



MASTER THESIS

Adding Speech to Dialogues with a Council of Coaches

Laura Bosdriesz
S1446673

MSC INTERACTION TECHNOLOGY

Faculty of Electrical Engineering Mathematics and Computer Science

EXAMINATION COMMITTEE

Dr. Ir. D. Reidsma (University of Twente, the Netherlands)

D.P. Davison MSc. (University of Twente, the Netherlands)

Prof. dr. D.K.J. Heylen (University of Twente, the Netherlands)

Dr. Ir. W. Eggink (University of Twente, the Netherlands)

Dr. Ir. H. op den Akker (Roessingh Research and Development, the Netherlands)

13 December 2020

UNIVERSITY OF TWENTE.

Abstract

With the ageing of the population, more diseases arise, putting pressure on the healthcare system. This requires a shift from treatment towards prevention of age-related diseases by stimulating the aging generation to take care of their own health. The Council of Coaches is a team of virtual coaches attempting help older adults to achieve health goals by offering insights and advice based on their expertise. Currently the user interacts with the coaches by selecting one of several predefined multiple-choice options. Although this a robust method to capture user input, it is not ideal for older adults. Spoken dialogues might offer a better user experience, but also comes with many complexities. The goal of this study is to adapt the COUCH system to support spoken interactions in order to answer the main question: *To what extent can spoken interaction offer a valuable addition to the multi-party virtual Council of Coaches application?*

User experiments are performed with the original text-based and the developed speech-based applications to research two different fields of interest: (1) the difference in experience between the two systems, and (2) the robustness of the speech implementation. In a controlled setting, 28 participants used both versions (i.e. a within-subjects design) for a limited amount of time in which the number of system errors was counted. Participants rated their experiences with both systems via questionnaires and open questions. This data was then analyzed to find differences between the two versions. During a one-week field study, the speech-based application is tested with 4 participants, who completed an interview in the end. These results are used to gain insights in the robustness of the application in a home setting.

Analysis of the collected data showed that the addition of speech led to a significant increase in some UEQ ratings (the novelty and stimulation scale). Additionally, the speech-version received significantly higher scores on several other items when explicitly comparing both systems. The field study revealed large fluctuations in user experiences, depending on the robustness of the speech recognition. In situations where the application worked properly, it was perceived relatively well. However, in situations where it worked insufficient, the application was perceived as cumbersome to use. Most but not all participants mentioned substantial problems in the speech recognition and the responsiveness of the system. Results from both experiments indicated the slow response speed of the application to be the main bottleneck of the experience, causing the feeling of miscommunication between human and machine.

Acknowledgement

This thesis marks the end of being a student at the University of Twente. I would like to take this opportunity to thank everyone who supported me during this master thesis and my master studies in general. First of all I wish to thank Dennis Reidsma and Daniël Davison, my academic supervisors from the University of Twente, who supervised me since the start of my graduation process. Your expertise on diverse fields allowed me to greatly improve the quality of my work. In particular I would like to thank Daniël for his patience with the application development, with which I had a difficult start. I also want to express my gratitude towards Dirk Heylen and Wouter Eggink, who agreed to join my examination committee. Last but not least I want to thank Harm op den Akker, my company supervisor, who offered me the possibility to carry out this assignment at Roessing Research and Development, even though it was a busy and difficult time due to Corona. Unfortunately I never had the chance to work on location and meet the entire RRD team, but I was very glad to have our weekly online meetings. Moreover, you provided me with extremely helpful feedback on my final report, especially on the structure and readability. During my project, I (online) met Dennis Hofs and Marian Hurmuz from Roessingh Research and Development, who I would like to thank for sharing their knowledge about the Council of Coaches and answering my questions.

Furthermore, I wish to express my gratitude towards all participants of the experiment, who took the effort completely voluntary participate in my experiment during a time that there was no need to be at the university.

Lastly, I would like to thank my family and friends for their support during this process. In particular, I want to thank Martijn and Birte, who were always willing to brainstorm and help with my project. Even though they were no experts in the subject, I could always approach them to hear their (outsider) view on the topic and discuss my ideas and developments during these isolated times. Their valuable opinions and feedback helped me to get to this final results.

Contents

1	Introduction	10
1.1	Council of Coaches	11
1.2	Why Speech?	12
1.3	Problem Statement	13
1.4	Approach	14
1.5	Document Structure	14
2	Theory	16
2.1	An Introduction in Conversational Interfaces	16
2.2	Conversation Mechanisms	17
2.3	The Technologies in Conversational Interfaces	17
2.3.1	Automatic Speech Recognition	18
2.3.2	Spoken Language Understanding	19
2.3.3	Dialogue Management	19
2.3.4	Response Generation	20
2.3.5	Text-to-Speech Synthesis	20
2.4	Limitations of Conversational Interfaces	21
2.4.1	Conversation Mechanisms	21
2.4.2	Naturalness of Speech	22
2.4.3	Speech Input Variations	22
2.4.4	Speech Synthesis for Older Adults	23
2.4.5	Expectations	23
2.4.6	Long-term Engagement	24
2.4.7	Privacy Issues	24
3	Related Work	25
3.1	In-home Social Support Agent	25
3.2	Kristina	26
3.3	Meditation Coach	26
3.4	Exercise Advisor	27
3.5	Implications for the Current Research	28
4	System design	29
4.1	The Council of Coaches Platform	29
4.1.1	The WOOL Dialogue Platform	29
4.1.2	The Coaches	30
4.1.3	The Council of Coaches Interface	31
4.1.4	The Dialogue Structure and Coaching Content	34
4.2	System Architecture	35
4.2.1	Automatic Speech Recognition	37
4.2.2	Dialogue Management	39

4.2.3	Spoken Language Understanding	41
4.2.4	Response Generation	41
4.2.5	Text-to-Speech Synthesis	41
4.3	Speech Synthesis Markup Language	42
4.4	Dialogue Design Strategies	43
4.5	Additional Features	44
4.6	Removed and Ignored Features	46
5	Methodology of Evaluation	47
5.1	Ethical permission	47
5.2	Controlled Experiment	47
5.2.1	Experimental Design	48
5.2.2	Hypothesis	48
5.2.3	Experimental setup	48
5.2.4	Measures	50
5.2.5	Participants	55
5.2.6	Procedure	55
5.2.7	Pilot Test	56
5.3	Field Experiment	57
5.3.1	Experimental Design	57
5.3.2	Experimental Setup	57
5.3.3	Interviews	57
5.3.4	Participants	58
5.3.5	Procedure	58
6	Results	60
6.1	Controlled Experiment	60
6.1.1	Observational Measures	61
6.1.2	User Experience	62
6.1.3	Explicit Comparison	67
6.1.4	Open Questions	69
6.2	Field Experiment	75
6.2.1	General Impressions	75
6.2.2	Practical Problems	76
6.2.3	Way of Interaction	77
6.2.4	Use and Recommendation of the Application	77
6.2.5	Suggestions for Improvements	77
6.2.6	Preference Version	78
7	Discussion	79
7.1	Discussion of the Sub-questions	79
7.2	Research Question 1	81
7.2.1	Answering Research Question 1	81
7.2.2	Discussion of the Results	81
7.2.3	Limitations of the Controlled Experiment	83
7.3	Research Question 2	84
7.3.1	Answering Research Question 2	84
7.3.2	Discussion of the Results	84
7.3.3	Limitations	85
7.4	Main Question	86
7.4.1	Answering the Main Question	86
7.4.2	Discussion of the Main Question	86

7.4.3	Comparison with Prior Research	87
7.5	Recommendations	88
7.5.1	Future Work	88
7.5.2	Application Improvements	89
8	Conclusion	92
	References	94
A	Questionnaires and interviews	99
A.1	Controlled Experiment	100
A.1.1	Intake Questionnaire	100
A.1.2	Observational Measurements	102
A.1.3	User experience	103
A.1.4	Explicit Comparison	104
A.2	Field Experiment	106
A.2.1	Intake Questionnaire	106
A.2.2	Interview questions	106
B	Forms and information	107
B.1	Controlled experiment	108
B.1.1	Information brochure	108
B.1.2	Informed consent	110
B.1.3	Coaches Sheet	111
B.2	Field Experiment	113
B.2.1	Invitation letter	113
B.2.2	Information brochure	114
B.2.3	Informed consent	116
B.2.4	Journal	117
C	Statistical Results	118
C.1	UEQ: Shapiro-Wilk test for normality check - text	119
C.2	UEQ: Shapiro-Wilk test for normality check - speech	119
C.3	UEQ: Paired samples t-test statistics for the comparison per scale	119
C.4	UEQ: Paired samples t-test results for the comparison per scale	120
C.5	Explicit Comparison: One-sample t-test statistics	120

List of Figures

1.1	The Council of Coaches living room user interface	11
2.1	The components of a spoken language conversational interface	18
3.1	In-home social support agent (Wizard of Oz)	25
3.2	The KRISTINA prototypes	26
3.3	The Meditation Agent	27
3.4	The FitTrack interfaces	28
4.1	The WOOL editor	30
4.2	The Council of Coaches scripted basic reply	32
4.3	The Council of Coaches scripted autoforward reply	32
4.4	The Council of Coaches scripted input reply	32
4.5	Council of Coaches Main Menu screen	33
4.6	First screen of the account creation process	33
4.7	The hierarchy of coaching topics	34
4.8	Visualization of the speech-based COUCH architecture	36
4.9	The basic reply sequence diagram	38
4.10	The autoforward reply sequence diagram	39
4.11	An example of the addition of a generic action-Statement to a reply option	39
4.12	An example of keyword tags, added to different answer options.	40
4.13	An example of keyword tags for positively and negatively phrased answers	40
4.14	The start and stop buttons	45
4.15	The record button	45
4.16	The megaphone button	45
5.1	The experimental setup of the controlled experiment.	50
6.1	Mean values per item ranked for the text-version.	64
6.2	Mean values per item ranked for the speech-version.	64
6.3	Visualization of the mean and variance of the UEQ scales	66
6.4	Boxplot of the explicit comparison results.	68
7.1	Automatically generating keywords example	90

List of Tables

4.1	The seven coaches from the Council	31
4.2	Descriptions of the coaching topics	35
5.1	Demographics of the participants.	55
6.1	The number of different error types and relative percentage	61
6.2	The descriptive statistics of the observational measurements.	62
6.3	UEQ scale mean and variance	63
6.4	Conbach's alpha coefficient	65
6.5	Results of the t-test performed for the explicit comparison	69
6.6	Reasons that were mentioned in favor of the text-version.	70
6.7	Reasons that were mentioned in favor of the speech-version.	70
6.8	Reasons that were mentioned for ease of use text-version	71
6.9	Reasons that were mentioned for ease of use speech-version	71
6.10	Reasons that were mentioned for most fun to use text-version	72
6.11	Advantages experienced in the text-version.	73
6.12	Advantages experienced in the speech-version.	73
6.13	Disadvantages experienced in the text-version.	74
6.14	Disadvantages experienced in the speech-version.	74

Abbreviations

ASR	Automatic Speech Recognition
COUCH	Council of Coaches
DM	Dialogue Manager
NLU	Natural Language Understanding
RG	Response Generation
RRD	Roessingh Research and Development
SDS	Spoken Dialogue System
SSML	Speech Synthesis Markup Language
SLU	Spoken Language Understanding
TTS	Text-to-Speech Synthesis
UEQ	User Experience Questionnaire
VPA	Virtual Personal Assistant
VUI	Voice User Interface
WER	Word Error Rate

Chapter 1

Introduction

This thesis is the continuation of the preliminary literature research 'Opportunities and Challenges for Adding Speech to Dialogues with a Council of Coaches' [1] that already investigated opportunities and challenges for adding speech to the Council of Coaches application. This preliminary research focused on the application's current limitations, solutions to overcome them, problems associated with speech implementation and pitfalls for speech recognizers. Since this thesis is a continuation on the work described in the literature research [1], it partly reuses Chapter 1 (introduction), 2 (theory) and 3 (related work). On the other hand, information that was not considered relevant for this thesis is removed from these chapters. Except from the main question, the problem statement (including sub-question and research questions) has changed, as well as the approach and document structure.

Population aging is a phenomenon that has been evident for several decades in Europe. The population of people older than 65 years is expected to increase from 101 million in 2018, to 149 million by 2050. This increase is even larger in the older population aged 75-84 years (60.5%), compared to the population aged 65-74 years (17.6%) [2]. Since aging increases the risk of age-related diseases and decline in function, many of these additional years will be lived with chronic diseases [3,4]. Additionally, older adults suffer from functional deterioration, revealed in decreased mobility, and vision and hearing loss [5]. The population aging and its related health-issues is likely to have a considerable impact on healthcare. This requires a shift from treatment towards prevention of age-related diseases by enabling the aging generation to stay independent longer and stimulate them to take care of their own health and condition. Research showed that innovative solutions in the area of electronic health (e-health) can be useful in personalizing the care provided [6-8].

With the advancements in digital healthcare, health coaching can be provided by virtual coaches. Virtual coaches can take the form of computer characters, running on web-based platforms or smartphone applications. Some examples of virtual health coaching systems are presented in Chapter 3. Personalized coaching uses strategies applied to the user's personal characteristics such as perceived barriers, personal goals and health status. Coaching interventions that are aimed at sending reminders, tracking goals, or providing feedback are designed for one individual [9]. Given that adequate coaching for older adults is important to reduce the pressure on the healthcare system, solving this by means of human coaches is not feasible and scalable to the required level. E-health technologies using virtual coaches, provide a good infrastructure for personalizing and tailoring the intervention.

Examples from literature show that personalized and virtual coaching in healthcare can be done by providing more health related information to older adults and using virtual personal

coaching, counseling and lifestyle advice to persuade and motivate them to change their health behaviors [10]. It has been investigated for some time already, especially to support patients with chronic conditions [11,12]. For example, Klaassen et al. [13] developed a serious gaming and digital coaching platform supporting diabetes patients and their healthcare professionals. Although those single coach systems has already shown a positive effect, a better performance in health coaching is expected to be achieved through a multi-agent virtual coaching system [10,14]. For this reason, the “Council of Coaches” (COUCH) revolved around the concept of multi-party virtual coaching.

1.1 Council of Coaches

Council of Coaches¹ is a European Horizon 2020 project developed by Roessingh Research and Development (RRD) to provide multi-party virtual coaching for older adults. The project aimed to improve their physical, cognitive, mental and social health and to encourage them to independently live healthy with help from a council of coaches [15]. The council consists of a number of coaches, all specialized in their own specific domain. They interact with each other, and also listen to the user, ask questions, inform, jointly set personal goals and inspire the users to take control over their health and well-being.

One of the objectives of the project was to develop coaches as interesting characters. This character design is reflected mainly by providing every coach with its own background story and related personalities. Any combination of specialized council members collaboratively covers a wide spectrum of lifestyle interventions, with the main focus on age-related impairments, chronic pain, and Diabetes Type 2. The project includes seven coaches and a robot assistant, who leads the interaction between the user and the system. All coaches and the robot assistant have their own place within the Council of Coaches living room based on their expertise (see Figure 1.1). Users are provided with an interface using buttons with scripted responses to interact with the coaches. Although this is a reliable way to capture input from the user, it is not ideal for older adults because they generally experience more difficulties reading and have less computer experience. This research attempts to discover if spoken dialogues within the Council of Coaches can offer a better user experience.



Figure 1.1: The Council of Coaches living room User Interface. From left to right: *peer support, physical activity coach, social coach, diabetes coach, cognitive coach, chronic pain coach, robot assistant and nutrition coach*. Figure reproduced from [16].

¹<https://council-of-coaches.eu/>

1.2 Why Speech?

As described in the previous section, COUCH is a text-based application using multiple written options to choose from in order to interact with the coaches. Other input options for such an e-health application could be free text or speech. Free text is different from the restricted approach of COUCH in the sense that users have the opportunity to type anything they want during an interaction. The benefit of the approach COUCH takes, is that coaches are certain about what they are responding to, which is more difficult when having to deal with free text or speech. There are some specific features that distinguish the language of conversation from the language of written text, causing some complex issues (discussed in more detail in Chapter 2). Nevertheless, at the same time speech brings many additional advantages, especially for older adults. Potential benefits can be found in the level of engagement, the maintenance of long-term relationships, to solve the loneliness problem, and to overcome physical barriers of the aging process.

Speech can contribute to one of the major challenges in e-coaching, which is to keep the user engaged for a longer period of time [17]. Turunen et al. [17] argue that if there is no long-lasting engagement, the health coach cannot have any further impact on the behavior change. They also found that building long-term relationships between the user and the system can benefit the level of engagement. Finally, Turunen et al. [17] obtained positive results for building a social and emotional human-computer relationship with a physical presence of an interface agent, using spoken, conversational dialogues. Other results from literature review in the field of virtual health coaching showed that speech-based virtual characters can improve user satisfaction and engagement with a computer system [18]. Both studies suggest a potential benefit for the implementation of a spoken conversational interface. Speech might increase engagement because it can make the human-computer interaction more natural, thereby improving the effects of coaching. On the other hand, it might decrease trust because it can create expectations of the system which may not be fulfilled.

Another area where speech can contribute is in the field of social companion and peer support. Loneliness is a common problem in today's older population, and since it is closely associated with depression [19], it is important for older adults not to feel lonely. COUCH contains a peer and support character, who also takes advice from the coaches and is there to share his experiences with the user from an equal friend viewpoint. Additionally, there is a social coach who can help the user with tips and advice on having a socially active life. Implementing speech in these characters might improve the effectiveness of their role as social companion. These characters can act as virtual friend when a more natural conversation in a home setting takes place, possibly decreasing the feeling of loneliness among older adults. Besides participating in conversations, virtual friends can read books or other long-form documents to help the users [20].

One last big advantage of implementing speech into an application for older adults, is to overcome barriers to access information. For many users, especially older adults, the ability to read and type decreases the usability and ease of use of an application. Conversational interfaces can bridge this gap by allowing them to talk to the system [20] and thereby avoiding manual input methods like the keyboard and mouse. This makes it a comfortable and efficient input method for older adults with physical disabilities and function-loss [21]. Additionally, younger and older adults differ in the way they interact with technology whereby the latter group generally faces more difficulties interacting with computers. Literature showed that speech could be one of the most natural and effective modalities to overcome older adult's problems related to their attitudes and experience with technology in general [22].

1.3 Problem Statement

Considering these promising advantages, especially for older adults, this research aims to assess the possibilities of implementing speech in the COUCH application which is not designed to function as speech interface. The biggest challenge lies in handling the dialogues. Adjustments to the dialogue structure are required to allow users to listen and speak to the coaches instead of reading and clicking to navigate through the dialogue. Thereby two problems need to be tackled. First, the design of an appropriate spoken dialogue need to be researched to create an application that might improve the user experiences. Second, the benefits and drawbacks of such a speech-based system need to be addressed, by obtaining user feedback. This research is an addition to the work described in Section 4.1 and does not aim to develop a conversational interface that is perfectly able to imitate human-to-human natural conversations, but instead, to investigate whether the state of art is developed enough to create a reliable and usable conversational system in a daily life setting. Hereby smart ways to handle the spoken dialogues in the context of COUCH are investigated that can contribute to the reliability and usability of the conversational system. It will attempt to find evidence that this is indeed an area of promise and that users experience additional advantages in a speech-based COUCH implementation. This goal is formalized in the following research question:

MQ: To what extent can spoken interaction offer a valuable addition to the multi-party virtual Council of Coaches application?

Spoken interfaces can be one-sided, which means that there is only audio input spoken by the user or only audio output spoken by the virtual character, but also two-sided, which means that both the user and the system can talk. This research focuses on the two-sided vocal interaction because we expect this type of interaction to be more interesting for the user. Therefore, a component to transcribe the spoken audio in the text, as well as a component to transform written text in output audio are required. The prior research [1] addressed the potential for such a system based on literature, but mainly focused on the recognition of spoken speech input via the automatic speech recognizer and not much on the spoken speech output via the text-to-speech synthesizer. Therefore, a couple of extra sub-questions regarding the text-to-speech synthesis are part of this thesis, in order to develop a system that is capable of showing the value of speech. Except from focusing on the speech recognition and speech synthesis, it is important for an application like this to investigate what additions to the graphical user interface are necessary and how the dialogues should be structured in order to function in the existing COUCH application. Therefore, we created the following sub-questions:

- SQ:1** What are current limitations in text-to-speech synthesis software and how can this problem be addressed?
- SQ:2** How robust is the current state of art in text-to-speech synthesis to create multiple humanlike voices necessary to ensure a natural interaction with all coaches?
- SQ:3** What additions to the graphical user interface are necessary to assure a pleasant user experience with the voice-based application?
- SQ:4** How should the dialogues be adjusted in order to function with the implementation of speech into the Council of Coaches?

These sub-questions are largely answered before the development phase and serve as base for the final system design. This final system is tested to answer the research questions, defined as:

- RQ:1** Does the addition of speech to the Council of Coaches application lead to an increase in user experience?
- RQ:2** How robust is the current state of art in speech recognition systems to create a usable and enjoyable system that can be used in a real-life (home) setting?

A controlled experiment is designed to compare the user experiences between the text- and speech-based applications and to provide an answer to research question 1. A field experiment is done to assess the user experiences when the application is used in a home-setting, where it is tested for its robustness, as defined by research question 2. The collection of these results helps to find and answer to the general research question.

1.4 Approach

In order to find answers to the questions posed in the previous section, this project will integrate an automatic speech recognizer and text-to-speech synthesizer into the existing multi-party Council of Coaches application that focuses on supporting older adults to live healthy. Additionally, smart strategies to design the spoken dialogues and manage the user interactions will be assessed. A detailed description of the development of this system can be found in Chapter 4. This system will then be put to the test in two user experiments. During the first user test, users will be presented with two versions of the system, one version where they can interact via speech, and one where they interact via reading and clicking. During the second user study, four older adults will use the speech-based system in-home for one week. Obtained user data and questionnaires filled out by all test subjects will then be analyzed in order to answer the research questions.

1.5 Document Structure

This section provides an overview for each of the following sections of the complete thesis.

2. Theory

In this chapter, first an introduction in conversational interfaces is given, including all its related concepts. This is followed by an introduction in conversation mechanisms and the chapter ends with a small review on the technologies involved in conversational interfaces and the challenges of implementing natural speech.

3. Related work

In this chapter, the practice of several virtual health coaching systems is explored. It includes a description of the advantages and limitations experienced in these related works. This information can help to anticipate pitfalls before the implementation of speech.

4. System Design

This chapter is divided in two parts. The first part gives an overview of the original COUCH application discussing the coaches, interface, coaching structure and content, and the WOOL platform, which is a simple, powerful dialogue platform for creating virtual agent conversations. The second part provides the technical details and design of the system developed for this research. It starts by introducing the system's architecture, followed by an explanation of each conversational component. It also described the strategies

for managing the dialogues and adjustments made to dialogues, guided by WOOL.

5. Methodology of Evaluation

In this chapter the methodology for two experiments is described. This methodology includes the experimental design, set-up, the questionnaires, the statistical analyses, an overview of the participants and the final procedure.

6. Results

In this chapter the results from the data analyses is presented for all collected data by both experiments. Then these results are used to draw any conclusions from the experiment and to answer the research questions.

7. Discussion

In this chapter, the outcomes of the experiments are discussed and positioned into existing literature. The results are interpreted to answer research question 1, 2 and the main question. Strengths, as well as limitations of the study are discussed, followed by future research directions.

8. Overall conclusions

The last chapter provides a summary of the main findings of this thesis and summarizes the answers to the research questions.

Chapter 2

Theory

The content of Chapter 2 is also for a large part reused from the literature research [1]. However, this preliminary research mainly focused on the automatic speech recognition component and less on the speech synthesizer component. Both components are important for the design of a conversational interface and therefore more research is done in the field of text-to-speech synthesizers. The added material includes the technical details and limitations regarding speech synthesizers. Furthermore, the literature research extensively investigated the state-of-art in automatic speech recognizers, while in this thesis one summarizing paragraph about the chosen speech recognizer is added.

In this chapter, an introduction in conversational interfaces and conversation mechanisms is provided. The remainder of the chapter gives an overview of the technologies involved in conversational interfaces and their difficulties and technical challenges of implementing natural speech. This section partly answers sub-questions 1 and 2.

2.1 An Introduction in Conversational Interfaces

Conversational interfaces enable people to interact with smart devices using spoken language in a natural way—just like engaging in a conversation with a person. A conversational interface is a user interface that uses different language components (see Section 2.3) to understand and create human language that can help mimicking human conversations [23].

The concept of conversational interfaces is not very new. According to McTear, Callejas and Griol [23], it already started around the 1960s with text-based dialogue systems for question answering. Somewhat later, around the 1980s, the concept of spoken dialogue systems has become important within the speech and language research. They mentioned that spoken dialogue systems (SDS) and voice user interfaces (VUI) are different terms for somewhat similar concepts (i.e. they use the same spoken language technologies for the development of interactive speech applications), although the difference is in their purpose of deployment. SDS have been developed for academic and industrial research, while at the same time VUIs have been developed in a commercial setting [23]. These systems are intelligent agents that use spoken interactions with a user to help them finish their tasks efficiently [24]. Academic systems often use embodied conversational agents (ECAs) [23], implemented with a more structured dialogue setting. ECAs are computer-generated characters with an embodiment, used to communicate with an user to provide a more humanlike and more engaging interaction. The main benefit of ECAs is that they allow human-computer interaction in the most natural possible setting, namely with gesture, body expressions and speech to enable face-to-face communication with

users [18]. A few examples from literature are described in Chapter 3. The commercial interfaces (VUIs) often not take a dialogue structure, but instead use a spoken command and response interface. This means that an interaction with these systems start with a request from the user, which is processed by the system. The system generates an answer and sends this answer back to the user. Examples of such commercial interfaces include Apple’s Siri, Google’s Assistant, Microsoft’s Cortana, Amazon’s Alexa, Samsung’s S Voice, Facebook’s M, Baidu’s Duer, and Nuance Dragon. For the ease of reading, we will use the term ‘*spoken dialogue systems*’ in the remaining of this report to comprise both the SDS and VUI.

2.2 Conversation Mechanisms

The main objective of a conversational interface is to support conversations between humans and machines. Understanding how human conversations are constructed is an important aspect in the development of a conversational interface. In general, participants take turns according to general conventions (turn-taking), they collaborate for mutual understanding (grounding), and they take measures to resolve misunderstandings (conversational repair) [23]. Some design issues for conversational interfaces come from the complexity of implementing these conversation mechanisms in the dialogue manager, explained in Section 2.3.3.

- | | |
|------------------------------|--|
| Turn-taking | Informally it can be described as “stretches of speech by one speaker bounded by that speaker’s silence – that is, bounded either by a pause in the dialogue or speech by someone else” [25]. |
| Grounding | The process of reaching mutual understanding between participants and keeping the conversation on track, for example by providing feedback or adding information [26]. In designing conversational interfaces it is important to know how understanding can be achieved, but also how misunderstanding may arise and how can be recovered from the communication problems. |
| Conversational repair | The process for repairing failures in conversations through various types of repair strategies, initiated by either the speaker or the interlocutor [23]. |

2.3 The Technologies in Conversational Interfaces

The major components of dialogue systems are: Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU), Dialogue Management (DM), Response Generation (RG) and Text-to-Speech Synthesis (TTS) [23, 27]. The steps involved in a conversational interface are as follows:

1. The ASR component processes the words spoken by the user in order to recognize them.
2. The SLU component retrieves the user’s intent from those words.
3. The DM component tries to formulate a response or, if the information in the utterance is ambiguous or unclear, the DM may query the user for clarification or confirmation.
4. The RG component constructs the formulated response, if desired.
5. The TTS component is utilized to produce the spoken response.

An overview of a complete spoken language conversational interface is shown in Figure 2.1.

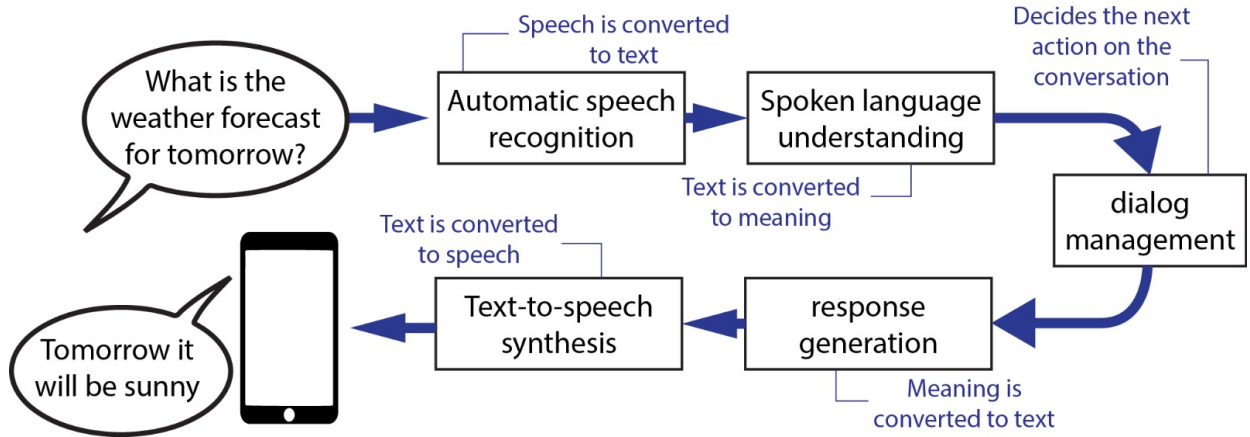


Figure 2.1: The components of a spoken language conversational interface. Figure constructed based on the steps described in [23].

2.3.1 Automatic Speech Recognition

Automatic speech recognition (ASR) is the process of recognizing what the user has said by transforming the speech into text [23]. This is done by decoding the audio input and come with a best guess transcription of the words being spoken. There are three types of ASRs: (a) speaker dependent, (b) speaker independent, and (c) speaker adaptable. Those systems differentiate in the amount of user training required prior to use. Speaker-dependent ASR requires training with voices of speakers prior to use, while speaker-independent does not. The latter systems are pre-trained during the system development with data from many speakers. In general, speaker-dependent ASR systems achieve better accuracy than speaker-independent systems, although the latter systems still have a relatively good accuracy when a user's speech characteristics fall within the range of the system's collected speech database. For example, a speaker-independent system trained with data from Western male voices, in the age range of 20-40, should work for e.g. any 30-year old Western men, but not for an Asian 85-year old women. Speaker-adaptable ASR is similar to speaker-independent ASR, but with the difference that adaptable ASR improves accuracy by gradually adapting to the user's speech [28]. However, since the primary goal of ASR research has been to create systems that can recognize spoken input from any speaker with a high degree of accuracy, and speaker-dependent systems are very time and effort consuming, most commercial systems are independent. Unfortunately, voices of older adults do often not fall within the range of collected speech used for the development of commercial systems. This may cause speaker-independent ASRs to not work optimally for the target group of our research. Group-dependent ASR might be a better approach, but requires a lot of data from older speakers to train the ASR [29]. Moreover, the focus of our research is not to develop the best performing ASR, but to test an implementation with a current state-of-art ASR.

We will not go into much detail in all possible ASR systems, since these are extensively researched already during the literature research preliminary to this thesis. During this previous study, the following conclusion had been drawn: "Due to the main goal involved, the considered ASR was chosen based on the following criteria: their performance, ease of use, documentation, availability of the system and language support in Dutch. Based on these criteria, the choice has fallen for the NLSpraak toolkit. This toolkit wins on ease of use, availability of the system and language support. There is not much research in its performance, but since the toolkit is build upon the Kaldi toolkit [30] that has shown to perform quite well, we expect NLSpraak's performance to be solid for this project".

2.3.2 Spoken Language Understanding

Spoken language understanding (SLU) involves the interpretation of the semantic meaning, conveyed by the ASR transformed spoken input. Traditional SLU systems contain an ASR and a natural language understanding (NLU) component. The ASR component splits the audio input into small frames and the NLU component, on its turn, provides a semantic label to each of these segments. This means that the fragments of the sentences are labeled with, for example, noun, verb, determiner, preposition, adverb or adjective. These semantic meaning of the different fragments are used to understand the complete input sentence, which can then activate the subsequent behavior in a human-computer conversation. Since the SLU component uses an ASR, its performance largely depends on the ASR's ability to correctly process speech, with a word error rate (WER) [31]. The WER is the percentage of words that are missed or mistranslated by the ASR. Completely parsing the grammar of a sentence only functions when the ASR is close to perfect with a very low WER. Nevertheless, the NLU performance can be made more robust against a bad performing ASR by shaping the NLU appropriately. In general, most practical conversational interfaces tried to achieve this robustness by using sets of attribute-value pair representations to capture the information relevant to the application from the speech input [23]. These attributes are pre-defined concepts related to the field of the application, while the values are the attribute specifications. For example, when looking for a health care institution with the specifications; "a dentist close to my house", the health institution 'type' and 'location' are the attributes, while the 'dentist' and 'near' are the value. This approach is robust as long as the right keywords can be retrieved.

2.3.3 Dialogue Management

Dialogue management (DM) relies on the fundamental task of deciding what action or response a system should take, based on the semantic meaning of the user input. This user input and output can be either a textual or vocal response [32], with the former being the case for the original COUCH application. It is an important part of the conversational interface design, given that this component entails all its dialogue structures and content. In addition, the dialogue manager is the main responsible for user satisfaction because its actions directly effect the user. Each DM tool depends on the specific characteristics of the dialogue system type it has been created for. These characteristics include the task, dialogue structure, domain, turns, length, initiative and interface. Additionally, DM tools differ in the way they can be authored. With some tools, the dialogue strategy is in the hands of the dialogue author, while in others it is in the hands of the programmer because it requires programming expertise to adjust some of the general built-in dialogue behaviors. For the development of the original COUCH application, RRD developed its own dialogue framework "WOOL" with as goal to make it accessible for non-technical dialogue authors. This platform is described in more detail in Section 4.1.1.

As already mentioned in Section 2.2, two frequent design issues within dialogue managers for conversational agents are the interaction and confirmation strategies [23]:

- Interaction Strategies** Determine who takes the initiative in the dialogue – the system or the user? There exist three types of interaction strategies: User-directed (i.e. user is leading the conversation), system-directed (i.e. system is leading the conversation by requesting user input) and mixed-initiative (i.e. both the user and system can take the initiative), all having their own advantages and disadvantages [32]. The advantage of user-directed is that the user can say whatever it wants to the system, creating the feeling of a natural conversation. On the other hand, the system might be prone to errors because it cannot handle and understand all conversation topics. This problem can be overcome by using a system-directed approach. While constraining the user's input, less errors will be made because the user has to behave according to the system's expectations. At the same time, this creates a less natural experience. Some middle way between the user- and system- directed strategies, is the mixed-initiative strategy, where the system can guide the user but where the user additionally can start new topics and ask questions.
- Confirmation strategies** Deal with uncertainties in spoken speech understanding. Two types of confirmation strategies exist: explicit (i.e. the system takes an additional conversation turn to explicitly ask for confirmation) and implicit confirmation (i.e. the system integrates part of the previous input in the next question to implicitly ask confirmation with its next question). The former confirmation has as disadvantage that the dialogue tends to get lengthy and interaction less efficient. The latter is more robust to this problem, but can cause more interpretation errors when the user did not catch the implicit confirmation request.

2.3.4 Response Generation

Response generation (RG) is the process following up the dialogue manager's response decision. The conversational interface has to determine the content of the response and an appropriate method to express this content. The content can be in the form of words or it can be accompanied by visual and other types of information. The simplest approach is to use predetermined responses to common questions. RG is commonly used for SDSs to retrieve structured information (e.g. Who is the king of the Netherlands?). This involves translating the structured information retrieved from a database into a form suitable for spoken responses. RG is more complex for systems like the Google assistant and less for a system like COUCH, which has relatively simple response generation. In the COUCH application, most responses are pre-scripted, using simple lookup tables and template filling. These templates can be dynamically filled with information about the interaction, where the coaches' possible text-to-speech sentences are considered.

2.3.5 Text-to-Speech Synthesis

Text-to-speech synthesis (TTS) is the process of synthesizing words generated by the RG component into spoken speech. TTS is closely related to ASR, since both systems need to accurately work together for a speech-based conversational interface to function effectively. A TTS is composed of two components: the front-end and back-end [33]. The back-end component is responsible for the normalization of words like numbers and abbreviations. The front-end component is responsible for assigning a phonetic transcription to parts of a word and then combines those to output a spoken sentence. There exist many different synthesizer technologies for this process, each trying to attempt naturalness (i.e. the similarity of output to human speech) and intelligibility (i.e. the ease with which the output is understood). TTS is used in applications where messages cannot be prerecorded but have to be synthesized in the moment [23].

Challenges of TTS can be divided in the text-normalization challenge and text-to-phoneme (a phoneme is a distinctive sound in a language) challenge [33]. The text-normalization challenge relies in deciding how to convert numbers and abbreviations, which both can be ambiguous dependent of its context. This challenge will be addressed in the current research by using the Speech Synthesis Markup Language (SSML) specifications¹, which is designed to provide a rich markup language for assisting the generation of synthetic speech in Web and other applications. This language allows the programmer to manually instruct the TTS about the required text-normalization in a specific context. Also exceptional mistakes can be adjusted via this language, whereas the rest of the text-to-speak is left the same. One important requirement for the implementation of SSML is that it has to be supported by the TTS. The text-to-phoneme challenge comprises the determination of correct pronunciation of a word based on its spelling, wherefore two basic approaches are used. The simplest dictionary-based approach is a matter of looking up each word in the dictionary and replacing the spelling with the pronunciation specified in the dictionary. The other rule-based approach works via pronunciations rules that are applied to words based on their spelling. This latter challenge is much more difficult to address because it is part of the implementation of the TTS.

The final requirements for the TTS necessary for this research included:

- The availability of six different voices
- The support of the Dutch language
- Available for free
- Although no hard requirement, the option for using SSML was appreciated.

2.4 Limitations of Conversational Interfaces

Although ASR technology is useful in a wide range of applications, it is never likely to be 100% accurate. One big difference between written language and spoken language is that spoken language is much more spontaneous compared to written text. Written text is grammatically correct, while spoken speech often is not. Complexities regarding the user characteristics of older adults, conversational mechanisms (i.e. the processes of turn-taking, grounding and conversational repairs), dialogue structure and speech input variations make the recognition of spoken language a complex process [25]. The limitations of conversational interfaces and its expected effects on the COUCH system are discussed in this section. Additionally, a suggestion on how to deal with the problem is given for every limitation.

2.4.1 Conversation Mechanisms

Spoken speech requires more conversation mechanisms than written text. Mechanisms such as turn-taking, grounding and conversational repair are much more complex to implement in a conversational interface. In the original COUCH application, users have to choose between several multiple choice text-input options. This eliminates the need for conversational repair and it simplifies grounding. The computer system always understands the user and when the user does not understand the computer, it can ask for repetition or more clarification via one of the prewritten input options. When implementing speech in such application, this process will become more complex. The ASR can misunderstand or not identify the spoken speech input. One way to improve the experience when such problems occur is to design good conversation mechanisms by including, for example, confirmation strategies (i.e. strategies to deal with uncertainties in spoken speech understanding). These strategies are implemented in the speech-based COUCH system and presented in Section 4.4.

¹<https://www.w3.org/TR/speech-synthesis11/>

2.4.2 Naturalness of Speech

The naturalness of spoken speech is difficult for conversational interfaces to deal with. COUCH takes a hierarchical dialogue structure (see Section 4.1.4 for a detailed description) where the dialogue follows a structured path based on the user input. When implementing speech to this dialogue system, the system should be able to deal with multiple spoken input sentences. For example, when a user asks about the positive effects of physical activity, the question can be phrased like:

- "Can you tell, eeh ..., tell me about the positive effects of physical activity, please?"
- "Why do I need to exercise more?"
- "What are the advantages of exercising?"
- "I am not into, .. I mean, don't like to be physically active, so why should I?"

One way to handle the speech is by retaining the multiple choice structure. In this way the system could "guess" which option is chosen by the user, based on its speech input. In the example above, the system has to understand that the option "positive effects" is chosen with each of the example inputs and that it has to mention the advantages of being active. However, this approach is complex since then the system has to understand all user input, also input which is not related to the coaching topics. An easier approach is to provide the user with a restricted list of input sentences, but this decreases the naturalness of the interaction.

Compared to the ASR and SLU components that experience problems with the naturalness of spoken input speech, the TTS has one of its fundamental limitations in the naturalness of the spoken speech output. Written text does not contain any emotions [6], constituting a complex domain for synthesized speech. There are no concrete parameters to classify the emotions expressed in synthesized word, as compared to ASR systems, that can use the WER as such a parameter [34]. Other difficulties found in mimicking natural speech from text input are the correct pronunciation of names and foreign words, and generating correct prosody [34]. Prosody plays an important role in transferring a full communication experience between the speaker and the listener. These latter limitations can, to some extent, be addressed by using SSML because it provides authors of synthesizable content a standard way to control aspects of speech such as pronunciation, volume, pitch and rate. How the SSML is integrated in COUCH is explained in more detail in Section 4.3.

2.4.3 Speech Input Variations

Conversational interfaces have to deal with the problem of handling speech input variation. Variations may cause the speech recognizer to incorrectly interpret the speech input or not recognize the speech at all (i.e. increasing the WER). This variation may be due to several factors such as age, speaking style, accents, emotional state, tiredness, health state, environmental noise and the microphone quality [35]. A few of these factors are considered important for the current project and will be elaborated on a bit more.

Environmental Noise	This is one of the big challenges in ASR systems since it interferes with the speech recognition of the user's voice. An experimental setup in a completely controlled laboratory environment can obtain very promising results for ASR systems, while at the same time, using the same system in a home setting can substantially increase the WER. Older adults who suffer from hearing loss might, for example, listen to loud radio and television at home, which increases the risk of ASR performance deterioration. One simple method for conversational interfaces to deal with noise is by providing the user with feedback about the noisiness in the environment. Simple feedback messages like, "Sorry I cannot hear you, it is too loud", can dramatically improve the user's experience because it shows understanding of the issue [36].
Speech characteristics	Speech of older adults is very different from younger adults in multiple ways, causing the WER of ASRs to be significantly higher for older adults [37]. The first issue related to older adults has to do with a naturally ageing of the voice. The characteristics of an aged voice found to be less easily recognized by standard ASR systems since these are often designed for the majority of the population, trained with speech of young adult speakers [22,35]. Second, literature suggests that a large segment of the older population experienced a past or present voice disorder [38]. People suffering from dysarthric speech or any other voice-related health problem tended to achieve lower ASR performance with the commercial applications [28].

2.4.4 Speech Synthesis for Older Adults

According to Kuligowska, Kisielewicz, and Włodarz [34], older adults have problems with understanding the synthesized speech, particularly older adults suffering from hearing problems. When they miss the contextual clues, such as hand gestures and lip movements, that compensate for weakened acoustic stimuli, understanding the speech can be very difficult. Fortunately this limitation can easily be addressed by offering the users the opportunity to use both written text and spoken speech in the interface.

2.4.5 Expectations

Speech can raise the expectations of the system [39]. The coaches from the council are not very smart, and for this reason designed as cartoon characters, communicating via text balloons. When users can talk to a application, they might expect the system to understand everything they say, also topics which are not related to the coaching. Research showed that especially older adults often use everyday language and their own words to formulate commands, even when explicit instructions regarding the required input are given [40]. When the system does not understand this, the user might experience more negative feelings leading to avoidance of the system in the worst case scenario. In this case, the coaches are not able to maintain a long-term relationship with the users and cannot provide coaching anymore. When implementing speech, cartoon-like characters are suggested. This is because cartoon images will lower customer expectations toward the skills of the characters, and match the technical abilities of the system [41]. Thus, the problem of high expectations might be overcome by keeping the coaches as they are, like dumb cartoon characters.

2.4.6 Long-term Engagement

To keep the user engaged for a longer period of time is challenging for conversational interfaces, but also for other technologies. It is important to consider that users might get bored when the system outputs exactly the same sentences multiple times because this can lead to interactions that become repetitive over a longer period of time. This was a problem mentioned in a study by Bickmore [42], for example. Results of the preliminary literature research [?] showed that there was too little content in the original COUCH application, which caused repetitiveness and boredom among users. Engagement might be improved by using speech because it can make the application more interesting, but keeping this engagement in the long-term remains challenging.

2.4.7 Privacy Issues

Although all technologies need to consider privacy issues, it is particularly important for conversational interfaces. Ethical and legal questions arise about what data is collected, who has access to it, how long the data is stored and where and what such data is used for [43]. The COUCH system is used in a safe home area where private conversations regarding physical, but also mental health, take place. Speech has to be recorded to participate in interactions, which may contain sensitive information.

Chapter 3

Related Work

This related work chapter is the last chapter that reused content from the preliminary literature research [1]. In this chapter, an extra section is added that describes the implications for the current research, based on findings from related work.

As described in Section 2.1, ECAs are computer-generated characters with an embodiment, used to achieve humanlike and more engaging interactions. Several ECAs have already been developed in virtual coaching systems to assist users making appropriate health-related decisions. The purpose of this chapter is to give a few relevant examples of these state-of-the-art virtual coaching systems and the advantages and disadvantages that they provide.

3.1 In-home Social Support Agent

The in-home social support agent [44] is a remote-controlled companion agent used in homes of the older adults. The remote Wizard of Oz research (see Figure 3.1) showed high levels of acceptance and satisfaction with the in-home social support agent, with many participants stating that it felt as a social companion. Older adults would like to tell stories and discuss the weather, their family and future plans with virtual companions. Participants spent most time on storytelling, indicating that this would be valued and utilized by older adults.

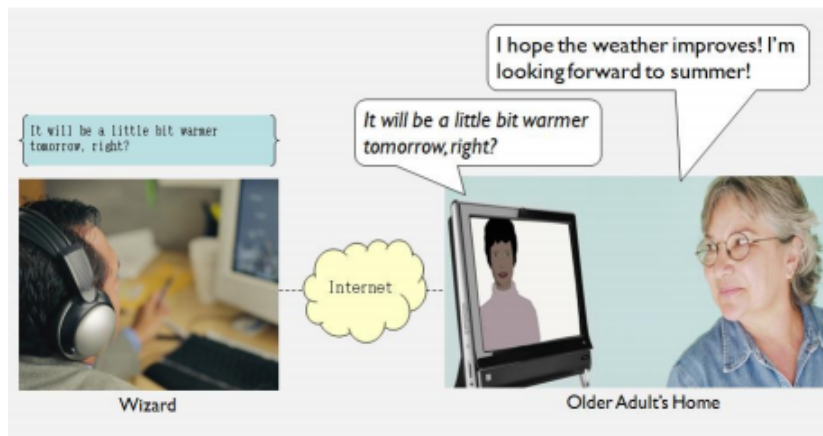


Figure 3.1: Wizard of Oz agent setup for in-home social support agent for older adults. Figure reproduced from [44].

3.2 Kristina

KRISTINA [45] is a conversational agent who provides healthcare advice, assistance and social companionship to older adults. The agent is composed of different modules that ensure multi-modal dynamic conversations with users. The system includes an ASR component for the speech recognition and a TTS component for the spoken surface output. For its non-verbal appearance, KRISTINA is realized as an ECA through a credible virtual character and offers different functionalities. First, the user can choose a scenario from a predefined list. Based on this scenario, users can converse with the virtual character by speech, although a back-up option for typed text is provided in case the conversational agent does not understand the spoken speech. The project underwent two iterations, with added features like a larger array of topics to talk about and more depth within the topics in the second iteration. Examples of the characters from the two iterations are shown in Figure 3.2.

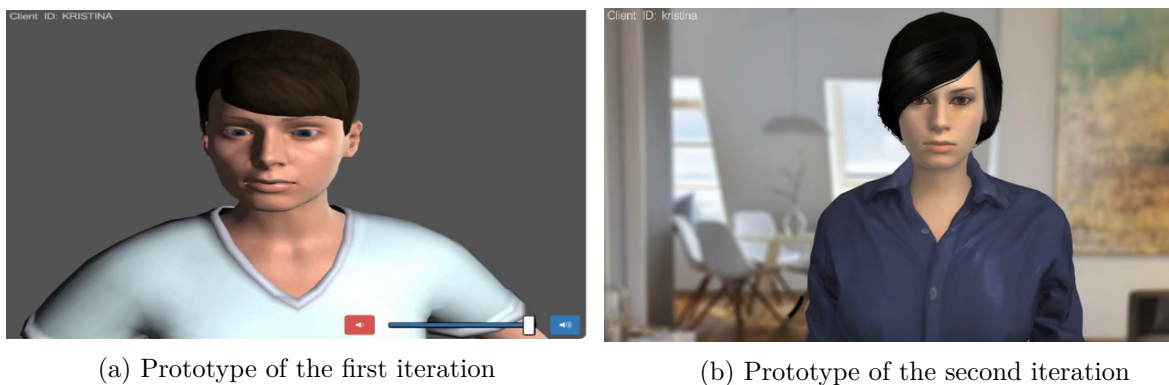


Figure 3.2: KRISTINA prototypes [45], captured from <http://kristina-project.eu/en/>

The final prototype provides a wide range of information and more advanced communication [46]. The system is able to generate proactive responses and dialogues to have more everyday-like conversations. The agent can ask for direct or indirect clarification when it detects more than one relevant topic to the user's input. KRISTINA scored high on points as trustworthiness, friendliness and professionalism. The design and behavior has tried to look natural to match the content and scenarios [46]. However, the gestures and facial expressions were considered as being too rigid and for this reason did not evoke empathy. The voice also did not evoke empathy: it is considered to be monotonous. One last major issue is the system latency, which is perceived as too long for a natural dialogue. These problem needs to be tackled in future systems.

3.3 Meditation Coach

The Meditation Coach [47, 48] is an ECA developed to guide users through a mindfulness meditation session and help them relax. The coach (shown in Figure 3.3) is made interactive by recording and processing data from a breathing sensor in the dialogue system. Participants appreciated that the system afforded tailored feedback and they experienced it as more effective in reducing anxiety than a videotaped meditation instructor. This finding was supported by the significantly stronger respiration regulation as measured by their respiration rate during meditation. The virtual coach was inhaling and exhaling in the rhythm of the participant's breath and it provided feedback about the pace of the breathing (e.g. "continue breathing at a slower pace"). These personal breathing instructions were based on the participant's measured breath duration and breathing rate. The results indicated that implementing a coach embodied

as conversational agent can achieve the coaching goal more effectively than a non-embodied conversational character. Nevertheless, participants were significantly more satisfied watching a video of a human meditation instructor. A major source of displeasure with the virtual coach was the lacking humanlike features, including its synthesized voice. A human voice is often preferred over a synthesized voice, and this is even more important for meditation applications.



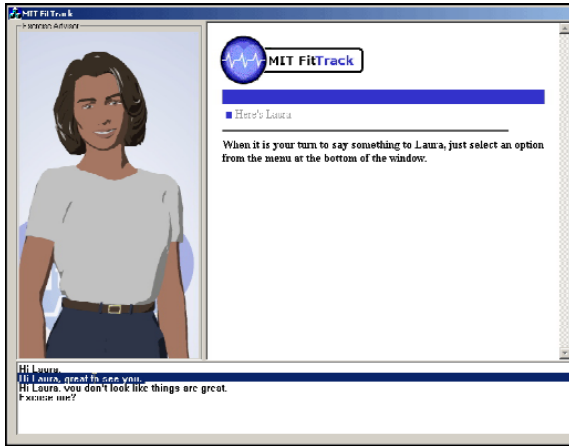
Figure 3.3: The Meditation Agent. Figure reproduced from [47].

3.4 Exercise Advisor

The exercise advisor [42, 49] is part of the FitTrack system which has been developed to investigate the ability of relational agents to establish and maintain long-term, social-emotional relationships with users, and to determine if these relationships could be used to increase the efficacy of health behavior change programs. The relational agent plays the role of exercise advisor that participants talk to about their physical activity. It is designed to be used on home computers on a daily basis. The agent uses synthesized speech and synchronized nonverbal behavior, but the user contributions to the dialogue rely on selecting items from multiple choice menus, dynamically updated based on the conversational context.

Two different versions of the exercise agent exist. A first system designed to work as exercise coach in general (see Figure 3.4a) and a second system that was adapted for older adults (see Figure 3.4b). This second system was designed to be easy to use, with a very consistent and intuitive user interface and an enlarged display area to accommodate visual impairments and it also contained an additional self-monitoring graph and educational content page, temporarily replacing the ECA.

Both evaluation studies of the exercise advisor system demonstrated the acceptance and usability of a relational agent by older adults. Participants found interacting with the agent to be relatively natural and this had a positive impact on the users' perceived relationship with the agent [42, 49]. However, some contradictory results were found. The former research [49] suggested that deploying conversational interfaces does not imply that natural language understanding must be used. The dynamic menu-based approach used in the FitTrack system provided many of the benefits of a natural language interface, such as naturalness and ease of use. As additional advantage, it is not necessary for the system to rely on error-prone understanding of unconstrained speech input. However, in the follow-up research [42], participants mentioned that they could not express themselves completely using the constrained, multiple-choice interaction. When asked, participants universally said they would have preferred speaking to the agent, rather than using a touch screen, but Bickmore et al. [42] mentioned that for future work, available systems first need to be thoroughly evaluated to ensure that they could provide high enough reliability given the variability and differences in voice quality in older adults. This



(a) Initial FitTrack interface with exercise advisor. Figure reproduced from [49].



(b) New FitTrack interface with exercise advisor. Figure reproduced from [42].

Figure 3.4: The FitTrack interfaces

is necessary so that users can engage in richer conversations and more freely express themselves while maintaining the ease of use of the multiple-choice selection input modality. Our research focuses on this future work proposal and investigates the possibilities of speech, while maintaining the ease of use of the multiple-choice approach.

3.5 Implications for the Current Research

This related work section provided insights in existing virtual characters created in the field of e-health, whereof a few findings are considered important for the current research. The in-home social support agent showed high levels of acceptance and satisfaction, indicating the relevance of such systems in general. KRISTINA offered users the backup option to type text in case the agent did not understand the spoken speech. Something similar can be considered for our project, in which users can click the textual option in case there is no response to the speech. Additionally, KRISTINA could ask for direct or indirect clarification when it detected more than one relevant topic. Such confirmation strategies are considered important for the current research because it can improve the naturalness of the conversation. A limitation found in both KRISTINA and the mediation coach was the synthesized voice: it was perceived as monotonous and lacking humanlike features. This shows the importance of agent voices that are somewhat humanlike for assuring a pleasant experience.

The exercise advisor by Bickmore et al. is quite similar to the COUCH application in context and dialogue structure and they share the goal to establish and maintain long-term relationships. The dialogue structures are similar in the selection of items from multiple choice menus, dynamically updated based on the conversational context. The applications differ in their input and output modalities that can be text, speech, or a combination of both. The original COUCH application only uses text input and output, the exercise advisor uses synthesized speech output, but no speech input, and the speech-based COUCH application uses synthesized speech input and output. Participants found interacting with the exercise advisor to be relatively natural, and this is expected to improve when implementing a two-directional (i.e. speech input and output) vocal interaction.

Chapter 4

System design

In this chapter, the design- and development process of the speech-based COUCH system is described. The original text-based COUCH application was not designed to support speech input and output, leaving the challenge of implementing it within the existing COUCH framework. To do this, the current Council of Coaches Platform is researched. The remainder of this chapter focuses on the implementation of the conversational interface components (as described in Section 2.3) into the application. First, an overview of the system architecture and communication is provided, followed by an explanation of each conversational component and its integration in the application. Furthermore, the adjustments in dialogue content and structure are described, as well as the features that were added or removed. This section partly answers sub-questions 2, 3 and 4 (see Sections 4.2.5, 4.5, and 4.2.2 respectively).

4.1 The Council of Coaches Platform

The Council of Coaches platform¹ consists of a few main components that are discussed in this section. First, the WOOL dialogue platform is discussed, which is used for handling and adjusting the dialogues according to our need. Second, an overview of the coaches is provided since these are the main characters of the application, providing all relevant information towards the user. Lastly, we explain the original interface and the dialogue structure and coaching content. This information serves as background for the reader and is relevant to understand the design- and development choices of the speech-based system.

4.1.1 The WOOL Dialogue Platform

As already mentioned in Section 2.3.3, different types of dialogue managers exist. The WOOL dialogue platform² is one such example, developed in order to easily manage dialogues. Figure 4.1 shows the WOOL editor with its relevant elements. The following terms defined by Beinema et al. [50] are important to understand.

Node:	A dialogue step that contains one <i>Statement</i> and a one or more <i>Replies</i> .
Agent:	A virtual speaker within a dialogue.
Statement:	Something an agent says.
Reply:	A possible reply that a user of the system can give.

¹<https://www.council-of-coaches.eu/>

²<http://www.woolplatform.eu/>

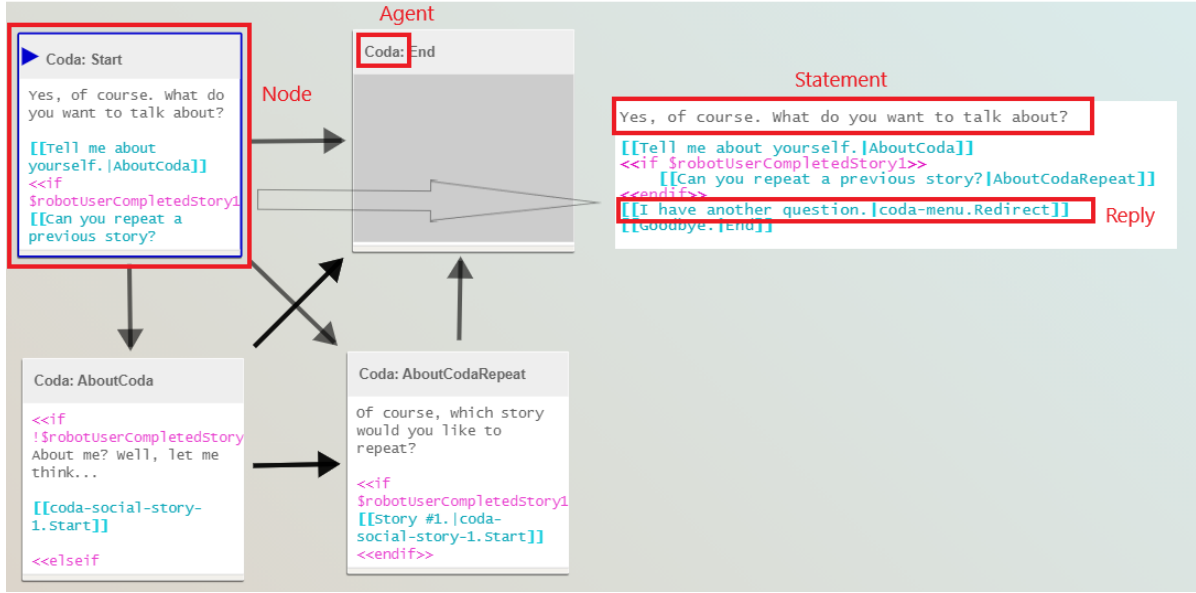


Figure 4.1: The WOOL editor, presenting a dialogue from *agent* Coda. Different *nodes* are shown, each containing a *statement* and multiple *replies*.

WOOL defined different types of replies: *basic*, *autoforward* and *input* replies. The basic reply is the standard reply option that links the user to a new node. An autoforward reply is a reply that does not contain a statement. In this case the user is not provided with multiple reply options, but instead is presented e.g. with a default 'continue' button that links to another node. An input reply is simply a reply in which the user is asked to enter input (such as names, numbers, etc.) in a text field. The WOOL dialogue platform is essentially a definition of a series of dialogue steps, represented by nodes and linked through user replies. These dialogue steps can be manipulated individually, without changing the structure of the complete hierarchy. More information about the details of the WOOL framework can be found in the work of Beinema et al [50].

The WOOL dialogue platform offers a few features that are particularly useful for the current research. First, it provides the possibility to manually add words to listen for to a statement. Second, it provides the possibility to add or remove intermediate dialogue steps without changing anything to the remaining of the dialogue. How these features help for the development of the speech-based system is discussed in more detail in Section 4.2.2). The third advantage that WOOL offers is its ease of use for people who are no expert in dialogue management tools. This can be particularly helpful for future work, as explained in Chapter 7 (the discussion).

4.1.2 The Coaches

One of the objectives of COUCH was to develop the coaches as interesting characters, with each coach having its own expertise and background. The complete set of coaches is shown in Table 4.1, including their role, name, nationality, gender and age. Besides the coaches, there is the robot Coda, the equivalent to an in-app 'menu'. Coda is designed to help with more technical functions like creating an account, logging in or out and changing settings.

Because COUCH contains many different coaches differing in gender and age, an extensive diversity of synthesized voices was required. For some roles that were obvious in their nationality (e.g. François who uses a lot of French words and talks about French wine and cheese), the intention was to show their nationality via their voices (e.g. French pronunciation of words like 'bonjour'). This design choice is elaborately explained in Section 4.2.5 and 4.3. Additionally,

the ages of the coaches were taken into account for the distribution of voices, by providing older sounding voices to Helen and Carlos and younger sounding voices to Emma and Francois. The chronic pain and diabetes coach were not considered for this research because of their very specific role, and therefore did not need a voice. For the interested reader, more information about the coaches' backstories can be found in the public deliverable D3.4 of the COUCH project [50], where an elaborate description of each coach is given, including their height, weight, place of birth, likes and dislikes, backstory, role, pointers for dialogue writing and coach selection blurb.

Role	Name	Nationality	Gender	Age
Physical Activity Coach	Olivia Simons	Dutch	Female	52
Nutrition Coach	François Dubois	French	Male	45
Social Coach	Emma Li	American	Female	28
Cognitive Coach	Helen Jones	British	Female	64
Peer Support	Carlos Silva	Portuguese	Male	67
Chronic Pain Coach	Rasmus Johansen	Danish	Male	33
Diabetes Coach	Katarzyna Kowalska	Polish	Female	45

Table 4.1: The seven coaches from the Council

4.1.3 The Council of Coaches Interface

Users can interact with the interface by clicking buttons with scripted content and response options. Users can determine themselves which coaches to interact with. When a coach is clicked, a text balloon with a statement appears. The first personal dialogue with each coach consists of a short introduction and a personal story or domain relation question (e.g. Francois asks the user whether he/she likes to cook). Starting from the second dialogue, users can choose to have a social conversation, do a coaching session or leave the conversation (see Figure 4.2 on the next page), which are examples of *basic* replies. An example of an *autoforward* reply in COUCH is shown in Figure 4.3 (on the next page), where the user is presented with a 'continue' button. An *input* reply example in COUCH is presented in Figure 4.4 (on the next page). It depends on the user input which route through the dialogue is taken. This setup limits user input, which has the strength of giving the coaches more clarity on what they respond to [51]. The coaches themselves can naturally keep an interaction going by supporting or contradicting each other to increase user engagement, active participation, reflection and critical thinking about their own health. Besides the interaction with coaches, users can setup the system and change settings by talking to the robot assistant.

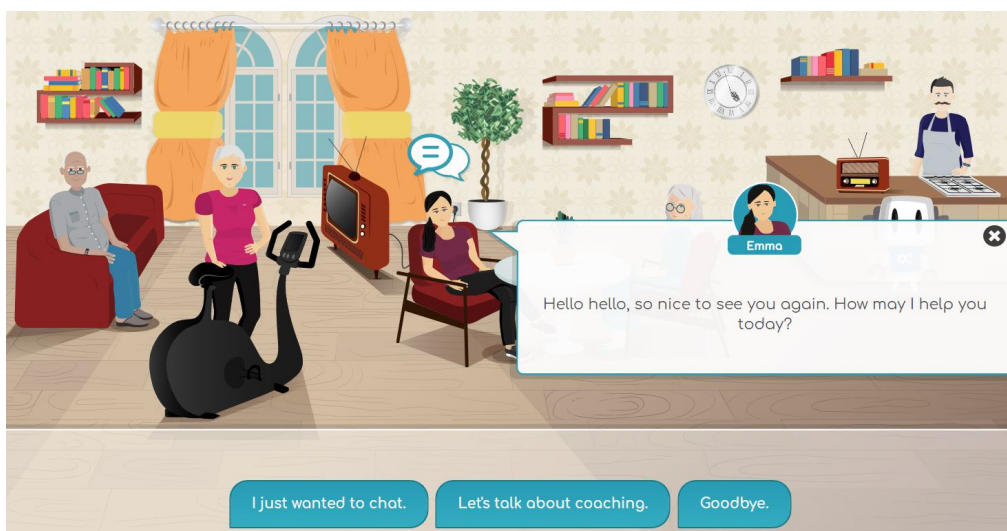


Figure 4.2: The Council of Coaches scripted basic reply¹.

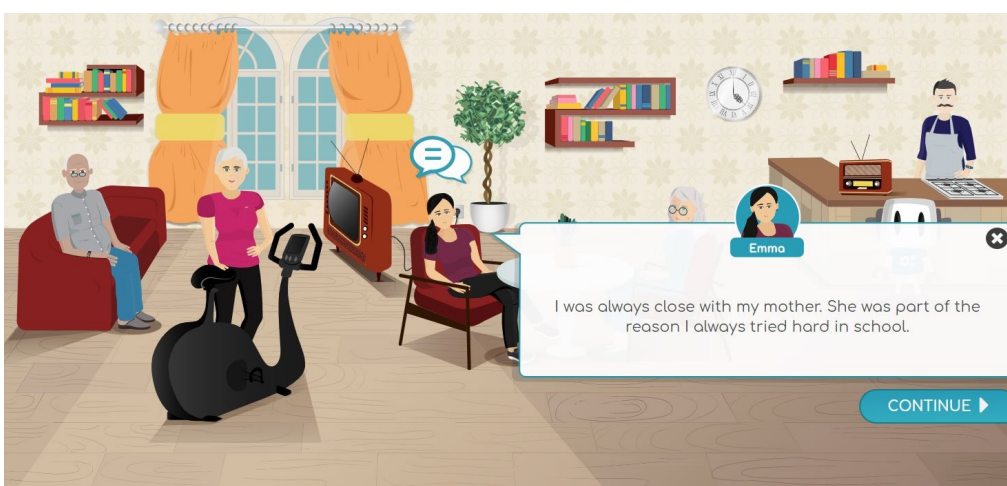


Figure 4.3: The Council of Coaches scripted autofoward reply¹.

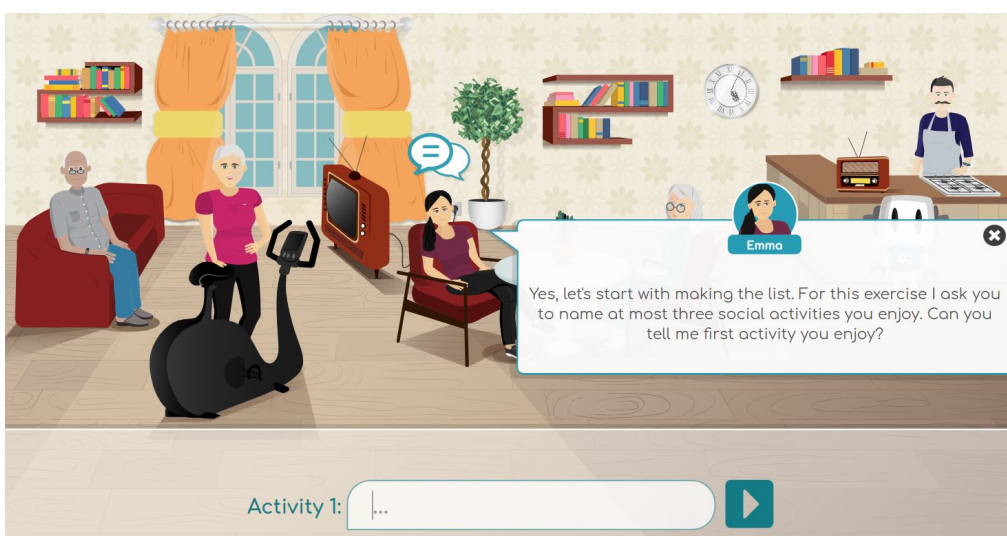


Figure 4.4: The Council of Coaches scripted input reply¹.

When users open the application, they are presented with the main menu Web interface, which has a welcoming and familiar visual of a building entrance (see Figure 4.5). It is an entrance to the “Council of Coaches” home. The functions available in this screen are: account creation, log in with an account that was created before and selection of preferred language. If not done before, users have to create an account, so their preferences and information can be stored, and dialogues can be personalized. This action, as well as the login action, allows the user to enter the Council of Coaches house. After pressing the button ‘create account’, Coda will guide users through this process (see Figure 4.6) and also introduces them to the COUCH system by teaching them how to interact with it.

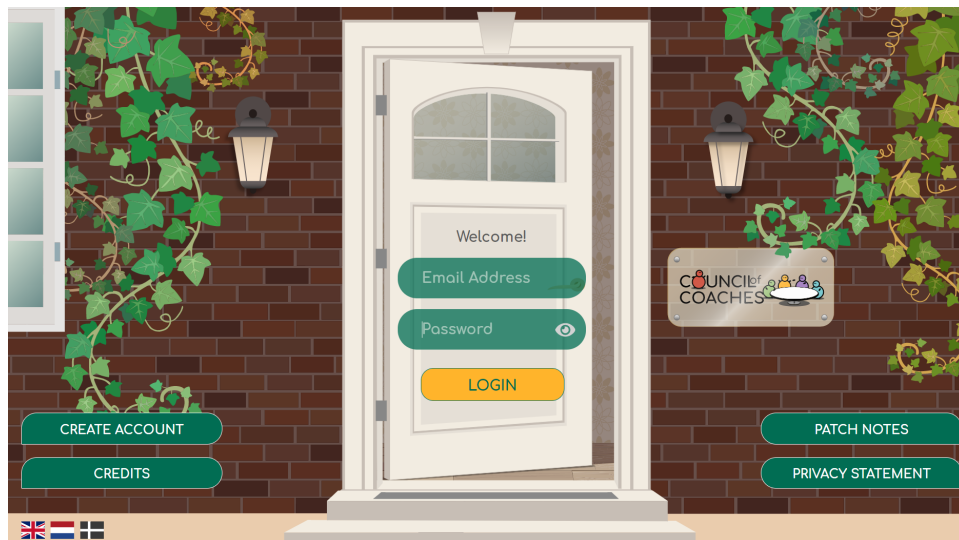


Figure 4.5: Council of Coaches Main Menu screen¹.

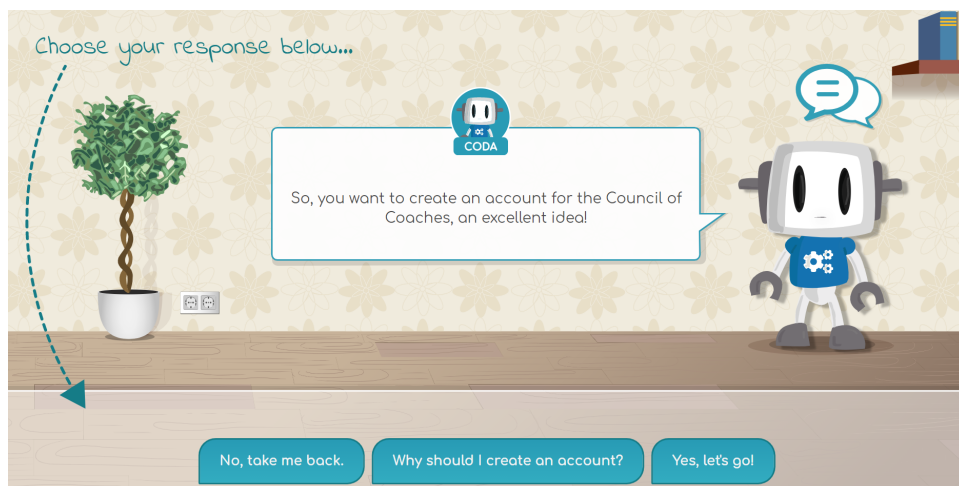


Figure 4.6: First screen of the account creation process, guided by Coda¹.

4.1.4 The Dialogue Structure and Coaching Content

The previous section showed dialogue examples of the original COUCH application and gave a feeling of such interaction. To give an idea about the high level structure of the conversation, this section describes the different content types and hierarchy of the dialogues. This information serves as background for the reader and explains the context of the current research.

To guide the user in their behavior change process, each coach has a number of dialogues available that can be used to discuss a specific topic. COUCH structured the topics that each coach can discuss with a hierarchy. The end-points of this hierarchy represent the topics for the dialogues. For the current research, we will describe one example structure (the physical activity coach) of such a coaching process (for more examples see the public deliverable D3.4 of the COUCH project [50]). The strategies for the other coaches are relatively similar. The topics structure starts with a 'Start' node that allows a choice between the 'Social' topics and the 'Coaching' topic (see Figure 4.7). The hierarchy is straightforward from the figure, with all grey nodes allowing a choice between topics and the blue nodes being final dialogue topics. A small description of each of the blue nodes is provided in Table 4.2.

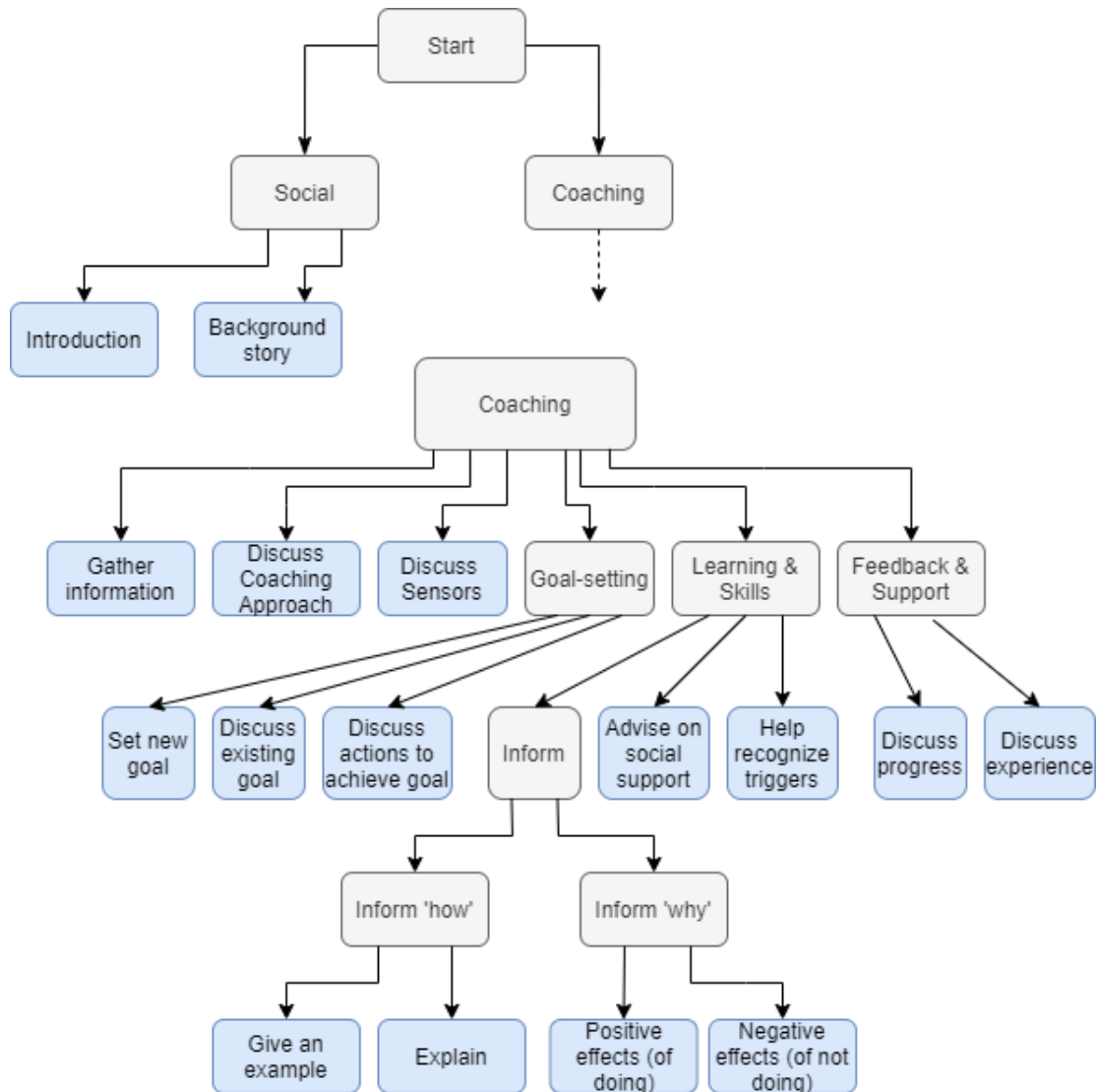


Figure 4.7: The hierarchy of coaching topics. Adapted from [50]. The blue nodes are topics and the grey nodes have subtopics themselves.

Topic	Description
Introduction	The dialogues in which the coach and the user introduce themselves.
Background story	The dialogues about the coach’s background story.
Gather information	The dialogues in which the coach asks the user about his preferences.
Discuss coaching approach	The dialogues in which the coach discusses with the user what coaching style he prefers.
Discuss sensors	Dialogues about the sensors (FitBit).
Set new goal	Dialogues about setting new goal long-term and daily goals.
Discuss existing goal	Dialogues about the users’ current goal and his experienced level of difficulty of the goal.
Discuss actions to achieve goal	Dialogues about what activities the user will do and when.
Advise on social support	Dialogues discussing how users might ask their family and friends to perform activities with them.
Help recognize triggers	Dialogues in which the user reflects on triggers, causing them to be less active.
Discuss progress	Dialogues about how the user feels about its process.
Discuss experience	Dialogues in which the user reflects on how it is going.
Give an example	Dialogues providing examples of how the user can be physically active.
Explain	Dialogues explaining the user how to be physically active.
Positive effects	Dialogues explaining users why they should be physically active.
Negatives effects	Dialogues explaining users why being physically inactive can be bad for their health.

Table 4.2: Descriptions of the coaching topics (the blue nodes in Figure 5.6).

4.2 System Architecture

To implement speech in the original text-based application, it is important to understand its structure and functioning. This applications’ WebClient is connected to the R2D2 server which has a submodule responsible for handling WOOL dialogues. The WebClient receives information about the dialogues from the R2D2 server and presents the statements with reply options to the user. Consequently, the chosen reply option is send back from the WebClient to the R2D2 server.

The speech-based system also uses a connection with the R2D2 server for handling and retrieving WOOL dialogues. Furthermore, our system broadly follows the conversational interface architecture, as described in Section 2.3, consisting of five main components: Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU), Dialogue Management (DM), Response Generation (RG) and Text-to-Speech Synthesis (TTS). While the ASR-, DM- and TTS component are connecting to external servers and clearly distinguishable in their function, the SLU- and RG component are less explicit components implemented in the WebClient. The WebClient is responsible for locally handling the keyword check (the SLU component) and constructing the content-to-speak for the TTS (the RG component). In this report, the term

'Keyword' describes the word to listen for which is linked to one reply.

The complete interaction process can be briefly described as follows. When the user speaks to the interface (i.e. the COUCH WebClient), the ASR component receives audio from the microphone, and transcribes this input to written text. The transcription is sent to the SLU component, while simultaneously information about the state of the current dialogue step is forwarded by the DM component. The SLU component compares the keywords received from the dialogue manager with the audio's transcription and returns the user reply for which the transcription matched the keyword. The dialogue manager consequently evokes the next dialogue step and sends the updated dialogue content to the RG component. The RG component uses the relevant content (i.e. the statement and voice of the coach) from this dialogue step and sends it in the correct request format to the Google Text-to-Speak API, which translates the statement in audio speech. The generated speech is sent to the interface, where the audio is played. The user receives the content via audio (from the TTS) and text balloons (from the DM component). Figure 4.8 visually presents this process and shows how the different components interact. All system components are more elaborately discussed in upcoming sections.

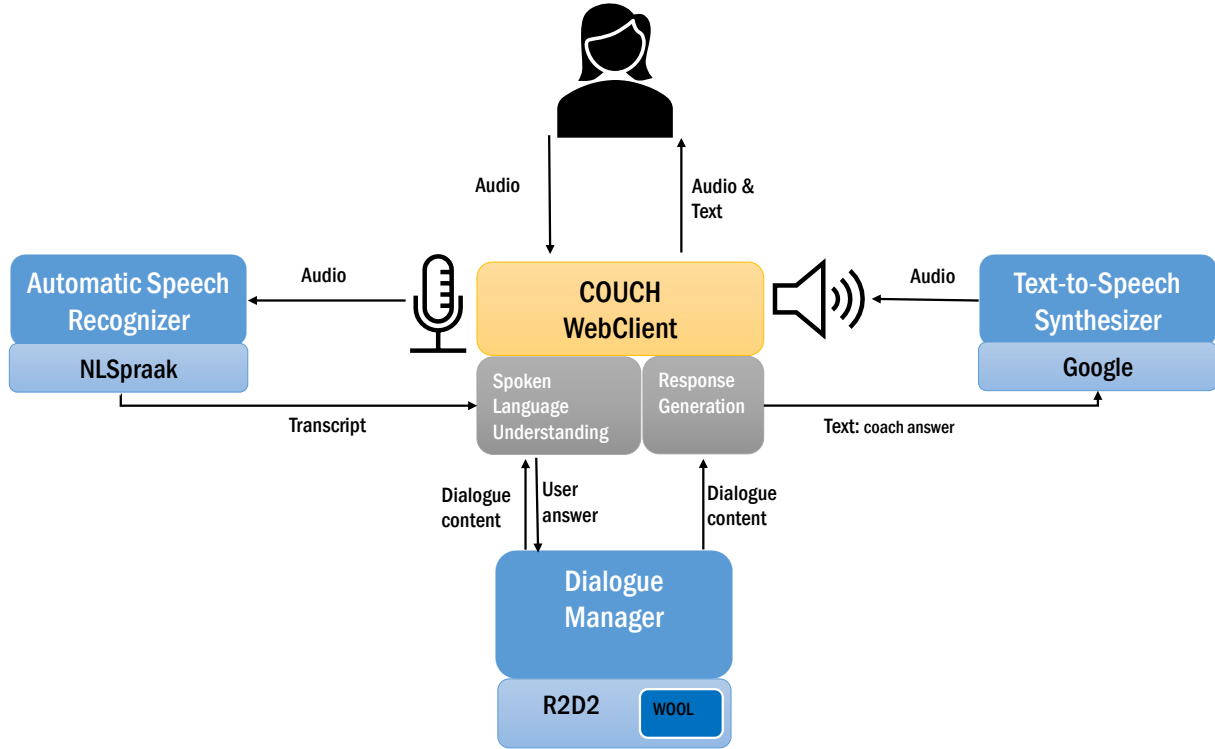


Figure 4.8: Visualization of the speech-based COUCH architecture. The system contains ASR, DM and TTS components that are connected with external servers, and SLU and RG components that are implemented in the WebClient and therefore colored differently.

The system is additionally described on dialogue step level because using the application essentially means a sequence of dialogue step repetitions. As mentioned in Section 4.1.3, three types of replies exists: *Basic*, *autoforward*, and *input* replies. Since input replies are not adjusted in the speech-based system, the focus was on basic and autoforward replies. The processes of giving these replies are presented in sequence diagrams. These diagrams show the interactions between the WebClient, R2D2 server, and NLSpraak server arranged in a time sequence. It depicts these three objects involved in the speech recognition during a dialogue step and the sequence of messages exchanged that were needed to carry out the functionality of that dialogue step. For simplicity, the TTS is not implemented in the diagrams.

Both types of replies are mixed together within the dialogues, causing the diagrams to have overlap. Therefore only the sequence diagram of a basic reply (Figure 4.9 on page 38) is presented from beginning to end. The sequence diagram of an autoforward reply (Figure 4.10 on page 39) is represented as smaller block that fits in the red square form the basic reply sequence diagram as alternative dialogue step within an interaction. The sequence diagram in Figure 4.9 can be globally described as follows. A user starts the interface by turning on the ASR. The server returns a request for permission to use the microphone. When the user gives permission, keywords are added for coach names and the ASR component starts listening. When the name of a coach is spotted, the coach starts an interaction and keywords are added for the related reply options. While the coach is speaking, the ASR is paused and not listening for user speech. When the TTS is finished synthesizing the coach's statement, the ASR is resumed and listening to the user again, searching for keywords. In case a keyword is spotted, content of the next dialogue step is retrieved and the process starts all over again. An alternative possibility is that the user speaks one of the keywords to end a conversation, so the interaction stops and the user returns to the living room. In case of an autoforward reply, the user does not speak any reply options, but instead the R2D2 server automatically

4.2.1 Automatic Speech Recognition

The 'NLSpraak' ASR system is used for transcribing the incoming speech into written text. For this speech recognizer, the Corpus Spoken Dutch³ was used to train Dutch models on the Kaldi framework [30]. Based on this Dutch Kaldi implementation, a web demonstration for this ASR was created⁴, which has been used as starting point for the current project. This demonstration was helpful because it already included a working JavaScript application. The server listened to the audio, which it received via an internet connection and returned the decoded transcription.

When the ASR system recognized a word that it never heard before, it printed this word as unknown: '<UNK>'. These unknown words could happen for two reasons: (1) the word was not in the vocabulary of the ASR system, so it could not be transcribed (e.g. the Dutch word 'coachingsessie'), or (2) a word was mumbled instead of clearly pronounced which caused the ASR system not to identify the spoken word and transcribe it as unknown word. The identification of unknown words is used for the implementation of dialogue management strategies (see Section 4.4)

³<http://lands.let.ru.nl/cgn/>

⁴<https://github.com/laurens75/SpeechAPIDemo>

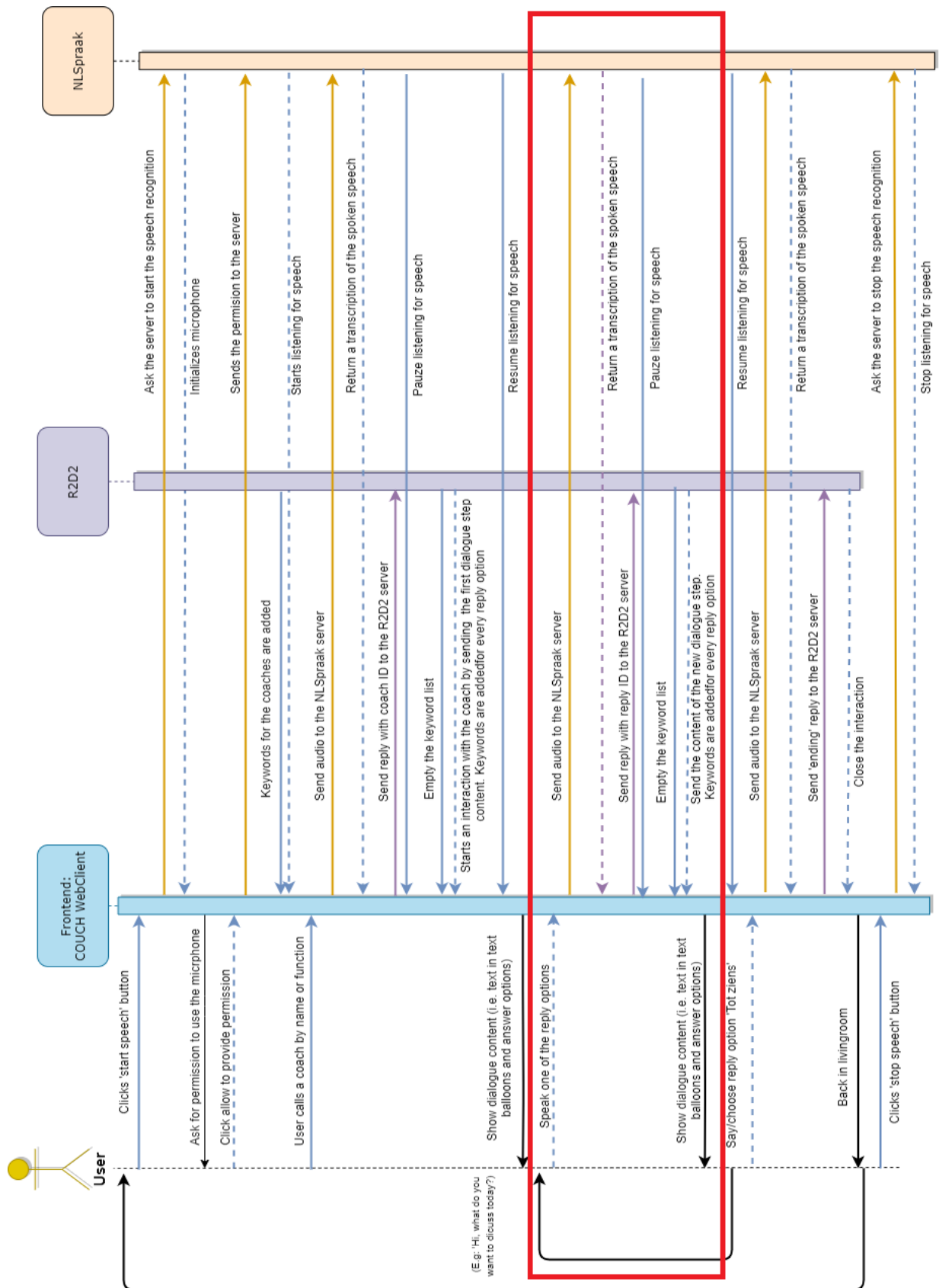


Figure 4.9: A sequence diagram presenting the process of answering with a basic reply in a dialogue step. The dashed lines represent return messages.

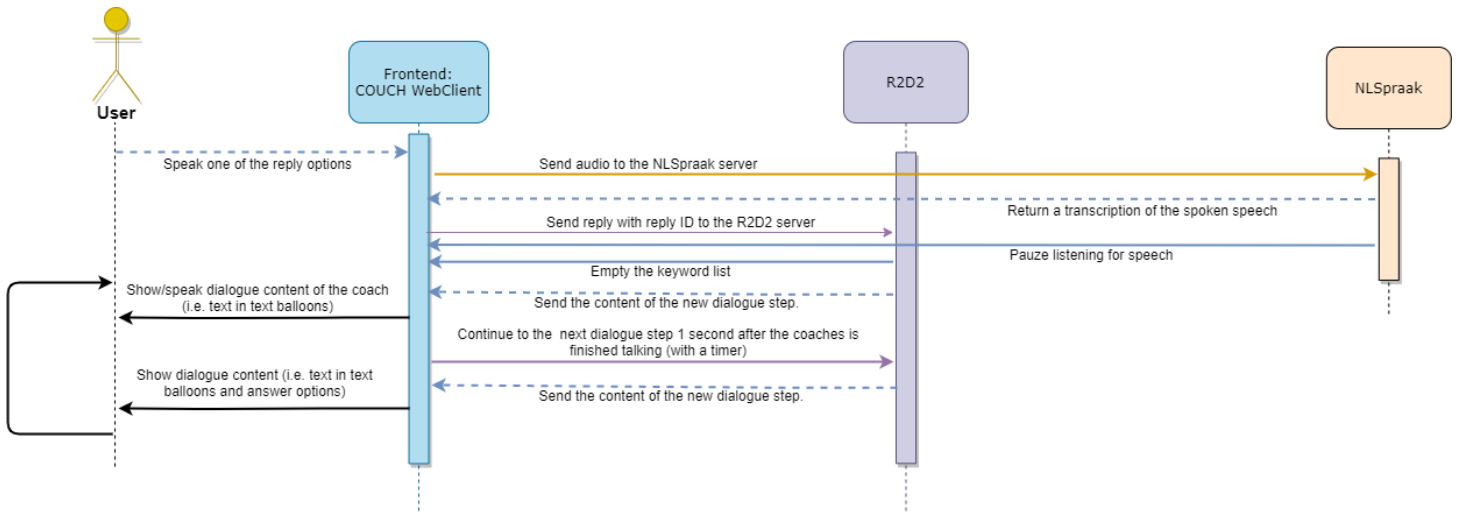


Figure 4.10: A sequence diagram presenting the process of an autofoward reply. Only part of the process is presented, that can replace the content of red square in Figure 4.9.

4.2.2 Dialogue Management

The DM component is responsible for responding to user behavior perceived via the ASR input and for generating the system behavior that is realized via TTS output. The dialogue manager controls the SLU and RG components, which are discussed in the successive sections.

The WOOL editor is used adjust the dialogues and the WOOL subcomponent of the R2D2 server was responsible for handling these adjusted dialogues. As mentioned in Section 4.1.1, WOOL offers some useful features that are used in the current research. For our project, the structure and content has not really been changed, but we used WOOL’s possibility to remove intermediate dialogue steps that were not relevant for the user studies (e.g. dialogues from the activity coach requiring a FitBit) without changing anything to the remaining of dialogues. Also, words that could not be pronounced correctly by the TTS were removed from the dialogues.

Furthermore, in WOOL it is possible to add an action-Statement to a reply. This means that each reply option can receive a specific action that is performed when that reply option is chosen. An example implemented in the original COUCH application is the response of the system to open the recipe book in case a user replies with its preference to check out recipes (see Figure 4.11).

`[[Show me the recipe options.|End|<<action type="generic" value="OPEN_RECIPE_BOOK">>]]`

Figure 4.11: An example of the addition of a generic action-Statement to a reply option

For the implementation of the speech-based system, this WOOL feature is used to provide keyword tags to different reply options. Keyword tags are created for each reply, as shown in Figure 4.12. This keyword list could be retrieved by the dialogue manager, specifying the action type and value in the request. The dialogue manager passed on the keywords to the SLU component that checked for the existence of one of these keywords in the speech transcription.

```

What can François do for you?
[[I just wanted to chat.|francois-social-menu.Start| <<action type="generic" value="ADD_SPECIFIC_KEYWORDS"
keywordlist = "praten, kletsen, zeggen">>]]
[[Let's talk about food.|francois-coaching-menu.Start| <<action type="generic"
value="ADD_SPECIFIC_KEYWORDS" keywordlist = "eten, voeding, voedsel">>]]
[[Goodbye.|End]]
<<endif>>

```

Figure 4.12: An example of keyword tags, added to different answer options.

Most replies were manually provided with keyword tags, chosen by the main message and important words of the reply. These main words were added as keywords, together with a set of similar words in order to increase the change for the ASR to detect a keyword and continue in the dialogue. For example, in the sentence 'I just want to talk', the keyword *talk* is added, as well as its synonyms *chat*, *speak* and *conversate* (see Figure 4.12). Due to the hierarchical content structure and fixed coaching topics, there was a lot of repetition in the dialogues. This offered the possibility to copy keyword tags to other dialogues. One common strategy in the keyword tagging was to provide a negative set of keywords to negatively phrased answer options and a positive set of keywords to positively phrased answer options. Figure 4.13 shows an example. For the reply that did not contain a keyword tag, the dialogue manager retrieved the last word from the normalized reply (i.e. the reply without stop words and punctuation marks) and added it as keyword. In the first place, the first word of a sentence as keyword, but the ASR showed to have more difficulties recognizing the first word then the last word. The retrieval of other words in a sentence was also considered, but this required a more complex implementation and because most keywords were tagged manually, it was not considered for the current implementation.

```

Today I want to talk about the importance of a social network. Are you interested in my story?
[[Okay, I'm curious.|Node_social_week1_1| <<action type="generic" value="ADD_SPECIFIC_KEYWORDS" keywordlist = "oké, prima,
goed, ja, best, heerlijk, zeker">>]]
[[No, sorry I am not.|Node_social_week1End| <<action type="generic" value="ADD_SPECIFIC_KEYWORDS" keywordlist = "nee, niet,
geen, sorry, helaas">>]]

```

Figure 4.13: An example of keyword tags for positively and negatively phrased answers

A difficulty experienced while tagging keywords was in sentences that were very similar but did or did not contain the word 'not'. For example: *'I am actually a big football fan'*, or *'I am not really a fan of football'* were difficult sentences to deal with. In those cases, the positive keywords (e.g. yes, big, always) were added for the first answer and the negative keywords (e.g. no, not, never) for the second answer. Contrary, it happened that sentences were very similar, but all answers resulted in the same next dialogue step. For example, the answers *'Oh, you play piano? Do tell!'*, *'I always love to listen to piano players'*, and *'How long have you been playing it for?'* all directed to the same follow-up step without setting parameters about the user's reply choice. In this case the main keyword 'piano' was only added for the first reply.

Besides the dynamic 'last word' and 'manual' keyword tags, we hard coded a set of keywords for two situations. The first hard coded set of keywords included all coach names and few alternative options such as the function of the coach, terms related to the function, and the number of the position of the coach in the living room (from left to right). For example, the keywords for Olivia included: *'Olivia'*, *'activity'*, *'sport'*, *'exercise'*, and *'two'*. These alternatives did not require users to remember all coach names and the provided an option in case the ASR did not recognize a name. The name 'Helen' was very difficult for the ASR to transcribe, making it difficult to start a conversation with the cognition coach. When observing the transcriptions that the ASR created when 'Helen' was spoken, words like 'heelen', 'alan', 'hellen', and 'ellen' came forward. To improve the recognition, these keywords have been added manually as keyword for the cognition coach. The second hard coded set of keywords included words to end

a conversation, such as: 'Stop', 'finish', 'end', and 'wrap up'. 'Bye' was a reply option available in many dialogue situations, but not in all, so it was added to this set of 'ending' keywords to give users the opportunity to always end a conversation by saying goodbye to the coaches. The only situation in which users were not allowed to end a conversation was in the middle of an autoforward reply.

4.2.3 Spoken Language Understanding

The SLU component used the transcription of the ASR and the keyword lists received from the dialogue manager to check if any keyword matched the content of the transcription. In case a keyword was spotted, the SLU component triggered an action that was dependent on that keyword. All actions are passed on to the DM component, so it could retrieve the successive dialogue step and all its specifications. A few actions were possible:

1. The transcription included a coach name or function, so the SLU component started an interaction with the coach (only when the user was not in a dialogue step yet).
2. The keyword belonged to one of the reply options, so the SLU component programmatically 'clicked' that reply option to continue to the next dialogue step.
3. The transcription included a word to end the conversation, so the SLU programmatically 'clicked' the cancel button and the user returned to the living room.

In case no keyword was spotted, the SLU checked if there was an unknown word transcribed. If this was the case, a clarification question was used (which is one of the dialogue management strategies described in Section 4.4). If no keyword and no unknown word were spotted, nothing happened and the SLU component waited for a new transcription from the ASR.

4.2.4 Response Generation

While the SLU component is responsible for the process between the incoming audio and the continuation to the next dialogue step, the RG component is responsible for the process between this next dialogue step and the outgoing audio. The RG component receives the updated specifications about the dialogue step from the dialogue manager. The job of the RG component is to retrieve the speaking coach and statement find the correct voice and text-to-speak. This information is formatted in an audio request the TTS could process (which is discussed in more detail in the next section). The resulting synthesized speech is send back to the RG component for playback. The WebClient presents the dialogue content in audio (received from the TTS) and text (received from the dialogue manager) to the user.

4.2.5 Text-to-Speech Synthesis

Task of the TTS

To convert the string received from the RG component into audio data, the request had to be send in the right format, specified in Google's protocol⁵. The input text, voice specifications (i.e. name and language code) and audio configurations were provided to the Google TTS, so the speech audio could be retrieved and sent back to the RG component for playback.

Choice for TTS

For synthesizing the coach statements to text, one option was to prerecord all context with different voices to create humanlike voices. Considering the limitations of KRISTINA [45] and the meditation ECA [47], this might improve the overall experience. However, for the current

⁵<https://cloud.google.com/text-to-speech/docs/create-audio>

research that includes a complete council of coaches, this approach was too time-consuming. Besides that TTS has improved in quality over the past years and is expected to sound sufficient.

Although a decision for an ASR system was already made during the preliminary literature research of this project, no TTS was chosen yet. This choice was quite easy due to the limited number of options. Since seven of the COUCH coaches were used for the user studies, we had to find a TTS that offered this many different voices. Other important criteria included the voices to be Dutch, the TTS to be free and when possible, to support the Synthesis Markup Language (SSML). All these criteria resulted in the choice for the Google Cloud Text-to-Speak⁶. Current commercially available voices from Google are powered by WaveNet⁷, software created by Google’s subsidiary DeepMind. The Google TTS differs from other voice synthesizers (including Apple’s Siri) by using machine learning to generate speech, while other synthesizers use concatenative synthesis. Concatenative synthesis is a technique for synthesizing sounds by concatenating individual syllables (i.e. sounds such as ‘ba’, ‘sht’, and ‘oo’) to form words and sentences. DeepMind claims that *‘WaveNets are able to generate speech which mimics any human voice and which sounds more natural than the best existing Text-to-Speech systems, reducing the gap with human performance by over 50%’* [52].

The TTS contains five Basic and five of these improved Wavenet voices (3 female, 2 male). Coaches Carlos, Olivia, Emma, Helen, Francois and Coda were included in the user studies, so six different voices were necessary. The five ‘human’ coaches received a Wavenet voice and the robot received a basic voice, because there was no need for the robot to sound more humanlike. The division was as follows:

- **Coda:** Standard-B
- **Carlos:** Wavenet-C
- **Olivia:** Wavenet-A
- **Emma:** Wavenet-D
- **Helen:** Wavenet-E
- **François:** Wavenet-B

The Google TTS provided the possibility to adjust the voices by pitch and speed, and to send SSML in the TTS request to allow for more customization in the audio response.

4.3 Speech Synthesis Markup Language

This section describes the customization of the audio responses via SSML. The essential role of the markup language is to provide authors of synthesizable content a standard way to control aspects of speech such as pronunciation, volume, pitch, rate. For this research the voices have not been adjusted by pitch or speed, because the voices were calm and clearly understandable for older adults. However, the pronunciation could be improved because some words turned out to be difficult to pronounce, such as: names, foreign words and slang. Additionally, the accentuation of the synthesized speech was somewhat strange in certain dialogue parts. Therefore, details on pauses, audio formatting for slang and the pronunciation of names and foreign words haven been implemented with SSML.

Some synthesizers (e.g. Amazon’s Alexa) include the option to pronounce parts of speech in a different language. This could be helpful in cases where François was talking about ‘moi’,

⁶<https://cloud.google.com/text-to-speech>

⁷<https://deepmind.com/blog/article/wavenet-generative-model-raw-audio>

or Carlos was greeting with 'Olá'. Unfortunately this 'lang' attribute was not available for the Google TTS. Instead, there was tried to give François a French accent by using one of the French voices for all dialogue content, but this voice was not able to pronounce Dutch sentences correctly so this idea was discarded. The 'phoneme' attribute makes it possible to pronounce words per small unit of the word and could be useful for words like 'wow' and 'huh', but was no option in the Google TTS either. Therefore, the more general '<sub>' attribute has been used instead. By replacing the text in the alias attribute value with the text for pronunciation, it was possible to provide a simplified pronunciation of a difficult-to-read word. This specification of word pronunciation worked for some words, such as; 'wow', 'Helen', 'Olá', 'Emilie', but not for words like 'hmm' and 'huh', which was pronounced by spelling all the letters separately. When the mispronunciation of a word could not be solved via SSML, it was removed from the dialogue.

4.4 Dialogue Design Strategies

A general strategy to manage dialogues was implemented in the autofoward replies. Autofoward replies contain parts of dialogues that stretches over multiple dialogue steps, without requiring a reply. For written text it is logical to include a continue button that a user can click to continue in this dialogue, because the moment to continue is dependent on the reading speed of the user. Contrary, for spoken dialogues it is more logical to automatically continue in the dialogue because it should happen after the coach's statement is spoken. Therefore, the 'continue' keywords and buttons were removed and the autofoward replies were automatically continued after one second. Only for purposes regarding the experimental setup of one experiment, the continue button was left in.

Furthermore, two design issues were considered for this research. As described in Section 2.3.3, two frequent design issues within DM for conversational agents are *interaction* and *confirmation* strategies [23]. Interaction strategies determine who takes the initiative in the dialogue – the system or the user? - and confirmation strategies deal with uncertainties in spoken speech understanding. Designing appropriate interaction strategies were not very relevant for the current project, due to the pre-defined structure. COUCH takes a mixed-initiative strategy, wherein the user can take the initiative first by choosing one of the coaches to talk to. Consequently the coach starts the conversation, wherein the coach can take initiative by proposing health-related information. The user is not free to respond with whatever he wants, but instead chooses the provided reply options to guide the dialogue. All possible dialogues implemented in the system were fixed in the sense that every dialogue node was connected to other dialogue nodes. Navigation happened by choosing reply options, that led the user to a next statement (see Section 4.1.4). In this way, the turn-taking was already implemented and this mixed-initiative strategy has not been changed when making the system voice-controllable.

Contrary, designing appropriate confirmation strategies were relevant for this research. No confirmation strategies were implemented in the original COUCH application because uncertainties only arise in spoken speech understanding and not in written speech understanding. Both explicit and implicit strategies has been investigated, but implicit strategies (i.e. in which the system integrates previous input in the next question) were not very suitable due to the restricted dialogue structure. Confirmation strategies (i.e. in which the system takes an additional conversation turn), on the other hand, turned out to be a suitable method for repairing the conversation. This strategy was necessary when the system did not catch a keyword from the spoken sentence and therefore got stuck in the conversation. For the implementation of such strategy, the Google design guidelines for dealing with conversational errors are used⁸. These guidelines distinguished between a 'no match' occurrence and a 'no input' occurrence. A No

⁸<https://designguidelines.withgoogle.com/conversation/conversational-components/errors.html>

Match error occurred when the system did not understand or interpret the user's response in context, while a No Input error occurred when the system did not detect a response from the user (e.g. because the user has not said anything while the microphone was open, or the user has not spoken loud enough). Two different methods for dealing with these errors are implemented.

For the No Match error, some explicit confirmation sentences and questions were hard coded. These sentences were drawn from the following design suggestions:

- Reiterate the question quickly and succinctly.
- Combine apologies with questions.
- Talk to the user like you're having a human-to-human conversation.

Which resulted in the following list of confirmation questions:

- 'Pardon me, I did not hear you properly, could you repeat the answer?'
- 'Excuse me, what did you say?'
- 'Sorry, I did not understand you well, can you say that again?'
- 'Could you repeat that?'
- 'Sorry, I was not wearing my hearing aid, can you repeat your answer?'

The ASR often transcribed spoken speech in small parts of sentences or even separate words, which would cause an overload of extra conversation turns taken by the system in case these questions were asked every time no keyword was spotted. Therefore, this strategy is implemented so that the ASR checked if there was no keyword spotted and there was an unknown word (<unk>, as described in Section 2.3.1) in the transcription. Then the system took an extra turn and asked the user for explicit confirmation. These unknown words are actually words that are not in the vocabulary of the ASR, but in practice an unknown transcription also occurred when the user was speaking unclear and therefore not understood.

The No Input error was dealt with by repeating a statement after 20 seconds in which no keyword was spotted. In this situation the system 'assumed' that the user was not present or paying attention, or did not hear what the coach just said. After 20 seconds it repeated itself in the hope the user would respond the second time. Of course this implementation contains some flaws, because it only assumes the user did not pay attention or did not hear the statement, while it can be that the user has spoken for 20 seconds but without mentioning any of the keywords. A better approach is proposed in Section 7.

4.5 Additional Features

A few extra features were added to the speech-based COUCH system, attempting to achieve a good user experience. Guidelines for designing voice user interfaces were used, including the advice to use a command-and-control approach for systems that have no idea the user might speak [53]. In a conversational interface the system cannot automatically distinguish between a random conversation or the start of a conversation with the system. With a command-and-control approach users must do something explicit to inform the system that they are going to speak [53]. For example, Siri requires the user to press the home button before speaking. When this happens the system typically responds with audio and/or visual feedback, so the user knows that the system is listening and the user can speak. In the speech-based COUCH system such approach is implemented by creating buttons to start and stop the speech recognition (see Figure 4.14), allowing users to decide for themselves whether they want to use the speech option or not. It assures better privacy because there is no need for the recognizer to keep listening for

speech when the application is open, but not used by the user. The 'start speech' button got blocked once it was clicked and the 'stop speech' button was blocked while initializing the page or clicking the 'stop button'. Additional verbal communication speaking 'speech recognition is started' or 'speech recognition is stopped' is used to update the user about the state of the ASR. One big difference of our implementation with Siri is that the speech recognition only needs to be started once, while Siri requires the home button to be pressed every time the user wants to speak. Besides the verbal feedback, we implemented a visual cue that indicated whether the ASR was turned on. When the user clicked the 'start speech' button, a recording button (see Figure 4.15) appeared on the left side to make the user aware that the recognizer is turned on.

The last added feature was the megaphone (see Figure 4.16), which replaced the small text balloons icons implemented in the original COUCH application. Interactivity was added by making these megaphones responsive to clicking and causing the coaches to repeat the previously spoken statement. This feature was only possible in basic replies and not in autofoward replies, because in this second scenario the coach automatically continued to the next dialogue step after a short moment.

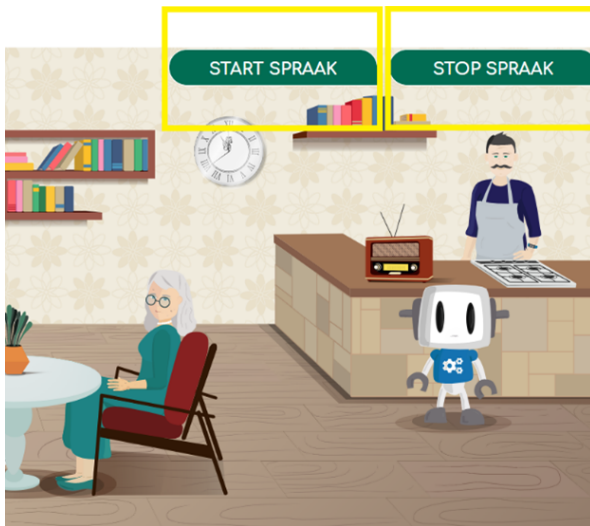


Figure 4.14: The buttons that turn on and off the ASR.



Figure 4.15: The record button that appears when the ASR is turned on.



Figure 4.16: The megaphone that can be clicked to make the coach repeat itself.

4.6 Removed and Ignored Features

Next to the addition of features in the speech-based system, it was necessary to remove or ignore some existing features. First, the radio was removed because this caused the ASR to capture and transcribe the music from the radio. Consequently, the system could spot a keyword in the transcription of the radio audio instead of the user audio, causing the system to continue to a next dialogue step without any action from the user. Another problem with the radio was the intensity of listening to the radio and at the same time listening to the coaches. Therefore we decided to remove the radio and all dialogue parts related to the radio from the system. Two other features which are not removed, but instead ignored by the ASR and TTS because of the irrelevance for the purpose of this study, included the recipe book and the text field input options. The recipe book could be reached by an interaction with François via voice, but it could not be navigated, opened or closed via speech. The same applied to conversations that contained text input fields. These parts could be reached by speech, but names or numbers that had to be entered as input could not be entered via speech. A physical click on the 'enter' is necessary when these text fields are filled.

Chapter 5

Methodology of Evaluation

This chapter presents the design of two separate experiments that were performed to supply participants with the system as described in Chapter 4. The first experiment is designed to explicitly address the differences in evaluation between the two systems, considering the number of errors that are made by the speech-based application. Therefore the external variables (i.e. the devices, the room, the environmental noise, etc.) are kept constant. This controlled study mainly focused on quantitative data and therefore a larger number of participants is required and is designed to answer Research Question 1. The second experiment is designed to test the robustness of the speech-based system in a home-setting and investigate the user experiences when used in such setting. A small number of participants is asked to use the speech-based application at home via a tablet and participate in an interview afterwards. This 'in-the-wild' study is designed to obtain qualitative data and answer Research Question 2. In the remaining of this report, the first controlled experimental study is referred to as *controlled experiment* and the second descriptive qualitative study as *field experiment*.

The chapter starts with a short description of the ethical procedure. Then it discusses per experiment the following subjects: a general description of the setup, a description of each of the measures, the participants, and the full procedure for a single participant. Additionally, a pilot test for the controlled experiment is discussed in Section 5.2.7.

5.1 Ethical permission

Before the start of the user studies, all required documents were send to the university's Ethics Committee for ethical approval. These documents are presented in Appendix B and included the information brochure, the consent form, and the ethics checklist. The consent form is a standard form, indicating that the participant is adequately informed about the experiment via the information brochure. Due to the COVID-19 crisis, the ethical approval was somewhat more complicated and included strict regulations. Extra precautionary measures were taken to comply to these regulations, such as: Guaranteeing 1.5m distance at any time, cleaning the equipment in between the sessions and conducting the interviews by telephone. An additional GDPR registration was done for both experiments because we collected personal data such as participant's demographics and voice recordings for the interviews. Both experiments were reviewed and approved by the Ethics Committee.

5.2 Controlled Experiment

The controlled experiment is designed to compare the speech-based version of COUCH with the original text-based application. During this experiment, participants are presented with both systems and asked to use them consecutively for a limited amount of time. The focus of

this study is to collect quantitative data via scale questionnaires and by counting the relative number errors. A few open questions are added to obtain better insights in the user preferences regarding the written- or speech-controlled interface.

5.2.1 Experimental Design

The experiment uses a within-subject design because of the low potential number of participants. This increases the chances of finding statistically significant results in the quantitative data. It also allows participants to make an explicit comparison between the two conditions in the final questionnaire, which may lead to additional insights. The independent variables are the two versions of the application (i.e. the original text-version and our developed speech-version) and the dependent variables are the observational measurements and user experiences ratings, measured via the user experience questionnaire (UEQ) and an explicit comparison questionnaire. To prevent the results being influenced by the order of the two conditions, we use a counter-balanced measure design. This means that the order in which the versions are presented to the participants is randomized, with half of the participants starting with the text-based application and half of the participants starting with the speech-based application.

5.2.2 Hypothesis

As introduced in Section 1.3, Research Question 1 asked:

Does the addition of speech to the Council of Coaches application lead to an increase in evaluations of the users' experiences?

Research Question 1 needs to be broken down into more specific hypothesis that can be validated or falsified through statistical analysis. Four hypothesis have been created for this purpose and are based on the findings described in Section 1.2 that different studies [17,18] suggest potential benefits for spoken conversational interfaces. Speech might be able to improve engagement because it can make the human-computer interaction more interesting and natural. The following hypothesis were thus formulated:

- H1 The interaction with the speech-based COUCH application leads to a significant increase in user evaluations of the UEQ scale *Stimulation*.
- H2 The interaction with the speech-based COUCH application leads to a significant increase in user evaluations of the UEQ scale *Novelty*.
- H3 The interaction with the speech-based COUCH application leads to a significant decrease in user evaluations of the UEQ scale *Efficiency*.
- H4 The interaction with the speech-based COUCH application leads to a significant increase in user evaluations measured via an explicit comparison questionnaire.

5.2.3 Experimental setup

To create a controlled experiment where the environmental conditions were the same for all participants, the experiment sessions were performed in the same lecture room in the Ravelijn building at the University of Twente. This lecture room proved to have a good acoustic, without background noise or resound. In this lecture room there were two desks placed side by side, with two chairs behind them. A computer was positioned at one of the desks and a separate laptop on the other desk. Participants used the COUCH systems on the laptop, which was connected with a snowball microphone to assure good quality speech input, and filled in the questionnaires on the computer. In this way, a distance of 1.5m between the researcher and participant could be preserved (e.g. when the participant filled in a questionnaire, the researcher could already set up the next system). The laptop and computer were cleaned between every session. The setup

of the experiment is pictured in Figure 5.1, where the laptop with microphone is positioned on the left desk, and the computer at the right desk.

Separate accounts were created by repeating the intake for every participant, making the dialogues more personalized. Names of participants could be used by the coaches to personally address the participant. Additionally, separate accounts kept the dialogue content constant for all participants, because no prior interaction had taken place via their account. The only interactions that were skipped for all participants included the introductory conversations wherein the coaches explained their role in the Council. Moreover, it was possible to retrieve log data from the interactions (e.g. date, time, coach, dialogue step, dialogue content), allowing us to check dialogue data used for the observational measurements. One account per user made this process much more structured.

Two documents were provided to the participants: (1) a sheet with an overview of the coach names and descriptions, and (2) a user manual. The user manual contained a general introduction and an explanation about the operation of the text-and speech-based application. Furthermore, this manual contained the steps that participants were asked to perform during the interaction with the speech-based system. These steps are also used as guideline during the interaction with the text-based system, although for this system the dialogues and interactions could not be manipulated, making it impossible to assure these exact options. Moreover, the steps are designed because they assure an interesting vocal interaction wherein the user is required to reply (i.e. there are not many autofoward replies) and without too many yes/no replies, which is less relevant in case of the textual interaction. The specified steps included the following:

1. Start an interaction with Coda and let him explain the interface to you.
2. Start an interaction with a coach of choice and go through with this conversation until finished and you are redirected back to the living room. Suggestions for interactions based on you interest are the following:

Active lifestyle advice: Olivia

Music instruments: Emma – I just want to talk

Soccer: Carlos – I just want to talk

Quiz about the memory: Helen – do a coaching session

Nutrition advice: Francois – talk about food

3. Start an interaction with a coach of choice and end it at a moment of choice.

Furthermore, participants were told that they could test two speech-specific functions:

- The repeat function of the megaphone
- The repetition of the coach by remaining quiet for a while

Participants were only required to have the interaction with Coda to explain the interface. The remainder of steps were not meant as hard requirements to follow, but more suggestions for interesting interactions and to get a broad idea about all system components. Participants were allowed to do interactions that were not on the list and to do steps in a different order.



Figure 5.1: The experimental setup of the controlled experiment.

5.2.4 Measures

This section describes the measures taken during the experiment, consisting of an intake questionnaire, observational measure, validated user experience questionnaire and a final questionnaire explicitly comparing the two systems. Questionnaires are presented via Google Forms, an online service. This allowed participants to fill out the surveys without requiring additional interaction with the researcher, and the data was directly available online. Complete versions of each questionnaire as presented to participants can be found in Appendix A.

Intake Questionnaire

The intake questionnaire is simply meant to collect demographics about each participant that could be relevant. Participants were asked to fill in their gender, age, level of education and working status. Additionally they were asked to indicate their level of experience with speech technologies. People with more experience in speech technologies (e.g. Siri, Google Home, Alexa) generally know how to interact with these kind of systems, possibly leading to a better understanding between human and machine and differences in the error rate [53].

Observational Measures

The first source of data is an observational measurement, in which the number of times participants are not understood or did not elicit a response from the system are counted. These measures are collected in realtime by keeping tally of different error types via the observational measurements sheet (presented in Appendix A.1.2) and are used give insights in the efficiency and reliability of the system. In case the system makes many errors in the recognition by misrecognizing speech or not recognizing speech at all, the system cannot be perceived as reliable and not efficient in use because the interaction often gets stuck.

The accuracy of the system is calculated with the number of times that there is no response from the system, which means the ASR fails in understanding and handling the spoken speech. However, since not all errors have the same effect on the system, it is important to distinguish between different types of errors. For the objective measurements done in this study, we defined different types of errors:

1. **ASR-errors:** the system is not recognizing the spoken speech and therefore is not responding to user input. The system cannot process to the next dialogue step.
2. **User-errors:** the user said something that was out of the vocabulary of the system and therefore could not be recognized.

The system partly dealt with ASR-errors by including conversation repair mechanisms (i.e. repeating the the statement after 20 seconds and querying one of the clarification questions). Although it was still an error by the system, it could be part of a humanlike conversation and is therefore called a ‘recovery’ error. Another type of error that occurred during the development of the system was in recognizing names of the coaches. However, this was not the most essential part of the conversation and not extensively researched in the first place, so these ‘name’-errors are treated separately and are not considered for the accuracy calculation. The accuracy calculation is based on the ‘fatal’ errors: errors that cause the dialogues to get stuck.

Just counting the number of errors to determine the accuracy was somewhat difficult due to the degree of freedom in interactions. Because not all participants did the same number of dialogue steps, counting misunderstandings by the system was more complicated. For this reason, a relative error percentage is calculated in which the number of misunderstandings is weighted against the number of dialogue steps. This approach could also account for autoforward replies, which are easier to get through because they do not expect a reply. The relative error percentage (p_{rel}) is calculated by dividing the absolute number of fatal errors(e_{abs}) by the total number of dialogue steps(n) and multiply it by 100. This calculation is defined in Formula 5.1.

$$p_{rel} = 100\% \frac{e_{abs}}{n} \quad (5.1)$$

Besides the ASR- and user-error types, one more type of error can be distinguished: dialogue-errors. In this situation the user says something which is correctly recognized by the system, but then continued to the wrong dialogue step. This means that the dialogue step was not labeled with correct keywords or multiple replies contained the same keyword and therefore a correct recognition did not result in the continuation to the intended dialogue step. Logging data can trace back to this type of error, but was not considered very important for the current research. Moreover, this type of error can be dealt with by carefully evaluating every reply option with its related keyword tag, especially for reply options that use the last word as keyword tag (although this might be a time-consuming process).

User Experience Questionnaire

The user experience is one of the most important concepts in designing any technological application. When user have no pleasant experience with an application, they are tended to never use it again. However, because this concept encompasses so many different aspects and can be very subjective and different for each user, it can be difficult to quantify user experience. One of the more commonly used tools for tackling this issues is the user experience questionnaire (UEQ) [54]. The original German version of the UEQ was created by a data analytical approach in order to ensure a practical relevance of the constructed scales. After extensive research and conducting many usability tests to test the reliability (i.e. the scales are consistent) and validity (i.e. the scales really measure what they intend to measure), the UEQ resulted in a questionnaire containing 6 scales with 26 items [55]:

- **Attractiveness:** Overall impression of the product. Do users like or dislike is?
Items: annoying/enjoyable, good/bad, unlikable/pleasing, unpleasant/pleasant, attractive/unattractive, friendly/unfriendly
- **Perspicuity:** Is it easy to get familiar with the product?
Items: not understandable/understandable, easy to learn/difficult to learn, complicated/easy, clear/confusing
- **Efficiency:** Can users solve their tasks with the product without unnecessary effort?
Items: fast/slow, inefficient/efficient, impractical/practical, organized/cluttered
- **Dependability:** Does the user feel in control of the interaction?
Items: unpredictable/predictable, obstructive/supportive, secure / not secure, meets expectations / does not meet expectations
- **Stimulation:** Is it exciting and motivating to use the product?
Items: valuable/inferior, boring/exiting, not interesting/interesting, motivating/demotivating
- **Novelty:** Is the product innovative and creative?
Items: creative/dull, inventive/conventional, usual/leading edge, conservative/innovative

These items are presented as seven-stage scale to reduce the well-known tendency bias for such types of items [54]. Each end of the scale contains one of the opposite items, as for example:

not understandable O O O O O O O understandable

The full list of questions is presented in Appendix A.1.3. The order of the positive and negative term for an item is randomized in the questionnaire. Per dimension half of the items start with the positive and half with the negative term [54]. Moreover, the order of the items is randomized per item, so the scales are not in sequence. Because this test has a strong psychological character, participants are asked not to spend a lot of time thinking about their answers, but rather to give responses based on their initial instincts.

The UEQ considers aspects of pragmatic and hedonic quality [56]. *Perspicuity*, *Efficiency* and *Dependability* are pragmatic quality aspects referring to the perceived usefulness, efficiency, and ease of use (so called utility and usability aspects). *Stimulation* and *Novelty* are hedonic quality aspects referring to the 'joy of use'. *Attractiveness* is only a dimension referring to the intrinsic attractiveness (i.e. positive qualities) or averseness (i.e. negative qualities). The items included in the attractiveness dimensions are almost all related to the appearance of the graphical interface. Since we barely made adjustments to the look of the application and items like 'attractive/unattractive', and 'friendly/unfriendly' were not very applicable, the attractiveness scale is removed from the questionnaire. No single items can be removed from a validated questionnaire like UEQ, but a complete dimension like attractiveness can [54]. By doing this, we also attempted to reduce the time demanded from the participants because the within-subject design already caused the experiment to be lengthy.

To analyze the data resulting from the UEQ, two different types of software were used: (1) an Excel-Tool for data analysis that was available free of charge via the UEQ website¹, (2) the statistical software SPSS. The second tool was used for analysis that were not available or incomplete in the provided Excel-Tool. Analysis performed via the Excel-Tool include:

- Data-pre-processing and assumption checking (for the t-test)
- The scale means and the means per item
- Cronbach's alpha

¹www.ueq-online.org

Analysis performed via SPSS include:

- Assumption checking (for the t-test)
- Paired samples t-test

The t-test available from the Excel-Tool was not used because this was an unpaired two-samples t-test, while we wanted to do a paired samples t-test. The paired samples t-test is used to determine whether the mean difference between two sets of observations is zero (i.e. there is no difference between the means). In a paired sample t-test, each subject is measured twice, resulting in two different sets of ratings for the same subject. As a parametric procedure (i.e. a procedure which estimates unknown parameters), the paired sample t-test makes several assumptions²: (1) the dependent variable must be continuous, (2) the observations are independent of one another, (3) the dependent variable should be approximately normally distributed, and (4) the dependent variable should not contain any outliers. Assumption (1) is met since the UEQ contains interval questions and for assumption (2) it can be reasonably assumed that the data collection process was random and all participants were independent of one another. Assumption (3) is checked in SPSS with the Shapiro-Wilk test for normality. All values with a significant value of $p > 0.05$ cannot reject the normality null hypothesis, and indicate a normal distribution. Assumption (4) is checked with the provided Excel-tool, which provides a simple heuristic checking how much the best and worst evaluation of an item in a scale differ [54]. Although such situations can also result from strong differences in opinions, or a misunderstanding of an item, it is seen as an indicator for a problematic data pattern when there is a big difference (>3) in the evaluation of an item. The occurrence of such a single case is not considered problematic, but when this is true for three or more scales, this might be an indication the response to be suspicious. The analysis tool suggested to remove answers from the data set that shows a critical value of 3 or higher. Assumptions (3) and (4) are checked during the data pre-processing (Section 6.1.2).

Besides comparing means and checking significance levels retrieved from the t-test, we were interested in the effect sizes; Cohen's D. A significant p-value tells us that there is a difference in user experience ratings, whereas an effect size tells us how big this difference is. SPSS does not support this statistical test, but the effect size for a paired-samples t-test could easily be calculated by dividing the mean difference by the standard deviation of the difference, as shown in Formula 5.2. An effect size around $d = 0.2$ is considered a small effect, $d = 0.5$ a medium effect and $d = 0.8$ a large effect.

$$d = \frac{\text{mean}}{SD} \quad (5.2)$$

Explicit comparison

The final questionnaire provided after interaction with the second system is designed to explicitly compare both COUCH versions. The questions posed here are not part of any existing questionnaire, but are instead formulated for the comparison of the text- and speech-based system. Inspiration for the questions came from an essay that described similar research comparing a coaching system employing plain text messages to deliver feedback with an ECA delivering the feedback [57]. Our explicit comparison questionnaire presented participants with 15 statements for which they had to indicate which of the two COUCH versions they felt the statement applied more strongly to, rated on a five-point scale. Values closer to -2 indicated a stronger association with the text-version, and values closer to +2 represent stronger associations with the speech-version. Zero corresponds with a neutral 'no preference' value. The statements related to different areas, some overlapping with items from the UEQ.

²www.statisticssolutions.com/manova-analysis-paired-sample-t-test/

For the following statements participants were asked: 'Please indicate which version you ...'

1. ... thought was more pleasant to use
2. ... thought was more efficient
3. ... thought was more interesting
4. ... thought was more credible
5. ... thought was more fun
6. ... would use for a longer period
7. ... would recommend to someone older then 55
8. ... would follow advice from more often
9. ... would rather spend money on
10. ... thought was more practical
11. ... thought was more cumbersome to use
12. ... thought was more monotonous
13. ... thought was more annoying
14. ... thought was more inconvenient to use
15. ... thought was more repetitive

The data is analyzed using SPSS. First, a one-sample t-test is used to determine if any of the mean values found is significantly above or below zero, the 'no preference' value. This indicates if for any question, there is a significant stronger tendency for the text-version (value close to -2) or the speech-version (value close to +2). Since we performed 15 different t-tests, the Alpha-coefficient is adjusted down, otherwise the number of false positives will become too high. Therefore we used the Bonferroni correction which is calculated by dividing the 'standard' alpha (0.05) by the number of t-tests. This resulted in a significance level of $p < 0.00333$. Similar to the UEQ questionnaire, we were also interested the effect sizes. Formula 5.2 is used to calculate the Cohen's D for the explicit comparison statements.

Besides asking participants to indicate their preference towards a version in a quantitative manner, we asked participants about their reasons why they had a specific preference via open questions. These questions were in the form of written open questions. The choice for written questions was made because the total number of interviews to conduct ($N=28$) was quite large. Additionally, due to the focus on quantitative data that was also acquired online, not many in-dept interview questions were asked. Lastly, participants for this experiment were mostly younger adults, who in general are able to easily read and write written text. The following questions were asked:

1. Which version did you prefer and why?
2. Which version was easiest for you to interact with and why?
3. Which version was most fun to use and why?
4. Which version would you recommend to someone older then 55?
5. What did you experience as advantage in the text-version?
6. What did you experience as disadvantage in the text-version?
7. What did you experience as advantage in the speech-version?
8. What did you experience as disadvantage in the speech-version?

To analyze the results of the open questions two methods were used. First, the number of participants that preferred a certain version in multiple aspects is investigated. Then the reasons why participants chose these versions are investigated. This is done by a thematic analysis in which the answers were bundled to common answers for every question. The results are presented

in separate tables, containing the reasons mentioned by participants and the number of times they mentioned that specific reason. The same happened with advantages and disadvantages, which are bundled to common answers and presented in a table, together with the number of times they are mentioned.

5.2.5 Participants

The study population for the controlled experiment consisted of both adults and older adults. The term older adult is defined as 55 years of age, and adult is defined as 18 years of age. Due to the nature of this experiment, which focused on obtaining quantitative data, a total of at least 25 participants was estimated to be sufficient. More participants is expected to give more reliable results, but due the controlled experimental setup, and the extra COVID-19 regulations, this is too time consuming.

The people approached to participate in the study were mainly students and fewer people from the target group. A total of 28 participants was approached via telephone or in real life at the university. The division of the group divided by gender, age, working status and experience with using speech technologies is presented in Table 5.1.

		19-24 (N=17)	25-30 (N=7)	55+ (N=4)	Total (N=28)
Gender	Female	11	5	2	18
	Male	6	2	2	10
Workstatus	employed	1	4	2	7
	retired	0	0	2	2
	student	16	3	0	19
Education	MBO	0	0	1	1
	HBO	0	1	1	2
	WO	17	6	2	25
Experience with speech technolo- gies	yes	1	1	2	4
	no	14	6	2	22
	moderate	2	0	0	2

Table 5.1: Demographics of the participants.

5.2.6 Procedure

This section discusses the procedure of the controlled experiment, which lasted a maximum of 35 minutes, with an average time of 25-30 minutes.

Recruiting phase

The recruiting phase for the controlled experiment was very simple and started by asking potential participants via the phone or in real life if they wanted to participate in the study. The potential participants recruited via phone were mainly friends and acquaintances of the researcher, while the people approached in real life were random students and employees present at the university. Due to COVID-19 crisis, approaching participants at university was difficult. Especially (generally older) employees worked from home, so could not be recruited at location. All potential participants received written or verbal information about the research, but not the complete information sheet yet. In case of a positive response, a moment is scheduled for the session.

Introductory phase

When entering the room, the participant is welcomed and provided with a small introduction about the experiment. At this point, the participant received the information sheet that contained a general explanation of the experiment, the system being tested, the data collected and the expectations that are placed on the participant. Moreover, the participant received the user manual, including the steps to perform during the experiment, and an overview with names and descriptions of all coaches. The participant is told that there is no right or wrong in performing the steps. Finally, the researcher asked the participant if there were any remaining questions about the experiment or the research in general. Once this procedure was done, a consent form is handed and the signing of this form by both the participant and researcher concluded the introduction. All forms and information sheets can be found in Appendix B. The introductory phase lasted approximately 5-10 minutes, but depended on the participant's reading speed.

Testing phase

The testing phase started with the participant completing the demographics questionnaire. After completion, the participant is asked to start using the first application and follow the steps as defined in the user manual. The researcher reported the numbers and nature of errors made by system on the observational measurement log sheet. The chosen dialogue paths and the number of dialogue steps were also added to this sheet. After the participant completed the steps for the first system, he received the first UEQ on the second device. In the meantime the researcher set up the second application. When the participant finished the first UEQ, he got back to the laptop and performed the steps for the second system. Then, the participant was asked to fill in the UEQ for the second system, and to continue with the explicit comparison and open questions. The testing phase took about 20-30 minutes.

Debriefing phase

During the debriefing phase of this experiment, the participant is thanked for his participation and asked if he had any questions or remarks. In case the participant is interested in receiving the results of the research, an email address is noted to send these in a later stadium.

5.2.7 Pilot Test

In order to test both the software and the procedure, a pilot test was performed prior to the actual experiment. No technical issues occurred during the pilot test, but it revealed that the experiment was too long. The initial list of steps to perform included six steps, from which three steps asked for a different type interaction (i.e. listen to a story, do a coaching session or listen to advice). Additionally, the pilot test indicated that the different dialogue types were not directly clear for the participant when using the application for the first time. The list of steps was reduced to four steps, from which one step included an interaction with a coach from beginning to end. Suggestions for interesting interactions were provided, instead of asking the user to have a specific interaction type. For example, the participant was suggested to interact with Carlos to talk about football when he was interested in football.

The initial plan for the observational measures was to record the time participants voluntarily played around, but the pilot test revealed that this measurement did not work. Participants already explored most options in the previous interactions and steps were not clearly performed one by one, but instead in random order. For this reason, timing the sessions was left out of the observational measurements, but the step to voluntarily play around remained included, so participants could decide themselves if they got a clear impression or wanted to see more.

5.3 Field Experiment

5.3.1 Experimental Design

Contrary to the first controlled experiment, the second experiment used an exploratory field research in which the system is tested in a natural setting. The field experiment is meant as 'long-term' study in which participants used the application in their own house for one week. It is used to test the robustness of the speech-based application in a natural environment, and thereby investigate the user experiences when used in such setting. To assess these differences, people with experience using the original text-based COUCH application by means of participation in a previous study [58] were approached for this experiment. The method to obtain data in this study was via qualitative in-depth interviews, attempting to obtain rich data.

5.3.2 Experimental Setup

The initial plan of the field experiment was to ask participants to use their own tablet, but some participants did not own a tablet that was able to run the speech application. For this reason they were provided with a Samsung Galaxy Tab A 10.1, which was delivered at home and picked up one week later. A shortcut to the web application was added to the tablet's home screen to make it easily accessible. All participants received the link to the application and a set of documents, including a user manual and a journal. The manual elaborately explained all functionalities of the system and contained a general introduction and help-desk contact details in case the participant ran into troubles. The journal (see Appendix B.2.4) was a compact form containing multiple text fields, used to write down the time spend with the system, technical problems experienced and general thoughts about the system. This journal is provided to help participants better remember their experiences and problems when discussing those during the interview. The participants without previous experience with the COUCH application additionally received the coach sheet that was used during the controlled experiment (Appendix B.1.3).

The interviews were recorded (audio only) using the recording function in telephone calls. After 24 hours these recordings were used to summarize the findings retrieved from the interviews. Complete transcripts were created, but only relevant comments by the participants were summarized per subject, analyzed per topic and presented in this report.

5.3.3 Interviews

The interviews were conducted by means of verbal communication, with as main reasons being the low number of participants, the nature of the experiment and the target group of older adults. In-depth interviews provide richer qualitative results which are potentially valuable. Additionally, older adults in general experience more difficulties reading and typing much text on a computer, so conducting interviews are more appropriate.

The interview questions were relatively informal and loosely structured and focused on understanding the target group's attitude towards the application, based on its robustness in a real-life setting. Additionally, one question regarding the user opinions compared to the original text-based system was added. The interview questions are listed in Appendix A.2.2, but since the interviews were not strictly organized, the final set of questions for every participant could differ somewhat. The general subjects that were discussed are the following:

- General impression of the system
- Practical problems experienced
- Way of interacting with the system
- Interest in using the application in the future
- Recommendation of the system
- Suggestions for improvements
- Preference version
- Additional comments

5.3.4 Participants

The study population consisted of older adults (i.e. 55 years of age). As described in Section 5.3.1, the intent of the field study was to recruit five participants from the group of people who participated in the original COUCH study [58]. This small number was enough because of the qualitative approach of the experiment. Unfortunately, only two ex-participants were recruited for the field experiment, which was even for qualitative data too little. For this reason, two other older adults without experience with the previous COUCH version were approached to participate. These differences between participants were carefully taken into account during the evaluation of data, but because of the qualitative nature of this experiment, we did not expect to experience any problems with this.

The final group of participants in the field study consisted of four participants between the age of 61 and 67 years. There were three female participants and one male participant, with an average age of 64 years. They all followed a higher education. Two participants were retired, one participant had a paid job and one participant did volunteer work. One participant had experience with speech technologies, while the other three had not.

5.3.5 Procedure

This section discusses the procedure of the field experiment. Because of the current COVID-19 regulations, physical contact was avoided as much as possible.

Recruiting phase

The recruitment of participants for the field experiment started with an advertisement in the newsletter from the original COUCH study. The potential participants who responded to this advertisement received general information about the research, but not the complete information sheet yet. When it became clear that no other participants would be recruited via the advertisement, two people in the age >55 were contacted via telephone to participate. At that moment, the information sheet, user manual and informed consent form were sent via email to all participants. The general information sheet contained the same type of information as the information sheet of the controlled experiment. The two participants who never used the COUCH application additionally received the coach sheet. With this email, the participants were also asked to confirm their participation. The participants who were able to digitally sign the consent, sent the signed form via email and the participants who experienced difficulties signing the form online did the consent procedure verbally during the introductory meeting. This consent form is similar to the one for the controlled experiment. The consent form, and coach- and information sheet can be found in Appendix B. Once participation had confirmed and eventual questions are answered, an appointment is made for the introductory meeting.

Introductory meeting

The introductory meeting lasted a maximum of 15 minutes and started with the verbal consent procedure (for the participants who had not given consent yet), by reading the description and items of the consent form. Participants were asked if they understood the content and if they agreed on participation. Afterwards, the procedure and setup and usage of the system was explained, in case that was not clear from the manual, as well as the journal which had to be filled in for each day of usage. Furthermore, an appointment for the final interview was planned about one week later than the introductory meeting. Finally, participants were asked to fill in the demographics questionnaire after the meeting. The two participants without previous experience with the original COUCH application were additionally asked to have one session with the text-based COUCH application to get insights in this version. Both links for the intake questionnaire and the text-based application were sent to the participants after the introductory meeting. When they finished, either with or without using the text-based system first, they could start the testing week. Completing these questionnaires and making an appointment for the debriefing concluded the introduction.

Debriefing phases

The debriefing is the final phase of the experiment, which took place after the testing period was completed. The researcher started a meeting via telephone to conduct the interview (discussed in Section 5.3.3). In the end, participants had the opportunity to ask questions or leave comments about the experiment or entire research. This debriefing phase lasted approximately 35 minutes.

Chapter 6

Results

This chapter presents the data analysis and results of the controlled study, designed to answer Research Question 1 (Section 6.1), and the field study, designed to answer Research Question 2 (Section 6.2). First, the results of the controlled experiment are discussed, starting with general observations made during the experiment. Then results of each source of collected data are presented: the observational measures, user experience questionnaires, explicit comparison questionnaire and answers provided to the open questions. In the second part of this chapter, the qualitative data collected from the field study is presented and discussed per question.

6.1 Controlled Experiment

A few observations were made during the experiment. First, participants who used speech technologies before seemed to 'try out' the system. These participants tried to say things that were related to the content written on the reply options, although not exactly matched and checked how far they could deviate from the reply options and still continue in the dialogue.

Second, the system made many errors in isolated and short words, as for example: 'oké', 'true', 'no', 'stop' and 'sure'. Most fatal ASR-errors were caused by speaking such words. This finding relates to our earlier approach to add the first word of a sentence as keyword to the list of keywords, but because the first word of a sentence was often not recognized, the implementation was changed by adding the last word to the keyword list. Besides short words, the system made many errors in the recognition of coach names, although this was dependent on the name and the person who pronounced it. Names like 'Emma' and 'Francois' were better captured than names like 'Helen' and 'Coda'. Moreover, the name Helen was difficult to pronounce for the TTS, suggesting that it might be a difficult name for computer processing in general. Because coach names were often spoken as isolated words, two problems were experienced simultaneously. When, for example, a sentence was spoken like '*Good to see you, Emma*', the response improved.

As last, it was noticeable that some participants directly stopped using the computer mouse while interacting with the speech-based system. Other participants hold their hands on the mouse to keep using it. As a consequence, some participants directly clicked on a coach or reply option when there was a slow response from the system. In general, when participants were told that they did not need the mouse, except from turning on or off the speech recognizer, the interaction got better. In this situation participants became more patient and waited for the system to respond, without clicking directly as alternative response mode.

6.1.1 Observational Measures

As described in Section 5.2.4, observational measures are performed to get insights in the accuracy of the speech recognizer. The error percentage is calculated by taking the number of dialogue steps, the number of fatal errors and the formula 5.1. These results are presented in Table 6.1, together with the number of different errors. Since user-errors are no errors caused by the ASR, these errors are not taken into account for the relative error percentage.

ID	Nr. of dialogue steps	ASR-error (no response)			User-error (no option)	Error Percentage
		Fatal	Names	Recovery		
1†	18	2	1	2	1	11.1%
2†	15	3	0	5	10	20.0%
3	21	2	0	0	1	9.5%
4	28	5	4	3	0	17.9%
5	21	4	0	2	1	19.0%
6	10	3	0	2	0	30.0%
7	20	4	0	0	0	20.0%
8	13	1	0	1	0	7.7%
9	12	3	0	4	0	25.0%
10	21	4	4	1	0	19.0%
11	35	8	1	0	0	22.9%
12	17	2	2	0	4	11.8%
13†	25	1	0	0	1	4.0%
14	16	4	0	1	1	25.0%
15	19	4	1	3	0	21.1%
16	24	3	1	0	0	12.5%
17	21	4	2	2	0	19.0%
18	13	3	0	2	1	23.1%
19	18	3	3	0	1	16.7%
20	15	3	4	1	0	20.0%
21	12	3	1	3	0	25.0%
22	25	7	3	0	0	28.0%
23	24	2	2	0	2	8.3%
24	18	8	0	1	0	44.4%
25	16	1	0	0	0	6.3%
26†	19	4	0	0	0	21.1%
27	25	2	0	1	1	8.0%
28	17	1	0	1	0	5.9%
Average	19.2	3.4	1.0	1.3	0.9	17.9%

Table 6.1: The number of different error types and relative percentage. Participants marked with the † symbol have experience in speech technologies.

Table 6.1 shows that a significant number of errors is made in the (correct) recognition of speech. The error percentages are especially high when compared with the text-version that (almost) never incorrectly or not responded to a user reply. The results of the observational measures show a wide spread of the error percentage and number of dialogue steps between participants (see Table 6.2). One participant (who did 25 steps) did not get a response from the system once, resulting in an error percentage of 4.0%, while a second participant (who did 18 steps) run into 8 fatal ASR-errors, resulting in an error percentage of 44.4%.

	N	Min	Max	Mean	Standard Deviation
Nr. of dialogue steps	28	10	35	19.21	5.520
ASR-errors	28	1	8	3.36	1.870
Names	28	0	4	1.04	1.401
Recoveries	28	0	5	1.25	1.378
User-errors	28	0	10	0.86	1.995
Error Percentage	28	4.0	44.4	17.94	8.91

Table 6.2: The descriptive statistics of the observational measurements.

Participants marked with the † symbol in Table 6.1 had experience using speech technologies (e.g. Google home, Siri and also one participants specifically with the NLSpraak recognizer). It is noteworthy that two participants with the most user-errors were participants with experience in speech technologies. This was also in line with the observation that these participants tried to test the boundaries and find out how far the vocabulary of the system reached.

6.1.2 User Experience

Data Pre-processing

As described in Section 5.2.4, the UEQ questionnaire is used to assess the differences in user experiences between the two system versions. However, to assess these differences with a t-test, still two assumptions had to be checked: the dependent variable should be approximately normally distributed, and the dependent variable should not contain any outliers.

First, we performed the Shapiro-Wilk test to check for a normal distribution of the dependent variable (i.e. the UEQ rating for each scale per system version). With the data from our text-version, a normal distribution was found for all scales. The data from the speech-version showed a normal distribution in all scales except the perspicuity scale, which resulted in a significance level of $p = 0.004$ ($p < 0.05$ indicates no normal distribution). The complete results are presented in Appendix C. Because this assumption is only violated for one variable we decided to continue and consider this when doing the t-test for the perspicuity scale.

The second assumption required that the dependent variable did not contain outliers. It can happen that not all participants answer all items seriously, although this problem is more often experienced when applying the UEQ as online questionnaire. Before performing the analysis, the data was checked for missing data. All participants responded to all questions in the questionnaires, leading to a complete dataset without missing values. Second, the data was checked for more or less random or not serious answers with the UEQ Excel-Tool. As mentioned in Section 5.2.4, big differences (>3) in the evaluation of an item are indicators for a problematic data pattern. In our study, no such critical value of 3 could be observed, indicating that no

random answers, response errors or misunderstandings were identified in any of the participants. This means that the overall UEQ had been understandable and the respondents had answered seriously. However, it is important to notice that in the current study the attractiveness scale is left out from the questionnaire, meaning that the number of scales is less than in the traditional UEQ version and the critical value benchmark might get lower. Nevertheless, only two participants per group (both text and speech) showed a critical value of 2, indicating that still the majority of respondents understood the questions and answered seriously.

Overview of System Evaluations

The mean and variances per UEQ scale (presented in Table 6.3a and Table 6.3b) are generated as starting point for the analysis. All values between -0.8 and 0.8 represent a neutral evaluation of the corresponding scale, while values > 0.8 represent a positive evaluation (values < -0.8 represent a negative evaluation). The results show a positive evaluation for the perspicuity, efficiency and dependability scales in both versions. An additional positive evaluation can be observed for the stimulation and novelty scales in the speech-version, while there is a neutral evaluation for these scales in the text-version. These findings already indicate a (small) difference in the evaluation of both systems, but this is statistically tested in a next sections.

UEQ scale	Mean	Variance
Perspicuity	↑ 1.722	0.28
Efficiency	↑ 1.120	0.61
Dependability	↑ 1.157	0.43
Stimulation	→ 0.426	1.00
Novelty	→ 0.343	1.58

(a) text-version

UEQ scale	Mean	Variance
Perspicuity	↑ 1.593	0.46
Efficiency	↑ 1.259	0.79
Dependability	↑ 0.963	0.44
Stimulation	↑ 1.259	0.51
Novelty	↑ 1.593	0.93

(b) speech-version

Table 6.3: UEQ scale mean and variance. A ↑ symbol indicates positive evaluations (values > 0.8) and the → symbol indicates neutral evaluations ($-0.8 < \text{value} < 0.8$).

Besides the means per scale, we looked into the system evaluations on an item-level. Figure 6.1 and Figure 6.2 show participant’s responses per item. The range of the scales is between -3 (horribly bad) and +3 (extremely good), but generally in real applications, only values in a restricted range are observed. It is, due to the calculation of means over a range of people with different opinions and answer tendencies (for example the avoidance of extreme answer categories), extremely unlikely to observe values above +2 or below -2. Thus, even a quite good value of +1.5 for a scale looks from the purely visual standpoint on a scale range of -3 to +3 not as positive as it really is. The same holds for negative values items.

As can be observed from Figure 6.1, negative items for the text-based system include that it is very usual (i.e. not in the leading edge) and boring to use. Figure 6.2 shows that the speed is ranked as negative item for the speech-based system. Nevertheless, both systems mainly received a positive evaluation for most items.

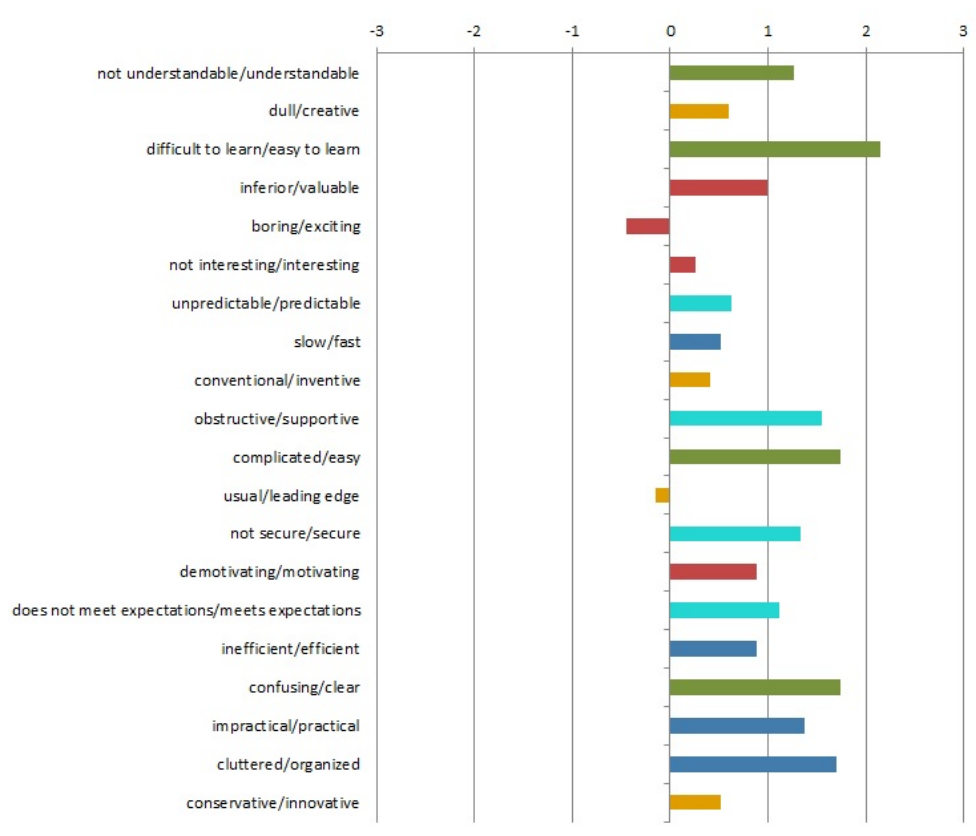


Figure 6.1: Mean values per item ranked for the text-version.

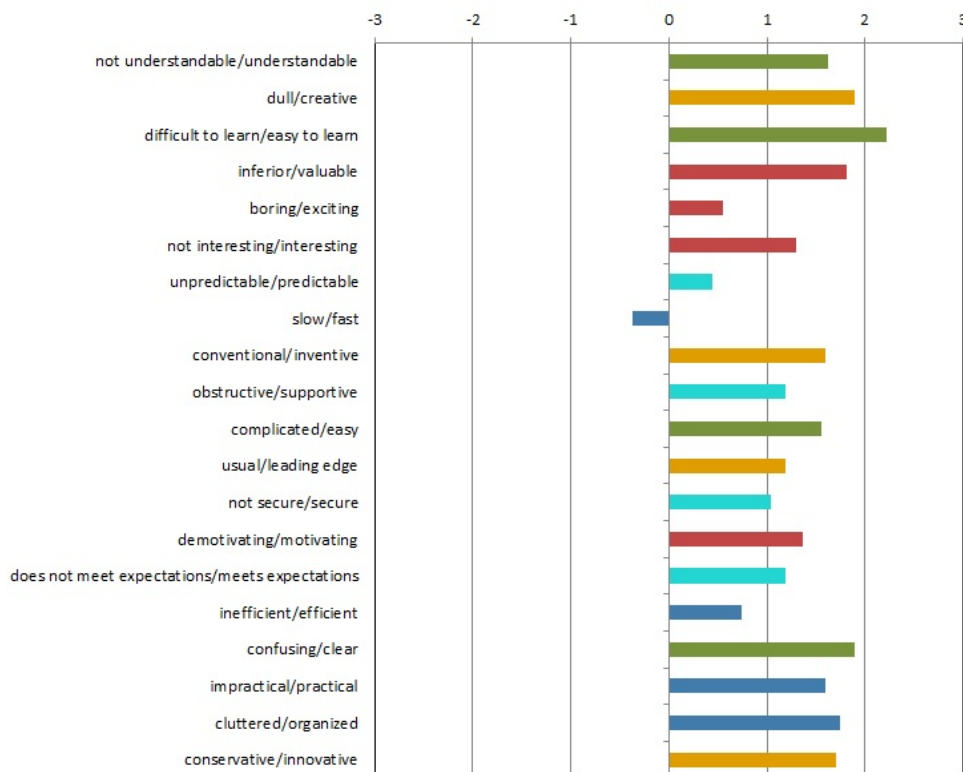


Figure 6.2: Mean values per item ranked for the speech-version.

Measure of Internal Consistency

A reliability analysis is performed to check the internal validity of the UEQ questionnaire per scale. Cronbach's Alpha Coefficient(α) is used as consistency measure and is calculated for each scale. The Alpha-coefficients (presented in Table 6.4) normally range between 0 and 1. The closer the coefficient is to 1, the greater the internal consistency of the items in the scale. There is no generally accepted rule how big the value of the coefficients should be, but a rule of thumb has been proposed by George and Mallery [59]: " $\alpha > 0.9$ – Excellent, $\alpha > 0.8$ – Good, $\alpha > 0.7$ – Acceptable, $\alpha > 0.6$ – Questionable, $\alpha > 0.5$ – Poor, and $\alpha < 0.5$ – Unacceptable". However, the value of the Alpha-Coefficient should be interpreted very carefully, especially in our case where we only have a small sample. In such cases a low Alpha-coefficient can result from sampling errors and may not be an indicator for a problem with the scale. Additionally, the UEQ handbook mentioned that scales which are irrelevant for a certain product may cause low Alpha-coefficients. In this case the responses might not be very consistent because participants have problems judging an UX quality aspect that is not important for the application [54].

Perspicuity		Efficiency		Dependability	
Items	Correlation	Items	Correlation	Items	Correlation
2, 4	0.19	2, 4	0.46	2, 4	-0.11
2, 13	-0.10	2, 13	0.18	2, 13	0.11
2, 21	0.15	2, 21	0.04	2, 21	0.12
4, 13	0.24	4, 13	0.44	4, 13	0.13
4, 21	0.20	4, 21	0.30	4, 21	0.11
13, 21	0.25	13, 21	0.38	13, 21	0.38
Average	0.15	Average	0.30	Average	0.12
Alpha	0.42	Alpha	0.63	Alpha	0.36

Stimulation		Novelty	
Items	Correlation	Items	Correlation
2, 4	0.49	2, 4	0.65
2, 13	0.60	2, 13	0.56
2, 21	0.49	2, 21	0.62
4, 13	0.80	4, 13	0.55
4, 21	0.34	4, 21	0.77
13, 21	0.40	13, 21	0.61
Average	0.52	Average	0.63
Alpha	0.81	Alpha	0.87

Table 6.4: Cronbach's alpha coefficient for the complete dataset (text and speech data).

As can be observed from Table 6.4, the Alpha-coefficients indicate the questionnaire to reach good reliability for the stimulation ($\alpha = 0.81$) and novelty ($\alpha = 0.87$) scales. This means that a high level of internal consistency for these scales can be observed from our data. The efficiency column presents a questionable value ($\alpha = 0.63$), while a very low Alpha-coefficient is obtained for the perspicuity scale ($\alpha = 0.42$) and dependability ($\alpha = 0.36$).

Since the UEQ questionnaire is a generally validated questionnaire, it is not good practice to remove any items from the data to improve the Alpha-coefficient. However, in such cases where a scale shows a massive deviation from a reasonable target value (e.g. 0.6 or 0.7), the corresponding scale should be interpreted very carefully. In our data this is the case for the perspicuity and dependability scale. While analyzing the results of the paired samples t-test (described in the next section), extra care should be taken for these low-scoring constructs.

Comparison of System Evaluations

To test Hypothesis 1, 2 and 3, we checked whether the scale means of the two measured systems differed significantly with a paired samples t-test. Table 3 and Table 4 in Appendix C show test results and statistic. P-values are used as indicator for significant results and Cohen's D values represent the effect size. A visualization of mean scores per scale for both systems is presented in Figure 6.3. The results will be explained per scale in more detail.

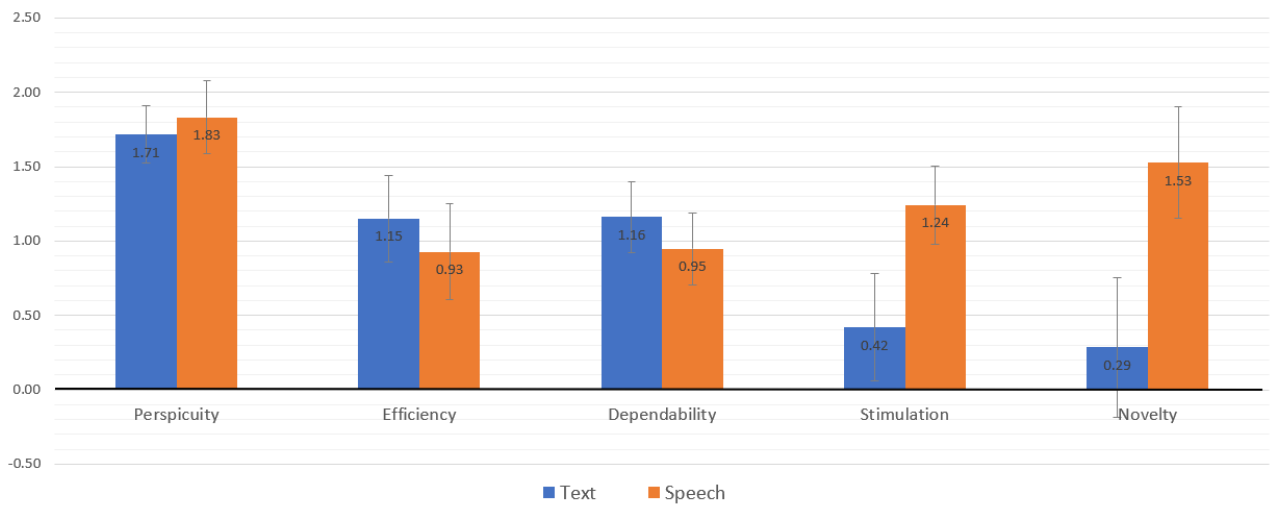


Figure 6.3: Visualization of the mean and variance of the UEQ scales.

Perspicuity:

There was no significant difference in the scores for the perspicuity scale of the text-version ($M = 1.71$, $DS = 0.52$) and speech-version ($M = 1.83$, $SD = 0.66$) conditions; $t(27) = -0.92$, $p = 0.366$. These results suggest that there is no difference in ratings of the perspicuity items between the two systems. Additionally, the effect size ($d = -0.174$) is very small.

Efficiency:

There was no significant difference in the scores for the efficiency scale of the text-version ($M = 1.15$, $SD = 0.79$) and speech-version ($M = 0.93$, $SD = 0.87$) conditions; $t(27) = 0.96$, $p = 0.346$. These results suggest that there is no difference in ratings of the efficiency items between the two systems. Additionally, the effect size ($d = 0.181$) is very small.

Dependability:

There was no significant difference in the scores for the dependability scale of the text-version ($M = 1.16$, $DS = 0.64$) and speech-version ($M = 0.95$, $SD = 0.65$) conditions; $t(27) = 1.677$, $p = 0.105$. These results suggest that there is no difference in ratings of the dependability items between the two systems. Additionally, the effect size ($d = 0.317$) is quite small.

Stimulation:

There was a significant difference in the scores for the stimulation scale of the text-version ($M = 0.42$, $DS = 0.98$) and speech-version ($M = 1.24$, $SD = 0.71$) conditions; $t(27) = -4.025$, $p < 0.001$. These results suggest that there is a difference in ratings of the stimulation items between the two systems. Additionally, the effect size ($d = -0.761$) is slightly below a large effect.

Novelty:

There was a significant difference in the scores for the novelty scale of the text-version ($M = 0.29$, $DS = 1.27$) and speech-version ($M = 1.53$, $SD = 1.01$) conditions; $t(27) = -5.753$, $p < 0.001$. These results suggest that there is a difference in ratings of the novelty items between the two systems. Additionally, the effect size ($d = -1.087$) is over a large effect.

Very low Alpha-coefficients were obtained for the perspicuity and dependability scales and the results of the t-test did not show any significant results. Therefore we can conclude that there is no reliable indication that the two systems are evaluated differently on these scales. Contrary, high Alpha-coefficients were obtained for the novelty and stimulating scales and the results of t-test did show significant results. The stimulation scale showed an effect size slightly below a large effect and the novelty scale an effect size over a large effect. This leads to the conclusion that participants found the speech-based system to be significantly and considerably more novel and stimulating. Therefore Hypothesis 1 and 2 can be confirmed: there is a significant increase in user evaluations measured with the UEQ for the novelty and stimulation scales. A decent alpha score was obtained for the efficiency scale, but no significant difference and a very small effect size was observed in participant's ratings regarding the efficiency of the system. Therefore we cannot conclude the text-based system to be more efficient than the speech-based system. This means that Hypothesis 3 cannot be confirmed: there is no significant decrease for the efficiency scale. In line with our expectations there is no significant difference in user evaluations of the perspicuity and dependability scales.

6.1.3 Explicit Comparison

As explained in Section 5.2.4, participants were asked to indicate which system applied more strongly to 15 different statements. The results of this measurement help to test Hypothesis 4. Table 5 in Appendix C shows the one-sample t-test statistics, including an overview of the mean values found for each statement. A visualization of these numbers is presented in Figure 6.4. From the graph we can see that the text-version scored higher on the positive attribute 'efficiency', but also on the negative attributes 'monotony' and 'repetitiveness'. On the other side, we see the speech-version to score higher on most of the positive statements (number one and three through nine), but also seems to be more cumbersome and inconvenient to use.

The results of the t-test are presented in Table 6.5. A number of significant results can be observed. A significantly stronger association with the speech-version is found for statements 3 till 9 (positive statements) and statement 12 (negative statement). A significantly stronger association with the text-version is found for statement 2. Most significant results are observed for the positive statements, while only one negative statement showed a significant difference. Participants believed that the text-version was monotonous while the speech-version was more interesting, credible and fun to use, and that they would use it for a longer period, recommend it to someone older than 55, follow advice from and would rather spend money on. On the other hand they experienced the speech-version to be less efficient and preferred the text-version in terms of efficiency.

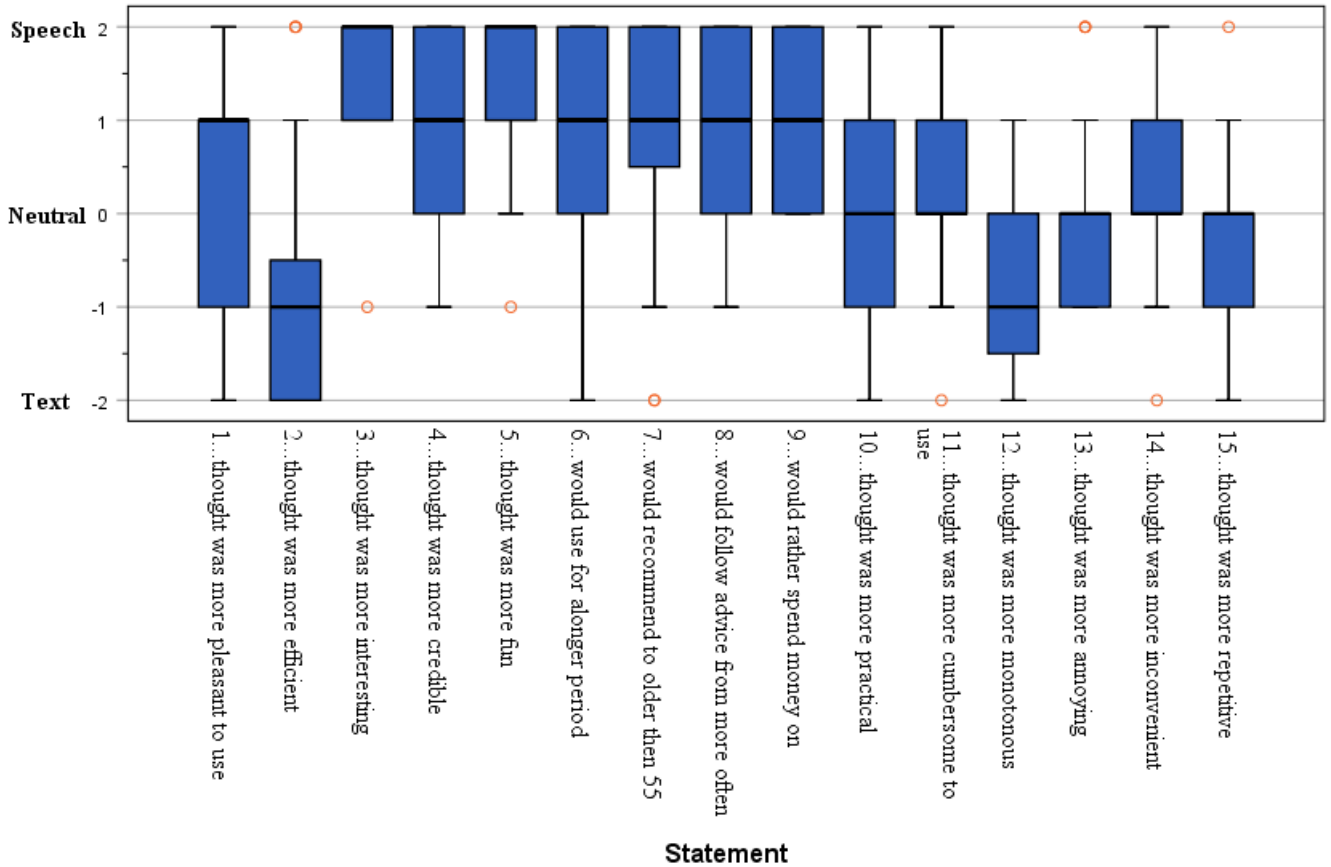


Figure 6.4: Boxplot of the explicit comparison results.

(Very) large effect sizes are observed for statements 3, 4, 5, 8, 9, and 12. This means that besides the statistically significant differences between the two versions for these statements, the differences in ratings were large, indicating that participants perceived the speech-version to be much more interesting, credible, fun and considerably less monotonous. Especially the statements 'interesting to use' and 'more fun to use' show very large effects, indicating a clear majority to be in favor for the speech-version regarding these statements. The statements 'would follow advice from more often' and 'would rather spend money on' also show large effects, indicating a strong tendency for the speech-version for these statements. The effect sizes for statement 2, 6, 7 can be classified between a medium and large effect, indicating a quite big (but somewhat smaller than previously discussed statements) difference between ratings.

The results suggest that Hypothesis 4 can for a large part be confirmed. The speech-based application leads to an increase in some user evaluations measured via the explicit comparison questionnaire, but not for all. Participants rated the speech-based system better on item 6 and 7, much better on item 3, 4, 5, 8, 9, 12 and much lower in terms of efficiency (item 2). Item 1, 10, 11, 13, 14 and 15 showed no significant difference between the two versions, indicating now explicit preference for those items. On the other hand, there was a decrease in the evaluations regarding the speech-based system's efficiency.

	Test value = 0						
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence interval of the difference		Cohen's D
					Lower	Upper	
1) ..thought was more pleasant to use.	1.072	27	0.293	0.286	-0.26	0.83	0.203
2) ..thought was more efficient.	-3.576	27	0.001	-0.857	-1.35	-0.37	-0.676
3) ..thought was more interesting.	12.813	27	0	1.643	1.38	1.91	2.421
4) ..thought was more credible.	4.674	27	0	0.857	0.48	1.23	0.883
5) ..thought was more fun.	10.003	27	0	1.5	1.19	1.81	1.89
6) ..would use for a longer period.	2.489	27	0.019	0.643	0.11	1.17	0.47
7) ..would recommend to someone older then 55.	3.855	27	0.001	0.929	0.43	1.42	0.729
8) ..would follow advice from more often.	5.612	27	0	1	0.63	1.37	1.061
9) ..would rather spend money on.	7.044	27	0	1.107	0.78	1.43	1.331
10) ..thought was more practical.	-1.216	27	0.234	-0.286	-0.77	0.2	-0.23
11) ..thought was more cumbersome to use.	1.987	27	0.057	0.357	-0.01	0.73	0.375
12) ..thought was more monotonous.	-4.804	27	0	-0.821	-1.17	-0.47	-0.908
13) ..thought was more annoying.	0.205	27	0.839	0.036	-0.32	0.39	0.039
14) ..thought was more inconvenient to use.	1.769	27	0.088	0.286	-0.05	0.62	0.334
15) ..thought was more repetitive.	-1.317	27	0.199	-0.25	-0.64	0.14	-0.249

Table 6.5: Results of the one-sample t-test performed with the explicit comparison statements. All bold numbers (and related statements) are significant values, based on $p < 0.00333$.

6.1.4 Open Questions

This section provides an overview of the topics from the set of open questions as discussed in Section 5.2.4 The results are discussed per question and the answers are bundled into a collection of most common answers.

Preference Version

A substantial majority of 85.7% of the participants ($N = 24$) preferred the speech-based system over the text-based system ($N = 3$). Hereby it is important to take into account that most participants were students and did not fall in the user target group. ==However, students were

asked to position themselves in the role of older adults and the four participants aged > 55 also preferred the speech-version over the text-version. Some quotes of reasons that participants mentioned in favor of the speech-version include: *'The speech-version gives a better impression of personal contact'*, and *'The speech-version was surprisingly fun. Hearing the answers out loud and having to say them back is a fun way of interacting with a computer.'* Other reasons mentioned in favor of the text-based system and the speech-based system are presented in Table 6.6 and Table 6.7 respectively. Two participants ($N = 2$, 7.1%) did not clearly state a preference for speech or text, but instead wrote: *"If speech recognition systems would develop and become better, I would say a speech-based system. Otherwise I would say a text-based system because for older adults it may be difficult to understand that they are restricted to the reply options and cannot use free speech"*, and *"The text-version is easier and faster in use. Speech is more innovative and might be convenient for people who cannot use their hands"*.

Reason	Number of times mentioned
No reason mentioned	1
Faster in use	2
Ability to determine your own speed	1
Easy to use	1

Table 6.6: Reasons that were mentioned in favor of the text-version.

Reason	Number of times mentioned
No reason mentioned	3
Easy to use (mainly the combination of written text and speech)	2
Keeps your focus	4
More innovative	2
More interactive	4
More fun to use	5
Interaction feels more natural and personal	6
Faster to use	2
Takes less effort for me to use	1
Stays interesting for a longer period of time	1
More convenient for older adults >55	2
You have the choice to speak or click	2

Table 6.7: Reasons that were mentioned in favor of the speech-version.

An important addition to the reason ‘interaction feels more natural and personal’ is that this natural interaction would improve when the speech recognizer responds faster. Participants who mentioned ‘easy to use’ as reason in favor of the speech-version mainly experienced the combination of written text and speech to be very helpful. This quality provided the option to answer by speaking as well as clicking. Moreover, it enabled participants to read and listen at the same time, making it easier to process the coaches’ information and enhancing focus.

Easiest to use:

Tables 6.6 and 6.7 showed the reason ‘easier to use’ in favor of both systems, which also resulted from the second question: ‘*What version do you think is easiest to use and why?*’ Although most people experienced the text-based system to be easier in use ($N = 18$, 57.1%), there were a number of participants who experienced the speech-based system to be easier ($N = 8$, 28.6%) and a few participants who experienced both system to be equally difficult ($N = 4$, 14.3%). A participant without preference mentioned: “*At this moment there is no difference, since only the text balloons can be spoken. When the speech-version is able to recognize an extensive vocabulary, this version will be easiest.*” According to this participant there is potential of the speech-based system in ease of use. Reasons regarding ease of use of the text-based system and the speech-based system are presented in Table 6.8 and Table 6.9 respectively.

Reason	Number of times mentioned
No reason mentioned	1
Restricted speech option	1
Speech was not always recognized	4
Speech took more time	1
You did not need any information from the beginning (e.g. coach names)	2
Less errors	3
Faster in responding	2
You did not have to pay attention if you were understood	1
You could use it in your own speed	2

Table 6.8: Reasons that were mentioned by people who experience the text-version to be easier.

Reason	Number of times mentioned
No reason mentioned	2
The application could be operated by voice and clicking	1
It is not necessary to click	2
You could give answers that were somewhat different from the answer options	1
Direct feedback	1
You remember more information when it is spoken	2

Table 6.9: Reasons that were mentioned by people who experience the speech-version to be easier.

One clear observation from these results is that in general the text-based system was perceived to work easier because more errors were made by the speech-based system. The system did not always respond to speech, while in the text-version (almost) no errors were made. Moreover, the text-based system was not very fast in responding, but adding speech to this version made the

system even slower, leaving participants in doubt if they were understood. Some participants thought the text-version to be easier because they could operate it in their own speed and did not have to wait for the coach to be finished speaking.

Most fun to use:

Although the text-based system won over the speech-based system in terms of easiness to use, all participants agreed that the speech-based system was more fun to use. The reasons that participants mentioned are presented in Table 6.10.

Reason	Number of times mentioned
No reason mentioned	3
More interaction	8
Suits better the to a conversation	1
More innovative	5
I felt like I could take more time in the interaction	1
More like talking to a person instead of a computer	5
More freedom in answering	1
Because of the sound, I felt like being spoken to	2
Interesting to test how much the coaches understood from what I was saying	1
More active	1
More interesting, because I never used speech technologies before	3
It is cool that it recognizes my voice	1

Table 6.10: Reasons that were mentioned why the speech-system was most fun to use.

The reason mentioned most by participants is that speech enhances interaction, leading to a conversation that feels much more like talking to a real person instead of a computer. Furthermore participants mentioned that speech technologies are quite novel and they did not use them a lot before, making it more interesting and fun for them to try out.

Recommendation for Older Adults

Because most participants were students aged <30 , one question was included which asked them about the version they would recommend to older adults (>55). A total of 23 participants (82.1%) would recommend the speech-version, while 2 participants (7.2%) would recommend the text-version and 3 participants (10.7%) recommended both versions. Although most participants did not indicate why they recommended a certain version, a few reasons were provided which are similar to reasons mentioned earlier, such as: *‘The natural interaction’*, *‘easier to use because no buttons need to be clicked’*, *‘more personal’*, *‘more effective in coaching’*, and *‘more fun’*. Noteworthy from the participants who recommended both systems was that their recommendation depended on the age of the user. Two participants recommended the text-version to adults >55 , but when adults become much older (e.g. >70), the speech-version was recommended. In contradiction, another participant recommended the speech-version to adults >55 and the text-version to the oldest adults, because: *‘the oldest adults may not understand that they are restricted in their answer options while using speech’*. Three participants recommended

the speech-version, provided that it will be improved with better recognition, a more extensive vocabulary and better response time.

Advantages and Disadvantages

Tables 6.11 and 6.12 provide an overview of the advantages of both systems as mentioned by the participants, and Tables 6.13 and 6.14 provide an overview of the disadvantages. Some advantages and disadvantages show overlap with earlier mentioned reasons.

Advantage	Number of times mentioned
It is simple to use	4
It is fast in use / Short reaction time	9
It is possible to use in noisy areas, and with music on the background	1
It is consistent and act like you expect	2
You can read/answer in your own speed	8
It makes less errors	5
Own interpretation of voices	1

Table 6.11: Advantages experienced in the text-version.

Advantage	Number of times mentioned
It is more interactive	5
It is more like a real conversation / a person is talking to me	10
It is much more fun	7
It costs less energy / is more relaxing	5
I take the information more seriously	3
It keeps your focus	2
You can communicate in your own way	2
You do not have to use the mouse/need hands for navigation	3
You do not have to look to the screen and in theory could do something else in the meantime	1
You do not have to read small letters on the screen	1

Table 6.12: Advantages experienced in the speech-version.

The biggest advantage of the text-version is related to speed. The text-based system is responding faster and participants can read and reply in their own speed, making it faster to navigate through. On the other hand, the text-version is experienced as boring, repetitive, impersonal and it is easy to get distracted from reading text.

Disadvantage	Number of times mentioned
There are little answer options / it is very predefined what you have to say	3
It is very repetitive	3
It is very usual/boring	9
You have to click a lot	2
It is slow	1
It feels impersonal	4
I tend to click through the messages very fast	1
Not suitable for people with visual impairments	2
It contains a lot of (unnecessary) text	3
It is easy to get distracted	3
Advices are less valuable	2
You easily read over (important) information	2

Table 6.13: Disadvantages experienced in the text-version.

Disadvantage	Number of times mentioned
Limited speech possibilities / small vocabulary	3
The reaction of a response (sometimes) takes long. This makes it unclear if you were understood.	11
Speech is not always recognized	8
More errors are made	1
Not suitable for people with hearing impairments	1
Wrong interpretation of the answer	1
You need to have a bit of foreknowledge in order to have an interaction	2
When what I said was not understood, it feels awkward to repeat myself	1
Voices are robotic and impersonal	1
Calling the coaches by name is not working optimal	1

Table 6.14: Disadvantages experienced in the speech-version.

The biggest advantage of the speech-version is that the conversation feels much more like talking to a real person, and for this reason participants tended to adopt information easier. This is an important finding for the purpose of COUCH that attempts to learn older adults to independently live happy and healthy. Another advantage is that using the speech-version costs less energy and easily keeps focus, which might lower the threshold for people to begin an interaction (e.g. after a busy and exhausting day). One last advantage is that the speech-version does not require to stay close to the screen or to use the mouse/hands for navigation. On the other hand, the speech-version responded too slow. Participants felt insecure about whether they

were understood by the system, which sometimes caused them to already repeat themselves even though they were understood. One participant mentioned: *“Waiting for the response of the system sometimes took too long, which gave the feeling that you were not understood, while this was almost never the case. The recognition worked very well!”* Although the recognition worked well according to this participant, a few others noticed as disadvantages that the speech was not always recognized.

One common disadvantage of both systems was the limited number of reply possibilities. Users would like to see options for more answers that are relevant to them, instead of a set irrelevant and predefined options. For speech this would mean that users would like to see more possibilities to use free speech and that they can define their own answers to a question or their own question for a coach.

6.2 Field Experiment

This section gives a summary of the interviews, which are analyzed per subject. Some questions were somewhat broad and related to each other, resulting in answers that were similar for multiple subjects and are therefore moved to different topics for analysis purposes.

6.2.1 General Impressions

Participants were first asked about their general impression of the system. They did not much appreciate the application in general, which made it more difficult to focus on their opinions towards speech. Participants indicated that the system became boring because the content was limited and not very diverse. In addition, using it for one week is limited because it is not possible to do multiple coaching sessions of one coach. Besides, participants did not appreciate not being able to provide their own answers (e.g. quote: *‘I could not even give the option - that sounds boring -, in a dialogue about the grandchildren of Carlos and instead had to choose between two irrelevant answers.* Furthermore, there was mentioned: *‘It is one protocol that you run, so that cannot really be called interactive’.* Lastly, errors in the dialogue (e.g. circled conversations, strange response to an answer, and errors in the statements) negatively influenced participants’ attitudes towards the system in general.

Besides the negative impression of the system in general, the system was received relatively well when focused on the speech implementation, although it had limitations (e.g. quotes: *‘It works in all its limitations’*, and *‘It works with speech, but it is a laborious process’*). Participants mentioned that they appreciated the ability of coaches to speak, especially the combination of written text and spoken sound was appreciated. This option allowed users to continue faster in cases the dialogue was not relevant and continued too slow. In addition, most participants experienced the voices to be very pleasant, adding value to the ability of coaches to talk. Participants mentioned that an interaction with speech was fun, but somewhat difficult. Having a vocal interaction was sometimes possible, but not always, while in a textual interaction it was clear what was possible and what no. Moreover, the response to speech was slow, which decreased the user experience and caused doubts about whether the user was understood. In case the system did not respond, the text buttons were used, which made the system similar to the original text-based system.

One participant had a negative experience with the system and perceived it as annoying and thought the voices to sound mechanic because the accentuation was not always correct.

6.2.2 Practical Problems

This section is divided in two parts. First, general difficulties experienced while running the experiment are described. Second, problems that participants experienced while using the system are discussed. These issues are considered for the analysis of results, because the participant who experienced most problems showed the most negative attitude towards the system.

General

Technical problems were considerably more present in the execution of the field experiment than the controlled experiment and was executed with a few major technical issues. Two participants managed to complete the testing period without any additional assistance or support, but the other participants contacted us about problems when running the application on an I-Pad. These problems could not be resolved in the given time frame, so an Android tablet was provided instead. This solved the problem for one participant, but the second participant continued to have problems because the provided tablet was old. It had a hard time running either the text- or speech-version of COUCH and caused a significant delay in both versions. Other problems with this tablet occurred with the microphone, which had to be manually reset every time the speech-version was started. A last issue came up when participants contacted us about the inability to log in, which was caused by an issue with the system certificates. Luckily, participants were able to use the system again one day later.

System

Besides general problems of the experiment, most participants experienced practical problems while using the application, although this was very different per participant. For one participant the system worked very well and it only happened a few times that there was no (correct) recognition. The system responded well and was particularly accurate in the recognition of coach names, even while it was used in a noisy room (i.e. the washing machine in the room was turned on). Other participants experienced more problems when using speech, but it depended on the moment they used it. This issue might be caused by the internet connection, which is not always stable in a home situation. Moreover, participants used their own equipment during the experiment, that would sometimes lead to technical issues.

It appeared that some words were more difficult to recognize than others. For example, saying 'stop' to end the conversation never worked, but saying 'end' did work. Furthermore, participants mentioned that the system response was slow. Although this was worst when using speech to interact, it was also experienced when using the scripted text-buttons. Most problems originated in François, who's response speed was slower compared to the other coaches. This sometimes caused the feeling of misunderstanding between human and machine, especially when using speech because there it resulted in errors. Moreover, one participant mentioned that the ASR was sensitive for the intonation of words. Simple words like 'oké' were understood well when it was pronounced in one way, but when spoken differently, the system did not respond at all.

Sometimes participants chose a reply option, but the resulting next dialogue step did not make sense for this reply. Occasionally this was an error by the speech-based system that captured a wrong keyword and therefore continued to the wrong dialogue step. Other times it was a strange dialogue transition (which was similar in the original application when checked by the participant). Another problem related to the dialogue content was that in some situations multiple answers contained the same word (e.g. the word 'tell' in two answer options): *'Then you have to stick very carefully to the text, because it is then unclear what the important word to say is'*.

6.2.3 Way of Interaction

One question asked in which way participants talked to the coaches and if they wanted to see this differently. Participants mentioned that they tried to test the boundaries of the vocabulary of the system by deviating from the reply options. When this elicited no response, they adhered to the given reply options. Some participants tried to use free speech, but quickly found out this did not work.

Participants wanted to see more dialogue content and an extensive set of reply options to make the interaction more like interacting with a real coach. Participants mentioned that the text-version of COUCH should already be implemented with free text, so users can type their own answers. In case the system understands a larger vocabulary, the user can provide many different replies and the system knows how to respond to each of them. More coaching content is required to appropriately respond to each reply. Such system would offer more possibilities to implement, to some extend, free speech, which is preferred to make the interaction more natural, interesting and useful. Moreover, free speech could be helpful for visually impaired people who cannot read the text on the screen.

One last problem with the interaction was that participants not appreciated the interactions where coaches did not properly greet or say goodbye. When a coach started the conversation by saying; *"hello, good to see you, how are you?"*, this was appreciated more then when a coach directly started with; *"do you want to know more about how to be more active?"*. Additionally, participants preferred the coaches to clearly say 'goodbye' to end the conversation.

6.2.4 Use and Recommendation of the Application

Participants were asked if they were interested in using the application in the future and would recommend it to others. All participants indicated not to want to use the application in the future, even though the speech recognition and responsiveness would be improved. The main issue mentioned was the little dialogue content and the limited number of reply options. However, this result is more related to the COUCH system as whole then to the speech implementation specifically. Participants mentioned that this disinterest might be caused by the limitations of the study, because they were not able to see all coaching content due to the time constraints of the application (e.g. the weekly coaching sessions) and to connect an activity tracker.

On the other hand, all participants would recommend the system to others, especially to people in specific user groups. The application is expected to be more fun and informative for less educated people and for people with less computer experience. Moreover, the application is suggested as helping hand for older adults following therapy, wherefore is mentioned: *These people can go to therapy and receive information there, but still have to apply it at home themselves. An application like this can help them in between the therapy sessions, especially when the therapy content is a lot to process and remember*'. Lastly participants mentioned the application to be useful for older adults living alone, who can benefit from the social function of the application.

6.2.5 Suggestions for Improvements

E-health applications like COUCH are considered important by the participants, but improvements are necessary to make the application more effective (e.g. quote: *"These types of systems have to be developed, but with these you quickly get stuck"*). Participants were asked for possible improvements to the system, resulting in a few concrete ideas.

Participants preferred a dynamic interface with animated coaches. A 3D interface with human-like characters is preferred most, but a 2D interface with characters showing facial expressions, hand gestures and lip movements to indicate they talk (e.g. like in an animated movies), is a good second option. Besides, there is suggested to sometimes move the coaches in the living room to train the memory of the user and keep the application interesting.

As already explained in Section 'The way of interaction' (Subsection 6.2.3), expanding the number of reply options and dialogue content is suggested. This improvement would allow people to participate in a more natural conversation and for visually impaired people to use the application. This is especially important since aging often occurs with visual deterioration. One suggestion to overcome this problem and avoid the complexity of free speech, is to verbally provide all possible answers. For example, the coach can say '*do you want to discuss something personal?*', or '*do you want to do hear tips about your diet?*'. This approach easily fits in the structured COUCH structure.

Another idea mentioned was to provide feedback about the speech that is recognized. In the current implementation it was clear that the ASR and microphone has turned on, but it did not indicate whether or not the incoming speech was transcribed. It was not perceived as problem that no keywords were spotted, as long it was clear that speech was transcribed. Although one participant indicated this preference, another participant expected this graphical feature to be unnecessary when the delay and slowness of the system was improved.

6.2.6 Preference Version

The interview ended by asking participants about their preference version. The two participants who participated in a previous COUCH study and had experience using the text-version, differed in opinion. One participant preferred the speech-version, because this gave the feeling of a real conversation and coaching-session. The second participant, who experienced a lot of general problems during the start of the study, was annoyed by using the speech-based system because it responded too slow and repetition was almost always necessary. The two participants without previous experience with the text-version (i.e. who only used the text-version once) both preferred the speech-version because speech improved the experience of a more humanlike interaction with the coaches. The combination of spoken speech and written text was appreciated, but there was also mentioned that for some older adults (especially the oldest ones) this combination might be confusing.

Chapter 7

Discussion

In this chapter, we discuss the results derived during this project. We start by presenting a complete overview of the answers to the sub-questions. Then we discuss the original research questions and main question and compare these conclusions with results derived from prior research. As last, recommendations regarding future work in this area are presented.

7.1 Discussion of the Sub-questions

A large part of the sub-questions were already addressed during earlier stages of the study (Chapters 2 and 4), but are repeated here for completeness:

SQ:1 *What are current limitations in text-to-speech synthesis software and how can this problem be addressed?*

The fundamental limitation of speech synthesizers is to create natural sounding synthetic speech. Despite the fact that modern TTS have reached a level of voice quality that no longer resembles robot-like voices, it remains very difficult for systems to vocally express emotion. Other limitations of TTS software is found in the correct pronunciation of names and foreign words, and generating correct prosody. Although the lack of vocal emotion is a limitation that is very hard to address, the correct pronunciation of words and names can be addressed by integrating SSML specifications in the synthesis request. The Google TTS supports the SSML '<sub>' attribute that is used for words like; 'wow', 'Helen', 'Olá' 'Emilie', but did not work for words like 'hmmm'.

SQ:2 *How robust is the current state of art in text-to-speech synthesis to create multiple human-like voices necessary to ensure a natural interaction with all coaches?*

The current state of art in text-to-speech synthesizers does not offer many options to create multiple humanlike voices. Most TTS software is paid, only offers one voice or is not available for the Dutch language. Only the Google TTS met all requirements, but also asked money when used on a larger scale. For this project we could use a try-out package that enabled us to use spend €300.- on synthesis requests in a time period of one year. In terms of robustness we can conclude that the TTS is already quite advanced. The voices are not identical to human voices yet, and there were some problems in the pronunciation of certain words, but overall the voices were perceived as clear and pleasant to listen to. Especially for an application like COUCH where all coaching content is static (i.e. the content is written by experts and does not involve dynamic generation of text), no new text has to be synthesized. Therefore TTS can be a robust solution.

SQ:3 *What additions to the graphical user interface are necessary to assure a pleasant user experience with the voice-based application?*

Some graphical user interface additions were thought in advance, following the design guidelines proposed by Pearl [53]. Other suggestions for adjustments were the result of the two user studies. Simple visual feedback in the form of a start- and stop-speech buttons were implemented to let the user know that the system was listening. This implementation gave verbal and visual feedback about the listening state of the system because users may get frustrated when it is unclear whether the system is listening. Some users liked to see more visual cues than only knowing that the microphone was turned on. They would have liked to see an indication that the system was really catching speech, but did not spot the correct keywords for example.

The controlled study showed that it was not clear for all users if they had to speak the 'continue' button in autoforward replies. The system automatically continued after a very short time, but because there was a button users thought they had to say or click 'continue' as well. Therefore this button was removed from the interface.

Cartoon-like characters as used in COUCH can lower customer expectations towards the technical abilities of the system [41]. However, participants of the field experiment indicated a preference for a less static application, wherein coaches are animated and show lip movements. Moreover, a more dynamic interface was suggested by changing the position of the coaches in the living room now and then. This better enhances the user's focus and keeps the memory sharp.

SQ:4 *How should the dialogues be adjusted in order to function with the implementation of speech into the Council of Coaches?*

We cannot conclude that there is only one way in which the dialogues should be adjusted in order to function with the implementation in the COUCH. However, there is one way (implemented during this research) that is most logical due to the structured dialogue architecture of the original COUCH application. The complete dialogue structure is build upon the WOOL framework, which is essentially a definition of a series of dialogues steps, linked through user replies. These dialogue steps were manipulated individually, by adding keyword tags as general actions to statements, without changing the structure of the complete hierarchy. This content could be retrieved by the COUCH Web Client and used to compare with the transcription of speech. When the keyword was spotted in the transcription, the Web Client could mimic a click event so the user continued in the dialogue and the content of the next dialogue step (i.e. the speaking coach, the coaching text, the reply options, etc.) could be retrieved. It was not necessary to adjust the structure and content of the dialogues, except from removing some irrelevant dialogues for the user studies. Also, words that could not be pronounced correctly by the TTS were removed from the dialogues.

To allow for free speech in the COUCH application, adjustments the WOOL-based dialogue structure are required. Users can define their own replies when using free speech, instead of speaking predefined replies. This requires the application to generate new coaching content based on user replies, otherwise the system does not know how to respond correctly. Changing the complete dialogue structure of COUCH is not the most convenient method to implement free speech into the application. Instead, another approach could be to link all user input to one of the reply options by using large keyword sets for every reply. Keywords sets can be automatically created based on the content of the reply, enabling for free speech (to some extent) and at the same time keeping the dialogues structure. This suggestion is explained in more detail in Section 7.5.2. The main drawback of this approach is that it is not always appropriate for dialogue content that cannot be related to any of the reply options.

7.2 Research Question 1

7.2.1 Answering Research Question 1

Research Question 1 stated:

Does the addition of speech to the Council of Coaches application lead to an increase in evaluations of the users' experiences?

Most participants (24 out of 28) preferred the COUCH speech-version over the text-version. Moreover, significant higher ratings of the novelty scale (with items: valuable, exciting, interesting and motivating) and stimulation scale (with items: creative, inventive, leading edge, innovative) are observed. The Alpha-coefficients for these scales were high, indicating that the construct was reliable and the items were measuring the same scale. Because the effect size was large for both constructs, the implementation of speech is expected to contribute to an increase in user evaluations of the stimulation and novelty scales, supporting Hypothesis 1 and 2 respectively. Hypothesis 3 cannot be supported, indicating that the addition of speech to the COUCH application does not lead to a significant decrease in user evaluations of the efficiency scale. No reliable Alpha-coefficient and significant test results could be observed for the efficiency scale.

Results of the explicit comparison showed that participants perceived the speech-version as more interesting, credible, fun to use and less monotonous than the text-version and that they would rather spend money on the speech-version, sooner recommend it to older adults, follow advice from and use it for a longer period of time. These results partly confirm Hypothesis 4, which expected a significant increase in user evaluations measured via the explicit comparison. No increase in evaluation is obtained for all items, but for 8 from the 15 items. One item, the efficiency, showed a significant decrease in evaluation of the speech-version and the remaining 6 items did not yield any significant results.

The open questions supported the results of the UEQ and explicit comparison. All participants experienced the speech-version to be more fun to use and most participants would recommend the speech-version to older adults. Advantages of the speech-version included: more interactive, exciting, interesting, and innovative. Two other important findings from the interview questions were that many participants mentioned that they really had the feeling of participating in a real conversation with a human-being, and that the interaction with the speech-version costed less energy and maintained focus.

7.2.2 Discussion of the Results

Most results derived from the UEQ, explicit comparison and open questions were in line. Multiple tests supported the finding that the text-version was very usual and boring to use, while the speech-version was perceived as interesting and entertaining. On the other hand, some contradicting results were obtained. The explicit comparison and UEQ results regarding efficiency were not in line. No significant difference in terms of efficiency could be observed, while this scale contains items like 'fast', 'efficient', 'practical', 'organized'. This finding contradicts the results of the explicit comparison, that showed a significantly stronger association with the text-version for the efficiency statement, and the open questions, that mentioned main advantages of the text-version to be: the short reaction time, the speed of use and the fact that it makes less errors. The original COUCH application already needed time to respond, but the implementation of

speech significantly decreased the response speed of the application. This problem was already experienced during the development phase, but was difficult to solve. It takes time to send the spoken speech to the NLSpraak server, wait for the transcription and to convert this transcription in a dialogue choice. The slow speed made the speech-based system more sensitive for errors.

No significant difference was observed in the ratings of perspicuity (*items: understandable, easy to learn, easy, clear*) and dependability (*items: predictable, supportive, secure, meets expectations*). This result was expected because the two versions are, apart from the textual or vocal interaction, very similar in the UEQ items (e.g. both systems are perceived to be very understandable and predictable). Additionally, the internal consistency measure resulted in very low Alpha-coefficients for the perspicuity and dependability scales, making it hard to draw conclusion in case the results were significant.

The number of errors made by the system differed a lot between participants. The relative error percentage ranged between 4.0% and 44.4%, meaning that the highest error percentage was more than 10 times the lowest error percentage. It is difficult to determine what caused this wide spread in errors. It might have technical causes, such as the load on the NLSpraak server at a specific time, or causes related to participant's speech characteristics. One of the limitations mentioned in Section 2.4 stated that ASRs have difficulties with speech input variations, for example caused by the characteristics of older voices. This might cause the ASR to deteriorate in the recognition of older voices and therefore decreasing the user experience. During the experiment the ASR did not seem to perform worse for older voices, but there might be other characteristics that caused the difference in errors (e.g. the loudness or pitch of the voice). Furthermore, younger and older participants differed in level of patience. In general, the older participants and participants who used speech technologies before, showed more patience while using the application. Therefore, less errors occurred in the system.

In addition to the wide spread of system errors, a wide spread in number of dialogue steps was observed. Participants were quite free in their interactions except from the interaction with Coda. Suggestions for interesting interactions were provided, but participants were not required to stick to these options. These suggestions were only based on interesting dialogue content and not on the length of dialogues. Moreover, not all participants showed the same level of interest in trying the application. These differences resulted in a wide spread in dialogue steps that users performed. However, since we calculated the relative error percentage, we did not expect this to be a big problem.

An important and surprising finding of the controlled experiment is that the speech-based application made it easier to maintain focus. It appeared to require participants less energy to interact via speech. This finding may positively influence the long-term engagement because it may lower the threshold to use the application after a long and exhausting day.

Closing Summary

Summarizing we can conclude that the addition of speech to the Council of Coaches application leads to an increase in user evaluations, even though the system is susceptible for errors and the efficiency is not optimal.

7.2.3 Limitations of the Controlled Experiment

For the interpretation of the results of Research Question 1, some limitations regarding the participants and experimental setup are considered.

Participant Demographics

The first limitation concerns the participant demographics that differ from the target user group and thus the generalization of results. The group of participants mainly consisted of students or young employees <30 years, only four participants were truly in the age category of older adults. This made it more difficult to conclude if older adults would also perceive a better user experience. In general, younger people (and especially high-educated students) have more affinity and experience with technologies, potentially creating bias in the evaluation of the more complex speech-version of COUCH. On the other hand, most younger participants did not have previous experience with speech technologies, while two of the four older adults had. Besides, this small group of older adults preferred the speech-based system over the text-based system. However, it is still more valid to use a participant group homogeneous to the target user group.

Number of Participants

A total of 28 participants participated in the study, which is a relatively small number for statistical calculations and increases the risk of finding false-positive reports [60]. A false positive is an error in which the test results incorrectly indicate the presence of a significant difference in user evaluation when in fact this is not present. However, since our study clearly showed a large and significant effect on the scales that were in line with our expectations and observations, we expect this risk to be minimal.

Experimental Setup

Some participants seemed to feel a bit insecure and uncomfortable during the experiment, probably because they were alone in a quiet room with the researcher, or did not exactly know what to expect. This made them a bit hesitant during the conversation, especially in the beginning. As a consequence, their voice was somewhat quiet and hesitant, making it more difficult for the system to catch the speech correctly. This problem might be overcome by adjusting the experimental setting, such that the participant is more at ease. This can be achieved by, for example, using a one-sided transparent screen for observation or by filming instead of observing all interactions.

Observational Measurement

The setup for the observational measurement can be improved because the researcher guided the entire experiment and counted the number of errors and dialogue steps, which was a lot to do at the same time. Logged history made it possible to retrieve information about the chosen interactions and reconstruct the dialogue path to count the number of dialogue steps, but this information was limited and not always clear. Logging information about the transcription created by the ASR and keywords added in each step could be a valuable addition because it was not always possible to see exactly what went wrong in the system (e.g. did the user say something that was not possible, or did the system incorrectly recognize it?).

7.3 Research Question 2

7.3.1 Answering Research Question 2

Research Question 2 stated:

How robust is the current state of art in speech recognition systems to create a usable and enjoyable system that can be used in a real-life (home) setting?

From the field study it is difficult to conclude whether the current state of art in speech recognition systems is robust enough to create a usable and enjoyable system that can be used in a real-life setting. It was very different between participants, and even within participants using the same device every time, how much problems in speech recognition were experienced. One participant could use the application with the washing machine turned on in the same room, while another participant mentioned that the coaches barely responded to their names, even without any environmental noise. These results indicate that the robustness of a speech-based application is not only depending on the ASR, but also depending on the performance of the device (i.e. the speed), the quality of the microphone and the stability of the internet connection.

The second reason why it was difficult to conclude whether the current state of art in speech recognition systems was robust enough to create a usable and enjoyable system, was that participants did not appreciate the COUCH application in general. However, most participants preferred the speech-version over the text-based application. Although none of the participants would use the application themselves, they would recommend it to others, especially to different user groups. In particular, the ability of coaches to talk was well received, especially because the voices were very clear and intelligible.

7.3.2 Discussion of the Results

A clear difference in attitude towards the speech-based system was observed between participants who experienced little and participants who experienced many problems during the experiment. The participant who experienced fewest problems in speech recognition was much more positive about the speech implementation than the participant who had many problems with the startup and during the experiment. One participant mentioned not to be a fan of speech technologies in general, but did appreciate the combination of speech and text made the speech-version.

The reason that participants did not appreciate the COUCH application in general was that it included too little coaching content and reply options. This was partly caused by the time constraints of the coaching sessions, that disabled users to continue with the next coaching sessions, which participants saw rather disappear (quote: *'In current society, it is not a practical reality to do this, users just want to decide for themselves and continue sooner'*). The time settings could have been adjusted for the field experiment to assure a better experience.

One participant of the field study mentioned that the responsiveness of the system depended on the intonation of the spoken words. This participant also mentioned that her voice normally is quite melodic and ranging from high to low pitch. The system often not captured her speech correctly, only when the intonation was exactly right. The need for a specific intonation was not experienced by other participants, so it might be caused by the fact that the ASR had to listen to a female, older and melodic voice, which is more complicated for an ASR than listening to a male and younger voice [35, 37]. Additionally, results from the observational measures in the controlled experiment showed that there was a wide spread in the number of errors made

by the system. This indicated that not all participants were equally good understandable for the system. However, during this experiment no clear distinction could be made between voices that were better understandable or less understandable (e.g. female vs male voices, young vs old voices, loud vs soft voices).

Closing Summary

Participants preferred the speech-version over the text-version, even though practical problems with the application were experienced. How well the system worked fluctuated a lot, making the current state of art in ASR systems and commercial devices (e.g. microphones in tablets and laptops) not robust enough for applications based on only speech in- and output. However, the combination of speech and text made the implementation of speech in Council of Coaches application robust enough. Generally, participants preferred a vocal interaction, but in case the system did not respond, the buttons could be consulted. Only the ability of coaches to speak (i.e. TTS) was already appreciated a lot.

7.3.3 Limitations

For the interpretation of the results of Research Question 2, again a few limitations need to be considered. The first limitation relates to the failing in finding five participants that participated in the original COUCH study, and the second limitation relates to the number of practical problems experienced in the field study.

Study Participants

The intention for the field study was to conduct it with five 'experienced' COUCH users that participated in the original study done by RRD, but only two people responded to the advertisement. Instead, two family members of the researcher were approached to participate, but they did not have experience with the COUCH application. To overcome this problem and be able to ask something about their preference version, they were asked to use the text-based COUCH application once, before starting with the speech-based application. It may be not that big issue because the field study is designed to test the robustness of such speech-based system in a home-setting, but it would be preferred to have a more homogeneous group of participants.

Practical Problems of the Field Study

As described in Section 6.2.2, there were some issues during the field study. Due to the Covid crisis, we tried to avoid any contact with the older participants and therefore participants were asked to use their own tablet for the experiment. The application was tested on a new Android tablet, but not on an I-Pad, where the application turned out not to run. A replacement was provided to two participants who used an I-Pad, but for one participant the tablet did not work either because it was too old to run any of both versions smoothly. Moreover, these participants tried to use the application on the day the server and application were down. All these issues might have influenced the user experience of these two participants.

7.4 Main Question

7.4.1 Answering the Main Question

The Main Research Question stated:

To what extent can spoken interaction offer a valuable addition to the multi-party virtual Council of Coaches application?

To answer our main research question it is important to distinguish between the difference in experience with the text- and speech-based version and the experience with the system as a whole. The controlled experiment addressed these differences between versions, while the field study focused on the experience as a whole. In general we can say that people preferred the speech-based application over the text-based application, even though we learned from the observational measurements that a lot of errors were made by the ASR in correctly recognizing speech, or recognizing speech at all. Additionally, using the application in a home-setting sometimes resulted in bad performance caused by bad quality microphones, old devices, or an unstable internet connection, decreasing user experience. Overall we expect spoken interactions to be a valuable addition to an e-health coaching application like the Council of Coaches.

7.4.2 Discussion of the Main Question

We have found evidence allowing us to give an affirmative answer to our original research questions. Therefore, our results do indicate that adding speech to e-health applications can contribute to a pleasant experience, but that a robust implementation and proper devices are important to assure this. Because the results were promising for a system like COUCH, it does not necessarily mean that this is true for all systems. In this section we discuss possible explanations for the results.

The structured dialogue setup of COUCH made the speech synthesis easier, allowing us to automatically create humanlike voices that were perceived relatively well by participants. Since the TTS is not prone to errors in the current application, only adding TTS as part of a spoken interaction already offered a valuable addition for such application. Solely implementing speech recognition without speech synthesis for the coaches would not fit an application like COUCH. Moreover, it is probably not much appreciated because of the large number of recognition errors in home situations. Besides facilitating the speech synthesis, the structured setup eased the speech recognition but thereby limited users in their responses. Users were not able to freely express themselves, which is not relatable to a human-to-human interaction. On the other hand, compared to the situation where users were limited with textual responses, the perception of a natural interaction increased a lot. Moreover, spoken interactions turned out to be more fun, interesting and interactive to use. These factors may increase the change for a long-term engagement with the coaches, allowing them to have further impact on the user's health behavior change, but therefore the system as a whole needs to become more interesting for all older adults.

In general, participants of the controlled experiment were more positive about the speech-based application than participants of the field experiment, which is probably caused by the fact that the application performed much better in the controlled experiment (although it differed per participant). Moreover participants only used the application for about 10 minutes in the controlled study, which might have caused the application to be perceived as novel and interesting. Preferences and attitudes in long-term interactions are likely to change, and novelty effects will wear out [61]. During the field study, where participants used the application for a longer period of time, the novelty effects might have worn out. None of these participants were interested in

using the application in the future, although it is important to consider that this observation was done for the application in general (with as main problem the content).

The robustness, efficiency and reliability of the system are far from perfect yet, but they can improve a lot with the development of technology, reducing the number of ASR-errors. Many technologies, such as better microphones with noise filtering already exist, but these are not very common yet to be implemented in devices for home use. Additionally, the application is not running on all devices, but this could be resolved by a better implementation. The biggest bottleneck of the current application was its speed and responsiveness. In many cases the ASR (almost) correctly transcribed the spoken speech, but because the response was very slow, participants started to repeat themselves. The text-version was already not very fast in responding, but by sending spoken audio to the ASR, waiting for a transcription from the ASR, sending text to the TTS, and waiting for this text to be transformed in audio, slowed down the application. This disadvantage was mentioned most, because it resulted in the feeling that users were not understood or listened to.

Closing Summary

From both this thesis and the preliminary literature research [1], we can conclude that spoken interactions indeed offer a valuable addition to a multi-party virtual Council of Coaches application, but that better technologies such as a better performing ASR, good quality microphones with noise filtering and a better, and especially faster implementation, would create a more robust application. However, even though people liked the speech-based application more, none of the participants would like to use an application like this in the future. Therefore the application needs to become more interesting in general, including more coaching content with a broader range of reply options. On the other hand, almost all participants would recommend the application to others, especially to specific user groups, such as older adults who feel lonely, have difficulties reading, are lower-educated and have little computer experience.

7.4.3 Comparison with Prior Research

This thesis contributes to existing literature in the insight that adding speech to an application like COUCH strongly increases the user experiences, even though much more errors are encountered in a speech-based system compared to a text-based system. Taking into account the original purpose of COUCH to engage users in humanlike interactions for a longer period of time, the findings that users perceived a vocal interaction as much more entertaining, personal, interactive and like a real conversation, is very important. These findings relate to the results of the study conducted for the original application, in where the text-based application scored low on the domains 'intention to use', 'recommend it to others', and 'enjoyment' [62]. Participants did not have the intention to use the system or recommend it to others and did not think the system was entertaining or exciting. Results obtained with the controlled experiment showed contrary results for the 'recommendation to others' and 'entertainment' field, where most participants recommended the speech-version to older adults and all participants experienced the speech-version to be more fun. We did not focus on 'intention to use' in the controlled experiment because the participants were generally not falling in the target group. On the other hand, the field experiment revealed similar results that participants did not have the intention to use the system, although they would recommend it to others. However, Hurmuz et al. [62] mentioned that participants expected to get more in depth and personal advises. "This gap between participants' expectations and the reality made participants probably less positive about the overall working of the system" [62]. It is important to notice that two of the participants for

the field experiment already participated in the original study [62] and therefore had different expectations for this follow-up study. This might have minimized this gap between expectations and reality.

Furthermore, this thesis contributes to existing literature by its findings regarding the speech synthesis. For the tightly structured COUCH application, the TTS with additional adjustments via SSML resulted in a positive experience of the voices and the ability of coaches to talk. Contrary, from Section 3 we learned that the voice of the KRISTINA agent was considered to be monotonous and that a major source of displeasure with the virtual meditation coach was the lacking humanlike synthesized voice [46, 48]. These issues contradict with the results of our study, in where the Google TTS voices were generally perceived as pleasant and clear to listen to. In this Related Work section we also discussed the exercise advisor from Bickmore et al. [42, 49]. The results of this research are in line with the results found in the two studies with the exercise advisor. Their initial research [49] stated that deploying conversational interfaces does not imply that natural language understanding must be used. This is in line with our finding that many participants appreciated the fact that coaches could talk and that only TTS already improved the naturalness of an interaction. However, in the exercise advisor’s followup research [42], participants mentioned that they could not express themselves completely using the constrained, multiple-choice interaction. This finding also resulted from our study, where participants could use their voice, but were still limited to the multiple-choice interaction. They preferred the option to draw their own answers and questions.

7.5 Recommendations

This section describes the future research directions and recommendations for the work described in this thesis. The section is divided in two parts. The first section (Section 7.5.1) provides research directions for future work. The second section (Section 7.5.2) provides some explicit suggestions for future work, based on the research directions.

7.5.1 Future Work

The current research showed the addition of speech to dialogues in an application like COUCH to be very valuable, but that users were very limited in their options to speak. Although users could deviate to some extent from the answer options, they experienced that the system only responded to the (almost) exact reply options. With the current implementation, this issue might be improved by manually adding large sets of keyword, but this is very time-consuming and inefficient. A recommendation for future work is to look into smart ways to automate the process of creating large keyword sets. Thereby research should be done in how dialogue authoring tools (e.g. the WOOL dialogue platform) can be improved to support this automatization and help construct better speech-based dialogue systems. These improvements allow users to use a larger vocabulary to navigate through the dialogues. Moreover, it can remove the written reply options, creating an interface that is more user friendly for older adults who more often deal with visual impairments. Because this thesis focused on the implementation of speech into an existing application, smart ways to adjust the dialogues were investigated. Apart from looking into smart ways to automate the process of keyword tagging, it might be of interest to investigate ways in which new dialogues can be created and tagged with keywords when a dialogue-based application is created from scratch.

From Chapter 1 we learned that keeping the user engaged for a longer period of time is one of the major challenges in e-coaching [17]. When the application does not offer a big variety in content, this can lead to interactions that become repetitive over a longer period of time. Our results suggested that speech-based interactions can positively influence long-term engagement because it showed to be much more innovative, entertaining and interesting to use. However, the novelty effect may disappear with time, negatively influencing long-term engagement. Future research should look into ways to overcome this problem with speech-based e-health applications.

The biggest bottleneck for a pleasant experience in our application was the slow responsiveness of the system. During the interaction it was often not clear for users whether or not they were understood by the system, only because it took a long time for the system to respond. This problem is probably caused by the combination of a slow ASR, as well as the speed of the COUCH application in general. Future work can test the speed of the NLSpraak ASR in a different application and investigate ways to improve the reaction speed of a speech-based interface.

Another suggestion for future work is to include more dialogue management strategies and more conversational feedback mechanisms. We already included the coaches' clarification questions and the automatic repetition after some time with no speech, but this did not cover all conversational errors. Especially the situation when there is no response from the system, although there is spoken input, can be improved. Now this only includes the clarification questions which are asked when no keyword and an unknown word is spotted, while it would improve the conversation a lot if feedback is also provided when there is spoken audio, but not (correctly) captured by the ASR.

A last interesting objective is to include free speech because this can make the interaction much more interesting and personal. Therefore it is necessary that the coaches have a larger vocabulary (manually or via automatic content generation) and can appropriately respond to different input. In this case the application can step of the text balloons and instead can provide better tailored coaching sessions, in where users are presented with personal advises based on subjects they lack knowledge about. For this research direction, more advanced ASRs (especially advancements in response speed), as well as in-home devices are necessary, as suggested by Bickmore [42]. Available systems need to ensure that they could provide high enough reliability given the variability and differences in voice quality in older adults. Free speech allows users to choose topics that are interesting or relevant for their personal situation and discuss those with the coaches. Additionally, free speech provides the option for older adults to tell stories and have small talk with conversational interfaces, which can help building relationships with coaches [44]. This might give the application an extra role as social companion for older adults suffering from loneliness.

7.5.2 Application Improvements

This section provides a list of suggestions that can be considered to improve the system developed for this thesis. Suggestions such as automatizing the keyword retrieval, improving the name recognition, dialogue management strategies and feedback mechanisms, but also some graphical user interface additions are provided. These ideas all differ in level of complexity, ranging from quite easy to very complex suggestions.

One simple suggestion is to follow the guidelines for designing voice user interfaces in their suggestion to make use of visual confirmation [53]. Since the text on the buttons is similar to what the user have to say, the system can simply highlight the button that is spoken. Then the user will know whether she has been correctly (or incorrectly) understood and that the continuation of the dialogue was right. Another option is to present the transcription of the speech

as one of the replies under the text balloons. This will not improve the recognition in general, but at least users know if and what the system captured from the sentence they spoke. It can also improve the way they speak to the system (and therefore reduce the error rate) because it allows users to see what words were misrecognized. This idea is similar to the functionality of VUIs in smartphones that show the text that is captured by the ASR on the screen. One last idea for the graphical user interface is to leave the interface 2D, but make the cartoonish characters animated, like characters in animated movies. In this way it is obvious for the user to see which coach is talking and it might improve the entertainment level of the application.

Another easy suggestion to make the application also suitable for visually impaired older adults, without changing the complete structure of the original application, is to make replying to the statements also voice directed. Instead of reading the reply options and speaking one of the options, the coach can ask the user what he wants to do or respond and consequently click the corresponding reply.

A more complex suggestion for improving the application is to automatically retrieve synonyms of the keywords. The WOOL dialogue platform makes it possible for health experts to write dialogues, without having any technical programming skills. Since these experts are responsible for the content of the dialogue, it would be helpful when they could decide about the keywords too. In this way they can write statements with multiple reply options, all containing a different main word (verb or noun), and add one distinctive keyword for each of these options. Based on this keyword, a complete list of synonyms can be retrieved, for example via the Open Dutch Wordnet¹, and shown as suggestion to the writer. This Corpus contains almost almost 118.000 Dutch synsets, which are in fact, conceptual representations of a word. For example the word 'dog' includes synsets like [dog, animal, pet]. In this way, the synset for a word can be retrieved from the Corpus and proposed as additional keywords, and dialogue writers can decide themselves which keywords to add by clicking the relevant proposed synonyms. An example of such a suggestion is provided in Figure 7.1. When a reasonable number of keywords all lead to a specific reply, this allows users to engage in richer conversations and more freely express themselves while maintaining the ease of use of the multiple-choice selection input modality. Considering the recommendations of Bickmore et al., even when taking such an approach it is important that the ASR ensures a high reliability, given the variability and differences in voice quality in older adults.

Speaker: Coda
Title: Start

```

1 Hi $userFirstName, how can I help you today?
2 [[Can you explain the interface again?|coda-explain-interface.Start|
3
4 [[About the Corona Virus.|Corona|
5
6 [[I just wanted to chat.|coda-social-menu.Start|
7
8 [[I want to log out.|LogoutConfirm|
9
10
11
12
13 [[Can you explain the interface again?|coda-explain-interface.Start|
14 <<action type="generic" value="ADD_SPECIFIC_KEYWORDS" keywordlist = "interface, user environment, application">>]]
15
16 [[About the Corona Virus.|Corona| <<action type="generic" value="ADD_SPECIFIC_KEYWORDS" keywordlist = "corona, virus, covid">>]]
17
18 [[I just wanted to chat.|coda-social-menu.Start| <<action type="generic" value="ADD_SPECIFIC_KEYWORDS" keywordlist = "talk, chat, interact">>]]
19
20 [[I want to log out.|LogoutConfirm| <<action type="generic" value="ADD_SPECIFIC_KEYWORDS" keywordlist = "log out, leave, sign off">>]]
21

```

Interface	User environment	Appearance	Application	Implementation
Epidemic	Virus	Corona	Covid	Crisis
Talk	Converse	Chat	Interact	Gossip
Log out	Leave	Quit	Sign off	Turn off

Figure 7.1: An example of how automatically generated keywords can be presented to the dialogue author.

¹<http://wordpress.let.vupr.nl/odwn/data/>

Because the recognition of the coach names performed very bad during the experiments, improving this recognition will reduce the number of errors made by the system. One very simple option is to include similar keywords for transcriptions that are often created, which we already included for Helen (i.e. 'helen', 'alan', 'hellen', 'ellen'), but not for the other coaches. Moreover, for Helen these four extra keywords were not enough. Words that are similar to the names can be added as keyword by checking what words the ASR often recognizes. A less manual implementation is to automatically compare a recognized word with the coach name. The edit distance can help to calculate how similar two strings are. "The minimum edit distance between two strings is defined as the minimum number of editing operations (operations like insertion, deletion, substitution) needed to transform one string into another" [63]. For example, the ASR transcribed the string '*gouda*' while the user spoke '*coda*'. The gap between these two words is 2 (remove an u, substitute g for c). This similarity measure can calculate how similar the spoken word is from one of the coach keywords and for example, by setting a maximum distance, a spoken word that shows less operations then this maximum can be interpreted as the connected coach keyword.

In the current application the NLSpraak ASR was used, which was a ready-made ASR. This caused the ASR to be a black box that recorded the spoken audio and returned a transcription of written text, but could not be improved for the purpose of the study. When the ASR is fully controlled, machine learning techniques can be used to train the ASR, for example by training the ASR for older adults voices, or by teaching it to retrieve new keywords from words that are often spoken in a specific context. Also, most ASRs works best for English because this language has most data. The performance of the ASR might be improved by training it with larger Dutch corpora.

From the controlled experiment it turned out that participants really enjoyed the radio feature implemented in the text-version, but removed from the speech-version. For future work it might be interesting to use speech filtering techniques to mitigate noise effects, so the radio can be turned on, without triggering the ASR component. Additionally, this filtering techniques allow participants to start the application without explicitly clicking the 'start speech', but instead trigger the application by speaking a specific sentence (e.g. the Google Home starts listening to the user when 'Hey Google' is spoken). This can be beneficial again for people experiencing difficulties with reading and clicking text balloons.

Chapter 8

Conclusion

This thesis aimed to explore the possibility of implementing speech to the Council of Coaches application. The intention was to create a speech-based application that can maintain a spoken conversation with a user by creating an engaging and humanlike experience. The application required the following implementations: it should be able to (1) recognize user input, (2) correctly transcribe the input to written text, (3) check the transcription for specified keywords, (4) correctly trigger the adequate reply to continue in the dialogue, and (5) transform the new dialogue content into output audio.

Even though the system made a significant number of recognition errors in some cases, many participants of the controlled experiment found the interaction fun, interesting and interactive, and preferred the speech-based application over the original text-based application. On the other hand, the field experiment resulted in a less enjoyable interaction, partly caused by the content of the application. Besides, the field experiment showed that speech technologies were not robust enough yet in order to be effectively used in a home-setting where participants used tablets with worse performing microphones.

The literature research preliminary to this thesis, answered five sub-questions, mostly related to the implementation of ASR systems. Since the focus of this thesis was to allow for speech interactions into both directions (i.e. user to coach, and coach to user), more research in the TTS was required. Therefore, the current research answered two additional sub-questions focusing on the state-of-art in TTS. Because speech was implemented in an existing application that used a fixed and self-contained dialogue structure, sub-questions about smart ways to handle the dialogues were added, as well as additional graphical user interface features. The sub-questions were for the most part answered before the system implementation, although the experimental studies provided additional insights.

To answer the research's main question *"To what extent can spoken interaction offer a valuable addition to the multi-party virtual Council of Coaches application?"*, two experiments are conducted with each answering a different research question.

RQ:1 The addition of speech to the Council of Coaches application leads to an increase in user experiences. Statistical analyses of the controlled experiment showed that participants rated the speech-version to be significantly more novel (items: valuable, exciting, interesting and motivating) and stimulating (items: creative, inventive, leading edge, innovative). Additionally, results showed the speech-version to be more interesting, credible, fun to use and less monotonous than the text-version and that participants would use the speech-version for a longer period of time, more often follow advice from it, sooner recommend it to older adults and rather spend money on the speech-version. Although the implementation of speech in the COUCH application did lead to a higher user evaluations for certain aspects,

a significant number of errors is made by the system, indicating that the technologies and implementation are far from perfect yet. The main problem of the speech-based system was its lack of responsiveness.

RQ:2 The current state of art in speech recognition systems and commercial technologies (i.e. devices with proper microphones), are not robust enough to create a usable and enjoyable system that can be used in a in-home setting. The field study showed large differences between participants in the amount of problems experienced. For some participants the system worked well without many errors, while other participants were hardly understood. In some cases it was dependent on the moment how well the speech was recognized. Although the speech recognition turned out not to be very robust, it still created system that was more enjoyable then the original text-version. In particular, the combination of written text and spoken speech was well received, because this improved the efficiency of the interaction in case the ASR left much to be desired. None of the users would like to use such an application in the future, but it can be useful for specific user groups.

Considering the finding that users appreciated the possibility for speech interactions much more then textual interactions, even though the application was far from robust, adding speech to such e-health applications offer good future prospects. With the rapid advancements in technology, it can be expected that in a few years commercial devices are developed enough to work properly in a home-settings. On the other hand, concepts like free speech for communication with devices is much more complex, though still very relevant for such e-health applications.

The next step for the speech-based COUCH application would be to improve the keyword retrieval through automatic keyword generation, which can decrease the workload of dialogue authors. Automatic keyword generation can be done by, for example, synonym retrieval from a Dutch corpus. The recognition of names can be improved by implementing a similarity measure that can determine if, and what name is spoken. The naturalness and fluency of the conversation can be improved by including more dialogue management strategies. Overall, we believe that speech can positively contribute to such e-health applications, even though the technologies are far from perfect.

Bibliography

- [1] Laura Bosdriesz, Dennis Reidsma, Daniel Patrick Davison, and Harm op den Akker. Opportunities and challenges for adding speech to dialogues with a council of coaches. 2020.
- [2] European Union. *Ageing Europe*. 2019.
- [3] Jorunn L. Helbostad, Beatrix Vereijken, Clemens Becker, Christop Todd, Kristin Taraldsen, Mirjam Pijnappels, Kamiar Aminian, and Sabato Mellone. Mobile health applications to promote active and healthy ageing. *Sensors (Switzerland)*, 17(3):1–13, 2017.
- [4] Christopher J.L. Murray, Theo Vos, Rafael Lozano, Mohsen Naghavi, Abraham D. Flaxman, Catherine Michaud, and Et Al. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, 380(9859):2197–2223, 2012.
- [5] Efraim Jaul and Jeremy Barron. Age-Related Diseases and Clinical and Public Health Implications for the 85 Years Old and Over Population. *Frontiers in Public Health*, 5(December):1–7, 2017.
- [6] Grigorios Karageorgos, Ioannis Andreadis, Konstantinos Psychas, George Mourkousis, Asimina Kiourti, Gianluca Lazzi, and Konstantina S. Nikita. The Promise of Mobile Technologies for the Health Care System in the Developing World: A Systematic Review. *IEEE Reviews in Biomedical Engineering*, 12(c):100–122, 2018.
- [7] Predrag Klasnja and Wanda Pratt. Healthcare in the pocket: Mapping the space of mobile-phone health interventions. *Journal of Biomedical Informatics*, 45(1):184–198, 2012.
- [8] Huan Li, Qi Zhang, and Kejie Lu. Integrating mobile sensing and social network for personalized health-care application. *Proceedings of the ACM Symposium on Applied Computing*, 13-17-April:527–534, 2015.
- [9] Michel C.A. Klein, Adnan Manzoor, Anouk Middelweerd, Julia S. Mollee, and Saskia J. Te Velde. Encouraging Physical Activity via a Personalized Mobile System. *IEEE Internet Computing*, 19(4):20–27, 2015.
- [10] Gerwin Huizing, Randy Klaassen, and Dirk Heylen. Designing and developing lifelike, engaging lifestyle coaching agents and scenarios for multiparty coaching interaction. *CEUR Workshop Proceedings*, 2338:25–29, 2018.
- [11] H. Hermens, H. op den Akker, M. Tabak, J. Wijsman, and M. Vollenbroek. Personalized Coaching Systems to support healthy behavior in people with chronic conditions. *Journal of Electromyography and Kinesiology*, 24(6):815–826, 2014.
- [12] Harm op den Akker, Valerie M. Jones, and Hermie J. Hermens. Tailoring real-time physical activity coaching systems: a literature survey and model. *User Modeling and User-Adapted Interaction*, 24(5):351–392, 2014.

- [13] Randy Klaassen, Rieks op den Akker, Pierpaolo Di Bitonto, Gert Jan Burger, Kim Bul, and Pam Kato. PERGAMON : A serious gaming and digital coaching platform supporting patients and healthcare professionals. *International Conference on ENTERprise Information Systems/International Conference on Project MANagement/International Conference on Health and Social Care Information Systems and Technologies, CEN-TERIS/ProjMAN/HCist 2016; Book of industry papers and abs*, (0):1–6, 2016.
- [14] Reshmashree B. Kantharaju, Alison Pease, Dominic De Franco, and Catherine Pelachaud. Is two beter than one? Effects of multiple agents on user persuasion. *Proceedings of the 18th International Conference on Intelligent Virtual Agents, IVA 2018*, pages 255–262, 2018.
- [15] Harm op den Akker, Rieks op den Akker, Tessa Beinema, Oresti Banos, Dirk Heylen, Björn Bedsted, Alison Pease, Catherine Pelachaud, Vicente Traver Salcedo, Sofoklis Kyriazakos, and Hermie Hermens. Council of coaches a novel holistic behavior change coaching approach. *ICT4AWE 2018 - Proceedings of the 4th International Conference on Information and Communication Technologies for Ageing Well and e-Health*, 2018-March(Ict4awe 2018):219–226, 2018.
- [16] Dennis Reidsma, Gerwin Huizing, Randy Klaassen, Daniel Davison, Kostas Konsolakis, Marcel Weusthof, Oresti Baños, Jorien van Loon, Merijn Bruijnes, Tessa Beinema, Silke ter Stal, Dennis Hof, and Donatella Simonetti. D7 . 5 : Final Council of Coaches Technical Prototype. Technical report, 2019.
- [17] Markku Turunen, Jaakko Hakulinen, Olov Ståhl, Björn Gambäck, Preben Hansen, Mari C. Rodríguez Gancedo, Raúl Santos De La Cámara, Cameron Smith, Daniel Charlton, and Marc Cavazza. Multimodal and mobile conversational Health and Fitness Companions. *Computer Speech and Language*, 25(2):192–209, 2011.
- [18] Mary Ellen Foster. Enhancing human-computer interaction with embodied conversational agents. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4555 LNCS(PART 2):828–837, 2007.
- [19] Christina R. Victor and Keming Yang. The prevalence of loneliness among adults: A case study of the United Kingdom. *Journal of Psychology: Interdisciplinary and Applied*, 146(1-2):85–104, 2012.
- [20] Matthew B. Hoy. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services Quarterly*, 37(1):81–88, 2018.
- [21] Heidi Horstmann Koester. Usage, performance, and satisfaction outcomes for experienced users of automatic speech recognition. *Journal of Rehabilitation Research and Development*, 41(5):739–754, 2004.
- [22] Thomas Pellegrini, Isabel Trancoso, Annika Hämäläinen, António Calado, Miguel Sales Dias, and Daniela Braga. Impact of age in ASR for the elderly: Preliminary experiments in European Portuguese. *Communications in Computer and Information Science*, 328 CCIS:139–147, 2012.
- [23] Michael McTear, David Griol, and Zoraida Callejas. *The Conversational Interface*. Springer, 2016.
- [24] Veton Kepuska and Gamal Bohouta. Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). *2018 IEEE 8th Annual Computing and Communication Workshop and Conference, CCWC 2018*, 2018-Janua(c):99–103, 2018.

- [25] Raquel Fernández. *Oxford Handbooks Online Dialogue*. Number August. 2015.
- [26] Herbert H. Clark and Edward F. Schaefer. Contributing to discourse. *Cognitive Science*, 13(2):259–294, 1989.
- [27] Victor W. Zue Glass and James R. Conversational interfaces: advances and challenges. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, 88(8):1166–1180, 2000.
- [28] Victoria Young and Alex Mihailidis. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive Technology*, 22(2):99–112, 2010.
- [29] Roeland van der Werff, Laurens Ordeman. Kaldi-NL scoort uitstekend op NBest evaluatie, 2016.
- [30] Daniel Povey, Arnab Ghoshal, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, Jan Silovsk, and Petr Motlicek. The Kaldi Speech Recognition Toolkit. *IEEE Signal Processing Society*, 2011.
- [31] Yao Qian, Rutuja Ubale, Patrick Lange, Keelan Evanini, Vikram Ramanarayanan, and Frank K. Soong. Spoken Language Understanding of Human-Machine Conversations for Language Learning Applications. *Journal of Signal Processing Systems*, 2019.
- [32] Jan Gerrit Harms, Pavel Kucherbaev, Alessandro Bozzon, and Geert Jan Houben. Approaches for dialog management in conversational agents. *IEEE Internet Computing*, 23(2):13–22, 2019.
- [33] Sadhana Gopal, Trishant Malik, and Seema Devi. A simple phoneme based speech recognition system. *International Journal of Modern Communication Technologies & Research*, 2(4), 2014.
- [34] Karolina Kuligowska, Pawel Kisielewicz, and Aleksandra Włodarz. Speech synthesis systems: Disadvantages and limitations. *International Journal of Engineering and Technology(UAE)*, 7(2):234–239, 2018.
- [35] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10-11):763–786, 2007.
- [36] Robert Jim Firby and Peter Graff. Environmental noise detection for dialog systems. 2017.
- [37] Ravichander Vipplerla, Steve Renals, and Joe Frankel. Longitudinal study of ASR performance on ageing voices. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2550–2553, 2008.
- [38] Nelson Roy, Joseph Stemple, Ray M. Merrill, and Lisa Thomas. Epidemiology of voice disorders in the elderly: Preliminary findings. *Laryngoscope*, 117(4):628–633, 2007.
- [39] Hee Rin Lee, Selma Šabanović, and Erik Stolterman. How Humanlike Should a Social Robot Be : A User-Centered Exploration. pages 135–141, 2016.
- [40] Stephen Anderson, Natalie Liberman, Erica Bernstein, Stephen Foster, Erin Cate, Brenda Levin, and Randy Hudson. Recognition of elderly speech and voice-driven document retrieval. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 1:145–148, 1999.

- [41] J T Luo, Peter Mcgoldrick, Susan Beatty, and Kathleen A Keeling. On-screen characters : their design and influence on consumer trust. *Journal of Services Marketing*, 2:112–124, 2006.
- [42] Timothy W. Bickmore, Lisa Caruso, Kerri Clough-Gorr, and Tim Heeren. 'It's just like you talk to a friend' relational agents for older adults. *Interacting with Computers*, 17(6):711–735, 2005.
- [43] Elayne Ruane, Abeba Birhane, and Anthony Ventresque. Conversational AI: Social and ethical considerations. *CEUR Workshop Proceedings*, 2563(December 2019):104–115, 2019.
- [44] Laura Pfeifer Vardoulakis, Lazlo Ring, Barbara Barry, Candace L. Sidner, and Timothy Bickmore. Designing relational agents as long term social companions for older adults. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7502 LNAI:289–302, 2012.
- [45] Leo Wanner, Elisabeth André, Josep Blat, Stamatia Dasiopoulou, Mireia Farrús, Thiago Fraga, Eleni Kamateri, Florian Lingenfelser, Gerard Llorach, Oriol Martínez, Georgios Meditskos, Simon Mille, Wolfgang Minker, Louisa Pragst, Dominik Schiller, Andries Stam, Ludo Stellingwerff, Federico Sukno, Bianca Vieru, and Stefanos Vrochidis. Kristina: A knowledge-based virtual conversation agent. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10349 LNCS:284–295, 2017.
- [46] Leo Wanner, Anna Ehmann, Marlen Brachthäuser, Gerhard Eschweiler, Louisa Pragst, Oriol Martínez, Dominik Schiller, Florian Lingenfelser, Bianca Vieru Mireia, Mireia Farrús, Simon Mille, Mónica Domínguez Josep Blat, Hermann Plass, Georgios Meditskos Benjamin, and Benjamin Schäfer. Final System Report. Technical report, 2018.
- [47] Ameneh Shamekhi and Timothy Bickmore. Breathe with me: A virtual meditation coach. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9238(Figure 1):279–282, 2015.
- [48] Ameneh Shamekhi and Timothy Bickmore. Breathe deep: A breath-sensitive interactive meditation coach. *ACM International Conference Proceeding Series*, (May):108–117, 2018.
- [49] Timothy W Bickmore and Rosalind W Picard. Establishing and Maintaining Long-Term Human- Computer Relationships. *ACM Transactions on Computer-Human Interaction*, 12(2):617–638, 2005.
- [50] Tessa Beinema, Harm op den Akker, Stephanie Kosterink, Silke ter Stal, and Janet van den Boer. D3 . 4 : Final coaching actions and content. Technical report, 2019.
- [51] Gerwin Huizing, Randy Klaassen, Reshmashree Bangalore Kantharaju, Silke ter Stal, Tessa Beinema, Harm op den Akker, and Merijn Bruijnes. D6 . 2 : Initial user interface design for Home UI and Mobile UI. Technical report, 2018.
- [52] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. pages 1–15, 2016.
- [53] Cathy Pearl. *Designing Voice User Interfaces*. O'Reilly, 2016.
- [54] Martin Schrepp. User Experience Questionnaire Handbook. pages 1–15, 2019.
- [55] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski Hochschule. Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. 8517(03), 2014.

- [56] Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and evaluation of a user experience questionnaire. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5298 LNCS:63–76, 2008.
- [57] J K Hendrix. *An Embodied Conversational Agent in a Mobile Health Coaching Application*. PhD thesis, 2013.
- [58] Marian Z.M. Hurmuz, Stephanie M. Jansen-Kosterink, Harm op den Akker, and Hermie J. Hermens. D7.6: Demonstration Protocol, Ethical. Technical report, 2020.
- [59] Darren George and Paul Mallery. *SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 Update*. Allyn and Bacon, 4th edition, 2003.
- [60] Wolfgang Forstmeier, Eric-jan Wagenmakers, and Timothy H Parker. Detecting and avoiding likely false-positive findings – a practical guide. 1954:1941–1968, 2017.
- [61] Kerstin Dautenhahn. Methodology & themes of human-robot interaction: A growing research field. *International Journal of Advanced Robotic Systems*, 4(1 SPEC. ISS.):103–108, 2007.
- [62] Marian Hurmuz, Tessa Beinema, Stephanie Jansen- Kosterink, Dominic De Franco, Silke ter Stal, and Harm op den Akker. D7 . 7 : Final Demonstration Results. Technical report, 2020.
- [63] Daniel Jurafsky and James H. Martin. Regular Expressions, Text Normalization, Edit Distance. In *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2019.

Appendix A

Questionnaires and interviews

A.1 Controlled Experiment

A.1.1 Intake Questionnaire

Algemene achtergrond

Fijn dat u wilt deelnemen aan dit onderzoek! Hieronder staan een aantal vragen met betrekking tot uw achtergrond. Vult u alstublieft deze vragen in zodat wij deze relevante informatie over u te weten kunnen komen.

Wat is u geslacht?

☐ Man

☐ Vrouw

☐ Zeg ik liever niet

Wat is uw leeftijd?

Jouw antwoord _____

Wat is uw hoogst genoten opleiding (met of zonder diploma)?

☐ Basisonderwijs

☐ Voortgezet onderwijs

☐ MBO

☐ HBO

☐ Universiteit

☐ Zeg ik liever niet

Welke van de volgende categorieën beschrijft het beste uw werkstatus?

- ☐ Student
- ☐ In loondienst
- ☐ Vrijwilliger/mantelzorger
- ☐ Met pensioen
- ☐ Overig
- ☐ Zeg ik liever niet

Heeft u ervaring met spraaktechnologieën ? (Denk aan: het gebruik van de spraakfunctie op een smartphone, Google Home, Siri, etc.)

- ☐ Ja
- ☐ Nee
- ☐ Gemiddelde

Verzenden

A.1.2 Observational Measurements

OBSERVATIONAL MEASUREMENTS

	Dialogue steps		Errors				
	Chosen path	Count of (normal) steps	User has to repeat/no reaction from the system	User said something that was not understandable	System is repeating itself	User clicks to the next dialogue step	Repeating questions
001 Tekst							
002 Spraak							
003 Tekst							
004 Spraak							
005 Tekst							

A.1.3 User experience

Gebruikerservaring

Maak dan nu uw evaluatie. Voor de beoordeling van het systeem vragen we u de onderstaande vragenlijst in te vullen. De vragenlijst bestaat uit twee tegengestelde eigenschappen die van toepassing zijn op het product. De rondjes staan voor verschillende gradaties. U kunt uw beoordeling geven door het rondje, die het meest uw indruk weerspiegelt, aan te vinken.

Graag uw eerste ingeving invullen. Wacht niet te lang met invullen om te voorkomen dat u gaat twijfelen over uw eerste ingeving. Soms bent u misschien niet helemaal zeker van uw antwoord of u vindt de eigenschap niet volledig van toepassing, kruis dan toch een rondje aan.

Het is uw mening die telt. Let op: er is geen goed of fout antwoord!

*Vereist

	1	2	3	4	5	6	7		
onbegrijpelijk	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	begrijpelijk	1
creatief	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	saai	2
makkelijk te leren	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	moeilijk te leren	3
waardevol	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	inferieur	4
vervelend	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	spannend	5
oninteressant	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	interessant	6
onvoorspelbaar	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	voorspelbaar	7
snel	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	langzaam	8
origineel	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	conventioneel	9
belemmerend	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	ondersteunend	10
complex	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	eenvoudig	11
gebruikelijk	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	nieuw	12
vertrouwd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	niet vertrouwd	13
motiverend	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	demotiverend	14
volgens verwachtingen	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	niet volgens verwachtingen	15
inefficiënt	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	efficiënt	16
overzichtelijk	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	verwarrend	17
onpraktisch	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	praktisch	18
ordelijk	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	rommelig	19
conservatief	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	innovatief	20

A.1.4 Explicit Comparison

Vergelijking COUCH versies

Gedurende het onderzoek heeft u twee verschillende versies van de COUCH applicatie gebruikt: één waarbij de conversatie plaatsvond door middel van spraak (de 'Spraakversie'), en één waarbij de conversatie gedaan werd door het aanklikken van de tekstballonnen (de 'Tekstversie'). In de volgende set vragen is het de bedoeling dat u een expliciete vergelijking maakt tussen deze twee softwareversies, door middel van uw ervaringen met beide systemen.

**Vereist*

Geef hieronder aan welke van de twee versies van het systeem u...

	Tekstversie		Geen verschil		Spraakversie
prettiger vond in gebruik	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
efficiënter vond werken	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
interessanter vond	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
geloofwaardiger vond	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
leuker vond	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
omslachtiger in	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
eentonig vond	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
liever voor een langere periode zou gebruiken	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
zou aanraden aan iemand met een leeftijd boven de 65	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
irriteranter vond	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
eerder adviezen van zou opvolgen	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
eerder geld aan uit zou geven	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
onhandiger vond	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
praktischer vond	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
repetitiever vond	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Welke versie gaat uw voorkeur naar uit en waarom? *

Jouw antwoord

Welke versie vond u het makkelijkst om te gebruiken en waarom? *

Jouw antwoord

Welke versie vond u het leukst om te gebruiken en waarom?

Jouw antwoord

Wat heeft u als voordelen ervaren in de tekst versie? *

Jouw antwoord

Wat heeft u als voordelen ervaren in de spraak versie? *

Jouw antwoord

Wat heeft u als nadelen ervaren in de tekst versie? *

Jouw antwoord

Wat heeft u als nadelen ervaren in de spraak versie? *

Jouw antwoord

Welke versie zou u aanraden aan iemand ouder dan 55?

Jouw antwoord

Verzenden

A.2 Field Experiment

A.2.1 Intake Questionnaire

Same questionnaire as Appendix A.1.1

A.2.2 Interview questions

1. Hoe vond u het om het systeem te interacteren, en dan niet wat betreft de inhoud van de coaches, maar gewoon heel algemeen de werking van het systeem?

Doelvragen: voorbeelden, hoe zou het beter kunnen volgens u?

Antwoorden zoals: makkelijk/moeilijk, leuk/stom

2. Zou u deze spraak-gebaseerde versie van de Council of Coaches aanraden aan anderen?

Ja: Waarom wel? Kunt u een voorbeeld geven?

Nee: Waarom niet? Kunt u een voorbeeld geven?

Heeft u nog andere voor- of nadelen ervaren?

3. Heeft u veel problemen ervaren met het gebruik van het systeem?

Antwoorden zoals: problemen dat het systeem mij niet verstond en niet doorging in het gesprek / systeem niet luisterden naar coachnaam / niet duidelijk hoe de applicatie bediend zou worden / systeem werkte te traag.

4. Op wat voor manier praatte u met de coaches?

Antwoorden zoals: ik zei letterlijk wat op de tekstballonnen stond / ik hield ongeveer aan wat op de tekstballonnen stond / ik vertelde graag alles wat in mijn hoofd omging.

5. Had u het idee dat u begrepen werd door de coaches?

Ja: Waarom wel? Kunt u een voorbeeld geven?

Nee: Waarom niet? Kunt u een voorbeeld geven?

6. Wat zou u graag willen veranderen aan de applicatie?

7. Zou u de applicatie zelf in de toekomst ook willen gebruiken?

Waarom wel/niet

Doelvragen: Hoe had de applicatie het nog leuker kunnen voor u om te blijven gebruiken?

8. Welke versie vond u prettiger om te gebruiken, de oorspronkelijk tekst gebaseerde versie of deze spraak gebaseerde versie?

Tekst: Waarom? Redenen dat spraak slecht werkt (traag, luistert niet goed, ingewikkeld, systeem doet niet wat ik verwacht)

Spraak: Waarom? Redenen dat tekst slecht werkt (veel lezen, saai)

Geen voorkeur: Waarom?

9. Er waren een aantal "conversation mechanisms" in de applicatie geïmplementeerd. Bijvoorbeeld: na 20 seconde in eenzelfde dialoogstap wordt die gesproken stap herhaald. En als er geen keyword is gespot en een onverstaanbaar woord was gesproken in de tekst, vroeg de coach aan de gebruiker om het te herhalen. Heb je zo een situatie ooit meegeemaakt?

Wat vond je hiervan?

Had je het idee dat je hierdoor beter/slechter/even goed begrepen werd door het systeem?

Appendix B

Forms and information

B.1 Controlled experiment

B.1.1 Information brochure

UNIVERSITEIT TWENTE.

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE

Informatieblad

INFORMATIEBLAD VOOR ONDERZOEK

‘Het Toevoegen van Spraak aan de Council of Coaches Applicatie’

Doel van het onderzoek

Dit onderzoek wordt geleid door Laura Bosdriesz. Dit onderzoek is onderdeel van een afstudeerproject waarbij een op spraak-gebaseerde versie van de Council of Coaches applicatie is gemaakt. Het doel van dit onderzoek is om een vergelijking te maken tussen het oorspronkelijke tekst-gebaseerde systeem en het nieuwe spraak-gebaseerde systeem. Hierbij wordt ook de robuustheid van het de spraakherkenning in het systeem getest.

Hoe gaan we te werk?

Het onderzoek vindt plaats op de universiteit, in een projectruimte in Carré of Design Lab. Hier wordt je welkom geheten en wordt er begonnen met een introductiegesprek en uitleg van het experiment. Gedurende dit gesprek en het gehele experiment worden de RIVM-regels in acht genomen. Tijdens het introductiegesprek worden de twee versies de Council of Coaches uitgelegd en wordt u gevraagd om twee vragenlijsten in te vullen met betrekking tot u geslacht, leeftijd, opleiding en werk, en met betrekking tot u ervaring met technologieën in te vullen. Deze vragenlijsten kunnen via een Google form worden ingevuld. Nadat u de vragenlijsten heeft ingevuld, krijgt u een document met de stappen die u moet doen met de applicaties. Wanneer u dat gelezen heeft kunt u beginnen met de eerste applicatie. Er is geen goed of fout in het experiment en het maakt niet uit als de beschreven stappen niet allemaal lukken. Als u klaar bent met het eerste systeem, dan kan de bijbehorende vragenlijst worden ingevuld. Hierna wordt het tweede systeem opgestart en wordt u gevraagd de al beschreven stappen te herhalen. Dit wordt vervolgens weer afgesloten met eenzelfde vragenlijst als voor het eerste systeem. Het onderzoek eindigt met een vergelijkende vragenlijst en een paar open vragen. Alle vragenlijsten kunnen online via een google form worden ingevuld.

Potentiële risico's en ongemakken

Er zijn geen fysieke, juridische of economische risico's verbonden aan uw deelname aan deze studie. U hoeft geen vragen te beantwoorden die u niet wilt beantwoorden. Uw deelname is vrijwillig en u kunt uw deelname op elk gewenst moment stoppen.

Vertrouwelijkheid van gegevens

Wij doen er alles aan uw privacy zo goed mogelijk te beschermen. Er wordt op geen enkele wijze vertrouwelijke informatie of persoonsgegevens van of over u naar buiten gebracht, waardoor iemand u zal kunnen herkennen. Voordat onze onderzoeksgegevens naar buiten gebracht worden, worden uw gegevens zoveel mogelijk geanonimiseerd.

UNIVERSITEIT TWENTE.

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE

Informatieblad

Uitsluitend ten behoeve van het onderzoek zullen de verzamelde onderzoeksgegevens worden gedeeld met bevoegde personen van de Universiteit Twente en Roessingh Research & Development.

De onderzoeksgegevens worden bewaard voor een periode van 10 jaar. Uiterlijk na het verstrijken van deze termijn zullen de gegevens worden verwijderd. De onderzoeksgegevens worden indien nodig (bijvoorbeeld voor een controle op wetenschappelijke integriteit) en alleen in anonieme vorm ter beschikking gesteld aan personen buiten de onderzoeksgroep.

Vrijwilligheid

Deelname aan dit onderzoek is geheel vrijwillig. U kunt als deelnemer uw medewerking aan het onderzoek te allen tijde stoppen, of weigeren dat uw gegevens voor het onderzoek mogen worden gebruikt, zonder opgaaf van redenen. Het stopzetten van deelname heeft geen nadelige gevolgen voor u of de eventueel reeds ontvangen vergoeding. Als u tijdens het onderzoek besluit om uw medewerking te staken, zullen de gegevens die u reeds hebt verstrekt tot het moment van intrekking van de toestemming in het onderzoek gebruikt worden. Wilt u stoppen met het onderzoek, of heeft u vragen en/of klachten? Neem dan contact op met de onderzoeksleider. Tot 24 uur na het onderzoek kunt u nog weigeren dat de data gebruikt wordt.

Contact informatie

Dennis Reidsma (afstudeerbegeleider)
of
Laura Bosdriesz (student)

Dit onderzoek wordt uitgevoerd vanuit de Universiteit Twente, faculteit Electrical Engineering Mathematics and Computer Science. Voor bezwaren met betrekking tot de opzet en of uitvoering van het onderzoek kunt u zich ook wenden tot de Secretaris van de Ethische Commissie van de faculteit Electrical Engineering Mathematics and Computer Science op de Universiteit Twente:

Contact Information

Ethics Committee, Faculty of EEMCS,
University of Twente
PO Box 217
7500 AE Enschede (NL)
Tel: +31(0)53.4892085
Email: ethics-comm-ewi@utwente.nl
Website: <https://www.utwente.nl/en/eemcs/research/ethics/>

Indien u specifieke vragen hebt over de omgang met persoonsgegevens kun u deze ook richten aan de Functionaris Gegevensbescherming van de UT door een mail te sturen naar dpo@utwente.nl.

Tot slot heeft u het recht een verzoek tot inzage, wijziging, verwijdering of aanpassing van uw gegevens te doen bij de Onderzoeksleider.

B.1.2 Informed consent

UNIVERSITEIT TWENTE.

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE

Toestemmingsverklaring Onderzoek

TOESTEMMINGSVERKLARING ONDERZOEK

Door dit toestemmingsformulier te ondertekenen erken ik het volgende:

1. Ik ben voldoende geïnformeerd over het onderzoek door middel van een separaat informatieblad. Ik heb het informatieblad gelezen en heb daarna de mogelijkheid gehad vragen te kunnen stellen. Deze vragen zijn voldoende beantwoord.
2. Ik neem vrijwillig deel aan dit onderzoek. Er is geen expliciete of impliciete dwang voor mij om aan dit onderzoek deel te nemen. Het is mij duidelijk dat ik deelname aan het onderzoek op elk moment, zonder opgaaf van reden, kan beëindigen. Ik hoef een vraag niet te beantwoorden als ik dat niet wil.
3. Ik geef toestemming dat bevoegde personen van de Universiteit Twente en Roessingh Research Development en bevoegde autoriteiten inzage kunnen krijgen in mijn onderzoeksgegevens. Ik geef toestemming om mijn gegevens te gebruiken voor de doelen die in het informatieblad genoemd zijn.

Naam proefpersoon:

Email-adres:

Datum:

Handtekening:

Ik verklaar hierbij dat ik bovenstaande proefpersoon volledig heb geïnformeerd over het genoemde onderzoek.

Naam onderzoeker: Laura Bosdriesz


Datum:

Handtekening:

B.1.3 Coaches Sheet

ALLES SELECTEREN VERDER ▶


 FRANÇOIS DUBOIS
Voeding
☒


 HELEN JONES
Cognitie
☒


 EMMA LI
Sociaal
☒


 CARLOS SILVA
Lotgenoot
☒


 OLIVIA SIMONS
Activiteit
☒


 RASMUS JOHANSEN
Chronische Pijn
☐


 KATARZYNA KOWALSKA
Diabetes
☐

! Carlos Silva

Beroep: Lotgenoot & Steun

Houdt van: Benfica, kaarten, tijd doorbrengen met zijn kleinkinderen

Houdt niet van: Mensen die hem zeggen wat hij moet doen

Carlos is geen coach. In feite luistert hij ook naar het advies van andere coaches. Hij worstelt ermee zijn leefstijl te verbeteren, en hij is hier om zijn ervaringen met u te delen.

ALLES SELECTEREN VERDER ▶


 FRANÇOIS DUBOIS
Voeding
☒


 HELEN JONES
Cognitie
☒


 EMMA LI
Sociaal
☒


 CARLOS SILVA
Lotgenoot
☒


 OLIVIA SIMONS
Activiteit
☒


 RASMUS JOHANSEN
Chronische Pijn
☐


 KATARZYNA KOWALSKA
Diabetes
☐

! Olivia Simons


Beroep: Coach fysieke activiteit

Houdt van: Met haar hond wandelen, fitness, shoppen, wat drinken met vrienden

Houdt niet van: Auto's, en vervuiling in het algemeen


Als u aan uw fysieke activiteit wilt werken, dan past Olivia perfect bij u. Ze zit boordevol enthousiasme, advies en kennis over hoe u een fysiek actieve levensstijl kunt bereiken. Zij helpt u doelen te stellen en u daaraan te houden.

ALLES SELECTEREN VERDER ▶



 FRANÇOIS DUBOIS
Voeding
☒


 HELEN JONES
Cognitie
☒


 EMMA LI
Sociaal
☒


 CARLOS SILVA
Lotgenoot
☒


 OLIVIA SIMONS
Activiteit
☒


 RASMUS JOHANSEN
Chronische Pijn
☐


 KATARZYNA KOWALSKA
Diabetes
☐

! Emma Li

Beroep: Sociale coach

Houdt van: Yoga, sushi, pianospelen, unieke schoenen verzamelen

Houdt niet van: Enge films kijken

Emma is een erg sociaal persoon die weet hoe het is om alleen te zijn. Emma vindt het sociale gezichtspunt voor elke situatie, en kan u helpen met tips en advies om een sociaal actief leven te leiden.

ALLES SELECTEREN

VERDER ▶

FRANÇOIS DUBOIS
Voeding

☒

HELEN JONES
Cognitie

☒

EMMA LI
Sociaal

☒

CARLOS SILVA
Lotgenoot

☒

OLIVIA SIMONS
Activiteit

☒

RASMUS JOHANSEN
Chronische Pijn

☐

KATARZYNA KOWALSKA
Diabetes

☐

! Helen Jones

Beroep: Cognitiecoach

Houdt van: Puzzels, wandelen in de natuur, tuinieren, breien, yoga, katten

Houdt niet van: Grote steden en lawaai

Helen is hier om u te helpen uw hersens flexibel en scherp te houden. Als cognitietrainer geeft ze advies en tips om cognitief fit te blijven in uw dagelijks leven door middel van eenvoudige taken en spelletjes.

ALLES SELECTEREN

VERDER ▶

FRANÇOIS DUBOIS
Voeding

☒

HELEN JONES
Cognitie

☒

EMMA LI
Sociaal

☒

CARLOS SILVA
Lotgenoot

☒

OLIVIA SIMONS
Activiteit

☒

RASMUS JOHANSEN
Chronische Pijn

☐

KATARZYNA KOWALSKA
Diabetes

☐

! François Dubois

Beroep: Dieetcoach

Houdt van: Eten, kaas, Franse muziek

Houdt niet van: Over zichzelf praten

Als dieetcoach kan François u helpen om gezonder te eten en drinken. Hij kan u helpen om dieetdoelen te stellen, en om u daaraan te houden. Met François kunt u uw wekelijkse eetpatroon bijhouden, en hij heeft een grote verzameling recepten waaruit hij u graag helpt kiezen, mocht u kookinspiratie nodig hebben.

B.2 Field Experiment

B.2.1 Invitation letter

Beste (oud-)deelnemer van de Council of Coaches,

Mijn naam is Laura Bosdriesz en ik studeer Interaction Technology aan de Universiteit van Twente. Momenteel ben ik bezig met mijn afstudeerproject waarbij ik de Council of Coaches heb veranderd in een applicatie die werkt via spraak. Dit betekent dat er nu gesprekken met de coaches gehouden kunnen worden door tegen ze te praten, waarop de coaches op hun beurt weer terugpraten. Voor het evalueren van dit systeem ben ik op zoek naar mensen die al ervaring hebben met het gebruik van Council of Coaches, om zo een goed beeld te krijgen van de voorkeuren van de gebruiker en de effectiviteit van een op spraak gebaseerd systeem.

Ik hoop dat u het leuk heeft gevonden om deel te nemen aan het eerste onderzoek en graag zou ik vragen of u nog een week wilt meedoen aan dit afstudeeronderzoek?

Wat houdt het onderzoek in?

Mocht u interesse hebben in het onderzoek, dan zal u rond half september een informatie brochure toegestuurd krijgen met alle details van het onderzoek. U zult via email benaderd worden voor het maken van een afspraak voor een online introductiegesprek. Tijdens dit introductiegesprek wordt de spraak-versie van de Council of Coaches uitgelegd en wordt u gevraagd om twee vragenlijstjes in te vullen met betrekking tot u geslacht, leeftijd, opleiding en werk, en met betrekking tot u ervaring met technologieën. Na dit gesprek kunt u zo vaak als u wilt, zo lang als u wilt en op momenten dat het u uitkomt het systeem gebruiken. Een week later wordt opnieuw een meeting via Skype ingepland waarbij u vragen krijgt over wat u van het systeem vond, en wordt er een vergelijking gedaan met het oorspronkelijke op tekst-gebaseerde systeem.

Hoe kunt u mee doen?

Mocht u interesse hebben of nog vragen hebben dan hoor ik graag van u via onderstaande contactgegevens:

- Naam: Laura Bosdriesz

Met vriendelijke groeten,

Laura Bosdriesz

B.2.2 Information brochure

UNIVERSITEIT TWENTE.

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE

Informatieblad

INFORMATIEBLAD VOOR ONDERZOEK

'Het Toevoegen van Spraak aan de Council of Coaches Applicatie'

Doel van het onderzoek

Dit onderzoek wordt geleid door Laura Bosdriesz. Dit onderzoek is onderdeel van een afstudeerproject waarbij een op spraak-gebaseerde versie van de Council of Coaches applicatie is gemaakt. Het doel van dit onderzoek is om een de robuustheid van dit systeem, en de bijbehorend voor-en nadelen in kaart te brengen.

Hoe gaan we te werk?

Het onderzoek begint met een online introductiegesprek. Tijdens dit introductiegesprek wordt de spraak-versie van de Council of Coaches uitgelegd en wordt u gevraagd om twee vragenlijsten in te vullen met betrekking tot u geslacht, leeftijd, opleiding en werk, en met betrekking tot u ervaring met technologieën. Deze vragenlijsten kunnen via een Google form worden ingevuld. Na dit gesprek kunt u zo vaak als u wilt, zo lang als u wilt en op momenten dat het u uitkomt het systeem gebruiken. U wordt verzocht om op het meegegeven dagboek formulier in te vullen wanneer u de applicatie gebruikt en wat de voornaamste zaken waren die opvielen tijdens de interactie. Een week later wordt opnieuw een meeting via Skype ingepland. Tijdens deze meeting wordt informatie vergaard door u te interviewen en uw antwoorden op te nemen via een Skype beeld- en audio-opname. Er zal ook een transcript worden uitgewerkt van het interview. Tijdens dit interview worden vragen gesteld over u ervaringen met het systeem, en wordt er een vergelijking gedaan met het oorspronkelijke op tekst-gebaseerde systeem.

Potentiële risico's en ongemakken

Er zijn geen fysieke, juridische of economische risico's verbonden aan uw deelname aan deze studie. U hoeft geen vragen te beantwoorden die u niet wilt beantwoorden. Uw deelname is vrijwillig en u kunt uw deelname op elk gewenst moment stoppen.

Vergoeding

U ontvangt voor deelname aan dit onderzoek geen vergoeding .

Vertrouwelijkheid van gegevens

Wij doen er alles aan uw privacy zo goed mogelijk te beschermen. Er wordt op geen enkele wijze vertrouwelijke informatie of persoonsgegevens van of over u naar buiten gebracht, waardoor iemand u zal kunnen herkennen. Voordat onze onderzoeksgegevens naar buiten gebracht worden, worden uw gegevens zoveel mogelijk geanonimiseerd.

Uitsluitend ten behoeve van het onderzoek zullen de verzamelde onderzoeksgegevens worden gedeeld met bevoegde personen van de Universiteit Twente en Roessingh Research & Development.

UNIVERSITEIT TWENTE.

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE

Informatieblad

De audio-opname wordt na 24 uur alleen gebruikt voor het maken van een transcript, daarna wordt deze verwijderd. Het transcript blijft vertrouwelijk en alleen voor analyse. De onderzoeksgegevens worden bewaard voor een periode van 10 jaar. Uiterlijk na het verstrijken van deze termijn zullen de gegevens worden verwijderd. De onderzoeksgegevens worden indien nodig (bijvoorbeeld voor een controle op wetenschappelijke integriteit) en alleen in anonieme vorm ter beschikking gesteld aan personen buiten de onderzoeksgroep.

Vrijwilligheid

Deelname aan dit onderzoek is geheel vrijwillig. U kunt als deelnemer uw medewerking aan het onderzoek te allen tijde stoppen, of weigeren dat uw gegevens voor het onderzoek mogen worden gebruikt, zonder opgaaf van redenen. Het stopzetten van deelname heeft geen nadelige gevolgen voor u.

Als u tijdens het onderzoek besluit om uw medewerking te staken, zullen de gegevens die u reeds hebt verstrekt tot het moment van intrekking van de toestemming in het onderzoek gebruikt worden. Wilt u stoppen met het onderzoek, of heeft u vragen en/of klachten? Neem dan contact op met de onderzoeksleider. Tot 24 uur na het onderzoek kunt u nog weigeren dat de data gebruikt wordt.

Contact informatie

Dennis Reidsma (afstudeerbegeleider)
of
Laura Bosdriesz (student)

Dit onderzoek wordt uitgevoerd vanuit de Universiteit Twente, faculteit Electrical Engineering Mathematics and Computer Science. Voor bezwaren met betrekking tot de opzet en of uitvoering van het onderzoek kunt u zich ook wenden tot de Secretaris van de Ethische Commissie van de faculteit Electrical Engineering Mathematics and Computer Science op de Universiteit Twente:

Contact Information

Ethics Committee, Faculty of EEMCS,
University of Twente
PO Box 217
7500 AE Enschede (NL)
Tel: +31(0)53.4892085
Email: ethics-comm-ewi@utwente.nl
Website: <https://www.utwente.nl/en/eemcs/research/ethics/>

Indien u specifieke vragen hebt over de omgang met persoonsgegevens kun u deze ook richten aan de Functionaris Gegevensbescherming van de UT door een mail te sturen naar dpo@utwente.nl.

Tot slot heeft u het recht een verzoek tot inzage, wijziging, verwijdering of aanpassing van uw gegevens te doen bij de Onderzoeksleider.

B.2.3 Informed consent

UNIVERSITEIT TWENTE.

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE

Toestemmingsverklaring Onderzoek

TOESTEMMINGSVERKLARING ONDERZOEK

Door dit toestemmingsformulier te ondertekenen erken ik het volgende:

1. Ik ben voldoende geïnformeerd over het onderzoek door middel van een separaat informatieblad. Ik heb het informatieblad gelezen en heb daarna de mogelijkheid gehad vragen te kunnen stellen. Deze vragen zijn voldoende beantwoord.
2. Ik neem vrijwillig deel aan dit onderzoek. Er is geen expliciete of impliciete dwang voor mij om aan dit onderzoek deel te nemen. Het is mij duidelijk dat ik deelname aan het onderzoek op elk moment, zonder opgaaf van reden, kan beëindigen. Ik hoef een vraag niet te beantwoorden als ik dat niet wil.
3. Ik geef toestemming dat bevoegde personen van de Universiteit Twente en Roessingh Research Development en bevoegde autoriteiten inzage kunnen krijgen in mijn onderzoeksgegevens. Ik geef toestemming om mijn gegevens te gebruiken voor de doelen die in het informatieblad genoemd zijn.

Naam proefpersoon:

Email-adres:

Datum:

Handtekening:

Ik verklaar hierbij dat ik bovenstaande proefpersoon volledig heb geïnformeerd over het genoemde onderzoek.

Naam onderzoeker: Laura Bosdriesz

Datum:

Handtekening:

B.2.4 Journal

Logboek

Datum:	
Tijd gebruikt:	
Positieve ervaringen:	
Ervaren problemen:	
Andere opmerkingen:	

Datum:	
Tijd gebruikt:	
Positieve ervaringen:	
Ervaren problemen:	
Andere opmerkingen:	

Datum:	
Tijd gebruikt:	
Positieve ervaringen:	
Ervaren problemen:	
Andere opmerkingen:	

Appendix C

Statistical Results

C.1 UEQ: Shapiro-Wilk test for normality check - text

	Statistic	df	Sig.
Perspicuity	0.938	28	0.100
Efficiency	0.953	28	0.242
Dependability	0.975	28	0.729
Stimulation	0.965	28	0.461
Novelty	0.965	28	0.461

C.2 UEQ: Shapiro-Wilk test for normality check - speech

	Statistic	df	Sig.
Perspicuity	0.879	28	0.004
Efficiency	0.946	28	0.156
Dependability	0.974	28	0.691
Stimulation	0.949	28	0.189
Novelty	0.950	28	0.203

C.3 UEQ: Paired samples t-test statistics for the comparison per scale

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Perspicuity_text	1.7143	28	.52137	.09853
	Perspicuity_speech	1.8304	28	.66337	.12537
Pair 2	Efficiency_text	1.1518	28	.78569	.14848
	Efficiency_speech	.9286	28	.87363	.16510
Pair 3	Dependability_text	1.1607	28	.64267	.12145
	Dependability_speech	.9464	28	.65390	.12357
Pair 4	Stimulation_text	.4196	28	.98143	.18547
	Stimulation_speech	1.2411	28	.70541	.13331
Pair 5	Novelty_text	.2857	28	1.26877	.23978
	Novelty_speech	1.5268	28	1.00770	.19044

C.4 UEQ: Paired samples t-test results for the comparison per scale

		Paired Differences					t	df	Sig. (2-tailed)	Cohen's D
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference					
					Lower	Upper				
Pair 1	Perspicuity	-.116	.668	.126	-.375	.143	-.919	27	.366	-0.147
Pair 2	Efficiency	.223	1.231	.233	-.254	.701	.959	27	.346	0.181
Pair 3	Dependability	.214	.676	.128	-.048	.476	1.677	27	.105	0.317
Pair 4	Stimulation	-.821	1.080	.204	-1.240	-.403	-4.025	27	.000	-0.761
Pair 5	Novelty	-1.241	1.142	.216	-1.684	-.798	-5.753	27	.000	-1.087

C.5 Explicit Comparison: One-sample t-test statistics

	N	Mean	Std. Deviation	Std. Error Mean
1) ...thought was more pleasant to use.	28	.29	1.410	.267
2) ...thought was more efficient.	28	-.86	1.268	.240
3) ...thought was more interesting.	28	1.64	.678	.128
4) ...thought was more credible.	28	.86	.970	.183
5) ...thought was more fun.	28	1.50	.793	.150
6) ...would use for a longer period.	28	.64	1.367	.258
7) ...would recommend to someone older than 55.	28	.93	1.274	.241
8) ...would follow advice from more often.	28	1.00	.943	.178
9) ...would rather spend money on.	28	1.11	.832	.157
10) ...thought was more practical.	28	-.29	1.243	.235
11) ...thought was more cumbersome to use.	28	.36	.951	.180
12) ...thought was more monotonous.	28	-.82	.905	.171
13) ...thought was more annoying.	28	.04	.922	.174
14) ...thought was more inconvenient to use.	28	.29	.854	.161
15) ...thought was more repetitive.	28	-.25	1.005	.190