Preliminary version of**: ter Heijden, N., and Brinkman, W.-P. (2011). Design and Evaluation of a Virtual Reality Exposure Therapy System with Automatic Free Speech Interaction.** ***Journal of CyberTherapy and Rehabilitation, 4*****(1). 41-55.**

Design and Evaluation of a Virtual Reality Exposure Therapy System with Automatic Free Speech Interaction

Niels ter Heijden and Willem-Paul Brinkman
Delft University of Technology
The Netherlands

slein13@gmail.com, w.p.brinkman@tudelft.nl

Corresponding Author
Willem-Paul Brinkman
Department of Mediamatics, Man-Machine interaction,
Delft University of Technology, Mekelweg 4, 2628 CD Delft, the Netherlands
Tel: +31 (0)15 2783534
Fax: +31 (0)15 2787141
Email:w.p.brinkman@tudelft.nl

Preliminary version of: **ter Heijden, N., and Brinkman, W.-P. (2011). Design and Evaluation of a Virtual Reality Exposure Therapy System with Automatic Free Speech Interaction.** ***Journal of CyberTherapy and Rehabilitation, 4*****(1). 41-55.**

Design and Evaluation of a Virtual Reality Exposure Therapy System with Automatic Free Speech Interaction

ABSTRACT

Research on Virtual Reality Exposure Therapy (VRET) to treat social phobia is not new. Still few studies focus on creating an elaborate conversation between the patient and characters in a virtual environment. This study focuses on techniques to run a semi-scripted conversation between virtual characters and a patient considering both manual and automatic speech response. Techniques evaluated are a speech detector and a speech recognizer. They were compared to a human control condition. We analyzed the flow and interaction individuals (N = 24) experienced and did a Turing like test. A case study with two phobic patients was also conducted. Both patients and therapist and their interaction with the system were observed. The study showed that the different automatic techniques had their (dis)advantages but often did not show any significant difference with the human control condition. A VRET system with semi-scripted conversations might therefore be suitable for the treatment of patients with social phobia. Using automatic speech response techniques might reduce the system workload demand placed upon therapists, allowing them to devote more attention towards monitoring the patient.

Keywords

Social phobia, public speaking, virtual reality exposure therapy, Turing test, dialogue, natural speech, therapist, patient.

1. INTRODUCTION

People that suffer from social phobia fear social situations in which they belief embarrassment may occur. This often leads to avoidance behaviour. They fear scrutiny and negative evaluation by others. Making a phone call, asking for assistance in a shop, or speaking in public, are all situations they might dread. Social phobia is one of the most common types of anxiety disorders, estimated to affect 13.3% of the US population in their life time (Kessler, et al., 1994). The disorder is associated with depression, substance abuse (e.g. alcoholism, drug abuse), restricted socialisation, and poor employment and education performance (Katzelnick, et al., 2001; Kessler, 2003). The disorder leads to intensive use of health services in the western world (Wiederhold & Wiederhold, 2005).

Exposure in vivo is the gold standard for the treatment of phobias. However, for social phobia this treatment might be difficult to arrange (e.g. arranging an audience), and for the therapist difficult to control (e.g. a patience or a hostile

audience). Exposure in virtual reality (VR) has therefore been suggested as an alternative with already some initial encouraging result (Klinger, et al., 2005; Robillard, Bouchard, Dumoulin, Guitard, & Klinger, 2010). Most virtual reality (VR) research focuses on one specific social situation i.e. speaking in front of a small group of virtual characters also called avatars (Anderson, Rothbaum, & Hodges, 2003; Klinger, et al., 2005; Pertaub, Slater, & Barker, 2001; Slater, Pertaub, & Steed, 1999). Still in the development of these settings the main focus is often on the body posture of the avatars (Anderson, et al., 2003; Herbelin, 2005; Klinger, et al., 2004; Slater, et al., 1999) and less on oral communication between the patient and the avatar. In work (Grillon, Riquier, Herbelin, & Thalmann, 2006) that does report on oral communication, implementations are often relatively limited in their flexibility to support free natural dialogues. This has motivated the here reported study into a public speaking scenario with virtual avatars that can ask patients questions and respond to their reply with follow up questions using automatic free speech interaction.

Nowadays, conversational software agents are used in several areas such as automated phone reservation systems (McTear, O'Neill, Hanna, & Liu, 2005), e-retail (McBreen & Jack, 2001) and real estate agents (Cassell, et al., 1999). Most of these dialogues are goal oriented or follow instructions from a user by scanning for expected keywords. They are not developed with the aim to strike a casual conversation. Still, attempts have been made to create such an agent. The earliest versions might be chatbots (Hutchens & Alder, 1999; Wallence, 2009; Weizenba, 1966). These type of conversational agents have their own contest ("The Loebner Prize in Artificial Intelligence," 1991) where they compete in a Turing test setting with the aim of developing a conversational agent of which the dialogue can not be distinguished from a human dialogue. Thus far, none have passed the test successfully and chatbot technology might therefore not be ready to simulate a conversation in a VR treatment setting. Most work seem to focus on Artificial Intelligence Markup Language AIML (Wallace, 2001) or variants (Galvao, Barros, Neves, & Ramalho, 2004) type bots, although other techniques have also been suggested such as sentence grammatical analysis (Araki & Doshita, 1995; Li, Zhang, & Levinson, 2000). Besides the dialogue reasoning component, the oral user input itself is also problematic. The perfect speech recognizer that converts human speech into computer understandable language does not yet exist (Jurafsky & Martin, 2009). Therefore, free speech seems too ambitious to realize with existing technology (Jurafsky & Martin, 2009; Martin, Botella, García-Palacios, & Osma, 2007). Instead semi-automatic alternatives seem more feasible. For example, therapists could control most of the avatars' behaviours, by simply listening to the patient and select appropriate responses. In this way, avatars might have a realistic (oral) behaviour; the drawback, however, might be the extensive workload forced up on the therapists. An alternative approach, which removes much of this workload, is to have patients read out loud one of the patient-responses from a short list of possible responses shown on the screen (Brinkman, van der Mast, & de Vliegher, 2008). This method can fairly successful be implemented with speech recognition. But with a list of

sentences on the screen it is unclear how this affects the level of presence and the associate anxiety response. Instead, an intermediating solution, using automatic keyword detection from the users' free speech (McTear, 2002), is explored here. This technique was combined with semi scripted dialogues that were controlled by a computer algorithm deciding what response the avatar should give to the patient. The focus on the research was to examine what level of technology sophistication would yield the best results taking the amount of work creating the dialogues in mind. Three automatic speech response techniques and a manual control condition were empirically compared in a controlled experiment and in a case study with two social phobia patients. The results suggest that a VR system with semi-scripted conversations might be a suitable exposure environment for social situations that have a conversational component. Using automatic response techniques are likely to reduce therapist workload, while giving individuals a similar level of presence.

2. SYSTEM DESIGN

An experimental virtual reality exposure therapy (VRET) system was created to provide a controlled environment for exposing patients to a public speaking situation. Patients were asked to talk for a few minutes about a specific subject in front of a small virtual audience after which virtual avatars asked the patients questions about the subject, and responded to the answers of patients in follow up questions.

2.1. System Set Up

Figure 1 shows the set up of the system used in the study. As the presence of the therapist might have an unwanted effect on the patient, a remote set up was developed by which patient and therapist were separated into two different rooms. The system was distributed over two computers; one handled the visualization of the virtual world including the avatars and all the audio input by the patient, while a second computer handled the interaction with the therapist and the reasoning logic of the dialogue. For the net communication and the therapist interface the Remote Delft Virtual Reality Exposure Therapy (RDVRET) framework was used. This framework handled the network communication between the two computers and the different software components, and also provided the basic graphic user interface components for the therapist user interface. The therapist could monitor the patient on a screen with speakers by a live video feed from a camera at the back of the patient room and an audio link with the patient room. The therapist could also open an audio channel to talk to the patient.

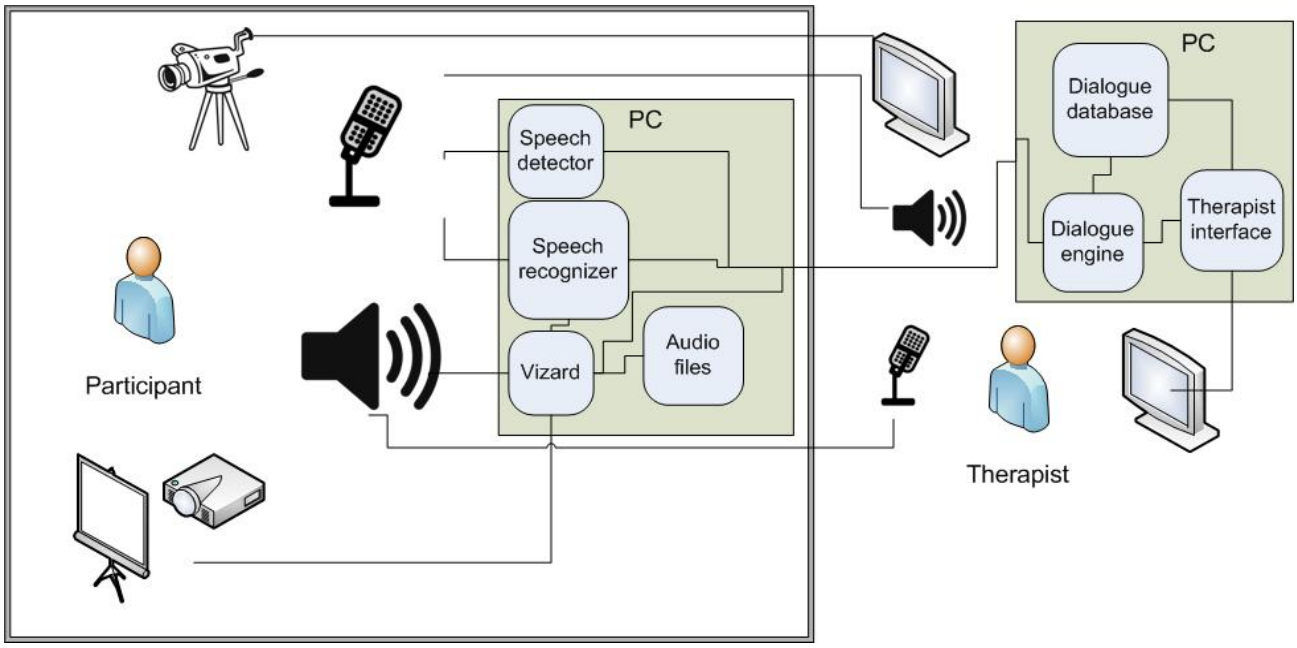

Figure 1: System set up

The software package Vizard was used for the visualization of the virtual room and the avatars. The avatars came from a commercial avatar package especially developed for Vizard. Extra animations for the avatars were modelled using key-framing. To support natural turn taking between the avatars and the patient, a software-routine was implemented ensuring that the avatars looked at the avatar that was talking or at the patient that was talking. Patients sat behind a table, with a microphone, and they wore a head set. They were sitting two meters away from the screen, on which a 3.5 by 2.5 meters virtual room was projected with a screen resolution of 1280 × 1024 pixels.

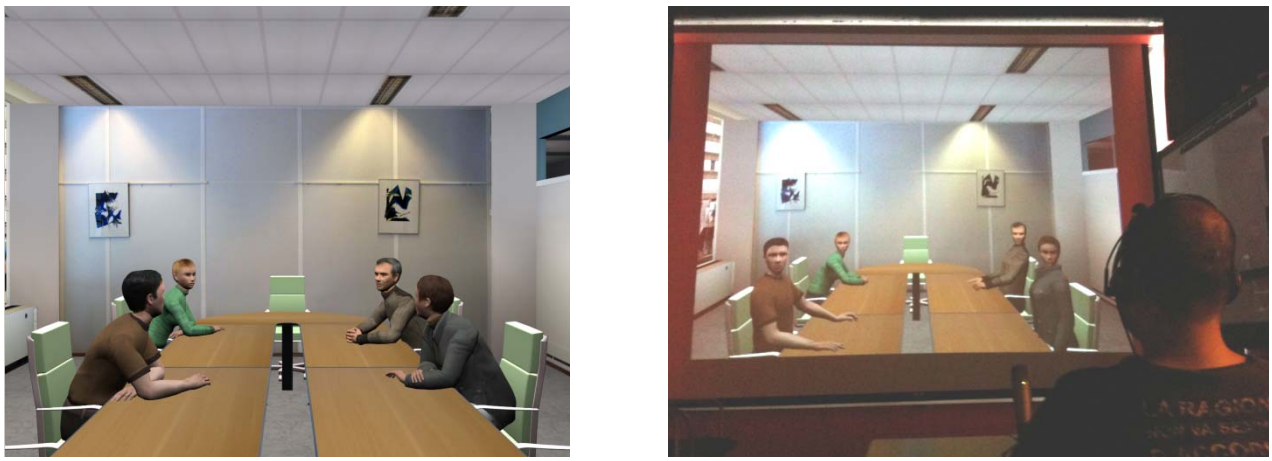

Figure 2: Virtual reality room. Right, an avatar is talking. Left, avatars look at the patient indicating they are listening to what the patient is saying.

To study different ways of interpreting and responding to patients, four different speech response conditions were implemented. The first condition did not use any speech recognition but solely used the amount of time the patient was talking to base its response on. The second condition did use the speech recognizer but only checked on a pre defined finite list of key-words that would be the same for all dialogues. The third condition added the ability to check for certain specific key-words appropriate for the specific point in the dialogue. Finally the last condition was a control condition, where a therapist selected the responses instead of a computer algorithm.

2.2. Dialogue Development

The system was designed to support an alternating computer-patient dialogue, in which the computer (i.e. the avatars) first asked the patient something, at which the patient replied, at which the computer on its turn replied again etc. As patients could speak freely, the computer should be able to react on various patient responses. Figure 3 shows an example of a part of such a dialogue. Each computer reply was linked with several potential patient replies. Although in theory there could be an infinitive number of different user replies, potential patient replies were grouped, and for each of these groups an appropriate computer reply was written. To control for the potential exponential growth of a dialogue tree, the depth of the tree was limited to a maximum of five turns after the computer leaf nodes were again brought together in a new opening question node by the computer, who always took the lead in the dialogue. In this way the final dialogue followed a repeating pattern of a widening tree merging back to a single node and widening again from there. A special editor tool, Editor3 (ter Heijden, Qu, Wiggers, & Brinkman, in press), was developed to support the development of the dialogues. The dialogues were saved to Sqlite database files that could be read later on by the dialogue engine (Figure 1).

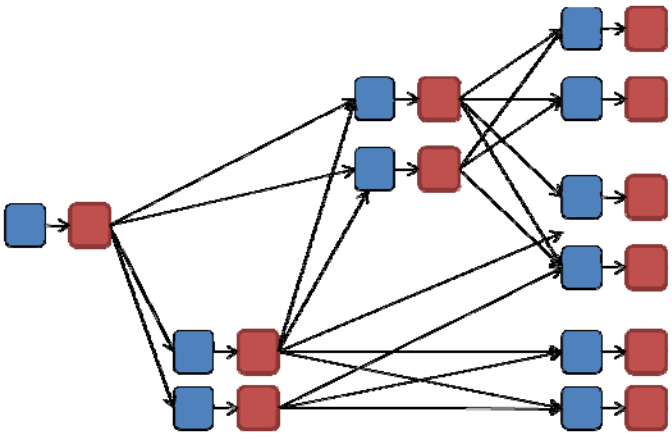

Figure 3: Example of dialogue structure in which the computer reply was linked with various patient responses.

For the study four dialogues were created on the subject: democracy, France, dogs, and penguins. For a Dutch target group these subjects were considered general enough that most people would have some knowledge and opinions on it, and allowing the avatars to pose some knowledge and opinion questions about it. Questions about recent events were avoided to make the dialogue more future proof. After the initial draft, the dialogues were evaluated and rewritten in iterative cycles. This was done by implementing the dialogue in a chatbot and asking users to chat with it. The results of chatbot sessions (ter Heijden, et al., in press) were used to evaluate the dialogues and to extract typical user replies, including keywords.

2.3. Avatar Response Categories

Besides the questions avatars could pose, special consideration was given to the patient reply. The grouping of these replies was based on two computational parameters: the length of the speech and the detection of specific keywords in the patient's reply by the speech recognizer. Potential patient responses were grouped into seven categories. The first category was the short response category. This was the most basic category. In the speech detection condition this was implemented by detecting whether a patient had given a reply that took less than one second, while in the two conditions with the speech recognizer an additional check was implemented to include only sentences with less than six words. Possible avatar response for this category were 'could you tell us something more?' or simple 'why?' The main aim of the avatars replies here was to engage patients in longer replies, and thereby overcoming potential avoidance behaviour. The second category was the Yes/No category. This category was used if the speech recognizer found the word "Yes" or "No" in the patient response. Possible avatar responses for this category were for example "why is that?" or more context related such as "What characteristics of birds do they have then?" on a patient yes reply on the question whether they consider a penguin a bird. Likewise, a no response would result in a reply such as 'Why not?'. The only additional action for a no response was a check for double negative, in which case that response was ignored for this category. The third category was the positive / negative category. Patient responses which included mainly positive words such as "of course" or "rightly" were considered positive replies. Responses which included mainly negative words such as "not" and "nothing" were considered negative replies. Replies on patients responses in this category included related questions that were based on the implications of acceptance or rejection of assumptions in the previous questions. For example, a patient's negative response on the question "What is your favourite penguin? would lead to the avatar asking "What is your favourite polar animal?" The fourth avatar response category was the don't know category. This included patients' response such as "I don't know" or "I don't have an opinion on that". Possible replies on this category were "Could you really not think of anything?" or "That is ok". The fifth category was the keywords category. This category included

responses in which a specific keyword was detected. Compared to the other categories, developing replies for responses in this category took more development effort. First a list of keywords had to be identified that patients would often use in their response. This was done based on a word frequency analysis on the replies obtained in chatbot evaluations. Second, for each keyword a specific avatar reply was written. In the other categories avatar reply was less dependent on the question leading up to patient response, and these replies could therefore be used at multiple occasions in the dialogue. This was not the case for patient responses in the keyword category. Here unique replies were written for each specific preceding avatar question. For example, when the patient response on the question "What kind of penguin do you know?" included the word emperor penguin, the avatar would continue the dialogue with the follow up question "What is the difference between an emperor penguin and a kings penguin?". The expected added value for this category was that patients might feel that the avatars were really listening and responding to their replies, making the social stressors more authentic. The sixth avatar response category was the general category. These were avatar responses that did not relate to what the patient might have said. This included opening questions, but also follow-up question on aspects patients might not have addressed yet, making them still natural to appear in dialogue. For example, a follow-up question for the question "What is the greatest threat to the penguin species?" was "The penguin is not a protected species. Do you think they should be?". Both questions have synergy with each other, while it is unlikely that the patients might have already addressed the issue of penguins being unprotected species in their initial reply. The seventh and last category is the end category. These are the avatar response to round up a question line with a remark such as "But we are straying too far away from the main topic, so back to the penguins", or shorter ones like "Ok" or "If you say so". Although these avatar responses could provoke a reply from the patient, the avatar did not directly respond on that. Instead another avatar started with a new question line.

Table 1 shows that not all avatar response categories were implemented in each speech response condition. No speech recognition technology was used in the speech detector condition. In this condition the computer only measured when a patient talked or stopped talking, and the amount of time a patient talked. On the other hand, in the (limited) speech recognition conditions, the speech recognisor software Nuance Naturally Speaking was used. Whereas the limited speech recognition condition only included the more content generic avatar response category, the speech recognition condition also included the keyword category. These two conditions were included to study the effect of including the keyword category as the development of this category is more labour intensive. In the human control condition the therapist selected an avatar reply from all the avatar response categories. For example, across the four dialogues, this meant the therapist on average could select from 4.35 avatar responses after the avatars had posed their opening question of a new question line.

Table 1: Implementation of avatar response categories in the four speech response conditions.

Avatar response category	Speech detector	Limited speech recognition	Speech recognition	Human Control
Short	X	X	X	X
Yes/No		X	X	X
Positive / Negative		X	X	X
Don't know		X	X	X
Keywords			X	X
General	X	X	X	X
End	X	X	X	X

3. METHOD OF THE EXPERIMENT

The experiment was set up as a within-subject design. All participants were exposed to the four speech response conditions: speech detector, limited speech recognition, speech recognition, and human control. To avoid possible learning effects about a subject, a participant talked about another subject in each condition. Furthermore the order of condition and the assignment of the subjects to conditions were balanced. This resulted in 24 sequence orders. Each participant was randomly assigned exclusively to one of these sequence orders.

3.1. Measures

Before the exposure, participants were asked to complete the Personal Report of Confidence as a Public Speaker (PRCS) questionnaire (Paul, 1966) to measure the participant's fear of public speaking. Furthermore, they were asked to complete a questionnaire to collect basic information such as gender, age etc and the level of proficiency and knowledge about computers, 3D techniques, and virtual reality. After the each condition, the participant sense of presence was measured with a modified version of the Igroup Presence Questionnaire (IPQ) (Schubert, Friedmann, & Regenbrecht, 2001) including only the general sense of being there question and the experienced realism questions to measure their subjective experience of realism in the virtual environment. In addition participants completed the Dialogue Experience Questionnaire (DEQ) (Table 5), to measure their experience of the dialogue and the avatars. After the four dialogues, participants were asked to complete the full IPQ questionnaire considering all four dialogues, and to complete three Turing test type questions. Here they were asked to rate the dialogues and place them in order of likeliness they were controlled by a computer or a human. As a check participants were also asked to order the dialogues on the smoothness of the conversation. Besides subjective data, behavioural data was also collected about the number of times participants

were interrupted by the avatars when they had not yet completed their answer. This also included situations where they had just paused for a moment and continued talking.

3.2. Procedure

At the start of the experiment participants received a short introduction about the overall aim of the study, and had to sign a consent form. Participants were, however, not informed about the different speech response conditions. After signing the form, they completed the PRCS and basic information questionnaire. Once this was complete, the speech recognizer was trained. The main part of the experiment consisted out of four sessions with the virtual audience, talking about four different subjects. To help them during the initial three minutes presentation about the subject, they were given a sheet with some general pointers to talk about, which however did not overlap with the question set of the avatars. The participants were also instructed not to pose questions to the avatars. The presentation phase lasted between 1.5 and 3 minutes, after which avatars would start the question and answer (Q&A) phase. This consisted out of 8 to 10 main questions with 1 to 5 follow up questions each (ter Heijden, et al., in press). After this, participants filled out IPQ and DEQ. Once completed, they received the presentation sheet to prepare themselves for the next session. Between the second and third session, participants were allowed to take a short break to drink something and walk around. After the four sessions, participants completed the Turing test type questions and were interviewed about the overall experience. The entire experiment took between 1.5 and 2 hours. After thanking the participants they received a small gift in the form of a chocolate bar as thanks for their participation.

3.3. Participants

Participants were recruited from the researchers own social network. The age of the 24 participants (7 female) ranged from 17 to 66 years (M = 32.6, SD = 13.4). A higher education was followed or completed by 14 participants, while 10 participants had no education, or had completed or were following an education below this level. From the 24 participants 21 had seen 3D stereoscopic images, but only nine had ever used a virtual reality system before. Participants had an average score of 9.8 (SD = 6.5) on the PRCS questionnaire, with three participants scoring 16, 25, and 26, falling into the category of subjects Paul (1966) had included in his anxiety treatment study.

4. RESULTS EXPERIMENT

4.1. Presence

The overall IPQ results were compared with the online IPQ data set[1] in a MANOVA which used the source as independent variable and the general presence and three factors as dependent variables. No clear significant ($F(4, 632) = 2.13$, $p. = 0.076$) difference was found between the online IPQ data set and IPQ rating from this experiment. However, univariate analysis found a significant effect ($F(1,635) = 7.79$, $p. = 0.005$) on the experienced realism factor. Participants in this experiment ($M = 2.5$, $SD = 1.13$) rated the experienced realism higher than ratings in the online IPQ data set (2.0, $SD = 0.83$). After each dialogue session participants had also filled out a modified version of the IPQ questionnaire that only included general presence and experienced realism factor questions. A MANOVA with repeated measure taking the speech response condition as within-subject variable and the two IPQ measures as dependent measures found no significant effect ($F(6,18) = 1.7$, $p. = 0.184$).

4.2. Dialogue Experience

The results of the DEQ were analysed for their internal consistency. Items that had low correlations with other factor items were removed. The mean Cronbach's α over the four conditions ranged from 0.78 to 0.90, above the 0.7 threshold. The participants mean score on the factor items were therefore used for future analysis.

Table 2: Cronbach's α results from reliability analysis on factors items of the DEQ.

Factors	Speech detector	Limited speech recognition	Speech recognition	Human control	Mean
Flow					
dialogue speed	0.83	0.81	0.79	0.82	0.81
interruption	0.91	0.87	0.86	0.95	0.90
correctness locally	0.86	0.61	0.88	0.76	0.78
correctness globally	0.76	0.81	0.74	0.82	0.78
Interaction					
involvement	0.80	0.73	0.88	0.87	0.82
discussion satisfaction	0.84	0.82	0.75	0.83	0.81
reality	0.79	0.77	0.83	0.76	0.78

[1] www.igroup.org/pq/ipq/data.sav downloaded on 12 Nov 2010

To examine the effect of the conditions on DEQ, a MANOVA with repeated measure was conduct. The speech response condition was taken as a within-subject variable and the four flow factors as dependent measures. The results revealed a signification overall effect ($F(12,12) = 5.63$, $p. = 0.003$) for the condition on the flow factors. Univariate analyses on the individual factors, only found a significant effect for speech response condition on dialogue speed ($F(3,69) = 14.74$, $p. <0.001$), and the correctness locally ($F(3,69) = 7.09$, $p. < 0.001$) factor. Figure 4 shows the mean factor scores. Pair wise comparisons with Bonferroni correction showed that for the dialogue speed rating the human control condition ($M = 4.3$, $SD = 1.36$) was significantly higher than the limited speech recognition ($M = 3.1$, $SD = 1.14$, $p = 0.001$) and the speech recognition ($M = 2.9$, $SD = 1.22$, $p < 0.001$) condition. Likewise the speech detector ($M = 3.9$, $SD = 1.25$) was also rated significant higher than the limited speed recognition ($p. = 0.016$) and speech recognition conditions ($p. = 0.003$). It seems therefore that the delay in the avatars response when using the speech recognition was noticeable for participants. For the factor correctness locally participants rated the human control condition ($M = 4.8$, $SD = 1.04$) higher than the speech detector ($M = 4.3$, $SD = 1.16$, $p. = 0.007$) and the limited speech recognition ($M = 3.8$, $SD = 0.90$, $p. < 0.001$) condition.

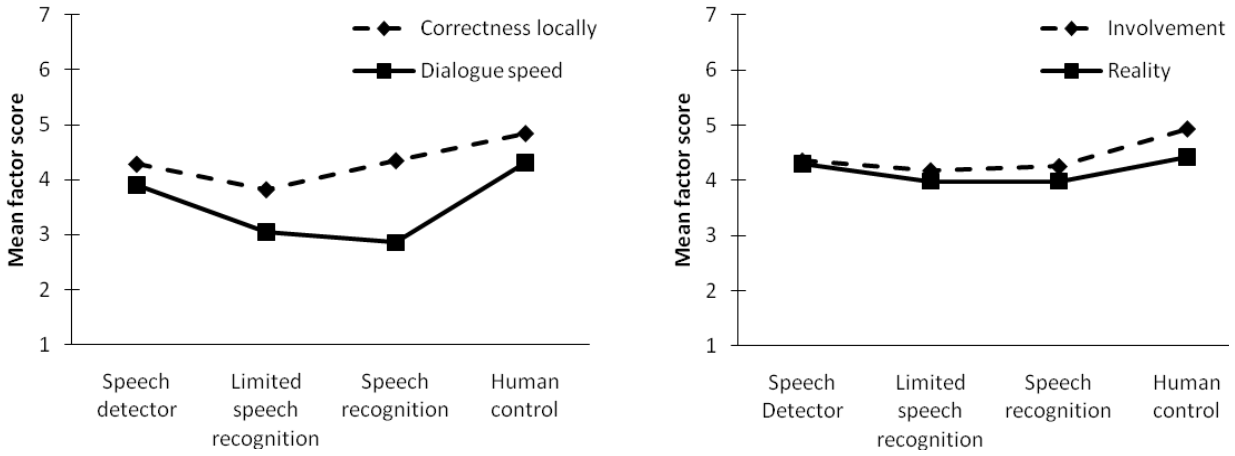

Figure 4: Mean factor score on DEQ factor on which significant effects where found for the speech response conditions.

A similar analysis was done on the three DEQ interaction factors. This analysis found a significant ($F(9,15) = 5.36$, $p. = 0.002$) overall effect for the conditions. Univariate analysis revealed a significant effect in the involvement factor (Greenhouse-Geisser correction: $F(2.2, 50.7) = 5.65$, $p. = 0.005$) and in the reality factor ($F(3, 69) = 3.56$, $p. = 0.019$). Pair wise comparisons with Bonferroni correction showed for the involvement factors that the participants had rated the human control ($M = 4.9$, $SD = 1.19$) significantly higher than any of the other conditions (speech detector: $M = 4.4$, $SD = 1.21$, $p. = 0.046$: limited speech recognition: $M = 4.2$, $SD = 1.12$, $p. < 0.001$; speech recognition $M = 4.3$, $SD = 1.40$, $p. =$

0.002). For the reality factor human control (M = 4.4, SD = 1.12) was only rated significantly higher than the limited speech recognition condition (M = 4.0, SD = 1.15, p. = 0.05).

4.3. Turing Test

During the experiment participants were not informed about the different speech response conditions. Because of the speech recogniser training, it seems likely they might have expected some computer control, if not all. After completing all dialogue sessions they were asked to rate their sessions on the likeliness that they were computer or human controlled. Figure 5 shows the mean ratings. An ANOVA found a significant effect (F(3, 69) = 3.35, p. = 0.024) for the speech response conditions. Pair wise comparison with Bonferroni correction showed that the human control condition (M = 4.4, SD = 1.53) was rated significantly (p. = 0.015) as more likely to be controlled by a human than the speech recognition condition (M = 2.7, SD = 1.85). Interestingly, the rating for human control condition did not significantly (t(23) = 1.33, p. = 0.195) deviate from the middle of the scale, even though participants were informed one of the four sessions was human controlled before they rated this question. A Friedman test on the ordering task of the sessions on the likeliness of computer or human control also revealed a significant effect (χ^2(3) = 7.85, p. = 0.049), as was also the case for the smoothness ordering task (χ^2(3) = 14.15, p. = 0.003). As can be seen in Figure 6, and was confirmed by Wilcoxon signed ranks pair wise comparison tests with Bonferroni correction, participants ranked the human control condition significantly higher than the speech recognition condition (Turing test: p. = 0.037; Smoothness: p. = 0.037).

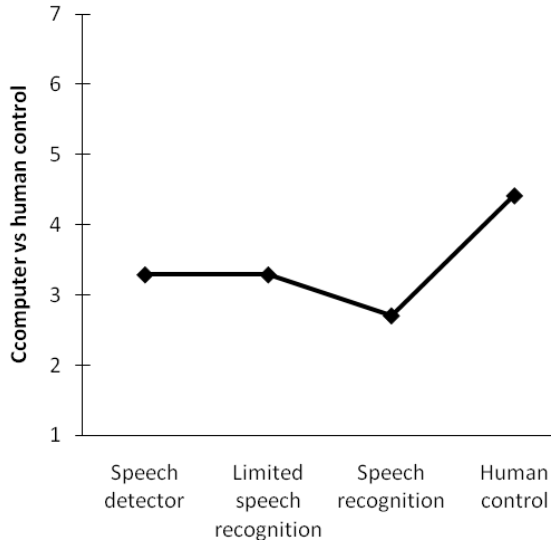

Figure 5: Mean rating on likeliness speech response was controlled by a computer or a human.

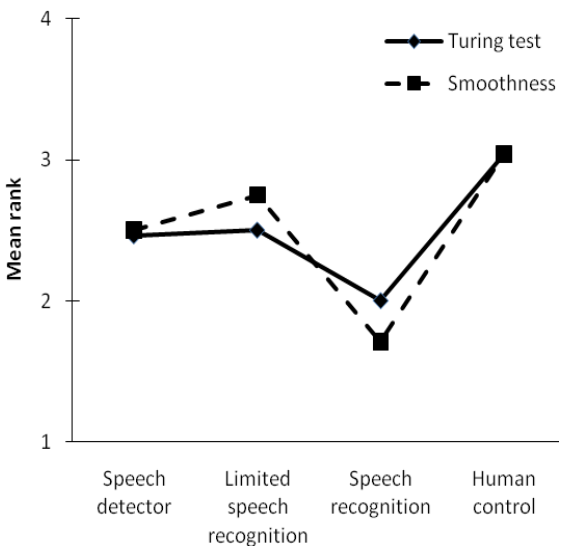

Figure 6: Mean ranking speech response condition on likeliness they were controlled by a computer or a human, and the smoothness of the dialogue

4.4. Avatar Interruptions

Figure 7 shows the number of times an avatar had interrupted the participants when they were still speaking. Interestingly, this was not only a problem for the automatic response conditions as it had also occurred in the human control condition. A Friedman test revealed a significant effect ($\chi^2(3) = 10.02$, $p. = 0.018$) for speech response condition. Wilcoxon signed ranks pair wise comparison tests with Bonferroni correction, found only that significantly ($p. = 0.034$) more interruptions were made in the speech detector ($Mdn = 2$) condition than in the human control condition ($Mdn = 0.5$).

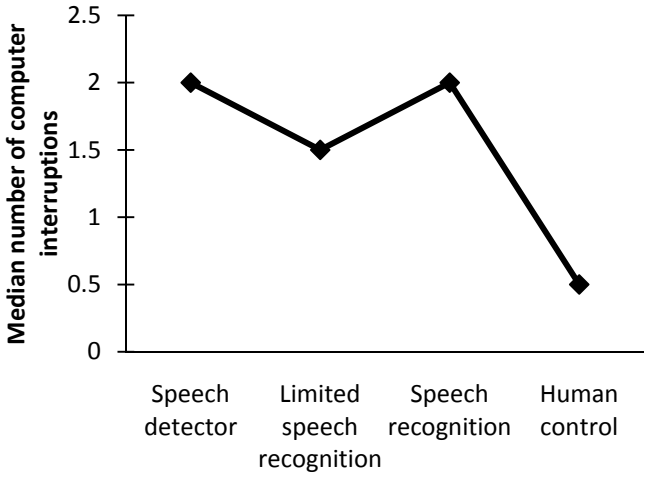

Figure 7: Median of the number of computer interrupts while participants had not finished their answer.

4.5. Discussion

Although several significant differences were found, more interesting is to notice that on several places no significant difference was found especially with the human control condition. A sample size of 24 for a One-sample *t*-test would give an 80% change of finding a significant effect ($d = 0.6$) with a size somewhere halfway between what Cohen classified as a large ($d = 0.8$) and a medium ($d = 0.5$) size effect. Therefore the absence of a significant difference could be interpreted that it is unlikely that a difference with a large effect would exists. Only on one measure did the human control condition outperform all the automatic speech response conditions –the DEQ interaction involvement factor. Clearly participants felt that avatars were more listening to them when they were controlled by a human. However, this did not seem to have had any effect on their overall feeling of presence as no difference was found between any of the speech response conditions. Likewise in the Turing test, the participants seemed to be unable to make a clear difference between the human control condition and some of the automatic speech response conditions. A drawback of the current

implementation of the speech recognisor was the noticeable response delay. If this were to be improved, it would be beneficial as the human control condition did not noticeably outperform the speech recogniser in correctly responding to a participant's answer or holding a conversation that could take place in real life. Ideally the automated system should combine the speed of the speech detector with the information the speech recognizer gives. Partial sentence information might even be enough for the limited speech recognizer because much of the attitude information can already be found in the first few words of the sentence.

5. CASE STUDY

As the experiment was conducted with non-patients, a case study was conducted to see how actual patients suffering from social phobia would behave and perceive the system. As this was a preliminary study, any potential treatment effect was not examined. Besides the patient, the case study also provided an opportunity to study how an actual therapist would use and perceive such a system, a user side of VRET systems only explored in a few reports (Brinkman, Sandino, & Van der Mast, 2009; Brinkman, Van der Mast, Sandino, Gunawan, & Emmelkamp, 2010; Wrzesien, Burkhardt, Alcañiz Raya, Botella, & Bretón López, 2010).

5.1. Method

One of the university student psychologists, and two patients, referred to here as client 1 and client 2, volunteered to participate in the case study. Client 1 was a 43 year old male MSc student that had not finished his final thesis project, as he avoided his thesis presentation to his supervisors. He wanted to participate because of his personal interest in the technology. Furthermore, it was an opportunity for him to enter the building of his faculty, something he had avoided for some time. Client 1 had a PRCS score of 25, and on the dichotomous version of the Tellegen Absorption Scale (TAS) he scored 11. Client 2 was a 25 year old male BSc student that had anxiety of presenting in public, believing that he would shake, stutter and talk incoherent when talking before a group of people. He was currently attending sessions from another psychologist. He had a PRCS score of 16, and a TAS score of 16. The therapist was a female psychologist that worked at the university and treated students from the university. The set up of the system and room was similar as in the experiment. Because of time constraints, the patients were not exposed to the limited speech response condition. The patients were invited for two sessions. The first session was used to introduce the patients to the system, and trained the speech recognizer on their voice. They also practised with the system using a few questions from the France dialogue, using a human response condition. In the second session, the patients were exposed to the democracy, dogs, and penguins dialogues. They started with the speech detection condition, followed by the human control condition, had a

small break, and finished with the speech recognizer condition. After each dialogue session they completed DEQ, and after the last session they completed IPQ. This was followed by an interview on how they had experienced the sessions. After this, the clients were thanked and for their participation received a small gift that had a value of about €10.

5.2. Results Patients

Whereas clients 2 presence score on IPQ did not differ from the participants from the experiment, client 1 gave the overall presence the maximal score of 6, which was significantly ($t(23) = -8.62$, $p. < 0.001$) higher than the participants average rating ($M = 3.5$, $SD = 1.44$). However, on IPQ sub-scales he scored significantly lower (spatial presence 0.4: M = 3.4, $SD = 1.14$, $t(23) = 12.87$, $p. < 0.001$; involvement 1.5: $M = 3.3$, $SD = 1.15$, $t(23) = 7.57$, $p. < 0.001$; experienced realism 1.75: $M = 2.5$, $SD = 1.13$, $t(23) = 3.29$, $p. = 0.003$). Client 2 scored the experienced realism as 2, which was also significantly ($t(23) = 2.21$, $p. = 0.037$) lower than the mean score obtained in the experiment.

The results of DEQ showed (Table 3) that client 1 was rating the flow factors and the interaction factor discussion satisfaction significant lower than participants had done after their dogs dialogue session with the human control speech response condition. This was the first time with a patient that the therapist controlled the system, which might partly have caused this. Client 2 gave a lower score to the flow factors than participants in the penguins dialogue with the speech detection condition. However, he gave a much higher rating for almost all DEQ factors for the dogs dialogue with the speech recognizer. Overall it seems therefore that both clients did not have a greater tendency than the participants in the experiment in experiencing the dialogue with human control as more positively, nor was the dialogue with the speech recognizer received more negatively. The data for the speed detector condition seems however less conclusive.

Table 3: DEQ results case study clients and experiment participants.

	Client 1 (experiment *M*)			**Client 2 (experiment *M*)**		
Factors	***Democracy – Speech detection***	***Dogs – Human control***	**Penguins – Speech recognizer**	**Penguins – Speech detection**	**Democracy – Human control**	**Dogs – Speech recognizer**
Flow						
dialogue speed	2.8 (3.3)	3.0 (5.4)*	3.0 (3.3)	2.0 (3.7)*	4.3 (4.1)	5.5 (2.4) ¥
interruption	3.4 (3.2)	6.0 (5.7)	4.2 (4.5)	5.6 (4.6)	5.8 (4.6)	3.4 (4.4)
correctness locally	3.8 (3.8)	3.3 (5.3)¥	3.8 (4.9)	2.8 (4.7)*	4.8 (5.0)	5.5 (4.0)*
correctness globally	4.0 (4.6)	3,8 (5.5)*	4.3 (4.5)	4.5 (4.2)	5.0 (5.1)	6.8 (4.8)*
Interaction						
involvement	4.4 (3.7)	4.2 (5.3)	4.0 (5.1)	4.2 (4.7)	4.8 (5.5)	5.6 (3.5)*
discussion satisfaction	3.4 (4.3)	3.6 (5.9)¥	4.4 (5.1)	5.6 (5.5)	6.2 (4.8)	5.0 (4.1)
reality	4.0 (3.4)	5.0 (5.2)	4.2 (5.0)	4.6 (4.1)	4.8 (4.6)	5.6 (3.2)¥

* Sign < 0.05; ¥ sign < 0.01

Table 4 gives an overview of the session times and the mean length of the answers. Client 1 and client 2 seem to have used different avoidance strategies. Where client 1 gave relative shorter answers, client 2 gave relative longer answers compared to participants in the experiment that received the same dialogue – speech response condition. With longer answers client 2 might have tried to control the conversation. His session times were also significant longer. Client 1 on the other hand, avoided watching the screen especially in the presentation phase. The other client also often watched the presentation sheet instead of the screen. Both could have been signs of avoidance behaviour. Interestingly, the participants in the experiment had also commented on the intense stare of the avatars when it was the participant's turn to talk. The avatars eye gaze might therefore be a phobic stressor, an element also controlled in other VRET environment for the treatment of social phobia (Grillon, et al., 2006).

Table 4: Session time and individual response time of clients and experiment participants in same dialogue – speech response condition.

		Session time in seconds		Individual response time in seconds	
Dialogue	**Speech response**	**Experiment *M*(*SD*)**	**Client**	**Experiment *M*(*SD*)**	**Client *M***
Client 1					
Democracy	Speech detection	423(66)	361	6.7(2.5)	4.4
Dogs	Human control	403(89)	385	8.6(4.4)	7.5
Penguins	Speech recognizer	587(49)	560	7.6(2.0)	4.7*
Client 2					
Penguins	Speech detection	389(71)	490*	4.8(2.5)	9.2¥
Democracy	Human control	690(292)	1191 ¥	15(11.7)	26.2
Dogs	Speech recognizer	427(65)	553 ¥	8.8(2.9)	13.0*

* Sign < 0.05; ¥ sign < 0.01

During the human control conditions, clients were also asked to rate their anxiety on the Subjective Unit of Discomfort (SUD) scale (Wolpe, 1958). Although the scores are relatively low, the SUD score of both clients went up in the Q&A phase (Figure 8). Both clients were also unable in the Turing test to point out the dialogue that was controlled by the therapist. Both rated the speech recognition (score 6) as most likely to be human controlled and the speech detection (score 2) condition as the most likely computer controlled, with the human control (score 4) being rated at the middle of the scale. In the session with the speech recognizer, client 2 was for the first time interrupted by the computer in the middle of his sentence. He suspected that the therapist had done this on purpose, which might therefore have influenced his rating of the Turing test. Both clients were enthusiastic about the program after the second session and they saw its uses for therapy. Client 2 afterwards felt more confident about the way he presented.

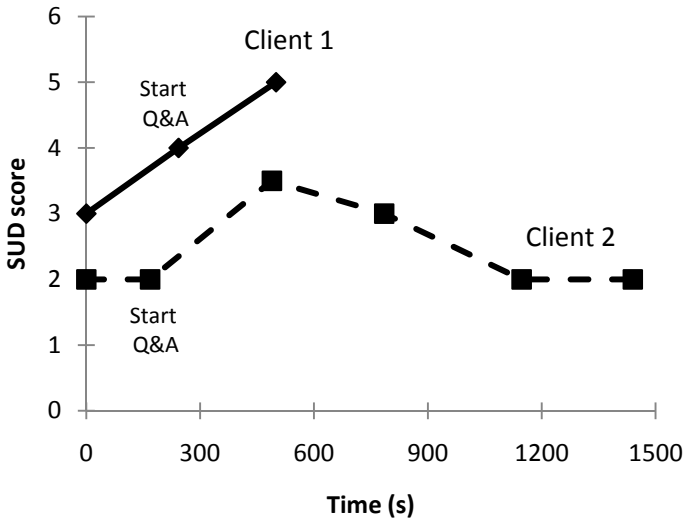

Figure 8: SUD score in the human control speech response condition.

5.3. Results Therapist

Observing the therapist it was apparent that she was more able to monitor the patient in the automatic response conditions than in the human control condition. In the latter, much of her attention focussed on selecting avatar responses. Instead of selecting the most fitting response, she often tried to select a response that would provoke the client. In the automatic speech response condition, the computer at points interrupted the clients in the middle of the sentence. This for her was very interesting. She saw it as a potential phobic stressor, and suggested to include this computer behaviour on purpose. In addition she saw this also as a strategy to respond to patients that talk for a long time as avoidance behaviour. She also wanted to have more control of the length of the presentation phase, being able to cut off the phase and move to the Q&A phase. She also preferred the order of phases not to be fixed. For some patients the Q&A phase could be a kind of introduction to the subject, with the presentation phase being the more anxiety provoking part, especially for patients with fear of presenting for a group. Also letting the patient stand instead of sit behind a desk while presenting without a bullet point sheet, she considered might be more anxiety provoking. Having more control over the type of avatars and the number attending the presentation she thought would also be an effective way to control the exposure. The intense eye gaze of the avatars might have triggered avoidance behaviours such as looking away by the clients, an observation also made by others when studying eye tracking behaviour of social phobic patients (Grillon, Riquier, & Thalman, 2007). The therapist therefore saw as future option the ability of the avatar to respond on this kind of behaviour. Overall the therapist was enthusiastic about the system and saw a future of using such a system for homework assignments.

6. CONCLUSION AND DISCUSSION

Manual speech response did not seem to outperform all automatic speech response techniques both for non-patients and phobic patients on factors such as presence, dialogue flow, discussion satisfaction, dialogue reality, and avatar interruption. The exception was, however, the ability to create the feeling that avatars are really listening. Here manual control seems still superior. Still, both non-patients and phobic-patients were on average unable to distinguish manual control from each the other types of automatic control. The benefit of automatic control was observed clearly on the therapist side, reducing system workload demands placed upon the therapists, and thereby allowing the therapist to devote more attention towards monitoring the patient.

The study also had a number of limitations that should be considered. First, to limit the range of potential patients' responses, the dialogues were designed with the computer taking the lead. However, some social phobic patients might also need exposure to situations where they have to take the lead. Second, the results of the speech recognizer condition might depend on the quality of the speech recognizer software used in this study. Other packages might give other results. Third, the social setting only focussed on public speaking, where social phobic patients might also fear other social situations. Exploring the techniques in other virtual scenes seems therefore interesting, especially whether they provide room for extensive exposure to conversations. Finally, treatment response was not explored. The case study only focussed on two therapy sessions. For an actual treatment more sessions seems needed, raising also the questions whether multiple dialogues would be required or even multiple social scenes.

The study also points out a number of potential social phobic stressors, such as: (1) avatar eye gaze directly towards or away from the patient; (2) responding on patient avoidance behaviour when looking away from the avatars; and (3) interrupting the patients when they are talking too long as avoidance behaviour. When varied, they might allow control of anxiety provoking elements in the exposure. Combining these with automatic speech response techniques and automatic anxiety measure instruments might reduce therapist workload demands even further, maybe even to a point where direct continuously therapist monitoring might not always be needed. This would open up options such as homework assignments, or a therapist monitoring multiple VRET sessions simultaneously. In the last case, an intelligent software agent in the VRET platform might support the therapist to cope with situations were multiple patients require attention simultaneously (Paping, Brinkman, & van der Mast, 2010). This study also demonstrates that therapist and patient might not have to be in the same room, as during the sessions they were in separate rooms communicating over a computer network. Still, for any of these technological solutions, it will remain essential to study them empirically,

demonstrating they provide at least a similar level of exposure with a reduced therapist workload. The results collected in this study seem to suggest this for automatic speech response techniques used in a VRET system to treat patient suffering from public speaking anxiety.

REFERENCES

Anderson, P., Rothbaum, B. O., & Hodges, L. E. (2003). Virtual Reality Exposure in the Treatment of Social Anxiety. *Cognitive and Behavioral Practice, 10*(3), 240-247.

Araki, M., & Doshita, S. (1995). Cooperative Spoken Dialogue Model Using Bayesian Network and Event Hierarchy. *Ieice Transactions on Information and Systems, E78d*(6), 629-635.

Brinkman, W.-P., Sandino, G., & Van der Mast, C. A. P. G. (2009). *Field Observations of Therapists Conducting Virtual Reality Exposure Treatment for the Fear of Flying*. Paper presented at the ECCE2009.

Brinkman, W.-P., van der Mast, C. A. P. G., & de Vliegher, D. (2008). *Virtual Reality Exposure Therapy for Social Phobia: a Pilot Study in Evoking Fear in a Virtual World*. Paper presented at the HCI2008 workshop - HCI for technology enhanced learning.

Brinkman, W.-P., Van der Mast, C. A. P. G., Sandino, G., Gunawan, L. T., & Emmelkamp, P. (2010). The therapist user interface of a virtual reality exposure therapy system in the treatment of fear of flying. *Interacting with Computers, 22*(4), 299-310.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H., et al. (1999). *Embodiment in Conversational Interfaces: Rea*. Paper presented at the CHI'99.

Galvao, A. M., Barros, F. A., Neves, A. M. M., & Ramalho, G. L. (2004). Adding Personality to Chatterbots Using the Persona-AIML Architecture. *LNAI 3315*, 963-973.

Grillon, H., Riquier, F., Herbelin, B., & Thalmann, D. (2006). *Use of Virtual Reality as Therapeutic Tool for Behavioural Exposure in the Ambit of Social Anxiety Disorder Treatment*. Paper presented at the Proceedings of the 6th International Conference on Disability, Virtual Reality and Associated Technology.

Grillon, H., Riquier, F., & Thalman, D. (2007). Eye-tracking as diagnosis and assessment tool for social phobia. *Virtual rehabilitation*, 138 - 145.

Herbelin, B. (2005). *Virtual Reality Exposure Therapy for Social Phobia.* Ecole Polytechnique Federale de Lausanne.

Hutchens, J. L., & Alder, M. D. (1999). Introducing MegaHAL. *Workshop on Human Computer Conversation*.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition* (2 ed.): Upper Saddle River, N.J.: Pearson Prentice Hall.

Katzelnick, D. J., Kobak, K. A., DeLeire, T., Henk, H. J., Greist, J. H., Davidson, J. R. T., et al. (2001). Impact of generalized social anxiety disorder in managed care. [Article]. *American Journal of Psychiatry, 158*(12), 1999-2007.

Kessler, R. C. (2003). The impairments caused by social phobia in the general population: implications for intervention. [Proceedings Paper]. *Acta Psychiatrica Scandinavica, 108*, 19-27.

Kessler, R. C., McGonagle, K. A., Zhao, S. Y., Nelson, C. B., Hughes, M., Eshleman, S., et al. (1994). Lifetime and 12-Month Prevalence of DSM-III-R Psychiatric-Disorders in the United-States - Results from the National-Comorbidity-Survey. [Article]. *Archives of General Psychiatry, 51*(1), 8-19.

Klinger, E., Bouchard, S., Legeron, P., Roy, S., Lauer, F., Chemin, I., et al. (2005). Virtual Reality Therapy Versus Cognitive Behavior Therapy for Social Phobia: A Preliminary Controlled Study. *CyberPsychology & Behavior, 8*(1), 76-88.

Klinger, E., Légeron, P., Roy, S., Chemin, I., Lauer, F., & Nugues, P. (2004). Virtual Reality Exposure in the Treatment of Social Phobia. *Studies in Health Technology and Informatics, 99*, 91-119.

Li, Y., Zhang, T., & Levinson, S. E. (2000). Word concept model for intelligent dialogue agents. *Text, Speech and Dialogue, Proceedings, 1902*, 445-449.

The Loebner Prize in Artificial Intelligence (1991). Retrieved 19-11-2008, 2008, from http://www.loebner.net/Prizef/loebner-prize.html

Martin, H. V., Botella, C., García-Palacios, A., & Osma, J. (2007). Virtual Reality Exposure in the Treatment of Panic Disorder With Agoraphobia: A Case Study. *Association for Behavioral and Cognitive Therapies, 14*(1), 58-69.

McBreen, H. M., & Jack, M. A. (2001). Evaluating humanoid synthetic agents in e-retail applications. [Article]. *Ieee Transactions on Systems Man and Cybernetics Part a-Systems and Humans, 31*(5), 394-405.

McTear, M. F. (2002). Spoken Dialogue Technology: Enabling the Conversational User Interface. *ACM Computing Surveys, 34*(1), 90-169.

McTear, M. F., O'Neill, I., Hanna, P., & Liu, X. (2005). Handling Errors and Determining Confirmation Strategies -An Object-based Approach. *Speech Communication, 45*(3), 249-269.

Paping, C., Brinkman, W.-P., & van der Mast, C. (2010). An explorative study into tele-delivered multi-patient virtual reality exposure therapy system. In K. Wiederhold (Ed.), *Coping with posttraumatic stress disorder in returning troops: Wounds of War II* (pp. 203-219): IOS press.

Paul, G. L. (1966). *Insight VS. Desensitization in Psychotherapy*. Stangord, CA: Standfort University Press,.

Pertaub, D. P., Slater, M., & Barker, C. (2001). *An Experiment on Fear of Public Speaking in Virtual Reality*. Paper presented at the Conference on Medicine Meets Virtual Reality 2001.

Robillard, G., Bouchard, S., Dumoulin, S., Guitard, T., & Klinger, E. (2010). Using virtual humans to alleviate social anxiety: preliminary report from a comparative outcome study. In B. Wiederhold, G. Riva & S. I. Kim (Eds.), *Annual review of cybertherapy and telemedicine 2010* (pp. 57-60). Amsterdam, The Netherlands: IOS Press.

Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The experience of presence: factor analytic insights. *Presence, 10*(3), 266-281.

Slater, M., Pertaub, D. P., & Steed, A. (1999). Public Speaking in Virtual Reality: Facing an Audience of Avatars. [Editorial Material]. *IEEE Computer Graphics and Applications, 19*(2), 6-9.

ter Heijden, N., Qu, C., Wiggers, P., & Brinkman, W. P. (in press). *Developing a dialogue editor to script interaction between virtual characters and social phobic patients*. Paper presented at the ECCE2010 Workshop on cognitive engineering for technology in mental health care and rehabilitation.

Wallace, R. (2001). Artificial Intelligence Markup Language Retrieved 12-12-08, 2008, from http://www.alicebot.org/TR/2001/WD-aiml/

Wallence, R. S. (2009). The anatomy of A.L.I.C.E. In R. Epstein, G. Roberts & G. Beber (Eds.), *Parsing the turing test* (pp. 181-210): Springer.

Weizenba, J. (1966). ELIZA - A Computer Program for Study of Natural Language Communication between Man and Machine. [Article]. *Communications of the Acm, 9*(1), 36-&.

Wiederhold, B. K., & Wiederhold, M. D. (2005). *Virtual reality therapy for anxiety disorders : advances in evaluation and treatment* (1 ed.): DC: American Psychological Association.

Wolpe, J. (1958). *Psychotherapy by Reciprocal Inhibition*. California: Stanford University Press.

Wrzesien, M., Burkhardt, J. M., Alcañiz Raya, M., Botella, C., & Bretón López, J. M. (2010). Analysis of distributed-collaborative activity during augmented reality exposure therapy for cockroach phobia. In B. K. Wiederhold, G. Riva & S. I. Kim (Eds.), *Annual Review of Cybertherapy and Telemedicine 2010 - Advanced Technologies in Behavioral, Social and Neurosciences* (pp. 134-139). Amsterdam, The Netherlands: IOS Press.

APPENDIX

Table 5: Items of the Dialogue Experience Questionnaire included in final analysis.

No	Statement (original Dutch text) ¥
Flow: dialogue speed	
1*	The discussion partners needed a long time to think (De gesprekpartners moesten lang na denken)
2*	The discussion partners often went quiet (De gesprekpartners lieten vaak stiltes vallen)
3*	On occasions I had to wait long for a reaction of the discussion partners (Ik moest soms lang wachten op een reactie van de gesprekpartners)
4*	The conversation did not run smoothly (De conversatie verliep stroef)
Flow: interruption	
1	I was always able to finish (Ik kon altijd volledig uitpraten)
2*	On occasions I was unable to tell everything that I would like to have told (Soms kon ik niet alles vertellen wat ik wilde vertellen).
3*	On occasions the discussion partners talked before their turn (De gesprekpartners praten soms voor hun beurt)
4*	On occasions, the discussion partners started talking while I was talking (De gesprekpartners praten soms door me heen)
5	I got enough time from the discussion partners to explain everything calmly (Ik kreeg genoeg tijd van de gespreksпartners om alles rustig te vertellen)
Flow: correctness locally	
1	The discussion partners addressed my answers (De gesprekpartners gingen in op mijn antwoord)
2	The discussion partner responded to my answers (De gesprekpartners reageerde op mijn antwoorden)
3	I got the feeling that the discussion partners understood my answers (Ik had het gevoel dat de gesprekpartners mijn antwoorden begrepen)
4	The questions had a logical order (De vragen volgde een logisch vervolg)
Flow: correctness globally	
1*	The discussion partners rambled (De gesprekpartners sprongen van de hak op de tak)

2* On occasion, the discussion partners asked things I had already answered. (Soms vroegen de gesprekpartners dingen die ik al beantwoord had)

3* On occasion, I had to repeat myself (Ik moest me zelf soms herhalen)

4* I did not always understood why things were asked (Ik snapte niet altijd waarom iets gevraagd werd)

Interaction: involvement

1 The discussion partners did listen to my answers (De gesprekpartners luisterde naar mijn antwoord)

2* The discussion partners acted detached (De gesprekpartners reageerden afstandelijk)

3 The discussion partners were interested in my answers (De gesprekpartners waren geïnteresseerd in mijn antwoorden)

4 I got a feeling that I was listened to (Ik kreeg het gevoel dat er naar me werd geluisterd)

5* On occasion, it felt like the discussion partners were not interested in me (Soms voelde het of de gesprekpartners niet geïnteresseerd in mij waren)

Interaction: discussion satifcation

1 The discussion was pleasant (Het gesprek verliep prettig)

2 I was left with a good feeling about the discussion (Ik heb een goed gevoel aan het gesprek overgehouden)

3 I have experienced the question round as pleasant (Ik heb de vragen ronde als prettig ervaren)

4* The questions of the discussion partners made me nervous (Ik werd nerveus van de vragen die de gesprekpartners stelden)

5* I did not feel comfortable during the question round (Ik voelde me niet op mijn gemak tijdens de vragen ronde)

Interaction: reality

1 I got the feeling that this type of conversation could happen in real life (Ik heb het gevoel dat dit soort gesprek ook in het echt kan voorkomen)

2 The discussion partners seemed natural (De gesprekpartners kwamen natuurlijk over)

3 I can imagine that this could happen to me in real life (Ik kan me voorstellen dat ik dit in het echt ook mee zou kunnen maken)

4* I had to adjust myself to the discussion partners (Ik moest mij zelf aanpassen aan de gesprekpartners)

5 The discussion partners seemed realistic (De gesprekpartners kwamen realistisch over)

¥Rated on 7-point Likert scale ranging from Strongly disagree to Strongly agree (helemaal oneens – helemaal eens); * Score revered