voorbeelden geven van activiteiten die u verlangt te doen bij een vakantie naar het strand?

5. **Vind u het belangrijk dat de locatie dichtbij is, of gaat u liever verder weg? Waarom?**

5.1 Op vakantie, zit je liever lekker knus dicht bij huis, of zoek je liever het avontuur ver weg van Nederland, en waarom?

5.2 Vind je het belangrijk dat de locatie dichtbij is, of ga je liever verder weg, en waarom?

5.3 Vind u het belangrijk dat de locatie dichtbij is, of gaat u liever verder weg? Waarom?

5.4 Reist u liever naar een locatie dichtbij, of gaat uw voorkeur uit naar een verdere locatie? Waarom is dat zo?

5.5 Reist u graag naar een locatie in of nabij Nederland, of beoogt u liever een meer verafgelegen locatie? Waarom heeft u deze mening?

6. **Vind u dat een gids van belang is voor een vakantie? Waarom wel/niet?**

6.1 Vind je het handig om een gids te hebben op vakantie? Waarom wel of niet?

6.2 Vind je dat een gids belangrijk is voor een vakantie? Waarom wel of niet?

6.3 Vind u dat een gids belangrijk is voor een vakantie? Waarom wel of niet?

6.4 Is een gids voor u van belang op vakantie? Waarom wel of niet?

6.5 Is het voor u van belang om op vakantie begeleid te worden door een gids? Waarom vind u van wel of van niet?

7. **Wat voor accommodatie vind u het fijnst en waarom?**

7.1 Wat soort plaats verblijf je het liefste in als je op deze vakantie gaat en waarom?

7.2 Wat voor vakantieverblijf vind je het fijnst voor deze vakantie en waarom?

7.3 Wat voor vakantieverblijf vind u het fijnst voor deze vakantie en waarom?

7.4 Wat voor verblijf heeft u graag op deze vakantie en waarom?

7.5 Op welke wijze van accommodatie verblijft u graag op deze vakantie en waarom?

8. **Zijn er enige eisen die vastzitten aan het kiezen van de locatie van uw accommodatie?**

8.1 Zijn er dingen die je echt niet kunt missen bij de plek van jouw verblijf?

8.2 Heb je bepaalde eisen die vastzitten aan het kiezen van de plaats van jouw accommodatie?

8.3 Heeft u bepaalde eisen die vastzitten aan het kiezen van de plaats van uw accommodatie?

8.4 Heeft u specifieke vereisten bij het kiezen van de locatie van uw accommodatie?

8.5 Heeft u specifieke vereisten aangaande het selecteren van de locatie van uw accommodatie?

9. **Is de lokale cultuur van uw bestemming belangrijk voor u? Waarom wel/niet?**

9.1 Als jij op vakantie gaat, is de cultuur van jouw bestemming belangrijk voor je? Waarom wel of niet?

9.2 Is de lokale cultuur van jouw bestemming belangrijk voor je? Waarom wel of niet?

9.3 Is de lokale cultuur van uw bestemming belangrijk voor u? Waarom wel of niet?

9.4 Is de lokale cultuur van uw bestemming van belang voor u? Waarom wel of niet?

9.5 Is de oorspronkelijke cultuur van uw bestemming van groot belang voor u? Waarom wel of niet?

10. **Wij zijn bij het laatste deel van de vragenlijst uitgekomen, hoe vond u deze ervaring met de stem assistent?**

10.1 We zijn jammer genoeg bijna klaar! Zou je me kunnen vertellen hoe je de ervaring met de stem assistent vond?

10.2 Wij zijn bij het laatste deel van de vragenlijst aangekomen, hoe vond je deze ervaring met de stem assistent?

10.3 Wij zijn bij het laatste deel van de vragenlijst aangekomen, hoe vond u deze ervaring met de stem assistent?

10.4 U heeft het laatste deel van deze vragenlijst bereikt. Zou u met mij kunnen delen hoe u deze ervaring met de stem assistent vond?

10.5 U nadert het einde van deze vragenlijst. Zou u kort kunnen evalueren hoe u deze interactie met de stem assistent ervaren heeft?

11. **Zijn er nog speciale wensen voor deze vakantie waar wij op moeten letten?**

11.1 Zijn er misschien nog wat extra’s waar ik rekening mee kan houden voor deze vakantie?

11.2 Zijn er nog speciale wensen voor deze vakantie waar wij naar kunnen kijken?

11.3 Zijn er nog speciale wensen voor deze vakantie waar wij op moeten letten?

11.4 Heeft u nog bijzondere wensen voor deze vakantie waar wij aandacht aan moeten besteden?

11.5 Heeft u nog specifieke wensen voor uw vakantie waar aanvullende aandacht aan besteed dient te worden?

12. **Dit is het einde van deze vragenlijst, ik vond dat het erg goed ging. Bedankt voor de tijd die u heeft genomen, de onderzoeker zal u nu verder begeleiden.**

12.1 We zijn bij het einde van de vragenlijst, ik vond dat het super goed ging! Echt heel erg bedankt dat je de tijd hiervoor hebt genomen, de onderzoeker zal je zo iets laten weten!

12.2 Dit is het einde van deze vragenlijst, ik vond dat her erg goed ging. Bedankt voor de tijd die je hebt genomen, de onderzoeker zal je nu verder begeleiden.

12.3 Dit is het einde van deze vragenlijst, ik vond dat het erg goed ging. Bedankt voor de tijd die u heeft genomen, de onderzoeker zal u nu verder begeleiden.

12.4 Dit is het einde van deze vragenlijst, naar mijn mening is het erg goed gegaan. Hartelijk bedankt voor de tijd die u heeft genomen, de onderzoeker zal u nu verder begeleiden.

12.5 Dit is het slot van de vragenlijst, naar mijn mening is alles in goede orde verlopen. Hartelijk bedankt voor de tijd die u heeft genomen, de onderzoeker zal u nu verder begeleiden.

Afsluiten protocol:

13. **Heel erg bedankt voor uw medewerking!**

13.1 Harstikke leuk dat je mee hebt gedaan aan dit onderzoek, heel erg bedankt!

13.2 Heel erg bedankt voor je medewerking!

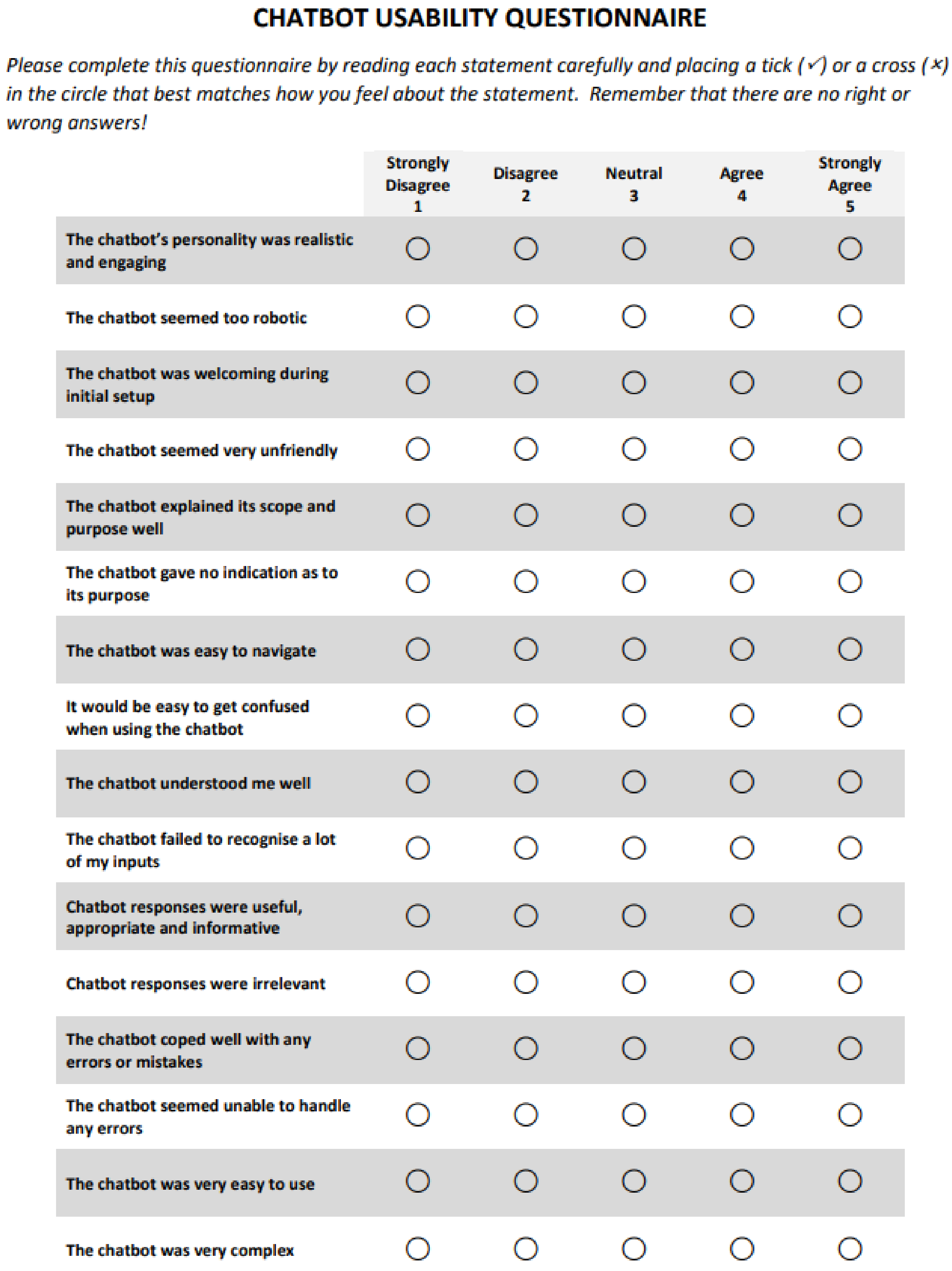
13.3 Heel erg bedankt voor uw medewerking!

13.4 Bedankt voor uw medewerking.

13.5 Ik dank u voor uw coöperatie.

APPENDIX B

Chatbot Usability Questionnaire (Holmes et al., 2019)



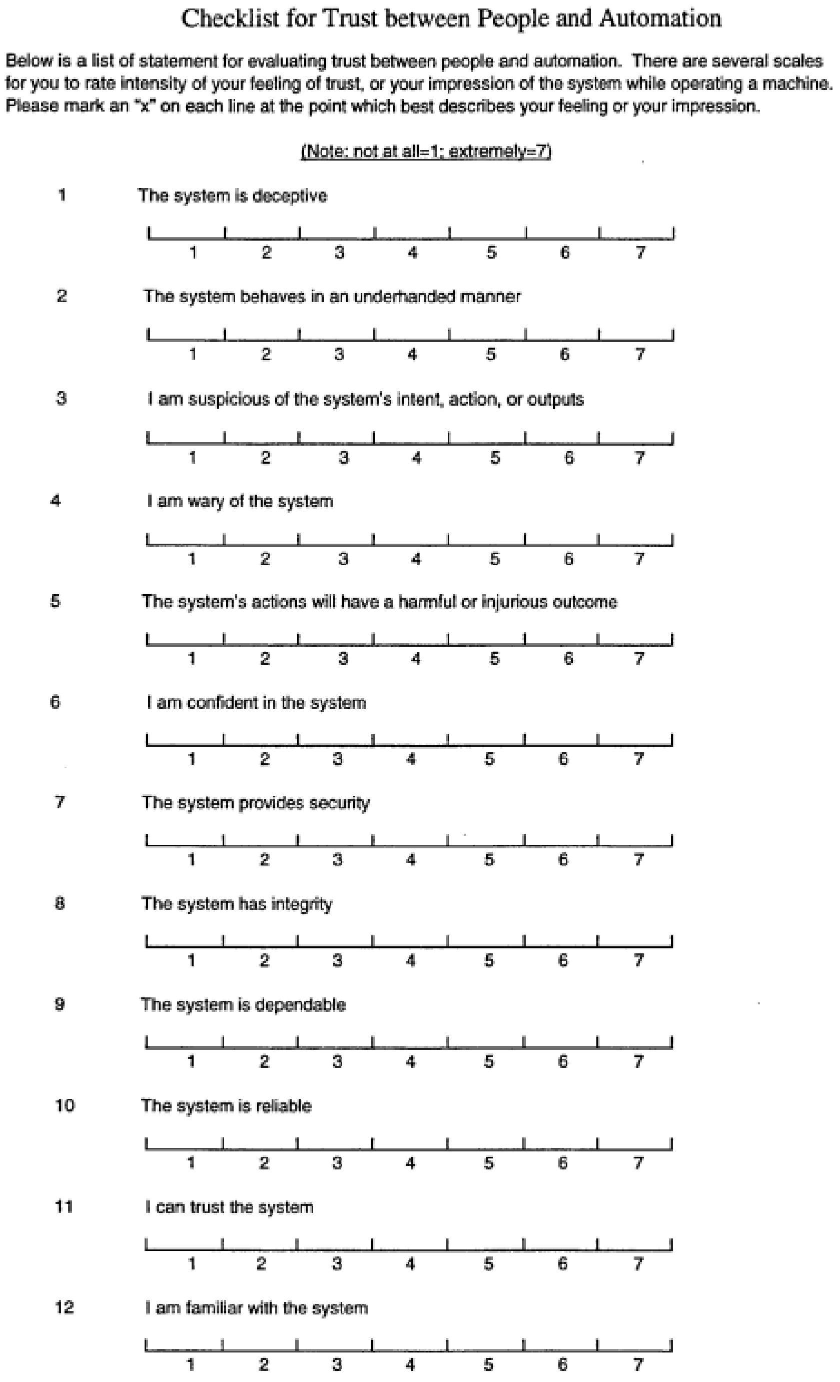
CHATBOT USABILITY QUESTIONNAIRE

Please complete this questionnaire by reading each statement carefully and placing a tick (✓) or a cross (×) in the circle that best matches how you feel about the statement. Remember that there are no right or wrong answers!

	Strongly Disagree 1	Disagree 2	Neutral 3	Agree 4	Strongly Agree 5
The chatbot's personality was realistic and engaging	○	○	○	○	○
The chatbot seemed too robotic	○	○	○	○	○
The chatbot was welcoming during initial setup	○	○	○	○	○
The chatbot seemed very unfriendly	○	○	○	○	○
The chatbot explained its scope and purpose well	○	○	○	○	○
The chatbot gave no indication as to its purpose	○	○	○	○	○
The chatbot was easy to navigate	○	○	○	○	○
It would be easy to get confused when using the chatbot	○	○	○	○	○
The chatbot understood me well	○	○	○	○	○
The chatbot failed to recognise a lot of my inputs	○	○	○	○	○
Chatbot responses were useful, appropriate and informative	○	○	○	○	○
Chatbot responses were irrelevant	○	○	○	○	○
The chatbot coped well with any errors or mistakes	○	○	○	○	○
The chatbot seemed unable to handle any errors	○	○	○	○	○
The chatbot was very easy to use	○	○	○	○	○
The chatbot was very complex	○	○	○	○	○

APPENDIX C

Checklist for Trust between People and Automation (Jian et al., 2000)



Checklist for Trust between People and Automation

Below is a list of statement for evaluating trust between people and automation. There are several scales for you to rate intensity of your feeling of trust, or your impression of the system while operating a machine. Please mark an "x" on each line at the point which best describes your feeling or your impression.

(Note: not at all=1; extremely=7)

1 The system is deceptive

1 2 3 4 5 6 7

2 The system behaves in an underhanded manner

1 2 3 4 5 6 7

3 I am suspicious of the system's intent, action, or outputs

1 2 3 4 5 6 7

4 I am wary of the system

1 2 3 4 5 6 7

5 The system's actions will have a harmful or injurious outcome

1 2 3 4 5 6 7

6 I am confident in the system

1 2 3 4 5 6 7

7 The system provides security

1 2 3 4 5 6 7

8 The system has integrity

1 2 3 4 5 6 7

9 The system is dependable

1 2 3 4 5 6 7

10 The system is reliable

1 2 3 4 5 6 7

11 I can trust the system

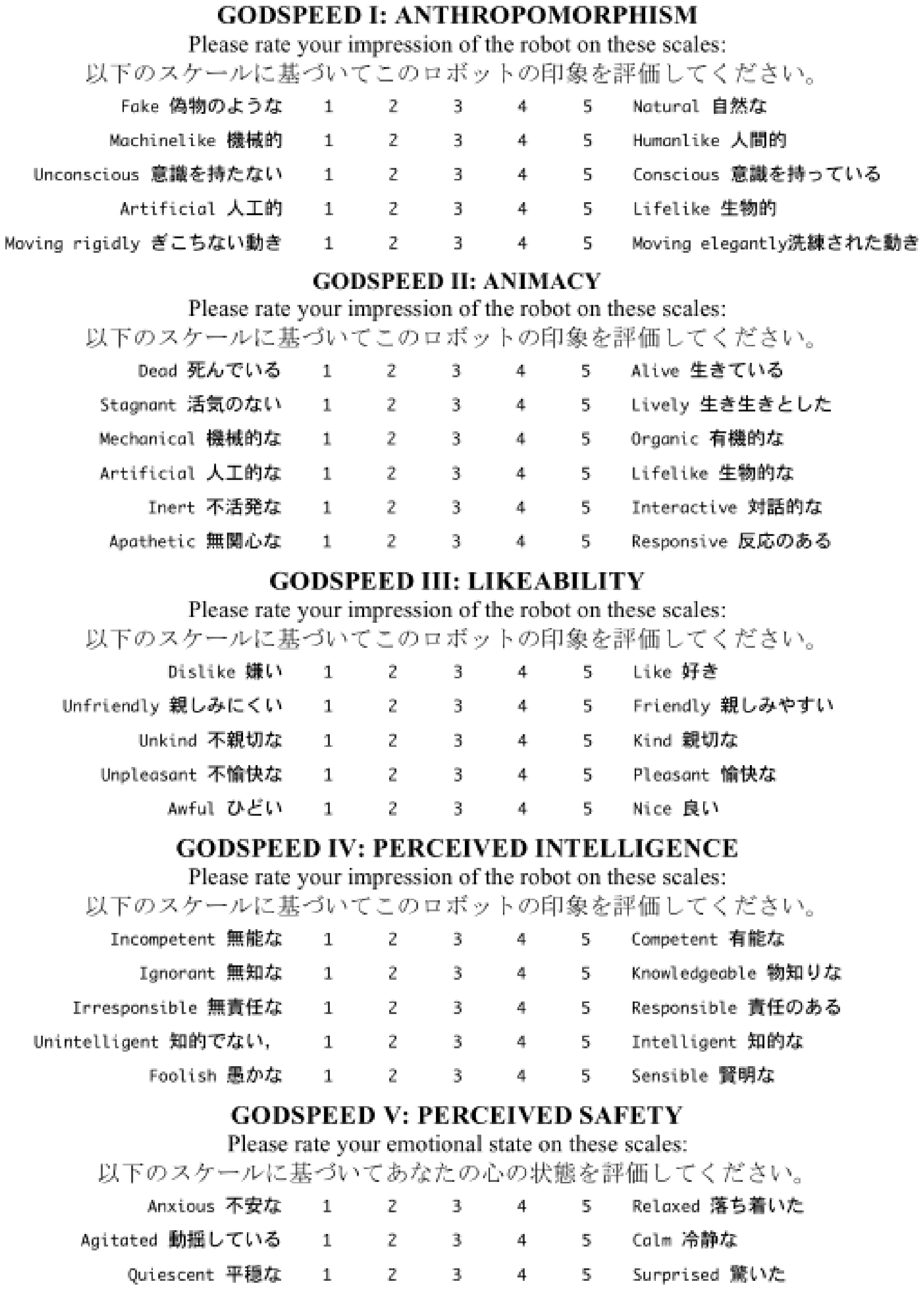
1 2 3 4 5 6 7

12 I am familiar with the system

1 2 3 4 5 6 7

APPENDIX D

Godspeed questionnaire (Bartneck et al., 2009)



GODSPEED I: ANTHROPOMORPHISM

Please rate your impression of the robot on these scales:
以下のスケールに基づいてこのロボットの印象を評価してください。

Fake 偽物のような	1	2	3	4	5	Natural 自然な
Machinelike 機械的	1	2	3	4	5	Humanlike 人間的
Unconscious 意識を持たない	1	2	3	4	5	Conscious 意識を持っている
Artificial 人工的	1	2	3	4	5	Lifelike 生物的
Moving rigidly ぎこちない動き	1	2	3	4	5	Moving elegantly洗練された動き

GODSPEED II: ANIMACY

Please rate your impression of the robot on these scales:
以下のスケールに基づいてこのロボットの印象を評価してください。

Dead 死んでいる	1	2	3	4	5	Alive 生きている
Stagnant 活気のない	1	2	3	4	5	Lively 生き生きとした
Mechanical 機械的な	1	2	3	4	5	Organic 有機的な
Artificial 人工的な	1	2	3	4	5	Lifelike 生物的な
Inert 不活発な	1	2	3	4	5	Interactive 対話的な
Apathetic 無関心な	1	2	3	4	5	Responsive 反応のある

GODSPEED III: LIKEABILITY

Please rate your impression of the robot on these scales:
以下のスケールに基づいてこのロボットの印象を評価してください。

Dislike 嫌い	1	2	3	4	5	Like 好き
Unfriendly 親しみにくい	1	2	3	4	5	Friendly 親しみやすい
Unkind 不親切な	1	2	3	4	5	Kind 親切な
Unpleasant 不愉快な	1	2	3	4	5	Pleasant 愉快な
Awful ひどい	1	2	3	4	5	Nice 良い

GODSPEED IV: PERCEIVED INTELLIGENCE

Please rate your impression of the robot on these scales:
以下のスケールに基づいてこのロボットの印象を評価してください。

Incompetent 無能な	1	2	3	4	5	Competent 有能な
Ignorant 無知な	1	2	3	4	5	Knowledgeable 物知りな
Irresponsible 無責任な	1	2	3	4	5	Responsible 責任のある
Unintelligent 知的でない,	1	2	3	4	5	Intelligent 知的な
Foolish 愚かな	1	2	3	4	5	Sensible 賢明な

GODSPEED V: PERCEIVED SAFETY

Please rate your emotional state on these scales:
以下のスケールに基づいてあなたの心の状態を評価してください。

Anxious 不安な	1	2	3	4	5	Relaxed 落ち着いた
Agitated 動揺している	1	2	3	4	5	Calm 冷静な
Quiescent 平穏な	1	2	3	4	5	Surprised 驚いた