Building Embodied Conversational Agents

Blomsma, Pieter A.

Publication date:
2023

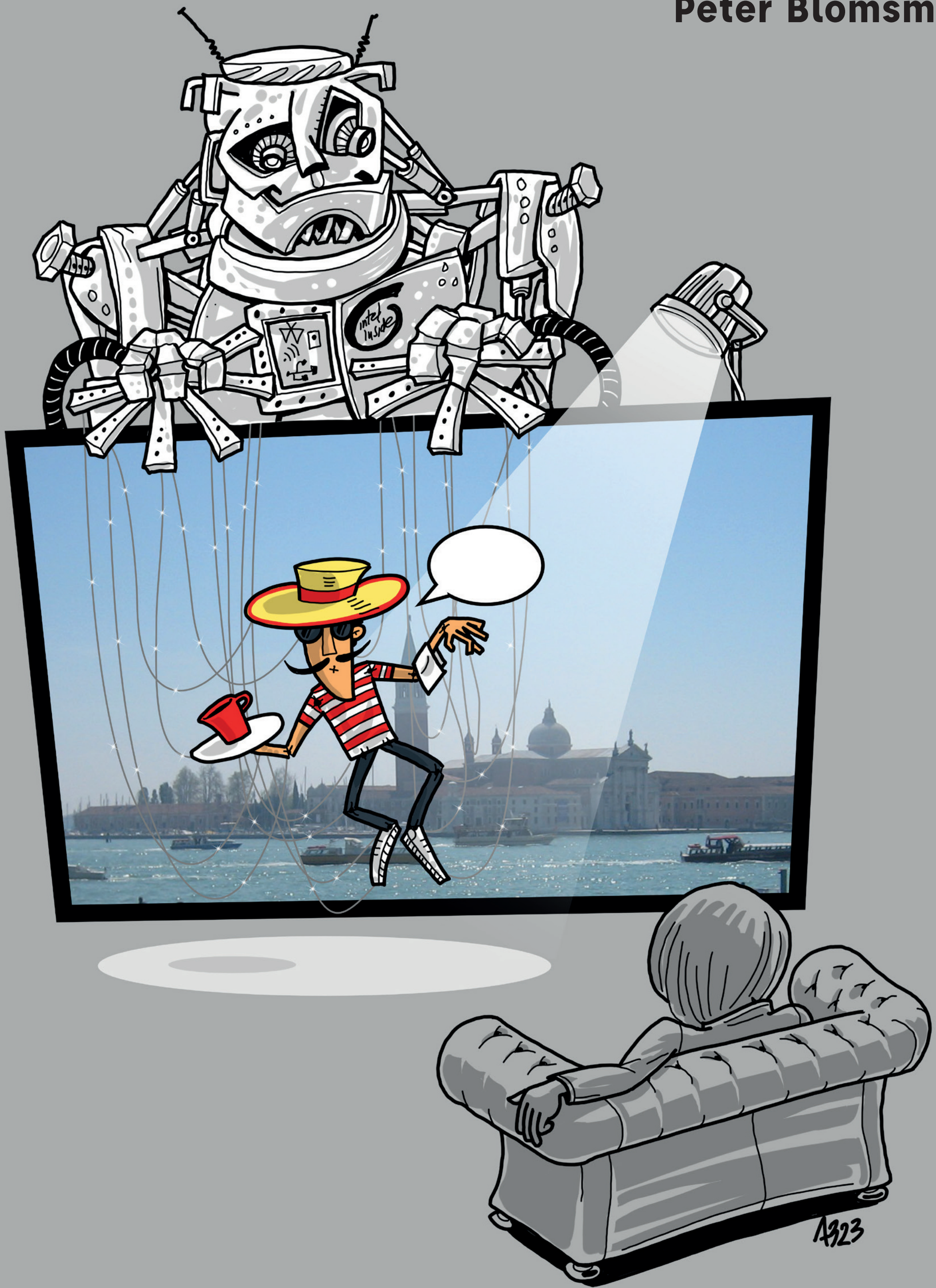
Document Version
Publisher's PDF, also known as Version of record

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Blomsma, P. A. (2023). *Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters*. [Doctoral Thesis, Tilburg University]. PrintPartners Ipskamp B.V. (SIKS Dissertation Series, 16).

Building Embodied Conversational Agents

Peter Blomsma



Building Embodied Conversational Agents

Observations on human nonverbal behaviour as a resource for the development of artificial characters

Peter Blomsma

Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters

Peter Blomsma
PhD Thesis
Tilburg University, 2023

TiCC PhD series No. 83

ISBN: 978-94-6469-343-0
Cover design: ABCartoons.nl

SIKS Dissertation Series No. 2023-16
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research school for Information and Knowledge Systems

Building Embodied Conversational Agents

Observations on human nonverbal behaviour as a resource for the development of artificial characters

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan Tilburg University
op gezag van de rector magnificus, prof. dr. W.B.H.J. van de Donk,
in het openbaar te verdedigen ten overstaan van
een door het college voor promoties aangewezen commissie in de
Aula van de Universiteit op dinsdag 20 juni 2023 om 10.00 uur

door

Pieter Alexander Blomsma geboren te Schoonrewoerd

Promotoren:
Prof. dr. M.G.J. Swerts, Tilburg University
Prof. dr. G. Skantze, KTH - Royal Institute of Technology
Prof. dr. J.H.M. Vroomen, Tilburg University

Leden promotiecommissie:
Prof. dr. E. André, Augsburg University
Prof. dr. C. Pelachaud, French National Centre for Scientific Research
Prof. dr. J. Beskow, KTH - Royal Institute of Technology
Prof. dr. E.J. Krahmer, Tilburg University

This thesis was (partly) accomplished with financial support from grants NSF-IIS-0416128, NSF-BCS-0826825, and PROJ-007246 from the European Union, OP Zuid, the Ministry of Economic affairs, the Province of Noord-Brabant, and the municipality of Tilburg awarded to Max M. Louwerse. Additional support from the Department of Cognition and Communication at the Tilburg School of Humanities and Digital Sciences was also greatly appreciated.

Contents

CHAPTER 1

1

Introduction

1.1 The evolution of user interfaces

"Wow this is so cool!" This is what I most probably yelled, back in the 90s, when my first computer program on our MSX computer turned out to do exactly what I wanted it to do. The program contained the following instruction:

```
COLOR 10            (1.1)
```

After hitting enter, it would change the screen color from light blue to dark yellow.

A few years after that experience, Microsoft Windows was introduced. Windows came with an intuitive graphical user interface that was designed to allow all people, so also those who would not consider themselves to be experienced computer addicts, to interact with the computer. This was a major step forward in human-computer interaction, as from that point forward no complex programming skills were required anymore to perform such actions as adapting the screen color. Changing the background was just a matter of pointing the mouse to the desired color on a color palette.

"Wow this is so cool!". This is what I shouted, again, 20 years later. This time my new smartphone successfully skipped to the next song on Spotify because I literally told my smartphone, with my voice, to do so. Being able to operate your smartphone with natural language through voice-control can be extremely handy, for instance when listening to music while showering. Again, the option to handle a computer with voice instructions turned out to be a significant optimization in human-computer interaction. From now on, computers could be instructed without the use of a screen, mouse or keyboard, and instead could operate successfully simply by telling the machine what to do.

In other words, I have personally witnessed how, within only a few decades, the way people interact with computers has changed drastically, starting as a rather technical and abstract enterprise to becoming something that was both natural and intuitive, and did not require any advanced computer background. Accordingly, while computers used to be machines that could only be operated by technically-oriented individuals, they had gradually changed into devices that are part of many people’s household, just as much as a television, a vacuum cleaner or a microwave oven.

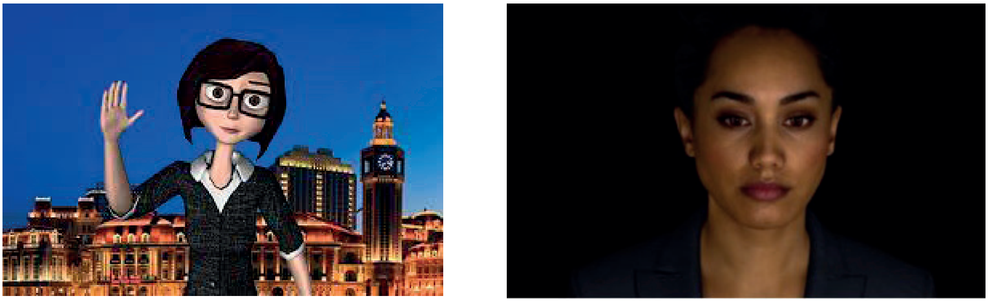
The introduction of voice control is a significant feature of the newer generation of interfaces in the sense that these have become more "antropomorphic" and try to mimic the way people interact in daily life, where indeed the voice is a universally used device that humans exploit in their exchanges with others. The question then arises whether it would be possible to go even one step further, where people, like in science-fiction movies, interact with avatars or humanoid robots, whereby users can have a proper conversation with a computer-simulated human that is indistinguishable from a real human. An interaction with a human-like representation of a computer that behaves, talks and reacts like a real person would imply that the computer is able to not only produce and understand messages transmitted auditorily through the voice, but also could rely on the perception and generation of different forms of body language, such as facial expressions, gestures or body posture. At the time of writing, developments of this next step in human-computer interaction are in full swing, but the type of such interactions is still rather constrained when compared to the way humans have their exchanges with other humans. It is interesting to reflect on how such future human-machine interactions may look like. When we consider other products that have been created in history, it sometimes is striking to see that some of these have been inspired by things that can be observed in our environment, yet at the same do not have to be exact copies of those phenomena. For instance, an airplane has wings just as birds, yet the wings of an airplane do not make those typical movements a bird would produce to fly. Moreover, an airplane has wheels, whereas a bird has legs. At the same time, an airplane has made it possible for a humans to cover long distances in a fast and smooth manner in a way that was unthinkable before it was invented. The example of the airplane shows how new technologies can have "unnatural" properties, but can nonetheless be very beneficial and impactful for human beings.

This dissertation centers on this practical question of how virtual humans can be programmed to act more human-like. The four studies presented in this dissertation all have the equivalent underlying question of how parts of human behavior can be captured, such that computers can use it to become more human-like. Each study differs in method, perspective and specific questions, but they are all aimed to gain insights and directions that would help further push the computer developments of human-like behavior and investigate (the simulation of) human conversational behavior. The rest of this introductory chapter gives a general overview of virtual humans (also known as embodied conversational agents), their potential uses and the engineering challenges, followed by an overview of the four studies.

1.2 Embodied Conversational Agents (ECAs)

1.2.1 Definition of ECAs

A virtual human that operates as an interface between a user and a machine is called an Embodied Conversational Agent ('ECA') [43]. ECAs are designed to understand and expressively react to verbal and non-verbal messages of their human interlocutors resulting, ideally, in meaningful conversations with the users [158]. Other terms that are used in the literature, and are closely related to ECAs, include Intelligent Virtual Agents ('IVA') and Socially Intelligent Agents ('SIA').



(a) SARA

(b) Digital Human from Soul Machines

Figure 1.1 Impression of a cartoonlike ECA on the left, and a photorealistic ECA on the right.

Definitions differ regarding their embodiment (IVAs are virtual, SIAs may be physical or robotic (see [133] for an elaborate overview regarding the terminology)). In this dissertation where the term ECA utilized, it refers to a realistic human-like character that is both visually and behaviorally indistinguishable from real humans.

Over the years, many ECAs have been developed. Cartoonlike characters include Greta, an agent that is used to research and develop models for multi-modal behavior and social competence [168], SARA, an agent that is used to research rapport-building capabilities between humans and ECAs [140], and REA [24], an ECA that represents a real-estate agent. Due to recent technological advances, more photorealistic ECAs are being developed. Companies like Soul Machines [1] and Uneeq [2] build ECAs to be 'employed' at companies as digital assistants or product specialists.

Although ECAs have greatly improved, there are no ECAs yet that are truly indistinguishable from real humans [203].

1.2.2 Features of ECAs

The appearance and performance of ECAs tend to be inspired by what can be observed in real human beings, especially in the way they communicate with others. During conversation, humans do not only rely on language to carry on a conversation, but also use different parts of the body to support or steer the interaction, such as through facial expressions, gestures, eye gaze, head nods and other non-verbal gestures. We raise our eyebrows at the end of a sentence to convey that we asked a question. We point to objects in our environment to indicate what we are talking about. We modulate the speech melody or intensity pattern of our voice to emphasize certain words.

Accordingly, we exploit multiple characteristics of our body to convey meaning and we communicate in a multi-modal way. This indeed could be considered the most natural form of communication, given that speakers and addressees tend to hear and see each other during most of their interactions, so that it is only logical that they take into each other's auditory and visual signals.

The language of computers on the other hand is composed of zeros and ones, in a way that is too complex and therefore impractical for (most) people to handle directly. Consequently, in order to make the interaction with computers feasible, users need an interface that translates information from the computer into information that a human can interpret, and vice versa. Since the inception of the computer, there has therefore been a lot of effort into optimizing the ways in which people can operate with computers, leading to the invention of mouse, keyboard, joystick, etc.. These inventions have been designed to make human-computer interaction more natural and efficient. These mechanical devices, however, can still be viewed and experienced as parts of a machine, where the computer is just a tool to be used for a specific purpose. We are now entering the next level of human-computer interaction with the attempt to create computers that act and communicate as a human; if successful, human-computer interaction will then shift from "using a tool" to "interacting with a living creature" or even "interacting with a friend". Studies have already shown that e.g. children build social bonds and feel relatedness with ECAs [35] and social robots [125] with whom they can learn a new language together [87] or share secrets with [22]. In establishing trust relationships with a robot, similar psychological mechanisms to those used in establishing a trust relationship with a human being come into play [58].

Clearer communication Multi-modal communication offers the possibility to communicate the same message via multiple communication channels or even add extra meaning or disambiguate the main message using multiple channels. For example, we can use gestures to clarify ambiguous words, or point to the type of ice-cream we would like to order. Facial expressions can reveal the extra connotation that comes with a certain message, e.g. whether it is a neutral, sad, happy or angry statement. Gestures can also convey the exact same message as our speech, e.g. when we put our thumbs up and say "ok" [95, 18]. Thus, adding multi-modality to human-computer interaction will add an extra layer of information to the interaction and can qualify the messages between speaker and addressee.

Minimal mental overhead Interacting with a computer requires a certain amount of mental effort. Interacting with a computer puts a certain cognitive load on the working memory of the user, as the user must translate his/her intentions into actions that fit within the computer paradigm. The amount of cognitive load that a user can handle is limited, as the working memory is limited. Thus, the less cognitive load is required for the computer interaction, the more cognitive resources remain available for the actual task[163]. Two known factors that minimize the mental overhead of computer interfaces are (1) if such interfaces are built upon familiar experiences and habits [154, 8] and (2) if users can interact in a multimodal fashion with an interface [221, 164]. Interestingly, ECAs tick both boxes. First, ECAs build upon a deeply rooted, familiar experience, namely human interaction. And secondly, ECAs take advantage of multimodal interactions. And indeed, researchs showed that ECA-based systems result in lower frustration levels than traditional text-based systems [97, 8]. Furthermore, [162] demonstrated that people are more efficient when using multimodal interfaces, when executing a spatial task participants showed less disfluencies and errors with a multimodal interface, compared to a unimodal interface. In the same line, [221] established that people playing a multimodal variant of game make less errors (and experience a lower cognitive load) compared to those who played the unimodal variant of the game. Thus, operating a computer via an ECA may facilitate an optimal interaction where a user could focus maximally on the task and spend the least cognitive resources possible on the interaction itself.

Rapport When two people (i) have a positive feeling towards each other, (ii) feel connected and synchronized, and (iii) have the feeling that they understand each other's ideas and communicate well, they have something that is called "rapport". Rapport feels good and seems to play a central role in successful relationships [200]. Although the rapport construct is historically defined in terms of human-human interactions, people can also experience something similar to rapport when interacting with an ECA [82, 46]. Like rapport in human-human conversations can have a positive influence in negotiation [33, 61] education [20] and healthcare [65], computer-systems able to establish rapport can have a positive influence on their users as well. One of the first examples was Rea, a real-estate ECA, that used rapport-seeking strategies to build rapport with extroverted people [24]. Another experiment with an ECA that gave educational instructions to math learners, [116] showed that if the ECA was actively building rapport, learners had better performance measures compared to an ECA that was not building rapport. In the healthcare context a social robot was to help children to teach self-management skills needed to cope with diabetes type 1. The experiment was executed in two different sets of conditions. The first condition was a neutral robot, the other robot was personalized and included actively seeking rapport. Children that interacted with the personalized version reported a more pleasurable interaction and had a higher learning outcome, compared to children that interacted with the neutral robot [91].

Would the increase of comprehensibility, the minimization of mental overhead and the ability to build social relationships suggest that all human-computer interactions would benefit from a computer that represents itself as human? Probably not. If a user knows how to interact effectively with a computer via shortcuts, a terminal or a text processor, some input tasks such as programming or writing a book may be accomplished faster via such terminal or text processor than via an ECA. Some researchers hypothesize that when a computer substitutes human tasks it could benefit from a human-like representation, while when the computer is used for computational tasks, it benefits from a 'machine'-like representation. The main reason is that humans seem to trust machines more than humans when it comes to rationality and efficiency [185]. Finally, using an ECA also comes with responsibility. Humans that interact with ECAs also (unconsciously) expect the ECA to understand them [172], and a human-like ECA may yield too high expectations in a user about the range of functionalities of the machine.

1.3 Relevance of studying ECAs

1.3.1 Scientific relevance

Human behavior is studied within a wide array of scientific disciplines, which results in a variety of explanations and models, relating to cognitive, social, linguistic, psychological, and philosophical aspects of human behavior. Building a human-like ECA could be the glue that ties multiple theories together. Building a human-like ECA implies that our knowledge of human behavior (e.g. how non-verbal communication functions) needs to be sufficiently explicit so that it can be implemented in an artificial character. Using this analysis-by-synthesis procedure, ECAs can serve as the ultimate test case to check to what extent our models of human behavior are realistic and applicable. Richard Feynman's quote, "If you cannot build it, you do not understand it", was aimed at synthetic biology, but also applies in this context as building an artificial human may lead to a deeper understanding of human behavior. Another scientifically relevant aspect of ECAs is that they can serve as useful tools to test specific hypotheses about human behavior, as they give the possibility to manipulate specific features while controlling others in an orthogonal design, so that the effect of specific variables can be accurately tested. When human actors (or confederates) are replaced by an ECA in a specific experiment, scientists get the opportunity to experiment with variables which were difficult or impossible to control by human participants [166]. For example, [77] conducted an experiment where human participants had to interact with an ECA. The experiment consisted of two versions, whereby the only difference between the versions was the pitch and speed rate of the interlocutor (which was an ECA in this case). Such an experiment would be less obvious with a human interlocutor. An actor who does everything the same way every time except the use of his/her pitch and speed rate is virtually impossible. [96] has experimented with ECAs that showed either short or long eye blinks while listening to the participant. Again, such an experiment would have been impossible with a human confederate, as a human confederate does not have this minute control over his/her eye blinks. Although actors naturally try to behave as similarly as possible in any experiment, there will always be uncontrollable factors. When using ECAs as confederates in an experiment, there will be more control over the experiment while maintaining the same ecological validity. Nonetheless, there is room to improve. [166] mentions e.g. that one of the current limitations of using ECAs in human experiments is the lack of spontaneous bi-directional interactions between human and ECA. Thus, by aiming at the creation of conversational artificial humans, we may not only test our knowledge of aspects of human behavior, but we can also utilize ECAs for maximizing control in an ecological valid setting to further deepening our knowledge.

1.3.2 Societal relevance

The combination of both computer-like and human-like properties give ECAs a unique position with specific properties that can be utilized in different societal applications. Compared to traditional computer-systems, ECAs are able to establish social bonds with their users [81]. Compared to their human counterparts, ECAs have a perfect memory and are always available [195]. In this section we discuss a few specific societal opportunities for ECAs.

Confidential interactions ECAs seem to be effective in getting information from people in contexts where people are commonly hesitant to share information [130, 82]. For example, [131] showed that participants (ex-soldiers) disclose more PTSD symptoms when interviewed by an ECA, compared to when they are interviewed by a real person or when they must fill out an anonymous questionnaire. Due to cultural stigma's, ex-soldiers do often not disclose all their PTSD symptoms. The anonymity of an ECA compared to a real person gives a participant freedom to share any symptoms. Compared to the anonymous questionnaire, an ECA has the rapport building skills needed to give participants confidence to open-up and speak their minds.

Pedagogical interactions One way to effectively learn new things is with a teacher or learning companion. An ECA could be a good substitute or supplement to the learning process. Compared to a human tutor, an ECA is always available, treats learners unbiased, is patient and will not be annoyed [94]. In general, the more a student is engaged in learning, the more a student learns. Students who learn with an ECA [11] or social robot[53] are generally highly engaged, which in turn could lead to positive learning outcomes. ECAs and social robots are able to memorize all historical learning performances of a learner and can in turn provide a learner with personalized activities and adaptive tutoring that lead to effective learning and positive learning outcomes

[57, 28]. Also, ECAs and social robots are able to motivate the student and stimulate essential learning behaviors [107, 86]. ECAs can be employed as teachers to for example learn social skills [196], mathematical skills [116] or to learn a (second) language [211]. More specifically, ECAs could especially be useful as interactive tools for people who sometimes find it hard to communicate with other human beings [151] or to learn skills necessary to cope with certain illnesses [91].

Multilingual interactions Another advantage of an ECA is that it could potentially speak all languages, such that each person interacting with the ECA could use his or her native language. This functionality could be especially useful for ECAs that act (i) as tourist guides [72], (ii) in multilingual healthcare related settings [157], (iii) in migration and integration contexts [208], and (iv) in language learning [138].

1.4 The challenges

Although current ECAs are getting better in terms of appearance and behavior, all ECAs are still distinguishable from real people [203]. This is largely because of two main challenges that make it difficult to create a human-like ECA: (1) humans are difficult to trick because of their sensitive perception and (2) believable human behavior is hard to generate within a timely manner due to its complexity.

1.4.1 Uncanny Valley

Human perception is sensitive for little mistakes in the interaction with an ECA. If an ECA looks sufficiently human-like, small errors in the behavior of that ECA can produce creepy, uncanny feelings in the observer. Those uncomfortable feelings are a sign of the so-called uncanny valley, the metaphorical lowland between the mountain of cartoonish looking avatars on the left side and the mountain of real humans on the right side [150]. The uncanny valley is a term coined by Masahori Mori back in 1970. In an essay, he predicted that if robots would appear and behave human-like, that those robots would elicit negative responses from people. This negative response would be caused by the robot that failed in fully replicating human behavior. Errors in the replication would invoke feelings of fright. Uncannyness is related to the horror genre. The German psychologist Ernst Jentsch described the uncanny feeling as a state of mind in which you cannot discern between what is alive and what is dead, between what is real and what is unreal. According to Jentsch we can experience the uncanny when we observe people that move abnormally, like a person having a seizure or a person that lost control over their bodily functions due to madness or a mental illness [105]. Masahori Mori stated that such experience can be expected with a human-like robot. The valley aspect of the uncanny valley originates from the graph that Mori supplied in his essay.

The graph shows that the higher the human-likeness of something is, the higher the perceived affinity. However, this otherwise positive related relationship contains a dip just before full human-likeness. This dip is the uncanny valley [150].

And indeed, people often get an uneasy feeling, or even a feeling of disgust, at the mere sight of a human-looking doll, computer-game character, or robot [216]. This poses a problem for the creation of ECAs that are indistinguishable from real humans. People seem to possess a doorkeeper that needs to be fooled or convinced that the human-like avatar in front of them is a real human. Why do we have this uncanny reaction to human-like creatures? Different theories exist. One theory hypothesizes that the uncanny feeling arises from our so called behavioral immune system [182], a system that detects dangerous peoples, like psychopaths [201] and virus-infected people [93], and once detected produces the uncanny feeling to nudge us to avoid those dangerous people. Another explanation is that the feeling sprouts from cognitive dissonance [216], e.g. we do not know if we must categorize this object as human or as non-human which causes the uncanny feeling.

As for the present, the uncanny valley has not yet been passed by current ECAs. The sensitive human perception that is able to detect small errors, and the uncanny feelings that result from such errors, is one of the major obstacles to actually create human-like ECA that are indistinguishable from real people.

1.4.2 Complexity

Human conversational behavior is complex. Multiple modalities (and channels) are involved in sending and receiving messages between interlocutors. When it comes to multi-modal behavior generation for ECAs there are a few concerns. For example: how to generate natural looking behavior? Human bodies are capable of all kinds of movements, but not all movements will be perceived as natural. And how to generate communicative behavior within the correct timeframe? Generation of behavior is time-sensitive that requires high computing power and speedy algorithms.

Natural behavior: Degrees of freedom The human body consists of about 600 muscles which can all be moved semi-independently. In addition, the human voice can utter a myriad of sounds on a spectrum of frequencies. Although some combinations of sounds with body movements will result in natural behavior, some behaviors will be perceived as unnatural [181, 39]. The human body allows for many degrees of freedom, which makes behavior generation complex. As a thought experiment, to grasp the intricacies of the complexity problem, one could imagine being a puppeteer of a human-like marionette with about 600 wires, one for each muscle. This marionette is placed into a conversation with a person that thinks that the marionette is a real person. When should one move which muscle (and with what speed and contraction) in order to come off as acting natural? This, together with operating the voice of the marionette, is essentially the challenge that a computer faces when simulating a human being.

Time sensitivity An ECA must not only generate meaningful behavior, it must also understand the (sometimes unpredictable) behavior from the other person to whom it has to react. This complex interplay between perception and generation must take place in a timely fashion. Time sensitivity takes place on multiple levels. On the generation side there is time alignment between the multiple modalities. For example, gestures should be shown within the same time of the utterance, there seems to be some wiggle room, but a gesture cannot be shown more then 160ms before the linguistic component is uttered or after the utterance (the timewindow of 160ms may be a bit longer depending on the concreteness of the gesture) [88]. On the perceptual side, some signals shown by the user are volatile, e.g. a short eye-blink [96] is about 200ms long. The epitome of rapid reactions in human conversations is the average silence between turns (i.e. when one persons stops talking, and the other one takes over), which is on average just 200ms [190].

In sum, behavior generation is a complex puzzle because behavior manifests itself across multiple channels, where a communicative element on each channel can keep the illusion of naturalness alive or take it away. A conversation is a continuous game and behavior must be generated within short time intervals, where every minuscule deviation from the expected sets off internal alarm bells for the user in the form of uncanny feelings.

1.5 The current dissertation

The four studies presented in this dissertation are all intended to reduce the aforementioned challenges and to pave the way towards the creation of an ECA that approaches the behavior of a real human.

Chapter 2 tries to mitigate the complexity of human behavior generation. The underpinning idea sprouted from the Pareto principle which states that 80% of the outcomes are produced by 20% of the causes. Which led to the conjecture that in order to curb the seemingly endless degrees of freedom of human behavior and to keep behavior generation within reasonable time-bounds, we perhaps should focus on the successful generation of the most frequent behaviors (the proverbial 20%). According to FACS, our face has 46 muscles. Each muscle can be contracted to 6 levels (0 is neutral, 5 is maximum contraction). Thus, the set of possible facial configuration according to

FACS is 6^{46}. However, is this large number of possible facial configurations actually used by humans, and should an ECA be able to move its face into 6^{46} different configurations? Of course, implementing 6^{46} is a time-consuming task for both the 3d artist and the testers. We showed that the largest portion of facial expressions was neutral (meaning, no contraction for any facial muscle). In cases that facial muscles were activated, most often those were only activated slightly. The results of this research show that indeed not all possible facial configurations are commonly used, or are used at all.

Chapter 3 describes another approach to curb the degrees of freedom in behavior generation. This time, the inspiration came from the concept of hidden attractors, an idea from dynamical systems theory. Simply put, a hidden attractor is an underlying pattern that exists in an ostensibly chaotic system. Multi-modal human behavior could be seen as a chaotic system that produces output over different semi-independent communication channels, such as facial expressions, gestures, and linguistic content that are all involved in behavior generation. In chapter 3 we show with cross recurrence analysis that there seem to be dependencies between the linguistic channel and the gesture channel. We show that some gestures often go together with certain linguistic parts (dialog acts), while there also seem to be 'mutually exclusive' relations, some linguistic parts never go together with certain gestures. This information can be used as an initial guidebook for an ECA developer to create a multi-modal behavior generation system that takes into account the dependencies between communication channels. Whereas Chapters 2 and 3 attempt to circumscribe the pool of possible human behaviors, Chapters 4 and 5 are about the variations of behaviors found in that pool.

Chapter 4 reports on the variability found in listening behavior. Listening behavior, more specifically backchanneling behavior, includes nodding, vocalizations, and facial expressions that listeners and speakers use to coordinate their interaction. We analyzed the behavior of 14 different participants that all interacted with the same (pre-recorded) speaker-stimulus. Participants were made to believe that they had a live skype-connection with the speaker. The speaker and the participant played a Tangram game. A jury consisting of 10 people identified the so-called backchannel opportunity points; moments in the speaker-stimulus that would allow for feedback from the listener. We analyzed the differences between the listening behavior of different participants, but also between different BOPs. We found out that variation between and within listeners exists with some listeners being more expressive than others, whereas some backchannel opportunity points trigger more responses than other.

Chapter 5 contains the results of an experiment based on the findings for chapter 4. In this follow-up research we investigated whether the variations in listening behavior correlate with personality perceptions, and if those variations can be used in ECAs.

Personality seems to be an important factor in conversational systems and specifically ECAs to create smooth and natural interactions with its user. We conducted two rating experiments in which participants judged the personalities (i) of human beings and (ii) of embodied conversational agents. The results show that personality perceptions of both humans and artificial communication partners are indeed influenced by the type of feedback behavior used. This knowledge could help developers on how to also include personality in the listening behavior of their ECAs, which in turn could generate a stronger sense of presence for the human interlocutor when interacting with the ECA.

All four studies presented in this dissertation are based on a few assumptions. All presented research assumes that the final goal of the work is to create perfect, human-like ECAs which closely resemble real humans and are modeled based on human behavior. Therefore, the chapters present analyses on human behavior or the synthesis of human behavior into an ECA. Other lines of research exist that try to exploit non-human behaviors in ECAs to optimize human-computer interaction, e.g. the possibility of co-embodiment and re-embodiment, which reflects the possibility that an ECA contains multiple identities or that multiple ECAs contain the same identities respectively [135]. However, in this work we strictly focus on the analysis of human behaviors. Each chapter in this dissertation is a self-contained study which is either already published or submitted to a peer-reviewed scientific journal (chapter 3 and 4) or a conference (chapter 1 and 2). Therefore, each chapter contains its own abstract, introduction, discussion and reference list. As a consequence, there is a minor overlap between the different chapters and there may be slight variations in jargon and stylistic elements, due to the different outlets.

CHAPTER 2

2

Spontaneous Facial Behavior Revolves Around Neutral Facial Display

Abstract

With forty-six Action Units (AUs) forming the building blocks in the Facial Action Coding System (FACS), millions of facial configurations can be formed. Most research has focused on a subset of combinations to determine the link between facial configurations and emotions. Despite the value of this research for psychological and computational reasons, it is not clear what the most common combinations of AUs are to form the most commonly expressed facial configurations. We used three diverse corpora with human coded facial action units for a computational analysis. The analysis demonstrated that the largest portion of facial behavior consists of the absence of AU activations, yielding only one specific facial configuration, that of the neutral face. These results are important for cognitive scientists, computer graphics designers and virtual human developers alike. They suggest that only a relatively small number of AU combinations are initially needed for the creation of natural facial behavior in Embodied Conversational Agents (ECAs). [1]

[1] This chapter is based on: Blomsma, P. A., Vaitonyte, J., Alimardani, M., & Louwerse, M. M. (2020). Spontaneous Facial Behavior Revolves Around Neutral Facial Displays. In Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk. ACM, New York, NY, USA, 8 pages .https://doi.org/10.1145/3383652.3423893

2.1 Introduction

The most natural form of communication is multimodal communication, in which verbal and non-verbal channels participate in the joint action of the dialog partners [50]. Among the non-verbal channels, the human face is widely investigated when it comes to social cues [206]. Facial displays reveal much about the people we talk to. The facial display could be seen as a dynamic information space that can adopt a myriad of different facial configurations, with each configuration potentially sending out a specific social signal [102]. A mere glance at a person's face can reveal information about their identity and gender [71], age [71], physical health [183], intention [3], emotion [4], eye gaze direction [76] and social traits [160]. Moreover, the face can encode linguistically relevant cues such as lip movements that can enhance language comprehension, particularly beneficial under difficult listening conditions [192]. Ideally, Embodied Conversational Agents (ECA) would allow humans to have face-to-face conversations with computers such that we can use our most natural form of communication [43]. As computer graphics techniques become more advanced and the realism of ECA faces increases, this reality is coming closer and questions regarding the development of ECA facial behavior become increasingly important. However, the generation of natural facial behavior for ECAs is still a major challenge.

In addition to the difficulty of all social cues that have to be taken into account in generating facial behavior, small inaccuracies in ECA behavior may result in feelings of uncanniness and frustration by the perceiver [181]. Herein lies a problem for the creation of ECAs; due to the precision needed to accurately mold a digital face to replicate each detail of its human counterpart [176], designing a natural facial configuration for an ECA is a very time-consuming task. Furthermore, the face allows for a nearly infinite number of different facial configurations. The combination of the above factors ensures that designing all possible facial configurations for an ECA, such that it can generate all desired facial behavior would take enormous amounts of time. As a compromise, to reduce those designing efforts, many ECA developers only implement a subset of facial configurations. Although this trade-off is a promising approach, it also begs the question which subset of facial configurations is the most optimal to implement. In other words, what are the most common facial configurations that humans employ? The current research aims to find answers to this question.

2.1.1 Uncanny Valley

The first factor that prevents developers from creating ECAs that are indistinguishable from real humans is the uncanny valley effect. Recent advances in animation and rendering technology (e.g. the AutoDesk Maya application, as well as Unity3D and Unreal engines) allow developers to design increasingly realistic, human-like avatars, both in academic and industrial virtual human projects [186, 198]. However, it turned out that tiny mismatches between real humans and the ECA's appearance or behavior can be perceived as frightening and evoke feelings of discomfort [150]. This so-called 'uncanny valley effect' might be explained biologically. Humans might have an internal behavioral immune system [182] that produces strong negative feelings once it detects dangerous people like psychopaths [201] or disease carriers [93]. Others have explained this uncanny valley effect as a consequence of cognitive dissonance [216]. Irrespective of the reason for the uncanny valley effect, correcting all the tiny mismatches that evoke the effect, is a laborious and meticulous process which currently prevents ECAs to be perceived as real humans [176].

2.1.2 Facial Action Coding System

The second factor that makes natural facial behavior generation for ECAs difficult is the nearly infinite number of possible facial configurations. The most commonly used reference framework to create facial configurations is the Facial Action Coding System (FACS) [204]. FACS is an instrument to objectively characterize facial movements in an anatomically-based manner [64], measuring the level of contraction of 46 facial muscle-endings, called Action Units (AUs). Encoding a facial configuration with FACS involves enumerating all AUs that are active in the face along with their corresponding intensity on a scale from 0 (i.e., no activation) to 5 (i.e., maximum activation). Consequently, FACS is able to describe 6^{46} possible combinations of AUs (albeit the case that in practice some combinations are very unlikely or even impossible). This number, equaling to 6.2e+35, is astronomically large, so large that it exceeds the average number of cells in a human body (i.e., 10e+16).

Given (i) the enormous number of facial configurations enabled by FACS, (ii) the photometric, kinematic and geometric complexities of the human face [219] and (iii) our sensitivity to inaccuracies in digital humans, one could come to the conclusion that digitalization of natural human-like faces is not feasible. A possible solution lies in reducing the number of facial configurations in the agent. There may be at least two solutions for such reduction. The first solution has been used in the psychological and computational literature so far, and involves using only prototypical facial configurations that go with extreme emotions. The other solution, which is the main subject of this paper, may lie in identifying which facial configurations are most commonly used in non-verbal communication.

2.1.3 Prototypical facial configurations

Researchers in the field of facial expressions of emotions are often interested in which specific facial expressions are related to specific emotions or social contexts [52]. Moreover, most ECA developers have focused on the implementation of prototypical emotions and corresponding facial behavior into an ECA [126, 169] and some have even focused on the implementation of prototypical expressions blended into more complex expressions [167]. Such implementations can be effective if the aim is to generate easily recognizable facial expressions of emotion for the perceiver, however ECAs that execute them do not necessarily replicate human-like facial behavior. According to literature, in natural situations, humans are not exposed equally to the above-mentioned prototypical emotions. We are exposed most frequently to happy expressions (31.0% of the time), in contrast with fearful expressions (3.4% of the time) [37]. If we apply Occam's Razor, it thus seems best to go for a happy rather than a fearful configuration. Granted that frequency may not be the only important factor in the choice of configurations (others may include deviation from the norm), an ECA that implements only the prototypical facial configurations should also take into account those ratios while generating facial behavior. However, taking into account those ratios still assumes that all facial behavior takes place within the confinements of prototypical facial configurations. In other words, we do not know what the distribution of facial behavior looks like when the option space outside these prototypical configurations is included.

The option space outside prototypical facial configurations may in fact be quite important. During a conversation, humans convey emotional content via the face not more than 18% of the time [69]. This is not surprising, as facial displays may serve a variety of functions, such as emphasizing words during a conversation or indicating that an utterance serves as a question. The point here is that the unbalanced distribution of prototypical facial configurations and the relatively limited emotional content that is conveyed in conversation may be indicative of the fact that we do not know much about the distribution of facial configurations in human-human conversation, let alone generating a representation of that distribution in human-ECA conversations.

2.1.4 Current Work

In this paper, we addressed the above-mentioned gap and investigated the distribution of human facial configurations in a natural setup. We selected three large non-posed human-coded facial behavior datasets that differed with regard to the type of spontaneous emotion elicitation used. Although the accuracy of automated FACS encoding software (.e.g. [127]) is impressive and is of added value for various applications, the precision of the produced encodings is lower than the encodings produced by certified FACS encoders, especially if the certified FACS encoders have a high interrater reliability. Many applications of automated encoding analyze faces to detect combinations of activated AUs to identify prototypical expressions of emotions. Because combinations of activated AUs are the main concern, the exact activation of each specific AU (and the associated encoding accuracy) are of less importance. However, the current research does analyse both combinations and single AU activations, therefore it is essential that each AU is accurately encoded. For that reason, in the current work we only selected manually encoded datasets with high interrater reliability scores (in contrast to other approaches like [191]). The number of available human FACS-encoded datasets containing non-posed facial configurations is rather small [119]. This is partly due to the fact that FACS encoding is a costly, labor-intensive process. It takes at least 100 hours of extensive training to become a certified FACS encoder and it takes two hours for such encoder to FACS encode one minute of video (depending on number of AUs and level of detail) [13]. Study 1 used a dataset with facial configurations from participants in a task-based multimodal communicative setting. Study 2 used a dataset with facial configurations from participants watching an emotion-eliciting video clip. Study 3 used a very specific dataset with pain-related facial behavior.

2.2 Study 1

Study 1 used a multimodal communication corpus of about 25 hours of dialog, with 48 students (30 female, 18 male; one Asian, 19 African-American, 28 Caucasian). Verbal and non-verbal channels of dyads were recorded while they participated in a Map Task scenario [6, 129]. The Map Task involves one person (the instruction giver), who has a map with a route and another person (the instruction follower) who has a similar map without a route. The goal of the task is to reproduce the route of the instruction giver's map onto the instruction follower's map. The interlocutors have full freedom of speech, however they cannot see each other's map (see [129] for details). Each facial movie was FACS-encoded at 250ms intervals by certified FACS coders for 11 specific AUs, resulting in 731,824 encoded frames (see Table 2.1 for an overview). AUs were encoded as either 0 (i.e., no activation) or 1 (i.e., activated). The coders had a high inter-observer agreement score, quantified by Cohen's κ (.78). The dataset is also used in [129, 29].

2.2.1 Method

The FACS encodings of the videos were imported as matrices into R. Each row corresponded with one encoded frame, thus the matrix contained 731,824 rows. Each column referred to one of the AUs and expressed the level of contraction for that specific AU, thus the matrix contained 11 columns. In order to deduce the number of distinct facial configurations present in the matrix, the unique rows in the matrix were identified by stripping out all duplicate rows in a copy of the matrix. Subsequently the number of occurrences of each unique row was counted in the original matrix[2].

[2] Code publicly available via: https://github.com/pblomsma/facial-behavior

2.2.2 Results and Discussion

As the dataset was encoded for 11 different AUs on a binary level, a total of 2^{11}(= 2,048) different facial configurations could have been expressed during the dialogs. Yet, only 67 different facial configurations were found in the 731,824 frames. From these 67 facial configurations, 15 were not shared among participants and were only expressed by one person.

By far the most common facial configuration was neutral (i.e. AUs having an activation level of zero), with 88.32% (SD=8.09, see Figure 2.1) of the encoded frames described by this facial configuration (646,347 frames). The 20 most occurring facial configurations are listed in Table 2.2.

Table 2.1 Zero-contraction frequency per AU per dataset. A dash indicates that that AU was not encoded for that dataset.

AU	Description	Map Task (Study 1)	DISFA (Study 2)	PAIN (Study 3)
1	Inner brow raiser	-	93.29%	-
2	Outer brow raiser	95.27%	94.37%	-
4	Brow lowerer	98.23%	81.20%	97.78%
5	Upper lid raiser	99.67%	97.91%	-
6	Cheek raiser	-	85.11%	88.52%
7	Lid tightener	-	-	93.05%
9	Nose wrinkler	-	94.55%	99.13%
10	Upper lip raiser	-	-	98.92%
12	Lip corner puller	97.86%	76.46%	85.77%
15	Lip corner depressor	-	93.99%	99.99%
17	Chin raiser	99.93%	90.12%	-
18	Lip puckerer	99.77%	-	-
20	Lip stretcher	-	96.54%	98.54%
23	Lip tightener	99.45%	-	-
25	Lip part	99.38%	64.80%	95.03%
26	Jaw drop	-	80.91%	95.68%
27	Mouth stretch	99.98%	-	99.96%
43	Eyes closed	-	-	98.00%
44	Squint	98.71%	-	-
45	Blink	99.23%	-	-
50	Speech	-	-	100.00%

Study 1 shows that the most-occurring facial configuration is neutral. Moreover, that for those cases where activation of an AUs was observed, the seven most-occurring facial configurations only had one of the eleven AUs activated. The neutral facial configuration combined with the seven most occurring configurations described 99.34% of the data. This indicates that only a small set of facial configurations with single AU activations is needed to describe the facial behavior in this dataset.

Table 2.2 Most frequent facial configurations in Map Task dataset.

%	Action Unit										
	2. Outer Brow Raiser	4. Brow Lowerer	5. Upper Lid Raiser	12. Lip Corner Puller	17. Chin Raiser	18. Lip Puckerer	23. Lip Tightener	25. Lip Part	27. Mouth Stretch	44. Squint	45. Blink
88.32%	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
4.34%	1	0	0	0	0	0	0	0	0	0	0
1.90%	0	0	0	1	0	0	0	0	0	0	0
1.45%	0	1	0	0	0	0	0	0	0	0	0
0.91%	0	0	0	0	0	0	0	0	0	1	0
0.70%	0	0	0	0	0	0	0	0	0	0	1
0.59%	0	0	0	0	0	0	0	1	0	0	0
0.49%	0	0	0	0	0	0	1	0	0	0	0
0.22%	0	1	0	0	0	0	0	0	0	1	0
0.22%	0	0	1	0	0	0	0	0	0	0	0
0.20%	0	0	0	0	0	1	0	0	0	0	0
0.14%	1	0	0	1	0	0	0	0	0	0	0
0.08%	1	0	0	0	0	0	0	0	0	1	0
0.08%	1	0	1	0	0	0	0	0	0	0	0
0.05%	0	0	0	0	1	0	0	0	0	0	0
0.04%	0	0	0	1	0	0	0	0	0	1	0
0.03%	1	0	0	0	0	0	0	0	0	0	1
0.02%	0	0	0	1	0	0	0	0	0	0	1
0.02%	0	1	0	0	0	0	1	0	0	0	0
0.02%	0	0	0	0	0	0	0	0	1	0	0

There are, however, at least three potential caveats in this study that need to be accounted for. The first is that we would like to assume that the Map Task corpus concerns natural conversation, but the question can be raised to what extent instructions on route navigations on a map elicit natural dialog and natural emotions. For instance, it is unlikely that sadness or disgust would be expressed in such a dialog setting. Secondly, in this corpus, 11 AUs were encoded on a binary level (activated or not). It may be the case that coders – despite their acceptable inter-rater reliability – marked minor activations of AUs as 0 and only major activations as 1. Therefore another dataset to determine whether the findings from Study 1 can be extended to a different setting is desirable. Finally, as the dataset is only coded in binary form, the activations cannot inform on the exact intensity of found facial configurations. For ECA developers it would be important to know the precise activation levels in order to exactly replicate the facial configurations into an ECA. Therefore, the next Study used a dataset that is encoded on the full FACS spectrum (i.e. 0-5).

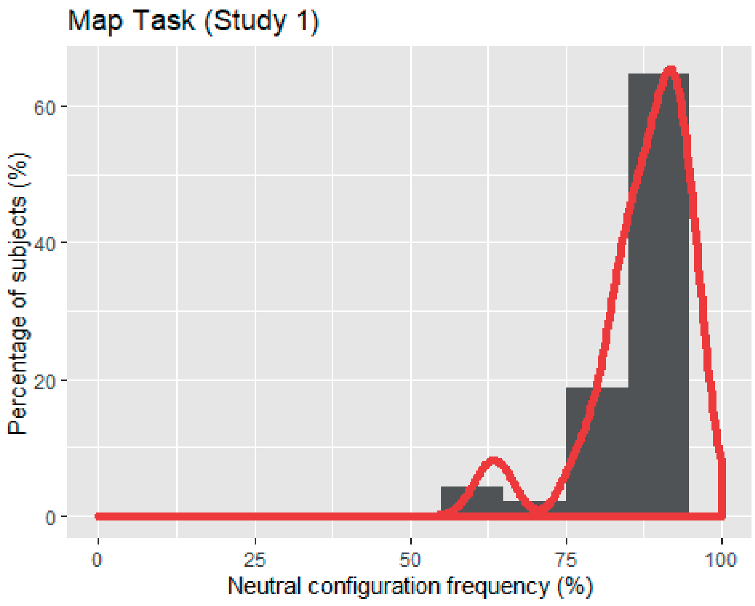


Figure 2.1 Distribution of neutral facial configurations (i.e., all AUs having zero contraction) among participants whose facial configurations were coded in Study 1.

2.3 Study 2

Study 2 used the Denver Intensity of Spontaneous Facial Actions (DISFA) dataset [141] consisting of facial movies of 27 subjects (15 male, 12 female; three Asian, two Hispanic, one African-American, 21 Caucasian) in the age range of 18-50 years. The dataset was created with the intention to examine AU activations in naturally occurring facial configurations. Each subject was filmed while they watched a 4-minute video that contained a collection of emotion-evocative clips designed to elicit a wide variety of spontaneous emotions, such as fear, disgust and surprise. Each frame of the collected videos was then FACS-encoded by certified FACS coder for a set of 12 specific AUs (see Table 2.1 for an overview). AUs were encoded on a scale from 0 (i.e., no activation) to 5 (i.e., maximum activation). In total, the DISFA dataset contains 130,814 encoded frames. The FACS encodings of the DISFA dataset had an inter-observer agreement score, quantified by an intra-class correlation coefficient [188] in the range between .80 and .94 per AU, which qualifies as a strong to very-strong inter-observer agreement score. The dataset is widely utilized within facial expression research (see e.g. [67, 202]). This set of encoded AUs is among the most studied in research regarding emotional expression and social interaction as described by [199].

2.3.1 Method

The method of analysis was identical to the one used in Study 1.

2.3.2 Results and Discussion

As the DISFA dataset is encoded for 12 AUs, all having a value between 0 (i.e., no activation) and 5 (i.e., maximum activation), 6^{12} (=2.2e+09) distinct facial configurations could be expressed by this AU selection. However, out of a total of 130,814 frames in the DISFA dataset, we only found 3,333 unique facial configurations. In addition, 2,664 (79.9%) of the 3,333 distinct facial configurations were 'idiosyncratic', i.e. those facial configurations were not shared among subjects and thus only expressed by one subject. This large number of idiosyncratic facial configurations described 18,533 frames (14.17%). Moreover, 949 of frames (0.73%) regarding those idiosyncratic facial configurations appeared in the dataset only once. Hence, these configurations only described one frame in the dataset.

Again, the most frequent facial configuration was the one described by all AUs having a contraction level of zero. This 'neutral' facial configurations described 48,616 (37.16%) of all frames with a SD = 20.60% over all participants (see Figure 2.2). Table 2.3 shows that the seven subsequent most present facial configurations only contained one activated AU. Thus, as the first eight configurations describe 50.48% of the frames, at least 50.48% of the frames were described by either a neutral facial configuration or a configuration with only one activated AU. Furthermore, all activation levels of twenty most occurring facial configurations are in the range of 0 and 3. In general, the intensity of those activation levels can be interpreted as 0, no activation; 1, trace of the action; 2, slight evidence; 3, pronounced; 4, severe; 5, maximum [63]. Hence, none of the AUs in the top twenty configurations was activated at a severe (4) or maximum level (5). This indicates that the more intense levels do not occur often. The histogram in Figure 2.3 shows the frequency of each activation level per AU. Indeed, all AU are mostly neutral (see also Table 2.1), and furthermore, when an AU is activated, it is most often only moderately activated.

Study 1 showed that most facial configurations in spontaneous communication are likely to be neutral. Similarly, Study 2 showed that even in a dataset derived from an experiment that was specifically designed to elicit emotional reactions from the participants, the most common facial configuration was neutral followed by some configurations with minor activations. However, one could still argue that in the first case the Map Task conversations were not emotional enough to produce many non-neutral facial configurations, and in the same line, that the second dataset did not use a strong enough video clip to elicit emotional reactions. Therefore, in our final study, we looked at a dataset that contains facial configurations elicited by a more direct elicitation method.

2.4 Study 3

The UNBC-McMaster Shoulder Pain Expression Archive (PAIN) Database [132] contains a total of 200 videos of 25 participant faces (12 male, 13 female). With the help of a physiotherapist, the participants were filmed while moving their painful shoulders. These 200 videos are comprised of a total of 48,398 frames, which are all FACS-encoded for a specific set of AUs by certified FACS coders. The PAIN dataset has an inter-coder percent agreement of 95% as determined by the Ekman-Friesen formula [132]. For this dataset, only those AUs that are potentially related to pain expression are encoded (see Table 2.1 for an overview).

2.4.1 Method

The method of analysis was identical to the one used in Study 1 and 2.

2.4.2 Results and Discussion

The PAIN dataset utilized 13 AUs. Each AU is encoded on a 6 level scale (0-5), except AU43 (eyes closed) which is encoded on a 2 level scale (i.e. present (5) or not(0)). Thus, $6^{12} * 2$ (=2.2e+09) distinct facial configurations could be expressed by this selection.

Table 2.3 Most frequent facial configurations in DISFA dataset.

%	Action Unit											
	1. Inner Brow Raiser	2. Outer Brow Raiser	4. Brow Lowerer	5. Upper Lid Raiser	6. Cheek Raiser	9. Nose Wrinkler	12. Lip Corner Puller	15. Lip Corner Depressor	17. Chin Raiser	20. Lip Stretcher	25. Lip Part	26. Jaw Drop
37.16%	0	0	0	0	0	0	0	0	0	0	0	0
2.99%	0	0	0	0	0	0	1	0	0	0	0	0
2.96%	0	0	0	0	0	0	0	0	0	0	2	0
2.26%	0	0	0	0	0	0	0	0	0	0	1	0
1.89%	0	0	2	0	0	0	0	0	0	0	0	0
1.18%	0	0	1	0	0	0	0	0	0	0	0	0
1.07%	0	0	3	0	0	0	0	0	0	0	0	0
0.97%	0	0	0	0	0	0	0	0	0	0	3	0
0.91%	0	0	0	0	0	0	0	0	0	0	2	1
0.87%	0	0	0	0	1	0	1	0	0	0	0	0
0.77%	0	0	0	0	0	0	0	0	1	0	0	0
0.72%	0	0	0	0	0	0	0	0	0	0	1	2
0.60%	0	0	0	0	0	0	0	0	0	0	1	1
0.60%	0	0	0	1	0	0	0	0	0	0	0	0
0.59%	0	0	0	0	0	0	2	0	0	0	0	0
0.54%	0	0	0	0	1	0	0	0	0	0	0	0
0.53%	0	0	0	0	0	0	1	0	0	0	2	0
0.50%	0	0	0	0	0	0	1	0	0	0	3	0
0.48%	0	0	0	0	0	0	0	0	0	0	0	1
0.44%	0	0	0	0	2	0	3	0	0	0	3	0

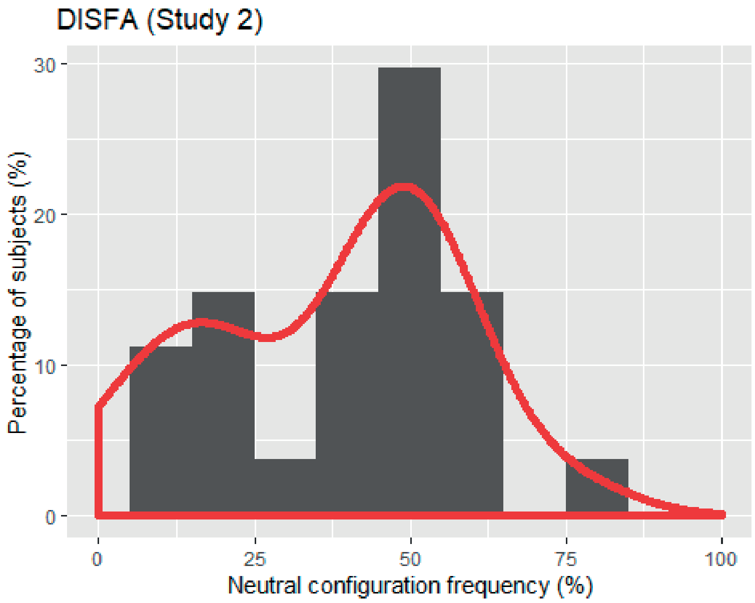


Figure 2.2 Distribution of neutral facial configurations (i.e., all AUs having zero contraction) among participants whose facial configurations were coded in Study 2.

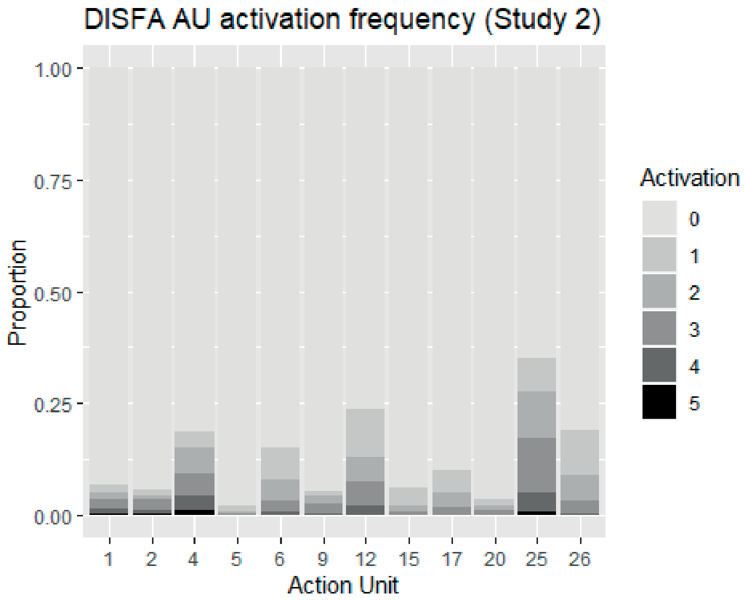


Figure 2.3 Frequency of the AU activation levels for each AU in the DISFA dataset.

Remarkably, the dataset contained only 476 distinct facial configurations. Despite the specific nature of the dataset, the most frequent facial configuration was the neutral one. 36,480 (75.4%) of all frames were described by the neutral facial configuration (see Figure 2.4 and Table 2.4). A total of 400 (84.2%) of the 476 distinct facial configurations were displayed by one subject only. These 'idiosyncratic' facial configurations described 3,620 frames (2.19%). Of the 400 'idiosyncratic' facial configurations (0.92%), 152 (.35%) are shown only in one frame.

Table 2.4 Most frequent facial configurations in PAIN dataset.

	Action Unit												
	4. Brow Lowerer	6. Cheek Raiser	7. Lid Tightener	9. Nose Wrinkler	10. Upper Lip Raiser	12. Lip Corner Puller	15. Lip Corner Depressor	20. Lip Stretcher	25. Lip Part	26. Jaw Drop	27. Mouth Stretch	43. Eyes Closed	50. Speech
75.38%	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
1.53%	0	0	0	0	0	1	0	0	0	0	0	0	0
1.16%	0	2	0	0	0	2	0	0	0	0	0	0	0
1.01%	0	0	0	0	0	0	0	0	0	2	0	0	0
0.95%	0	2	0	0	0	3	0	0	0	0	0	0	0
0.94%	0	0	0	0	0	2	0	0	0	0	0	0	0
0.85%	0	0	0	0	0	0	0	0	1	0	0	0	0
0.85%	0	0	0	0	0	3	0	0	0	0	0	0	0
0.84%	0	0	1	0	0	0	0	0	0	0	0	0	0
0.77%	0	1	0	0	0	1	0	0	0	0	0	0	0
0.71%	0	0	0	0	0	0	0	0	0	0	0	5	0
0.64%	0	0	0	0	0	0	0	0	2	0	0	0	0
0.56%	0	3	0	0	0	3	0	0	0	0	0	0	0
0.55%	0	0	1	0	0	1	0	0	0	0	0	0	0
0.43%	0	1	0	0	0	2	0	0	0	0	0	0	0
0.42%	0	0	0	0	0	0	0	0	0	1	0	0	0
0.41%	0	0	0	0	0	0	0	0	0	4	0	0	0
0.40%	0	0	2	0	0	0	0	0	0	0	0	0	0
0.34%	0	0	0	0	0	3	0	0	0	4	0	0	0
0.33%	0	4	3	0	0	4	0	0	0	0	0	0	0

Table 2.4 shows the twenty most frequent facial configurations in the dataset. It is notable that, in this dataset, unlike in the previous two studies, the third most frequent

expression already contains multiple AUs activated. Furthermore, the most frequent facial configurations contain activation levels from 0 to 4. Please note that AU 43 is encoded on a binary scale, thus the 5 for facial configuration 11 in Table 2.4 denotes eyes closed instead of an extreme contraction. Admittedly, in contrast to the previous two studies, the setting of the PAIN dataset is probably less in line with a situation in which an ECA is likely to be employed. Therefore, it is probably not useful to reproduce the exact facial configurations from this dataset onto an ECA. However, this dataset also confirms that neutral is the most common facial configuration, even while such an extreme elicitation method is used.

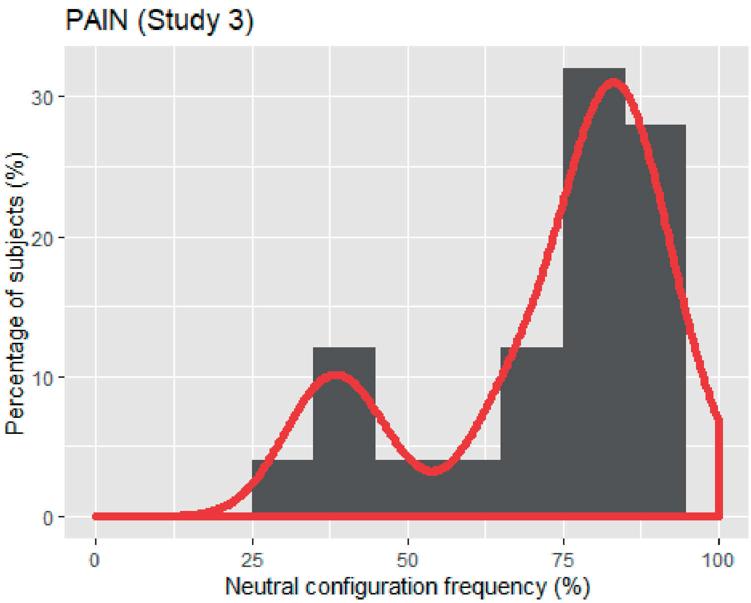


Figure 2.4 Distribution of neutral facial configurations (i.e., all AUs having zero contraction) over subjects in Study 3.

2.5 General Discussion

The goal of the present research was to determine the set of most common human facial configurations to guide research and development of ECAs. Creating ECAs that mimic human behavior with such precision that they are able to cross the uncanny valley is still a major challenge. On the one hand, the complexity of the face [102], which allows for a seemingly infinite number of distinct facial configurations according to facial action coding system (FACS), poses a combinatorial explosion of facial configurations that need to be implemented in the ECA. On the other hand, the high sensitivity of users towards small errors in ECAs' facial behavior demand for an accurate implementation of each facial configuration, which in turn is a time-consuming process. The combination of these two factors demand for a nearly infinite amount of time to implement all possible facial configurations with high accuracy. As a solution to this dilemma, developers often choose to limit ECAs' facial behavior to a small set of prototypical facial configurations [126, 169]. Although such limitation simplifies the problem of generating the appropriate facial behavior at the right time, it also poses the question if such a small set of prototypical facial configurations is the most common (i.e. most optimal) set of configurations to implement.

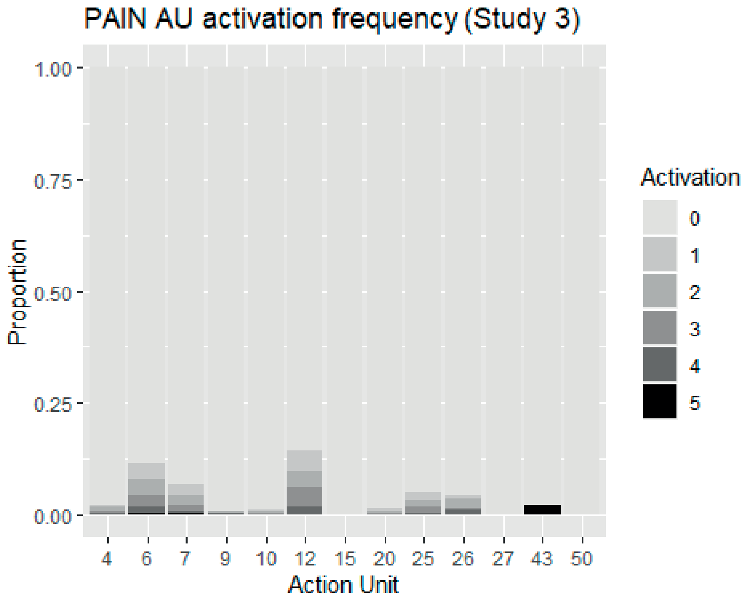


Figure 2.5 Frequency of the AU activation levels for each AU in the PAIN dataset.

In order to determine the distribution of human facial configurations, three human coded FACS-encoded datasets were analyzed. All three datasets were encoded by multiple human raters, with a high inter-rater reliability. Findings showed that the neutral facial configuration constitutes the largest part of the facial behavior in all three analyzed datasets. The main reason for such a large proportion of neutral faces may sprout from the fact that the face is not always involved in the display of emotional content [69]. Moreover, the display of some affective states may produce a neutral face as indicated by [142] who found that faces of bored students are indistinguishable from students with a neutral face. Furthermore, most of the time, emotions are not expressed or conveyed dramatically. In other words, emotional content is not always exhibited through the facial displays, in that other modalities such as vocal [75, 180] and bodily postures [49] are also involved in the expression of emotional content.

Given the large number of neutral facial configurations and the fact that humans use contextual information to attribute meaning to perceived facial configurations [178], facial configurations are perhaps ambiguous and not a strong predictor of a person's internal state [12]. Age and culture probably also played a role. Younger children are more expressive when it comes to emotions, than older children [193], thus if we would have analyzed facial configurations of children, we would probably have found a lower percentage of neutral facial configurations than with the current datasets, which all contained facial behavior of adults. Moreover, cultural display rules could also have played a role. Some cultures are more expressive than others due to social norms related to that cultural group. Those so-called cultural display rules informally describe how members of a culture group should express their emotions [139].

In addition, the neutral facial expression may be a transitional state that occurs in-between other visible emotional states reflected in the face. Another reason for the large amount of neutral facial displays could be the absence of an audience in the collected datasets. According to [68], people smile more when they are in a social environment (either real or imagined), compared to when they are alone. While the subjects of the Map Task dataset were recorded in a face-to-face setting, and the subjects of the PAIN dataset were recorded under supervision of a physiotherapist, the subjects of the DISFA dataset were indeed recorded in solitude. Thus, a proportion of the neutral faces in the DISFA dataset may be due to the lack of an audience. Following this line of reasoning, one might expect that these datasets that did include an audience would then contain less neutral faces than DISFA. However, that was not the case - on the contrary, DISFA contained the smallest proportion of neutral faces. A possible explanation is that of the three datasets, only DISFA was created with the intention to capture emotional facial displays from its participants. Thus, despite the fact that participants were alone, that intention may led to more variety in terms of facial display in the DISFA dataset than in the other datasets. Another factor that potentially caused a large number of neutral configurations, related to the AU coding scale, is the potential insufficient sensitivity of FACS. Due to the fact that FACS assumes only five levels of facial muscle contraction, facial configurations that contain facial muscle contractions below the first level are encoded as zero activation. Further research could inform us whether an extended version of FACS with a more sensitive encoding would be able to distinguish between neutral and non-neutral faces with more precision. Given the results of this research, what set of facial configurations should therefore be implemented in ECAs? Study 1 showed that in the minor case that facial behavior is non-neutral, most often AU2 (outer brow raiser), AU12 (lip corner puller) and AU4 (brow lowerer) are activated (see Table 2.1). According to [63], the outer brow raiser (AU2) is used during conversation to emphasize words or to highlight question marks, the lip corner puller (AU12) may be related to smiling and the brow lowerer (AU4) is part of the prototypical expressions of sadness and anger. However, the exact social and communicative functions of these frequent facial configurations can only be speculated for now and requires further research. Furthermore, from Study 2 we can derive that non-neutral facial configurations often only contain light activation intensities. Study 3 showed that pain-related facial behavior may have more pronounced facial configurations. Although the setting used to conduct the PAIN dataset is probably less comparable to a setting where an ECA will be deployed, Study 3 shows that in situations with a pre-defined context, specific facial displays may be required. Nevertheless, in generic situations, the facial behavior of ECA should rely on the most commonly-observed facial configurations, and when specific signalling is needed (e.g. in an emotional environment), the precedence should be given to that particular signal. A recent work by Stratou et al. has reduced the AU space by using factors that describe combinations of correlated AU activations [191] ; such approach implies that ECA facial displays can be manipulated with a small number of factors and a change in a specific factor would signify a change in all AUs related to that factor. However, results of this study hint towards the contrary; the most frequent facial displays are neutral or have only a single activated AU instead of a combination of AUs. Further research is needed to compare the approach of Stratou et al. [191] with the current study. Moreover, in case an ECA is deployed within a specific domain or context or with a specific purpose, the methods described in this paper can be used to analyze datasets that contain relevant facial behavior for such deployment.

One of the limitations of this study, is the small number of participants the analysis are based on (less than 100). This is because of the low availability of hand-coded FACS datasets with spontaneous facial behavior. In addition, each study analyzed a different set of AUs (see Table 2.1). An analysis based on a combination of the three datasets would not provide any generalizable results, as only AU4, AU12 and AU25 were encoded for all three datasets. Future research may include an extension of the current analysis with the output of automatically analyzed datasets of spontaneous facial displays. This would both increase the number of subjects and the AU overlap of the datasets. An additional benefit of such extension is that any results that specifically surfaced due to the experimental settings of the used datasets will automatically fade into the background. Another important point is that the current study used the frame as a unit of analysis, which provides insight into the most common facial configurations. However, due to this unit of analysis, the study does not inform about the dynamical element of facial expressions, therefore it remains an open question what the most natural sequence of facial configurations would be.

In conclusion, spontaneous facial behavior seems to revolve around the neutral facial configuration. It is therefore recommended, at least from an ECA development point of view, to start focusing on the most common facial configurations instead of the prototypical facial configurations. Restricting an ECA's facial behavior to a subset of facial configurations seems to be a promising approach, only if this set contains a reflection of real human facial behavior instead of uncommon prototypical configurations. Implementing a subset with most common facial configurations saves time and effort and would likely result in more realistic facial behavior, bringing crossing the uncanny valley closer.

2.6 Acknowledgements

This research has been funded by grants NSF-IIS-0416128, NSF-BCS-0826825, and PROJ-007246 from the European Union, OP Zuid, the Ministry of Economic affairs, the Province of Noord-Brabant, and the municipality of Tilburg awarded to Max M. Louwerse. The usual exculpations apply.

CHAPTER 3

3

Intrapersonal dependencies in multimodal behavior

Abstract

Human interlocutors automatically adapt verbal and non-verbal signals so that different behaviors become synchronized over time. Multimodal communication comes naturally to humans, while this is not the case for Embodied Conversational Agents (ECAs). Knowing which behavioral channels synchronize within and across speakers and how they align seems critical in the development of ECAs. Yet, there exists little data-driven research that provides guidelines for the synchronization of different channels within an interlocutor. This study focuses on intrapersonal dependencies of multimodal behavior by using cross-recurrence analysis on a multimodal communication dataset to better understand the temporal relationships between language and gestural behavior channels. By shedding light on the intrapersonal synchronization of communicative channels in humans, we provide an initial manual for modality synchronisation in ECAs.[3]

[3] This chapter is based on: Blomsma, P. A., Vaitonyte, J., Linders, G. M., & Louwerse, M. M. (2020). Intrapersonal dependencies in multimodal behavior. In Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), October 19–23, 2020, Virtual Event, Scotland Uk. ACM, New York, NY, USA, 8 pages .https://doi.org/10.1145/3383652.3423893

3.1 Introduction

Natural communication is multimodal [50]. It includes tightly interwoven verbal (i.e., speech) and non-verbal (i.e., facial expressions, eye gaze, body posture, and gestures) channels. The use of multiple channels facilitates communication in allowing speakers to express messages more clearly. For instance, the advantage of multimodality is apparent in co-speech gestures, hand and arm movements that spontaneously occur with spoken language [144].

Gestures can help disambiguate information in the spoken modality [95]. One can gesture while saying "The cup was *this* big" or point to a certain glass on the table while uttering "Can you pass me *that* one?". While gestures can convey complementary information to speech, they can also be redundant [18]. Gestures can thus vary in their semantic relationship to speech. Spoken and gestural modalities are not only related on a semantic level, they are also coupled on a temporal level [85, 212], such that the most meaningful part of gesture slightly precedes speech content [88]. Thus, on the production side, speech and gestures are related and on the comprehension side, listeners draw on both linguistic utterances and gestures to understand a speaker's message [109].

Studying the interplay of speech and gestures is important for understanding human-human interaction, but even more so for human-machine interaction. That is, when humans interact with Embodied Conversational Agents (ECAs), these agents display similar verbal and non-verbal competences to humans [43]. However, generating ECAs' non-verbal behavior, including gestures that need to be aligned with the generated speech, is not trivial. In order to do so, several aspects need to be considered. For example, should gestures be generated on the basis of speech, or should speech and gestures be two independent systems? Are there any contingencies and constraints as to which types of gestures occur with speech? If so, how are they aligned across time?

In human interlocutors, temporal dependencies between different multimodal behaviors have been shown to exist in face-to-face settings [129]. In the current study, we ask the question whether similar dependencies between spoken and gestural modalities also exist within a single speaker. Thus, to what extent are verbal and non-verbal behaviors aligned within an individual? The answer to this question is crucial in the development of ECAs. Without knowing which types of behavior go together, the generation of naturally-looking multimodal cues is comparable to looking for a needle in a haystack. First, this is because the proverbial haystack of possible behaviors is the combinatorial explosion of all the degrees of freedom of each modality. Without guidance, the number of possible behaviors is semi-infinite. Another difficulty involves selecting behaviors that come across as natural. Without guidance, such selection is not feasible. Therefore, we suggest that, to limit the number of possible behaviors, it is worthwhile to look at the relations between the spoken and gestural modalities at the intrapersonal level. It is plausible that certain speech-gesture combinations are not possible, meaning they do not occur in human communication. Such information would help diminish the size of the haystack and, in turn, simplify the process of finding the needle. Thus, a better understanding of the mechanisms that govern multimodal behavior within one individual could directly inform the development of ECAs.

In this study, we use a cross-recurrence quantification analysis to investigate speech-gesture dependencies within a speaker during face-to-face communication. The cross-recurrence analysis is applied to a multimodal communication corpus of about 25 hours of dialog that is encoded for speech (thirteen categories of dialog acts) and gestures (five types of gestures) at 250 ms intervals. The main goal of this study is to show that cross-recurrence quantification analysis is useful to unravel potential underlying patterns in intrapersonal multimodal behavior. We focus specifically on the correlations between different dialog acts and gesture behaviors. These results can be used by agent developers to increase the multimodal realism of their dialogue systems.

3.2 Background

3.2.1 Embodied Conversational Agents

An important part of ECAs' multimodal behavior generation is the temporal coordination of speech and gestures. Gesture plays a prominent role in conveying information in human communication [165] and humans are able to notice whether speech and gesture of ECAs are consistent or not [66]. In general, ECAs use rule-based approaches to produce gestures that are based on the speech that has been uttered [207]. However, it is difficult to construct precise rules about how the different modalities interact in communication as the relationships between different modalities on the production side remain unclear. Furthermore, large richly-annotated multimodal corpora are scarce due to the time-intensive and difficult labor required to create such corpora [113].

Also for this reason, most available corpora are specific to a domain or purpose. Hence, there are very few general-purpose multimodal corpora. Currently, rules lack the precision on a temporal level that data-driven techniques could provide, as rules are usually manually extracted from these multimodal corpora and implemented in ECAs [175]. BEAT was one of the first rule-based systems and it uses syntactic and semantic information to generate gestures and eye gaze behavior during speech [44]. Similar systems have been designed subsequently, such as the NUMACK system [114], the NVBGenerator [121] and more recently, the Cerebella system [124]. With the increase in multimodal corpora and computing power, researchers slowly started to investigate extracting patterns from data automatically. The first data-driven approaches focused on data from individual speakers [156].

These approaches were quickly extended to include data from multiple speakers. For example, [218] collected a dataset consisting of TED videos with subtitles and used an end-to-end learning approach to learn the relation between gesture and speech. Others have used prosody to automatically learn mappings between prosody and gestures [47] or a deep learning approach that learns gesture behavior from prosody, syntax and semantics [48]. Some researchers have used hybrid approaches that combine a rule-based approach with a data-driven one [19, 179]. Focusing only on iconic gestures, [19] uses a Bayesian network to decide which gesture to use, after which it is further specified by a set of rules. Finally, in a recent approach gestures were generated in an adaptive way through optimizations, based on feedback from human participants [36].

3.2.2 Dialog acts and gestures

Dialog acts are used to represent the pragmatic (contextual) meaning of user utterances in dialog. Although the communicative intention of a speaker is not marked explicitly in speech, dialog acts capture such intentions [10, 184]. Dialog acts and gestures could potentially be coupled in ECAs to generate realistic multimodal behavior. Just like prosodic, syntactic and lexical units, dialog acts are linguistic units that may relate to gestures, however wether such relationship exists, and if so, to what extent is an open question.

There are many theoretical arguments for the existence of such a relationship. [143] was the first to note that gestures also have a pragmatic function, next to their propositional content. [15] introduced a new class of hand gestures, called *interactive gestures*, stating the importance of (hand) gestures for the organization of discourse and turn-taking. On a pragmatic level, next to a contribution to the organization of the discourse, [110] distinguishes three other pragmatic functions. Gestures can provide the way speech should be interpreted, add to the meaning of speech by being an operator and also elucidate the speech act being used [110].

From an empirical point of view, however, there is little evidence for such relationship. [101] tried to establish an empirical relationship between dialog acts and gestures by analyzing the distribution of gestures across dialog acts in a small multimodal corpus of non-task-based conversations between three participants and found that turn-keeping utterances and fillers contain relatively the most hand gestures, while backchannels contain the least [101]. Moreover, self-touch motions occurred most with backchannels and laughter. This research helps to understand the different distributions of gestures among dialog acts, however the results do not necessarily imply a direct relationship between gesture and dialog acts. The data could simply reflect the fact that gestures help in dividing the information into small parts and thus help in conceptualizing information [111]. More concretely, gestures are more frequent during high cognitive load [112] and turn-keeping phrases intuitively have a higher cognitive load than backchannels. These results are interesting from a conversational agent point of view, since an ECA needs to know when to gesture. A similar study, looking at instances of hand gestures, observed relationships between dialog acts and certain interactive gestures [177]. So, while theoretically a relationship between gestures and dialog acts is not unlikely, current empirical evidence is for their relationship is unclear. In this study, we hope to shed more light on this relationship by investigating it with cross-recurrence quantification analysis.

3.2.3 Cross-recurrence Quantification Analysis

No default framework exists to scrutinize the temporal relationships between the different behavioral channels. We looked at frameworks utilized in studies regarding behavioral coordination between speakers and selected cross-recurrence quantification analysis (CRQA) [51]. Many other analytic frameworks disregard the temporal organization of the relationships and primarily aggregate data over the temporal dimensions of analysis. Such frameworks calculate event frequencies, rates or magnitudes [89]. Although such calculations have produced many insights, in the current research, we are primarily interested in the time-related patterns across the intention and gesture channels. If two events, i.e., a gesture activity and a dialog activity, often happen simultaneously, we could quantify this relationship by calculating the correlation between the two channels representing the activities. However, since the multimodal production system is more complex, this view would be too simplistic and a simple correlation would not be able to capture potentially complex relationships between the two channels. For example, it might be the case that events on two channels never happen at the exact same time, but follow or precede each other. This means that, next to the strength of the correlation, we also do not know the timing of this correlation.

In order to find out how two events relate temporally to each other, we need a measure that can quantify the relationship between them at different time shifts. One way to do this is to look at the co-occurrence of the events in time. For this, we use cross-recurrence analysis. This measure looks at the correlations in time, the recurrences of discrete events. The main idea behind cross-recurrence analysis is to determine how often events of one time series are succeeded (or preceded) by events of another time series, which is expressed in a proportional measure, called recurrence rate. This measure is obtained for a specific delay as follows: one of the time series is delayed, i.e., shifting all values of that time series a number of steps into the past. This means that the result of delaying a time series with one time step is that all values shift one step into history, such that the original value of the first time step is removed and the last element of time series is deleted. Thus, the length of the time series is one step shorter. In order to compare the shifted time series with a non-shifted time series, the last element of the non-shifted time series is also removed, such that both time series have the same length. The recurrence rate is obtained by creating a new time series that contains a 1 for each time point if both time series contain 1 for that time point as well. The recurrence rate is equal to the sum of the resulting vector divided by the length of that vector. The recurrence rate for a specific delay indicates how often an event on one channel co-occurs with an event on the other channel. Analyzing the recurrence rate for multiple delays informs us on how often events co-occur and within which time-frames [137].

3.3 Method

3.3.1 Dataset

A multimodal communication corpus of about 25 hours of dialog is used. 48 students from the University of Memphis participated (30 female, 18 male; 1 Asian, 19 African-American, 28 Caucasian) in a Map Task scenario [129]. This scenario requires two persons: an instruction giver, who has a map containing a route, and instruction follower, with a slightly different map without a route. The aim of the task is to reproduce the route of the instruction giver's map onto the instruction follower's map. The interlocutors can freely interact. However, they cannot see each other's map.

The corpus consisted of 13 encoded verbal and 10 encoded non-verbal behavior channels per interlocutor at 250 ms time intervals, with a total of 731,824 encoded intervals. The non-verbal behavior of each interval was encoded for five types of gestures according to the gesture coding scheme proposed by [144]: beat, deictic, iconic, metaphoric and symbolic gestures (emblems). Some gesture types were subdivided over multiple channels. The coders had a high inter-rater agreement score, quantified by Cohen's κ (.82).

Beat gestures are rhythmic hand movements that do not directly convey meaning but help marking discourse and organizing speech. Beat gestures were encoded over two channels: *beat single* contained the isolated, standalone beat events and *beat multiple* channel contained the sequences with multiple connected beat gestures. Deictic gestures referred to locations in space (i.e., pointing at something or someone). This space can be real (pointing at present objects and people) as encoded in the *deictic concrete* channel, or methaporical (pointing at abstract ideas, located in space), as encoded in the *deictic abstract* channel. Iconic gestures conveyed information about actions and object attributes by bearing partial resemblance to them, e.g., gestures depicting the path of the movement or the shape of the object. The behavior was encoded for iconic gestures, related to landmarks (*iconic landmark*) and those related to route information (*iconic route*). Metaphoric gestures were similar to iconic ones, in that they are also pictorial, but instead of bearing a resemblance to concrete entities, metaphoric gestures depict abstract ideas. Metaphoric gestures were captured in three different channels: *metaphoric level action*, gestures related to actions; *metaphoric meta-action*, metaphoric gestures related to meta-actions and *metaphoric metaphor*, metaphoric gestures that do not belong to the action and meta-action categories. Symbolic gestures (also called emblems) were conventionalized gestures (e.g., "thumbs up") and were least dependent on the speech content, as they do not require speech to be disambiguated. All symbolic gestures were encoded in one channel: *symbolic*. The communicative intention was encoded by thirteen different dialog act types that are typically used for Map Task scenarios [38]. The types included dialog acts that communicate new information (*instruct*, *explain*, *check* and *align*), responses to previous dialog (*reply-yes*, *reply-no*, *reply-what*, *acknowledgement*, *clarification*), dialog acts related to preparations in the experiment (*ready*) and an *unknown* category for unclassifiable dialog acts. Dialog acts were encoded at utterance level, thus each conversational turn could potentially consist of a sequence of dialog acts. The coders had a good inter-rater agreement score, quantified by Cohen's κ (.67).

The 23 behavior channels were encoded as binary time series. If a behavior channel contained an event at a certain interval (e.g., the person was making an iconic gesture related to the route), the value at that time point was encoded as 1. If no event was measured at that interval (e.g., person was not making an iconic gesture), the value was encoded as 0. The dataset was also used in [129] and [30].

3.3.2 Analysis

The dataset was loaded into the statistical computing software R as a matrix, such that each row represented a specific interval and each column, a specific behavior channel. Hence, the matrix comprised 731,824 rows and 23 columns. Subsequently, the recurrence rate was calculated for every possible combination of behavior channels. Thus, 23 behavior channels times 22 behavior channels resulted in 506 recurrence rate calculations. Each recurrence rate calculation involved calculating the recurrence rate per experiment (session) for every delay between 0 and 240 intervals (as each interval is 250 ms, this range corresponds with 0 and 60 seconds). The final result for each of the

506 recurrence rate calculations was the mean of the recurrence rates for each interval over all sessions for that specific combination. The final results were then plotted as a recurrence plot. As an example, the recurrence plots for *deictic concrete* with *beat multiple* and *deictic concrete* with *beat single* are shown in Figure 3.1. For the calculation of the recurrence rates, we used a custom-made script [4].

This script produced exactly the same results as the widely-used CRQA analysis package for R [51]. The only difference with the CRQA analysis package is that, due to some minor optimizations in the custom made script, it allowed us to use longer time series as input and to shorten the calculation times significantly. Finally, because the recurrence rate is a proportional value that is difficult to interpret on its own, we also calculated the random (chance level) recurrence rate for each combination to facilitate a contrast and to determine the significance of the results at a later stage. Unlike other works that have facilitated such contrast by randomized baseline patterns [129], which can be wobbly and difficult to interpret by the human eye, this work uses the random recurrence rate, which produces straight baselines that make interpretation more accessible for human raters. The random recurrence rate (RRR) was calculated by dividing the number of events of the second behavior channel by the length (number of intervals) of the first behavior channel, as in Equation (3.1), where:

- $T1_s$ is the number of events of the first behavior channel.
- $T1_l$ is the length of the first behavior channel.
- $T2_s$ is the number of events of the behavior channel.
- $T2_l$ is the length of the second behavior channel.

$$RRR = \frac{T1_s}{T2_l}\,\frac{T2_s}{T1_l} \tag{3.1}$$

3.3.3 Results selection

To minimize the risk of both Type 1 error, i.e., classifying the recurrence plot as having the effect where there is none, and Type 2 error, i.e., selecting a wrong time window, such that the plot contains a real effect within a different time range, the results were selected in three stages, both automatically and manually.

Each stage analyzed only the results selected in the previous stage. The first selection automatically discarded recurrence plots without any significant difference between the recurrence rate and the random recurrence rate. To calculate if such significant difference exists for a combination, the delay with the largest difference between the

[4] Code publicly available via: https://github.com/pblomsma/CRQA-turbo

recurrence rate and the random recurrence rate was taken as input for a significance analysis. The significance for this delay was calculated with a mixed-regression analysis with actual recurrence rate for that delay against the random cross-recurrence rate (see Equation 3.1), with role, session and dialog as fixed factors. The second selection automatically discarded recurrence plots where 20% or more of the time delays had a zero cross-recurrence rate, as this indicates data sparsity. The final selection was done by four raters, on an individual basis. For each recurrence plot, each rater judged whether the recurrence plots contained an effect or not. Plots that were not unanimously classified as having an effect were discussed in a group meeting. Plots for which at least three raters agreed upon a certain classification were then reclassified, while the others were identified as conflicted and removed from the analysis. In sum, all selected results had a significant peak or valley, contained less than 20% zero cross-recurrence rates and the cross-recurrence plot of those results were classified by at least three raters as having an effect.

3.4 Results and discussion

Cross-recurrence rates were analyzed for 506 combinations (23 times 22 behavior channels). Of the 506 combinations, 133 have been found significant by automatic analysis (described in Section 3.3.3), i.e., the largest distance between the random recurrence rate and the actual recurrence rate was significant for those 133 combinations. Four human raters classified 130 of those results unanimously as having an effect. An example of a recurrence plot with a significant effect is shown in Figure 3.1. In this figure, the temporal relationship between deictic gestures and beat gestures is examined from the perspective that beat gesture(s) happen at the same time or after the deictic gesture has started. Both types of gesture show the largest effect at time point 0, which means that both gestures most often occur simultaneously. It is also possible that beat gestures most often occur before deictic gestures, but that cannot be derived from Figure 3.1 as it only shows one side of the recurrence plot.

The results of the remaining 130 combinations were then further analyzed. For every recurrence plot, we identified the delay that corresponded with largest difference between random recurrence rate and actual recurrence rate between 0 and 1250ms. This time window was chosen to maximize the possibility that two behaviors are somehow related to each other. If the largest difference resulted from the actual recurrence rate being below the random recurrence rate, then such relation was classified as a mutual exclusive relation ("Mutex"), as this indicates that an event on the first behavior channel is not followed by an event on the second behavior channel. In other words, those events do not go together for that specific delay. On the other hand, recurrence plots where the largest difference resulted from the actual recurrence rate being higher than the random recurrence rate, were classified as "synchronized", as an event of the first behavior channel was followed above average by an event on the second behavior channel for that specific delay. These results are summarized in Figure 3.2. Due to the low number of participants, we deviated from good statistical practice by not splitting our dataset in two sets (i.e., one for the selection of effects, and one for the significance calculation), but used the same data for both the selection and the selective analysis. A potential consequence of this circular dependency is that the reported results could show inflated correlations[117]. The next paragraphs discuss the notable aspects of Figure 3.2, grouped by details regarding the gesture and dialog act modality, and the relations between both modalities.

3.4.1 Within gesture modality

One would have expected that (1) a person could only produce one gesture at a time and (2) that gesture events in general would not occur often. Thus, one would not expect to find any temporal synchronizations between gestures, and expect only mutual exclusive relations. However, because of the second conjecture it is likely that the number of events would not result in any significant relations, and we would therefore find only non-significant mutual exclusive relations between the gestures (resulting in grey squares Figure 3.2). Indeed, the results do not show any mutually exclusive relationship within the gesture modality. This could sprout from a data sparsity problem: i.e., not enough gesture events were available in the data to produce significant cross-recurrence valleys around time point zero. However, this non-exclusivity could also be explained by the fact that the boundaries between different gesture types are less clearly delineated than they are for the spoken modality. While it is possible to encode and classify gestures, they are not discrete and categorical the way words are [134]. Except for emblems that are conventional gestures, gestures are not “frozen” and can take on many forms. In consequence, being spontaneously produced, gestures are representative of thought processes or mental representations of an event [40], suggesting that gestures are not stored in the mental lexicon and happen on the fly.

However, against what one would expect, deictic and beat gestures seem to synchronize with a peak at 0 ms. Thus, speakers often start a deictic and beat gesture (both the *beats single* and *beats multiple* version) at the same time. Beat gestures are often found superimposed over other types of gestures [42], which could explain this finding. It could also be an encoding artifact. The starting phase of a beat gesture and a deictic gesture are quite similar, this similarity could have elicited uncertainty on the part of the encoders. Indeed, an inspection of the data reveals that out of the total 1562 intervals that were encoded as *beat single*, 987 (63.4%) were also encoded as deictic gestures. 705 (61.8%) of the 1141 intervals, that were encoded as *beat multiple*, were also encoded as deictic gestures. Thus, further research is needed to investigate if gestures in general do not happen together (with the exception of the aforementioned superimposed beat gestures).

3.4.2 Within dialog act modality

Regarding the relations among different dialog acts, one would expect to only find significant mutual exclusive relations between dialog acts with a peak delay at 0 ms. First of all, a speaker can only express one dialog act at a time, which implies that no dialog act combination can be synchronized and thus dialog acts should be mutually exclusive. Secondly, every utterance in the in the corpus (that contained about 25 hours of dialog) was encoded with a specific dialog act, the corpus should have enough dialog act data points to reap significant results. Finally, as a speaker most probably does not pause between dialog acts, one would expect to see 0 ms peaks.

Indeed, there is no synchronization between any of the dialog act combinations. However, contrary to the expectations, not all combinations have significant mutually exclusive relations either. Furthermore, not all significant mutually exclusive relations have a peak delay at 0 ms, Why is the peak of all the significant mutually exclusive relations not at the 0 ms delay? There are several possible explanations. First, small interruptions by the interlocutor with e.g., a backchannel or a short acknowledgement could cause the speaker to stop speaking for a moment to subsequently continue with a different dialog act. Secondly, it could be an artifact of the data, since the start of a new dialog act, after a previous one ends is often within 250 ms [190]. The reason that some combinations have non-significant results may, in hindsight, sprout from a data sparsity problem. Not all dialog acts have a high frequency in the data. As a result, some combinations of dialog acts do not occur often enough to reap significant results.

3.4.3 Between modalities

As indicated in the background section, several researchers have given theoretical arguments as to why a relationship between gestures and pragmatic functions (which are represented by dialog acts) should exist. However, empirically, little of this relationship has been verified yet. Figure 3.2 shows multiple significant mutually exclusive and synchronized relations. Most synchronized relations show that dialog acts lead gestures. In other words, a gesture starts after a dialog act has started. For example, the *deictic abstractness concrete* gesture starts after an *instruct* or *explain* dialog acts has started (among others). This is contrary to what has been argued in the literature [128].

However, the *beat single* gesture often starts before a *yes-no* or *what query*. Most mutually exclusive relations are led by the gesture, or in case of the *acknowledge* dialog act and the *beat single* gesture, are at the same time (0 ms). Deictic gestures do not go together with the dialog acts *ready*, *acknowledge*, and *reply-yes* and *reply-no*. Interestingly, those dialog acts are also often short and, according to literature, do not involve high cognitive effort [112]. However, further research is needed to claim with more certainty to what extent deictic gestures and cognitive effort are related. Why are *acknowledge* and *beat single* gestures mutually exclusive? Some studies show that beat gestures are used by the speaker to emphasize certain cues [59]. Thus, the likely explanation for why *acknowledge* and *beats* are mutually exclusive is that the former is mainly used in listening while the latter has a role in speaking – for emphasis and discourse structuring [145]. Thus, it seems logical that neither go together, since one is mainly used while speaking and the other while listening.

3.4.4 Precision of event generation

The results have shown the existence of several temporal dependencies between combinations of dialog acts and gestures. How can those dependencies, both mutual exclusive and synchronized, be translated to exact multimodal behavior generation rules for ECAs? Figure 3.2 can be used to draw up a first set of rules about which dialog acts and gestures do or do not go together. However, if we specifically want to know within which time interval both events should take place, then Figure 3.2 does not provide enough detail. Figure 3.2, only providing the peak delay of the recurrence, is not informative regarding the delay at which the recurrence starts and ends. Knowledge about the exact start and end delays can be translated into specific rules, that take into account how tightly or loosely, the two channels are coupled and thus the time frame during which two events should be generated together. Therefore, we report those start and end delays (which can be interpreted as a measure for spread) for every significant combination in Table 3.1. The spread is specified by the start and end intervals of the cross-recurrence effect. The start interval indicates the interval at which the cross-recurrence rate starts to rise above the random recurrence rate, and the end interval indicates the latest interval before the cross-recurrence rate crosses the random recurrence rate again. In case a combination signifies a mutually exclusive relation, it is exactly the opposite: the start interval indicates the first interval below the baseline and the end interval indicates the latest interval before the cross-recurrence rate intersects with the baseline again. For example, *beat single* gesture events have been found above baseline 5.50 seconds before an *explain* dialog act starts, *explain* dialog act events have been found 11 seconds before a *beat single* gesture starts.

Table 3.1 Spread indications

Channel A	Channel B	Start	End	Peak	
Beat multiple	Deictic	-21.00	21.75	0.00	+++
	Iconic landmark	-8.50	11.00	-0.25	+++
	Explain	-3.25	12.50	0.25	+++
Beat single	Beat multiple*	-0.50	0.00	0.00	- - -
	Deictic concrete	-50.00	60.00	0.00	+++
	Iconic landmark	-32.00	13.25	-0.25	+++
	Instruct	-37.75	8.00	0.25	+++
	Explain	-5.50	11.00	0.50	+++
	Acknowledge	-2.00	3.25	0.00	- - -
	Query-YN	-1.50	27.00	0.00	+++
	Query-W*	-2.75	0.00	-0.50	+++
	Clarify*	0.00	9.00	2.25	+++
Deictic concrete	Ready*	-2.25	0.00	-0.75	- - -
	Instruct	-37.50	60.00	0.50	+++
	Explain	-1.00	1.75	0.00	+++
	Check*	-18.50	0.00	0.00	+++
	Acknowledge	-19.00	2.75	0.00	- - -
	Query-YN	-2.00	2.50	0.25	+++
	Reply-Y	-6.50	6.00	-0.25	- - -
	Reply-N*	-4.00	0.00	-1.00	- - -
	Query-W*	-1.75	0.00	0.00	+++
	Check*	0.00	13.00	0.50	+++
	Align*	0.00	11.25	1.00	+++
	Query-W*	0.00	2.00	0.50	+++
	Reply-W*	0.00	9.75	0.50	+++
Iconic landmark	Reply-W	-0.75	4.25	1.25	+
	Reply-Y*	0.00	24.50	1.75	- - -
Iconic route	Iconic landmark*	-11.25	0.00	0.00	- - -
	Acknowledge*	-4.25	0.00	0.00	- -

Note. Pluses and minuses mark positive and negative regression coefficients.

Positive coefficients correspond with peaks, negative coefficients with valleys

(mutex). The number of symbols indicates p-level: +++ <0.001, ++<0.01, + <0.05. Combinations marked with a * had only direction that was taken into consideration for this table, as the other direction was filtered put by the selection process.

3.5 Conclusion

In the current study, we investigated the temporal dependencies between verbal and nonverbal behaviors. We specifically looked at which dialog acts and gestures do (or do not) occur in the same time frame.

The results provide multiple insights. First of all, significant effects for relationships within the gesture modality exist, such as deictic and beat gestures, and iconic and beat gestures being synchronized while relationships within dialog acts are mutually exclusive. Secondly, significant effects between dialog acts and gestures exist. These results function partly as a proof for the theoretical conjectures that the pragmatic function of speech is linked to the accompanying gestures [110], but also show that agent developers should not ignore those dependencies as they can help build more accurate multimodal behavior generation systems.

Interestingly, and contrary to the existing literature [128], we found that some dialog acts start before the gesture starts (e.g., people start with their explanation first, and then start to use deictic gestures). Furthermore, we found some instances of dialog acts and gestures that do not go together. Especially deictic gestures do not go with specific dialog acts (*ready*, *acknowledge*, *reply-no*, *reply-y*), which probably relates to the low information density of a typical instance of one of those dialog acts. In other words, deictic gestures are probably more appropriate with dialog acts that are more information-rich.

However, more research is needed to further analyze this relation. Acknowledgments and beat gestures do not go together either. We encourage ECA developers to compare their behavior generation rules with the measurements as presented in Figure 3.2 and Table 3.1 in order to get ECAs behavior generation closer to human behavior. Please note that this work solely reports the patterns found in the human speaker. What the perceptual effects are of the found patterns, and especially the effects of breaking such patterns, is an open question for future research. Finally, we hope that not only this article provides insights into the relationships between intentions and gestures, but that it also shows the agent community the possibilities of using cross-recurrence analysis to translate human behavioral data into patterns to be used for agent dialog systems.

3.6 Acknowledgements

This research has been funded by grants NSF-IIS-0416128, NSF-BCS-0826825, and PROJ-007246 from the European Union, OP Zuid, the Ministry of Economic affairs, the Province

of Noord-Brabant, and the municipality of Tilburg awarded to Max M. Louwerse. We thank the raters of the recurrence plots for their time and effort. The usual exculpations apply.

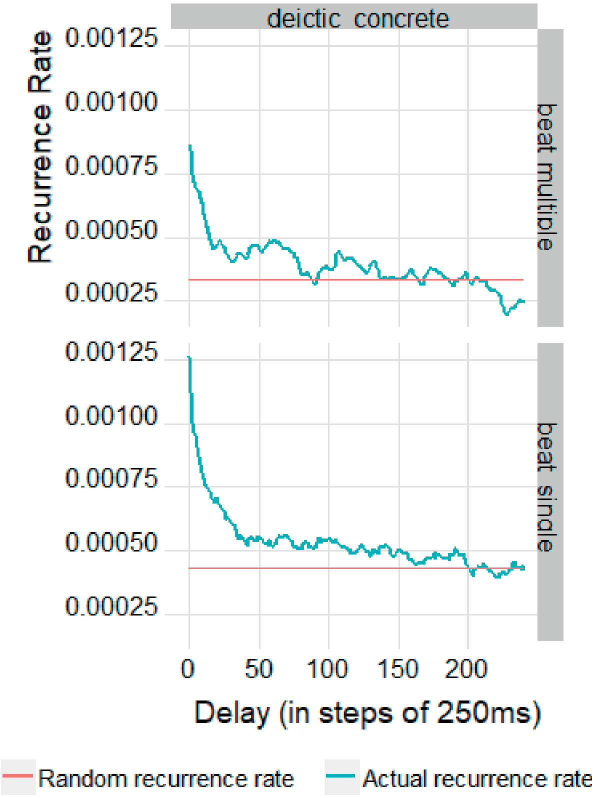


Figure 3.1 Recurrence plots for the behavior channel beat multiple and deictic concrete, and beat single and deictic concrete.

The graph can be interpreted as follows: the actual recurrence rate for e.g., the delay 50 equals 0.00045 in the upper graph. This signifies that in 0.045% of the total analyzed time an event occurred at the behavior channel beat multiple, 50 intervals (of 250 ms) after an event occurred at the deictic concrete behavior channel. As the recurrence rate does not take into account the number of total events of both behavior channels, the rate is difficult to interpret on its own, therefore the random recurrence rate (see Equation 3.1) is also plotted to contrast the recurrence rate and facilitate its interpretation.

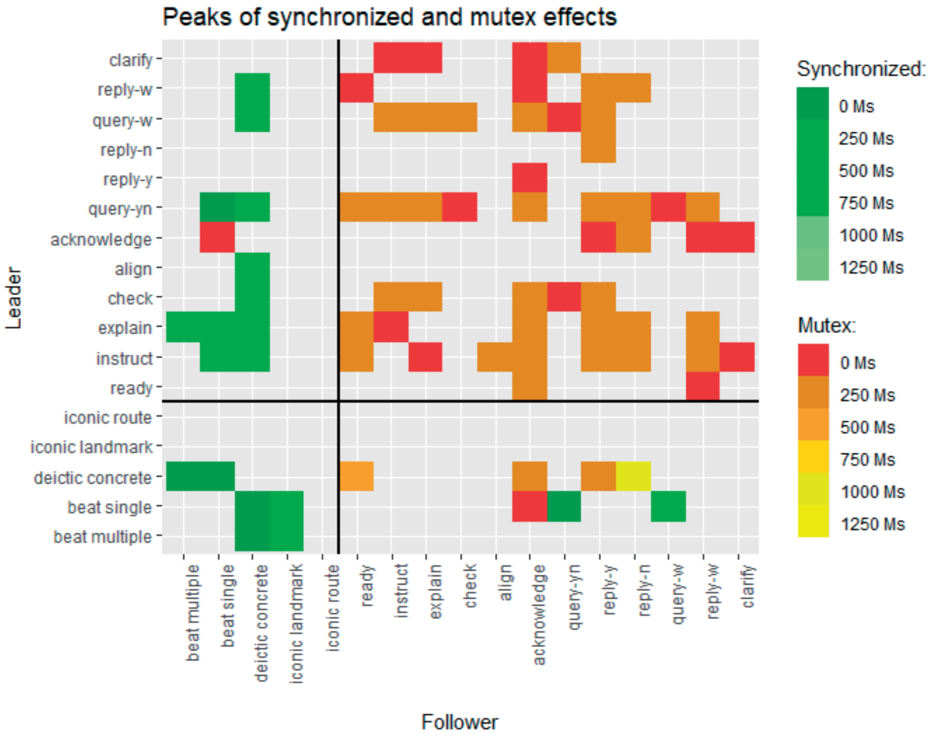


Figure 3.2 Overview of combinations of channels which had their most significant peak or valley within 0 ms and 1250 ms.

Combinations that had their most significant peak or valley outside of this timeframe, or had no significant difference at all, were left out of this overview.

CHAPTER 4

4

Variability between and within addressees in how they produce audiovisual backchannels

Abstract

In spoken conversations, speakers and their addressees constantly seek and provide different forms of audiovisual feedback, also known as backchannels, which includes nodding, vocalizations, and facial expressions. It has previously been shown that addressees backchannel at specific points during an interaction, namely after a speaker provided a cue to elicit feedback from the addressee. However, addressees may differ in frequency and type of the feedback that they provide, and likewise speakers may vary the type of cues they generate to signal the backchannel opportunity points. Research on the extent to which backchanneling is idiosyncratic is scant. In this article, we quantify and analyze the variability in feedback behavior of fourteen addressees who all interacted with the same speaker stimulus. We conducted this research by means of a previously-developed experimental paradigm that generates spontaneous interactions in a controlled manner. Our results show (1) that backchanneling behavior varies between listeners (some addressees are more active than others), and (2) backchanneling behavior varies between backchannel opportunity points (some points trigger more responses than others). We discuss the relevance of these results for models of human-human and human-machine interaction.[5]

[5] This paper has been submitted, but not yet published

4.1 Introduction

A spoken conversation can be operationalised as a highly interactive form of cooperative activity between at least two individuals. In that sense, it is more than an exact data transfer process, whereby a sender simply transmits information to a receiver, who then decodes the incoming message. The latter characterization of a spoken interaction does not do justice to the observation that an addressee is often more than a passive listener and is, in fact, co-responsible for a successful exchange of information [50]. Indeed, communication via speech can sometimes be a fuzzy endeavour, e.g. because of a noisy channel or the fact that a speaker may not correctly estimate a listener's prior knowledge about a specific state of affairs. As a result, it is typically the case that speakers and addressees seek and provide feedback on the smoothness of the interaction, to check whether information has successfully arrived at the other end of the communication chain. Accordingly, there is a growing interest in current models of spoken interaction regarding the systematicity of various types of feedback behavior.

In this paper, we are specifically interested in the brief responses, called *backchannels* [217] that addressees return during an interaction. Such backchannels, which can be verbal and non-verbal, serve as cues to show to a speaker that an addressee is engaged and listening. Backchannels thus convey attention and interest to the speaker and they can also regulate turn-taking [84]. While verbal backchannels include vocalizations (laugh, sigh, etc.), paraverbals ('mm-hmm', 'uh-huh', etc.), and short utterances ('really', 'yeah', 'okay'), non-verbal backchannels consist of facial expressions, nodding, eye gaze, and gestures. It has been shown that there is a marked difference between signals that serve as "go-on" cues i.e., to make clear that the addressee has correctly processed the incoming message and signals that highlight a possible communication problem so that a speaker-sender may have to repair a potential error [115, 187, 80].

In the literature, backchannels are distinguished from turn-taking cues. The intention of a speaker, when backchanneling, is to signal that the current speaker is still in charge of the turn, while the intention of a turn-taking cue is to interrupt the speaker and to take the speaking turn. Thus, backchannels can be viewed as a form of cooperative overlap, or from a turn-taking perspective, as a turn-yielding cue [21].

4.1.1 Backchannel-inviting Cues

It has been shown that the timing of backchannels is crucial to guarantee a smooth interaction [81, 170]. For instance, [81] demonstrated that a wrongly-timed head nod from a listener can disrupt a speaker, which suggests that addressees typically are efficient at producing backchannels at the right points in an interaction. Indeed, research shows that backchannels occur at specific points in a conversation, e.g., after the speaker gives a so-called backchannel-inviting cue [84], also called backchannel-preceding cues [123]. The specific behaviors that the speaker produces to transmit backchannel-inviting cues to elicit backchannel behavior from an addressee occur in different forms, including the usage of specific prosodic patterns. [83] demonstrated that speakers use rising and falling intonations to elicit feedback, [45], [210] showed that listeners often provide a backchannel after speakers have lowered their pitch for at least 110 ms, [45] showed that pauses in the speaker's speech, and also certain parts of speech are predictive of backchannels. [17] revealed that mutual gaze often occurs prior to a backchannel being produced. Speakers may not be aware of sending out backchannel-inviting cues, but listeners and observers are capable of picking up on those signals. [16] showed that listeners are even able to provide backchannels at the right moment when not attending to the content of the speech.

4.1.2 Backchannel Opportunity Points

Although speakers provide backchannel-inviting cues, it is up to the addressee to pick up on these cues and identify relevant moments in a conversation to produce backchannels. Those moments in a conversation, where it's appropriate for an addressee to provide some kind of listener feedback, are referred to as backchannel opportunity points (BOPs) [81]. BOPs, which are also known as jump-in points [149] and response opportunities [55], are points in the interaction where an addressee could or would want to provide feedback in reaction to the speaker [54].Prior studies show that not all BOPs are used by addressees to provide a backchannel [108, 170]. The potential reason why not every BOP is seized is that people's listening behavior is idiosyncratic. However, research to which listening behavior is idiosyncratic is scant.

4.1.3 Current Work

The goal of this study is to shed light on the variation that exists in backchannel behaviors across addressees and within an individual addressee. Specifically, we ask (i) what types of behaviors are utilized by addressees to give feedback during BOPs?; (ii) how does feedback behavior differ across different addressees?; (iii) to what extent differs the behavior of addressees for the same BOP?

The fact that we expect there to be variability between and within addressees in their feedback behavior is in line with previous findings that human beings do not have a fixed communication style. Speakers have been shown to adapt their way of speaking depending on the situational context, such as the type of addressee or the specific environment. Typically, speakers talk differently to children or adults, and switch to a different style when they notice that their partner experiences some problems of understanding (e.g., because that person is not a native speaker) [32]. Along the same lines, there may be differences across addressees, e.g., depending on personality traits or the mere fact that some addressees have more developed communicative skills [213]. It could be expected that addressees may vary in how they produce backchannel behaviors, with some spots in the interaction eliciting stronger or more backchannels than others (e.g., because such a cue is felt to be more needed). Also, some addressees may be more extravert or engaged so that one could expect differences across addressees as well. Insight into the variability of audiovisual backchannel behavior is not only informative to understand how human-human communication proceeds, but it is also relevant for practical applications, such as models of human-computer interaction, specifically, social robots and embodied conversational agents

(ECAs) [43], also known as Socially Interactive Agents (SIAs) [133]. In a similar manner to human-human interaction, it could be useful for ECAs to vary in the extent to which they backchannel, e.g., depending on the type of user, context and application. It it also likely that inducing variability may render the interaction style of an ECA more natural and less monotonous, similarly to the efforts to synthesize variability in speech and language generation systems [73]. However, modelling natural backchannel behavior for artificial entities is a non-trivial task for at least two reasons. One of the difficulties lies in detecting and appropriately responding to backchannel-inviting cues. Another difficulty is that due to backchannel behavior being idiosyncratic, it is not easy to define what a typical backchannel behavior should consist of for an ECA. To investigate variation in backchannel behaviors and to answer the research questions above, we conducted a computational study based on the data collected in a human experiment that used the so-called o-cam paradigm [79]. The o-cam paradigm was set up to allow comparisons between multiple addressees who are exposed to identical conversational data from the same speaker stimulus. The computational study consisted of two analyses. Analysis I examines the speaker stimulus, specifically the identification of BOPs, the categorization of those BOPs and the prosodic properties of the backchannel-inviting cues preceding the BOPs. Analysis II investigates the addressee behavior during the BOPs. We compared the behavior of the addressees across multiple channels (i.e., facial expressions, head movement and vocalizations) to examine the degree of variability between and within addressees.

4.2 Dataset

This study employed the materials of a database previously recorded during an experiment conducted by [34]. The database consisted of (1) one video recording of the stimulus, henceforth the ‘speaker’, and (2) the video recordings of 14 participants who were filmed during the experiment, henceforth ‘addressees’. Each video was 8.42 minutes long and contained 6.25 minutes of conversation, the remaining time was used for game-related tasks such as preparing and answering questions (see explanation below). The number of participants is comparable to similar backchannel studies, including [118] and [171].

The recorded experiment was based on the o-cam paradigm [79], an experimental design that combines the advantages of online paradigms (i.e., highly controllable environment, easy to run) with the advantages of offline settings (i.e., high ecological validity). The core concept of the o-cam paradigm is that a participant thinks that s/he is having a computer-mediated conversation with another participant (i.e., an interaction via a video conferencing setting), while, in reality, the other participant is a confederate whose video is pre-recorded. Certain manipulations are used in the setup to make a participant think it is a real life conversation [79]. The o-cam paradigm has been previously utilized to, for example, study the relationship between gender and leadership capabilities [98], and investigate the influence of smiling behavior [152].

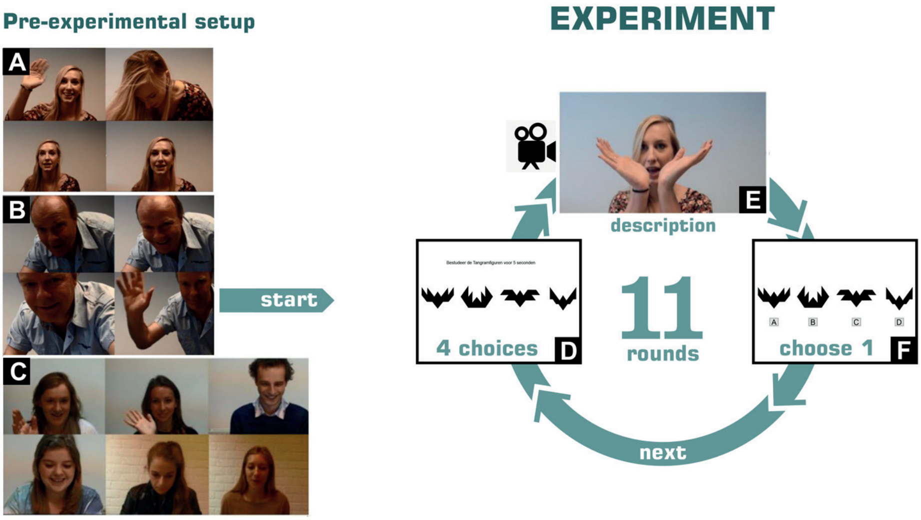


Figure 4.1 Visual impression of the o-cam experiment.

First the participant is prepared (A-C), after that 11 rounds are played: In each round the participant is shown 4 figures (D), followed by a description of 1 of those figures (E) after which participant indicates which figure is described (F).

The experiment reported in [34] was aimed to elicit feedback-behavior from the participants. Each addressee played a Tangram-game with the speaker (who was a pre-recorded confederate) via computer mediated connection. During the experiment, the addressee was presented with 4 Tangram figures for 5 seconds, followed by a description of one of those Tangrams provided by the speaker. The participant's task was to choose the figure from the 4 Tangram figures based on the description by the speaker. See Figure 4.1 for a visual illustration of the experiment. The experiment consisted of 11 rounds in which each time a different quadruple of Tangram figures would be used. The participants were told that the experiment was related to abstract thinking and that they were not allowed to ask questions since asking questions would make the game too simple. The confederate (the speaker) was not informed about the goal of the study in order to keep the experiment as ecologically valid as possible. After the experiment, participants were asked if they suspected that instead of a live interaction they were presented with a pre-recorded video of another person.

The data of five participants were discarded because they answered positively, whereas one participant asked a question during the experiment and thus their data was also discarded.

4.3 Analysis I: Speaker Behavior

The first analysis regards only the speaker behavior to identify the backchannel opportunity points ('BOPs') and to analyze the audiovisual behavior of the speaker during the backchannel-inviting cues preceding the BOPs. The identified BOPs are subsequently used in Analysis II to investigate the addressee feedback behavior. An obvious approach to identify the BOPs would be to annotate the backchannel behavior for each of the addressee videos separately. However, such an approach comes with at least two disadvantages. As addressees do not necessarily utilize all BOPs to provide feedback, analyzing the addressees would thus not necessarily result in the identification of all BOPs. Furthermore, using the same data for selection and selective analysis would result in a circular analysis also known as 'double dipping' [117]. Therefore, we identified the BOPs based on the speaker stimulus.

4.3.1 Methods

BOP identification

We used parasocial consensus sampling [90, 100], which takes the advantage of the fact that humans, especially as a third-party observer, can aptly point out BOPs in a conversation [55]. The approach consisted of two steps: identification of possible BOPs by a jury of multiple judges, followed by the aggregation of the output of the jury to determine genuine BOPs. Genuine BOPs are those BOPs that are identified by at least a certain percentage of judges.

For the identification of BOPs, we used a human jury that consisted of 10 judges. Each judge watched the speaker video and identified each moment that s/he thought was appropriate to backchannel. Each judge was instructed in the same way. First, they were explained what backchanneling behavior is, namely, the listening signals one gives during a conversation that includes head nods and sounds like 'uh-uh', 'hmm', and 'hm hm' and combinations of nods and sounds. Next, they were asked to watch the speaker video and to make a sound (e.g., 'yes') when s/he thought it was appropriate to backchannel, either verbally, non-verbally or both. The audio of the judge was recorded.

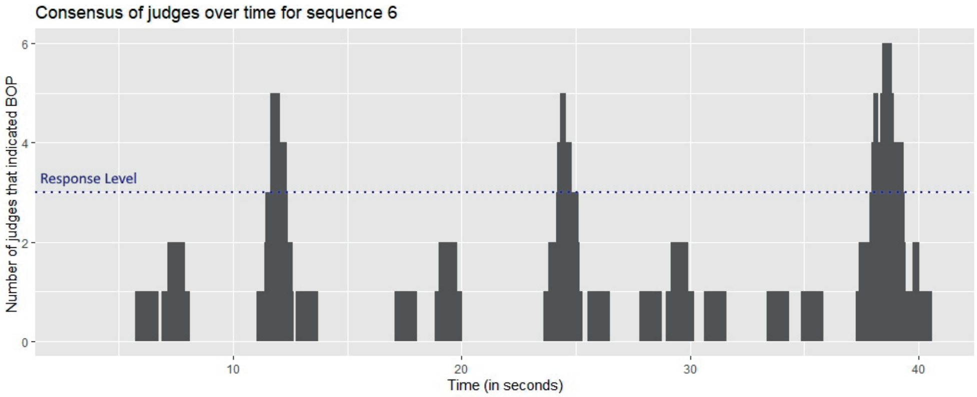


Figure 4.2 Illustration of a part of the speaker stimulus, with at each point in time the number of judges that indicated the presence of a BOP.

If three judges or more indicated a BOP at a certain point, then this point is considered as a genuine BOP.

The aggregation of all the recordings of judges allowed us to determine for each data point in the stimulus, the percentage of judges that thought that a specific moment was a BOP. BOPs that were agreed upon by a minimum percentage of judges were classified as a genuine BOP and selected for a further analysis. The minimum percentage is based on the expected numbers of backchannels in the recording. [170] states that one could expect from 6 to 12 backchannels per minute. Since our recording was 6.25 minutes, we therefore expected between 38 and 77 backchannels.

All the recordings of judges were preprocessed with Audacity [9]: We used a Noise Gate filter (250 ms attack and -12.50 dB grate threshold) to remove background noise and a 20 dB audio amplification to ensure that a judge was audible. Each recording was then converted to a binary timeseries with a resolution of 25 frames per second (FPS), in such a way that frames that contained a sound with an amplitude above 0.1 were converted to 1, and otherwise, to 0. Although [100] used a resolution of 10 FPS, we decided to use 25 FPS as this matched with the FPS of both our video recording and the FaceReader encodings (as described in the subsequent section).

Because judges had to vocally indicate visual backchannels, which start on average 202 ms before a vocal backchannel [215], the onset of each indication was set to 202 ms before the actual onset in order to correct for a potential delay. Each onset of a judge's indication was converted to a potential BOP of the length of 1000 ms in line with [100]. Finally, a timeseries was created with a resolution of 25 FPS, where each frame (i.e., sample) contained the number of judges that indicated a BOP for that frame.

BOP types: Continuer and End-of-turn

To gain further insight into whether specific BOPs or BOP types affect average addressee behavior, we subdivided the BOPs into two categories. Although each BOP functions as a moment for the addressee to acknowledge certain information, we conjecture that the urge to acknowledge is the strongest at the end of each game-round. After all, no further information will follow the last BOP of a game-round, and thus the addressee should have enough information to answer the question at that point. And if not, the addressee should indicate that at that last BOP. Therefore, we estimate that the most expressive addressee behaviors will be observable at the last BOP of a round. Hence, we have created the following categories: (1) all BOPs that are the last of a round, we called this category end-of-turn (EOT), and (2) all other BOPs that are placed during a round, we called this category Continuer. Given this categorisation, the EOT category contained 11 cues and the Continuer category contained 42 cues.

Backchannel-inviting cues

To verify that indeed the (visual) prosody is different for backchannel-inviting cues compared to the prosody used during remaining part of the conversation, we analysed the pitch properties, facial behavior and head movement of the speaker's backchannel-inviting cues that preceded the identified BOPs. The cues were isolated by selecting the last 1000 ms of the speaker stimulus sound before the start of each BOP. Although there is no consensus on the length of such samples in literature, e.g. [189] analysed the last 200ms of the voiced region for pitch, while [123] reports longer sample lengths including 1000ms.

We choose 1000ms to be on the safe side of finding a voiced part in the sample. The pitch properties were extracted with Praat [31]. Of each sample the F0 values (i.e. the fundamental pitch values), were extracted with a precision of 100 frames per second. Trailing and leading frames that did not contain pitch information were discarded. For each sample the average, minimum, maximum, F0-range (which is the maximum minus the minimum), average and form were obtained. The form was calculated by subtracting the average pitch of the second half of the sample from the average pitch of the first half of the sample, such that a negative number for form means an increasing pitch and a positive number means a decreasing pitch.

The facial behavior and head movements were analyzed based on the output of FaceReader 8 software [159]. The stimulus video was encoded with Action Units as based on the Facial Action Coding System [64]. Every frame of the videos was encoded with the following Action Units: 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 18, 20, 23, 24, 25, 26, 27, 45 as well as X, Y and Z coordinates were extracted for head orientation. Each AU can be scored for intensity on an ordinal scale from 0 (i.e., absence of an AU) to 5 (i.e., maximum intensity). For some frames in the dataset, FaceReader was unable to detect a face, and thus also unable to encode head position and/or AU activations.

Head nods were quantified for all backchannel-inviting cues following [161]. Specifically, for head nods we extracted amplitude and frequency. Amplitude equals the maximum tilt angle, i.e. the difference between the minimum and maximum X rotation angle. Frequency is the sum of upward and downward peaks per second of the X rotation angle. To prevent that small noise-related changes in elevation direction would influence the frequency, we ignored upward and downward peaks that differed a maximum of 1 degree. In order to verify if the backchannel-inviting cues differed from non-backchannel-inviting cues, each backchannel-inviting cue was paired with a randomly selected voice sample from the speaker stimulus. Paired T-tests were conducted between the obtained pitch properties, head movements and the average AU activation of the backchannel-inviting cues and the non-backchannel-inviting cues. The Bonferroni correction was applied for the multiple pairwise comparisons.

Subsequently, the analyzed properties of the backchannel-inviting cues of the EOT category were compared with the Continuer category. The two categories were compared with the Welch's T test for significance, and also corrected with Bonferroni.

4.3.2 Results

BOP identification

The number of identified backchannels per response level is depicted in Figure 4.2. Genuine (i.e. definite) BOPs were based on a consensus-level of 30% (3 coders) such that 53 BOPs were taken into account. The average duration of the identified BOPs was 934 ms (*SD* = 403).

Backchannel-inviting cues

The backchannel-inviting cues had a higher maximum pitch and larger F0-range, compared to the random selected samples. There were no significant differences for average pitch, minimum and form. The highest pitch observed in backchannel-inviting cues was on average 350.36 Hz (*SD* = 106.94 Hz), while the highest pitch in the random samples had a lower average of 201.30 Hz (*SD* = 70.88 Hz). The F0-range for backchannel-inviting cues was on average 156.07 Hz (SD = 111.84 Hz) , while the random samples had a lower average F0-range of 102.34 Hz (*SD* = 71.77 Hz). See Table 4.1 for all the results. The speaker's head movements and facial behavior did not differ significantly between cues and non-cues, and also not between EOT and Continuer related inviting-cues (see Table 4.3 and 4.4).

Table 4.1 Pitch properties of backchannel-inviting cues, compared to non-cues.

	Cue (1)	Non-Cue (2)	Diff (1)-(2)	df	Cohen's d
Average	246.95 (37.23)	250.72 (40.63)	-3.77	48	0.10
Min	194.29 (43.44)	303.64 (54.23)	-109.35	48	0.14
Max	350.36 (106.94)	201.30 (70.88)	149.06*	48	0.51
F0-range	156.07 (111.84)	102.34 (71.77)	53.73*	48	0.57
Form	16.10 (52.87)	16.66 (49.74)	-32.76	48	0.45

Statistics are based on paired t test analysis. All values are in Hertz. The Diff score is the result of subtracting the mean Cue value of the mean random value.

*p<.05, ** p<.01, *** p<.001

The backchannel-inviting cues that preceded BOPs from the EOT category had a significant lower average pitch, as compared to the cues that preceded the Continuer category. The form was also markedly different, EOTs have a downward going pitch on average, while the other cues had a upward going pitch on average. There were no significant differences for minimum, maximum and F0-range. See for an overview of the results, Table 4.2.

Table 4.2 Pitch properties of backchannel-inviting cues that precede EOTs vs cues that precede Continuers

	EOT (1)	Continuer (2)	Diff (1)-(2)	df	Cohen's d
Average	218.68 (13.05)	253.68 (38.02)	-35.00***	43.30	1.00
Min	159.37 (46.51)	202.61 (38.79)	-43.24	12.16	1.07
Max	318.66 (117.47)	357.91 (104.37)	-39.25	12.60	0.37
F0-range	159.28 (129.43)	155.30 (108.99)	3.98	12.22	0.04
Form	23.25 (21.97)	-25.47 (53.88)	48.72***	36.72	0.98

Statistics are based on Welch's t test analysis. All values are in Hertz.

*p<.05, ** p<.01, *** p<.001

Table 4.3 Averages of different channels (and standard deviations) over backchannel-inviting cues and non-backchannel-inviting cues.

	Cue (1)	Non-Cue (2)	Diff	df	Cohen's d
Head Movement					
Frequency	4.80(2.41)	4.65(3.37)	0.15	49	0.05
Amplitude	11.39 (7.43)	9.99 (7.37)	1.41	49	0.19
Facial Gestures					
Inner Brow Raiser (AU1)	0.33 (0.49)	0.32 (0.48)	0.02	49	0.04
Outer Brow Raiser (AU2)	0.23 (0.52)	0.09 (0.29)	0.14	49	0.34
Brow Lowerer (AU4)	0.53 (0.84)	0.38 (0.65)	0.15	49	0.19
Upper Lid Raiser (AU5)	0.05 (0.24)	0.02 (0.14)	0.03	49	0.16
Cheek Raiser (AU6)	0.64 (0.73)	0.40 (0.56)	0.24	49	0.36
Lid Tightener (AU7)	0.00 (0.03)	0.06 (0.22)	-0.05	49	0.34
Nose Wrinkler (AU9)	0.00 (0.00)	0.00 (0.00)	0.00	49	0.00
Upper Lip Raiser (AU10)	0.00 (0.00)	0.00 (0.00)	0.00	49	0.00
Lip Corner Puller (AU12)	0.06 (0.23)	0.03 (0.14)	0.03	49	0.16
Dimpler (AU14)	0.66 (0.91)	1.05 (1.20)	-0.39	49	0.39
Lip Corner Depressor (AU15)	0.00 (0.00)	0.00 (0.00)	0.00	49	0.00
Chin Raiser (AU17)	0.07 (0.27)	0.05 (0.21)	0.01	49	0.06
Lip Puckerer (AU18)	0.02 (0.12)	0.04 (0.16)	-0.01	49	0.08
Lip stretcher (AU20)	0.00 (0.00)	0.01 (0.06)	-0.01	49	0.20
Lip Tightener (AU23)	0.00 (0.00)	0.00 (0.02)	-0.00	49	0.20
Lip Pressor (AU24)	0.00 (0.00)	0.00 (0.00)	0.00	49	0.00
Lips part (AU25)	0.00 (0.03)	0.01 (0.06)	-0.01	49	0.17
Jaw Drop (AU26)	2.43 (0.89)	2.53 (0.87)	-0.10	49	0.12
Mouth Stretch (AU27)	0.09 (0.27)	0.17 (0.45)	-0.07	49	0.20
Eyes Closed (AU43)	0.00 (0.00)	0.00 (0.00)	0.00	49	0.00

Statistics are based on paired t test analysis. The Diff score is the result of substracting the mean BOP value of the mean non-BOP value. * p <.05, ** p<.01, *** p<.001

Table 4.4 Averages of different speaker channels (and standard deviations) of backchannelinviting cues that precede EOTs vs cues that precede Continuers

	EOT (1)	Continuer (2)	Diff	df	Cohen's d
Head Movement					
Frequency	5.27 (2.72)	4.68 (2.34)	0.60	14.30	0.25
Amplitude	13.70 (8.33)	10.76 (7.15)	2.94	14.30	0.40
Facial Gestures					
Inner Brow Raiser (AU1)	0,50 (0,58)	0,29 (0,47)	0,21	4,20	0,44
Outer Brow Raiser (AU2)	0,44 (0,65)	0,17 (0,47)	0,26	12,94	0,52
Brow Lowerer (AU4)	0,96 (1,22)	0,41 (0,68)	0,54	11,73	0,67
Upper Lid Raiser (AU5)	0,00 (0,00)	0,06 (0,27)	-0,06	39,00	0,27
Cheek Raiser (AU6)	0,96 (0,90)	0,55 (0,67)	0,41	13,17	0,56
Lid Tightener (AU7)	0,00 (0,00)	0,01 (0,04)	-0,01	39,00	0,18
Nose Wrinkler (AU9)	0,00 (0,00)	0,00 (0,00)	0,00	0,00	0,00
Upper Lip Raiser (AU10)	0,00 (0,00)	0,00 (0,00)	0,00	0,00	0,00
Lip Corner Puller (AU12)	0,00 (0,00)	0,07 (0,25)	-0,07	39,00	0,31
Dimpler (AU14)	0,98 (1,12)	0,58 (0,84)	0,40	13,28	0,45
Lip Corner Depressor (AU15)	0,00 (0,00)	0,00 (0,00)	0,00	0,00	0,00
Chin Raiser (AU17)	0,02 (0,07)	0,08 (0,30)	-0,06	48,72	0,21
Lip Puckerer (AU18)	0,02 (0,07)	0,03 (0,13)	0,00	30,90	0,03
Lip stretcher (AU20)	0,00 (0,00)	0,00 (0,00)	0,00	0,00	0,00
Lip Tightener (AU23)	0,00 (0,00)	0,00 (0,00)	0,00	0,00	0,00
Lip Pressor (AU24)	0,00 (0,00)	0,00 (0,00)	0,00	0,00	0,00
Lips part (AU25)	0,00 (0,00)	0,00 (0,03)	0,00	39,00	0,18
Jaw Drop (AU26)	2,53 (0,82)	2,40 (0,92)	0,13	17,60	0,15
Mouth Stretch (AU27)	0,25 (0,44)	0,06 (0,18)	0,19	10,97	0,75
Eyes Closed (AU43)	0,00 (0,00)	0,00 (0,00)	0,00	0,00	0,00

Statistics are based on paired t test analysis. The Diff score is the result of substracting the mean BOP value of the mean non-BOP value.

p <.05, ** p<.01, *** p<.001

4.4 Analysis II: Addressee Behavior

In the following subsection, we first compare audiovisual feedback behavior at BOP and non-BOP spots in the spoken messages. Then, we focus on BOPs only to see to what extent we can observe variability in audiovisual feedback behavior within and between addressees.

4.4.1 Methods

Semi-automated measures of audiovisual behavior

The videos from the addressees were all encoded for facial expressions, head movements, and vocal backchannels as follows. The head movements and facial behavior were analyzed analogue to how the backchannel-inviting cues were analyzed (see Section 4.3.1). The vocal backchannels of the addressee videos were manually encoded by 1 coder with ELAN 6.0 encoding software [214]. The coder indicated the moments that an addressee made a sound and its duration. The vocal backchannels were quantified for all identified BOPs as follows; if an addressee made a sound during a BOP, the BOP was represented by 1 for the addressee, or else by 0.

Comparisons of audiovisual behavior in BOPs vs non-BOPs

To understand whether the behavior of addressees differed between the BOPs and the rest of the conversation, we paired each BOP with a random non-BOP of the same length. A non-BOP is a moment in the conversation for which none of the judges thought it was a BOP. We compared the behavior of all addressees for a specific BOP with the behavior exhibited at the same non-BOP. Paired t-tests were carried out over all encoded channels. Pairs that contained frames that FaceReader was unable to encode were discarded. To determine how backchannel behavior differs across addressees, we calculated the average behavior per addressee, and reported the average behavior across all addressees. The Bonferroni correction was applied for the multiple pairwise comparisons.

BOP types: Continuer and End-of-turn

The differences of behavior betweeen Coninuer BOPs and EOT BOPs were quantified with the Welch's t-test, corrected with the Bonferoni method.

4.4.2 Results

Overall, the behaviors during BOPs and non-BOPs differed markedly, except that we did not find any differences regarding minute facial expressions related to the action units (see Table 4.5, and Figures 4.3, 4.4 and 4.5). Even though the standard deviations for amplitude and frequency were high, there was a significant difference between the head movement of an addressee during a BOP and a non-BOP. On average, the frequency of head movement during a BOP was 3.43 upward/downward peaks per second, being 0.68 higher than the frequency in a non-BOP. The average amplitude was 5.95 degrees, which was 1.87 higher compared to a non-BOP. Across all BOP instances, 28% of the time, vocalizations were produced, while during non-BOPs this occurred only 3% of the time. The behavior of the facial muscles was generally the same during BOPs and non-BOPs (see Table 4.5 for an overview) and contained no significant differences.

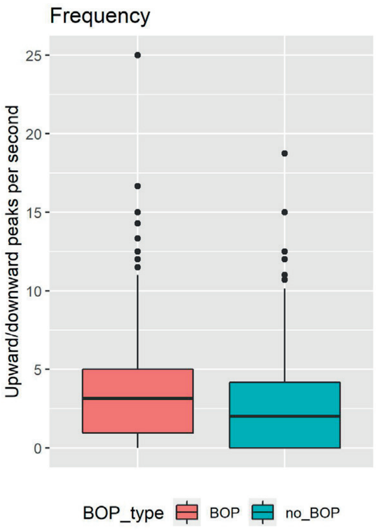


Figure 4.3 Head movement frequency during BOPs and outside of BOPs

Variation of backchannel behaviors across addressees

There was a substantial variation regarding different behaviors across addressees. Head movement differed among the addressees. Although the mean frequency of head movement was 3.46 upward/downward peaks per second across addressees, the most nodding addressee showed 5.47 upward/downward peaks per second on average, compared to 1.49 upward/downward peaks per second on average for the least nodding addressee. Amplitude was on average 5.97 degrees, with the addressee on the lowest end having an amplitude of 3.34 degrees on average, while the addressee on the highest end showed on average 9.65 degrees amplitude. Addressees vocalized 28% of the BOPs on average, while the least vocal addressee only vocalized 4% of the BOPs and the most vocal addressee vocalized 59% of all BOPs. AU activations also varied, for example, the AU with the highest variation (*SD* = 1.50) was Eyes Closed (AU43), followed by Lip Corner Puller (AU 12) (*SD* = 1.00). See table see Table 4.6 for a complete overview and Figure 4.6 for a visual inspection.

Table 4.5 Averages of different channels (and standard deviations) over BOPs and non-BOPs

	BOP (1)	non-BOP (2)	Diff	df	Cohen's d
Head Movement:					
Frequency	3.43 (3.09)	2.75 (2.76)	0.68***	705	0.23
Amplitude	5.95 (5.69)	4.07 (3.74)	1.87***	709	0.39
Vocalisations	0.28 (0.48)	0.03 (0.18)	0.25***	741	0.72
Facial Gestures:					
Inner Brow Raiser (AU1)	0.01 (0.12)	0.01 (0.08)	0.00	712	0.01
Outer Brow Raiser (AU2)	0.00 (0.03)	0.00 (0.02)	0.00	712	0.02
Brow Lowerer (AU4)	0.07 (0.26)	0.08 (0.27)	0.00	712	0.01
Upper Lid Raiser (AU5)	0.00 (0.04)	0.01 (0.06)	0.00	712	0.04
Cheek Raiser (AU6)	0.32 (0.65)	0.39 (0.75)	-0.07	712	0.09
Lid Tightener (AU7)	0.11 (0.28)	0.19 (0.28)	0.01	712	0.05
Nose Wrinkler (AU9)	0.00 (0.00)	0.00 (0.00)	0.00	712	0
Upper Lip Raiser (AU10)	0.11 (0.30)	0.08 (0.27)	0.03	712	0.09
Lip Corner Puller (AU12)	0.78 (1.00)	0.87 (1.10)	-0.09	712	0.08
Dimpler (AU14)	0.01 (0.08)	0.02 (0.14)	-0.01	712	0.11
Lip Corner Depressor (AU15)	0.01 (0.10)	0.01 (0.09)	0.01	712	0.05
Chin Raiser (AU17)	0.23 (0.55)	0.23 (0.57)	0.00	712	0.00
Lip Puckerer (AU18)	0.00 (0.04)	0.01 (0.09)	0.01	712	0.09
Lip stretcher (AU20)	0.00 (0.02)	0.01 (0.08)	-0.01	712	0.09
Lip Tightener (AU23)	0.02 (0.16)	0.02 (0.12)	0.01	712	0.04
Lip Pressor (AU24)	0.17 (0.39)	0.15 (0.38)	0.02	712	0.05
Lips part (AU25)	0.31 (0.85)	0.30 (0.88)	0.01	712	0.01
Jaw Drop (AU26)	0.05 (0.30)	0.02 (0.17)	0.02	712	0.09
Mouth Stretch (AU27)	0.00 (0.02)	0.00 (0.04)	0.00	712	0.03
Eyes Closed (AU43)	1.39 (1.50)	1.35 (1.49)	0.04	712	0.03

Statistics are based on paired t test analysis. The Diff score is the result of substracting the mean BOP value of the mean non-BOP value. * p <.05, ** p<.01, *** p<.001

Variation within addressees

The average addressee behavior also differed across the different BOPs. Figure 4.7 shows the distribution of behavior per BOP. On average, the frequency was 3.42 upward/downward peaks per second across BOPs. However, BOP 35 elicited an average frequency of 1.10 upward/downward peaks per second, while at BOP 51, addressees show an average of 6.25 upward/downward peaks per second. The amplitude also varied, the mean amplitude across all BOPs was 5.96, while the minimal average amplitude was 0.65 degrees at BOP 51, and the maximum average amplitude was 14.1 degrees at BOP 11. Some BOPs (e.g., 12, 16, 17) were never vocalized, while other BOPs were vocalized by 93% of the addressees (e.g., BOP 26). The effect of addressee-dependent behavior can also be visually inspected in Figure 4.8. For a full overview of the numbers, see Table 4.7.

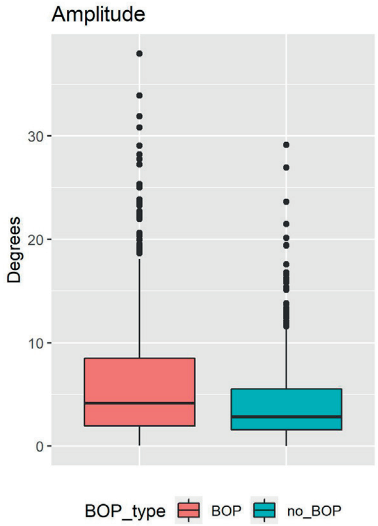


Figure 4.4 Head movement amplitude during BOPs and outside of BOPs

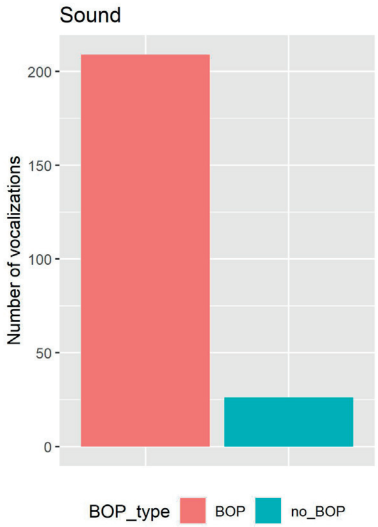


Figure 4.5 Number of times addressees vocalized BOPs versus the number of vocalized non-BOPs

Table 4.6 Differences in feedback behavior between addressees

	Mean	SD	Min	Max
Head Movement:				
Frequency	3.45	1.21	1.49	5.45
Amplitude	5.97	2.10	3.34	9.65
Vocalisations	0.28	0.14	0.04	0.59
Facial Gestures:				
Inner Brow Raiser (AU1)	0.01	0.02	0	0.06
Outer Brow Raiser (AU2)	0.00	0.00	0	0.01
Brow Lowerer (AU4)	0.07	0.21	0	0.79
Upper Lid Raiser (AU5)	0.00	0.01	0	0.04
Cheek Raiser (AU6)	0.32	0.33	0	1.14
Lid Tightener (AU7)	0.11	0.14	0	0.44
Nose Wrinkler (AU9)	0	0	0	0
Upper Lip Raiser (AU10)	0.11	0.20	0	0.65
Lip Corner Puller (AU12)	0.77	0.51	0.19	1.82
Dimpler (AU14)	0.01	0.02	0	0.05
Lip Corner Depressor (AU15)	0.01	0.02	0	0.07
Chin Raiser (AU17)	0.25	0.41	0	1.29
Lip Puckerer (AU18)	0.00	0.01	0	0.05
Lip stretcher (AU20)	0.00	0.00	0	0.01
Lip Tightener (AU23)	0.02	0.06	0	0.20
Lip Pressor (AU24)	0.17	0.25	0	0.91
Lips part (AU25)	0.32	0.47	0.02	1.57
Jaw Drop (AU26)	0.05	0.15	0	0.58
Mouth Stretch (AU27)	0.00	0.00	0	0.01
Eyes Closed (AU43)	1.39	1.40	0	4.06

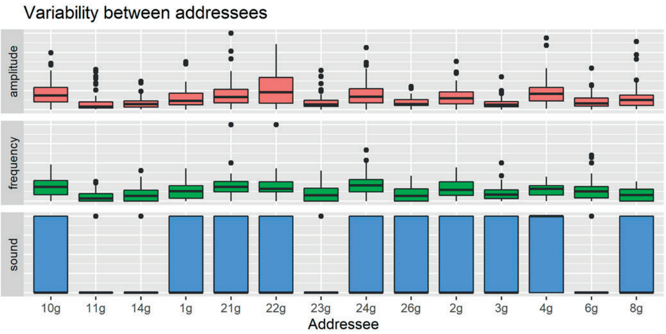


Figure 4.6 Values for head movement, vocalisations and Dimpler (AU 10) values for each addressee. Frequency, Amplitude and AU values are scaled, such that 1 represents the maximum value and 0 the lowest value.

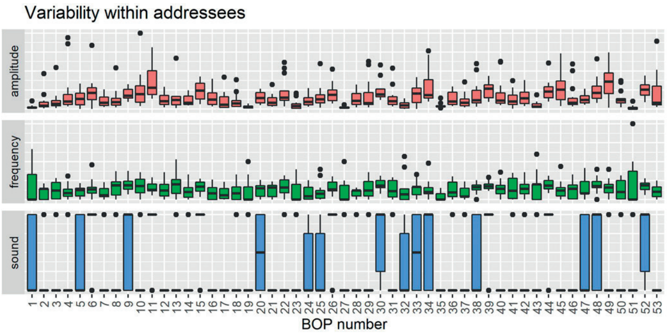


Figure 4.7 Values for head movement, vocalisations and Dimpler (AU 10) values for each BOP. Frequency, Amplitude and AU values are scaled, such that 1 represents the maximum value and 0 the lowest value.

Table 4.7 Differences in feedback behavior within addressees

	Mean	SD	Min	Max
Head Movement:				
Frequency	3.42	1.01	1.10	6.25
Amplitude	5.96	3.11	0.65	14.1
Vocalisations	0.28	0.28	0	0.929
Facial Gestures:				
Inner Brow Raiser (AU1)	0.01	0.03	0	0.19
Outer Brow Raiser (AU2)	0.00	0.01	0	0.05
Brow Lowerer (AU4)	0.07	0.04	0	0.15
Upper Lid Raiser (AU5)	0.00	0.01	0	0.07
Cheek Raiser (AU6)	0.32	0.30	0	1.56
Lid Tightener (AU7)	0.11	0.08	0	0.27
Nose Wrinkler (AU9)	0	0	0	0
Upper Lip Raiser (AU10)	0.11	0.07	0	0.31
Lip Corner Puller (AU12)	0.79	0.53	0.14	2.51
Dimpler (AU14)	0.01	0.02	0	0.11
Lip Corner Depressor (AU15)	0.01	0.02	0	0.07
Chin Raiser (AU17)	0.24	0.22	0.05	0.45
Lip Puckerer (AU18)	0.00	0.01	0	0.05
Lip stretcher (AU20)	0.00	0.01	0	0.03
Lip Tightener (AU23)	0.02	0.05	0	0.17
Lip Pressor (AU24)	0.17	0.08	0.04	0.42
Lips part (AU25)	0.32	0.34	0.00	2.03
Jaw Drop (AU26)	0.04	0.07	0	0.25
Mouth Stretch (AU27)	0.00	0.00	0	0.03
Eyes Closed (AU43)	1.38	0.185	0.85	1.73

Variation within addressees for different BOP types

The BOPs that are marked as EOT BOP elicit higher nodding amplitudes from the addressees than the Continuer BOPs; furthermore, EOTs let to more vocalisations, on average 60% of the time while during the remaining BOPs, addressees vocalized 20% of the time, on average. Nodding frequency is not different between the two types of BOPs. See for all the results Table 4.8.

Table 4.8 Averages of head movement and vocalisations over continuer BOPs and End of Turns

	EOT (1)	BOP (2)	Diff	df	Cohen's d
Frequency	3.19 (2.58)	3.06 (3.03)	0.13	509.23	0.04
Amplitude	6.78 (5.99)	4.56 (4.48)	2.22***	372.78	0.46
Vocalisations	0.60 (0.49)	0.20 (0.40)	0.40***	208.74	0.97

Statistics are based on Welch's t test analysis. The Diff score is the result of substracting the mean BOP value of the mean End Of Turn (EOT) value.

* p <.05, ** p<.01, *** p<.001

4.5 Discussion

In this study, we were interested in a computational examination of the variability in backchannel behaviors among addressees. We looked at whether and how behavior varied during backchannel opportunity points (BOPs) across and within addressees, specifically focusing on head movement, vocalizations, and facial expressions produced by fourteen addressees in a Tangram-game. The game setup used the o-cam paradigm, meaning that each addressee was exposed to exactly the same behaviors produced by the speaker. We showed that in general head movement and vocalization behavior significantly differed between BOPs and non-BOPs.

Nodding behavior and vocalizations were most pronounced during BOPs, compared to non-BOP instances. However, it is notable that the amount of facial activity was generally the same during BOPs and non-BOPs, characterized by most AUs being activated at low-intensity levels. These low-intensity levels may be a consequence of the experimental setup, namely that addressees did not exhibit higher AU intensities because of the nature of interaction that the experimental setup (o-cam paradigm) allowed. However, it is more likely that low facial activity during both BOPs and non-BOPs was the result of a general pattern, which is that during natural interactions people rarely produce exaggerated facial expressions [30].

Further dissection of behavior during BOPs showed that there was a person-specific variability. This between-addressee variability indicates that not every addressee demonstrated the same feedback behavior during BOPs. Some individuals were more discrete with their feedback behavior than others. In addition, the analysis indicated BOP-related differences. Some BOPs manifested more expressive behavior on average than others. Thus, in general, the timing of feedback behavior seems to adhere to certain rules. All addressees showed consistently different behavior during the BOPs than outside of the BOPs. However, the exact behavior seemed to be influenced by person-specific and BOP-related variables.

4.5.1 Variability between addressees

There was also variability between addressees. While all addressees nodded and vocalized during BOPs more than outside of them, there was variability in the extent to which addressees produced nodding and vocalizations during BOPs.

Interestingly, the most vocal addressee produced a sound during more than half of the BOPs, a substantial difference from the least vocal addressee, who vocalized 14 times less. Likewise, the addressee with the smallest amplitude (addressee 14, with an average amplitude of 2.9) differed substantially from the person with the most pronounced amplitude (addressee 22 with an average amplitude of 7.4).

Given that all addressees were subject to the same experimental paradigm, the most likely source of this variation in backchannel behavior was the addressee's tendencies related to personality characteristics. In other words, while most BOPs were amenable to nods and vocalizations, addressees differed in the manifestation of their listening behaviors. Prior research shows that backchannel behavior can be linked, to some extent, to the personality characteristics of a person as measured through the Big-Five traits [205]. In addition, different backchannel behaviors can systematically engender certain personality perceptions as perceived by the third party observers [56]. In the study by [56], backchannel behaviors were applied on a virtual agent, showing that, for example, higher frequency of backchannels was related to the perception of extroversion. Other factors could include gender, research showed that women tend to backchannel with a higher frequency than men, and that backchanneling occurs more frequently in Japanese than in American English [60, 136, 70].

In a future experiment, it would be valuable to take into account the characteristics of the addressee, such as personality, gender, and cultural background to identify factors that may play a role in producing the person-specific variability of feedback behavior. In addition, it would be beneficial to extend the length of the experiment to harvest more behavioral data from each addressee, which would allow to also shed light on potential intrapersonal variability, unrelated to BOP or person-specific characteristics. Although it is currently unknown what the time limits are of an o-cam experiment, we hypothesize that a longer experiment would result in more addressees that would find out that the speaker is pre-recorded.

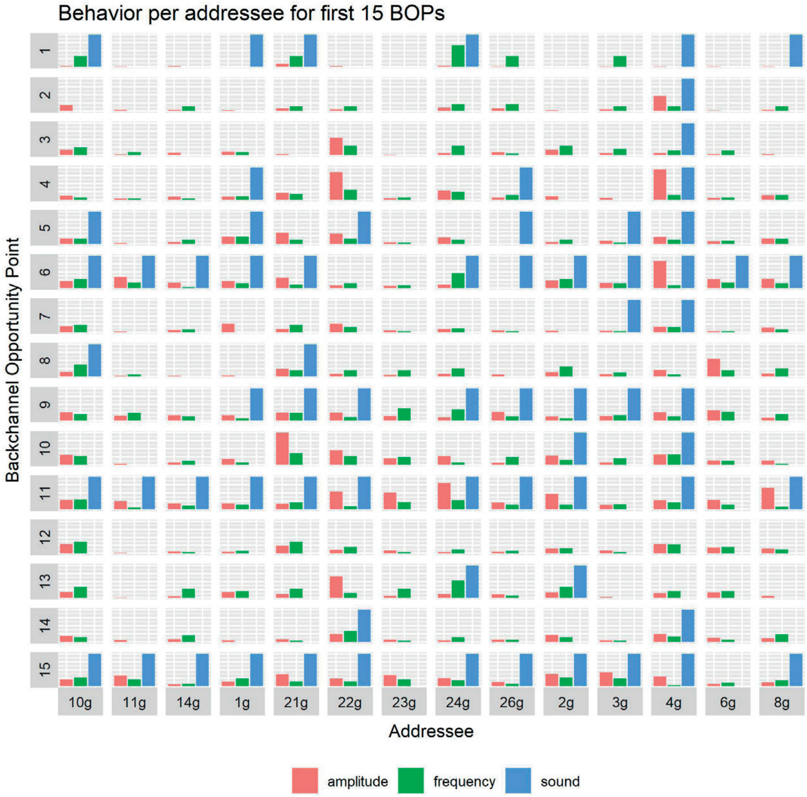


Figure 4.8 Behavior for each addressee per BOP. Only the first 15 BOPs due to visualisation restrictions.

4.5.2 Variability between BOPs

While nodding and vocalizations characterize spontaneous listening behavior, the high standard deviations regarding nodding behavior (i.e., amplitude and upward/ downward peaks per second) suggest that different BOPs lead to the differing amount of nodding. This can be seen in Figure 4.8.

Regarding the current data, differing nodding patterns based on a BOP may partially be related to the fact that some Tangrams may have been more difficult to understand than others. That is, if an addressee quickly understood the description of a figure, they may have nodded more energetically compared to those instances where they doubted and hence nodded in a less pronounced fashion. This insight is related to the early research on non-verbal behavior conducted by [26], who showed that based on both the frequency of nods and their duration the involvement of an addressee was communicated differently. In particular, a single nod of 400 ms in duration acted as a strong affirmation of the speaker's behavior while a nod of 800 ms or longer signaled disbelief and even elicited interruptions on the part of the speaker. Overall, this demonstrates that the nature of backchannels varies as the interaction unfolds. Our research also put forward a difference between behavior showed during the last BOPs of a round, and BOPs that were located during a round. The last BOP of a round may have acted as a feedback point, but also as marking the end of a round. The addressee was signaled at this BOP that the moment of choosing the correct Tangram was near, and therefore the function of the BOP was perhaps different than the other BOPs. The speaker was more 'asking' for a confirmatory signal from the addressee, than an acknowledging feedback signal. Indeed also the backchannel-inviting cues from the speaker were clearly different when signaling the last BOP of a round, compared to other BOPs. The speaker was using a downward inflection when signaling the last BOP of the round, compared to an upward inflection when signaling other BOPs, and used a lower pitch rate on average. In return, addressees were more expressive during the EOTs, in the sense that they vocalized more often and showed a higher amplitude in their nodding behavior. That backchannel-inviting cues have a lower pitch at the end of a round, and have a downward inflection is in line with [74], which shows that speakers 'reserve' the low pith to mark the end of a turn, while keeping using a higher pitch in other cases to prevent that the turn is taken over by the opponent.

Given the variability in audiovisual behavior between various BOPs, we looked at a few cases in more detail to gain insight into possible reasons for the differences. In particular, we did some speculative analyses of BOP 26, which was vocalized by 92% of the addressees and received relatively frequent head nods (4.36), versus BOP 16 which was not vocalized by any of the addressees and not frequently marked by head nods (2.21). Comparing these two instances yields the impression that the strength of the feedback cue (in terms of nodding and auditory backchanneling) is related to the degree to which the speaker signals that the information she provided is complete. BOP 26 occurs at the end of round 5, just after the speaker said "That's the one you have to pick. So, a square chimney and a triangle from the side of the house". During this BOP of 1000 ms the speaker is completely silent. The speaker appears to cue that she provided all the information the addressee needs to pick the correct Tangram figure, and therefore expects a strong affirmative backchannel. BOP 16, on the contrary, occurs at the beginning of round 4, just at the end of short sentence from the speaker "These are more like birds..", where it is clear that more details from the speaker are needed to be able to identify the Tangram she is describing. At this stage, a strong feedback cue from the addressee would seems less appropriate, given that the provided information is still incomplete, but an addressee may acknowledge that s/he is listening to the speaker and awaiting further details. Obviously, future work is needed whether these impressions would generalise to more conversational contexts.

4.5.3 Division of labour

Given the results described above, it is interesting to compare the audiovisual behaviour of the speaker with that of the addressee. Admittedly, given that we only recorded one speaker, our claims related to her role would have to be explored further in future work, but based on our analyses so far, it appears that our speaker more consistently makes use of auditory than visual cues to elicit feedback from her addressees. Indeed, while we find some prosodic differences between BOPs and non-BOPs, there are no significant differences in facial activity. Conversely, the addressees appear to exploit visual cues more regularly than vocalisations to return feedback after BOPs. In other words, given the broader set of audiovisual cues that function within an interaction, these results suggest that a speaker is more often using auditory features, and the listener is more often making use of silent, visual cues, except for BOPs that occur at the final edge of a turn where a speaker is basically signalling that she has arrived at the end of her turn and will this stop talking.

While this would have to be explored further in the future, these results point to a division of labour between auditory and visual cues in the feedback mechanism of a conversation, with the former being more typical for the speaker and the latter for the addressee. The advantage of being able to access multiple channels, is that their use can be distributed over conversation partners so that they can exchange information in parallel. For instance, while one person is talking, the other can return visual feedback, such as affirmative head nods or expressions of surprise or misunderstanding, that do not interfere with the speech produced by the other as these are produced in silence. If instead dialogue partners were to produce speech simultaneously, miscommunication might well result from the overlapping sound streams, because the speech by one person might mask that of the other [194].

4.5.4 Embodied Conversational Agents

Understanding variation in backchannel behaviors across addressees is important for applications in embodied conversational agents. If for a large portion of backchannels nodding and vocalizations can be produced to show that one is engaged and listening, future research could investigate the conditions under which these behaviors are necessarily produced and vice versa the conditions when there is a slim chance that either a nod or a vocalization will occur. Understanding this balance between variability and stability of backchannel behaviors across a human-human conversation can help make artificial agents that can give flexible feedback and that come across natural in human-computer conversations. Moreover, person-specific variability may be used in an ECA to augment gender, personality, and cultural characteristics. In other research, we have shown that indeed specific backchannel behavior in an ECA can elicit specific personality perceptions by its audience. We copy-synthesized the feedback behavior of different addressees during various BOPs onto an ECA and asked participants to indicate the perceived personality characteristics of the ECA. Among other conclusions, we found that a higher nodding amplitude during a BOP is perceived as more extroverted than a smaller nodding amplitude.

Previous studies show that when listening behaviors are missing or are poorly timed, the communication is negatively affected and can go off the rails [16]. The current findings suggest that there is no 'one listening behavior', but a variety of behaviors across different BOPs and across different addressees. And although nods and vocalizations are characteristic of spontaneous interactions, the degree to which they will be produced varies between addressees.

Acknowledgments

We thank Melanie Brugel for her help with recording the audiovisual materials, Lieke van Maastricht for having served as a speaker for the experiment, Alex Blomsma for the visual impression of the experiment and the various people from our social network who participated as judges to obtain the BOPs. Julija Vaitonyté is a member of the VIBE project.

CHAPTER 5

5

Backchannel behavior influences personality perception

Abstract

Different applications or contexts may require different settings for a conversational AI system, as it is clear that e.g. a child-oriented system would need a different interaction style than a warning system used in emergency situations. The current article focuses on the extent to which a system's usability may benefit from variation in the personality it displays. To this end, we investigate whether variation in personality is signaled by differences in specific audiovisual feedback behavior, with a specific focus on embodied conversational agents. This article reports about two rating experiments in which participants judged the personalities (i) of human beings and (ii) of embodied conversational agents, where we were specifically interested in the role of variability in audiovisual cues. Our results show that personality perceptions of both humans and artificial communication partners are indeed influenced by the type of feedback behavior used. This knowledge could inform developers of conversational AI on how to also include personality in their feedback behavior generation algorithms, which could enhance the perceived personality and in turn generate a stronger sense of presence for the human interlocutor [6].

[6] This chapter is based on: Blomsma, P., Skantze, G., & Swerts, M. (2022). Backchannel Behavior Influences the Perceived Personality of Human and Artificial Communication Partners. Frontiers in Artificial Intelligence, 5.

5.1 Introduction

5.1.1 Personality perception

Personality refers to the consistent behavioral responses of a person and is often expressed in terms of the Big Five theory [106]. There is a growing scientific interest in rendering conversational AI systems with various types of personality as this may help to make interactions with such systems more natural, and would allow to tune their interaction style to different situations or users. Our current paper will tackle this issue in view of the further development of so-called Embodied Conversational Agents (ECAs), i.e., computer interfaces that are graphically represented as a human body or human face, in order to allow users to interact face-to-face with computers in a way that resembles that of their interactions with real humans [146, 41].

Technically speaking, ECAs are nothing more than a collection of algorithms that together orchestrate the interaction with the interlocutor. To create the illusion that the interlocutor is not conversing with just some mindless algorithms, but with a partner who has thoughts and emotions, there have been attempts to render such conversational AI systems with a specific personality. This may enhance a feeling of social presence, and therefore increase the experience of dealing with a system that truly understands the intentions and feelings of the user [122, 153, 25]. Furthermore, it may be useful if conversational AI systems adapt their personality and conversational style to the specific application or intended audience. For instance, a conversational AI implemented for a playful environment would typically demand a different interaction style than one which is put to use in a crisis or emergency context [78]. Likewise, a conversational AI may have to adjust its behaviour depending on whether it addresses a child or an adult, or a person with specific communicative deficiencies [213]. Adaptation of its personality to the personality of the interlocutor could also increase conversational quality by exploiting the mechanics behind similarity-attraction, as indeed people have been shown that people feel more attraction towards people or systems that match their personality [122]. Therefore, personality adaptation has the potential to result in conversations that are more engaging and enjoyable [104]. A study conducted by [197] demonstrated that therapy sessions were more engaging when a robot's conversational behavior was matched to the personality traits of the patient, as compared to a robot that had mismatched its conversational behavior. Additionally, personality adaptation could lead to a more favorable perception of the system. [7] found that matching a robot's gaze behavior with participants resulted in a more positive attitude towards the robot, as compared to mismatched gaze behavior.

While personality potentially may help facilitating social presence and conversational quality, it would also seem to be a requirement to create next-level conversational AI systems in yet another sense. One of the factors that prevents the creation of human-like systems whose appearance is perceived as being similar to that of real humans is related to the uncanny valley effect, the phenomenon that small errors in behavior generation of the system can evoke feelings of fright discomfort in the interlocutor. While various theories exist regarding the cause of this effect, some explain it by cognitive dissonance [216], i.e., the discomfort that arises because it is unclear to a user or observer if the conversational system should be perceived as human- or system-like. A conversational AI system that lacks a personality, and may therefore generate inconsistent behavior, could increase the feelings of unease on the part of the user, as he or she may feel uncertain on how to deal with a conversation partner who displays deviant interactive behaviour [220].

5.1.2 Variability in feedback behavior

The current paper focuses on variability in feedback. In particular, we look at backchannel behavior, which refers to the feedback dialogue partners give each other on the smoothness of the information exchange process [50]. While one person is talking, the addressee typically returns brief responses, called backchannels, which can be auditory (e.g., “uhuh”) or visual (e.g., head nod) in nature [62]. Backchannels serve as cues to signal how the information was received at the other end of the communication channel, where one could roughly make a distinction between “go-on” or “do-not-go-on” signals. Although feedback behavior is person-dependent, backchannels are more expected at certain points in the conversation, namely during backchannel opportunity points (BOPs) [81], moments when its appropriate to give feedback [217]. BOPs are signaled by the speaker with a so-called backchannel-inviting cue [84], signals via the prosodic channel, such as rising and falling intonations [83], low pitch ranges [209] and short pauses [45]. However, addressees may vary regarding the extent to which they react to such speaker-initiated cues and utilize BOPs.

Behavior during those BOPs differs significantly from behavior outside of the BOPS. Specifically, speakers’ nodding behavior, vocalisations and the use of the upper lip raiser (AU 10) is more apparent during BOPs than outside of those moments. During BOPs, people diverge considerably in timing, frequency and type of audiovisual feedback behaviour. There are both differences between how people react to the same BOP (within-people diversity) and how the same person react to different BOPs (within-person diversity) [27]. Also people can diverge considerably in timing, frequency and type of audiovisual feedback behaviour. It is intuitively clear that, likewise, different conversational AI systems may have to vary regarding the degree, the type and the frequency of backchanneling. For instance, an “emphatic” tutoring system that has to assist learners to acquire a specific new skill may have to produce more supporting cues than a more neutral system that is consulted to give legal advice or specific route directions.

The current article therefore focuses on whether personality perception is influenced by the aforementioned variability in feedback behavior a person gives. This question is inspired by the outcome of our previous study, which led to the impression, though not explicitly tested in that earlier study, that differences in interaction style generated variable perceptions of the personality of the people whose feedback behaviour was being recorded. While some participants appeared to come over as introvert and somewhat uninterested, others gave the impression of being extravert and lively, suggesting that there may exist a relation between the type of feedback behavior and the perceived personality of a person.

There are reasons to believe that this variability in behavior would lead to different personality perceptions. Although existing research into this question is scant, there are some studies that point into an affirmative direction. For example, [99] analyzed the personalities of backchannel coders and the relation between that personality and the number of identified backchannel opportunity points. They found that a high number of identified backchannel opportunity points are related to high values for Agreeableness, Conscientiousness and Openness. [103] showed that extraverted people have a higher tendency to use multimodal backchannels (e.g. a combination of utterance and nod), compared to introverted people who tend to rely on unimodal backchannels and show as well that this difference was perceived by human interlocutors when this behavior was re-enacted by a digital human-like robot. [23] found that extraverted persons produce more backchannels than introverted persons. Albeit that feedback behavior is person-dependent and related to personality, we still lack knowledge on the perceptual impact of those person-dependent behaviors (e.g. what the perceived difference is between a passive vs. a dynamic addressee). If indeed the perception of personality is related to the type of feedback behavior a person produces, then this would give developers another opportunity to perpetuate the personality of a conversational AI system. Moreover, if feedback behavior is chosen randomly, such behaviors might conflict with the desired perceived personality of such conversational AI system.

5.1.3 Current study

In this study, we thus would like to gain insight into the relation between perceived personality and feedback behavior. We will utilize participant recordings from a previously conducted o-cam based experiment [34, 79], where each participant was made to believe to have a real-life conversation with another person, but who in reality is a pre-recorded speaker. In our first experiment, parts of these participant recordings are shown to observers who are asked to rate perceived character traits of the participant in the recording.

Those ratings are analyzed in relation to the listening behavior of those participants (i.e. the auditory and visual backchannels during the recordings), to establish whether perceived character traits correlate with patterns in audiovisual backchannel behavior. In a second experiment, the same stimuli are re-enacted by a conversational AI (i.e. on a virtual Furhat robot), to verify whether the assessment of the personality of these synthetic characters is likewise affected by their feedback behavior.

5.2 Materials

For our study we utilized the dataset that was (partly) generated in a previous study described in [27]. The materials consisted of video recordings of 14 participants (henceforth called original addressees) of an o-cam experiment and the identified backchannel opportunity points (BOPs) of those recordings. BOPs are moments during the conversation that allow for listener feedback from the original addressee [81]. For every BOP, the following behavior was encoded: (1) vocalisations: did an original addressee vocalize during the BOP or not, (2) the nodding behavior of the original addressee, quantified by amplitude (the maximum head movement angle in head-pose elevation direction) and frequency (number of upward and downward peaks per timeframe) and (3) the average contraction of a facial muscle called the upper lip raiser, as defined by the Facial Action Coding System as Action Unit (AU) 10 [63]. Our previous study, found that those four variables (amplitude, frequency, vocalisations and AU10) were the more important ones for original addressees during BOPs, as compared to the rest of the interaction. The videos were recorded during an experiment in which a participant plays a game with ostensibly another participant (a confederate) via a video connection. However, in reality there is no live video connection and the original addressee plays the game with a pre-recorded video of the confederate ('the speaker'). The illusion of a real connection, which typifies o-cam experiments, is facilitated by the use of specific techniques, see e.g. [79]. Contrary to experiments that involve physical presence of confederates, the o-cam paradigm allows for a tightly controlled environment where each participant is subjected to exactly the same speaker-stimulus, while having a highly ecologically valid setting. This particular o-cam experiment was executed by [34] and aimed at eliciting listening behavior from the original addressees. Each original addressee played a Tangram game with the speaker, which consisted of 11 rounds. In each round, the original addressee was first shown 4 different abstract pictures (tangram figures) for 5 seconds, subsequently, the speaker gave a description of one of those four tangram figures, and after that the original addressee had to choose which of the four shown tangram figures was described by the speaker. Each round contained 4 new tangram figures. See Figure 4.1 for an illustration of the Tangram game and experiment.

The experiment resulted in the video recordings of 14 original addressees, who each interacted with exactly the same speaker-stimulus. Each video was 8 minutes and 42 seconds long, and contained 6 minutes and 15 seconds of interaction. All videos were analysed with Facereader [159] to annotate the videos for a number of action units and head position. Head movements (nods) were then derived from the head pitch over time and quantified in terms of amplitude (maximum distance between y coordinates of the head position) and frequency (number of peaks and valleys of head position per timeframe). Sound was annotated by 1 coder in a binary fashion, 1 for presence of sound, 0 for being silent. The BOPs were identified by a panel of 10 judges. Each judge separately indicated the points in the speaker-stimulus that s/he thought were a BOP. With help of the parasocial consensus sampling method [100], all judgements were aggregated. All BOPs that were indicated by at least 3 judges were marked as genuine BOPs, which resulted in 53 different BOPs. See [27] for a detailed explanation of the data annotation process. The following two experiments make (indirect) use of the materials collected in that earlier dataset. Experiment 1 explores to what extent the perceived personality of the recorded human participants is determined by variation in their feedback behaviour as compared to the personality perceived by appearance only. Experiment 2 tests to what extent findings from the first experiment generalizes to the perception of artificial avatars, whose feedback behaviour was modelled based on the outcome of experiment 1.

5.3 Experiment 1: perceived personality of real humans

5.3.1 Method

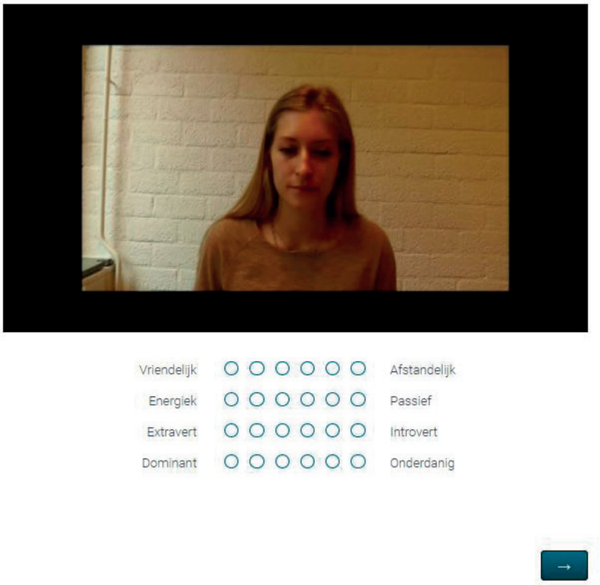
Participants

Eighty-two students from Tilburg University were recruited from the Tilburg University subject pool to participate in the first experiment in exchange for course credits. Seven students did not finish the experiment for unknown reasons and were discarded. The remaining seventy-five students did complete the experiment (11 male, 63 female and 1 other, Age: mean 21.31, SD = 3.17). The Research Ethics and Data Management Committee of the Tilburg School of Humanities and Digital Sciences approved the experiment under identification code REDC#2021/33. All participants provided their consent before they participated in the experiment.

Stimuli

The stimuli consisted of 14 pictures, one of each original addressee, and 42 video clips of 8 seconds length. Pictures were included to get the baseline personality indication from participants (called *preconception score*), as personality impressions are likely to be based on mere appearance [155]. In order to get personality perception scores for all original addressees, while keeping duration of the experiment within reasonable limits, we choose to include the behavior of all original addressees while including only three specific BOPs.

The pictures were created by exporting the first neutral frame from each original addressee recording. Neutral here means that original addressee did not have any facial muscle contractions, in other words, all annotated AUs had a value of 0 for the exported frame. For one of the original addressees (1g) such neutral frame was not available in the dataset, as AU43 ('eyes closed') was annotated as contracted for many of the frames. The selected picture for this original addressee was the first neutral frame, with ignorance of AU43. See Figure 5.1 for a representative example of such neutral picture.



Vriendelijk ○ ○ ○ ○ ○ ○ Afstandelijk

Energiek ○ ○ ○ ○ ○ ○ Passief

Extravert ○ ○ ○ ○ ○ ○ Introvert

Dominant ○ ○ ○ ○ ○ ○ Onderdanig

Figure 5.1 Impression of a neutral picture question as part of experiment 1

The video clips were extracted as follows: for each of the 14 original addressee recordings, three stimulus-videos of 8 seconds length were cut out, such that each video included a specific BOP. The stimulus-videos were cut such that the middle frame corresponded to the middle frame of the BOP. This resulted in 42 stimulus videos, each containing the behavior of one of the 14 original addressees for one of the three specific BOPs. See

Figure 5.2 for an impression of a stimulus-video. The three selected BOPs were chosen from a pool of 53 BOPs. The decision for those three BOPs was guided by the desires that (1) each BOP would have a different characteristic and (2) that each original addressee would exhibit different behaviors among themselves during the selected BOPs. After manually trying out different combinations, the following BOPs were selected: 16, 47, and 49.

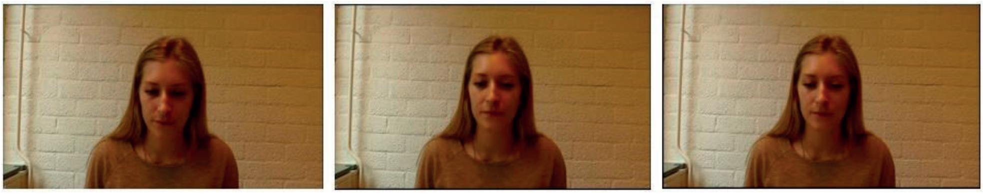


Figure 5.2 A few still images from an original addressee during BOP 49

BOP 16 takes place at the start of round 4, just after a short sentence of the speaker “These are more like birds.." and has a duration of 1160 ms. The speaker has not shared much information at this point, and it is clear that she will need to share more information to enable the listener to identify the correct Tangram figure. Therefore, it may be inappropriate for the original addressee to provide a strong (vocalized) feedback signal as more information is coming and simply acknowledging to the speaker that one is listening seems to suffice. Interestingly, none of the original addressees vocalized during BOP 16. BOP 47 occurs near the end of round 10, just after the speaker utters “The person looks to the left and the arms also point to the left." and before the speakers says “I think you can figure it out by now.". BOP 47 is 920 ms long. At this feedback point it should be clear that the speaker shared all information that is needed to determine the correct Tangram figure, and thus a stronger feedback signal from the original addressee could be appropriate to signal that all information is understood. Six original addressees gave vocalized feedback signal at this BOP. BOP 49 is approximately halfway round 11 (the last round) and has a length of 960 ms. Just after the speaker explained “You must have the one with the very lowest passage." and before the speaker said “Thus, it’s four buildings, all four with a sort of passage through in the middle ...". At this point in time the speaker explained the main hint that’s needed for choosing the correct Tangram figure, however, the explanation is a bit ambiguous, hence the explanation that follows BOP 49. For an overview of the behavior of all original addressees during these three BOPs, see Figure 5.3.

Procedure

Participants took part in an online experiment using the environment of Qualtrics (Qualtrics, 2021). Before the start of the experiment, participants read the instructions, signed the consent form and familiarized themselves with the task with one practice picture and one practice video. The practice stimuli were taken from the stimuli used in [202] and were not in any way related to the stimuli used in the rest of the experiment. Successively, the stimuli (pictures and videos) were presented to the participants. First, the 14 pictures were presented in random order, followed by the presentation of the 42 video clips in random order as well. For each stimulus, participants were asked to rate the perceived personality of the original addressee in that stimulus for four dimensions on 6-point bipolar Likert scales: Friendliness (1:Friendly 6:Distant), Activeness (1:Active - 6:Passive), Extraversion (1:Extravert 6:Introvert) and Dominance (1:Dominant - 6:Submissive). See Figure 5.1 for an example of a presented stimulus.

The experiment concluded with two general questions: the participant was asked to fill in their age (open field) and indicate their gender (options: Male, Female, Other, Don't want to say). On average it took 42 minutes and 39 seconds to complete the experiment (SD = 151 minutes and 42 seconds).

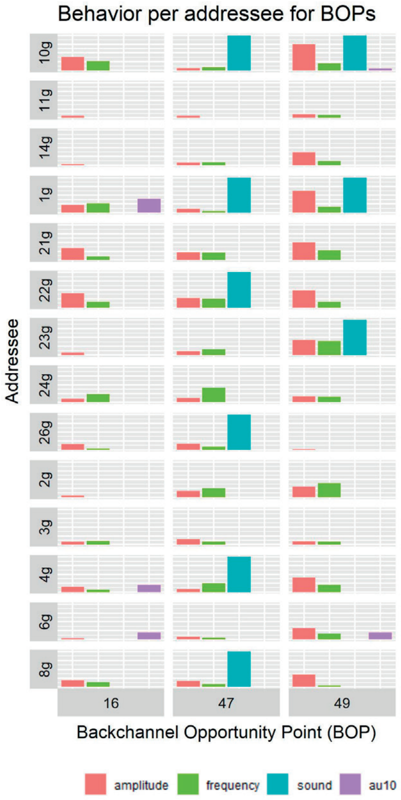


Figure 5.3 Quantification of the feedback behavior of the original addressees during the BOP 16, 47 and 49.

Statistical analyses

Statistical analyses were conducted in R Studio (version 1.1.456; [174]). Linear mixed-effects models were used to fit each of the four personality dimensions with the lme4 package [14]. The participant's response for each dimension served as the dependent variable. The main goal was to determine if behavior related to feedback contributed to the perception of personalities of original addressees or not. Therefore we analyzed both the perceived personality of the original addressee in the static picture (called the preconception score), and the contribution of the audiovisual behavior during the BOP (sound, head movement (frequency and amplitude) and AU10). The fixed effects that entered the model were preconception score (6 levels: 1 - 6), sound (2 levels: sound, no-sound), frequency (number of nods per frame, value between 0 and 1), amplitude (maximum amplitude of nod per BOP in degrees, values between 0 and 28) and AU10 (mean contraction of AU10 during BOP, values between 0 and 5). Participants, BOPs and original addressees were treated as random effects, with random intercepts, in all models. Degrees of freedom and Satterthwaite approximation for p-values for all main effects were obtained from the lmerTest package [120]. For every dimension we fitted two models, one with *preconception score* to see to what extent the static picture effects perception, and one without *preconception score* to see to what extent the more dynamic audiovisual features can explain the perceptual results on their own. Table 5.2 contains the descriptive statistics of the perception scores per dimension, Figure 5.4 shows the perception scores per dimension and original addressee, and Table 5.1 contains the estimates for all fixed effects.

5.3.2 Results

Friendliness

The average score given for the videos for the friendly-distant dimension was 3.05 (SD = 1.39), the videos of original addressee 4g were perceived as most friendly (mean = 1.91, SD = 1.02), while the videos of original addressee 11g was perceived as most distant (mean = 4.25, sd = 1.27). The *preconception score*, i.e. the score of the pictures of the original addressees, were rated on average with a 3.25 score (SD = 1.28). Similar to the results for the videos, the picture of original addressee 4g was perceived as most friendly (mean = 2.27, sd = 0.90), while original addressee 21g was perceived as most distant (mean = 4.23, SD = 1.10). The model with *preconception score* showed that *preconception score* had a significant effect on this dimension (*b*=0.220, *SE*=0.018, *df*=3129.000, *t*=11.971, *p* <0.001). However, also *amplitude* (*b*=-0.027, *SE*=0.006, *df*=871.700, *t*=-4.657, *p* <0.001), *sound* (*b*=-0.243, *SE*=0.070, *df*=2554.000, *t*=-3.472, *p* <0.001) and *au10* (*b*=-0.246, *SE*=0.073, *df*=2859.000, *t*=-3.383, *p* <0.001) had a significant effect on the perception of friendliness. The model without *preconception score* showed significant effects for the same variables: *amplitude* (*b*=-0.028, *SE*=0.006, *df*=816.900, *t*=-4.547, *p* <0.001), *sound* (*b*=-0.244, *SE*=0.072, *df*=2520.000, *t*=-3.412, *p* <0.001) and *au10* (*b*=-0.246, *SE*=0.074, *df*=2857.000, *t*=-3.322, *p* <0.001). From this we can conclude that, although appearance as tested with the *preconception score* played a significant role, that the amplitude, sound and AU10 during a BOP influence the perception of the friendly-distant dimension, such that a higher amplitude, usage of sound and more contraction of AU 10 correlates with a higher friendliness score. See also Figure 5.5 for a visual representation.

Activeness

On average, participants rated the videos 3.61 (SD = 1.42) on the active-passive dimension. Like with the friendliness dimension, original addressee 4g was perceived most active (mean = 2.35, SD = 1.51), while original addressee 11g was rated as most passive (mean = 4.93, SD = 1.05). The mean *preconception score* was 3.63 (SD = 1.27), where the picture of original addressee 4g was rated as most active (mean 2.41, score = 0.94) and original addressee 26g as most passive (mean = 4.44, SD = 1.03). The model that included *preconception score* produced significant effects for *preconception score* (*b*=0.235, *SE*=0.018, *df*=23.350, *t*=12.726, *p* <0.001), thus appearance had a significant influence on the score. Next to that, *amplitude* (*b*=-0.027, *SE*=0.006, *df*=115.200, *t*=-4.874, *p* <0.001), *frequency* (*b*=-0.821, *SE*=0.273, *df*=1589.000, *t*=-3.001, *p* <0.001), *sound* (*b*=-0.311, *SE*=0.068, *df*=988.800, *t*=-4.542, *p* <0.001) had significant effects. When *preconception score* was ignored, the model also resulted in significant effects for *amplitude* (*b*=-0.027, *SE*=0.006, *df*=105.400, *t*=-4.762, *p* <0.001), *frequency* (*b*=-0.821, *SE*=0.280, *df*=1529.000, *t*=-2.928, *p* <0.001), *sound* (*b*=-0.312, *SE*=0.070, *df*=937.600, *t*=-4.448, *p* <0.001). So nodding behavior (amplitude and frequency) and sound both correlate with the perception of activeness. Frequent nodding, a higher amplitude and vocalisations during BOPs result in a higher activeness score.

Table 5.1 Overview of fixed effects for all 12 models.

	Human (all)				Human (without preconception score)				Avatar			
Friendly Distant	*b*	*SE*	*df*	*t*	*b*	*SE*	*df*	*t*	*b*	*SE*	*df*	*t*
Intcpt	2.627	0.225	18.438	11.650***	3.341	0.230	16.676	14.501***	3.986	0.212	6.374	18.777***
Pre. s.	0.220	0.018	3128.955	11.971***	-	-	-	-	-	-	-	-
Amp.	-0.028	0.006	871.686	-4.657***	-0.028	0.006	816.973	-4.547***	-0.033	0.007	1654.109	-5.064***
Freq.	0.144	0.279	2797.100	0.515	0.145	0.285	2786.093	0.509	-0.417	0.308	2506.289	-1.354
Sound	-0.243	0.070	2554.217	-3.472***	-0.244	0.072	2519.544	-3.412***	-0.100	0.078	2677.430	-1.290
AU10	-0.492	0.145	2858.921	-3.383***	-0.493	0.148	2857.302	-3.322***	-0.206	0.080	2434.854	-2.578*
Active Passive	*b*	*SE*	*df*	*t*	*b*	*SE*	*df*	*t*	*b*	*SE*	*df*	*t*
Intcpt	3.176	0.198	23.348	16.058***	4.029	0.213	17.664	18.919***	4.389	0.174	8.275	25.295***
Pre. s.	0.235	0.018	3138.523	12.726***	-	-	-	-	-	-	-	-
Amp.	-0.027	0.006	115.228	-4.874***	-0.027	0.006	105.419	-4.762***	-0.037	0.006	1095.954	-6.031***
Freq.	-0.821	0.274	1589.229	-3.001**	-0.821	0.280	1528.568	-2.928**	0.042	0.286	2237.498	0.147
Sound	-0.311	0.068	988.773	-4.542***	-0.311	0.070	937.573	-4.448***	-0.404	0.072	2394.651	-5.593***
AU10	-0.207	0.143	1828.159	-1.452	-0.207	0.146	1776.930	-1.415	-0.120	0.074	2160.367	-1.608
Extravert Introvert	*b*	*SE*	*df*	*t*	*b*	*SE*	*df*	*t*	*b*	*SE*	*df*	*t*
Intcpt	3.057	0.164	25.761	18.680***	4.056	0.190	17.038	21.370***	4.315	0.136	7.991	31.687***
Pre. s.	0.273	0.017	3131.610	15.943***	-	-	-	-	-	-	-	-
Amp.	-0.020	0.005	3131.610	-3.747***	-0.019	0.005	48.743	-3.608***	-0.030	0.006	530.140	-5.410***
Freq.	-0.391	0.259	1055.201	-1.510	-0.387	0.269	961.971	-1.437	0.228	0.264	1011.031	0.863
Sound	-0.264	0.065	568.042	-4.087***	-0.266	0.067	512.057	-3.961***	-0.495	0.067	1355.399	-7.381***
AU10	-0.040	0.135	1258.826	-0.295	-0.042	0.140	1167.998	-0.301	-0.115	0.069	929.692	-1.684
Dominant Submissive	*b*	*SE*	*df*	*t*	*b*	*SE*	*df*	*t*	*b*	*SE*	*df*	*t*
Intcpt	2.968	0.115	49.393	25.836***	4.020	0.138	18.585	29.218***	3.990	0.125	11.195	31.945***
Pre. s.	0.293	0.017	2975.232	17.188***	-	-	-	-	-	-	-	-
Amp.	-0.009	0.004	2586.582	-2.272*	-0.009	0.004	2988.443	-2.201*	-0.013	0.006	290.697	-2.273*
Freq.	-0.260	0.234	2267.288	-1.112	-0.241	0.247	2898.954	-0.977	0.508	0.272	693.490	1.870
Sound	-0.101	0.057	2233.669	-1.780	-0.103	0.060	2887.389	-1.706	-0.389	0.069	906.322	-5.630***
AU10	-0.111	0.122	2049.164	-0.905	-0.114	0.129	2825.401	-0.880	-0.004	0.070	638.318	-0.055

The number of asterisks indicates p-level: *** <0.001, **<0.01, * <0.05.

Table 5.2 Average scores per dimension for experiment 1 and 2

	Experiment 1				Experiment 2	
	Preconception		Video		Video	
Dimension:	Mean	SD	Mean	SD	Mean	SD
Friendly - Distant	3.25	1.28	3.05	1.39	3.61	1.45
Active - Passive	3.63	1.27	3.61	1.42	3.99	1.37
Extroversion - Introversion	3.66	1.32	3.78	1.32	3.98	1.29
Dominant - Submissive	3.59	1.25	3.89	1.20	3.87	1.31

Extroversion

The videos were rated on average 3.78 (SD = 1.32) for the extroversion-introversion dimension. The videos of original addressee 4g were rated as most extrovert (score = 2.35, SD = 1.15), while the videos of original addressee 14g were rated as most introvert (mean = 4.85, SD = 1.03). The *preconception score* was 3.66 (SD = 1.32) on average, where original addressee 4g was perceived as most extrovert (mean = 2.84, SD = 1.21) and 11g as most introvert (mean = 5.11, SD = 0.90). The model extroversion-introversion score that included the *preconception score* produced significant results for *preconception score* (*b*=0.273, *SE*=0.017, *df*=3132.000, *t*=18.680, *p* <0.001), but also for the behavior related variables: *amplitude* (*b*=-0.020, *SE*=0.005, *df*=56.900, *t*=-3.747, *p* <0.001), *sound* (*b*=-0.264, *SE*=0.065, *df*=568.000, *t*=-4.087, *p* <0.001). The model for the extroversion-introversion score (intercept: 4.056, SE: 0.190) without preconception score produced also significant effects for *amplitude* (*b*=-0.019, *SE*=0.005, *df*=48.740, *t*=-3.608, *p* <0.001) and *sound* (*b*=-0.266, *SE*=0.067, *df*=512.100, *t*=-3.961, *p* <0.001).

Thus amplitude and sound influence, next to the appearance of the person, the extroversion - introversion score. A higher amplitude, and the presence of sound correlate with a higher score for extroversion.

Dominance

The score for the videos was 3.89 (SD = 1.20) for the dominant-submissive dimension. Original addressee 21g was perceived as most dominant (mean = 3.10, SD = 1.06), as based on the videos. Original addressee 11g was perceived as most submissive (mean = 4.88, SD = 1.09). Preconception score was on average 3.59 (SD = 1.25), with original addressee 21g being most dominant (mean = 2.37, SD = 1.11) and original addressee 11g most submissive (mean = 4.79, SD = 0.87). The model with *preconception score* showed significant effects for *preconception score* (*b*=0.932, *SE*=0.017, *df*=2975.000, *t*=17.188, *p* <0.001) and *amplitude* (*b*=-0.009, *SE*=0.004, *df*=2587.000, *t*=-2.272, *p* <0.05). The model without *preconception score* showed also significant results for *amplitude* (*b*=-0.008, *SE*=0.004, *df*=2988.000, *t*=-2.201, *p* <0.05). While appearance has a significant correlation, amplitude also correlates with dominance: higher amplitude correlates with a higher perceived dominance.

5.3.3 Discussion

Our first experiment thus brought to light that the feedback behaviours significantly influenced the perceived personality of recorded participants, albeit that the relative importance of the variables we entered in our model varied as a function of the personality dimension we explored. Importantly, we showed that a person's personality is not merely based on the first impression we get from a still image, e.g. whether an individual shown in a picture at first sight looks friendly or dominant, but that this perception is modulated by more dynamic auditory of visual cues of that person. Our next experiment tests whether the findings based on analyses of real humans can be reproduced with avatar stimuli, in which various feedback behaviours are implemented.

5.4 Experiment 2: perceived personality of avatars

5.4.1 Method

Participants

Eighty-four students from Tilburg University were recruited from the Tilburg University subject pool to participate in the second experiment in exchange for course credits. Ten students did not complete the experiment for unknown reasons. Seventy-four students completed the experiment (20 male, 55 female, Age: mean 21.12, SD = 2.25). None of those participants had participated in experiment 1. The experiment was approved by the Research Ethics and Data Management Committee of the Tilburg School of Humanities and Digital Sciences under the same identification code as experiment 1 (REDC#2021/33). All participants gave their consent before participation.

Stimuli

The avatar experiment contained 42 videos, and contrary to the human experiment, did not include any still pictures. This was done because all videos in this experiment contained the same avatar, thus no preconception score was required. The content of the stimulus-videos of this experiment was exactly the same as those of the first experiment, except that the behavior of the original addressee in the original movie is now acted out by an avatar. The audio was copied from the original recordings. The avatar videos were created with the Furhat SDK [5], which provides a virtual simulation of the physical Furhat robot. First, the facial behavior of the original addressee was transferred onto the virtual Furhat robot. This was done by playing the original videos on a computer screen while having them analyzed on an IPhone with Live Link Face app (version 1.1.1). Live Link Face analyzes 62 different properties of facial behavior (including head movement) on a 60 frames per second basis. The output of Live Link Face was then converted with the Furhat Gesture Capture app (version 4.3.6) and played out on the virtual Furhat using Furhat SDK 2.0.0[7]. The Furhat SDK offered a collection of 10 different avatar-faces, so called textures. All sequences were played out on the default texture. Having the same face for all sequences had the advantage that texture-specific effects did not have to be taken into account. The default texture was, in our opinion, the most gender-neutral option from the collection, such that it would work for sequences originating from both genders. In addition, compared to other textures, the default texture has a rather cartoonish appearance which would minimize the chance of a uncanny valley related experience among the participants. Furhat was recorded with the OBS screencapture tool (version 27.0.1)[8]. The 14 recordings were then synchronized and merged with the sound of the original video with ShotCut (version 21.01.29)[9]. From here, we used the same method as in the human experiment. We cut out three videos per avatar video, each containing the original addressees behavior during exactly one BOP.

Figure 5.6 shows a few still images from Furhat as appearing in the stimuli.

Procedure

Participants again took part in the online experiment using the online environment of Qualtrics (Qualtrics, 2021). Before the start of the experiment, participants read the instructions, signed the consent form and familiarized themselves with the task with two practice videos. The practice video clips of Furhat were created in the same fashion as the stimuli for the experiment, but using a different BOP (BOP 21). Participants were asked to watch the video clips and indicate how they judged the personality of the original addressee in the same way as described for Experiment 1. The 42 video clips were shown in random order. On average it took 18 minutes and 46 seconds to complete the experiment (SD = 19 minutes and 34 seconds).

Statistical analyses

The results of experiment 2 are analyzed in the same way as experiment 1. However, as experiment 2 did not include stimuli to obtain a *preconception score*, the results contain only one model.

[7] https://www.furhat.io
[8] https://www.obsproject.com
[9] https://www.shotcut.org

5.4.2 Results

Friendliness

On average, Furhat received a score of 3.61 (SD=1.45) for the friendly-distant score. The transferred behavior of original addressee 10g was perceived as most friendly (mean = 2.75, SD = 1.33), and 14g as most distant (mean = 4.24, SD = 1.28). The model for the friendly-distant score (intercept: 3.986, SE: 0.212) produced significant effects for *amplitude* (*b*=-0.033, *SE*=0.007, *df*=1654.000, *t*=-5.064, *p* <0.001) and *au10* (*b*=-0.206, *SE*=0.080, *df*=2435.000, *t*=-2.578, *p* <0.05).

Activeness

The videos were rated, on average, with a score of 3.99 (SD = 1.37) on the active-passive scale. The behavior of original addressee 10g was perceived as most active (mean = 3.10, SD = 1.36), while that of original addressee 14g was perceived as most passive (mean = 4.57, SD = 1.15). The model for the active-passive score (intercept: 4.389, SE: 0.174) produced significant effects for *amplitude* (*b*=-0.037, *SE*=0.006, *df*=1096.000, *t*=-6.031, *p* <0.001) and *sound* (*b*=-0.404, *SE*=0.072, *df*=2395.000, *t*=-5.593, *p* <0.001).

Extroversion

The mean score for extroversion-introversion was 3.98 (SD = 1.29). The most extraverted behavior was that of original addressee 10g (mean = 3.30, SD = 1.32) and the behavior of original addressee 3g was perceived as most introverted (mean = 4.48, SD = 1.11). The model for the extroversion-introversion score (intercept: 4.315, SE: 0.136) produced significant effects for *amplitude* (*b*=-0.030, *SE*=0.006, *df*=530.100, *t*=-5.410, *p* <0.001) and *sound* (*b*=-0.495, *SE*=0.067, *df*=1355.000, *t*=-7.381, *p* <0.001).

Dominance

The mean perception for dominant-submissive was 3.87 (SD = 1.31). Original addressee 1g was perceived as most submissive (mean = 3.46, SD = 1.44) and that of original addressee 3g as most dominant (mean = 4.47, SD = 1.16). The model for the extroversion-introversion score (intercept: 3.990, SE: 0.124) produced significant effects for *amplitude* (*b*=-0.013, *SE*=0.006, *df*=290.697, *t*=-2.273, *p* <0.05), *sound* (*b*=-0.389, *SE*=0.069, *df*=906.322, *t*=-5.630, *p* <0.001).

5.4.3 Discussion

Our second experiment with judgments of avatars is in line with the results of our first experiment in which human beings were being rated, in the sense that variable feedback behaviors again led to differences in perceived personality of the avatars. However, we also noticed that the results of both experiments were slightly at variance regarding the significance and strength of specific auditory and visual cues. Although we have not included a statistical comparison between the models in this paper, we will discuss the differences between the models in more detail in the general discussion section.

5.5 General discussion and conclusion

We have reported about two perception experiments, both consisting of 42 8-second video clips that showed a human or artificial listening original addressee. Participants were asked to rate the perceived personality of that original addressee in terms of different dimensions. The first experiment also presented participants with still images of the original addressees in the clips, who were likewise rated regarding the different personality traits. In the first experiment, the video clips contained 14 different original addressees during 3 different backchannel opportunity points (the moments in conversation that allow for feedback). In the second experiment, the same stimuli were shown, except that they were re-enacted by a virtual Furhat robot. The results of the first experiment showed that backchannel behavior influences personality perception, which modulated the first impressions that people obtained from the still pictures. The results of the second experiment show that comparable effects could be achieved when such behavior is re-enacted by a conversational AI system. In the following, we first detail more specific resemblances and differences between the outcomes of the two experiments, and then discuss the outcomes in a broader perspective.

An overview of the significant results from the various models can be found in Table 5.1. While the results are quite analogous for ratings of human and artificial stimuli, we also observe some variability. Regarding the Friendly Distant dimension, we see that the model of for the human condition produced significant results for amplitude, sound and AU10, while the avatar model only did so for amplitude and AU10. Moreover, the estimate for AU10 for the human condition (-0.493) is more than double the estimate for the avatar condition (-0.206). For the Active - Passive dimension, the human model produced significant results for frequency, but the avatar model did not generate any significant results. For the Extroversion - Introversion dimension, both models produced significant effects for the variables amplitude and sound, even if sound had a higher estimate for the avatar condition (-0.495) than for the human condition (-0.266). And finally, when looking at the Dominant - Submissive dimension, we see that the human condition only had a significant effect for amplitude, while the avatar model also had a significant effect for sound as well (next to amplitude).

So while the results appear to be quite consistent over the two experiments, it may be worthwhile to reflect somewhat on the differences between conditions, especially regarding the variable effect of sound. First, it is important to note that we focused on the effect of four sets of features in both human and artificial stimuli on personality perception, namely sound, amplitude, frequency and AU10. But while only these features were varying in the avatar data, the human data also contained additional variation that we had not investigated further (e.g., other facial expressions, hand gestures and body posture) that nonetheless could have affected the perception results, if only because they made the human data more natural. In that sense, the conditions are not entirely comparable, as judgments of human data may be closer to what people do in their daily life than judgments of artificial creatures.

Also, note that the audio variable is different from the visual features in that this one was identical in both conditions, whereas the visual features were modelled via the avatar settings, and therefore only a computational approximation of the human data. Yet, despite the similarity regarding the audio feature, it is interesting to observe that this variable does not always have similar effects in the human and avatar data on personality judgments. For instance, the audio data increase the perception of friendliness when human original addressees are judged, but not when the avatar data are scored. Maybe this could be due to the fact that participants, when rating this dimension of friendliness in avatar stimuli, are unsure about their judgments. Indeed, of the 9 avatar videos that contained vocalisations, the friendliness perception scores of three of those videos are highly variant. Maybe this is due to the fact that judges have some difficulty to relate the natural voice with an artificial visual appearance of the avatar, so that they have problems taking the audio variable into account for this variable. Other factors may include the effects related to the mismatch between the human voice and the human-looking (but rather cartoonish) avatar, as non-human systems endowed with real human voice may lead to expectations mismatch [148]. Moreover, it may have been somewhat confusing for participants to note that, although the visual appearance of the avatar was the same in all stimuli, the choice of voices changed.

Conversely, we observe a significant effect of the audio variable on the judgments of dominance with avatars, while this effect is absent in the judgments of human data. According to the literature, the dominance-submission is, in general, perceived through multiple channels: Facial expressions related to anger and aggression are perceived as highly dominant, while fearful expressions are related to submission [92], direct eye contact and upward head tilt express dominance, contrary to downward head tilt and averted gaze which are perceived as submissive [147]. Voice frequency is related to dominance as well, as men rate male voices with a lower frequency as more dominant [173]. In that sense, the avatar data may have represented a relatively poor approximation of this dimension, as many variables mentioned above were not included in the stimuli, so that participants may have relied to a larger extent on the audio cue, compared with their judgments of the human data. Also, it is important to note that voice dominance is gender specific, as low voice has been shown to lead to perception of male voices only, whereas the Furhat character seems to look rather gender-neutral, which could have influenced perception as well. In a future study we could also include facial AUs related to expressions of fear and anger to see if those influenced the perception. How can the insight that personality is perceived through backchannel behavior be integrated into the behavior of an ECA? In case the backchannel behavior of the ECA is modeled based on human data, we recommend selecting the humans based on the desired personality of the ECA, rather than modeling the behavior based data originating from humans with random personalities. So e.g. if the desired personality of an ECA is introverted, utilize the data of people that are perceived as introverted. Additionally, in case of adjusting an existing backchannel generation algorithm, we would focus on increasing or decreasing the amplitude. As amplitude showed a significant effect for all four personality dimensions, we expect that magnifying the head movements of the avatar would lead to a more extravert, friendly, active and dominant perceived ECA, while reducing the head movements would lead to the opposite perception. The exact magnitudes to increase or decrease the amplitude is part of future research. Note that the non-BOPs are picked randomly. In future research, it would be interesting to pick less random non-BOPs, e.g. moments such as end of a sentence, or breath-based patterns that are recurring during the conversation, but do not co-incide with a BOP.

In conclusion, personality perception is indeed influenced by the behavior a person exhibits during backchannel opportunity moments. Especially the utilised amplitude for head nodding behavior correlates with multiple personality dimensions. These results suggest that it could be useful for conversational AI, and ECA developers in particular, to start implementing feedback behavior generation algorithms that take into account the reported variables (amplitude, frequency, sound and AU10) to strengthen the personality perception of their avatar in order to create more natural interactions and induce a stronger social presence with its interlocutors.

Acknowledgments

We thank Julija Vaitonyté for her valuable insights on the experimental design, Alex Blomsma for the visual impression of the experiment, and Martijn Goudbeek for his statistical insights.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

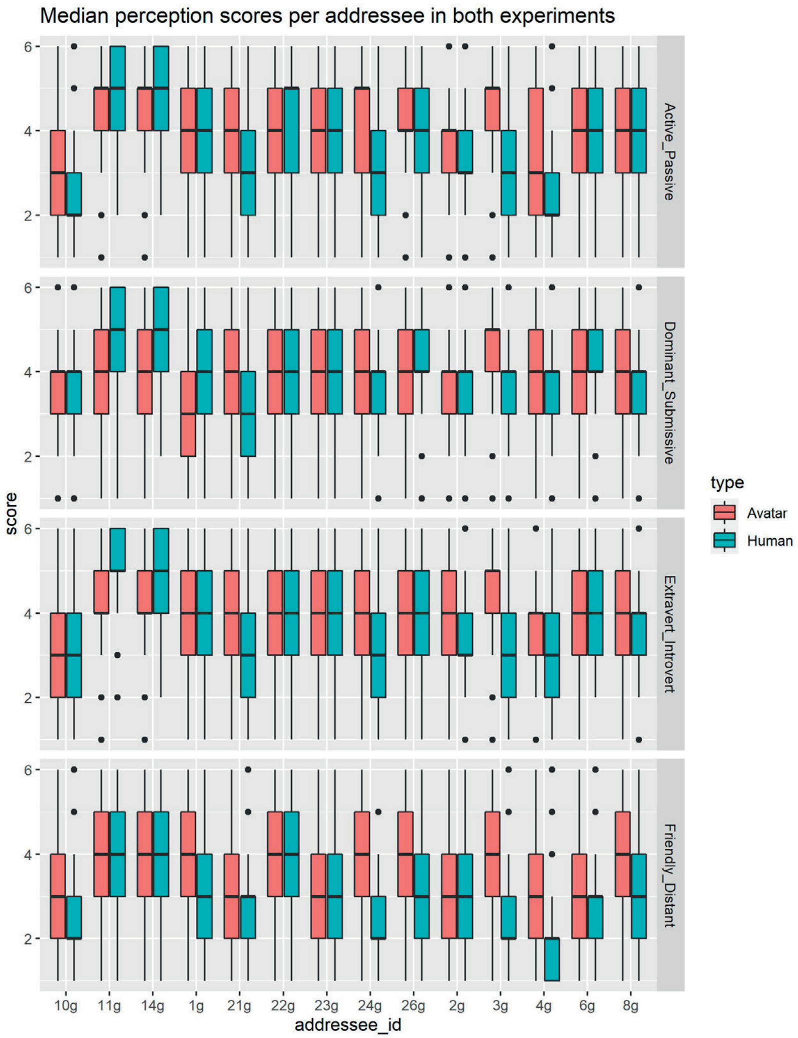


Figure 5.4 Overview of perception scores for all original addressees in experiment 1 and 2

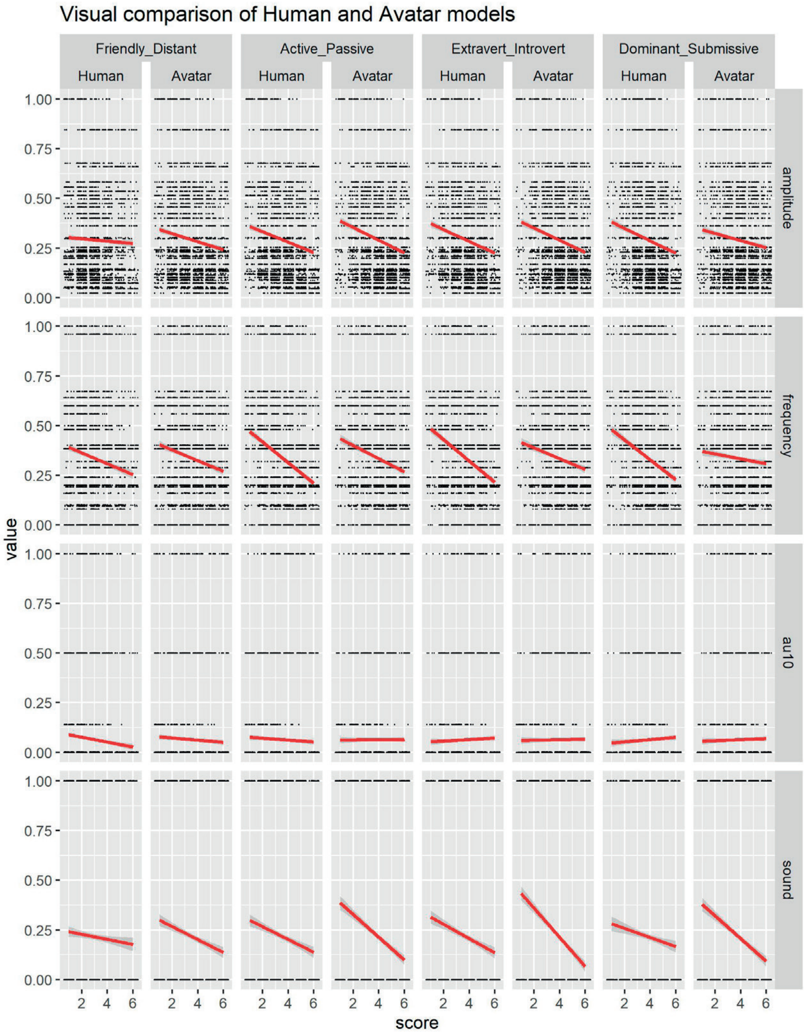


Figure 5.5 Visual indication of distributions for all variables available in Avatar model

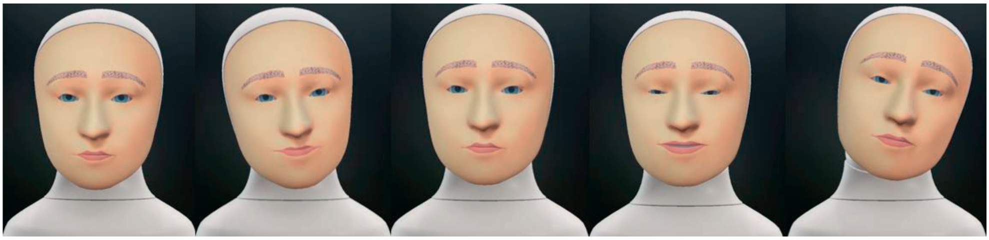


Figure 5.6 Visual impressions of the visual Furhat robot, as used during experiment 2

CHAPTER 6

General discussion and conclusion

The previous chapters presented 4 different studies that were all related to the same question: How can insights into the way human beings interact with each other inspire and eventually be utilized by developers to create Embodied Conversational Agents ('ECA') that act like real humans. Each study took a different perspective to this question. The studies focused on different aspects of people's communicative strategies and properties, with a main interest in their facial expressions, gestures and dialog acts, and the way people backchannel feedback. We examined how the complexity of human behavior could be simplified, could be transferred into the artificial intelligence of an ECA, and how ECAs simulating human-like behavior are perceived by observers. In this final chapter we will discuss the presented studies in a broader perspective, reflect on the theoretical implications of the findings, and give directions for future research. This chapter starts with a short summary of the different sub-studies.

6.1 Summary of studies

In the first study, reported in chapter 2, entitled "Spontaneous Facial Behavior Revolves Around Neutral Facial Displays", we investigated to what extent human behavior can be described by using only the most frequent behavioral patterns. Building an ECA capable of simulating all (theoretical) possible human behavior, including rarely occurring patterns, is valuable, but comes with a few drawbacks, including the time-consuming effort that is required to build such an ECA. We asked ourselves the question if we could simplify the functionalities of an ECA by only implementing the types of human behavior most frequently observed. This approach would not result in an ECA that is omnipotent, but could potentially result in an ECA that is able to behave natural in the most frequent situations, and can be created within a reasonable amount of time, as ECA developers could focus their attention and time to a limited set of patterns.

In our study we specifically looked at facial behavior. Facial behavior is traditionally expressed with the use of the facial action coding system ("FACS"). FACS quantifies facial behavior in terms of 46 different muscle(endings) that each can be contracted on a scale from 0 (relaxed) to 5 (fully contracted). Accordingly, 46^6 different facial configurations can be expressed within FACS, which implies an astronomically large number, equal to 6.2e+35. To identify the most frequently occurring facial configurations we analyzed three different FACS-encoded datasets. In particular, we used an extreme dataset (people that were recorded while their painful shoulder was moved), an emotional dataset (people that were recorded while watching an emotion eliciting movie) and a social dataset (people that were playing a maptask game). For each dataset, we identified a number of specific facial configurations that occurred and subsequently counted how often each facial configuration appeared in each dataset. The results showed that the most frequent facial configuration in each dataset was the neutral facial configuration, that is, a facial configuration where each encoded facial muscle is relaxed (has zero contraction). The most frequent facial configurations that did contain muscle contractions, included oftentimes only slight activations (1 or 2) of either the outer brow raiser, the lip corner puller or the brow lowerer. This work shows that a large percentage of the datasets can be described by only taking the most frequent facial configurations into account. Thus, it seems that ECA developers could in turn implement only a small set of facial behaviors to simulate the most frequent patterns.

Chapter 3 describes the second study called "Intrapersonal dependencies in multimodal behavior", which centers on the question if behavior generation for ECAs could be simplified by utilizing dependencies between different behavior channels. Just like the face can theoretically show a semi-infinite amount of facial expressions, other channels, such as language, tonality, and gestures, also allow for semi-infinite expressive possibilities. The combination of all channels and their expressive possibilities makes behavior generation a challenging task. Discovering dependencies between channels could help limiting the number of possibilities, and therefore help in simplifying behavior generation. In this study we focused on two channels: the hand movements of the speaker (the Gesture channel) and the intentions of the speaker (the Dialog Acts channel). The dataset utilized in this study contained 25 hours of encoded human dialog, and included encodings for Gestures and Dialog Acts. Gestures were encoded for 5 different gestures (i.e. beat, deictic, iconic, metaphoric and symbolic gestures) over 10 different variables. Dialog Acts were encoded for 13 different dialog acts (instruct, explain, check, align, reply-yes, reply-no, reply-what, acknowledgement, clarification, ready and unknown). For the analysis, a technique of cross recurrence quantification analysis ('CRQA') was used to identify possible intrapersonal dependencies. CRQA involves calculating how often two channels contain activity at the same time, or with a delay on one of the channels. This percentage of co-occurring activity is called recurrence rate. As such, CRQA is able to e.g. identify how often a beat gesture starts 500ms after an Acknowledgement dialog act started. We analyzed 506 combinations (23 times 22 channels) and found 130 significant results. Those results both showed a significant difference between a random recurrence rate and the actual recurrence rate, and four human raters classified them as having an effect. The results included both synchronized and mutual-exclusive relationships between different gestures and dialog acts. The results shown in this work, can aid ECA developers to build more accurate multimodal behavior generation systems. The last two studies concerned listening behavior and how such human listening behavior could be utilized by ECAs.

The study in chapter 4, called "Backchannel behavior is idiosyncratic", reported on the variability found in human backchannel behavior. Although people provide backchannels at specific moments during a conversation, how often a person utilizes such moment to provide feedback and in what way seems to be person-specific, e.g. differing between individuals who are quite expressive and others who reveal less behavior. However, not much research has been devoted to such idiosyncratic variability. Developers could utilize insights into backchannel variability to mimic natural feedback behavior into ECAs. Backchanneling behavior, includes head nods, 'hmms' and 'uhuhs' that conversational partners provide during an interaction. Listeners provide backchannels during backchannel opportunity points ("BOP"), moments that are signaled by the speaker with a backchannel-inviting cue to indicate that a listener may provide feedback. A backchannel-inviting cue can occur in several forms, e.g. through a lowering of the pitch for a short time or by making eye-contact with the listener. The research utilized a previously gathered dataset consisting of 14 video recordings of addressees who were listening to figure descriptions from a speaker. The recordings were made during an o-cam paradigm based experiment, where each participant (addressee) played a game via a skype-like setting with the speaker. With the help of scripted manipulations the participants were made to believe that they were involved in a live interaction with the speaker, while in reality the speaker was a pre-recorded stimulus. In our approach we first asked a jury of 10 judges to identify the BOPs in the speaker stimulus. We analyzed facial, nodding and vocal behavior of each addressee during each BOP. We found that the behavior during BOPs was different than in other (non-BOP) locations. During BOPs addressees tend to have a higher nodding frequency and amplitude. However, facial behavior was equivalent during and outside BOPs and did not seem to be related to backchanneling. In addition, behavior during BOPs differed significantly between addressees in terms of nodding frequency and amplitude, and vocal behavior. On average 28% of BOPs were vocalized, with the least vocal participant only vocalizing in 4% of the BOPs, and the most vocal participant vocalizing in 58% of the BOPs. At the same time, certain BOPs appeared to elicit on average different behavior from the addressees, than other BOPs. For example, on average the frequency was 3.42 upward/downward peaks per second, while certain BOPs had an average frequency of 1.10 upward/downward peaks per second. We found that especially BOPs at the end of a game round received more expressive behavior in terms of nodding amplitude and vocalization, than the other BOPs. This research suggests that in order to equip ECAs with natural, believable backchannel behavior, we should take into account that not all BOPs are equal. And that backchannel behavior is idiosyncratic.

Chapter 5 is built on the insights regarding variability that we observed and discussed in chapter 4. Impressionistically, the data suggested that backchanneling behaviors elicited different perceptions of personality, i.e. that the frequency and type of backchanneling had an effect on how a specific individual would be perceived in terms of hs/her personality. Thus, in the two experiments presented in this chapter, we have tested how backchannel behavior is correlated with personality perception. Multiple studies from the past show that in order to have a natural, believable ECA, that ECA should behave according to some consistent, congruent personality. Therefore, if backchanneling behavior would correlate with the perception of certain personality characteristics, insight into this relation would give ECA developers more control over the personality they want to implement in their ECAs. For this study, we have selected three different BOPs from the study described in chapter 4. Each BOP elicited a different type of behavior. We have shown short movie clips containing the behavior during the BOP to participants and asked how those participants to judge the personality of the person shown in the movie clip. The participants had to judge the personality on 4 different dimensions, being Friendly-Distant, Active-Passive, Extroversion-Introversion and Dominant-Submissive. In our analysis we have looked at correlations between their scores on a certain dimension and the observed nodding, vocal and facial behavior. The result was that indeed feedback behaviors significantly influence the perceived personality of participants, although the variables differed for each dimensions. In a follow-up experiment, the behavior of the stimuli was transferred onto an avatar - the same avatar for each stimulus - in order to make sure that the personality perceptions were not influenced by other factors such as a person's looks, gender or clothing. In the same fashion, the movies of the avatar were shown to participants who now had to judge the personality of the avatars. On a general level, the results of the human and the avatar experiment were equivalent in the sense that indeed, different feedback behaviors led to different personality perceptions.

6.2 Implications and Future work

While each chapter contains its own specific recommendations for future research, in this section we would like to reflect on the broader implications of our research. As mentioned in the introduction, multiple challenges prevent us from creating human-like ECAs, including (i) how to trick the sensitive perception of the human interlocutor and (ii) how to deal with the complexity of human behavior. The central theme in this dissertation is to the degree of variability in human non-verbal behavior. We would like to discuss future research by means of two sides of this variability: The first theme, the uniformity of human behavior, deals with the constraints that reduce the seemingly limitless number of possible behavioral combinations. The second theme, the variability in human behavior, reflects on the other side of the same coin, and focuses on the variability found within that constraint range of human behavior.

6.2.1 Uniformity of human behavior

The human body allows for many degrees of freedom. The set of possible human behaviors is nearly-infinite. The human face can theoretically show a semi-infinite number of facial configurations (chapter 2). Humans use multiple gestures in combinations with different dialog acts (chapter 3). And backchanneling behavior can be expressed via different channels, including nodding behavior and vocalizations (chapter 4). However, those three studies all reveal that humans in reality use only a small subset of the larger set of all theoretically possible behaviors. Indeed, chapter 2 shows that although the human face has many facial muscles which can be combined to make different facial configurations, most often humans show a neutral facial configuration. In the cases that the facial configurations were not neutral, most often only slight activations of one single facial muscle were shown. Chapter 3 shows a similar picture. The dialog acts and gestures of a speaker do not operate independently. The speaker's intentions and the shown gestures are inter-dependent. Some gestures are never shown with certain intentions, while other gestures are often shown with certain intentions. Chapter 4 shows, albeit that the main focus of the study was on the variations in listener feedback, that the listener's feedback behavior plays out within a certain range of behavioral possibilities. Indeed, the frequency of a listener's nodding behavior has a maximum, the amplitude has a maximum. In that sense, the results on non-verbal behavior are in line with what has been observed for other aspects of spoken communication. For instance, the repertoire of sounds that a language uses to form spoken words is incredibly limited compared to what speakers are able to produce with their vocal apparatus.

Distilling the minimum set of variables that are required to generate natural behavior for an ECA seems to be a fruitful way to face the complexity challenge. Therefore, future research could focus on further determining the demarcations of natural human behavior. What is the range of human behavior, and what are the edges of this range. Future research could be executed in line with chapter 2: one could look at facial behavior in other datasets and analyze what facial configurations are found in those datasets. Currently, manually-encoded FACS datasets containing spontaneous facial behavior are only sparsely available. However, automatic facial encoding software is getting more and more precise which could pave the way to encode large swaths of video recordings (e.g. from YouTube) containing spontaneous facial behavior.

This would make it possible to further define the range of facial configurations that is utilized, across cultures and contexts, which in return would enable ECA developers to even better define the degrees of freedom that are required in ECAs to show human natural facial behavior. Just as languages can differ in the settings and choice of phonemes, e.g. as evidenced by the fact that some phonological contrasts (e.g. the l-r difference) are not exploited by all languages across the globe, it could be interesting to use this approach to gain further insight into cultural and linguistic variability in non-verbal behavior. It could also be interesting to apply chapter 2's approach on other parts of behavior to identify the most occurring behaviors for other channels, such as gestures and head movements. For example with newly-developed software for semi-automatic gesture encoding which would help to analyze large numbers of recordings of spontaneous conversational behavior. Likewise, chapter 3's approach could also be utilized on other behavior channels. Cross-recurrence quantification analysis can also be exploited to further identify potential correlations between different modalities. Here again, automatic encoding of linguistic content (speech to text), facial behavior, head movements and gestures would potentially allow a large study into the intrapersonal dependencies of those modalities.

6.2.2 Variability in human behavior

Even when the previous subsection stressed the point that humans use only a limited set of theoretically possible non-verbal behaviors, the research in this dissertation has also shown that there may still be variability within these constraints, some of which could be due to person-dependent factors. Chapter 2 revealed that some facial configurations were only shown by one specific person. Chapter 4 and 5 presented results on how people vary in how they give feedback during conversations. While not studied in this dissertation, in addition to person-related factors, some variability is likely to be caused by circumstantial factors, such as stress, understanding and rapport-seeking strategies of dialogue partners.

What are the consequences of such idiosyncratic differences for developers that would like to create ECAs that approach the behavior of real humans? One option could be to only implement the most frequently occurring behaviors in the artificial characters, thereby ignoring the variability found in human data. To test to what extent this would be a valid choice, one could conduct a perception experiment (as a counterpart to what we did in chapter 5) where facial configurations are constrained by the variability as observed in the data of chapter 2. That would shed light on how an ECA is assessed when only the most frequently observed behaviors are implemented.

One of the challenges in creating lifelike ECAs is the sensitive perception of human beings able to spot tiny mistakes in an ECAs behavior, which could lead to uncanny valley feelings. As discussed in the previous section, utilizing only a small number of variables to generate human behavior would simplify behavior generation. However, would that small number of variables be enough to trick the sensitive perception of the onlooker? And would it be possible to express idiosyncratic behavior with such a small number of variables? Could we find a bliss point, a sweet-spot in the number of variables that is both small enough to face the complexity challenge, and large enough to give room for all expressions needed to make a natural, and believable ECAs able to cross the uncanny valley? Given the idiosyncrasy found in the multiple studies, we conjecture that ECA developers may be better off not focusing on implementing the average behavior of a large group of people as such implementation would wipe-out all idiosyncratic behaviors, or if not, would combine idiosyncratic behaviors of multiple people. The question is if such an amalgam of behavior is still perceived as natural and believable behavior by onlookers. Again, this should be researched. What we do think would work out is to model the behavior of one specific human, and transfer that behavior into an ECA. This would ensure that all idiosyncratic behaviors are instilled in the ECA. Related to this approach, it would be interesting to research facial behavior of one specific person for a long(er) time. Such approaches are already common in the realms of speech synthesis, where models are trained on large datasets containing the voice of only one person. Studying multiple hours of recordings of one person, or even equipping a person with a facial tracker (such as a indie headcam system) [10] for an extensive amount of time could result in a better, more natural, model to feed the ECA. Modeling one specific person would maybe simplify the task as person-dependent variables such as personality may be ignored, and don't have to be modelled into the ECA. However, person-related variability that may be caused by factors such as context and stress may have to be taken into account. An extra challenge of copying the exact visuals and behavior of one specific person into an ECA is that it will potentially be more difficult to trick the perception of onlooker that also know the human counterpart of an ECA, as they will probably not only detect flaws in the naturalness of the ECA, but also specific differences between the ECA and its human counterpart.

6.3 Conclusion

Human conversational behavior is complex, but far less complex than what a human body (with all its potential possibilities for movement and sound) could in principle show. Our dissertation has revealed that human beings tend to use only a subset of all possible behaviors. One possible implication of this outcome is that it facilitates the work of those who would like to generate human-like ECAs. Accordingly, in order to simplify behavior generation, ECA developers could limit the option space of an ECA to the limits of human conversational behavior, instead of the limits of the human body. Such approach would simplify behavior generation while still having the ability to generate natural-looking behavior. However, future research could investigate whether this would indeed be a fruitful alley. On the one hand, limiting the possible nonverbal behaviors of an ECA may make such a character appear very machine-like. So that it would not do justice to the variability that can still be observed in human data, that

[10] See https://facewaretech.com/cameras/indie-headcam/

reflect different communicative styles related to personality and other factors. It could be researched how such variability can be used in order to make ECAs become more human-like, even when that may be challenging in view of the uncanny valley problem. As an alternative, it could be interesting to explore under what circumstances an ECA may produce non-verbal features that humans do not normally display, but which are nonetheless functional, just as other machines have features that are not necessarily inspired by observations on humans.

Bibliography

[1] (2022). Soul machines ltd. https://www.soulmachines.com. Accessed: 2022-10-16.

[2] (2022). Uneeq. https://digitalhumans.com. Accessed: 2022-10-16.

[3] Adams, R. B., Ambady, N., Macrae, C. N., and Kleck, R. E. (2006). Emotional expressions forecast approach-avoidance behavior. *Motivation and emotion*, 30(2):177–186.

[4] Adolphs, R. (2002). Recognizing emotion from facial expressions: psychological and neurological mechanisms. *Behavioral and cognitive neuroscience reviews*, 1(1):21–62.

[5] Al Moubayed, S., Beskow, J., Skantze, G., and Granström, B. (2012). Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems*, pages 114–130. Springer, Berlin, Heidelberg.

[6] Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The hcrc map task corpus. *Language and speech*, 34(4):351–366.

[7] Andrist, S., Mutlu, B., and Tapus, A. (2015). Look like me: matching robot personality via gaze to increase motivation. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3603–3612, Seoul.

[8] Aneja, D., Hoegen, R., McDuff, D., and Czerwinski, M. (2021). Understanding conversational and expressive style in a multimodal embodied conversational agent. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–10.

[9] Audacity Team (2021). *Audacity 3.1.3*. Audacity Team.

[10] Austin, J. L. (1962). *How to do things with words*. Oxford University Press, Oxford.

[11] Ayedoun, E., Hayashi, Y., and Seta, K. (2019). Adding communicative and affective strategies to an embodied conversational agent to enhance second language learners' willingness to communicate. *International Journal of Artificial Intelligence in Education*, 29(1):29–57.

[12] Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., and Pollak, S. D. (2019). Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68.

[13] Bartlett, M. S., Hager, J. C., Ekman, P., and Sejnowski, T. J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, 36(2):253–263.

[14] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

[15] Bavelas, J. B., Chovil, N., Lawrie, D. A., and Wade, A. (1992). Interactive gestures. *Discourse processes*, 15(4):469–489.

[16] Bavelas, J. B., Coates, L., and Johnson, T. (2000). Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941.

[17] Bavelas, J. B., Coates, L., and Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, 52(3):566–580.

[18] Beattie, G. and Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? an experimental investigation. *Semiotica*, 123(1-2):1–30.

[19] Bergmann, K. and Kopp, S. (2009). Gnetic–using bayesian decision networks for iconic gesture generation. In *International Workshop on Intelligent Virtual Agents*, pages 76–89. Springer.

[20] Bernieri, F. J. and Rosenthal, R. (1991). Interpersonal coordination: Behavior matching and interactional synchrony. In *Fundamentals of nonverbal behavior*, page 401–432. Cambridge University Press.

[21] Bertrand, R., Ferré, G., Blache, P., Espesser, R., and Rauzy, S. (2007). Backchannels revisited from a multimodal perspective. In *Auditory-visual Speech Processing*, pages 1–5, Hilvarenbeek, Netherlands.

[22] Bethel, C. L., Stevenson, M. R., and Scassellati, B. (2011). Secret-sharing: Interactions between a child, robot, and adult. In *2011 IEEE International Conference on systems, man, and cybernetics*, pages 2489–2494. IEEE.

[23] Bevacqua, E., De Sevin, E., Hyniewska, S. J., and Pelachaud, C. (2012). A listener model: introducing personality traits. *Journal on Multimodal User Interfaces*, 6(1):27–38.

[24] Bickmore, T. and Cassell, J. (2001). Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 396–403.

[25] Biocca, F. (1999). The cyborg's dilemma: Progressive embodiment in virtual environments. *Human Factors in Information Technology*, 13:113–144.

[26] Birdwhistell, R. L. (1970). *Kinesics and Context: Essays on Body Motion Communication*. University of Pennsylvania Press.

[27] Blomsma, P., Vaitonyte, J., Skantze, G., and Swerts, M. (2022). Variability between and within addressees in how they produce audiovisual backchannels. [Manuscript submitted for publication].

[28] Blomsma, P. A. (2018). Eagle eye: a progress measure for intelligent tutoring systems. Master's thesis, Master's thesis, Utrecht University.

[29] Blomsma, P. A., Linders, G. M., Vaitonyte, J., and Louwerse, M. M. (2020a). Intrapersonal dependencies in multimodal behavior. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*.

[30] Blomsma, P. A., Vaitonyte, J., Alimardani, M., and Louwerse, M. M. (2020b). Spontaneous facial behavior revolves around neutral facial displays. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8.

[31] Boersma, P. and Weenink, D. (2022). *Praat: doing phonetics by computer (Version 6.2.10)*.

[32] Bortfeld, H. and Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23(2):119–147.

[33] Bronstein, I., Nelson, N., Livnat, Z., and Ben-Ari, R. (2012). Rapport in negotiation: The contribution of the verbal channel. *Journal of Conflict Resolution*, 56(6):1089–1115.

[34] Brugel, M. (2014). *Het effect van de eye gaze en lach van de spreker op het uitlokken van feedback bij de ontvanger*. PhD thesis, Master's thesis, Tilburg University.

[35] Burger, F., Broekens, J., and Neerincx, M. A. (2016). Fostering relatedness between children and virtual agents through reciprocal self-disclosure. In *Benelux conference on artificial intelligence*, pages 137–154. Springer.

[36] Cafaro, A., Ravenet, B., and Pelachaud, C. (2019). Exploiting evolutionary algorithms to model nonverbal reactions to conversational interruptions in user-agent interactions. *IEEE Transactions on Affective Computing*, pages 1–12.

[37] Calvo, M. G., Gutiérrez-García, A., Fernández-Martín, A., and Nummenmaa, L. (2014). Recognition of facial expressions of emotion is related to their frequency in everyday life. *Journal of Nonverbal Behavior*, 38(4):549–567.

[38] Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J. C., and Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational linguistics*, 23(1):13–31.

[39] Carr, E. W., Hofree, G., Sheldon, K., Saygin, A. P., and Winkielman, P. (2017). Is that a human? categorization (dis) fluency drives evaluations of agents ambiguous on human-likeness. *Journal of Experimental Psychology: Human Perception and Performance*, 43(4):651.

[40] Cartmill, E. A., Beilock, S., and Goldin-Meadow, S. (2012). A word in the hand: action, gesture and mental representation in humans and non-human primates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1585):129–143.

[41] Cassell, J. (2001). Embodied conversational agents: Representation and intelligence in user interfaces. *AI Magazine*, 22(4):67.

[42] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420.

[43] Cassell, J., Sullivan, J., Churchill, E., and Prevost, S. (2000). *Embodied conversational agents*. MIT press.

[44] Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2001). Beat: the behavior expression animation toolkit. In *Proceedings of SIGGRAPH '01: The 28th International Conference on Computer Graphics and Interactive Techniques*, pages 477—-486, New York, NY, United States. Association for Computing Machinery.

[45] Cathcart, N., Carletta, J., and Klein, E. (2003). A shallow model of backchannel continuers in spoken dialogue. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 51–58. Association for Computational Linguistics. 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL) 2003 ; Conference date: 12-04-2003 Through 17-04-2003.

[46] Cerekovic, A., Aran, O., and Gatica-Perez, D. (2016). Rapport with virtual agents: What do human social cues and personality explain? *IEEE Transactions on Affective Computing*, 8(3):382–395.

[47] Chiu, C.-C. and Marsella, S. (2011). How to train your avatar: A data driven approach to gesture generation. In Vilhjálmsson, H. H., Kopp, S., Marsella, S., and Thórisson, K. R., editors, *Intelligent Virtual Agents*, pages 127–140, Berlin, Heidelberg. Springer Berlin Heidelberg.

[48] Chiu, C.-C., Morency, L.-P., Marsella, Stacy", e. W.-P., Broekens, J., and Heylen, D. (2015). Predicting co-verbal gestures: A deep and temporal modeling approach. In *Intelligent Virtual Agents*, pages 152–166. Springer International Publishing.

[49] Chouchourelou, A., Matsuka, T., Harber, K., and Shiffrar, M. (2006). The visual analysis of emotional actions. *Social Neuroscience,*, 1(1):63–74.

[50] Clark, H. H. (1996). *Using language*. Cambridge university press, Cambridge.

[51] Coco, M. I. and Dale, R. (2014). Cross-recurrence quantification analysis of categorical and continuous time series: an r package. *Frontiers in psychology*, 5:510.

[52] Cowen, A. S., Keltner, D., Schroff, F., Jou, B., Adam, H., and Prasad, G. (2021). Sixteen facial expressions occur in similar contexts worldwide. *Nature*, 589(7841):251–257.

[53] de Haas, M., Vogt, P., van den Berghe, R., Leseman, P., Oudgenoeg-Paz, O., Willemsen, B., de Wit, J., and Krahmer, E. (2022). Engagement in longitudinal child-robot language learning interactions: Disentangling robot and task engagement. *International Journal of Child-Computer Interaction*, page 100501.

[54] de Kok, I. and Heylen, D. (2010). Differences in listener responses between procedural and narrative task. In *Proceedings of the 2nd international workshop on Social signal processing*, pages 5–10.

[55] de Kok, I. A. (2013). *Listening heads*. University of Twente.

[56] De Sevin, E., Hyniewska, S. J., and Pelachaud, C. (2010). Influence of personality traits on backchannel selection. In *International Conference on Intelligent Virtual Agents*, pages 187–193. Springer.

[57] de Wit, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., Krahmer, E., and Vogt, P. (2018). The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 50–58. IEEE.

[58] Di Dio, C., Manzi, F., Peretti, G., Cangelosi, A., Harris, P. L., Massaro, D., and Marchetti, A. (2020). Shall i trust you? from child–robot interaction to trusting relationships. *Frontiers in Psychology*, 11:469.

[59] Dimitrova, D., Chu, M., Wang, L., Özyürek, A., and Hagoort, P. (2016). Beat that word: How listeners integrate beat gesture and focus in multimodal speech discourse. *Journal of Cognitive Neuroscience*, 28(9):1255–1269.

[60] Dixon, J. A. and Foster, D. H. (1998). Gender, social context, and backchannel responses. *The Journal of social psychology*, 138(1):134–136.

[61] Drolet, A. L. and Morris, M. W. (2000). Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology*, 36(1):26–50.

[62] Duncan Jr, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in society*, pages 161–180.

[63] Ekman, P., Friesen, W., and Hager, J. C. (2002). *Facial Action Coding System, A Human Face*. Weidenfeld Nicolson, Salt Lake City, UT.

[64] Ekman, P. and Friesen, W. V. (1978). *Facial action coding system: Investigator's guide*. Consulting Psychologists Press.

[65] English, W., Gott, M., and Robinson, J. (2022). The meaning of rapport for patients, families, and healthcare professionals: a scoping review. *Patient Education and Counseling*, 105(1):2–14.

[66] Ennis, C., McDonnell, R., and O'Sullivan, C. (2010). Seeing is believing: body motion dominates in multisensory conversations. *ACM Transactions on Graphics (TOG)*, 29(4):1–9.

[67] Fabian Benitez-Quiroz, C., Srinivasan, R., and Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570.

[68] Fridlund, A. J. (1991). Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of personality and social psychology*, 60(2):229. [69] Fridlund, A. J. (2014). *Human facial expression: An evolutionary view*. Academic Press.

[70] Furo, H. (2000). Listening responses in japanese and us english: Gender and social interaction. In *Social and cognitive factors in second language acquisition: Selected proceedings of the 1999 Second Language Research Forum. Somerville, MA: Cascadilla*, pages 445–457.

[71] Ganel, T. and Goshen-Gottstein, Y. (2002). Perceptual integrality of sex and identity of faces: Further evidence for the single-route hypothesis. *Journal of Experimental Psychology: Human Perception and Performance*, 28(4):854.

[72] Garrido, P., Seron, F., Barrachina, J., and Martinez, F. (2017). Smart tourist information points by combining agents, semantics and ai techniques. *Computer Science and Information Systems*, 14:1–23.

[73] Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

[74] Geluykens, R. and Swerts, M. (1994). Prosodic cues to discourse boundaries in experimental dialouges. *Speech communication*, 15(1-2):69–77.

[75] Gendron, M., Roberson, D., van der Vyver, J. M., and Barrett, L. F. (2014). Cultural relativity in perceiving emotion from vocalizations. *Psychological science*, 25(4):911–920.

[76] George, N. and Conty, L. (2008). Facing the gaze of others. *Neurophysiologie Clinique/Clinical Neurophysiology*, 38(3):197–207.

[77] Gijssels, T., Casasanto, L. S., Jasmin, K., Hagoort, P., and Casasanto, D. (2016). Speech accommodation without priming: The case of pitch. *Discourse Processes*, 53(4):233–251.

[78] Goetz, J., Kiesler, S., and Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, pages 55–60. Ieee.

[79] Goodacre, R. and Zadro, L. (2010). O-cam: A new paradigm for investigating the effects of ostracism. *Behavior Research Methods*, 42(3):768–774.

[80] Granström, B., House, D., and Swerts, M. (2002). Multimodal feedback cues in human-machine interactions. In *Speech Prosody 2002, International Conference*.

[81] Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R. J., and Morency, L.-P. (2006). Virtual rapport. In *International Workshop on Intelligent Virtual Agents*, pages 14–27. Springer.

[82] Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. (2007). Creating rapport with virtual agents. In *International workshop on intelligent virtual agents*, pages 125–138. Springer.

[83] Gravano, A. and Hirschberg, J. (2009). Backchannel-inviting cues in task-oriented dialogue. In *Tenth Annual Conference of the International Speech Communication Association*, pages 1019–1022.

[84] Gravano, A. and Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

[85] Graziano, M. and Gullberg, M. (2018). When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in psychology*, 9:879.

[86] Gulz, A. (2004). Benefits of virtual characters in computer based learning environments: Claims and evidence. *International Journal of Artificial Intelligence in Education*, 14(3, 4):313–334.

[87] Haas, M. d., Baxter, P., de Jong, C., Krahmer, E., and Vogt, P. (2017). Exploring different types of feedback in preschooler and robot interaction. In *Proceedings of the companion of the 2017 acm/ieee international conference on human-robot interaction*, pages 127–128.

[88] Habets, B., Kita, S., Shao, Z., Özyurek, A., and Hagoort, P. (2011). The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of cognitive neuroscience*, 23(8):1845–1854.

[89] Haywood, S. L., Pickering, M. J., and Branigan, H. P. (2005). Do speakers avoid ambiguities during dialogue? *Psychological Science*, 16(5):362–366.

[90] Heldner, M., Hjalmarsson, A., and Edlund, J. (2013). Backchannel relevance spaces. In *Nordic Prosody XI, Tartu, Estonia, 15-17 August, 2012*, pages 137–146. Peter Lang Publishing Group.

[91] Henkemans, O. A. B., Bierman, B. P., Janssen, J., Looije, R., Neerincx, M. A., van Dooren, M. M., de Vries, J. L., van der Burg, G. J., and Huisman, S. D. (2017). Design and evaluation of a personal robot playing a self-management education game with children with diabetes type 1. *International Journal of Human-Computer Studies*, 106:63–76.

[92] Hess, U., Blairy, S., and Kleck, R. E. (2000). The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal behavior*, 24(4):265–283.

[93] Ho, C.-C., MacDorman, K. F., and Pramono, Z. D. (2008). Human emotion and the uncanny valley: a glm, mds, and isomap analysis of robot video ratings. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 169–176.

[94] Hobert, S. and Meyer von Wolff, R. (2019). Say hello to your new automated tutor–a structured literature review on pedagogical conversational agents. In *14th International Conference on Wirtschaftsinformatik, Siegen*. [95] Holler, J. and Beattie, G. (2003). Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener? *Gesture*, 3(2):127–154.

[96] Hömke, P., Holler, J., and Levinson, S. C. (2018). Eye blinks are perceived as communicative signals in human face-to-face interaction. *PloS one*, 13(12):e0208030.

[97] Hone, K. (2006). Empathic agents to reduce user frustration: The effects of varying agent characteristics. *Interacting with computers*, 18(2):227–245.

[98] Hong, A., Schaafsma, J., van der Wijst, P., and Plaat, A. (2014). Taking the lead: Gender, social context and preference to lead. In *European Conference on Management, Leadership & Governance*, pages 445–451. Academic Conferences International Limited.

[99] Huang, L. and Gratch, J. (2012). Crowdsourcing backchannel feedback: understanding the individual variability from the crowds. In *Feedback Behaviors in Dialog*, Portland, Oregon.

[100] Huang, L., Morency, L.-P., and Gratch, J. (2010). Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, AAMAS '10, page 1265–1272, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[101] Ishi, C., Mikata, R., and Ishiguro, H. (2018). Analysis of relations between hand gestures and dialogue act categories. In *Proceedings of the 9th International Conference on Speech Prosody 2018*, pages 473–477.

[102] Jack, R. E. and Schyns, P. G. (2015). The human face as a dynamic tool for social communication. *Current Biology*, 25(14):R621–R634.

[103] Jain, V., Leekha, M., Shah, R. R., and Shukla, J. (2021). Exploring semi-supervised learning for predicting listener backchannels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, New York, NY.

[104] Janowski, K., Ritschel, H., and André, E. (2022). Adaptive artificial personalities. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*, pages 155–194. Association for Computing Machinery.

[105] Jentsch, E. (1997). On the psychology of the uncanny (1906). *Angelaki: Journal of the Theoretical Humanities*, 2(1):7–16.

[106] John, O. P., Naumann, L. P., and Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. *Handbook of Personality: Theory and Research*, pages 114––158.

[107] Johnson, W. L. and Lester, J. C. (2016). Face-to-face interaction with pedagogical agents, twenty years later. *International Journal of Artificial intelligence in education*, 26(1):25–36.

[108] Kawahara, T., Yamaguchi, T., Inoue, K., Takanashi, K., and Ward, N. G. (2016). Prediction and generation of backchannel form for attentive listening systems. In *Interspeech*, pages 2890–2894.

[109] Kelly, S. D., Barr, D. J., Church, R. B., and Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of memory and Language*, 40(4):577–592.

[110] Kendon, A. (2017). Pragmatic functions of gestures: Some observations on the history of their study and their nature. *Gesture*, 16(2):157–175.

[111] Kita, S. (2000). *How representational gestures help speaking*, page 162–185. Language Culture and Cognition. Cambridge University Press.

[112] Kita, S. and Davies, T. S. (2009). Competing conceptual representations trigger co-speech representational gestures. *Language and Cognitive Processes*, 24(5):761–775.

[113] Knight, D. (2011). The future of multimodal corpora. *Revista Brasileira de Linguística Aplicada*, 11(2):391–415.

[114] Kopp, S., Tepper, P., and Cassell, J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 97–104.

[115] Krahmer, E., Swerts, M., Theune, M., and Weegels, M. (2002). The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates. *Speech communication*, 36(1-2):133–145.

[116] Krämer, N. C., Karacora, B., Lucas, G., Dehghani, M., Rüther, G., and Gratch, J. (2016). Closing the gender gap in stem with friendly male instructors? on the effects of rapport behavior and gender of a virtual agent in an instructional interaction. *Computers & Education*, 99:1–13.

[117] Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535.

[118] Krogsager, A., Segato, N., and Rehm, M. (2014). Backchannel head nods in danish first meeting encounters with a humanoid robot: The role of physical embodiment. In *International Conference on Human-Computer Interaction*, pages 651–662. Springer.

[119] Krumhuber, E. G., Skora, L., Küster, D., and Fou, L. (2017). A review of dynamic datasets for facial expression research. *Emotion Review*, 9(3):280–292.

[120] Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. (2017). lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82(1):1–26.

[121] Lee, J. and Marsella, S. (2006). Nonverbal behavior generator for embodied conversational agents. In *International Workshop on Intelligent Virtual Agents*, pages 243–255. Springer.

[122] Lee, K. M. and Nass, C. (2003). Designing social presence of social actors in human computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 289–296, Fort Lauderdale, Florida.

[123] Levitan, R., Gravano, A., and Hirschberg, J. (2011). Entrainment in speech preceding backchannels. In *ACL (Short Papers)*, pages 113–117.

[124] Lhommet, M. and Marsella, S. C. (2013). Gesture with meaning. In *International Workshop on Intelligent Virtual Agents*, pages 303–312. Springer.

[125] Ligthart, M., Neerincx, M., Hindriks, K. V., et al. (2019). Getting acquainted for a long-term child-robot interaction. In *International Conference on Social Robotics*, pages 423–433. Springer.

[126] Lisetti, C., Amini, R., and Yasavur, U. (2015). Now all together: overview of virtual health assistants emulating face-to-face health interview experience. *KI-Künstliche Intelligenz*, 29(2):161–172.

[127] Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., and Bartlett, M. (2011). The computer expression recognition toolbox (cert). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305. IEEE.

[128] Louwerse, M. M. and Bangerter, A. (2010). Effects of ambiguous gestures and language on the time course of reference resolution. *Cognitive Science*, 34(8):1517–1529.

[129] Louwerse, M. M., Dale, R., Bard, E. G., and Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive science*, 36(8):1404–1426.

[130] Lucas, G. M., Gratch, J., King, A., and Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100.

[131] Lucas, G. M., Rizzo, A., Gratch, J., Scherer, S., Stratou, G., Boberg, J., and Morency, L.-P. (2017). Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI*, 4:51.

[132] Lucey, P., Cohn, J. F., Prkachin, K. M., Solomon, P. E., and Matthews, I. (2011). Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Face and Gesture 2011*, pages 57–64.

[133] Lugrin, B., Pelachaud, C., and Traum, D., editors (2021). *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, volume 37. Association for Computing Machinery, New York, NY, USA, 1 edition.

[134] Lupyan, G. and Thompson-Schill, S. L. (2012). The evocative power of words: activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, 141(1):170.

[135] Luria, M., Reig, S., Tan, X. Z., Steinfeld, A., Forlizzi, J., and Zimmerman, J. (2019). Re-embodiment and co-embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, pages 633–644.

[136] Maltz, D. N. and Borker, R. A. (2018). A cultural approach to male-female miscommunication. In *The matrix of language*, pages 81–98. Routledge.

[137] Marwan, N., Romano, M. C., Thiel, M., and Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics reports*, 438(5-6):237–329.

[138] Massaro, D. W., Ouni, S., Cohen, M. M., and Clark, R. (2005). A multilingual embodied conversational agent. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 296b–296b. IEEE.

[139] Matsumoto, D. (1990). Cultural similarities and differences in display rules. *Motivation and emotion*, 14(3):195–214.

[140] Matsuyama, Y., Bhardwaj, A., Zhao, R., Romeo, O., Akoju, S., and Cassell, J. (2016). Socially-aware animated intelligent personal assistant agent. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 224–227.

[141] Mavadati, S. M., Mahoor, M. H., Bartlett, K., Trinh, P., and Cohn, J. F. (2013). Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160.

[142] McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., and Graesser, A. (2007). Facial features for affective state detection in learning environments. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, volume 29, pages 467–472.

[143] McNeill, D. (1985). So you think gestures are nonverbal? *Psychological review*, 92(3):350.

[144] McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

[145] McNeill, D. (2006). Gesture: a psycholinguistic approach. *The encyclopedia of language and linguistics*, pages 58–66.

[146] McTear, M. (2020). Conversational ai: Dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies*, 13(3):1–251.

[147] Mignault, A. and Chaudhuri, A. (2003). The many faces of a neutral face: Head tilt and perception of dominance and emotion. *Journal of nonverbal behavior*, 27(2):111–132.

[148] Moore, R. K. (2017). Appropriate voices for artefacts: some key insights. In *1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*, Skvöde, Sweden.

[149] Morency, L.-P., De Kok, I., and Gratch, J. (2008). Predicting listener backchannels: A probabilistic multimodal approach. In *International Workshop on Intelligent Virtual Agents*, pages 176–190. Springer.

[150] Mori, M. (1970). Bukimi no tani [the uncanny valley]. energy 7, 33–35. transl. kf macdorman and n. kageki 2012. *IEEE transactions on robotics and automation*, 19:98–100.

[151] Mower, E., Black, M. P., Flores, E., Williams, M., and Narayanan, S. (2011). Rachel: Design of an emotionally targeted interactive agent for children with autism. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE.

[152] Mui, P. H., Goudbeek, M. B., Roex, C., Spierts, W., and Swerts, M. G. (2018). Smile mimicry and emotional contagion in audio-visual computer-mediated communication. *Frontiers in psychology*, 9:2077.

[153] Natale, S. (2020). To believe in siri: A critical analysis of ai voice assistants. *Communicative Figurations Working Papers*, 32:1–17.

[154] Naumann, A., Hurtienne, J., Israel, J. H., Mohs, C., Kindsmüller, M. C., Meyer, H. A., and Hußlein, S. (2007). Intuitive use of user interfaces: defining a vague concept. In *International Conference on Engineering Psychology and Cognitive Ergonomics*, pages 128–136. Springer.

[155] Naumann, L. P., Vazire, S., Rentfrow, P. J., and Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and social psychology bulletin*, 35(12):1661–1671.

[156] Neff, M., Kipp, M., Albrecht, I., and Seidel, H.-P. (2008). Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)*, 27(1):1–24.

[157] Nguyen, T.-T., Sim, K., Kuen, A. T. Y., O'donnell, R. R., Lim, S. T., Wang, W., and Nguyen, H. D. (2021). Designing ai-based conversational agent for diabetes care in a multilingual context. *arXiv preprint arXiv:2105.09490*.

[158] Niewiadomski, R., Bevacqua, E., Mancini, M., and Pelachaud, C. (2009). Greta: an interactive expressive eca system. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 1399–1400. Citeseer.

[159] Noldus (2019). *FaceReader: Tool for automated analysis of facial expression: Version 8.0*. Noldus Information Technology, Wageningen, Netherlands.

[160] Oosterhof, N. N. and Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, 105(32):11087–11092.

[161] Otsuka, K. and Tsumori, M. (2020). Analyzing multifunctionality of head movements in face-to-face conversations using deep convolutional neural networks. *IEEE Access*, 8:217169–217195.

[162] Oviatt, S. (1997). Multimodal interactive maps: Designing for human performance. *Human–Computer Interaction*, 12(1-2):93–129.

[163] Oviatt, S. (2006). Human-centered design meets cognitive load theory: designing interfaces that help people think. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 871–880.

[164] Oviatt, S., Coulston, R., and Lunsford, R. (2004). When do we interact multimodally? cognitive load and multimodal communication patterns. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 129–136.

[165] Pally, R. (2001). A primary role for nonverbal communication in psychoanalysis. *Psychoanalytic Inquiry*, 21(1):71–93.

[166] Peeters, D. (2019). Virtual reality: A game-changing method for the language sciences. *Psychonomic bulletin & review*, 26(3):894–900.

[167] Pelachaud, C. (2009). Modelling multimodal expression of emotion in a virtual agent. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3539–3548.

[168] Pelachaud, C. (2017). Greta: A conversing socio-emotional agent. In *Proceedings of the 1st acm sigchi international workshop on investigating social interactions with artificial agents*, pages 9–10.

[169] Pérez, J., Cerezo, E., Gallardo, J., and Serón, F. J. (2018). Evaluating an eca with a cognitive-affective architecture. In *Proceedings of the XIX International Conference on Human Computer Interaction*, pages 1–8.

[170] Poppe, R., Truong, K. P., and Heylen, D. (2011). Backchannels: Quantity, type and timing matters. In *International Workshop on Intelligent Virtual Agents*, pages 228–239. Springer.

[171] Poppe, R., Truong, K. P., Reidsma, D., and Heylen, D. (2010). Backchannel strategies for artificial listeners. In *International Conference on Intelligent Virtual Agents*, pages 146–158. Springer.

[172] Porcheron, M., Fischer, J. E., Reeves, S., and Sharples, S. (2018). Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12.

[173] Puts, D. A., Gaulin, S. J., and Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and human behavior*, 27(4):283–296.

[174] R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

[175] Rehm, M. and André, E. (2008). From annotated multimodal corpora to simulated human-like behaviors. In *Modeling Communication with Robots and Virtual Humans*, pages 1–17. Springer.

[176] Riemer, K., Kay, J., et al. (2019). Mapping beyond the uncanny valley: A delphi study on aiding adoption of realistic digital faces. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

[177] Rieser, H. (2011). Gestures indicating dialogue structure. In *Proceedings of SEMdial 2011, 15th Workshop on the Semantics and Pragmatics of Dialogue*, pages 9–18.

[178] Russell, J. A. and Dols, J. M. F. (1997). *The psychology of facial expression*, volume 131. Cambridge university press Cambridge.

[179] Sadoughi, N. and Busso, C. (2019). Speech-driven animation with meaningful behaviors. *Speech Communication*, 110:90–100.

[180] Sauter, D. A., Eisner, F., Ekman, P., and Scott, S. K. (2015). Emotional vocalizations are recognized across cultures regardless of the valence of distractors. *Psychological science*, 26(3):354–356.

[181] Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., and Frith, C. (2012). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social cognitive and affective neuroscience*, 7(4):413–422.

[182] Schaller, M. and Park, J. H. (2011). The behavioral immune system (and why it matters). *Current directions in psychological science*, 20(2):99–103. [183] Scheib, J. E., Gangestad, S. W., and Thornhill, R. (1999). Facial attractiveness, symmetry and cues of good genes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1431):1913–1917.

[184] Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge University Press, Cambridge.

[185] Seeger, A.-M., Pfeiffer, J., and Heinzl, A. (2017). When do we need a human? anthropomorphic design and trustworthiness of conversational agents. In *Proceedings of the Sixteenth Annual Pre-ICIS Workshop on HCI Research in MIS, AISeL, Seoul, Korea*, volume 10.

[186] Seymour, M., Evans, C., and Libreri, K. (2017). Meet mike: epic avatars. In *ACM SIGGRAPH 2017 VR Village*, pages 1–2.

[187] Shimojima, A., Katagiri, Y., Koiso, H., and Swerts, M. (2002). Informational and dialogue-coordinating functions of prosodic features of japanese echoic responses. *Speech communication*, 36(1-2):113–132.

[188] Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

[189] Skantze, G. (2012). A testbed for examining the timing of feedback using a map task. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*.

[190] Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

[191] Stratou, G., Van Der Schalk, J., Hoegen, R., and Gratch, J. (2017). Refactoring facial expressions: An automatic analysis of natural occurring facial expressions in iterative social dilemma. In *2017 Seventh international conference on affective computing and intelligent interaction (ACII)*, pages 427–433. IEEE.

[192] Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215.

[193] Swerts, M. (2011). Correlates of social awareness in the visual prosody of growing children. *Laboratory Phonology*, 2(2):381–402.

[194] Swerts, M. and Krahmer, E. (2020). Visual Prosody Across Cultures. In *The Oxford Handbook of Language Prosody*, pages 477–485. Oxford University Press.

[195] Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., and Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *Journal of medical Internet research*, 22(3).

[196] Tanaka, H., Negoro, H., Iwasaka, H., and Nakamura, S. (2017). Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PloS one*, 12(8):e0182151.

[197] Tapus, A. and Mataric, M. J. (2008).*´User Personality Matching with a Hands-Off Robot for Post-stroke Rehabilitation Therapy*, pages 165–175. Springer Berlin Heidelberg, Berlin, Heidelberg.

[198] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2019). Face2face: Real-time face capture and reenactment of rgb videos. *Communications of the ACM*, 62(1):96–104.

[199] Tian, Y.-I., Kanade, T., and Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115.

[200] Tickle-Degnen, L. and Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological inquiry*, 1(4):285–293.

[201] Tinwell, A. and Grimshaw, M. (2009). Bridging the uncanny: an impossible traverse? In *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era*, pages 66–73.

[202] Vaitonyte, J., Blomsma, P. A., Alimardani, M., and Louwerse, M. M. (2019). Generating facial expression data: Computational and experimental evidence. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, IVA '19, page 94–96, New York, NY, USA. Association for Computing Machinery.

[203] Vaitonyte, J., Blomsma, P. A., Alimardani, M., and Louwerse, M. M.˙ (2021). Realism of the face lies in skin and eyes: Evidence from virtual and human agents. *Computers in Human Behavior Reports*, 3:100065.

[204] van der Struijk, S., Huang, H.-H., Mirzaei, M. S., and Nishida, T. (2018). Facsvatar: An open source modular framework for real-time facs based facial animation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 159–164.

[205] Vinciarelli, A., Chatziioannou, P., and Esposito, A. (2015). When the words are not everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls. *Frontiers in ICT*, 2:4.

[206] Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759.

[207] Wagner, P., Malisz, Z., and Kopp, S. (2014). Gesture and speech in interaction: An overview.

[208] Wanner, L., Klusch, M., Mavropoulos, A., Jamin, E., Marín Puchades, V., Casamayor, G., Cernockˇy, J., Davey, S., Domínguez, M., Egorova, E., et al.` (2021). Towards a versatile intelligent conversational agent as personal assistant for migrants. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 316–327. Springer.

[209] Ward, N. (1996). Using prosodic clues to decide when to produce back-channel utterances. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1728–1731. IEEE.

[210] Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.

[211] Wik, P. and Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. *Speech communication*, 51(10):1024–1037.

[212] Willems, R. M. and Hagoort, P. (2007). Neural evidence for the interplay between action, gesture and language: a review. *Brain Language*, 101:278–289.

[213] Williams, G. L., Wharton, T., and Jagoe, C. (2021). Mutual (mis) understanding: Reframing autistic pragmatic "impairments" using relevance theory. *Frontiers in Psychology*, 12:1277.

[214] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: A professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.

[215] Wlodarczak, M., Buschmeier, H., Malisz, Z., Kopp, S., and Wagner, P. (2012). Listener head gestures and verbal feedback expressions in a distraction task. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog, INTERSPEECH2012 Satellite Workshop*, pages 93–96.

[216] Yamada, Y., Kawabe, T., and Ihaya, K. (2013). Categorization difficulty is associated with negative evaluation in the "uncanny valley" phenomenon. *Japanese Psychological Research*, 55(1):20–32.

[217] Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting, 1970*, pages 567–578.

[218] Yoon, Y., Ko, W.-R., Jang, M., Lee, J., Kim, J., and Lee, G. (2019). Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4303–4309. IEEE.

[219] Zakharov, E., Shysheya, A., Burkov, E., and Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9459–9468.

[220] Zibrek, K., Kokkinara, E., and McDonnell, R. (2018). The effect of realistic appearance of virtual characters in immersive environments-does the character's personality play a role? *IEEE transactions on visualization and computer graphics*, 24(4):1681–1690.

[221] Zimmerer, C., Krop, P., Fischbach, M., and Latoschik, M. E. (2022). Reducing the cognitive load of playing a digital tabletop game with a multimodal interface. In *CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Summary

In the future, we might talk to computers like we talk to other people, using our voice and body language. Computers will represent themselves like a person, will be able to understand and use language, gestures, facial expressions, and tone of voice to communicate. Creating such a human-like representation of the computer is challenging because human communication is a speedy process, and thus computers do not have much time to process and react to the communication. Also, people don’t like it when human-like representations make mistakes because it can make the computer seem weird or creepy. This dissertation revolves around the theme of how insights in how human beings interact with each other can be utilized to create embodied conversational agents (ECA, i.e. a human-like representation of a computer) that interact like real humans. The studies focused on different aspects of people’s communicative strategies and properties, with a main interest in their facial expressions, gestures and dialog acts, and the way people backchannel feedback. We examined how the complexity of human behavior could be simplified, could be transferred into the artificial intelligence of an ECA, and how ECAs simulating human-like behavior are perceived by observers.

The first study entitled "Spontaneous Facial Behavior Revolves Around Neutral Facial Displays" investigated whether human behavior can be described using only the most frequent patterns, focusing on facial behavior. Based on the Facial Action Coding System (FACS) the most frequently occurring facial configurations in three different datasets were identified. We found that the most frequent facial configuration in each dataset was the neutral facial configuration, and the most frequent facial configurations that did contain muscle contractions included only slight activations of a few facial muscles. This suggests that ECA developers may implement only a small set of facial behaviors to simulate the most frequent patterns.

The second study, "Intrapersonal dependencies in multimodal behavior", explored whether the behavior generation for ECAs could be simplified by identifying dependencies between different behavior channels. The study focused on the dependecies between hand movements of the speaker (Gesture channel) and the intentions of the speaker (Dialog Acts channel), using cross recurrence quantification analysis to identify possible intrapersonal dependencies. The study found 130 significant results of synchronized and mutual-exclusive relationships between different gestures and dialog acts, which could aid ECA developers in building more accurate multimodal behavior generation systems.

The third study, “Variability between and within addressees in how they produce audiovisual backchannels” examines the variability found in human backchannel

behavior, which includes head nods, 'hmms,' and 'uhuhs' that conversational partners provide during an interaction. The study found that, although listeners all tend to backchannel during specific points during an interaction, how often and in what way a person utilizes backchannels is idiosyncratic, differing between individuals.

The fourth study, "Backchannel behavior influences personality perception", is built upon the third study. In this study participants rated the personality of humans and avatars that showed certain backchannel behavior. The study found that feedback behaviors significantly influence the perceived personality of participants. These results can be used by ECA developers to instill a stronger sense of personality into their ECAs.

This study found that people tend to use only a small part of all the possible ways they could communicate. This means that developers who want to create computer programs that act like humans can limit the program's abilities to match only what humans typically do, instead of trying to make the program do everything a human can physically do. This could make it easier to create more natural-looking computer programs. But more research is needed to see if this is a good idea.

Samenvatting

Het is te verwachten dat we in de toekomst met computers praten zoals we met mensen praten. Op een computerscherm zien we een als mens uitziende verschijning, een avatar, met bewegingen en stemgeluid gelijk aan die van een mens. De computer begrijpt dan niet alleen wat we zeggen, maar ook onze non-verbale communicatie, zoals intonatiepatronen, gezichtsuitdrukkingen en handgebaren.

Het is een uitdaging om zo'n menselijke avatar te bouwen, omdat menselijke communicatie snel en complex is. Voor een computer is het lastig om met dezelfde snelheid de boodschap te interpreteren en adequaat, op een menselijke wijze, te reageren. Daarnaast blijkt dat als een op een mens lijkende avatar zich niet helemaal menselijk gedraagt, de consequentie is dat het raar of griezelig op mensen over kan komen.

Het thema van dit proefschrift is hoe inzichten in menselijke communicatie kunnen worden gebruikt om menselijke avatars te bouwen. De vier studies in dit boek richten zich op verschillende aspecten van menselijke communicatie, met name op gezichtsuitdrukkingen, handgebaren en de wijze waarop mensen luisteren. We hebben onderzocht hoe de complexiteit van menselijk gedrag kan worden vereenvoudigd, hoe dit gedrag kan worden gebruikt in het maken van een als mens uitziende avatar, en welke effecten dit heeft op toeschouwers. In de eerste studie is gekeken naar de meest frequente gezichtsuitdrukkingen in drie verschillende datasets. De meest voorkomende gezichtsuitdrukking bleek een neutrale gezichtsuitdrukking te zijn, terwijl andere veel voorkomende gezichtsuitdrukkingen slechts enkele spieren in het gezicht gebruikten. Dit suggereert dat ontwikkelaars van avatars de complexiteit van gedragsgeneratie kunnen verkleinen door slechts een kleine set aan gezichtsuitdrukkingen te implementeren.

In de tweede studie hebben we gekeken of we gedragsgeneratie konden versimpelen door mogelijke afhankelijkheden in kaart te brengen tussen twee verschillende gedragskanalen, namelijk de intentie van een spreker (zogenaamde dialog acts) enerzijds en de handgebaren die de spreker gebruikt anderzijds. Hieruit is gebleken dat sommige handgebaren vaak samengaan met bepaalde dialog acts, maar ook dat bepaalde handgebaren nauwelijks samengaan met bepaalde dialog acts. Ontwikkelaars kunnen deze afhankelijkheden mogelijk gebruiken om sneller accuraat gedrag te genereren. In de derde studie hebben we de variabiliteit onderzocht in luistergedrag waaronder hoofdknikken en 'hmms' en 'uhuhs' die gesprekspartners tijdens een interactie geven. Uit deze studie bleek dat, hoewel luisteraars allemaal de neiging hebben om op specifieke momenten tijdens een interactie feedback te geven, de frequentie en manier van feedback geven persoonsafhankelijk is. In de vierde studie hebben we de resultaten van de derde studie verder onderzocht door proefpersonen te laten kijken naar zowel luisterende mensen als luisterende avatars, en beide te beoordelen op persoonlijkheid. Uit deze studie bleek dat het type feedbackgedrag de waargenomen persoonlijkheid aanzienlijk beïnvloedt. Ontwikkelaars kunnen met deze resultaten een sterker gevoel van persoonlijkheid meegeven aan hun avatar door passend luistergedrag te implementeren, wat hoogstwaarschijnlijk leidt tot een succesvollere interactie tussen de avatar.

De studies in dit proefschrift laten zien dat mensen slechts een klein deel gebruiken van alle mogelijke manieren waarop ze zouden kunnen communiceren. Dit suggereert dat ontwikkelaars van avatars gedragsgeneratie kunnen versimpelen door m.n. te focussen op frequent gedrag. Daarnaast zijn er juist individuele gedragsverschillen tussen mensen die door ontwikkelaar zouden kunnen worden ingezet om de avatar een sterkere persoonlijkheid te geven.

Acknowledgements

Zonder twijfel was mijn tijd als promovendus een van de meest dynamische en verrijkende fases in mijn leven. Het was een tijd gevuld met hoogte- en dieptepunten. Zonder de steun en het vertrouwen van bepaalde mensen was deze tijd onmogelijk geweest.

Marc, heel erg bedankt voor je bijzonder goede begeleiding! Onze wekelijkse meetings op maandag om 13.00 waren ongetwijfeld een van de hoogtepunten. Ik heb je creatieve inzichten en de tijd die je hebt genomen om mijn schrijfsels steeds weer van feedback te voorzien enorm gewaardeerd. Bedankt dat je me onder je vleugels hebt willen nemen.
Gabriel, thanks a lot for sharing your technical knowledge and expertise and the usage of the digital Furhat robot.
Jean, dank je wel voor je sturende inzichten en goede feedback op de eerste versie van deze dissertatie. Ik waardeer het enorm dat ik met zulke intelligente wetenschappers heb mogen samenwerken.

I want to express my deep appreciation to the members of the Commission, namely Prof. dr. E. André, Prof. dr. C. Pelachaud, Prof. dr. J. Beskow, and Prof. dr. E.J. Krahmer, for agreeing to review my dissertation, offering me constructive feedback, providing references and insights to improve my dissertation's quality, and for being present during the defense. I am extremely grateful for your support and assistance. Thank you so much!

I want to express my gratitude to all the reviewers who generously dedicated their time to evaluate the four articles in my dissertation. While some articles were accepted right away, others required a few revisions before being accepted. I truly appreciate the feedback I received from the reviewers as it has been instrumental in advancing my thought process and helping me grow as a scientist. Thank you so much!

Ik wil mijn vrienden en familie bedanken voor hun steun.Mark D. en Arthur, ik wil jullie graag bedanken voor de geweldige vriendschap die we hebben. De afgelopen jaren zijn bijzonder geweest en ik waardeer het enorm dat jullie er altijd waren om me te laten lachen, zelfs in de moeilijkste tijden. Ik ben dankbaar voor alle verhalen die we samen hebben gecreëerd en natuurlijk dat jullie er voor me zullen zijn tijdens mijn verdediging, voor het geval ik flauwval.

Anna, zonder jou was dit absoluut niet mogelijk geweest. Waar moet ik beginnen. Jij hebt me op zoveel vlakken ondersteund, daar ben ik je eeuwig dankbaar voor. Sommige van mijn beslissingen hadden een grote impact op ons leven, maar jij aarzelde nooit om me te helpen. Jouw scherpe en grondige feedback was van onschatbare waarde voor mij. Naast een geweldige moeder, bleek je ook een geweldig hoofd juridische zaken te zijn.

Aart, zonder jouw hulp, inzichten en motivatie had ik nooit dit dankwoord kunnen schrijven (want dan had ik mijn promotietraject voortijdig afgebroken). Zoals je weet, zie ik je als een van de onofficiële begeleiders van mijn promovendus traject. Dank voor je tomeloze energie, je voorbeeld om door te gaan, ook als het zwaar is. Ik vond het een eer om zoveel tijd met je door te mogen brengen, inzicht te krijgen in je gedachtenprocessen.

Maarten, dank je vriendschap, voor al je ideeën en het reviewen van meerdere teksten. Jan, dank voor de vele lunchwandelingen, alle koffie en je strakke drumritmes. Hoe mooi is het dat we - als buren op de Damstraat in Utrecht - elkaar in Soesterberg hebben leren kennen en dat we uiteindelijk zo samen het PhD leven op verschillende universiteiten met elkaar konden delen.

Ik wil mijn (schoon)familie bedanken. In het bijzonder wil ik mijn kinderen en mijn nichtjes bedanken. Uiteindelijk zijn jullie mijn grootste inspiratie. Ik wil mijn ouders bedanken. Jullie staan altijd voor me klaar, welke richtingen of keuzes ik ook maak. Dat waardeer ik enorm.
Mark, dank je wel voor je altijd oprechte interesse en het delen van de je waardevolle gedachten. Ate en Marianne, bedankt ook voor jullie steun aan mij en aan ons gezin. Brecht, dank voor het vele leuke samenwerken tijdens corona. Meïr dank voor het goede gesprek in de Ceuvel.
In het bijzonder wil ik mijn opa bedanken voor het voorbeeld dat hij heeft gegeven voor wat betreft het najagen van zijn passies en interesses. Ook ben ik dankbaar voor de "bullshit detector" die ik van hem heb geërfd.

I would like to thank all the members of the VIBE team. Max Louwerse, thank you for giving me the chance to start my PhD. Maryam thank you for your valuable input during the meeting. Katherine, thanks for your good humor, your creative freedom and your great taste in music, and for starting the reading club together. Hossein, thank you for your intriguing stories and for the initial spark that led to my first publication. Julija, thank you for your loyal friendship, the bunch of “gezelligheid”, and for the countless times you have reviewed my texts. Guido, Kiril and Anna van Limpt, thank you for your companionship during work and the many enjoyable "krentenbol" walks we took on campus and in the woods.

Nog speciale dank voor Emiel, Ruben en Saar. Heel erg bedankt voor de warme ontvangst bij DCC. Dank Anouck voor de interessante gesprekken over chatbots en het initiatief voor het gezamenlijke paper.

List of publications

- **Blomsma, P. A.,** Skantze, G., & Swerts, M. (2022). Backchannel Behavior Influences the Perceived Personality of Human and Artificial Communication Partners. Frontiers in Artificial Intelligence, 5.

- Braggaar, A., Tomas, F., **Blomsma, P.,** Hommes, S., Braun, N., Van Miltenburg, E., ... & Krahmer, E. (2022, July). A reproduction study of methods for evaluating dialogue system output: Replicating Santhanam and Shaikh (2019). In Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges (pp. 86-93).

- Vaitonyte, J., **Blomsma, P. A.,** Alimardani, M., & Louwerse, M. M. (2021). Realism of the face lies in skin and eyes: Evidence from virtual and human agents. Computers in Human Behavior Reports, 3, 100065.

- **Blomsma, P. A.,** Vaitonyte, J., Alimardani, M., & Louwerse, M. M. (2020, October). Spontaneous Facial Behavior Revolves Around Neutral Facial Displays. In Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (pp. 1-8).

- **Blomsma, P. A.,** Linders, G. M., Vaitonyte, J., & Louwerse, M. M. (2020, October). Intrapersonal dependencies in multimodal behavior. In Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (pp. 1-8).

- Vaitonyte, J., **Blomsma, P. A.,** Alimardani, M., & Louwerse, M. M. (2019, July). Generating Facial Expression Data: Computational and Experimental Evidence. In Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (pp. 94-96).

SIKS dissertation series

2016

1. Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
2. Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
3. Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
4. Laurens Rietveld (VU), Publishing and Consuming Linked Data
5. Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
6. Michel Wilson (TUD), Robust scheduling in an uncertain environment
7. Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
8. Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
9. Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10. George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11. Anne Schuth (UVA), Search Engines that Learn from Their Users
12. Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13. Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14. Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15. Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16. Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17. Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18. Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19. Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20. Daan Odijk (UVA), Context & Semantics in News & Web Search
21. Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22. Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23. Fei Cai (UVA), Query Auto Completion in Information Retrieval
24. Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25. Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26. Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27. Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28. Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29. Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30. Ruud Mattheij (UvT), The Eyes Have It
31. Mohammad Khelghati (UT), Deep web content monitoring
32. Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33. Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34. Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35. Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36. Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37. Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38. Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39. Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40. Christian Detweiler (TUD), Accounting for Values in Design
41. Thomas King (TUD), Governing Governance: A formal framework for analysing institutional design and enactment governance
42. Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43. Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44. Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45. Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46. Jorge Gallego Perez (UT), Robots to Make you Happy
47. Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks

48. Tanja Buttler (TUD), Collecting Lessons Learned
49. Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50. Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017

1. Jan-Jaap Oerlemans (UL), Investigating Cybercrime
2. Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
3. Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
4. Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
5. Mahdieh Shadi (UVA), Collaboration Behavior
6. Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
7. Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
8. Rob Konijn (VU), Detecting interesting differences:data mining in health insurance data using outlier detection and subgroup discovery
9. Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10. Robby van Delden (UT), (Steering) Interactive Play Behavior
11. Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12. Sander Leemans (TUE), Robust Process Mining with Guarantees
13. Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14. Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15. Peter Berck (RUN), Memory-Based Text Correction
16. Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17. Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18. Ridho Reinanda (UVA), Entity Associations for Search
19. Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20. Mohammadbashir Sedighi (TUD), Fostering engagement in knowledge sharing: the role of perceived benefits, costs and visibility
21. Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22. Sara Magliacane (VU), Logics for causal inference under uncertainty

23. David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24. Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25. Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26. Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27. Michiel Joosse (UT), Investigating positioning and gaze behaviors of social robots: people's preferences, perceptions and behaviors
28. John Klein (VU), Architecture Practices for Complex Contexts
29. Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30. Wilma Latuny (UvT), The Power of Facial Expressions
31. Ben Ruijl (UL), Advances in computational methods for QFT calculations
32. Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33. Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34. Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35. Martine de Vos (VU), Interpreting natural science spreadsheets
36. Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37. Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38. Alex Kayal (TUD), Normative Social Applications
39. Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40. Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41. Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42. Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43. Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44. Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45. Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46. Jan Schneider (OU), Sensor-based Learning Support
47. Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48. Angel Suarez (OU), Collaborative inquiry-based learning

2018

1. Han van der Aa (VUA), Comparing and Aligning Process Representations
2. Felix Mannhardt (TUE), Multi-perspective Process Mining
3. Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
4. Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
5. Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
6. Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
7. Jieting Luo (UU), A formal account of opportunism in multi-agent systems
8. Rick Smetsers (RUN), Advances in Model Learning for Software Systems
9. Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10. Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11. Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
12. Xixi Lu (TUE), Using behavioral context in process mining
13. Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14. Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
15. Naser Davarzani (UM), Biomarker discovery in heart failure
16. Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17. Jianpeng Zhang (TUE), On Graph Sample Clustering
18. Henriette Nakad (UL), De Notaris en Private Rechtspraak
19. Minh Duc Pham (VUA), Emergent relational schemas for RDF
20. Manxia Liu (RUN), Time and Bayesian Networks
21. Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
22. Eric Fernandes de Mello Araujo (VUA), Contagious: modeling the spread of behaviours, perceptions and emotions in social networks
23. Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24. Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis
25. Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
26. Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
27. Roelof Anne Jelle de Vries (UT), Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

28. Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29. Yu Gu (UVT), Emotion Recognition from Mandarin Speech
30. Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019

1. Rob van Eijk (UL), Comparing and Aligning Process Representations
2. Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
3. Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
4. Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
5. Sebastiaan van Zelst (TUE), Process Mining with Streaming Data
6. Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
7. Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
8. Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
9. Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems
10. Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11. Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12. Jacqueline Heinerman (VU), Better Together
13. Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14. Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15. Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16. Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17. Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18. Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19. Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20. Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21. Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22. Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture

23. Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24. Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25. Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description
26. Prince Singh (UT), An Integration Platform for Synchromodal Transport
27. Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses
28. Esther Kuinderman (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29. Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30. Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31. Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics
32. Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33. Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34. Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and
35. Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming
36. Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills
37. Jian Fang (TUD), Database Acceleration on FPGAs
38. Akos Kadar (OUN), Learning visually grounded and multilingual representations

2020

1. Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
2. Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
3. Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
4. Maarten van Gompel (RUN), Context as Linguistic Bridges
5. Yulong Pei (TUE), On local and global structure mining
6. Preethu Rose Anish (UT), Stimulation architectural thinking during requirements elicitation - an approach and tool support
7. Wim van der Vegt (OUN), Towards a software architecture for reusable game components
8. Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
9. Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10. Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining
11. Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
12. Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13. Marco Virgolin (CWI), Design and application of gene-pool optimal mixing evolutionary algorithms for genetic programming
14. Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15. Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16. Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17. Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18. Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19. Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20. Albert Hankel (VU), Embedding Green ICT Maturity in Organisations
21. Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be
22. Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
23. Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24. Lenin da Nobrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25. Xin Du (TUE), The Uncertainty in Exceptional Model Mining
26. Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27. Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
28. Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29. Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30. Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst
31. Gongjin Lan (VU), Learning better – From Baby to Better
32. Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising
33. Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation
34. Anna Bon (MU), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

35. Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021

1. Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
2. Rijk Mercuur (TUD), Simulating Human Routines:Integrating Social Practice Theory in Agent-Based Models
3. Seyyed Hadi Hashemi (UVA), Modeling Users Interacting with Smart Devices
4. Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
5. Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
6. Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
7. Armel Lefebvre (UU), Research data management for open science
8. Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
9. Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
10. Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11. Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12. Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13. Bob R. Schadenberg (UT), Robots for autistic children: understanding and facilitating predictability for engagement in learning
14. Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15. Onat Ege Adali (TU/e), Transformation of value propositions into resource re-configurations through the business services paradigm
16. Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
17. Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
18. Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
19. Roberto Verdecchia (VU), Architectural Technical Debt: Identification and Management
20. Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
21. Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
22. Sihang Qiu (TUD), Conversational Crowdsourcing
23. Hugo Manuel Proen (LIACS), Robust rules for prediction and description
24. Kaijie Zhu (TUE), On Efficient Temporal Subgraph Query Processing

25. Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26. Benno Kruit (CWI & VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27. Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28. Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022

1. Judith van Stegeren (UT), Flavor text generation for role-playing video games
2. Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
3. Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
4. Unal Aksu (UU), A Cross-Organizational Process Mining Framework
5. Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
6. Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
7. Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
8. Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
9. Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10. Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11. Mirjam de Haas (UvT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
12. Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13. Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
14. Michiel Overeem (UU), Evolution of Low-Code Platforms
15. Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
16. Pieter Gijsbers (TU/e), Systems for AutoML Research
17. Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
18. Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
19. Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation

20. Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
21. Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
22. Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
23. Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
24. Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
25. Anna L.D. Latour (LU), Optimal decision-making under constraints and uncertainty
26. Anne Dirkson (LU), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
27. Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
28. Onuralp Ulusoy (UU), Privacy in Collaborative Systems
29. Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
30. Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
31. Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
32. Cezara Pastrav (UU), Social simulation for socio-ecological systems
33. Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
34. Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
35. Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

2023

1. Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
2. Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
3. Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
4. Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
5. Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
6. António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment

7. Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
8. Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
9. Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
10. Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11. Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12. Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
13. Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
14. Selma Čauševiˇc (TUD), Energy resilience through self-organization´
15. Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
16. Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters

TiCC PhD Series

1. Pashiera Barkhuysen. Audiovisual prosody in interaction. Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 3 October 2008.
2. Ben Torben-Nielsen. Dendritic morphology: Function shapes structure. Promotores: H.J. van den Herik, E.O. Postma. Co-promotor: K.P. Tuyls. Tilburg, 3 December 2008.
3. Hans Stol. A framework for evidence-based policy making using IT. Promotor: H.J. van den Herik. Tilburg, 21 January 2009.
4. Jeroen Geertzen. Dialogue act recognition and prediction: Explorations in computational dialogue modelling. Promotor: H. Bunt. Co-promotor: J.M.B. Terken. Tilburg, 11 February 2009.
5. Sander Canisius. Structured prediction for natural language processing: A constrained satisfaction approach. Promotores: A.P.J. van den Bosch, W. Daelemans. Tilburg, 13 February 2009.
6. Fritz Reul. New architectures in computer chess. Promotor: H.J. van den Herik. Co-promotor: J.W.H.M. Uiterwijk. Tilburg, 17 June 2009.
7. Laurens van der Maaten. Feature extraction from visual data. Promotores: E.O. Postma, H.J. van den Herik. Co-promotor: A.G. Lange. Tilburg, 23 June 2009 (cum laude).
8. Stephan Raaijmakers. Multinomial language learning: Investigations into the geometry of language. Promotores: W. Daelemans, A.P.J. van den Bosch. Tilburg, 1 December 2009.
9. Igor Berezhnoy. Digital analysis of paintings. Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 7 December 2009.
10. Toine Bogers. Recommender systems for social bookmarking. Promotor: A.P.J. van den Bosch. Tilburg, 8 December 2009.
11. Sander Bakkes. Rapid adaptation of video game AI. Promotor: H.J. van den Herik. Co-promotor: P. Spronck. Tilburg, 3 March 2010.
12. Maria Mos. Complex lexical items. Promotor: A.P.J. van den Bosch. Co-promotores: A. Vermeer, A. Backus. Tilburg, 12 May 2010 (in collaboration with the Department of Language and Culture Studies).
13. Marieke van Erp. Accessing natural history: Discoveries in data cleaning, structuring, and retrieval. Promotor: A.P.J. van den Bosch. Co-promotor: P.K. Lendvai. Tilburg, 30 June 2010.
14. Edwin Commandeur. Implicit causality and implicit consequentiality in language comprehension. Promotores: L.G.M. Noordman, W. Vonk. Co-promotor: R. Cozijn. Tilburg, 30 June 2010.
15. Bart Bogaert. Cloud content contention. Promotores: H.J. van den Herik, E.O. Postma. Tilburg, 30 March 2011.

16. Xiaoyu Mao. Airport under control: Multiagent scheduling for airport ground handling. Promotores: H.J. van den Herik, E.O. Postma. Co-promotores: N. Roos, A. Salden. Tilburg, 25 May 2011.
17. Olga Petukhova. Multidimensional dialogue modelling. Promotor: H. Bunt. Tilburg, 1 September 2011.
18. Lisette Mol. Language in the hands. Promotores: E.J. Krahmer, A.A. Maes, M.G.J. Swerts. Tilburg, 7 November 2011 (cum laude).
19. Herman Stehouwer. Statistical Language Models for Alternative Sequence Selection. Promotores: A.P.J. van den Bosch, H.J. van den Herik. Co-promotor: M.M. van Zaanen. Tilburg, 7 December 2011.
20. Terry Kakeeto-Aelen. Relationship Marketing for SMEs in Uganda. Promotores: J. Chr. van Dalen, H.J. van den Herik. Co-promotor: B.A. Van de Walle. Tilburg, 1 February 2012.
21. Suleman Shahid. Fun & Face: Exploring Non-verbal Expressions of Emotion during playful Interactions. Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 25 May 2012.
22. Thijs Vis. Intelligence, Politie en Veiligheidsdienst: Verenigbare grootheden? Promotores: T.A. de Roos, H.J. van den Herik, A.C.M. Spapens. Tilburg, 6 June 2012 (in collaboration with the Tilburg School of Law).
23. Nancy Pascall. Engendering Technology Empowering Women. Promotores: H.J. van den Herik, M. Diocaretz. Tilburg, 19 November 2012.
24. Agus Gunawan. Information Access for SMEs in Indonesia. Promotor: H.J. van den Herik. Co-promotores: M. Wahdan, B.A. Van de Walle. Tilburg, 19 December 2012.
25. Giel van Lankveld. Quantifying Individual Player Differences. Promotores: H.J. van den Herik, A.R. Arntz. Co-promotor: P. Spronck. Tilburg, 27 February 2013.
26. Sander Wubben. Text-to-text Generation by Monolingual Machine Translation. Promotores: E.J. Krahmer, A.P.J. van den Bosch, H. Bunt. Tilburg, 5 June 2013.
27. Jeroen Janssens. Outlier Selection and One-class Classification. Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 11 June 2013.
28. Martijn Balsters. Expression and Perception of Emotions: The case of depression, sadness and fear. Promotores: E.J. Krahmer, M.G.J. Swerts, A.J.J.M. Vingerhoets. Tilburg, 25 June 2013.
29. Lisanne van Weelden. Metaphor in Good Shape. Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 28 June 2013.
30. Ruud Koolen. Need I say more? On Overspecification in Definite Reference. Promotores: E.J. Krahmer and M.G.J. Swerts. Tilburg, 20 September 2013.
31. J. Douglas Mastin. Exploring Infant Engagement, Language Socialization and Vocabulary Development: A Study of Rural and Urban Communities in Mozambique. Promotor: A.A. Maes. Co-promotor: P.A. Vogt. Tilburg, 11 October 2013.

32. Philip C. Jackson. Jr. Toward Human-level Artificial Intelligence – Representation and Computation of Meaning in Natural Language. Promotores: H.C. Bunt, W.P.M. Daelemans. Tilburg, 22 April 2014.
33. Jorrig Vogels. Referential Choices in Language Production: The Role of Accessibility. Promotores: A.A. Maes, E.J. Krahmer. Tilburg, 23 April 2014 (cum laude).
34. Peter de Kock. Anticipating Criminal Behaviour. Promotores: H.J. van den Herik, J.C. Scholtes. Co-promotor: P. Spronck. Tilburg, 10 September 2014.
35. Constantijn Kaland. Prosodic Marking of Semantic Contrasts: Do Speakers adapt to Addressees? Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 1 October 2014.
36. Jasmina Maric. Web Communities, Immigration and Social Capital.´ Promotor: H.J. van den Herik. Co-promotores: R. Cozijn, M. Spotti. Tilburg, 18 November 2014.
37. Pauline Meesters. Intelligent Blauw. Promotores: H.J. van den Herik, T.A. de Roos. Tilburg, 1 December 2014.
38. Mandy Visser. Better use your Head: How People learn to signal Emotions in Social Contexts. Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 10 June 2015.
39. Sterling Hutchinson. How Symbolic and Embodied Representations work in Concert. Promotores: M.M. Louwerse, E.O. Postma. Tilburg, 30 June 2015.
40. Marieke Hoetjes. Talking hands: Reference in speech, gesture and sign. Promotores: E.J. Krahmer and M.G.J. Swerts. Tilburg, 7 October 2015
41. Elisabeth Lubinga. Stop HIV/AIDS. Start Talking? The Effects of Rhetorical figures in Health Messages on Interpersonal Discussions among South African Adolescents. Promotores: A.A. Maes, C.J.M. Jansen. Tilburg, 16 October 2015.
42. Janet Bagorogoza. Knowledge Management and high Performance: The Uganda financial Institutions Models for HPO. Promotor: H.J. van den Herik. Co-promotores: A.A. de Waal, B.A. Van de Walle. Tilburg, 24 November 2015.
43. Hans Westerbeek. Visual Realism: Exploring Effects on memory, Language Production, Comprehension, and Preference. Promotores: A.A. Maes, M.G.J. Swerts. Co-promotor: M.A.A. van Amelsvoort. Tilburg, 10 February 2016.
44. Matje van de Camp. A link to the past: Constructing historical social networks from unstructured data. Promotores: A.P.J. van den Bosch, E.O. Postma. Tilburg, 2 March 2016.
45. Annemarie Quispel. Data for all: How professionals and non-professionals in design use and evaluate information visualizations. Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 15 June 2016.
46. Rick Tillman. Language Matters: The Influence of Language and Language use on Cognition. Promotores: M.M. Louwerse, E.O. Postma. Tilburg, 30 June 2016.
47. Ruud Mattheij. The Eyes have it. Promoteres: E.O. Postma, H. J. Van den Herik, P.H.M. Spronck. Tilburg, 5 October 2016.
48. Marten Pijl. Tracking of Human Motion over Time. Promotores: E. H. L. Aarts, M. M. Louwerse. Co-promotor: J. H. M. Korst. Tilburg, 14 December 2016.

49. Yevgen Matusevych. Learning Constructions from Bilingual Exposure: Computational Studies of Argument Structure Acquisition. Promotor: A.M. Backus. Co-promotor: A. Alishahi. Tilburg, 19 December 2016.
50. Karin van Nispen. What can People with Aphasia communicate with their Hands? A Study of Representation Techniques in Pantomime and Co-speech Gesture. Promotor: E.J. Krahmer. Co-promotor: M. van de Sandt-Koenderman. Tilburg, 19 December 2016.
51. Adriana Baltaretu. Speaking of Landmarks: How Visual Information influences Reference in Spatial Domains. Promotores: A.A. Maes, E.J. Krahmer. Tilburg, 22 December 2016.
52. Mohamed Abbadi. Casanova 2: A Domain Specific Language for General Game Development. Promotores: A.A. Maes, P.H.M. Spronck, A. Cortesi. Co-promotor: G. Maggiore. Tilburg, 10 March 2017.
53. Shoshannah Tekofsky. You are Who you Play you are: Modelling Player traits from Video Game Behavior. Promotores: E.O. Postma, P.H.M. Spronck. Tilburg, 19 June 2017.
54. Adel Alhuraibi, From IT-business Strategic Alignment to Performance: A moderated Mediation Model of Social Innovation, and Enterprise Governance of IT. Promotores: H.J. van den Herik, B.A. Van de Walle. Co-promotor: S. Ankolekar. Tilburg, 26 September 2017.
55. Wilma Latuny. The Power of Facial Expressions. Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 29 September 2017.
56. Sylvia Huwaë. Different Cultures, different Selves? Suppression of Emotions and Reactions to Transgressions across Cultures. Promotores: E.J. Krahmer, J. Schaafsma. Tilburg, 11 October 2017.
57. Mariana Serras Pereira, A Multimodal Approach to Children's deceptive Behavior. Promotor: M.G.J. Swerts. Co-promotor: S. Shahid. Tilburg, 10 January 2018.
58. Emmelyn Croes. Meeting Face-to-Face online: The Effects of Video-mediated Communication on Relationship Formation. Promotores: E.J. Krahmer, M. Antheunis. Co-promotor A.P. Schouten. Tilburg, 28 March 2018.
59. Lieke van Maastricht. Second Language Prosody: Intonation and Rhythm in Production and Perception. Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 9 May 2018.
60. Nanne van Noord. Learning Visual Representations of Style. Promotores: E.O. Postma, M. Louwerse. Tilburg, 16 May 2018.
61. Ingrid Masson Carro. Handmade: On the Cognitive origins of Gestural Representations. Promotor: E.J. Krahmer. Co-promotor: M.B. Goudbeek. Tilburg, 25 June 2018.
62. Bart Joosten. Detecting Social Signals with Spatiotemporal Gabor Filters. Promotores: E.J. Krahmer, E.O. Postma. Tilburg, 29 June 2018.

63. Yan Gu. Chinese Hands of Time: The Effects of Language and Culture on Temporal Gestures and Spatio-temporal Reasoning. Promotor: M.G.J. Swerts. Co-promotores: M.W. Hoetjes, R. Cozijn.Tilburg, 5 June 2018.
64. Thiago Castro Ferreira. Advances in Natural Language Generation: Generating varied Outputs from Semantic Inputs. Promotor: E.J. Krahmer. Co-promotor: S. Wubben. Tilburg, 19 September 2018.
65. Yu Gu. Automatic Emotion Recognition from Mandarin Speech. Promotores: E.O. Postma, H.J. van den Herik, H.X. Lin. Tilburg, 28 November 2018.
66. Francesco Di Giacomo, Metacasanova: A High-performance Meta-compiler for Domain-specific Languages. Promotores: P.H.M. Spronck, A. Cortesi, E.O. Postma, Tilburg, 19 November 2018.
67. Ákos Kádár. Learning Visually grounded and Multilingual Representations. Promotores: E.O. Postma, A. Alishahi. Co-promotor: G.A. Chrupala. Tilburg, 13 November 2019.
68. Phoebe Mui. The Many Faces of Smiling: Social and Cultural factors in the Display and Perception of Smiles. Promotor: M.G.J. Swerts. Co-promotor: M.B. Goudbeek. Tilburg, 18 December 2019.
69. Véronique Verhagen. Illuminating Variation: Individual Differences in Entrenchment of Multi-words Units. Promotor: A.M. Backus. Co-promotores: M.B.J. Mos, J. Schilperoord. Tilburg, 10 January 2020.
70. Debby Damen. Taking Perspective in Communication: Exploring what it takes to change Perspectives. Promotor: E.J. Krahmer. Co-promotores: M.A.A. van Amelvoort, P.J. van der Wijst. Tilburg, 4 November 2020.
71. Alain Hong. Women in the lead: Gender, Leadership Emergence, and Negotiation Behavior from a Social Role Perspective. Promotor: J. Schaafsma. Co-promotor: P.J. van der Wijst. Tilburg, 3 June 2020.
72. Chrissy Cook. Online gaming and trolling, Promotores: J. Schaafsma, M.Antheunis. Tilburg, 22 January 2021.
73. Nadine Braun, Affective Words and the Company They Keep: Investigating the interplay of emotion and language. Promotor: E.J.Krahmer. Co-promotor: M.B. Goudbeek. Tilburg, 29 March 2021.
74. Yuegiao Han. Chinese Tones: Can You Listen with Your Eyes? The Influence of Visual Information on Auditory perception of Chinese Tones. Promotor: M.G.J. Swerts. Co-promotores: M.B.J. Mos, M.B. Goudbeek.Tilburg, 18 June 2021.
75. Tess van der Zanden. Language Use and Impression Formation: The Effects of Linguistic Cues in Online Dating Profiles. Promotor: E.J. Krahmer. Co-promotores: M.B.J. Mos, A.P. Schouten. Tilburg, 22 October 2021.
76. Janneke van der Loo. Mastering the Art of Academic Writing: Comparing the Effectiveness of Observational Learning and Learning by Doing. Promotor: E.J. Krahmer. Co-promotor: M.A.A. van Amelsvoort. Tilburg, 1 December 2021.

77. Charlotte Out. Does Emotion shape Language? Studies on the Influence of Affective State on Interactive Language Production. Promotor: E.J. Krahmer. Co-promotor: M.B. Goudbeek. Tilburg, 16 December 2021.
78. Jan de Wit. Robots that Gesture, and their Potential as Second Language Tutors for Children. Promotor: E.J. Krahmer. Co-promotor: P.A. Vogt.Tilburg, 28 January 2022.
79. Ruben Vromans. Communicating personalized risks to patients with cancer. Promotores: E.J. Krahmer, L.V. van de Poll-Franse, S.C. Pauws.Tilburg, 8 July 2022.
80. Chris van der Lee. Next Steps in Data-to-Text Generation: Towards Better Data, Models, and Evaluation. Promotor: E.J. Krahmer. Co-promotor: S. Wubben. Tilburg, 25 November 2022.
81. Annemarie Nanne. Social Drivers of Visual Brand-Related User Generated Content: Creation, Content, and Consequences. Promotores: M.L. Antheunis, G. van Noort, E.O. Postma. Tilburg, 2 December 2022.
82. Yan Xia. Why? Because... Socio-psychological and syntactic variables affecting causal attribution in interpersonal verbs. Promotor: A.A. Maes.Co-promotor: R. Cozijn. Tilburg, 10 March, 2023.
83. Peter Blomsma. Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters. Promotores: M.G.J. Swerts, J.H.M. Vroomen, G. Skantze. Tilburg, 20 June, 2023.