

# ECActive

## Embodied Conversational Agent for Mental Health Intervention

by

Mehedi Anam Sarder

to obtain the degree of Master of Science  
at the Delft University of Technology,  
to be defended publicly on Wednesday August 29, 2018 at 1:00 PM.

Student number: 4747313  
Project duration: February 5, 2018 – August 5, 2018  
Thesis committee: Dr. W. P. Brinkman, EWI-Interactive Intelligence, TU Delft  
Dr. J. Broekens, EWI-Interactive Intelligence, TU Delft  
Dr. N. Tintarev, EWI-Web Information Systems, TU Delft  
F. Burger, EWI-Interactive Intelligence, TU Delft  
H. Spliethoff, COO, WonderMedia

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.



# Preface

You are about to read this note which is the final touch on my Master Thesis. This work is the result of an intense semester of hard work both in terms of academic obligation and industrial activity. In fact, as required by EIT Digital Master School, the thesis internship was carried out in company. More specifically, Wonder Media Productions, an animation solution and content production company with office in Antwerp, showed interest in endowing their virtual characters with Artificial Intelligence to partially take over the tasks of the existing puppeteering system. During this relatively brief experience I was able to combine my tech enthusiasm, typical of an Engineer, with my ready concern about people health and well being. It has been a stage of intense learning for me, not only in the scientific field, but also on a personal level. Therefore, I hope this work can leave a mark, at least in a small way, providing insightful sources to my colleagues for further developments.

As a proficient entrepreneur knows that a successful startup cannot be founded without the effort of an efficient team, an ambitious scientist recognizes that a legitimate research cannot be carried out without the assistance of a group of great professionals.

First, among the TU Delft academic staff, I would like to thank my project main supervisor, Dr. Willem-Paul Brinkman for his guidance towards a more realizable version of the system. He provided me with the necessary tools to choose the right research direction and not be lost in pure theoretical visions of the research. Second, I would particularly like to single out my daily supervisor Franziska Burger. Her continuous assistance and feedback throughout these months were fundamental for reviewing this work. I would like to thank also Dr. Joost Broekens. During his course "Social Signal Processing" he inspired me to undertake and approach this research area for the upcoming Master Thesis project.

During my internship period in Wonder Media, Herman Spliethoff, became for me an important mentor and advisor. I would like to express my appreciation for sharing industrial expertise which provided me with a business perspectives of this work. Last but definitely not least, I would like to acknowledge Peter Verswyvelen for his "technical" assistance. Even with busy company schedules he contributed with his coding skills to get over my struggles with system implementation.

Every proudest accomplishment in life is reached through repeated renovation of commitment and self-motivation. And this refurbishment process is guaranteed (sometimes indirectly) by the people close to us (not by physical distance). I mention you here with few words as I could not say enough to express my appreciation. I would like to start thanking my Mom, who genuinely nurtured me without even knowing exactly what I was going through. I would like to thank also my Dad, who sacrificed his own ambitions to see fulfill my dreams. A special thank goes to my caring Panda for loving me and holding my hand through my ups and downs. Finally I would like to mention my genuine friends, for their support and concern despite my partial absence in this period.

Knowledge is not only scholastic or book-learned. It is often a result of human experiences and relationships. I would to express gratitude to all the people I met this year. First, my colleagues in company and room-mates who welcomed me and made me feel like home from day one. Second, my course mates who were always available for questions regarding formal procedures and helped me with experimental setup. Last, all the volunteers who participated to the experimental study and every single person who dedicated their precious time for keeping me in their prayers. *Alhamdulillah.*

*Mehedi Anam Sarder  
Delft, The Netherlands  
August 2018*



# Abstract

*Embodied Conversational Agents (ECA) seek to provide a more natural means of interaction for a user through verbal and non-verbal properties of human face-to-face communication. For this reason, these systems are found to bring benefits in different mental health related interventions. However, a key challenge in developing agents to replace the human interlocutor in a dyadic conversation, is to simulate appropriate attentive listening behaviors. In this thesis work, we explored different backchannel strategies and studied their effects in terms of likability and engagement. We built a fully embodied conversational agent with three different levels of backchannel strategies and ran a within-subject study with a convenience sample of 24 participants. The results showed that the amount of emotional words in the speech of users increased if the attentive listening capabilities of the agent were improved. In addition, the capability to trigger both verbal and nonverbal backchannels with proper timing was found to be a relevant feature in terms of improved speech rate and emotional words. Contrary to our hypothesis, backchannels based on actual emotion and sentiment analysis of the speech content were not found to be significantly influential on the quality of interaction. Multi-modal approaches are suggested for future works in order to overcome limitations of this work due to potential lack in emotion detection accuracy.*



# Contents

<b>List of Figures</b>	<b>5</b>
<b>List of Tables</b>	<b>7</b>
<b>Acronyms</b>	<b>9</b>
<b>1 Introduction</b>	<b>11</b>
1.1 Background & Motivation . . . . .	11
1.2 Research Question . . . . .	12
1.3 Overview Structure . . . . .	12
<b>2 Related Work</b>	<b>15</b>
2.1 Attentive Listening . . . . .	15
2.2 Embodied Conversational Agents . . . . .	15
2.3 Attentive Listening in ECAs . . . . .	17
2.4 Hypotheses . . . . .	18
<b>3 System</b>	<b>19</b>
3.1 Design . . . . .	19
3.1.1 Requirements . . . . .	20
3.1.2 System Model . . . . .	21
3.2 Implementation . . . . .	21
3.2.1 Audio Input . . . . .	21
3.2.2 Dialogue Engine . . . . .	23
3.2.3 Agent Output . . . . .	25
3.3 Pilot Study . . . . .	26
3.3.1 Results . . . . .	27
<b>4 Evaluation</b>	<b>29</b>
4.1 Methods . . . . .	29
4.1.1 Experimental Setup . . . . .	29
4.1.2 Subjects . . . . .	30
4.1.3 Measurements . . . . .	31
4.1.4 Procedure . . . . .	31
4.1.5 Data Analysis . . . . .	32
4.2 Results . . . . .	33
4.2.1 Likability . . . . .	33
4.2.2 Engagement . . . . .	34
4.3 Discussion . . . . .	35
4.3.1 Limitations . . . . .	36
<b>5 Discussion &amp; Conclusion</b>	<b>37</b>
5.1 Main Findings . . . . .	37
5.2 Limitations & Future Works . . . . .	37
5.3 Conclusion . . . . .	38
<b>A Appendix</b>	<b>39</b>
<b>B Appendix</b>	<b>41</b>
<b>C Appendix</b>	<b>43</b>
<b>D Appendix</b>	<b>45</b>
<b>Bibliography</b>	<b>47</b>





# List of Figures

2.1	Examples of ECAs for mental health interventions . . . . .	16
3.1	The interaction model. . . . .	21
3.2	The magnitude and envelope of a signal . . . . .	22
3.3	The effect of Uncanny Valley . . . . .	25
3.4	The final visualization . . . . .	26
4.1	Pictures of experimental setup. . . . .	30
4.2	Likability distribution by condition . . . . .	33
4.3	Engagement distribution by condition (part 1) . . . . .	34
4.4	Engagement distribution by condition (part 2) . . . . .	35



# List of Tables

3.1 System parameters . . . . .	28
4.1 Multilevel analyses of the outcome measurements . . . . .	33
4.2 Statistic summary of the outcome measurements . . . . .	33



# Acronyms

**AI** Artificial Intelligence.

**API** Application Programming Interface.

**ASR** Automatic Speech Recognition.

**BMI** Brief Motivational Intervention.

**CRS** Comparative Research Study.

**DEQ** Dialogue Experience Questionnaire.

**ECA** Embodied Conversational Agent.

**FSM** Finite State Machine.

**HMM** Hidden Markov Model.

**MHL** Mental Health Literacy.

**SDK** Software Development Kit.

**STT** Speech-To-Text.

**UI** User Interface.

**USC** University South California.

**VAD** Voice Activity Detection.



# 1

## Introduction

*The objective of this chapter is to present the main research question and the motivations behind it. Research approach and overview on thesis structure are also presented.*

### 1.1. Background & Motivation

Currently around 450 million people suffer from mental health related disorders in the world and 43.8 million American adults experience mental illnesses in a given year [1]. Most of the time, these conditions arise from mild disturbances before growing into severe impairments. The latter substantially interfere with one or more major life activities and can lead to substance abuse, chronic diseases or, in worst scenario, suicides. Yet suicidal behaviors and other mental chronic disorders are preventable when diagnosed early and treated properly. Therapies for early conditions are available, but of all adults with any mental illness in 2015 in U.S., only 43% received mental health services [2]. The causes are related to stigma, discrimination and neglect, which prevent care and treatment from reaching the target patients. In addition, for a large portion of the global population, therapy sessions are too expensive. The cost is related to a deeper and crucial problem: shortage of mental health care professionals. In fact, even if we consider an advanced country like USA, roughly 11 psychiatrists for every 100,000 people are available according to a recent report[5].

Potential solutions to this under-staffing of the specialty could be provided by e-mental health interventions based on Embodied Conversational Agents (ECAs). These autonomous agents are found to be suitable for support and self-help in different mental health interventions [22]. Two recent Randomized Controlled Trials ([91] and [86]) indicate that conversational agents lead to significantly positive stress management behaviors and psychological well-being in patients. Two other RCTs show that Internet-based treatments are effective in reducing depressive symptoms ([3], [30]). Lisetti et al.[85] evaluate with a Comparative Research Study (CRS) how an empathic virtual conversational agent could change the way traditional Brief Motivational Intervention (BMI) is provided. Last but not least, in [90] it is found that embodied conversational agents could be also involved for educational purpose, in particular to change people's attitudes to mental health disorder, reduce stigma and improve Mental Health Literacy (MHL).

Despite the benefits of this type of unguided e-health interventions, there are also problems related to the effectiveness. In fact, meta-analyses have repeatedly shown guided e-health interventions, such that involve a human supporter or caregiver, to be more effective than unguided ones [4]. One reason could be the lack of adherence, which is a major issue in Internet-based interventions [23]. Dropout from treatments with no human support is another concern [31]. While this renders human involvement desirable, guidance is also more expensive and less scalable than autonomous design. Therefore, researchers are investigating the possibilities of employing automated virtual agents to provide a proxy for the "human touch".

In their most advanced form, ECAs are used to simulate counseling sessions ([88], [89], [84]) as they present the general strategies used to tackle different types of mental issues. This mostly relies on the fact that mental health counselors, regardless of the specialization (stress management, anxiety control, behavior check) and their training level (clinical practitioner or unlicensed coach), share similar methods of face-to-face dyadic psychotherapy sessions. Examples of counseling skills, which are assumed to be effective when adopted by ECAs, include alternating close-ended, open-ended, probing questions, using basic concepts of coping theory and listening actively ([89], [84]).

Among the active listening skills, an important role is played by backchannels (BCs) as they provide meaningful feedback to the users. These behavioral cues are a fundamental part of counseling sessions to establish rapport between speaker and listener [24]. They are also relevant for the listener to display clear and unambiguous signs which show that he or she is actually listening. This ensures that the speaker does not conclude that what they are talking about is uninteresting. Moreover, people suffering from mental disorders, in particular depressive ones, have reduced willingness to speak [57]. Therefore, if some “appropriate” feedback is provided coherently, the speaker might feel more at ease and communicate more easily, openly and honestly, remaining engaged with the ECA for longer period of time [42].

## 1.2. Research Question

Unguided mental health related interventions, based on Embodied Conversational Agents, are explored as a valid alternative to traditional approach to support people affected by mental conditions, such as stress, anxiety and depression. The quality of the interaction with these virtual agents is related to their ability to resemble a real-world face-to-face communication. The main research interest in this work is to study the importance of simulating verbal and nonverbal attentive listening behaviors during a dyadic human-agent communication. More specifically the following research question is defined:

**RQ:** *“Do appropriate backchannel strategies have positive influences on human speakers in the context of a face-to-face human-agent interaction?”*

## 1.3. Overview Structure

Here we provide an overview on our research strategy and planning with reference to the thesis structure.

After presenting our research interest and the motivations behind it (Chapter 1), we studied the related works in order to build arguments supporting our research hypotheses (Chapter 2). We first described the importance of attentive listening for mental health related interventions and investigated which characteristics of Embodied Conversational Agents rendered them suitable for e-mental health interventions. Then we analyzed the existing strategies for implementing backchannels into ECAs. Last, we formulated our two research hypotheses to translate the position taken in this thesis, which is that user’s perception is influenced positively by both timing and emotional content of agent’s responses.

As a first step in testing these hypotheses, an ECA was endowed with three different backchannel strategies (Chapter 3). Each of them constituted a specific requirement directly derived from each of the hypotheses. We also defined other common system requirements and then realized the model of our conversational agent. Later we provided implementation details for each of the units, including input analysis, dialogue engine and agent output. We finally operated a pilot study to check if the system was correctly functioning and collect qualitative data to adjust the system parameters.



Afterwards, we designed an experimental study to test the research hypotheses, (Chapter 4). We ran an within-subject study where an opportunity sample of 24 participants was exposed to three different backchannel conditions. We decided to operationalize engagement and likability with the output measurements. Multi-model analysis was adopted to check if there was a significant relationship between independent and dependent variables. A planned comparison was also used to determine where exactly the effect observed in main analysis was occurring.

Main findings of the study provided the verdicts on each of the hypotheses (Chapter 5). In particular, it was shown that timing was a significant feature influencing the quality of human-agent interaction in a positive way. Then we reflected also on the limitations of this research work, which resided mostly in the system implementation. Finally, strategies based on multimodality were suggested for future works in order to improve the system performance.



# 2

## Related Work

*The objective of this chapter is to build the argumentation supporting the research hypotheses, which state that user's perception is influenced by both timing and emotional content of agent's responses. First, it examines how human expresses attentive listening through backchannel signals. Next, it shows why embodied conversational agents can be considered a suitable interface for simulating these signals. Related works on ECAs implementing backchannels are analyzed and, finally, based on their limitations, two separate research hypotheses are formulated to translate the position taken in this thesis.*

### 2.1. Attentive Listening

In the context of mental health related interventions, the therapist is required to fully understand, respond to and remember the context. In fact, professional figures, such as clinical psychologists and occupational psychotherapists, are encouraged to master the capabilities of attentive listening. Attentive (or active) listening involves listening with all senses, concentrating on the content rather than passively “hearing” the speaker. Nonverbal mimicry [35] and verbal paraphrase or summary are important practices in promoting awareness and detecting potential discrepancies between the patient's current behavior and his or her actual beliefs[36]. Other visible and unambiguous, signs such as using of silence to open up communication, leaning forward, nodding, smiling and using fillers, make the patient feel more comfortable to keep talking. These behavioral cues are defined as regulators[33]. These types of “listener responses”, which are sometimes deliberately displayed to show that the listener is actually listening, are also called backchannels (BCs).

These signals are performed by the listener as “*non-intrusive acoustic and visual signals provided during the speaker's turn*” [38]. Multimodal backchannels communicate perception, attention, interest, comprehension, belief, and acceptance towards the speaker and what he or she is saying [26]. These behavioral cues, if generated with proper timing and content, are found to provide information about the basic communicative functions towards what is being said to the conversational agent [39]. Timing is identified to be fundamental for a better likability and similarity to human interlocutor [62].

### 2.2. Embodied Conversational Agents

One potential solution to poor adherence of current automated treatments for pervasive mental health disorders is provided by Embodied Conversational Agent (ECA). Current literature adopts different terminologies but ECAs are commonly defined as *autonomous agents with digital embodiments capable to communicate and to exchange informations with the users using verbal and non-verbal modes of communication* ([67], [13]).

One of the advantages of ECA is that it can operate without any human support. It is endowed with computer models that give it knowledge-based capabilities to reason over the input data derived from the communication with real-world users. These models assure also that the agent is able to carry out an interaction autonomously through dialogue states and meaningful responses.

**Dialogue Management Initiative** One of the main features to consider when analyzing and implementing a dialog management system of a conversational agent is the type of initiative. In fact, in every state of the dialogue, the agent reacts based on the type of initiative. If it decides to adopt a strong *system initiative* approach, user actions are only allowed at certain points and agent reactions are generated based on prefixed rules. On the contrary, in the case of a *user initiative* policy, any type of agent reaction, such as backchannel, is built on user's input data. The latter could be speech or nonverbal data depending on the number of input modalities.

**Multi-modality & Embodiment** Early chatbots, such as Eliza[6], Parry[8] and Alice[9], with their pure text-based interfaces, are ideal to test dialogue managing capabilities. However, as they do not accomplish a visual representation of themselves, they are not eligible for nonverbal communication. This has an effect on the overall quality of system feedback. For example, gestures are fundamental for the listener as they help comprehension and memory ([25], [26]). One of the features that distinguishes ECAs from other conversational agents is their embodiment. This visual component is found to convey additional informations such as emotion and reinforce the belief that it is a social entity [13]. In Figure 2.1 we show a few examples of human embodiment in ECAs used for mental health intervention. Lastly, the integration of multi-modality is necessary also in input as is proved to produce superior results in human behavior analyses ([58], [59]) and has shown positive effects in cognitive researches ([26], [28], [27]).



Figure 2.1: Examples of Embodied Conversational Agents for mental health interventions. From top-left clockwise: Gabby, Ellie (SimSensei), Justina, Greta, Bill Ford and Laura

## 2.3. Attentive Listening in ECAs

The shortage of mental health care professionals has encouraged the researchers to explore the possibilities of simulating the attentive listening behaviors through autonomous systems. Characteristics such as flexibility in dialogue management strategies and possibility of multi-modality in both input and output, make Embodied Conversational Agent (ECA) a suitable interface to deliver attentive listening behaviors, such as backchannels.

Several studies ([40], [39], [41]) have been conducted in order to detect prototypical back-channel signals to be used in a listener module for an ECA. Backchannels can take many different forms and be expressed using different modalities. In [40], visual backchannels such as gaze, smile, nodding are analyzed and it is found that a combination of them alters the perceived meaning of the backchannel in a context-free dialogue. In [39], vocal utterings, paraverbals, words or short sentences and their placement in sequences of talk are analyzed. It is found that specific responses are tightly connected to the speaker's narrative (e.g. sentence completion, brief restatements) and generic responses are mere signals of continued attention (e.g. requests for clarification, short vocalizations, head nods). In [41] it is highlighted that also a listener in conversation can have influence on the talk, not through contributions to the development of the topic but through minimal backchannels. These responses include such items as acknowledgements, brief agreements and continuers (i.e. *Yeah, Mm hm, Uh huh and Mm*).

Apart from the specific "lexicon" and the role of BCs in conversation, researchers have focused on identifying context of user-initiative BCs. It is noted that backchannels are often produced after rhythmic pulses in the speaker's speech ([43], [44]). Corpus-based research showed that backchannels are also often generated when the vocal message finishes with a rising or falling pitch [45], after "meaningful" words (i.e. nouns, verbs and adverbs) [47]. In addition, BCs are found to be often preceded by a short pause in the speaker's discourse [46].

Several authors have addressed real-time prediction of BC timing using either machine learning ([48], [49]) or rule-based algorithms ([50]). The former can automatically determine decision rules from labeled training features (supervised learning). This corpus of dialog data, used to model human behavior ([61]), is strongly context-dependent and hard to generalize. For this reason manually defined rule-based algorithms that predicts BCs based on specific features are broadly applied in many applications. Different techniques make use of different features, but they all have in common simplicity given the real-time requirements. Continuous lower pitch is one of the rules used to predict BCs and it is reported to work better than random [50].

In the current literature, several works have attempted first approaches to the implementation of a backchannel model as part of ECA design. One of the first characters to be able to interact with speakers using vocal and non verbal signals was the talking head Gandalf [51]. REA [14], while showing different houses to the users, is also able to display para-verbals, head nods and short verbal statements. In [52], an agent provides only visual signals and in [53] the authors improve that by introducing multi-modal analysis of the speaker. Multi-modality is a key element in [19] and in the state-of-art prototype, SimSensei, developed by University South California (USC) for mental health intervention [66]. Finally, attempts are made to trigger backchannels reasoned over the users' intentions and beliefs [54].

The existing mental health ECAs simulate only a subset of backchanneling strategies, considering some characteristics of face-to-face communication and neglecting others. Most of them ([14], [52], [53]) generate verbal and nonverbal without any or enough understanding of the content, especially the emotional significance of the message delivered by the speaker. Some others ([48], [49], [50]) only focus on determining the correct timing and improving the predictability of BCs but ignores the quality of displayed visual signals which are important for improving the quality of interaction. For example, the speakers are usually sensitive to the listener gaze: if they begin to formulate a sentence and they realize the listener is not looking at them, they can stop, restart and repeat or rephrase what they already said [60]. In addition, the perception and the likability of the conversation are also influenced by the fluidity and synchronization in avatar animations [55].

## 2.4. Hypotheses

ECAs are considered a suitable interface for simulating attentive listening behaviors, such as backchannels. Studies have been done to identify the lexicon ([40], [39], [41]) and the role of BCs in a conversation ([26], [39]). Other works have addressed real-time prediction of BC timing ([48], [49], [50]) and multimodal simulation of BC signals ([14], [52], [53]). However, very few of them studied the effects of simulating multimodal BCs using “appropriate timing” on the overall interaction quality [62]. Lastly, to the best of our knowledge there is no investigation yet in current literature focusing on the effects of simulating BCs based on “emotional understanding”, that is the emotional and sentiment weight (e.g. positive or negative) of the users’ speech content.

Based on these argumentations, we formulate our hypotheses regarding this research topic:

**H1:** *The users find the user-initiative verbal and non-verbal backchannels displayed with appropriate timing, more likable and engaging compared to the ones generated at random timing.*

**H2:** *The users find the user-initiative verbal and non-verbal backchannels displayed with appropriate timing which show understanding of the emotional content, more likable and engaging compared to the ones generated with appropriate timing but no understanding of the speech content.*

Formulation of the research hypotheses is fundamental to make decisions on the design, implementation and evaluation of our system. In fact, in order to evaluate our hypotheses, we need to build an ECA capable of simulating attentive listening behaviors, predicted with different rule-based algorithm. These rules are based on timing and emotional understanding. Finally, we run an experimental study to test if there is an increase in level of engagement and likability in a user in the setting of a face-to-face dyadic conversation.

# 3

## System

*The objective of this chapter is to present the system that was built to test the hypotheses formulated in the previous chapter. First the system requirements are derived from analysis of problem space and the hypotheses. Then it is presented the design of the interaction model and the implementation of different modules. Finally, a pilot study is run to check if the system was correctly functioning and to adjust the system parameters.*

### 3.1. Design

As a first step in testing the hypotheses, we implemented an Embodied Conversational Agent to run an experimental study and to collect valuable measurements. However, as we observed from our state-of-art research, the current literature offers multitudinous possibilities for attentive listening realization from both technological and intervention perspective. We made decisions during the design phase to have a fully functioning system without overloading it with superfluous features. These decisions included the visual aspect of the agent, animation repository, input-output modalities, dialogue management policies and, most importantly, backchannel strategies.

From the point of view of the interaction with the agent, three different levels of backchannel “intelligence” were designed and then implemented. These could be considered requirements directly specified by our research hypotheses:

- **Level 1:** The agent randomly displays neutral verbal backchannels (e.g. “Aha! That’s very interesting. Tell me something more about it”) accompanied by nonverbal regulators (e.g. head nodding, head scratching, smile) and nonverbal illustrators (e.g. hand movement). Complete list of movements with respective durations and screen-shots is provided in Appendix B.
- **Level 2:** The agent detects user’s silence and coherently displays same verbal and nonverbal cues from previous level. Here *coherently* means with a well-defined timing.
- **Level 3:** The agent responds to the emotional content of user’s speech, detects user’s silence and coherently displays neutral or positive or negative verbal backchannels (e.g. “It seems you did (not) like that experience. Tell me something more about it”) accompanied by the same nonverbal behavioral cues as in the previous levels.

As it is noticeable, level 2 and 3 were built as an improvement and enrichment of the respective previous ones. In that sense, level 1 and 2 are required to test hypothesis 1 (**H1**) and level 2 and 3 are required to test hypothesis 2 (**H2**). Regardless of the attentive listening level, other common functionalities were required by the system in order to deliver the different backchannels to the users.

### 3.1.1. Requirements

As any other system, our proposed conversational agent presented some technological requirements to be fulfilled in order to achieve sufficient reliability on the overall system. The functionality requirements of our agent were divided into two categories, front-end and back-end requirements, and they were prerequisites for all the attentive listening levels.

#### Front-end requirements:

- Embodiment
- Coordinated animations
- Multimodal interface
- Real-time responsiveness

One key element was the embodiment, which could have been human or animal and both digital and physical but in our context we excluded the latter as it required robotic systems. The use of embodiment brought issues related to the visual component to the conversational agents. First of all the correct level of visualization details was investigated as higher quality of visualization and realism did not always correspond to higher credibility and engagement. An example was given by the effect of "uncanny valley" [69]. The term referred to the limit of realism and the point where the increase in similarity of a virtual agent to human appearance and movement resulted into a strong drop in believability and comfort. In addition, a particular attention should have been paid on animation details for the 3D character such as idle behavior, gaze level, the timing of different actions or movements in order to convey the correct message and emotion[21]. A smile displayed in a wrong moment, for example when the speaker was talking about a bad accident, could have disrupted the rapport with the virtual agent. As a modality is a single separated channel of sensory data produced by the system, our agent could be considered to generate multimodal outputs. This required also the agent to be delivered using multimodal interface, which allows the user to interact with the ECA and share informations. With the spread of electronic devices among the population, nowadays the interfaces can range from web-based applications for on line platforms to mobile applications for end-user devices. In addition to visual and audio data exchange, we required the system to present a "button" to allow the user to trigger the start of conversation and keep track of the total interaction time. Last but not least, real-time condition was mandatory for simulating backchannels in the context of real-world dyadic interaction.

#### Back-end requirements:

- Low computational cost
- Retrieval-based answer generation
- Animations based on the existing database of rigged characters

The first requirement was directly derived from our real-time condition imposed by the context of attentive listening behaviors. In order to keep computational demand low, we preferred to generate *retrieval-based* responses using pattern matching rules implemented as a Finite State Machine (FSM). This was opposed to *generative-based* strategies, where the answer sentences were formed through machine learning methods. Although the latter presented high scalability, the method was also time-consuming, therefore we opted for a more context-dependent solution. This was possible as we decided the specific topic for the conversation and chose "a priori" the list (i.e. dictionary) of possible verbal and nonverbal behavioral cues. The last requirement was justified for the following reason: as we were using proprietary 3D characters, the predefined set of animations was not scalable in number.



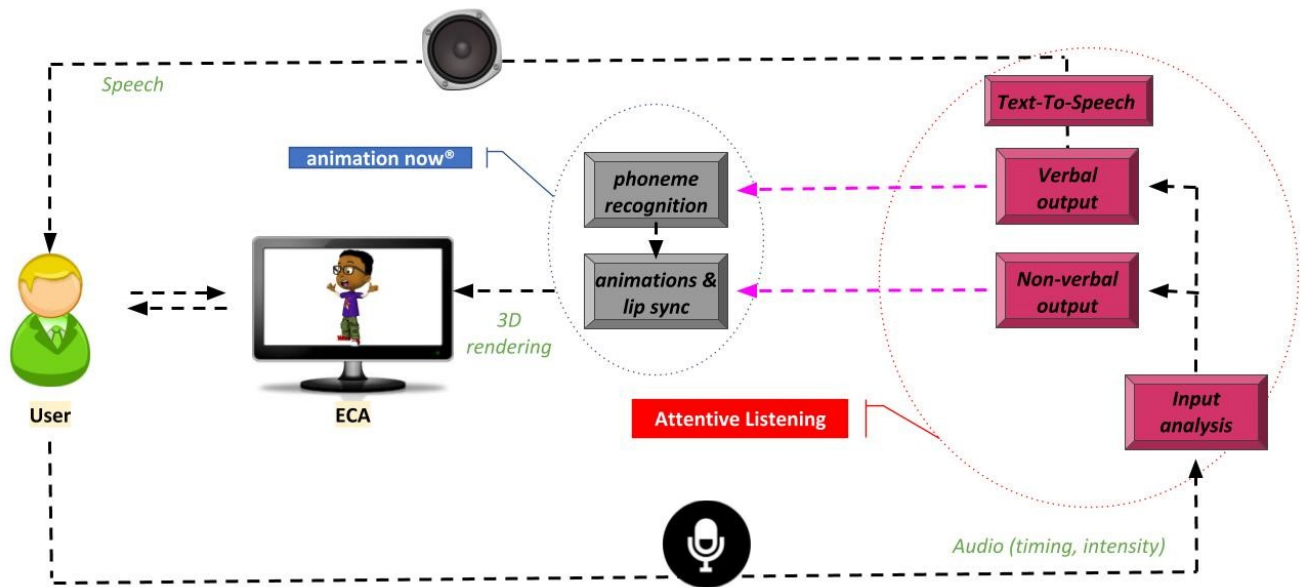


Figure 3.1: The interaction model.

### 3.1.2. System Model

Here we present a simplistic version of our interaction model and explain its different parts. In figure 3.1 we consider the agent capable of providing backchannels with a complete feedback loop and fulfilling all the system requirements.

In the assumption of a dyadic conversation a human took the role of the main speaker, and the ECA, as an interlocutor performed attentive listening behaviors. User's speech, recorded through simple input device (e.g. microphone) was analyzed on the fly by a cloud based recognition engine and processed locally for pause detection. The system analyzed the input informations and decided the necessary response signals, both in terms of audio and video. Verbal backchannels were delivered through direct output device (e.g. headset) but the stream of audio information was fed into also the AnimationNow<sup>1</sup> tool for phoneme detection and viseme conversion for proper lip synchronization. AnimationNow is the proprietary software from WonderMedia<sup>2</sup>, creative production company, which source code provided us with necessary libraries for phoneme to viseme conversion and animation generation. The final interface was a monitor for displaying the 3D character and its reactions to the user.

## 3.2. Implementation

In this section, we provide more details on each of the building block of the system model, starting from input analysis to output synthesis. The selected programming language was C# as adopted by AnimationNow developers. Nevertheless C++ was also adopted and wrapped into the main file as it provided higher flexibility and major compliancy to open-source libraries.

### 3.2.1. Audio Input

The sound input was very important for the correct functioning of the application as the only feedback the system got from the user is through his/her speech. Two main components of the system used the audio input to generate their data. The first one was speech detector, which task was to find out if the person was talking or not and it was partially achieved using Voice Activity Detection. The second part was the speech recognizer that converted the spoken phrases to text strings used during pattern-matching step. Only the first component was sensitive to background noise, which was preferable to be reduced at minimum.

<sup>1</sup><http://animationnow.org/>

<sup>2</sup><https://wondermedia.tv/>

**Voice Activity Detection (VAD)** One of the challenges that we had to overcome when creating an automatic system for free speech which was always "active", was to figure out how to model users' silence in order to detect it efficiently, and react properly as soon as possible. This was because we were not employing any "hard" button which could have been pressed when the user finished to talk as, for example, it was common among text-based chatbots. We simulated with high fidelity a real-world face-to-face conversation without any video input which meant that the system could not detect when the mouth opening was small enough to consider the speaker in pause. For this reason, this step resulted to be not so trivial.

A voice activity detector was charged with three main functions:

1. Extract the envelope of the 2D audio signal.
2. Choose a suitable threshold (loudness) level.
3. Detect timings where the absolute magnitude value of the envelope exceeds the predefined threshold.

An envelope of an audio signal is the curve that outlines its extreme values over time, independently of how its frequency content makes it oscillate (see Figure 3.2).

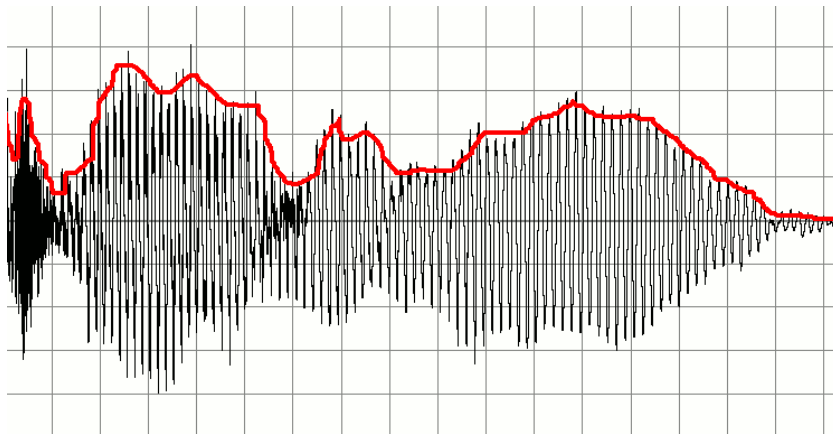


Figure 3.2: The magnitude of a 2D audio signal and its envelope highlighted in red.

To avoid too sharp boundaries between talking and silence, each sample value was replaced by an average of its  $N$  previous neighbors before the check with the threshold value. This approach is called moving average and it is a simple and low demanding way to low-filter the whole signal. This step is also visible in Figure 3.2 as the red signal (envelope) do not cross exactly the signal high picks and it has low sensibility to the impulse-based sounds.

To make the system even more robust to background noise, we combined the Voice Activity Detection (VAD) technology with recognition event. In other words, in order to change the system state into listening state, we also checked if the first detected sound could have been recognized as an English word. In this way if the user generated an undesired loud sound (e.g. cough) the system would not have considered the user speaking. This strategy also worked for detecting filled pauses, which are defined as a sound produced during speech that represents a silent period but filled by a vocalization (*umm, ehh*). This type of pauses is usually harder to interpret compared to "standard" unfilled pauses ([71]).

**Speech Recognition** As our ultimate goal was to convert real-time users' speech into strings which were than analyzed for keyword spotting, we required to use an Automatic Speech Recognition (ASR). Also known as Speech-To-Text (STT), Speech Recognition is the set of methodologies and techniques used to translate spoken natural language into text string using a machine. It fundamentally functions as a pipeline that converts PCM (Pulse Code Modulation) digital audio from a sound card into recognized speech. The PCM digital audio is first transformed into a better acoustic representation. Then a "grammar" is usually applied so the speech recognizer knows what phonemes to expect. A grammar could be anything

from a context-free grammar to full-blown Language. Afterwards the system detects which phonemes are spoken and, following a conversion table, translates them into words.

This technology has a very long history but only recently became fundamental in commercial applications. Different existing libraries included both open-source (CMU Sphinx<sup>3</sup>, Microsoft SAPI<sup>4</sup>, Kaldi<sup>5</sup>) and licensed solutions (BING Speech<sup>6</sup>, Google Cloud<sup>7</sup>, IBM Watson<sup>8</sup>). The formers usually followed statistical approaches, such as Hidden Markov Model (HMM) and the commercial ones were based on learning methods such as neural network and took advantage of a big amount of training data. Among all different libraries we identified the ones which provided Software Development Kit (SDK) in either C# or C++ as they were both compatible languages with the remaining system architecture. We first opted for Microsoft SAPI following the "open-source path". However its high sensibility to noise forced us to adopt Bing Speech API for our final implementation.

Bing Speech API, used also by intelligent assistant Cortana<sup>9</sup>, is part of a collection of machine learning algorithms that Microsoft has developed to solve problems in the field of Artificial Intelligence (AI). This set, now called Cognitive Services but previously known as "Project Oxford", can be used by developers using Windows machines, through standard REST calls over the Internet. REST (Representational State Transfer) is an architecture style for designing networked applications and it relies on HTTP protocol for calls and requests.

Once registered to Microsoft Azure<sup>10</sup> portal, a free subscription to the service offered the possibility to carry out 5,000 utterance transactions per month which was a sufficient limit for our testing purposes. Instead of REST API we used client libraries available for Windows machine with different language compatibility including C#. Unlike REST which made direct calls using HTTP, client libraries used websocket-based protocol with possibility to get intermediate results for continuous and long audio streams. The system could get real-time streaming, which meant that when our application (client) sent audio to the service (server), it received partial recognition results back. This was implemented in an asynchronous state machine, which means that the user could respond to the agent anytime and the system should have been able to call a thread which detected the start of audio stream as soon as possible. The necessary references to the libraries were added using NuGet<sup>11</sup> package manager and they were built using Visual Studio 2017 Community edition<sup>12</sup> on a Windows 10 machine.

### 3.2.2. Dialogue Engine

The dialogue management system was mixed-initiative and realized through asynchronous Finite State Machine (FSM) with retrieval-based answer generation. It was mixed-initiative because, although the overall condition and structure of the interaction was controlled by the agent, the user had influence on the timing and on the emotional content of the agent response. In addition, an asynchronous "event-driven" FSM meant that the transition between states was controlled by the event inputs, so that the FSM did not need to wait for a clock signal input.

There were three different dialogue states: Idle, Speaking and Listening. Here *Idle* state was not to be confused with "idle" pose explained in 3.2.3. The start of the conversation was triggered by a manual command which shifted the agent state from *Idle* to *Speaking*. The *Idle* state could be reached only with the end of the conversation which could be triggered again by a manual click or automatically as established by a timer *timerGlobal*. In this case, in order to not interrupt users' stream of thoughts (level 2 and 3), we prompted the users about the ending of the conversation ("*Thanks for your time. Unfortunately our time is over. See you*

<sup>3</sup><https://cmusphinx.github.io/>

<sup>4</sup><https://docs.microsoft.com/en-us/previous-versions/windows/desktop/ms723627>

<sup>5</sup><http://kaldi-asr.org/>

<sup>6</sup><https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/>

<sup>7</sup><https://cloud.google.com/speech-to-text/>

<sup>8</sup><https://www.ibm.com/watson/services/speech-to-text/>

<sup>9</sup><https://www.microsoft.com/en-us/cortana>

<sup>10</sup><https://azure.microsoft.com/en-us/>

<sup>11</sup><https://www.nuget.org/>

<sup>12</sup><https://visualstudio.microsoft.com/>

*next time!*”) only when they were in silence. For the rest of the interaction, switch between *Speaking* and *Listening* happened based on user audio input and speech detection module described in the section 3.2.1. The system went to *Listening* state whenever the user started speaking and it went back to *Speaking* when it detected a certain amount (*timerSilence*) of silence. In order to avoid continuous speech by the user without any interruption (e.g. monologues) we also imposed a maximum amount of words (*maxWord*) that the user could say without any pause. The most suitable values of *maxWord* and *timerSilence* were found empirically through a small Pilot Study (3.3).

As we previously stated, the answers were generated through a pattern-matching strategy. This meant that at every switch from *Listening* to *Speaking* state, the system checked the pattern in user’s speech and decided accordingly the vocal response. First thing to be analyzed was the number of total words and if it was lower than a threshold, *minWords* (pre-defined empirically by the Pilot Study), then the agent realized that the user had not spoken enough and *he* repeated the previous statement and invited the speaker to take the initiative (i.e. *“Don’t be shy. Please tell me something.”*). Same was applied for the situation where the user did not speak at all for a certain amount of time (*timerSilence2*).

If the speaker had spoken sufficiently, than for level 3 we needed to perform a keyword spotting step. At every partial speech recognition event, the recognized string was split into words and saved temporarily into a vector list. Then each of the element of the list was compared with two other arrays of words, one containing positive sentiment keywords and another with negative ones. These two lists were result of a detailed annotation work done by Bing Liu and Minqing Hu [80][70] using opinion content on social media.

Nevertheless we modified the two lists to create the knowledge base for our study to reduce the computational demand of this module. First, we manually eliminated all the misspelled word which were added as they were frequently appearing in social media platforms. In our case they were not relevant as our Speech Recognition system always converted to correctly spelled English words. Furthermore, as the complexity of the template-matching step ( $O(n)$ ) was influenced by the number of total keywords, we reduced the length by eliminating “redundant” parts of the words. In other words, for Regular Expression Matching it was enough to provide the core of a string. For example in order to detect the keywords “NICE” and “NICELY” it was enough to provide the string “NICE”. However doing this we had to pay attention on which were the first and second members of the string comparison. In fact we needed to take each of the recognized input words and search for it inside the vector of positive and/or negative keywords. Otherwise the repetition of the same word (or its trunk) would not have been counted multiple times.

Two counters tracked the total number of positive and negative words mentioned by the speaker. At the end, if *countPos* was greater than *countNeg* we generated a response that highlighted the positive sentiment of the experience told by the user and viceversa. Example of response in case of positive emotional content could have been: *“So overall it seems you liked it.”*. This kind of response is typical of a Reflective Listening strategy which promotes awareness and detects potential discrepancy between people statements and his/her real opinion.

In a later phase, more specifically after the Pilot Study, we introduced an “offset” value which was added when comparing the integer value of the two counters. This was because most of the time a majority of only an unit value was not sufficient to determine the overall sentiment level and the positive category was used only if positive words were detected such as “naturally” or “rightly”. This “softer” margin performed better as the agent displayed a “safer” neutral reaction whenever there was no clear majority of one sentiment category compared to another.

So far, we considered the implementation for level 2 and 3. For level 1 the dialogue management engine described previously was disabled. An alternative and much simpler dialogue policy is taken into consideration. In fact for level 1 we simply needed to generate neutral attentive listening behavioral cues at random time. The *timerLevel1* decided how often these verbal and nonverbal signals were generated and the most suitable value was inferred from the Pilot Study data.

### 3.2.3. Agent Output

**Audio Output** The audio output was one of the two communication channels used by the program to respond to the users input in multimodality. The character was able to speak using synthesized speech created using Microsoft Speech API (SAPI 5.3) and lip synchronization was realized through Animation Now internal code. However, the latter was realized using Annosoft<sup>13</sup> viseme coding which was different from SAPI visemes. Therefore we first implemented a real-time conversion from SAPI phoneme to Annosoft phoneme. Only afterwards we realized the viseme mapping using Annosoft 17 viseme. The complete table of visemes is presented in Appendix A where they are shown at “full open” and sorted by similarity. The viseme targeted for our character should have been somewhat exaggerated from real mouth positions but a blending functionality tended to weaken everything somewhat. The vowels were condensed to fewer visemes and still had realism.

The synthesized voice was Microsoft David and we kept the pitch and volume at default level. The user could adjust the latter by simply using the headphone buttons. An issue arises when the system itself generated sounds and could potentially “listen to itself”. To avoid this, the user was given a headphone that will limit the interference from the system on itself.

**Visualization** The User Interface (UI) including the main window and buttons was developed in C# language. Jimmy, the virtual character, was chosen among the ones created by the production team in WonderMedia as it presented the greatest number of possible behavioral cues. In fact this 3D character was able to perform more than 50 different gestures and movements. The most relevant of these non-verbal behaviors are listed in Appendix B with highlights on regulators (backchannels) and illustrators which accompanied the generated verbal outputs. In this appendix also the durations of each animation are provided (in seconds). In addition to the animations, during the idle pose (that was when no animation is triggered), the character performed natural movements such as breathing, random eye blinking and head glitch.

Furthermore Jimmy presented an animal embodiment, more specifically of a monkey. This choice was based on studies reporting increases in empathy and better awareness in a human in front of animals [68]. Although there are many complex factors which determine views toward animals, in general human actions toward animals are rooted in perceptual concepts concerning the intrinsic nature of being animal himself/herself. In other words, animal embodiment could have been a suitable setting for delivering attentive listening behaviors. In addition, this choice of embodiment could have also avoided issues related to the “uncanny valley” of realism [69]. As reported in Figure 3.3 by Mori [69], stuffed toys with animal appearance come just before a drop point in credibility. In addition, we assumed that the presence of an animal could eliminate potential prejudice or stigma due to human conventions such as gender, age, ethnicity.

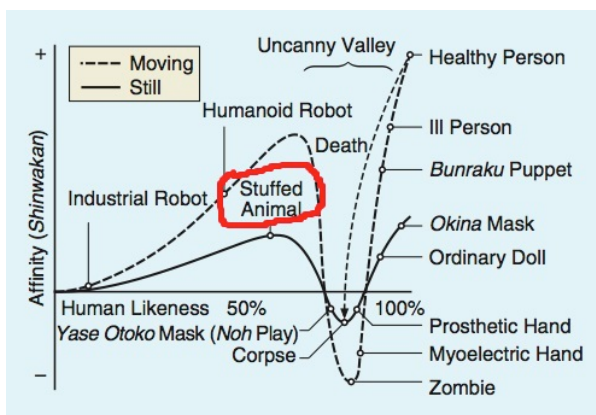


Figure 3.3: Uncanny valley and the position of stuffed animal as presented in [69].

<sup>13</sup><http://www.annosoft.com/lipsync-tool>

Although the avatar was the focal point of the users, the background world could not have been overlooked. We chose a familiar home environment and filled the room with just the minimum necessary amount of details without overloading the graphic requirements. The final visualization is shown in Figure 3.4.

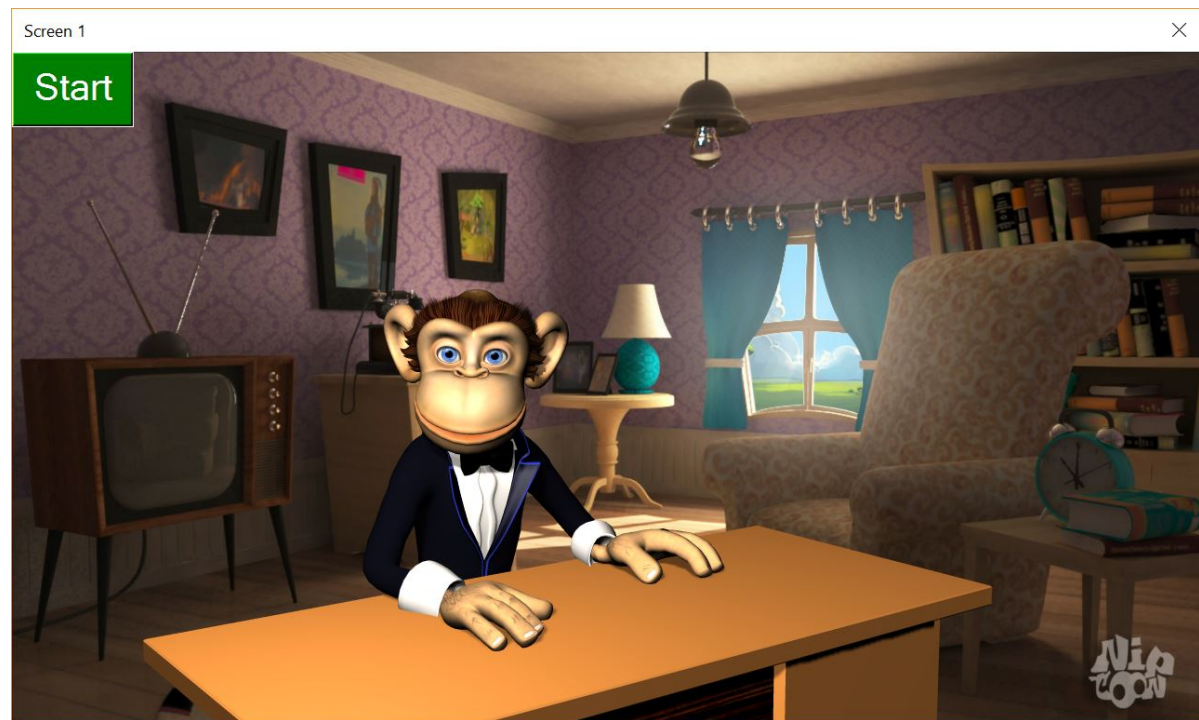


Figure 3.4: The final visual output with the virtual character Jimmy and the background setting of a room.

### 3.3. Pilot Study

We chose to run a pilot experiment for different purposes. First, we needed to operate a software testing to verify the overall performance of the system prototype and the functioning of the architecture. The verification process was used to check if the agent responses fit on the user reactions, ensuring a natural conversation flow. Secondly, as already anticipated in the previous section, we needed to collect interaction data from real-world user in order to fix some parameters, including values for four timers, two thresholds and one offset. Last, the Pilot Study was a also necessary step to test the setup and procedure for the final experimental study.

Eight participants for this study were formed by an opportunity sample of WonderMedia employees. Each of them interacted twice with the agent integrated with level 3 of "intelligence" as it was the most complete version of the system. A second conversation was appropriate in order to confirm the findings from the first one. After deciding a date, the users were invited to join at any time of the day which was most suitable to them. They were asked to talk about a "vacation story" for each of the interaction. We decided to use the recording studio room where the 3D agent is rendered on a 32 inch screen. The participants were asked to sit with a distance of 1.5 meters from the agent and to wear a pair of headphones with integrated microphone. The distance was switched between 1.5 (social distance) meter and 1 meter (personal distance) for each participant in order to observe users' reaction and gain knowledge about the most suitable value. It is stated that social interactions (among acquaintances) are usually carried out between 1.2 and 2.1 meters [72]. These interpersonal distances are highly variable (i.e. by cultural background) and mostly referred to "physical" presence of two or more interlocutors (human-human or human-robot [73]).

Initially the total time of interaction was fixed to 5 minutes (*timerGlobal*). The other initial values are listed in the Table 3.1.

While the participants were interacting with the agent, the experimenter followed the agent behavior through a console application to control remotely the overall situation to avoid major technical issues. The audio was recorded for further qualitative analysis. Immediately after each interaction, the users were asked to provide some constructive feedbacks about the overall quality of the interaction, their perception and feelings about the predefined topic. As all the participants were part of the production team in WonderMedia, they were quite familiar with the 3D character Jimmy and the technicality behind its animations. For this reason they were asked to not focus excessively on the agent appearances rather on its reactions while providing their personal opinions.

### 3.3.1. Results

Qualitative data was collected based on participants' impressions and experimenter's manual annotations. A rough analysis of speech rate (words per minute) was done wherever it was necessary. First we noticed that changes in distance between the screen and the participants did not effect severely the general interaction fluency and personal perceived quality. However almost everyone did find it too "invasive" when it was set to 1 meter. Therefore we chose the final value as the minimum required for a social interaction: 1.2 meters.

We noticed that the total conversation time of 5 minutes resulted to be too long for the users as majority of them said to "have nothing more relevant things to say" after around 3 minutes. One of the participants even dropped out from the interaction by clicking "STOP" button after that amount of time. Forcing the users to speak more could influence the flow of a real making it a priori fabricated and unnatural. Therefore we establish a new value of *timerGlobal* equal to 180 seconds, which was commenced when the user click the start button.

We observed that often the users went to a sort of "monologue mode". In fact a couple of participant got very few and scattered feedbacks as they were speaking without any pause. This was probably due to the fact that one participant was a native English speaker who did not need any time to organize the speech and the dialogue flow was profoundly fluent. In general this cannot allow the agent to capture silences and especially if the user is a mental state where they do not expect to be responded, he or she will keep talking. For this we introduced the parameter *maxWord* which was the maximum amount of words that the user could say without any pause. Based on the average Speech Rate and total content, we set the value of it to 150 words.

Based on these observations in addition to data provided by the participants about the specific topic and their perception, we are able to determine the total number of behavioral cues that the conversational agent needed to generate for simulating an attentive listening approach. Once established the amount of backchannels we could calculate how often they needed to be synthesized (*timerLevel1*) by simply spreading the cues pseudo-randomly throughout the entire interaction time. Instead of deciding a fixed value for this timer, we defined an interval between which it was selected after every previous backchannel generation. The lower limit of this interval was 30 seconds and the upper limit was bounded by 40 seconds.

One of the most challenging task was to change the appropriate value of *offset*. So far in order to infer the sentiment, we considered the whole speech between two pauses, counted the total number of positive and negative opinion words and then compared them. Of course "hard" margin was not the most appropriate. For example, consider the following statement: *"The hiking at the Dolomites was not that enjoyable"*. The system identified *"enjoyable"* as a positive valence but ignored the word *"not"* which beard same weight of opposite sentiment. As we kept *offset* at default value of zero, this statement was judged to be positive when it was clearly not. One solution could have been to simply add the word to the list of negative opinion keywords, but as we decided to not manually add any more data, we simply increment the value of the *offset*. This was still not a perfect fit as with increase of the number of total utterances the reliability declined considerably. Here one extreme scenario: *"That morning we were very energized and invigorated by an enjoyable and lukewarm weather. However while we were driving through those scenic and astonishing views, we got involved in a car accident and my boyfriend got injured. This ruined our whole honeymoon. which was supposed*

to be *unforgettable*". We can count 7 positive words and only 3 negative keywords. Here the system would consider the overall experience as "likable" which evidently was not. We did not consider this extreme cases as probably a different value of *offset* would not improve the result in these circumstances. Other strategy of emotion analysis, for example taking in consideration what emotion was expressed later in the sentence could have been relevant to response in these extreme cases.

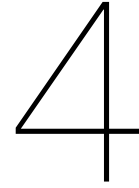
Finally *minWord*, minimum number of words that a user needed to say between two pauses, and *timerSilence*, amount of silence needed to consider the user in pause, were slightly modified. The former was reduced from 10 words to 8 words and the latter was increased of half second. The other timer, *timerSilence2* never reached zero during the different 16 interactions, as the user pronounced at least a word between two pauses. However we did not remove the timer for simple purpose of generality and we kept the value established previously.

An overall view of the final values of the parameters compared to the initial ones, is provided in Table 3.1.

Table 3.1: System parameters

Parameters		Value	
<i>Label</i>	<i>Description</i>	<i>Initial</i>	<i>Final</i>
timerGlobal	total time of interaction	300 sec	180 sec
timerSilence	await time for user's silence	1 sec	1.5 sec
timerSilence2	await time for user's reaction	15 sec	15 sec
timerLevel1	await time for generating BCs ( <i>level 1</i> )	<i>Not Applicable</i>	rand(30, 40) sec
minWord	words required between two pauses	10	8
maxWord	words allowed between two pauses	<i>Not Defined</i>	150
offset	used in emotion detection ( <i>level 3</i> )	0	1





# Evaluation

*The objective of this chapter is to present the methodology of the experimental study used to test the research hypotheses. In addition it provides the results obtained and their interpretation.*

## 4.1. Methods

The aim of the experiment was to test our research hypotheses (Section 2.4) that is to check whether participants found a virtual agent to be more likable and more engaging with higher levels of listening intelligence (Section 3.1). A within-subjects design was chosen, in which participants had to tell the avatar three vacation stories, one per condition, and the listening intelligence of the avatar was altered between the conditions. Therefore, for every condition (1, 2 and 3) the system displayed BCs corresponding to a specific level of intelligence (1, 2 and 3). In summary, in condition 1 the backchannels were generated at random timing, in condition 2 they were generated when the users were in silence like in condition 3, but in the latter condition, BCs were also based on emotional content of the user's speech.

To mitigate the carryover effect caused by factors, such as order, that could bias repeated-measures designs, we counter-balanced orderings in two ways: first, the order in which the user was exposed to the conditions was selected randomly just before the interaction. Second, also the vacation story to be told was assigned randomly. Ethical approval was granted for the study by the TU Delft Human Research Ethics Committee in June 2018.

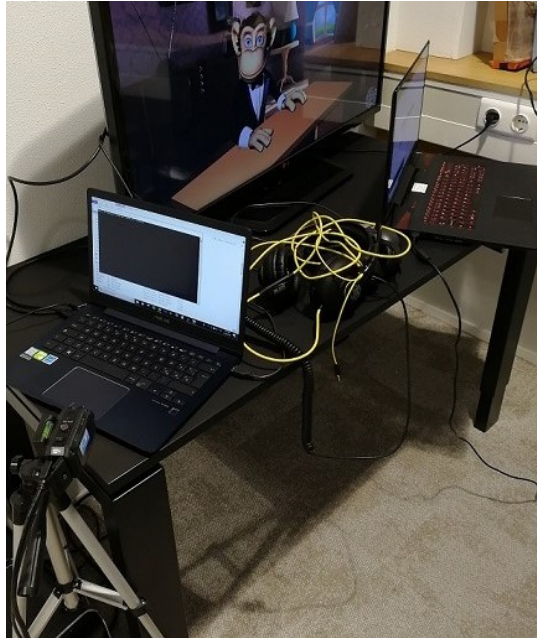
### 4.1.1. Experimental Setup

Two different locations were chosen for the study. The first one was located at Wonder Media company's headquarter in Antwerp, Belgium and the second one on the TU Delft university campus. For both of the locations, the experimental setting was kept as much similar as possible. A quiet room was selected, which in first scenario resulted in a recording studio at the first location and case in a laboratory (Insyght Lab) of the Faculty of Electrical Engineering, Mathematics and Computer Science at the second location. The final implementation of the system was released in three different executables (for the three different conditions) to be run on a laptop with the following specifications:

*CPU Intel i7-8550 1.99 GHz, RAM 8 GB, 64-bit, SSD 512 GB, NVIDIA GeForce MX150 with 2GB dedicated.*

The Embodied Conversational Agent, Jimmy, was displayed either on a 32-inch or a 39-inch monitor based on the specific location. In first case (Figure 4.1b), it was placed around 1.2 meters from where the speaker was asked to take a seat on a swivel chair. This distance was then increased to 1.4 meters when the screen size was found to be larger in the second case. In this way we could keep the same proportion between visualization area and human-agent distance. The proportion was inferred from qualitative analysis of our Pilot Study, where some participants found discomfort if the distance was reduced enough to invade "personal" space and greater distance brought more distractions from the surrounding environment.

A headset was provided in order to interact with the agent. The user could adjust the output volume using two buttons on the headset. Considering other instruments, a mouse was provided in order to trigger the start of the conversation clicking the button “Start”. An HD camera was used on top of a tripod to record the output command lines to check remotely the system behaviors and error managing. Last but not least, a second touchscreen laptop was provided to the users in order to fill in a questionnaire. Figure 4.1a shows all the tools and devices used during the experimental study.



(a) Tools used for experimental study



(b) Participant interacting with Jimmy

Figure 4.1: Pictures of experimental setup.

### 4.1.2. Subjects

The participants were formed by an opportunity sample of 24 English speaking adults over 18 years old. The volunteers (mean age: 26; men:15, women:9) were from one of the following academic locations: University of Antwerp (Belgium) and Delft University of Technology (Netherlands). As our dialogue system was in English, the potential people to be contacted were chosen among Master Students to assure an adequate proficiency level in both speaking and listening. In fact, both universities required a minimum of B2 score (Intermediate Level) for admission to their respective Master programs, which was assumed to be sufficient for interacting properly with the virtual conversational agent about the specific “vacation” topic, and in particular, to speak naturally for a longer duration of time without having to search for words.

The potential list of users were first contacted informally through word of mouth and then enlisted through an online doodle<sup>1</sup> module. Here they used a personal contact email address to register, to get a quick overview on the experimental study and to choose their preferred time slots (30 minutes each). Three days prior to the first experiment day at each location (Antwerp or Delft), individual emails were sent to all the participants of that specific region to provide detailed informations about the study. The number of interactions as well as the topic of “vacation” stories was revealed to the participants. They were also asked to prepare three stories with titles prior to arriving for the experiment. They were also instructed on how to reach the specific location of the experiment.

<sup>1</sup><https://doodle.com/it/>

### 4.1.3. Measurements

Likability was operationalized with the Dialogue Experience Questionnaire (DEQ) questionnaire. This questionnaire, attached in Appendix D, was developed to measure participants' experience of dialogues in dyadic conversations with avatars[74]. Each participant completed this questionnaire after each of the interaction conditions. The questionnaire was formed by 32 questions measuring the user experience on seven sub-dimensions: dialogue speed, interruption, correctness locally, correctness globally, involvement, discussion satisfaction and reality. The items were scores on a 7-point Likert scale ranging from strongly disagree (1) to strongly agree (7). To calculate the likability global score for each participant and condition, the arithmetical mean was calculated.

In addition to DEQ, another important stream of data has been collected for each participant: audio recordings. The speech data was used to count the number of words spoken. For this purpose, a free and open-source digital audio editor and recorder was used to collect data from the microphone <sup>2</sup>. To optimize the storage demand, we chose the Sound Activated Recording option with threshold of 10 decibel (dB) (i.e. barely audible). In other words, the audio stream was saved only when the user was actually speaking. At any other time the recording was switched to pause mode. The speech data from the interaction was processed through a software which made use of similar source code as the Speech Recognition module of the Embodied Conversational Agent. It had capabilities to read the *wav* format data files (three for each participant) and converted this to strings using client libraries from Bing Speech API. Then three main counting operations were done simultaneously: total number of words (*TotWords*), total number of emotional words (*TotEmotionalWords*) and total number of positive sentiment words (*TotPosWords*). For the last two counting operations, the keyword-spotting algorithm presented in section 3.2.2 was adapted. These data were then used to calculate Speech Rate, expressed in words per minute (*wpm*), Emotional Level, expressed in emotional words per minute (*e-wpm*), Emotion Ratio (*EmoRatio*), relating the total number of emotional words to the total number of spoken words, and Positive Emotion Ratio (*PosRatio*), relating the total number of emotional words with a positive valence to the total number of emotional words spoken. In the following the detailed calculation method of these dependent variables used to operationalize the level of user engagement. The increase in their values suggested an increase in the level of user engagement.

- $wpm = \frac{TotWords}{TotMinutes}$
- $e-wpm = \frac{TotEmotionalWords}{TotMinutes}$
- $EmoRatio = \frac{TotEmotionalWords}{TotWords}$
- $PosRatio = \frac{TotPosWords}{TotEmotionalWords}$

### 4.1.4. Procedure

On the day of the study, once the participants were welcomed at the reception desk, they were requested to fill in an Informed Consent form, which is provided in Appendix C. In order to keep the collected data anonymous, a participant ID was assigned to each of them. They were asked to have three interaction sessions in English with the virtual agent and topics of the discussions were the same: "vacation story" as it was already revealed to them through individual emails. Before the start of the experiment, they provided a list of titles to the researcher. The researcher chose one of the (remaining) titles for each condition at random. More specifically two operations of randomization were performed using an online random sequence generator <sup>3</sup>. First each of title was randomly matched with a specific condition. Then also the order of the condition was picked randomly.

Afterwards the participants were invited to the experimental room, asked to take a seat, to wear the headphones and to make themselves comfortable. After a brief microphone check, the experimenter launched the first executable and the conversational agent was immediately displayed on the main monitor. The researcher then left the room after telling the participants

<sup>2</sup><https://www.audacityteam.org/>

<sup>3</sup><https://www.random.org/sequences/>

which story he or she should talk about with the agent. The users were able to start the interaction by clicking "start" whenever they felt ready. The participants were encouraged to describe the story in as much detail as possible and to speak fluently in order to help the listener comprehend. During the interaction, the researcher could monitor live the output logs of the system and the agent visualization by using a camera facing both the laptop and the main screen. In this way the experimenter had always knowledge about the content of the conversation (thanks to the real-time Speech-To-Text software), the state of the agent, its behaviors, and he could manually annotate if something unexpected occurred. Furthermore, the experimenter was also informed by the system when each interaction time was over so he could return to the experiment room accordingly.

At the end of each interaction, while the participants were filling out the DEQ questionnaire using the second touchscreen laptop, the researcher set up the following interaction and made himself available for questions regarding the questionnaire. However, if the participants had any doubt or presented inquiry about the overall study and interrogated about the purpose of the study, they were requested to wait until they have concluded all the three interactions. They were also reminded that they were allowed to take any breaks or withdraw from the study anytime. When the final questionnaire was completed, clarifications and, if requested, the personal audio recordings, were provided. At the end of the experiment, participants were thanked for their time and received a cookie as a token of appreciation. No additional compensation was provided.

#### 4.1.5. Data Analysis

Apart from an Exploratory Data Analysis (EDA), the real potential of the collected measurements resided in the statistical evaluation. First we identified multi-level model [75] as the most suitable fit for our repeated measures as all the observations were nested within participants. In fact, each participant had influence on the final outcome value of each interaction. Real world cause of this effect could have been due to different levels of perception or personality. As these factors could not have been eliminated, they were accounted for the statistical analyses.

Multilevel or hierarchical model are an extension of standard linear models. We used the free statistical computing software R<sup>4</sup> for building the models. We started by fitting a null-model (Model 0) for each of the dependent variables. This baseline model did not take into consideration the influence of different conditions (1, 2 and 3) in order to predict new possible measurements and only considers the random effect of participants as random intercept, hence modeling each participant as their individual mean across conditions. We compared this model with an extended one presenting mixed effects, which added to random intercept due to participant, the fixed effect (slope) of conditions (Model 1). Given our hypothesis of greater likability of the agent and higher engagement with the agent for conditions with more intelligent listening behavior, we expect the fixed effect of condition to significantly predict likability and engagement with a positive coefficient.

When significant effects emerged in the model comparison, we also operated planned comparisons to determine where exactly the effect observed in main analysis was occurring and what the magnitude of differences was between the groups. Planned comparisons were adopted as we had made specific predictions about which group means should differ before any data was collected. In our case based on our hypotheses, we performed test on two different pair combinations: condition 1 in comparison with condition 2 and condition 2 compared to condition 3. The former one checked whether time was relevant in generating backchannels for improved interaction quality (**H1**). The latter checked whether emotion based backchannels were relevant for enhanced interaction quality (**H2**). As we assumed normal distributions, we chose to perform only parametric t-test. To have a quantitative measure of the magnitude of the difference, we calculated Cohen's *d* as it was an appropriate effect size for the comparison between two means [76].

---

<sup>4</sup><https://www.r-project.org/>

## 4.2. Results

The results of the multi-level analysis of all primary outcome measures are provided in Table 4.1. The essential statistical summary for different conditions are shown in Table 4.2.

Table 4.1: Multilevel analyses results of primary outcome measures (Model 0 vs. Model 1)

Variable	df	$\chi^2$	p-value
Likability (deq)	5	33.7	<.001
Engagement (wpm)	5	71.6	<.001
Engagement (e-wpm)	5	74.6	<.001
Engagement (EmoRatio)	5	32.4	<.001
Engagement (PosRatio)	5	1.8	0.408

Table 4.2: Statistic summary of the outcome measurements

Attributes	Condition 1		Condition 2		Condition 3	
	Mean	SD	Mean	SD	Mean	SD
Likability (deq)	3.39	0.64	4.55	0.81	4.44	0.78
Engagement (wpm)	69.04	21.15	116.96	27.31	110.04	31.61
Engagement (e-wpm)	2.71	1.20	5.96	2.01	6.21	1.79
Engagement (EmoRatio)	4.20	1.41	5.17	1.42	5.91	1.55
Engagement (PosRatio)	76.83	12.59	74.58	13.46	78.96	11.29

### 4.2.1. Likability

Likability was chosen to be represented by the average 7-point Likert score on the Dialogue Experience Questionnaire. The addition of fixed effect due to different conditions to a null-model, significantly improved the model's fit. This confirmed that the model was able to better predict the likability if it had knowledge about which condition was applied for specific measurements.

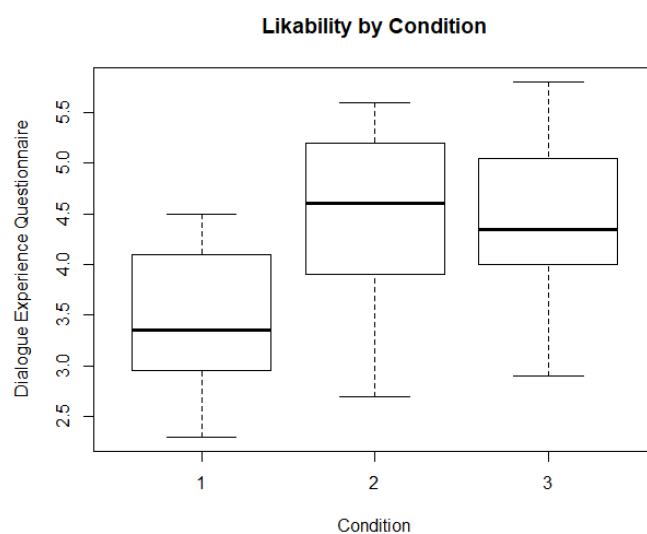


Figure 4.2: Likability distribution by condition

As model comparison showed significant effects for condition, we continued with planned comparisons. The pair-by-pair comparison confirmed that there was a significant difference between condition 1 and 2 in terms of likability ( $t(23) = 5.8, p < .001, d = 1.6$ ) but not when comparing condition 2 and 3 ( $t(23) = -0.6, p = 0.52, d = -0.1$ ). For a better understanding of the underlying data distribution, Table 4.2 provides the statistical summary and Figure 4.2 visualizes the distribution in a boxplot.

#### 4.2.2. Engagement

**Speech Rate** The extension to Model 1 showed significant improvement. As it can be observed from Figure 4.3a, the distribution relationship between condition 1 and 2 was similar to that of likability and a t-test showed that there was a large effect on Speech Rate ( $t(23) = 9.9, p < 0.001, d = 1.9$ ) However from the boxplots is not clear if the backchannels produced in condition 3 has effects on users' speech rate compared to the ones generated in condition 2. A paired t-test between samples from conditions 2 and 3 ( $t(23) = -2.5, p = 0.018, d = -0.2$ ) showed that there there was actually a significant difference in speech rate, but with only a small effect.

**Emotional Level** The model which considered different class categories to predict new possible outcomes performed better than Model 0 (Table 4.1). Planned comparisons showed a significant and large effect on Emotional Level when comparing condition 1 and 2 ( $t(23) = 9.9, p < 0.001, d = 2.0$ ). A paired t-test was used to verify that there was no significant difference between the mean value of condition 2 and mean value of condition 3 ( $t(23) = 0.8, p = 0.42, d = 0.13$ ). These results can also be seen from Figure 4.3b.



Figure 4.3: Distribution of each class category (Speech Rate and Emotional Level)

**Emotion Ratio** For the percentage of emotional words in proportion to total number of words spoken by each users, the addition of fixed effect due to different conditions, improved significantly the model's fit. The computed values of paired t-tests also confirmed the significant difference between condition 1 and 2 ( $t(23) = 4.1, p < 0.001, d = 0.6$ ) and between condition 2 and 3 ( $t(23) = 2.6, p = 0.013, d = 0.5$ ). Figure 4.4a illustrates the distributions.

**Positive Emotion Ratio** For the other ratio scale measure, *PosRatio*, Model 1 did not have condition as a significant predictor and the extension to it did not show any significant improvements in the model's fit. Therefore detailed analyses were not conducted. In Figure 4.4b it could be observed that the distributions are largely overlapping.

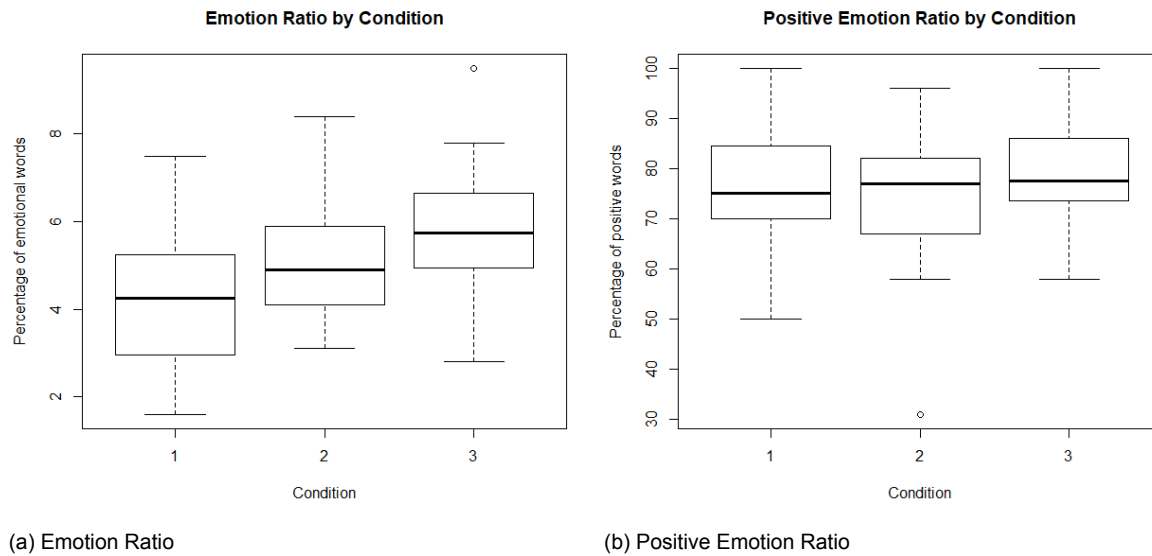


Figure 4.4: Distribution of each class category (Emotion Ratio and Positive Emotion Ratio)

### 4.3. Discussion

So far we reported all the quantitative results obtained through the experimental study. In this section we focus on detailed interpretation of the results and the main limitations behind this evaluation.

First, according to the results, the users found the interaction more likable and engaging if the backchannels were generated with correct timing. This could be justified as in real world scenario if the listener talks over the speaker, it could be a signal that the speaker is not being “listened to” by the conversational partner. Furthermore, if not perfectly timed (condition 1), the head nods and shakes were more disruptive than helpful. As Chiu et al. point out [77], this could have been due to the fact that speakers could not ignore these feedbacks and their concentration on the story decreased.

Second, the extension to emotion understanding (condition 3) brought either mixed or no significant effect on users (compared to condition 2) in terms of likability and engagement. The most clear hypothesis supporting trend for engagement was found in the positive effect on emotion ratio of words. However, the increase between conditions 2 and 3 was caused by a relative decrease in speech rate rather than a relative increase in emotional words. A possible explanation for the decrease in speech rate may be that the agent did not always interpret the sentiment correctly, potentially breaking the flow of participants stories and causing them to have to rethink, explain themselves, and possibly decreasing their desire to disclose more. For example, if someone at certain point of the conversation had remembered a bad accident he or she had during the last vacation and the agent had replied “*it seems you liked the experience*”, the user would have reacted adversely. In fact, during the debriefing after the experiment, few subjects pointed out that they intentionally kept their stories short when the listener seemed to not listen to them, pointing out most probably the exposure to the condition 1 or condition 3.

In order to have an understanding on what were the chances that the agent could have done a wrong sentiment analysis, we checked the sub-dimensions of DEQ questionnaire. In particular we focused on *correctness globally*, that was correctness of the entire dialogue and consistency between the different question lines, *correctness locally*, which took into account the correctness of the reactions from the agent on the user’s responses and *discussion satisfaction*, i.e. the feeling the user got during the question phase and how the user experienced the answers and attention from the avatar. Although there was no relevant difference between condition 2 and 3 in the mean of scores for the first sub-dimension, for the latter one

there was an evident decrease for condition 3. In fact the average score for the *correctness locally* sub-dimension decreased from 4.6 to 4.0 when the agent displayed feedbacks based on users' emotional content. This could have been caused by mis-interpretation of the overall positive and negative weight of the vacation story.

The Positive Emotion Ratio was entirely uninfluenced by the backchanneling strategy. This latter measurement variable was also independent on different backchannel strategies. Furthermore, it was the only measurement of engagement level which also in case of attentive listening cues generated at random timing did not change significantly.

Finally, from point of view of methodology, the multi-model analyses and planned comparisons assumed the distributions to be normally distributed, which was not always legitimate. In fact, running the Shapiro-Wilk test we noticed that for Emotional Level, the difference between the samples of the two groups were not normally distributed. In this case the multilevel modeling could not have been considered suitable and for the detailed testing we should have adopted the non-parametric paired (Wilcoxon) t-test. However, in terms of findings, it was not a limitation as the results obtained through Wilcoxon test were the same.

### 4.3.1. Limitations

According to the evaluation, condition 2 presented the best performance and could be considered the preferred setup for our conversational agent. However, the overall results showed some limitations compared to a real-world interaction. For example, if we consider the dependent variable speech rate for a spontaneous conversation in English, its value is found to be between 5 and 6 syllables per second ([78]) or between 150 and 190 words per minute ([79]) where the main interquartile range of our results spreads from 100 to 140 wpm. Although there could have been many factors influencing these results, we point out one of our limitation. During the system design phase, we decided to identify the user's silence through Voice Activity Detection and Speech Recognition. These modules were designed to detect "instant" silence and the use of a simple timer (*timerSilence*) was not most appropriate to understand the "quantity" of silence needed to consider the user in pause. This was probably caused by the fact that the chosen parameter was highly variable due to different personalities, cultural backgrounds and also the level of English proficiency. In fact, a user with low proficiency could have needed more time to search for words to describe a specific part of the story and therefore, the timer value should have been greater.

Another important parameter that had been modified after the pilot study (Section 3.3), was *timerGlobal* which expressed the total time of interaction. We remark that the choice to reduce its value from five to three minutes was probably a poor choice. We noted that in the case of the pilot study, the fact that users "*did not have anything more to say*" was apparently linked to the fact that they did not have enough time to collect their memories and prepare the story to be told. On the contrary, some of the participants who were recruited for the experimental study admitted that the interaction sessions seemed "*to be quite short*" as they "*did not have time to say everything*". One could observe that this could have been due to the fact that they were instructed about the topic of the conversation a priori and probably some of them prepared the structure of their vacation stories beforehand. However, in the confirmation email, we also mentioned that "*each session with the virtual agent will take approximately three minutes*". Anyway, it is possible that the choice of reducing *timerGlobal* was not suitable, as the participants of the pilot study were not exposed to the exact same conditions as the ones from the experimental study.

Finally, in 4.3 we analyzed the sub-dimensions of the Dialogue Experience Questionnaire in order to have an idea on how often the agent did not correctly analyze users' emotion (condition 3). However, for a more precise check, a post-experimental analysis should have been run with the entire recording, including the pauses, from each interaction session. Unfortunately this was not possible in our case as the recording software cut all the "not audible sounds" (i.e. silences) as a Sound Activated Recording was imposed (see 4.1.1). This choice was initially made to optimize storage requirements, but we realized in a later phase that the impact would have been imperceptible, compared to the potential benefits in the post-analysis phase.



# 5

## Discussion & Conclusion

*The objective of this chapter is to provide the main findings of the evaluation and verdicts on research hypotheses. It also presents the limitations behind this work, proposes direction for future works and finally, draws the conclusions.*

### 5.1. Main Findings

In the context of mental health related disorders, such as depression, anxiety and stress, self-help interventions based on Embodied Conversational Agents are expected to bring great benefits ([86], [22]). More specifically in the setting of a dyadic face-to-face interaction, the category of influence of appropriate attentive listening behaviors, in particular backchannels, was interrogated in this research. We hypothesized that these feedbacks, if were generated with suitable timing and correct emotional content understanding, would have a positive impact on the quality of the interaction in terms of user likability and engagement. These two factors were operationalized through five different dependent variables: Dialogue Experience Questionnaire score, Speech Rate, Emotional Level, Emotion Ratio and Positive Emotion Ratio.

Analyzing these outcome measurements, we observed that these rule-based backchannels did influence human speakers while interacting with the conversational agent [62]. However, the benefit of increasing backchannel intelligence could not be observed beyond the condition where the backchannels were generated at correct timing.

Correct timing, which was defined as a suitable moment when the user was silent, was found to be an important feature to be considered when implementing attentive listening behaviors into an ECA system as we hypothesized (**H1**). In fact, both likability and engagement measurements (apart from the Positive Emotion Ratio), showed significant increase in the condition where the users were exposed to correctly timed backchannels compared to the condition where these signals were randomly generated [50].

Our second hypothesis (**H2**) could be rejected on the basis of the evidence gathered in this study. While we had hypothesized an increase in both likability and engagement, no difference was found in likability and engagement either decreased (speech rate) or remained constant across the two conditions (emotional level, positive emotion ratio).

### 5.2. Limitations & Future Works

For detecting user silences and the emotional valence of their speech content, our approach could benefit from the extension to a multimodal one. In the field of computer vision, mouth opening and gaze interpretation could enhance the accuracy of the Voice Activity Detection module implemented in our conversational agent ([80]). Furthermore, multimodal strategies could be adopted for detecting other facial and nonverbal features relevant for correct sentiment analysis ([33], [58]). Alternatively, in the field of machine learning, a system could be trained with dyadic conversational data regarding specific topic, such as vacation, hobby and family, to perform sentiment analysis based on different conversational patterns.

Choices made to simplify the system requirements, such as fixing the total conversation time, may have reduced the external validity of our study. In fact, real-world interlocutors require a certain time interval, which is variable from person to person, to establish some degree of mutual trust and rapport before reaching their average speaking rate [81]. Some other choices on experimental setup, such as using a quiet room, keeping the human-agent distance and the conversational topic fixed, may have influenced our results, too. However, these choices could be justified for simulating attentive listening behaviors in the context of a dyadic psychotherapy session. Future works could test our system and methodology in this context by studying the effectiveness on a systematic sample of clinical population.

Finally, although we considered our dialogue manager mixed-initiative, for each dialogue state, our conversational agent asked closed-ended questions which did not give the user freedom in his or her response. The system always expected the user to talk about the predefined topic, in a predefined structure. Although for our research purpose it was assumed to be acceptable, further works can improve the agent capabilities to pursue “*small talk*”, which is expected to build rapport and credibility [81]. We believe that a more user-centered approach might lead to a better level of engagement with the users without forcing them to speak, which *per se* is unnatural. For this target, it is recommended to integrate the current dialogue engine with libraries from Google DialogFlow<sup>1</sup> or Microsoft Bot APIs<sup>2</sup> for complex dialogue management. These natural language processing units provide tools for developing complex “chatbot-style” systems, scalable with data for automatic learning and adaptable to our conversational agent.

### 5.3. Conclusion

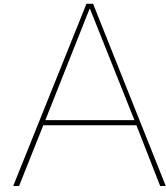
The main research purpose of this thesis was to study if attentive listening behaviors, more specifically backchannels, could influence the overall quality of human-agent interaction. In particular, it was hypothesized that there would be a positive correlation between the backchannel intelligence and interaction quality. The results showed that backchannels generated with proper timing had a positive impact on both user likability and engagement. On the other hand, backchannels based on emotional content of users speech had either mixed effects or no significant influence on the quality of the interaction.

In conclusion, this work considered relevant aspects in the field of self-help mental health related interventions based on Embodied Conversational Agents. This research used a multimodal approach to backchannel generation and studied their influence on an opportunity sample of users in the setting of a dyadic human-agent interaction. Additional developments could be pursued in future works, but this thesis provides relevant evidence of the importance of timing when simulating verbal and nonverbal attentive listening.

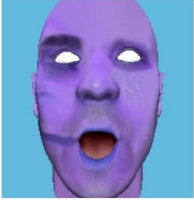

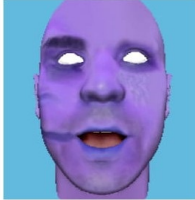


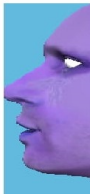
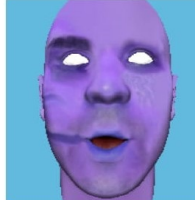



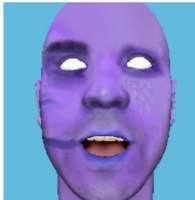
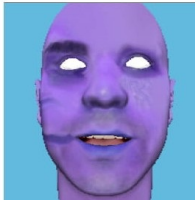
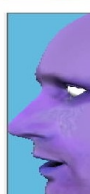
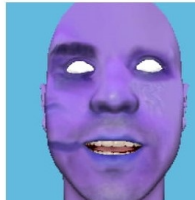

---

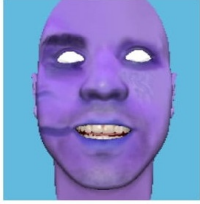

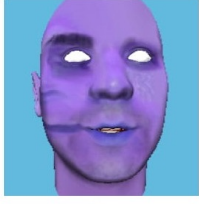


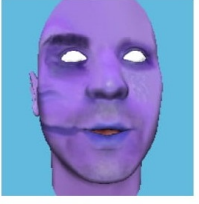

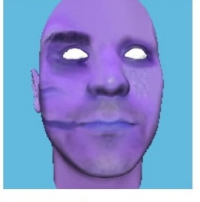

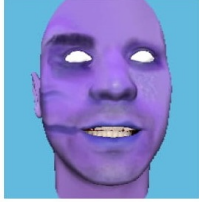

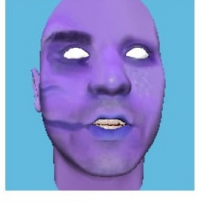

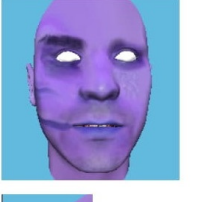
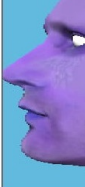

<sup>1</sup><https://dialogflow.com/>

<sup>2</sup><https://dev.botframework.com/>



# Appendix

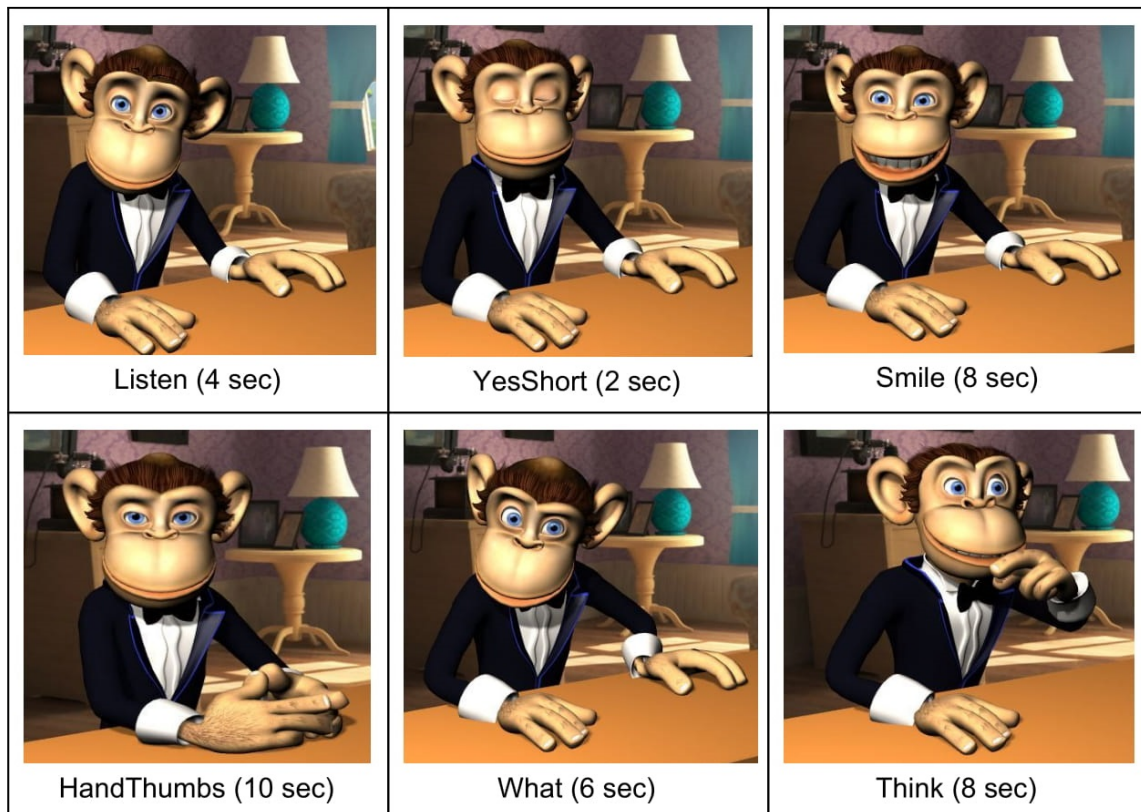
 	 	 	 
AA - <b>odd</b> , <b>adah</b>	AH, h - <b>adapt</b> , <b>marshah e</b>	AO - <b>score</b> , <b>ought</b>	AW OW - <b>cow, oats</b>
 	 	 	 
OY UH UW- <b>toy, tou</b> gh, two	EH, AE - <b>Ted</b> , <b>cat</b>	IH,AY - <b>hit</b> , <b>Hide</b>	EY - <b>ate, gate</b>

  y, IY - yes, yum, eat	  r, ER- ranger	  l - loud, unload	  w - would, unwind
  m,p,b	  n,NG,DH,d,g,t, z,ZH,TH,k, s	  CH, j, SH	  f,v
 X			

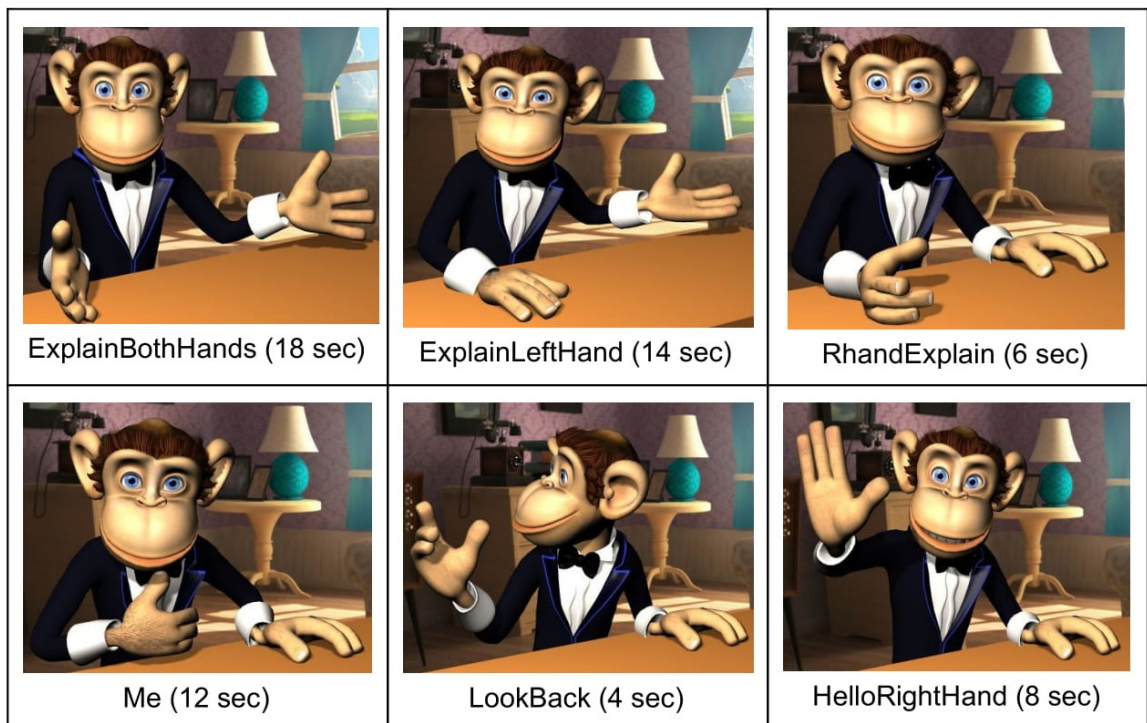
# B

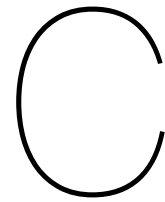
## Appendix

### Regulators (back-channels)



## Illustrators





# Appendix

## Informed Consent (1/2)

### TITLE OF STUDY

Virtual Conversational Agent for Mental Health Intervention.

### PRINCIPAL INVESTIGATOR

Mehedi Anam Sarder

Electrical Engineering, Mathematics and Computer Science - TU Delft

Van Mourik Broekmanweg 6, 2628 XE Delft

+39 3479163318 - [m.a.sarder@student.tudelft.nl](mailto:m.a.sarder@student.tudelft.nl)

### PURPOSE OF STUDY

You are being asked to take part in a research study. Before you decide to participate in this study, it is important that you understand why the research is being done and what it will involve. Please read the following information carefully. Please ask the researcher if there is anything that is not clear or if you need more information. The purpose of this study is to observe the interaction between a human participant and a virtual agent in the setting of a dyadic conversation.

### STUDY PROCEDURES

You are asked to interact three times for approximately 5 minutes with a virtual avatar displayed on a monitor. You are requested to put on the headphones and start the interaction by simply clicking on “Start”. A microphone will be used to process real-time your verbal input to improve the interaction with the virtual agent. The speech data will also be recorded for data analysis. At the end of each interaction, you will be asked to fill in a questionnaire.

### RISKS

We do not expect any risks to be associated with the participation in this experiment.

### BENEFITS

There will be no direct benefit to you for your participation in this study.

### CONFIDENTIALITY

Your responses to the final survey will be anonymous. Please do not write any identifying information on your paper. Every effort will be made by the researcher to preserve your confidentiality.

### Informed Consent (2/2)

The following measures are also taken to ensure confidentiality:

- You will be assigned a CODE number and that will be used on all research notes and documents.
- The audio material will be available only to the researcher, the supervisor and, if necessary, the thesis committee. If requested, you can have access to this (.wav) file right after you complete the experiment.
- After the project is over (September 2018), all the audio recordings will be deleted.
- NO video material will be recorded!

### CONTACT INFORMATION

If you have questions at any time about this study, or you experience adverse effects as the result of participating in this study, you may contact the researcher whose contact information is provided on the first page.

### VOLUNTARY PARTICIPATION

Your participation in this study is voluntary. It is up to you to decide whether or not to take part in this study. If you decide to take part in this study, you will be asked to sign this consent form. After you sign the consent form, you are still free to withdraw at any time and without giving a reason. Withdrawing from this study will not affect the relationship you have, if any, with the researcher. If you withdraw from the study before data collection is completed, your data will be returned to you or destroyed.

---



---

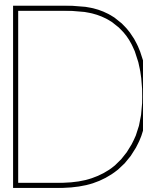
### CONSENT

*I have read and I understand the provided information and have had the opportunity to ask questions. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving a reason and without cost. I understand that I will be given a copy of this consent form if I request it. I voluntarily agree to take part in this study.*

Participant's signature \_\_\_\_\_ Date \_\_\_\_\_

Investigator's signature \_\_\_\_\_ Date \_\_\_\_\_





# Appendix

## EnglishDEQ

### Flow: dialogue speed

No	Statement	Strongly disagree	1	2	3	4	5	6	7	Strongly agree
1*	The discussion partners needed a long time to think		0	0	0	0	0	0	0	
2*	The discussion partners often went quiet		0	0	0	0	0	0	0	
3*	On occasions I had to wait long for a reaction of the discussion partners		0	0	0	0	0	0	0	
4*	The conversation did not run smoothly		0	0	0	0	0	0	0	

### Flow: interruption

No	Statement	Strongly disagree	1	2	3	4	5	6	7	Strongly agree
1	I was always able to finish		0	0	0	0	0	0	0	
2*	On occasions I was unable to tell everything that I would like to have told		0	0	0	0	0	0	0	
3*	On occasions the discussion partners talked before their turn		0	0	0	0	0	0	0	
4*	On occasions, the discussion partners started talking while I was talking		0	0	0	0	0	0	0	
5	I got enough time from the discussion partners to explain everything calmly		0	0	0	0	0	0	0	

### Flow: correctness locally

No	Statement	Strongly disagree	1	2	3	4	5	6	7	Strongly agree
1	The discussion partners addressed my answers		0	0	0	0	0	0	0	
2	The discussion partner responded to my answers		0	0	0	0	0	0	0	
3	I got the feeling that the discussion partners understood my answers		0	0	0	0	0	0	0	
4	The questions had a logical order		0	0	0	0	0	0	0	



# Bibliography

- [1] <https://www.nimh.nih.gov/health/statistics/mental-illness.shtml>
- [2] Center for Behavioral Health Statistics and Quality. (2016). 2015 National Survey on Drug Use and Health: Detailed Tables. Substance Abuse and Mental Health Services Administration, Rockville, MD.
- [3] Spek V, Cuijpers P, Nyklicek I, Riper H, Keyzer J, Pop V. Internet based cognitive behaviour therapy for symptoms of depression and anxiety: a metaanalysis. *Psychol Med.* 2007 Mar; 37(3):319–28
- [4] Richards D, Richardson T. Computer based psychological treatments for depression: a systematic review and metaanalysis. *Clin Psychol Rev.* 2012 Jun; 32(4):329–42.
- [5] The Silent Shortage: How Immigration Can Help Address the Large and Growing Psychiatrist Shortage in the United States. Retrieved from: <https://research.newamericaneconomy.org/report/the-silent-shortage-how-immigration-can-help-address-the-large-and-growing-psychiatrist-shortage-in-the-united-states/>
- [6] Joseph Weizenbaum. ELIZA—A Computer Program For the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM* Volume 9, Number 1 (January 1966): 36-35.
- [7] Michael L. Mauldin (1994). CHATTERBOTS, TINYMUDS, and the Turing Test: Entering the Loebner Prize Competition. Conference: Conference: Proceedings of the 12th National Conference on Artificial Intelligence, Seattle, WA, USA, July 31 - August 4, 1994, Volume 1.
- [8] Margaret Boden (2006). *Mind as Machine: A History of Cognitive Science*, University of Sussex. ISBN: 9780199241446
- [9] Wallace, R. S. (2007). The Anatomy of A.L.I.C.E. Parsing the Turing Test Retrieved 12-12-08, 2008, from <http://www.alicebot.org/anatomy.html>
- [10] Wallace, R. S. - Be Your Own Botmaster. ALICE A. I. Foundation, 2004.
- [11] 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.
- [12] Hutchens, J. L., & Alder, M. D. (1999). Introducing MegaHAL. Workshop on Human Computer Conversation.
- [13] Isbister K, Doyle P. The blind men and the elephant revisited evaluating interdisciplinary ECA research. In: Ruttkay Z, Pelachaud C, editors. From brows to trust evaluating embodied conversational agents. Dordrecht, Netherlands: Springer; 2004. pp. 3–26.
- [14] 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, Pittsburgh, PA.
- [15] Je Rickel and W. Lewis Johnson (1998). STEVE: A Pedagogical Agent for Virtual Reality. Information Sciences Institute & Computer Science Department, University of Southern California.

- [16] Kenny, P., Parsons, T.D., Gratch, J., Leuski, A., Rizzo, A.A.: Virtual Patients for Clinical Therapist Skills Training. In: Pélachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 197–210. Springer, Heidelberg (2007).
- [17] Boye, J. and Gustafson, J. (2005) How to do dialogue in a fairy-tale world. In Proceedings of the 6th SIGDial workshop on discourse and dialogue.
- [18] Bernsen, N. and Dybkjær, L. (2004) Evaluation of Spoken Multimodal Conversation. In Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI'04), Pennsylvania, USA, pages 38-45.
- [19] Radosław Niewiadomski, Elisabetta Bevacqua, Maurizio Mancini, Catherine Pelachaud (2009). Greta: an interactive expressive ECA system. Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009), Decker, Sichman, Sierra, and Castelfranchi (eds.), May, 10–15., 2009, Budapest, Hungary.
- [20] Bickmore, T. (2003) Relational Agents: Effecting Change through Human–Computer Relationships. Thesis (PhD), MIT Media Arts and Science.
- [21] Corradini, A., Fredriksson, M., Mehta, M., Königsmann, J., Bernsen, N. O., & Johansson, L. (2004). Towards believable behavior generation for embodied conversational agents. In M. Bubak, G. D. VanAlbada, P. M. A. Sloot & J. J. Dongarra (Eds.), Computational Science - Iccs 2004, Pt 3, Proceedings (Vol. 3038, pp. 946-953). Berlin: Springer-Verlag Berlin.
- [22] F. Matcham, L. Raynera, J. Hutton, A. Monk, C. Steel, M. Hotopf. Self-help interventions for symptoms of depression, anxiety and psychological distress in patients with physical illnesses: A systematic review and meta-analysis. *Clinical Psychology Review*. Volume 34, Issue 2, March 2014, Pages 141-157.
- [23] Christensen, H., Griffiths, K.M., Farrer, L., 2009. Adherence in internet interventions for anxiety and depression: systematic review. *J. Med. Internet Res.* 11 (2), e13.
- [24] T. Kawahara, Miki Uesato, Koichiro Yoshino, Katsuya Takanashi. Toward Adaptive Generation of Backchannels for Attentive Listening Agents. Kyoto University, Japan.
- [25] G. Merola, “The Effects of the Gesture Viewpoint on the Students’ Memory of Words and Stories,” Proc. Gesture Workshop, pp. 272- 281, 2007.
- [26] I. Poggi, *Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication*. Weidler Buchverlag, 2007.
- [27] S. Campanella and P. Belin, “Integrating Face and Voice in Person Perception,” *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 535-543, 2007.
- [28] R. Campbell, “The Processing of Audio-Visual Speech: Empirical and Neural Bases,” *Philosophical Trans. Royal Soc. London—B Biological Sciences*, vol. 363, no. 1493, pp. 1001-1010, 2007.
- [29] Hollon SD, Muñoz RF, Barlow DH, Beardslee WR, Bell CC, Bernal G, et al. Psychosocial intervention development for the prevention and treatment of depression: promoting innovation and increasing access. *Biol Psychiatry* 2002; 52(6):610-630.
- [30] Warmerdam L, van Straten A, Twisk J, Riper H, Cuijpers P. Internet-based treatment for adults with depressive symptoms: randomized controlled trial. *J Med Internet Res* 2008;10(4):e44.
- [31] Eysenbach, G. (2005). The law of attrition. *Journal of Medical Internet Research*, 7(1), e11.

- [32] Azy Barak, Liat Hen, Meyran Boniel-Nissim & Na'ama Shapira (2008) A Comprehensive Review and a Meta-Analysis of the Effectiveness of Internet-Based Psychotherapeutic Interventions, *Journal of Technology in Human Services*, 26:2-4, 109-160.
- [33] Ekman, P. (1999). Emotional and Conversational Nonverbal Signals. In Messing, L. S. & Campbell, R. (Eds.), *Gesture, Speech, and Sign* (pp. 45-55). New York, NY: Oxford University Press.
- [34] Ronald Poppe, Khiet P. Truong, Dennis Reidsma, and Dirk Heylen; Backchannel Strategies for Artificial Listeners; Human Media Interaction Group, University of Twente, 2010.
- [35] Bavelas, J. B., Black, A., Lemery, C. R., and Mullett. Experimental Microanalysis of Addressees in Face-to-face Dialogue. J. Department of Psychology, University of Victoria, Columbia, Canada (1986).
- [36] Rogers, C. R. A theory of therapy, personality, and interpersonal relationships, as developed in the client-centered framework. In S. Koch (Ed.), *Psychology: A Study of a Science: Vol. 3* (pp. 184-256). New York, NY: McGraw-Hill. (1959).
- [37] Cuff, B., Brown, S. J., Taylor, L., & Howat, D. (2014). Empathy: A review of the concept. Department of Psychology and behavioural Sciences, Coventry University, Priory Street.
- [38] Yngve, V.: On getting a word in edgewise. In: *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pp. 567-577 (1970).
- [39] Allwood, J., Nivre, J., Ahlsn, E.: On the semantics and pragmatics of linguistic feedback. *Semantics* 9(1) (1993).
- [40] Heylen, D., Bevacqua, E., Tellier, M., Pelachaud, C.: Searching for prototypical facial feedback signals. In: Pelachaud, C., Martin, J.-C., Andr e, E., Chollet, G., Karpouzis, K., Pel e, D. (eds.) *IVA 2007. LNCS (LNAI)*, vol. 4722, pp. 147-153. Springer, Heidelberg (2007).
- [41] Gardner, R.: Between Speaking and Listening: The Vocalisation of Understandings. *Applied Linguistics* 19(2), 204-224 (1998)
- [42] Bavelas, J.B., Coates, L., Johnson, T.: Listeners as co-narrators. *Journal of Personality and Social Psychology* 79(6), 941-952 (2000).
- [43] Dittmann, A.T., Llewellyn, L.G.: The phonemic clause as a unit of speech decoding. *Journal of Personality and Social Psychology* 6(3), 341-349 (1967).
- [44] Duncan Jr., S.: On the structure of speaker-auditor interaction during speaking turns. *Language in Society* 3(2), 161-180 (1974).
- [45] Gravano, A., Hirschberg, J.: Backchannel-inviting cues in task-oriented dialogue. In: *Proceedings of Interspeech*, Brighton, UK, pp. 1019-1022 (September 2009).
- [46] Cathcart, N., Carletta, J., Klein, E.: A shallow model of backchannel continuers in spoken dialogue. In: *Proceedings of the Conference of the European chapter of the Association for Computational Linguistics*, Budapest, Hungary, vol. 1, pp. 51-58 (2003).
- [47] Bertrand, R., Ferr e, G., Blache, P., Espesser, R., Rauzy, S.: Backchannels revisited from a multimodal perspective. In: *Proceedings of Auditory-visual Speech Processing*, Hilvarenbeek, The Netherlands, pp. 1-5 (August 2007).
- [48] Noguchi, H., Den, Y.: Prosody-based detection of the context of backchannel responses. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, pp. 487-490 (November 1998).
- [49] Okato, Y., Kato, K., Yamamoto, M., Itahashi, S.: Insertion of interjectory response based on prosodic information. In: *Proceedings of the IEEE Workshop Interactive Voice Technology for Telecommunication Applications*, Basking Ridge, NJ, pp. 85- 88 (1996).

- [50] Ward, N., Tsukahara, W.: Prosodic features which cue backchannel responses in English and Japanese. *Journal of Pragmatics* 32(8), 1177–1207 (2000).
- [51] Thorisson, K.R.: *Communiative Humanoids: A Computational Model of Psychosocial Dialogue Skills*. PhD thesis, MIT Media Laboratory (1996)
- [52] Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating rapport with virtual agents. In: Pelachaud, C., Martin, J.-C., Andr e, E., Chollet, G., Karpouzis, K., Pel e, D., et al. (eds.) *IVA 2007. LNCS (LNAI)*, vol. 4722, pp. 125–138. Springer, Heidelberg (2007).
- [53] Morency, L.-P., de Kok, I., Gratch, J.: A probabilistic multimodal approach for predicting listener backchannels. In: *Autonomous Agents and Multi-Agent Systems* (2009).
- [54] Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., Stocksmeier, T.: Modeling embodied feedback with virtual humans. In: Wachsmuth, I., Knoblich, G. (eds.) *ZiF Research Group International Workshop*.
- [55] Jeremy N. Bailenson, Kim Swinth, Crystal Hoyt, Susan Persky. *Embodied-Agent Appearance and Behavior on Self-Report, Cognitive, and Behavioral Markers of Copresence in Immersive Virtual Environments*. Massachusetts Institute of Technology. August 2005.
- [56] Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L. and Stent, A. (2001) *Toward Conversational Human-Computer Interaction*. In *AI Magazine*, 22(4), pages 27–37.
- [57] Talavera, J. A., S az-Ruiz, J., & Garc a-Toro, M. (1994). Quantitative measurement of depression through speech analysis. *European Psychiatry*, 9(4), 185-193.
- [58] J. Russell, J. Bachorowski, and J. Fernandez-Dols, “Facial and Vocal Expressions of Emotion,” *Ann. Rev. Psychology*, vol. 54, no. 1, pp. 329-349, 2003.
- [59] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, “A Survey of Affect Recognition Methods: Audio, Visual and Spontaneous Expressions,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39-58, Jan. 2009.
- [60] Goodwin, C. (1981). *COnversational Organization: interaction between speakers and hearers*. New York: Academic Press.
- [61] Martin, J.C., Paggio, P., Kuehnlein, P., Stiefelhagen, R., Pianesi, F.: Introduction to the special issue on multimodal corpora for modeling human multimodal behavior (2007).
- [62] Poppe, R., Truong, K.P. & Heylen, D. Perceptual evaluation of backchannel strategies for artificial listeners. *Autonomous agents and multi-agent systems*. Page: 235-253. (2013).
- [63] Kenny, P., Hartholt, A., Gratch, J., Swartout, W., Traum, D., Marsella, S., Piepol, D.: *Building Interactive Virtual Humans for Training Environments*. (2007)
- [64] Cassell, J., Vilhj almsson, H.H., Bickmore, T.: *BEAT: the Behavior Expression Animation Toolkit*. In: Prendinger, H., Ishizuka, M. (eds.) *Life-Like Characters*, pp. 163–185. Springer, Heidelberg (2004).
- [65] Bickmore, T.W., Picard, R.W.: Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput. Hum. Interact.* 12, 293–327 (2005)
- [66] Morency L. P., Stratou A., DeVault D., Hartholt A., Lhommet M., Lucas G., Morbini F., Georgila K., Scherer S., Gratch J., Marsella S., Traum D., Rizzo A. (2015). *SimSensei Demonstration: A Perceptive Virtual Human Interviewer for Healthcare Applications*. Institute for Creative Technologies, University of Southern California.
- [67] Cassell, Justine; Prevost, Scott; Sullivan, Joseph; Churchill, Elizabeth (2000), *Embodied Conversational Agents*, Cambridge, MA: MIT Press.

- [68] Lawrence, E.A. (1985). Human perceptions of animals and animal awareness: The cultural dimension. In M.W. Fox & L.D. Mickley (Eds.), *Advances in animal welfare science 1985/86* (pp. 285-295). Washington, DC: The Humane Society of the United States.
- [69] Mori, M., Karl F. MacDorman, Norri Kageki. The Uncanny Valley. June 2012. *IEEE ROBOTICS & AUTOMATION MAGAZINE*.
- [70] Liu B., Hu M. and Cheng J. Opinion Observer: Analyzing and Comparing Opinions on the Web. Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.
- [71] Arim, Eva, Francisco Costa & Tiago Freitas. 2003. An empirical account of the relation between discourse structure and pauses in Portuguese. Paper at Prosodic Interfaces, Acoustique, Acquisition, Interpretation, Nantes.
- [72] Hall, Edward T. (1966). *The Hidden Dimension*. Garden City (NY). Anchor Books. ISBN 0-385-08476-5.
- [73] Mumm Jonathan, Bilge Mutlu (2011). Human-robot proxemics: physical and psychological distancing in human-robot interaction. University of Wisconsin-Madison, Madison, WI, USA. Proceedings of the 6th international conference on Human-robot interaction. Pages 331-338.
- [74] Brinkman, W.P., and ter Heijden, N.(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.
- [75] Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods*, 18(3), 385. doi:10.1037/a0032964
- [76] Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. Hillsdale, NJ: L Erlbaum; 1988.
- [77] Chiu, C., Y. Hong, et al. (1995). Gaze direction and fluency in conversational speech: unpublished manuscript.
- [78] Deese, James (1984). *Thought into speech: The psychology of a language*. Englewood Cliffs, NJ: Prentice-Hall.
- [79] Tauroza, Steve & Desmond Allison. 1990. Speech rates in British English. *Applied Linguistics* 11. 90-105.
- [80] Liu, P., and Wang, Z. Voice activity detection using visual information. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. Department of Electronics Engineering Tsinghua University.
- [81] T. Bickmore, J. Cassell. *Small Talk and Conversational Storytelling In Embodied Conversational Interface Agents*. (1999). Gesture and Narrative Language Group - MIT Media Laboratory.
- [82] Begoli, E. (2014). Procedural Reasoning System (PRS) architecture for agent-mediated behavioral interventions.
- [83] Bickmore, T., Schulman, D., & Shaw, G. (2009) DTask and litebody: Open source, standards-based tools for building web-deployed embodied conversational agents. In: Vol. 5773 LNAI. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 425-431).
- [84] Kavakli, M., Li, M., & Rudra, T. (2012). Towards the design of a virtual sociologist on aborigines substance abuse: A coping-theory perspective.

- [85] Lisetti, C., Amini, R., Yasavur, U., & Rische, N. (2013). I can help you change! An empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems*, 4(4). doi:10.1145/2544103
- [86] Ly, K. H., Ly, A. M., and Andersson, G. (2017). A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet Interventions*, 10, 39-46. doi:10.1016/j.invent.2017.10.002
- [87] Morbini, F., DeVault, D., Sagae, K., Gerten, J., Nazarian, A., & Traum, D. (2014). FLoReS: A Forward Looking, Reward Seeking, Dialogue Manager. In J. Mariani, S. Rosset, M. Garnier-Rizet, & L. Devillers (Eds.), *Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems into Practice* (pp. 313-325). New York, NY: Springer New York.
- [88] Oh, K. J., Lee, D., Ko, B., & Choi, H. J. (2017). A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation.
- [89] Rudra, T., Li, M., & Kavakli, M. (2012). ESCAP: Towards the design of an AI architecture for a virtual counselor to tackle students' exam stress.
- [90] Sebastian, J., & Richards, D. (2017). Changing stigmatizing attitudes to mental health via education and contact with embodied conversational agents. *Computers in Human Behavior*, 73, 479-488. doi:10.1016/j.chb.2017.03.071
- [91] Shamekhi, A., Bickmore, T., Lestoquoy, A., & Gardiner, P. (2017). Augmenting group medical visits with conversational agents for stress management behavior change (Conference Paper) (Publication no. 10.1007/978-3-319-55134).
- [92] Yasavur, U., Lisetti, C., & Rische, N. (2013) Modeling brief alcohol intervention dialogue with MDPs for delivery by ECAs. In: Vol. 8108 LNAI. *Lecture Notes in Computer Science* (including subseries *Lecture Notes in Artificial Intelligence* and *Lecture Notes in Bioinformatics*) (pp. 92-105).