Virtual Conversation

Enabling conversation between a virtual human and a patient with social phobia

Niels ter Heijden
TU Delft
Version 1.0

Contents

1. Introduction 3
   1.1 Problems 4
   1.2 Research goal 4
   1.3 Specifications 5
2. Listening and observing 6
   2.1 Talking to a wall 6
   2.2 Problematic speech recognition 7
   2.3 Increasing recognition robustness 8
   2.4 Error recovery 9
   2.5 Side conversation 9
   2.6 Speech recognition packages 10
3. Understanding 11
   3.1 Understanding text 11
   3.2 AIML 12
   3.3 Scripting 12
   3.4 Stroop 13
   3.5 Guiding the conversation 13
4. Responding 15
   4.1 Anxiety provoking 15
   4.2 Emotional bots 15
   4.3 Text to speech 16
   4.4 Mouth to eye 17
5. Conclusion 18
6. Abbreviations 20
7. References 21

1. Introduction

Virtual reality (VR) has a wide range of uses from simulator training to treatment of certain phobias. In the last decade the use of virtual reality in the treatment of certain phobias, for example fear of flying or fear of heights; have successfully been developed and implemented. An area where a lot of research still needs to be done is the treatment of social phobias with the help of VR. Social phobias are some of the most common phobias [4, 5] that people suffer from and often it persists during a patients entire lifetime if left untreated. The effect of social phobias can be severe for example not taking or completing certain classes if it involves presenting in front of a group, or even not applying for or taking a job advancement [6, 7].

Virtual reality is useful for treating patients with phobias [3, 8-12] because it gives therapists a world they can completely control and does not require physically visiting certain locations. This can include places like an airport, a tall building or a podium. Because of the complete control the therapist has, the anxiety level of the patient can gradually be altered until the patient would feel confident to accomplish the task in real life. Social situations in VR are somewhat harder to accomplish and therefore the use of VR in treatment is very limited. Fear of flying for example already uses VR for regular treatment whereas social phobia is still limited to lab experiments. These experiments have already shown that people feel the same anxiety in a virtual environment [13, 14] as in en vivo when exposed to situations such as presenting in front of a group or interacting with an unknown virtual person. Up until the present interaction by the virtual humans (VH) with the patient was limited to automated movements or basic verbal reactions initiated by the therapist [15]. Also the realism of the situations is still very limited whereas for fear of flying the environment is very detailed, even, for example, the cup holders are shown in the plane. On the other hand in the social phobia environment the 3D virtual humans are very block-like and rooms are very basic and pretty much empty (see Figure 1).

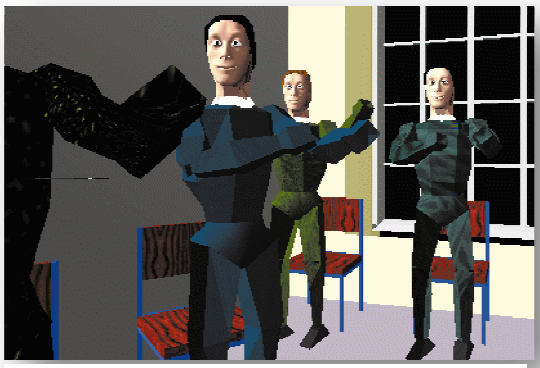

Figure 1: applauding audience from [2]

Experiments using the virtual environment (VE) for social treatment often only used avatars that sat in place and showed some sort of standard programmed reaction or behavior. A few examples are turning around and looking to the patient if he comes near or booing or cheering when "listening" to a presentation given by the patient. Only a few experiments put the avatar in a situation that it really has to respond to the patient. In all situations the therapist had to control all actions or responses making the therapist more a puppet master then a doctor.

1.1 Problems

Virtual environments for treating social phobias need social actors while situations like presenting in front of a group can be handled by relatively simple avatars that only have to show scripted movements. In situations where the patient needs to interact with a virtual human (VH) a more complete social interaction is needed. This could be achieved by the therapist however this involves a lot of work and thus distracts the therapist from the treatment. Another problem is the pause between the patient uttering a sentence and the therapist choosing or typing the response. This might slow down the conversation or breakdown the feeling of presence [16] felt by the patient. Most of these problems might be solvable with a degree of automation. A few problems that make it hard to automate are the lack of standard input devices in a VE. There is no keyboard for the patient to type its interactions and therefore the system needs to listen and try to understand the patient. Another option is that the patient can be limited in their responses by choosing options with their navigation tracker (joystick). While a virtual keyboard or selected responses will most likely break the feeling of presence a microphone might increase the feeling of presence. Listening (via microphone) is a good option because this is the modality also used in a conversation between two humans.

After listening to what the patient said one must understand and select the proper response that avatar needs to give. For this problem no easy and complete solutions exist. Analyzing and understanding sentences is a field of research still in its starting phase and is not yet at a stage that it can generate the required level of presence needed for the treatment. Advances are being made but most research focuses on deceiving the user into believing there is some sort of intelligent and understanding on the computers side, while this is not the case. Attempts to make a computer "understand" what is being said never made it to a usable state [17]. Most attempts were limited to simple sentences or filling in query requirements to access a database with information. At this moment it is impossible for an automated avatar to participate in an intelligent conversation with a human. For this reason some limitations or conversational guidance is needed for the computer to be able to take over this task from the therapist.

Finally when a suitable response is selected it needs to be conveyed to the patient in a natural way. Speech seems to be the most obvious way to do so, but is not without its own limitations. Sound needs to be synchronous with lip and head movements and possibly certain gestures. The intonation of the spoken words also needs to be correct to convey certain meanings. For example the simple sentence "Oh great" can have multiple meanings just by the way it is spoken.

1.2 Research goal

As stated before the research area of social phobia treatments using VR is still fairly new and unexplored. Research has shown that VR can be helpful and even an effective treatment with fairly basic virtual environments and avatars. As for the real social aspect, communicating with others, is still largely left aside because of technical difficulties and the amount of active involvement needed from the therapist to operate the system. The research goal is therefore to automate the social aspect, communicating with the patient, in a way that frees up time for the therapist and still gives him "absolute" control over the conversation. Communication is done on multiple levels between humans but the verbal component is one of the most conscious of the channels used and therefore will be the initial focus. This presents three main problem areas where solutions need to be found. The first problem is listening and converting spoken text into computer comprehensible code. After this the avatar needs to think up a response that is suitable to the situation. This can be done either by comprehending the sentence spoken by the patient or by trickery with preprogrammed responses. Finally the avatar needs to respond to the patient in a way that feels natural to humans and therefore does not break the feeling of presence.

How to convert speech to computer comprehensible code is being actively researched by many people. Many methods and techniques exist and no one at this point in time has found the holy grail, a program that can in any circumstance convert speech to text almost flawlessly, with or without speech training [18]. If this does not work correctly the automated avatar would fail from the start. In order to relieve this problem a bit some techniques have to be used to improve the recognition, for example by training the system or limiting the number of words or sentences that need to be recognized. To solve this problem an already existing and trained package will be used. This means workarounds for the problems and limitations of this package need to be found and implemented.

The main focus of the project will be on comprehending the patients and selecting the appropriate response to return to the patient. To realize this a comparison needs to be made between different techniques to see which one gives the best or most feeling of presence to the patient and also gives the therapist the means to influence the anxiety level created. The chosen method also needs to be able to work with and disguise the weak points of the speech recognition and speech synthesis systems. The total package should deliver a socially interacting virtual human that responds realistically to the patient and therefore creates a high feeling of presence in the virtual environment so that it can be used for treating social anxieties.

1.3 Specifications

Making an initial system that would work in all cases and everywhere in the world would be next to impossible and therefore the end system needs to be somewhat limited to a domain and test group. As stated before the main goal is to relieve the side tasks of the therapist by automating the avatar in its conversational skill while still giving the therapist control over the anxiety levels generated by the system. The first important limitation is the test group since speech recognition success rates crumble to almost nothing when someone is trying to speak a non native language [19]. It would be unwise to use an English system with at least a majority of the initial patient group being Dutch. This means that the most advanced and latest speech recognizers that are not ported or trained on Dutch are unusable. The speech recognizer itself need to be easy to integrate into the other parts of the system and needs to be able to figure out when someone is speaking. It also needs to be able to discover the sentences said from a limited amount of possible sentences or able to locate and identify certain keywords in a sentence.

The dialog part cannot be totally free but needs to be limited or even guided in a domain to make “intelligible” responses possible. The Loebner contest [20] has shown that at this point in time a computer system responding like a human is not possible yet. The big question is how to guide or limit the conversation. To rate certain possibilities the conversation freedom factor might be a nice guideline.

To avoid spending most of the time making a virtual representation of a human with all the movements and details needed to make it convincing a prebuilt package named Vizard will be used. This way most of the time can be spend on the actual conversational aspect of the problem. The same goes for the difficult task of speech recognition where the development and training of the recognizer would take too much time and therefore a prebuilt and trained package will be used.

2. Listening and observing

The first avatars that were used in social phobia treatment or studies did not use a lot of automated animation the therapist was task with controlling all actions of the computer characters [1, 21, 22]. A reason why automation might not have been implemented was the lack of speech recognition because this is the only real input channel the patient used during the treatment. Recognizing speech is a difficult task for a computer and even the most recent speech recognizers of today do not have a human level of recognition or error rate [23].

This chapter will elaborate a bit about avatars and techniques already used in test treatments in Talking to a wall (2.1) to show what the state-of-the-art is and what the limitations or problems are after which the limitations and problems with speech recognition are described in 2.2. Of course there are things that can be done to decrease the recognizer error rate (2.3) but errors will occur and need to be handled well by the program (2.4). Finally there should also be an escape from the virtual world so that the patient can talk to the therapist without the system running wild (2.5). Because speech recognition is not the main focus of this project an already build package will be used. What the selection criteria are and what package is most suited will be discussed in 2.6.

2.1 Talking to a wall

Avatars used in recent research for social phobia treatments were almost completely controlled by the therapist with only a few automated and looped movements done by the computer [14]. This also meant that the therapist was limited to the preprogrammed actions and also needed to take into account a lag between pushing a button for an action and the actual action itself [14]. This meant that timing often was wrong with the effect that avatars began cheering in mid sentences of the speaker [14]. Also the severity of the response was fixed meaning that in the experiments there were three options a neutral control audience, a happy cheerful or interested audience and finally the angry, unhappy or rude audience [14, 24, 25]. Often switching between these three modes was not even possible during the running session. A second option to vary the anxiety level was by changing the situation such as talking to an empty room or podium closed off with curtains or talking to an audience of avatars [21, 26]. A audience was used for treating anxiety when giving a speech for a group of people (see Figure 2) [8, 11, 21, 24, 26] in this situation the virtual audience

Figure 2: Conference room from [3]

does not have to understand the speech and ask questions in the end they only need to show some basic positive or negative behavior completely controlled by the therapist. No real speech recognition was needed in these cases but does basically mean the patient was talking to a wall be it a wall consisting out of non- or badly responding avatars.

The virtual audience case is one of the simpler situations to create and is the focus of most early research only later on a more direct interaction between avatar and patient was tried out. Firstly simple by just tracking movement of the patient if he came virtually near the VH. Responses from the VH was limited to turning and looking to the patient [13] or in some cases uttering some unrecognizable words that sounded annoyed or at least had to make clear he was entering the virtual personal space [13]. Uttering something unrecognizable was done on purpose because the researches where afraid that uttering something the user could understand would engage him into conversation and no understanding or even recognition of speech was implemented. So the avatars in the experiment by Garau [13] where unable to respond verbally. A step more advance was a situation where the patient had to interact with VH in a bar (see Figure 3: virtual bar from [1]) [1] where different VH had different attitudes toward the patient. Not all avatars where programmed with responses therefore if the patient moved to and talked to the wrong avatars he would get no response. The most obvious avatar to talk to, the barman, had a series of responses to redirect the user towards the correct group of avatars that were programmed with responses. What the precise responses and utterances wore and how these were generated was not elaborated in the paper. But more than likely all were pre recorded sentences released with a push of a button by the researcher. Meaning also these avatars had no speech recognition or understanding in them and also the number of responses had to be limited.

Figure 3: virtual bar from [1]

The lack of examples of automated virtual humans would almost suggest it is too hard or impossible to accomplish. However it is not the speech recognition holding the development back because examples of successful speech recognition are available [27]. Multiple speech recognition programs are available for control of systems or computers and also to convert dictated speech into text. Therefore it should possible to make an avatar that can listen to real human speech as long as some limitations are used to overcome the speech recognition weak points. These limitations will be further worked out later in this chapter.

2.2 Problematic speech recognition

To work out what the limitations are for speech recognition software available today the first step would be to work out what situations gives speech recognizers problems. The most obvious factor that gives speech recognizers problems is background noise [28]. Even for well trained listeners a loud audio source can give enough interference making listening to a conversation impossible. For

example try to talk to each other during a rock concert or an airplane taking off. Speech recognizers (SR) are even more sensitive to background noise than humans even a hum of a computer fan near the microphone can be enough to make recognition very hard. With social phobia treatments the environment is very controllable limiting or even eliminating background noise is in the realm of possibilities. Therefore the problem of background noise is not a big problem here. If certain background sounds are needed to help the feeling of presence; headphones can be used to deliver it to the patient without interfering with the speech recognition.

Next SR systems also have problems with two different speech sources where humans can repress the second source if needed or at least distinguish between the two and follow one of the conversations it is very hard for SR to do so. A solution could be implementing voice recognition so that the SR software knows who is talking but this still does not solve the problem of two sources talking at the same time because the SR software still will try to view the combined sound as one source. In the situation of treatment multiple sources talking at the same time would not happen too often since only the patient and therapist are in the room. The next problem when does the patient speak to the system and when does his speech end [29] need to be solved. Another problem is the patient uttering none words like “uhmm”, “ehh”, ect could interfere with correct recognition. Speech recognition suffers a lot if the speaker is not speaking in its native tongue; reduction of 50% of the recognition is possible [19]. This means in this particular case that Multilanguage recognition needs to be used or one language needs to be selected else a lower recognition value needs to be accepted and maybe compensated [30]. Also a problem is the fact that SR is not perfect therefore it might happen that not everything is recognized correctly. Therefore the system has to ask if the patient could repeat what he said. This could be a tiresome and time wasting business and is not part of treating a patient. It also could have an adverse effect on the feeling of presence of the patient. Therefore the robustness of the SR software needs to be increased especially because patient will not have a lot of time to training the system on their speech because time spend training the system is less time spend on the actual therapy. Increasing robustness is an important task and therefore will be worked out later in this chapter.

Because of all these weaknesses in recognition it might be needed to build a failsafe into the program so that the therapist can intervene if recognition completely fails in a certain situation. Of course this should be avoided as much as possible because it will undo the automation that has been provided.

2.3 Increasing recognition robustness

Speech recognizers have a hard time correctly processing information in certain situations therefore to increase the robustness of the recognition certain strategies can be applied. Firstly reducing the background noise increases the correct recognition rate. But also limiting the vocabulary that can be uttered or need to be recognized at a certain moment in time helps increase recognition rate. Also a small recognition training session at the start would help the system. Another factor is the hardware used because it has an influence on the recognizer because it influences the quality of the inputted sound and that needs to be as good as possible. A microphone somewhere in the room might also pick up to much background noise or if the patient does not speak loud enough register nothing. Because the patient already uses a head mounted display (HMD) for the visual representation of the world it would be foolish not to (miss)use it to place a microphone as close to the patient mouth as possible thereby reducing background noise interference. Finally avoiding the patients having to speak something else than their native tongue can have a substantial influence on the recognition rate. When the system finds its way in real treatments the treatment would also be given in the patient native tongue. So the system should use a speech recognizer trained for this language.

2.4 Error recovery

With the use of speech recognition errors will occur as it is unavoidable with the state of the art of speech recognition [31]. Therefore some error recovery needs to be implemented to maintain the feeling of presence when the system gets stuck on a set wrongly understood or none understood words. One solution has already be mentioned the therapist could intervene if the system does not know what to do anymore. Of course trying to avoid all the pit falls in speech recognition and making it as robust as possible will help reduce the number of errors.

Later on a technique called frame based dialog systems [32] will be elaborated. Here another safety measure for wrongly understood speech is given [27, 31]. The technique revolves around the system trying to double check everything important the user has said. This could be done directly or even better by indirect means. For example if the user names a hotel and the system is not 100% certain it understood the right name of the hotel it could ask the user to repeat the name or simply wait or ask for the area where the hotel should be found and check against a database of hotel names in that area and so indirectly verify if he heard the correct name. Of course the easiest solution could be to simulate human responses too badly heard speech something in the trend of "hee?" or "sorry?" as long as it is not too long and overly friendly. Sentences like "sorry I did not hear that right could you please repeat it?" is not something humans are used to in everyday life [33] and therefore would feel very unnatural to the patient if the system would use it.

2.5 Side conversation

The system is not the only social actor during a therapy with which a patient might want to interact (figure 4). Therefore it would be useful to take into account that the patient might want to say some things to the therapist without the system trying to interpret it for its own use. This because it might have unexpected influences on the virtual world and shatter the feeling of presence if the avatars would react on "sorry I really cannot do this" with "oke 10 apples it will be". This problem is unique for this domain because it combines two separate pieces of technology namely speech recognition and treatment in a virtual environment. Speech recognition research avoids multiple speech sources because of the problems mentioned before. As treatments in the virtual environment did not use automated avatars yet a new solutions need to be found. Possible solutions to this problem could be as simple like an on/off switch for the microphone that can be used by the therapist or the patient. A somewhat more advanced solution could be certain keywords that would switch the system on and off that needs to be uttered by the patient before he starts to talk to the therapist. Problem with both these solutions is that it requires a certain action before the patient start talking to the therapist these actions could easily be forgotten if the patient is in a panicky state or is just making a quick remark.

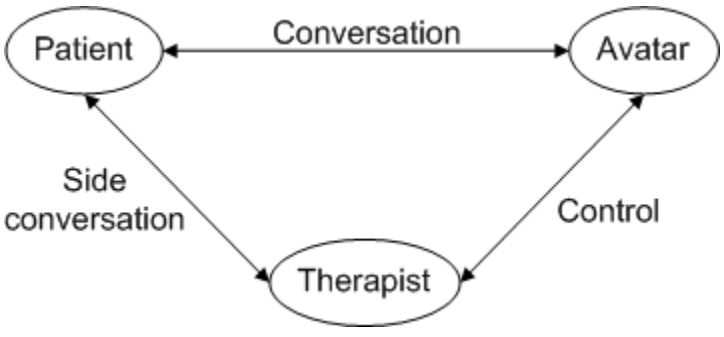

Figure 4: Communication/Control schema

It would be best if the system could figure out that certain sentences are completely wrong for the conversational domain and therefore must be meant for someone else or just background noise. Or if the systems detect certain keywords like the therapist name it could consider the uttered sentence as irrelevant for the program. Another modality that could be used to detect if the patient is talking to the system or the therapist is checking the head direction. Often

people look to the person they are talking to and therefore a patient looking away from the avatar might not be talking to it and therefore the sentences could be safely ignored by the system. All this has the advantage that it does not require any specific action from the patient to tell the system he is not talking to it.

2.6 Speech recognition packages

The field of speech recognition is still actively evolving with a lot of research and new techniques popping up, but because almost everyone sees the importance of speech recognition [29, 34] for a multimodal user interface with computer systems it is not only left at theoretical papers. Companies and research groups have developed working speech recognition systems that could be used in a project like this. Big names in the speech recognition business are Nuance (Dragon) and Microsoft that sell their product to consumers and in Microsoft case even provide api's for programmers to use. This project has certain limitations stated earlier that needs to be taken into account firstly the speech recognition needs to work with Dutch because of this the Microsoft version already is eliminated as a viable option. Secondly the program needs to be semi open so that it can be optimized for the situation because of this Nuance product is not a viable option anymore because of its close source nature.

Alternative to commercial product are the research packages like "Sonic" and "HTK" that are more freely available but designed to conduct research in the field of speech recognition. Because of this they are somewhat more complicated to set up and are not coming with standard trained databases or user friendly training programs. Once again because the speech recognition needs to be in Dutch and completely retraining of the system would be needed what requires extensive time and effort. Luckily this faculty is not a stranger in the field of speech recognition and an already trained system is available for Sonic. Sonic also provides all the needed options like real time recognition and keyword like recognition with standard non word filtering and silence detection. But because it's a package for research into speech recognition it comes with a lot of unneeded functions and flexibility. Sonic also only runs on Linux and there is no real support or active community using or maintaining the package.

3. Understanding

When the computer program has received the "correct" input from the speech recognizer it has to do something useful with it. This could be called "understanding" what is said and act on it with an appropriate response. This understanding could be done by letting the computer analyze the sentence and dissect its words into grammatical categories and derive the subject and possible meaning of the sentence. A much more successful method is used by chatter bots in the Loebner contest [20]. Pattern matching is used to select a response for the input and skipping the whole understanding part [35-37]. A strategy to make understanding easier by guiding the conversation. This way the program can limit the domain of the topic and guess what will be the most likely response from the user. To guide a conversation a goal can be useful. A goal could be a dataset of information from the user.

The content of this topic will be structured in the following way. In understanding text (3.1) different techniques for analyzing the input will be discussed. Which is followed by a more detailed description of AIML (3.2) and techniques to make it easier for a program to have a successful discussion (3.3). The techniques to guide the conversation or to make it easier for the recognizer could even serve a therapeutic purpose this is discussed in the section on stroop (3.4). In 3.5 the actual techniques to guide a conversation are elaborated on.

3.1 Understanding text

Understanding something the patient said is not as trivial for a computer system to accomplish. There are two main ways to handle it.

The first technique is to try to give the system some sort of understanding about what is said. This can be done with logic and a good description and analysis of the sentence and words [38]. The meaning and subject from a sentence are extracted from it by locating the verbs and pronounces. This is then used to generate an appropriate response to the input sentence. This response is also generated in a way like the analysis is done by placing verbs and pronounces in the right spots. Things that need to be taken into account in the progress are plurals and subject for words like he, it or they. This gives the sentence its little details that are extremely important if the response of the system needs to be on the subject discussed. Systems that use this approach are often slow because of the huge databases needed and the difficult analysis of sentences [17]. Also the level of responses is not good enough yet to keep a conversation going. This is mostly because of the little inner meanings of sentences or the many exceptions on the rules people make in sentence they speak like words double meaning or sarcastic undertone.

The second technique that can be used to "understand" text is by generating a reaction database. In this database all responses are defined for certain input patterns. A way to design such a database is by using the Artificial Intelligence Markup Language (AIML) that is used by almost all successful internet chatter bots [35]. By just matching input from the user to a certain output sentence the understanding of the sentence is redirect back to a human. Only creating a database with enough responses is already a labor intensive task. Also this way of understanding text has the problem of the lack of context. The responses are predefined to a general input sentence this means the reaction might not be as topic specific or with a miss matched input string even complete wrong. Never the less these kinds of interaction bots are the most effective kind in the Loebner contest [20]. The effectiveness of such chatter bots increase when the range of the topic area discussed is limited. Also in certain roles they are more effective such as the role of a therapist or a schizophrenic person. Because such roles disguises the weaknesses a chatter bot has like no prior or very limited knowledge of the discussion topic before the last sentence, but also the sometimes erratic changes of topic because a default response sentence fired on a none patterned input sentence. Another

application where the simple chatter bots seem to do a acceptable job is in simple question and answer system where people can ask in a conversation like way certain questions at the system that then tries to produce the correct answer [39]. Here the chatter bot only needs information about the domain it works in and also the questions people will tend to ask will be fairly similar making the bot an automated FAQ system.

3.2 AIML

AIML [40] is a markup language used by the chatter bot community. It is based on XML and defines pattern and response sets. These sets are called categories and can consist out of something like "Hello *" → "Hi there how are you" where the "*" mark means anything may be placed here. The language defines more language specific symbols that can be used in the patterns but also the output line. Symbols in the output line are used to place user saved variables like name, address or gender. These variables start off with a default value and are filled in when the information comes up in the conversation. This can work as long as the user does not use obscure sentence constructions or ask the information before it is filled in [41]. So it can happen that the bot response with "Your name is UNKNOWN" if asked "Do you know my name".

The AIML language and the way it is used does not use any reasoning but just purely looks up the correct category and put out the response. Little adaptations are bots that change topic when the correct response (or input pattern) is unavailable. This means that the default category is used that defines a series of output sentences where one is picked out at random. There are adaptations to AIML to give it a bit more functionality [42]. So is persona-AIML [43] an adaptation that adds personality to the bots responses by keeping track of the bots state. This means that besides the input pattern also the bots state has to match persona-AIML defined states for a certain response. The added data in the database does mean a lot more work for the programmer but might be just like AIML automated or distributed over a lot of participants. The way the original AIML database started off was by hand filling certain responses while this means that the database remains semi consistent it also limits the number of input-response patterns. Soon the developers of AIML based chatter bots made them "learn" new patterns by releasing them on the internet and let the visitors chat with them and generate the new patterns. This increased the database significantly but also meant that the bots replies are less consistent. Another problem is that visitors tend to be abusive [37, 44-46] or at least are less restrained when talking to the bot with the effect that the bot also learned those patterns and responded in natura. This meant that the newly generated categories still needed some screening on these types of responses. It also shows that the chatter bots are not that advanced yet that they fool the user into believe they are not a computer ran program [47], and therefore the interaction is not limited by social standards.

3.3 Scripting

While AIML based chatter bots are fairly simple and their responses misses a lot of things like context and history they are without a doubt the most effective into deceiving people into believing that they are not a bot. Winning all prices for years in a row. Sadly this does not mean they are convincing enough to win the Loebner price. A way to make the chatter bots a bit more effective is by limiting the domain they talk about. This will mean that when the user deviates from this domain the answers given by the bot will make even less sense. To reduce this problem the users could be instructed to only talk about a certain domain and thereby limiting the number and kind of sentences they can give to the bot. This technique or limitation is not without its problems either because limiting the domain also means the extensive AIML database generated by the users cannot be used or all the responses not applicable to the domain should be filter out. Still the learned patterns will be limited and often not very specific or elaborate in the domain. The problem of the system is the unpredictability of the user and the extreme flexibility natural language gives to

generate sentences. If this flexibility is largely removed the chatter bots might function a lot better. This does mean the conversation is reduced to a partially scripted sequences but this might be acceptable in certain cases.

3.4 Stroop

Limiting the sentences a user or patient can say by giving him or her a limited set of options to choose from might even be beneficial for the treatment of the patient. By forcing them to say things they are uncomfortable with they are forced to confront that fear [48]. The therapist can also make sure that the patient cannot use situation avoiding things like very short sentences or minimal interaction. Finally it might also be possible to influence the level of anxiety that is generated by the patient by altering the things he needs to say that may or may not generate a higher level of fear in the patient.

Just using words to treat patient is not new, research done on the stroop paradigm by Masia [48] has show that patient reading certain words that they associate anxiety with helps treat this anxiety. Initially stroop showed that people have difficulty telling the color of a word like “red” or “green” when this word is written in a different color then its meaning. This effect then was tried on words other then color names like social anxiety words as “meeting”, “party” or “presentation” and it was shown in the paper of Masia that people with social anxiety disorders needed a longer time to say the color. Later the effect on the anxiety itself was research by Masia and shown that people that did this test over hundreds of words and a couple of days had a decrease in anxiety rating for social situations. So in effect confronting patient with words that associate negative emotions and putting this in a better perspective or positive emotion by the therapist would help the patient overcome this anxiety. Success in the experiments by Masia where limited to patients that first did not want to go to group therapy and after using the stroop tests did enroll.

This just shows that even forcing a patient to read and say certain words and sentences might already help them in their treatment. Supporting the conversation restricting or guiding techniques that could make the task of the chatter bot much easier and might make it possible for them to be convincing enough to humans to be seen as full social actor.

3.5 Guiding the conversation

Because open and free conversation between the patient and the virtual human is not possible with the limitations in speech recognition and understanding of these sentences by a computer, discussed earlier in this chapter and chapter 2, a more limited way of conversing is needed. Guiding the conversation in a scripted way removes a lot of the limitation and problems such as speech recognition. By limiting the options that the speech recognizer needs to recognize the error rate could become more manageable. Also the problems with understanding the sentences are reduced to choosing written down responses to the input sentence if using AIML. The only real variable might be the order or the position in the scripted conversation tree. The choices made by the patient still means that the tree could become massive so merging branches back to a main conversation line might be needed to make all things manageable. AIML might be very suited for this task because its conversation tree is never bigger than the last input or two and the response options it has to that. The whole dialog system would therefore consist out of groups of input sentence, possible responses and possible responses the user could give on that. Still careful planning is needed to avoid loops and make the conversation progress semi natural to an end state after which the scene is played out.

A alternative way to avoid making and planning a conversation tree and all its difficulties is by using a frame based dialog system [27]. These systems are mainly used in automated planning or

registering systems. An often used example is a system that can be called to book hotels and plan a night out to a theater [31]. Here the system has a clear goal gathering certain information from the user after the user initiated a domain like booking a hotel room. Frame based systems are build to gather the information necessary from the user either by analyzing what the user is telling the system or actively enquire for the information. This has the effect that the user is free to say things in what order he wants while the system still sticks to a limited domain and tries to direct the conversation in a certain direction. Frame based systems also uses tactics to verify information from the user without asking him to repeat everything multiple times. This will catch the errors made by the speech recognition system and makes sure that the conversation keeps moving forwards to the end goal [31].

This technique has the potential to be very useful in making a social actor in a virtual world by giving the patient a goal he needs to accomplish like buy new shoes and giving him absolute freedom in what he want to say and how he wants to accomplish the goal. While the system is still able to manage it all by only listening for certain keywords and trying to fill its own goal of gathering all needed points of information and by doing so also guides the conversation in a way that it keeps being comprehensible for the system. Also the verifying techniques could be used to lengthen the conversation somewhat and limit the responses like “could you repeat that” or “I did not understand” from the system by just ignoring the sentences or guessing the most likely thing that was said and later on in the conversation verifying it.

4. Responding

The last step for the virtual human to make is actually responding on the input of the patient. This could be just an animated response or some "intelligible" audio response. The purpose of the virtual human in social phobia therapy is to generate a controlled level of anxiety in the patient. Generating this anxiety can be done in different ways such as in the early studies used simple animations of clapping or walking away or taking a certain posture [1, 21, 22]. In rare cases the avatar could even say some pre recorded sentence showing his emotion to the user. To let the avatar talk back to the user speech synthesizers are needed or some sorts of pre recorded response database both have their advantages and disadvantages [49].

This chapter starts with describing the anxiety levels that needs to be provoked in the patient in 4.1 followed by emotional bots (4.2) where the ways a virtual human can show emotion is discussed. Next techniques for text to speech (4.3) are elaborated and what techniques will be used in this project and the reasons why. Followed by mouth to eye (4.4) that discusses the final details of the talking animation that is needed for realism and how this will be implemented in this project.

4.1 Anxiety provoking

The virtual humans are the social actors in the VE that interact with the patient and therefore probably the main source that generates the anxiety in the patient. These anxiety levels need to be managed the therapist must be able to influence the VH in such a way that they generate more or less anxiety with the patient. In early experiments that looked at the question if VH actually generated similar anxiety in the patient as the real thing used low level animated avatars that looked or followed the test subject in virtual reality. Test revealed that anxiety was present in patients when they where immerged in the VE [5, 25]. The question then became if the anxiety levels could be influence and used as treatment these experiments used three anxiety levels that where generated by giving the virtual humans different animations and sound effects [11]. First level was neutral level that was accomplished by making all the VH static in the world. This level was used as control for the other two levels the second level was an ecstatic audience that responded very positive on the test subject. The virtual humans would stand and applaud or shout positive remarks to the test subject. The last level was the negative public that would walk away or shout negative remarks and should provoke a higher anxiety level with the test subject [2].

This crude division into three levels showed that the effect on the test subject were significant and that in this way the anxiety level of the patient could be influence and used as treatment. An study [1] that focused more on the one-on-one interaction between the avatar and patient used a similar division where an avatar would be considered negative if it made snappy remarks back or responded uninterested. The positive level was accomplished by a much easier going conversation where the avatar would respond positive and interested on every remark from the patient. The few case studies of real attempted treatment used virtual reality more as exposure therapy where the different level played a lesser role and realism was much more important [6]. Of course derived from other phobia treatment fields the use of control on the anxiety level should be available for the therapist and based on the research focus should go from positive encouragement to the negative responding bots.

4.2 Emotional bots

Humans show emotion in a multitude of modalities from body movement, behavior and speech [50]. Some of the emotions are shown deliberately maybe even a bit exaggerated to envisage the emotion while others are only shown by very subtle clues and might not even be shown on purpose. Some emotions are very conscious while others might not even be as apparent to the person having them.

Virtual humans show no emotion at the level of real humans, attempts are being made by giving virtual humans the correct body movements, postures or facial expressions [51, 52]. Virtual humans are also limited by the amount of detail they can show sometimes this is because of the monitor not being able to show all the small ripples on a face but often it is just because the model of the virtual human would become so complex that rendering it real time would become impossible with normal computer power available. Avatars will therefore be limited by showing the basic postures and facial expressions most of the emotional clues will have to come from other modalities such as speech [53]. With speech you can show emotion by altering the tone but also the content of what the virtual human is saying. Because of the relative ease of showing emotions this way it will be one of the primary channels of showing emotion and influencing the anxiety level of the patient.

4.3 Text to speech

Converting text, that is the main means of creating a conversational database and handle the conversation in a computer program; to speech that goes from the speakers to the patients ears where it will be converted and processed the human way still is not a straight forward task. The generated speech needs to convey not only the basic information of the text but also the emotion of the virtual human. Therefore tone, speed and volume of the speech have to be influenced by the program. Techniques used to convert text to speech or to give a virtual human the ability to communicate in the literature can be grouped into two categories. The first is the virtual humans that used speech synthesizers most of them uses the Festival [54] package that is build on active research into correct and believable speech synthesis. The second option is to use pre recorded sound so instead of converting text to speech the text is compared to pre recorded sound tracks that then are played [49]. Both methods have their advantages and disadvantages that should be considered when deciding what method are the best to use.

Speech synthesizers like Festival have as big advantage that it does not require any recording of actions before it can be used. It can also “say” anything from the moment the package is installed without using a huge database. Of course this flexibility comes at a price because the synthesizers is made to simulate the human way of producing speech the result also sounds simulated or computer generated. The produced speech does not sound a lot like normal human speech it is understandable but it is not perfect [55]. Also generating any emotion in the speech is near impossible because the generated speech is accomplished by fine tuning a multitude of algorithms and generating a slightly different speech that still sounds similar but has a different emotion in it would require the same or even more tuning and is therefore not available in the package yet.

Pre-recording the speech to be played at the correct moments by the program has the main advantage that of having absolute control on how it sounds. There cannot be anything more real that the recorded real thing but that also highlights the biggest problem, have to record everything beforehand last minute changes to the conversation will not be easy to accomplish and also the initial task of recording all sentences would take up a considerable amount of time. It might be the only way to produce real emotion in the speech of a virtual human and therefore the only real option if emotional speech is needed.

A hybrid method between the two is also possible where instead of completely computer generated speech the program uses pre recorded fragments of speech that it then clued together. The fact that no standard package is available that uses this technique might indicate the success rate of this method. It would also probably suffer from the same inability to generate realistic emotional speech as the speech synthesizers [49].

4.4 Mouth to eye

The last step of a speaking avatar is the actual movement of talking. Producing the sound alone is not enough to generate a realistic image to the patient of a talking virtual human. Lip movement is just as important [56] and in the real interaction between two humans lip movement might even help make the message clearer. It is a fact that people can read what is said only by looking at the lip movement and maybe a little context of the situation. Virtual humans are not capable in generating that level of realism at this point in time and still be real time computable [57]. Ways to improve the realism is by recording the lip movement from humans while they are also recording the sound and then let these two channels be played again by the virtual human in sync. An easier option is by using cartoon like speech where mouths simulate the basic syllables and approach the real lip movement of the speech. Increasing believability could be accomplished by making more transitional animation between different syllables. An important factor still remains the synchronizing of the speech and the lip movement so that every syllable is show on time this will require increasing or decreasing the animation speed during the uttering of the sentence.

For this project Vizard will be used to generate the visuals of the program and as stated before the finer parts of the visual representation could be better left to professionals in that field. Vizard comes with standard lip movement that is similar to the cartoon like speech. It is more advanced than just open and closing the mouth but less advanced then pre recorded lip movements. Because the focus of the project is not on the actual conveying of the message itself but more how the virtual human should select what he want to convey the standard Vizard lip movement will be used.

5. Conclusion

Virtual reality has proven itself to be useful for the treatment of many different phobias such as fear for flying and fear of heights but still needs to find its way into the social phobia category. Studies have already shown that social phobias can be treated with virtual reality but the programs used often stick to the somewhat easier kinds of simulations. An example of these simulations is talking in front of a group of interested or disinterested virtual humans that express this by simple utterances or movements. A harder segment of the social phobias to simulate is the one-on-one interaction because here the avatar needs to be very flexible to follow the conversation and more often than not the therapist has to keep the conversation going in a wizard of oz style control on the virtual human.

This creates an opportunity for improvement and automation so that the therapist is relieved from this task and the avatar gets more flexibility to converse with the user. This requires some important steps starting with the actual input the virtual human needs from the patient. The input method used should be the most natural way of communicating with the patient as to optimize the feeling of presence. This method is speech and therefore a robust speech recognition program needs to be integrated into the avatar. Next is understanding and selecting the correct response on the utterance of the patient either by analyzing the sentence said or matching it to patterns in a response database. Finally the avatar needs to talk to the patient to convey its response in such a manner that does not break the feeling of presence of the patient.

Speech recognition is an important and difficult part of the interactive virtual human. With the state-of-the-art of speech recognition a freely listening and understanding program does not exist. Speech recognition suffers from a lot of problems ranging from background noise to not being able to handle accents other than those in the training set. This means that techniques need to be used to make the recognition more robust. The first step is making the trained language similar to that of the test group, in this case Dutch. Another step is limiting the number of words the recognizer needs to find at a certain moment by guiding the conversation and limiting its domain.

Understanding and selecting the appropriate response will be the main focus of this research mainly because it is here that the conversation needs to be kept going. Not only does the response need to be appropriate for the sentence uttered by the patient but it also needs to evoke the right anxiety for the patient. A crude division would be the friendly, nice and encouraging responses versus the rude and bored responses. A therapist needs to be the one that influences what kinds of responses are selected so that they can fine tune the therapy. To match the input to a correct output certain techniques can be used such as the techniques used by online chatter bots like A.L.I.C.E. These bots have proven themselves fairly successful in the Loebner contest but they might not be able to generate the required level of presence for the treatment of patients. The responses of chatter bots improve as soon as the domain of the conversation is somewhat limited or if the chatter bot need to assume a certain role. Combining the chatter bot technique with the conversation guiding needed for the speech recognition might give the realistic result needed. A second option is using the frame based conversation techniques used in automated telephone helpdesks. Here the program tries to acquire certain information and has a database of sentences it can utter to ask for this information. The user has the freedom to give this information in the order he wants to without having to follow a strict algorithm. Here the speech recognizer only listens to keywords in sentences and discards all other information and as back-up. The program tries to directly or indirectly verify if what it heard is correct. This technique was especially developed to cope with the problems speech recognizers have when listening to random people that did not train the system. The biggest limitation of this technique is the need of a clear goal for the system consisting out of a set of data points that needs to be filled in for example name, address, date and

place. Therefore it will not be able to have small talk with a user without an information gathering purpose. Since the goal of the project is to create a social situation between the patient and the virtual human the information gathering task might break the feeling of presence.

The last step is uttering a response to the patient in a realistic way. An option would be to use a speech synthesizer, for example Festival, to convert the plain text used in the program and database into speech. A real advantage in this is that any sentence can be uttered and last minute changes in the database can easily be accomplished. Disadvantages are that the speech generated does not sound very human, it is understandable but not convincing. The utterance is emotionless and therefore a problem for a virtual human that needs to evoke some sort of anxiety in the patient by what and how it says things. The only option left is pre-recording all possible utterances the avatar can make and just playing these recordings to the patient. This does remove the flexibility of the system but enables the emotional tone of the utterance. The remaining task is synchronizing the sound with the virtual image so that lip movement and possible facial and body expressions are in sync. For the visualization Vizard will be used to provide the visual representation of virtual humans and a limited set of movements, for example lip movements. Combining the various elements into a single virtual human should create a one-on-one capable social actor needed for the treatment of certain social phobias that cannot be treated using virtual reality today.

6. Abbreviations

VR	Virtual Reality
VE	Virtual Environment
HMD	Head Mounted Display
AIML	Artificial Intelligence Markup Language
SR	Speech Recognizers
VH	Virtual Human

7. References

[1] L. James, C.-Y. Lin, A. Steed *et al.*, "Social Anxiety in Virtual Environments: Results of a Pilot Study," *CyberPsychology & Behavior,* vol. 6, no. 3, pp. 237-243, 2006.

[2] M. Slater, D. P. Pertaub, and A. Steed, "Public Speaking in Virtual Reality: Facing an Audience of Avatars," *IEEE Computer Graphics and Applications,* vol. 19, no. 2, pp. 6-9, 1999.

[3] E. Klinger, P. Légeron, S. Roy *et al.*, "Virtual Reality Exposure in the Treatment of Social Phobia," *Studies in Health Technology and Informatics,* vol. 99, pp. 91-119, 2004.

[4] S. G. Hofmann, "Special Series: Innovations in Cognitive Behavioral Treatments of Anxiety Disorders Of Treatments and Technologies," *Cognitive and Behavioral Practice,* vol. 6, pp. 221-222, 1999.

[5] E. Klinger, S. Bouchard, P. Legeron *et al.*, "Virtual Reality Therapy Versus Cognitive Behavior Therapy for Social Phobia: A Preliminary Controlled Study," *CyberPsychology & Behavior,* vol. 8, no. 1, pp. 76-88, 2005.

[6] H. V. Martin, C. Botella, A. García-Palacios *et al.*, "Virtual Reality Exposure in the Treatment of Panic Disorder With Agoraphobia: A Case Study," *Association for Behavioral and Cognitive Therapies,* vol. 14, no. 1, pp. 58-69, 2007.

[7] S. Roy, E. Klinger, P. Légeron *et al.*, "Definition of a VR-Based Protocol to Treat Social Phobia," *CyberPsychology & Behavior,* vol. 6, no. 4, pp. 411-420, 2003.

[8] S. R. Harris, R. L. Kemmerling, and M. M. North, "Brief Virtual Reality Therapy for Public Speaking Anxiety," *CyberPsychology & Behavior,* vol. 5, no. 6, pp. 543-550, 2002.

[9] W.-P. Brinkman, C. A. P. G. v. d. Mast, and D. d. Vliegher, "Virtual Reality Exposure Therapy for Social Phobia: a Pilot Study in Evoking Fear in a Virtual World," in HCI for technology enhanced learning, 2008, pp. 85-89.

[10] B. Herbelin, "Virtual Reality Exposure Therapy for Social Phobia," Institut des systèmes informatiques et multimédias, Ecole Polytechnique Federale de Lausanne, 2005.

[11] M. M. North, S. M. North, and J. R. C. Clark, "Virtual Reality Therapy: An Effective Treatment for the Fear of Public Speaking," *The International Journal of Virtual Reality,* vol. 3, no. 3, pp. 1-6, 1998.

[12] T. D. Parsons, and A. A. Rizzo, "Affective Outcomes of Virtual Reality Exposure Therapy for Anxiety and Specific Phobias: A Meta-analysis," *Journal of Behavior Therapy and Experimental Psychiatry,* vol. 39, pp. 250-261, 2006.

[13] M. Garau, M. Slater, D.-P. Pertaub *et al.*, "The Responses of People to Virtual Humans in an Immersive Virtual Environment," *Presence,* vol. 14, no. 1, pp. 104-116, 2005.

[14] D. P. Pertaub, M. Slater, and C. Barker, "An Experiment on Fear of Public Speaking in Virtual Reality," in Conference on Medicine Meets Virtual Reality 2001, Newport Beach, Ca, 2001, pp. 372-378.

[15] H. Grillon, F. Riquier, B. Herbelin *et al.*, "Use of Virtual Reality as Therapeutic Tool for Behavioural Exposure in the Ambit of Social Anxiety Disorder Treatment," in Proceedings of the 6th International Conference on Disability, Virtual Reality and Associated Technology, Esbjerg, Denmark, 2006, pp. 105-112.

[16] K. R. Thórisson, "On the Nature of Presence," in AISB 2005 Symposium, Presence Cues for Virtual Humanoids, Hertfordshire, UK, 2005, pp. 17-24.

[17] J. Barnett, K. Knight, I. Mani *et al.*, "Knowledge and Natural Language Processing," *Communications of the ACM,* vol. 33, no. 8, pp. 50-71, 1990.

[18] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication,* vol. 22, no. 1, pp. 1-15, 1997.

[19] C. Teixeira, I. Trancoso, and A. Serralheiro, "Recognition of non-native accents," in In Eurospeech '97, Rhodes, Greece, 1997, pp. 2375-2378.

[20] "The Loebner Prize in Artificial Intelligence," 19-11-2008, 2008; http://www.loebner.net/Prizef/loebner-prize.html.
[21] M. Slater, D.-P. Pertaub, C. Barker *et al.*, "An Experimental Study on Fear of Public Speaking Using a Virtual Environment," *CyberPsychology & Behavior,* vol. 9, no. 6, pp. 627-633, 2006.
[22] E. Klinger, S. Bouchard, P. Légeron *et al.*, "Virtual Reality Therapy Versus Cognitive Behavior Therapy for Social Phobia: A Preliminary Controlled Study," *CyberPsychology & Behavior,* vol. 8, no. 1, pp. 76-88, 2005.
[23] J. L. Gauvain, and L. Lamel, "Large-vocabulary Continuous Speech Recognition: Advances and Applications," *Proceedings of the IEEE,* vol. 88, no. 8, pp. 1181-1200, 2000.
[24] P. Anderson, B. O. Rothbaum, and L. E. Hodges, "Virtual Reality Exposure in the Treatment of Social Anxiety," *Cognitive and Behavioral Practice,* vol. 10, no. 3, pp. 240-247, 2003.
[25] P. Anderson, E. Zimand, L. F. Hodges *et al.*, "Cognitive Behavioral Therapy for Public-speaking Anxiety Using Virtual Reality for Exposure," *Depression and Anxiety,* vol. 22, no. 3, pp. 156-158, 2005.
[26] B. Herbelin, F. Riquier, F. Vexo *et al.*, "Virtual Reality in Cognitive Behavioral Therapy : a Study on Social Anxiety Disorder," in 8th International Conference on Virtual Systems and Multimedia, Gyeongju, Korea, 2002.
[27] M. F. McTear, "Spoken Dialogue Technology: Enabling the Conversational User Interface," *ACM Computing Surveys,* vol. 34, no. 1, pp. 90-169, 2002.
[28] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication,* vol. 16, no. 3, pp. 261-291, 1995.
[29] J. Bing-Hwang, and S. Furui, "Automatic Recognition and Understanding of Spoken Language - A First Step Toward Natural Human-machine Communication," *Proceedings of the IEEE,* vol. 88, no. 8, pp. 1142-1165, 2000.
[30] A. Rudzionis, K. Ratkevicius, T. Dumbliauskas *et al.*, "Control of computer and eletric devices by voice," *Elektronika Ir Elektrotechnika*, no. 6, pp. 11-16, 2008.
[31] M. McTear, I. O'Neill, P. Hanna *et al.*, "Handling Errors and Determining Confirmation Strategies -An Object-based Approach," *Speech Communication,* vol. 45, no. 3, pp. 249-269, 2005.
[32] I. O'Neill, P. Hanna, X. Liu *et al.*, "Implementing Advanced Spoken Dialogue Management in Java," *Science of Computer Programming,* vol. 54, no. 1, pp. 99-124, 2005.
[33] G. Skantze, "Exploring human error recovery strategies: Implications for spoken dialogue systems," *Speech Communication,* vol. 45, no. 3, pp. 325-341, 2005.
[34] G. Ferguson, and J. F. Allen, "TRIPS: An Integrated Intelligent Problem-Solving Assistant," in American Association for Artificial Intelligence, Madison, WI, 1998, pp. 567-572.
[35] R. S. Wallace. "The Anatomy of A.L.I.C.E.," 12-12-08, 2008; http://www.alicebot.org/anatomy.html.
[36] F. Roberts, and B. Gülsdorff, "Techniques of Dialogue Simulation," *Intelligent Virtual Agents*, pp. 420-421, 2007.
[37] A. D. Angeli, and S. Brahnam, "I hate you! Disinhibition with Virtual Partners," *Interacting with Computers,* vol. 20, pp. 302-310, 2008.
[38] J. L. Hutchens, and M. D. Alder. "The NonI Conversation Simulator," 15-12-08, 2008; http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.8968.
[39] R. P. Schumaker, M. Ginsburg, H. Chen *et al.*, "An evaluation of the chat and knowledge delivery components of a low-level dialog system: The AZ-ALICE experiment," *Decision Support Systems,* vol. 42, pp. 2236-2246, 2006.
[40] R. Wallace. "Artificial Intelligence Markup Language," 12-12-08, 2008; http://www.alicebot.org/TR/2001/WD-aiml/.

[41] M. L'Abbate, U. Thiel, and T. Kamps, "Can Proactive Behavior turn Chatterbots into Conversational Agents?," in Proceedings of the 2005 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, France, 2005, pp. 173-179.
[42] V. Maya, M. Lamolle, and C. Pelachaud, "Influences on Embodied Conversational Agent's Expressivity: Toward an Individualization of the ECAs," in Proceedings of AISB 2004 convention. Symposium on Language, Speech and Gesture for Expressive Characters Leeds, UK, 2004, pp. 75-85.
[43] A. M. Galv˜ao, F. a. A. Barros, A. e. M. M. Neves *et al.*, "Adding Personality to Chatterbots Using the Persona-AIML Architecture," *LNAI 3315*, pp. 963-973, 2004.
[44] A. D. Angeli, S. Brahnam, and P. Wallis, "Abuse: The Dark Side of Human-computer Interaction," in Interact 2005 Adjuct Proceedings, Rome, 2005, pp. 91-92.
[45] S. Brahnam, "Gendered Bots and Bot Abuse," in CHI 2006 Montreal, Canada, 2006.
[46] A. D. Angeli, G. I. Johnson, and L. Coventry, "The Unfriendly User: Exploring Social Reactions to Chatterbots," in Proceedings of The International Conference on Affective Human Factors Design, Singapore, 2001, pp. 27-29.
[47] A. D. Angeli, "To the Rescue of a Lost Identity: Social Perception in Human-chatterbot Interaction," in In: Proceedings of AISB'05 joint symposium on Virtual Social Agents Hatfield, UK, 2005, pp. 7-14.
[48] C. L. Masia, D. W. McNeil, L. G. Cohn *et al.*, "Exposure to Social Anxiety Words: Treatment for Social Phobia Based on the Stroop Paradigm," *Cognitive and Behavioral Practice,* vol. 6, no. 3, pp. 248-258, 1999.
[49] D. O'Shaughnessy, "Interacting with computers by voice: Automatic speech recognition and synthesis," *Proceedings of the IEEE,* vol. 91, no. 9, pp. 1272-1305, Sep, 2003.
[50] A. Nijholt, "Issues in multimodal nonverbal communication and emotion in embodied (conversational) agents," *6th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2002)/8th International Conference on Information Systems Analysis and Synthesis (ISAS 2002)*, pp. 208-215, Jul 14-18, 2002.
[51] H. Pengyu, W. Zhen, and T. S. Huang, "Real-time speech-driven face animation with expressions using neural networks," *IEEE Transactions on Neural Networks,* vol. 13, no. 4, pp. 916-927, 2002.
[52] D. Heylen, "Challenges Ahead: Head Movements and Other Social Acts During Conversations," in Proceedings of the Joint Symposium on Virtual Social Agents, 2005, pp. 42-52.
[53] I. R. Murray, and J. L. Arnott, "Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion," *Journal of the Acoustical Society of America,* vol. 93, no. 2, pp. 1097-1108, Feb, 1993.
[54] "Festival," 19-11-2008, 2008; http://www.cstr.ed.ac.uk/projects/festival/.
[55] R. I. Damper, Y. Marchand, M. J. Adamson *et al.*, "Evaluating the pronunciation component of text-to-speech systems for English: a performance comparison of different approaches," *Computer Speech and Language,* vol. 13, no. 2, pp. 155-176, Apr, 1999.
[56] Y. Fu, R. X. Li, T. S. Huang *et al.*, "Real-time Multimodal Human-avatar Interaction," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 18, no. 4, pp. 467-477, Apr, 2008.
[57] N. M. Thalmann, P. Kalra, and M. Escher, "Face to virtual face," *Proceedings of the IEEE,* vol. 86, no. 5, pp. 870-883, May, 1998.